Introductory Statistics for the Life and Biomedical Sciences

Derivative of
OpenIntro Statistics
Third Edition

Original Authors

David M Diez
Christopher D Barr
Mine Çetinkaya-Rundel

Contributing Authors

David Harrington
[Briefly Describe Contribution]

Julie Vu
[Briefly Describe Contribution]

Alice Zhao
[Briefly Describe Contribution]

Copyright © 2015. Third Edition.

This textbook is available under a Creative Commons license. Visit openintro.org for a free PDF, to download the textbook's source files, or for more information about the license.

Contents

Preface

This book provides an introduction to statistics and its applications in the life sciences, and biomedical research. It is based on the freely available *OpenIntro Statistics, Third Edition*, and, like *OpenIntro* it may be downloaded as a free PDF at **Need location**. The text adds substantial new material, revises or eliminates sections from *OpenIntro*, and re-uses some material directly. Readers need not have read *OpenIntro*, since this book is intended to be used independently. We have retained some of the exercises from *OpenIntro* that may not come directly from medicine or the life sciences but illustrate important ideas or methods that are commonly used in fields such as biology.

Introduction to Statistics for the Life and Biomedical Sciences is intended for graduate and undergraduate students interested in careers in biology or medicine, and may also be profitably read by students of public health. It covers many of the traditional introductory topics in statistics used in those fields, but also adds some newer methods being used in molecular biology. Statistics has become an integral part of research in medicine and biology, and the tools for displaying, summarizing and drawing inferences from data are essential both for understanding the outcomes of studies and for incorporating measures of uncertainty into that understanding. An introductory text in statistics for students considering careers in medicine, public health or the life sciences should be more than the usual introduction with more examples from biology or medical science. Along with the value of careful, robust analyses of experimental and observational data, it should convey some of the excitement of discovery that emerges from the interplay of science with data collection and analysis. We hope we have conveyed some of that excitement here.

We have tried to balance the sometimes competing demands of mastering the important technical aspects of methods of analysis with gaining an understanding of important concepts. The examples and exercises include opportunities for students to build skills in conducting data analyses and to state conclusions with clear, direct language that is specific to the context of a problem. We also believe that computing is an essential part of statistics, just as mathematics was when computing was more difficult or expensive. The text includes many examples where software is used to aid in the understanding of the features of a data as well as exercises where computing is used to help illustrate the notions of randomness and variability. Because they are freely available, we use the R statistical language with the *R Studio* interface. Information on downloading R and *R Studio* is may be found in the Labs at **openintro.org**. Nearly all examples and exercises can be adapted to either SAS, Stata or other software, but we have not done that.

Textbook overview

The chapters of this book are as follows:

1. **Introduction to data.** Data structures, variables, summaries, graphics, and basic data collection techniques.
2. **Probability (special topic).** The basic principles of probability. An understanding of this chapter is not required for the main content in Chapters **??**-**??**.
3. **Distributions of random variables.** Introduction to the normal model and other key distributions.
4. **Foundations for inference.** General ideas for statistical inference in the context of estimating the population mean.
5. **Inference for numerical data.** Inference for one or two sample means using the normal model and t distribution, and also comparisons of many means using ANOVA.
6. **Inference for categorical data.** Inference for proportions using the normal and chi-square distributions, as well as simulation and randomization techniques.
7. **Introduction to linear regression.** An introduction to regression with two variables. Most of this chapter could be covered after Chapter **??**.

8. Multiple and logistic regression. An introduction to multiple regression and logistic regression for an accelerated course.

The remainder of this section requires revision

OpenIntro Statistics was written to allow flexibility in choosing and ordering course topics. The material is divided into two pieces: main text and special topics. The main text has been structured to bring statistical inference and modeling closer to the front of a course. Special topics, labeled in the table of contents and in section titles, may be added to a course as they arise naturally in the curriculum.

Examples, exercises, and appendices

Examples and within-chapter exercises throughout the textbook may be identified by their distinctive bullets:

● **Example 0.1** Large filled bullets signal the start of an example.

Full solutions to examples are provided and often include an accompanying table or figure.

⊙ **Guided Practice 0.2** Large empty bullets signal to readers that an exercise has been inserted into the text for additional practice and guidance. Students may find it useful to fill in the bullet after understanding or successfully completing the exercise. Solutions are provided for all within-chapter exercises in footnotes.[1]

There are exercises at the end of each chapter that are useful for practice or homework assignments. Many of these questions have multiple parts, and odd-numbered questions include solutions in Appendix **??**.

Probability tables for the normal, t, and chi-square distributions are in Appendix **??**, and PDF copies of these tables are also available from **openintro.org** for anyone to download, print, share, or modify.

[1] Full solutions are located down here in the footnote!

OpenIntro, online resources, and getting involved

OpenIntro is an organization focused on developing free and affordable education materials. *OpenIntro Statistics*, our first project, is intended for introductory statistics courses at the high school through university levels.

We encourage anyone learning or teaching statistics to visit **openintro.org** and get involved. We also provide many free online resources, including free course software. Data sets for this textbook are available on the website and through a companion R package.[2] All of these resources are free, and we want to be clear that anyone is welcome to use these online tools and resources with or without this textbook as a companion.

We value your feedback. If there is a particular component of the project you especially like or think needs improvement, we want to hear from you. You may find our contact information on the title page of this book or on the About section of **openintro.org**.

Acknowledgements

This project would not be possible without the dedication and volunteer hours of all those involved. No one has received any monetary compensation from this project, and we hope you will join us in extending a *thank you* to all those volunteers below.

The authors would like to thank Andrew Bray, Meenal Patel, Yongtao Guan, Filipp Brunshteyn, Rob Gould, and Chris Pope for their involvement and contributions. We are also very grateful to Dalene Stangl, Dave Harrington, Jan de Leeuw, Kevin Rader, and Philippe Rigollet for providing us with valuable feedback.

[2] Diez DM, Barr CD, Çetinkaya-Rundel M. 2012. `openintro`: OpenIntro data sets and supplement functions. http://cran.r-project.org/web/packages/openintro.

Chapter 2

Probability

What are the chances that a woman with an abnormal mammogram has breast cancer? What is the likelihood that an overweight male teenager with high blood pressure will develop cardiovascular disease by the age of 50? What is the probability that two parents who are unaffected carriers of a genetic mutation that causes cystic fibrosis will have a child that suffers from the disease. All of these questions use the language of probability, and despite how easy it is to ask these questions, answers are not always easy to come by. Probability also forms the foundation for data analysis and statistical inference, since nearly every conclusion to a study should be accompanied by a measure of uncertainty. In the publication of LEAP study discussed in Chapter 1, the manuscript included the probability that the results of the study could have been due simply to chance variation (a very small probability, as will be seen later in the text).

Like all mathematical tools, probability becomes easier to understand and work with when the important concepts and language have been formalized. With the right tools, seemingly difficult problems can be solved in a series of reliable, reproducible steps. This chapter introduces that formalization, using two types of examples. One set of examples uses familiar terms using settings most people have seen before – the outcomes of rolling dice or picking cards from a deck of playing cards. The second type of examples are drawn from medicine, biology or public health, and reflect the context and language used in those fields. The approaches to solving both types of problems are surprisingly similar,

once the problem has been posed clearly.

2.1 Defining probability

2.1.1 Some examples

Some of these dice examples can be dropped, but leaving them for now in case they are reference later.

We begin with some familiar examples.

Example 2.1 A "die", the singular of dice, is a cube with six faces numbered 1, 2, 3, 4, 5, and 6. What is the chance of getting 1 when rolling a die?

If the die is fair, then the chance of a 1 is as good as the chance of any other number. Since there are six outcomes, the chance must be 1-in-6 or, equivalently, 1/6.

Example 2.2 What is the chance of getting a 1 or 2 in the next roll?

---

1 and 2 constitute two of the six equally likely possible outcomes, so the chance of getting one of these two outcomes must be $2/6 = 1/3$.

Example 2.3 What is the chance of getting either 1, 2, 3, 4, 5, or 6 on the next roll?

100%. The outcome must be one of these numbers.

Example 2.4 What is the chance of not rolling a 2?

Since the chance of rolling a 2 is 1/6 or $16.\bar{6}\%$, the chance of not rolling a 2 must be $100\% - 16.\bar{6}\% = 83.\bar{3}\%$ or 5/6.

Alternatively, we could have noticed that not rolling a 2 is the same as getting a 1, 3, 4, 5, or 6, which makes up five of the six equally likely outcomes and has probability 5/6.

Example 2.5 Consider rolling two dice. If $1/6^{th}$ of the time the first die is a 1 and $1/6^{th}$ of those times the second die is a 1, what is the chance of getting two 1s?

If $16.\bar{6}\%$ of the time the first die is a 1 and $1/6^{th}$ of *those* times the second die is also a 1, then the chance that both dice are 1 is $(1/6) \times (1/6)$ or $1/36$.

Here is an example from genetics.

Example 2.6 Cystic fibrosis (CF) is a life-threatening genetic disorder characterized by the buildup of thick mucus in the lungs and pancreas, caused by mutations in the *CFTR* gene located on chromosome 7. Defective copies of *CFTR* can result in the reduced quantity and/or function of the CFTR protein, which transports sodium and chloride across cell membranes. CF is an autosomal recessive disorder – an individual only develops CF if they have inherited two affected copies of *CFTR*. Individuals with one normal (wild-type) copy and one defective (mutated) copy are known as carriers; they do not develop CF, but may pass the disease-causing mutation onto their offspring.

Suppose that both members of a couple are CF carriers. What is the probability that a child of this couple will be affected by CF? The problem sounds a bit more complicated than calculating probabilities for the outcome of rolling a die, but can be solved with the same simple methods. We show two solutions.

Solution 1: Enumerate all of the possible outcomes and exploit the fact that the outcomes are equally likely, as in **??**. During reproduction, each parent passes along one copy of the *CFTR* gene, with each copy passed along with probability 1/2. Figure 2.1 shows the four possible genotypes for a child of these parents, with the paternal chromosome in blue, the maternal chromosome in green, chromosomes with the wild-type and mutated version of CFTR marked with + and−. Each of the four outcomes (wild-type CFTR, wild-type CFTR), (wild-type CFTR, CFTR mutation) (CFTR mutation, wild-type CFTR) and (CFTR mutation, CFTR mutation), so the child will be affected with probability 1/4. It is important to recognize that the child being an unaffected carrier consists of two distinct outcomes, not one.

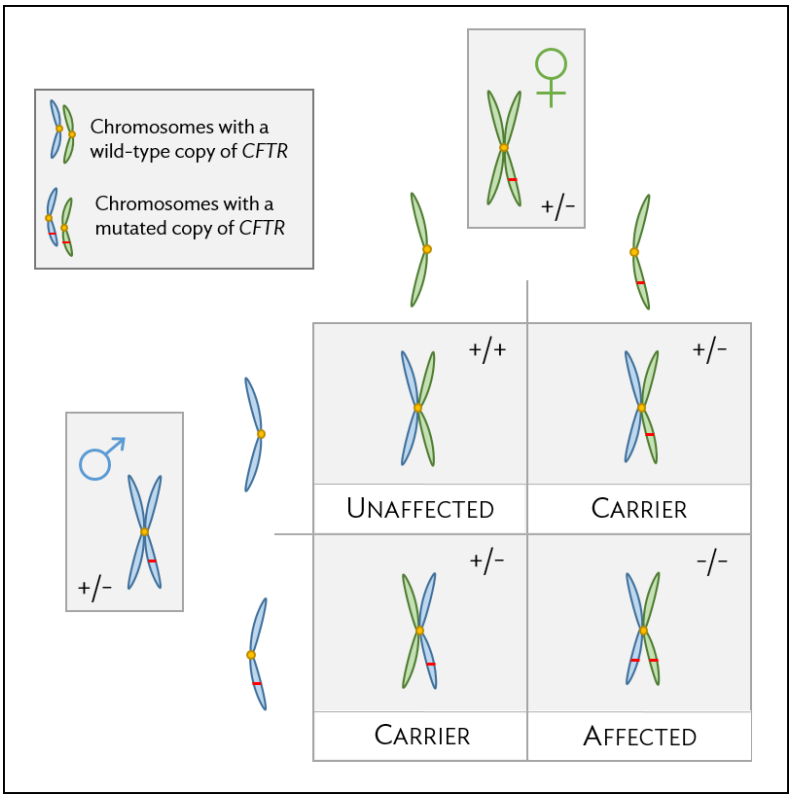


Figure 2.1: Pattern of inheritance of a child of two unaffected carriers of CFTR

Solution 2: Calculate the proportion of outcomes that produce an affected child, as in 2.1. During reproduction, half of the time, the mother will pass along an affected gene. When the child receives an affected gene from the mother, about half of those times, the father will have passed along an affected gene. So the proportion of times the child will be affected is $(1/2) \times (1/2) = 1/4$.

Guided Practice 2.7 Suppose the father is affected by CF and the mother is an unaffected carrier. What is the probability that their child will be affected by the disease?

Solution: Since the father is affected, he will always pass along a defective copy of the gene. Since the mother will pass along a defective copy half of the time, the child will be affected half of the time, or with probability 1/4.

2.1.2 Probability

Probability is used to assign a level of uncertainty to outcomes of phenomena that are happen randomly (rolling dice, passing along a defective gene during reproduction), or appear random because of a lack of understanding about exactly how the phenomenon occurs (an obese teenager with high blood pressure developing cardiovascular disease later in life).In either case, the interpretation is the same – the chance that some event will happen in the future – and modeling these complex phenomena as random can be useful.

Mathematicians and philosophers have struggled for centuries (literally) to arrive at a clear statement of how probability is defined, or what it means. In this text we use the most common definition, which also has the clearest interpretation.

Probability

The **probability** of an outcome is the proportion of times the outcome would occur if the random phenomenon could be observed an infinite number of times.

Probability is defined as a proportion, and it always takes values between 0 and 1 (inclusively). It may also be displayed as a percentage between 0% and 100%.

It is easy to imagine rolling dice a large number of times to observe the law of large numbers, but for examples like the CF example, the interpretation of probability is more hypothetical, since family sizes are typically small. But it is not too difficult to imagine a thought experiment in which two parents have many children. If the two parents are unaffected carriers, approximately 25% of their off spring will suffer from CF.

This definition of probability can be illustrated by rolling a die many times. Let $\hat{p}_n$ be the proportion of outcomes that are 1 after the first n rolls. As the number of rolls increases, $\hat{p}_n$ will converge to the probability of rolling a 1, $p = 1/6$. Figure **??** shows this convergence for 100,000 die rolls. The tendency of $\hat{p}_n$ to stabilize around p is described by the **Law of Large Numbers**.

The behavior shown in 2.2 matches most people's intuition about probability, but proving mathematically that the behavior is always true is surprisingly difficult and is

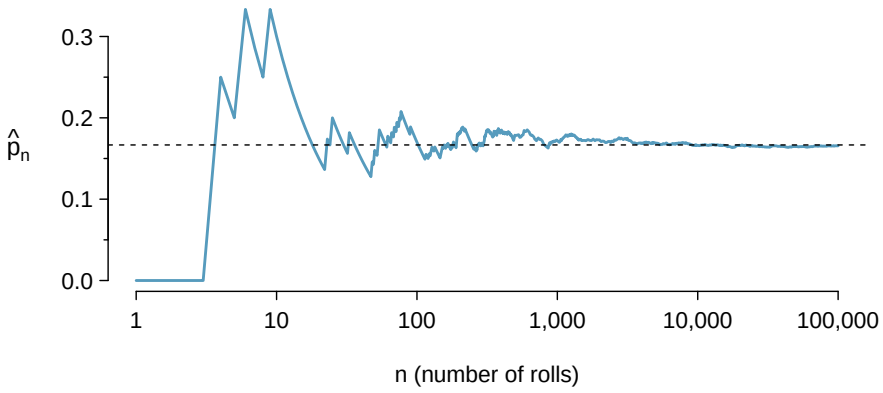


Figure 2.2: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

beyond the level of this text. Mathematicians call the result *The Law of Large Numbers*, which is used to justify mathematically this intuitively appealing definition.

Law of Large Numbers

As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the probability p of that outcome.

Occasionally the proportion will veer off from the probability and appear to defy the Law of Large Numbers, as $\hat{p}_n$ does many times in Figure **??**. However, these deviations become smaller as the number of rolls increases.

The notation p is the probability of rolling a 1. We can also write this probability as

$$P(\text{rolling a } 1)$$

$P(A)$
Probability of outcome A

As we become more comfortable with this notation, we will abbreviate it further. For instance, if it is clear that the process is "rolling a die", we could abbreviate $P(\text{rolling a } 1)$ as $P(1)$. We also have a notation for an event itself, so the event A of rolling a 1 will be written as $A = \{\text{ rolling a } 1\}$, with associated probability $P(A)$.

2.1.3 Disjoint or mutually exclusive outcomes

Two outcomes are called **disjoint** or **mutually exclusive** if they cannot both happen. When rolling a die, the outcomes 1 and 2 are disjoint since they cannot both occur. In the cystic fibrosis example, the two outcomes of a wild-type gene from the mother and a mutated gene from the father and a mutated gene from the mother, wild-type from the father are disjoint. In the die example, the outcomes 1 and "rolling an odd number" are not disjoint since both occur if the outcome of the roll is a 1. The outcomes of a child being affected and having at least one mutated copy of CFTR and not disjoint. The terms *disjoint* and *mutually exclusive* are equivalent and interchangeable.

Calculating the probability of disjoint outcomes is easy. When rolling a die, the outcomes 1 and 2 are disjoint, and we compute the probability that one of these outcomes will occur by adding their separate probabilities:

$$P(1 \text{ or } 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3$$

What about the probability of rolling a 1, 2, 3, 4, 5, or 6? Here again, all of the outcomes are disjoint so we add the probabilities:

$$\begin{aligned} &P(1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5 \text{ or } 6) \\ &\quad = P(1) + P(2) + P(3) + P(4) + P(5) + P(6) \\ &\quad = 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1. \end{aligned}$$

The probability that a child will be an unaffected carrier in the CF example is $(1/2) = (1/2) = 1/4$.

The **Addition Rule** guarantees the accuracy of this approach when the outcomes are disjoint.

Addition Rule of disjoint outcomes

If A_1 and A_2 represent two disjoint outcomes, then the probability that one of them occurs is given by

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

If there are many disjoint outcomes A_1, ..., A_k, then the probability that one of these outcomes will occur is

$$P(A_1) + P(A_2) + \cdots + P(A_k) \tag{2.8}$$

Probability problems rarely consider individual outcomes and instead consider *sets* or *collections* of outcomes. Let A represent the event where a die roll results in 1 or 2 and B represent the event that the die roll is a 4 or a 6. We write A as the set of outcomes $\{1, 2\}$ and $B = \{4, 6\}$. These sets are commonly called **events**. Because A and B have no elements in common, they are disjoint events. A and B are represented in Figure 2.3.

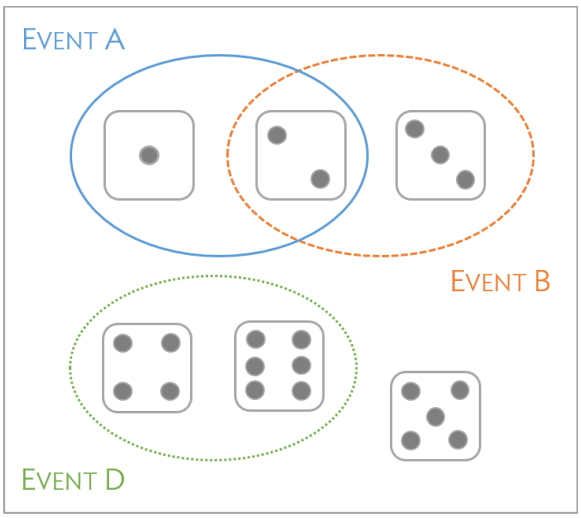


Figure 2.3: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common.

The Addition Rule applies to both disjoint outcomes and disjoint events. The probability that one of the disjoint events A or B occurs is the sum of the separate probabilities:

$$P(A \text{ or } B) = P(A) + P(B) = 1/3 + 1/3 = 2/3$$

Guided Practice 2.9 (a) Verify the probability of event A, $P(A)$, is 1/3 using the Addition Rule. (b) Do the same for event B.[1]

Guided Practice 2.10 (a) Using Figure 2.3 as a reference, what outcomes are represented by event D? (b) Are events B and D disjoint? (c) Are events A and D disjoint?[2]

Guided Practice 2.11 In Guided Practice 2.10, you confirmed B and D from Figure 2.3 are disjoint. Compute the probability that event B or event D occurs.[3]

should we add more genetics problems here? I have removed the email example because of the possible confusion between events involving sampling from a population vs a study sample. If we think we can make that clear, we can use examples from famuss, perhaps by posing a problem of sampling members from the study participants. Note also that this is moving more slowly than the Stat 102 notes, but we did show some of this material on the blackboard. If we use this chapter in 102, perhaps we can move quickly to more complicated examples.

2.1.4 Probabilities when events are not disjoint

Let's consider calculations for two events that are not disjoint in the context of a regular deck of 52 cards, represented in Table 2.4. If you are unfamiliar with the cards in a regular deck, please see the footnote.[4]

Guided Practice 2.12 (a) What is the probability that a randomly selected card is

[1] (a) $P(A) = P(1 \text{ or } 2) = P(1) + P(2) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$. (b) Similarly, $P(B) = 1/3$.

[2] (a) Outcomes 2 and 3. (b) Yes, events B and D are disjoint because they share no outcomes. (c) The events A and D share an outcome in common, 2, and so are not disjoint.

[3] Since B and D are disjoint events, use the Addition Rule: $P(B \text{ or } D) = P(B) + P(D) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

[4] The 52 cards are split into four **suits**: ♣ (club), ♢ (diamond), ♡ (heart), ♠ (spade). Each suit has its 13 cards labeled: 2, 3, ..., 10, J (jack), Q (queen), K (king), and A (ace). Thus, each card is a unique combination of a suit and a label, e.g. 4♡ and J♣. The 12 cards represented by the jacks, queens, and kings are called `face cards`. The cards that are ♢ or ♡ are typically colored red while the other two suits are typically colored black.

2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣	A♣
2♢	3♢	4♢	5♢	6♢	7♢	8♢	9♢	10♢	J♢	Q♢	K♢	A♢
2♡	3♡	4♡	5♡	6♡	7♡	8♡	9♡	10♡	J♡	Q♡	K♡	A♡
2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠	A♠

Table 2.4: Representations of the 52 unique cards in a deck.

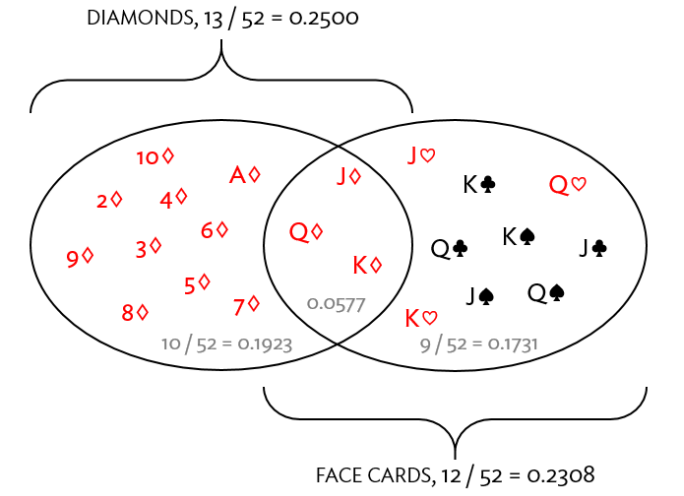


Figure 2.5: A Venn diagram for diamonds and face cards.

a diamond? (b) What is the probability that a randomly selected card is a face card?[5]

Venn diagrams are useful when outcomes can be categorized as "in" or "out" for two or three variables, attributes, or random processes. The Venn diagram in Figure 2.5 uses a circle to represent diamonds and another to represent face cards. If a card is both a diamond and a face card, it falls into the intersection of the circles. If it is a diamond but not a face card, it will be in part of the left circle that is not in the right circle (and so on). The total number of cards that are diamonds is given by the total number of cards in the diamonds circle: $10 + 3 = 13$. The probabilities are also shown (e.g. $10/52 = 0.1923$).

Let A represent the event that a randomly selected card is a diamond and B represent the event that it is a face card. How do we compute $P(A \text{ or } B)$? Events A and B are not disjoint – the cards $J\diamondsuit$, $Q\diamondsuit$, and $K\diamondsuit$ fall into both categories – so we cannot use the

[5] (a) There are 52 cards and 13 diamonds. If the cards are thoroughly shuffled, each card has an equal chance of being drawn, so the probability that a randomly selected card is a diamond is $P(\diamondsuit) = \frac{13}{52} = 0.250$. (b) Likewise, there are 12 face cards, so $P(\text{face card}) = \frac{12}{52} = \frac{3}{13} = 0.231$.

Addition Rule for disjoint events. Instead we use the Venn diagram. We start by adding the probabilities of the two events:

$$P(A) + P(B) = P(\diamondsuit) + P(\text{face card}) = 12/52 + 13/52$$

However, the three cards that are in both events were counted twice, once in each probability. We must correct this double counting:

$$\begin{aligned} P(A \text{ or } B) &= P(\text{face card or } \diamondsuit) \\ &= P(\text{face card}) + P(\diamondsuit) - P(\text{face card and } \diamondsuit) \\ &= 13/52 + 12/52 - 3/52 \\ &= 22/52 = 11/26 \end{aligned} \tag{2.13}$$

Equation (2.13) is an example of the **General Addition Rule**.

General Addition Rule

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{2.14}$$

where $P(A \text{ and } B)$ is the probability that both events occur.

TIP: "or" is inclusive

When we write "or" in statistics, we mean "and/or" unless we explicitly state otherwise. Thus, A or B occurs means A, B, or both A and B occur.

⊙ **Guided Practice 2.15** If A and B are disjoint, describe why this implies $P(A \text{ and } B) = 0$. (b) Using part (a), verify that the General Addition Rule simplifies to the

simpler Addition Rule for disjoint events if A and B are disjoint.[6]

Guided Practice 2.16

In areas of the developing world, the human immunodeficiency virus (HIV) and tuberculosis (TB) are infectious diseases that affect substantial proportions of the population. Individuals sometimes have both diseases (are co-infected); children of HIV-infected mothers may have HIV (be HIV$^+$) and TB can spread from one family member to another. In a mother child pair, let A = { the mother has HIV }, B = { the mother has TB }, C = { the child has HIV }, D = { the child has TB }. Write out the definitions of the events A or B, A and B, A and C, A or D.

2.1.5 Probability distributions

A **probability distribution** is a table of all disjoint outcomes and their associated probabilities. Table 2.6 shows the probability distribution for the sum of two dice.

Dice sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Table 2.6: Probability distribution for the sum of two dice.

Rules for probability distributions

A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes listed must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

Chapter **??** emphasized the importance of plotting data to provide quick summaries.

[6] (a) If A and B are disjoint, A and B can never occur simultaneously. (b) If A and B are disjoint, then the last term of Equation (2.14) is 0 (see part (a)) and we are left with the Addition Rule for disjoint events.

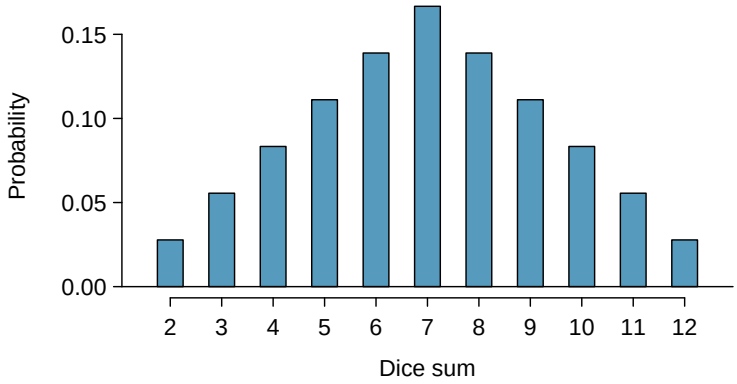


Figure 2.7: The probability distribution of the sum of two dice.

Probability distributions can also be summarized in a bar plot. The probability distribution for the sum of two dice is shown in Table 2.6 and plotted in Figure 2.7.

In this bar plots, the bar heights represent the probabilities of outcomes. If the outcomes are numerical and discrete, it is usually (visually) convenient to make a bar plot that resembles a histogram, as in the case of the sum of two dice.

A graph of probability distribution can convey important information about a distribution quickly.

The distribution of birth weights for 3,999,386 live births in the United States in 2010 is shown in figure 2.8. The data are available as part of the US CDC National Vital Statistics System [7]. The graph of the distribution shows that most babies born weighed between 2000 and 5000 grams (2kg to 5 kg), but there were both small (less than 1000 grams) and large (greater than 5000 grams) babies. Pediatricians think of normal birthweight as between 2.5 and 5 kg.

2.1.6 Complement of an event

S
Sample space

Rolling a die produces a value in the set {1, 2, 3, 4, 5, 6}. This set of all possible outcomes is called the **sample space** (S) for rolling a die. We often use the sample space to examine the scenario where an event does not occur.

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the

[7] http://205.207.175.93/vitalstats/ReportFolders/reportFolders.aspx

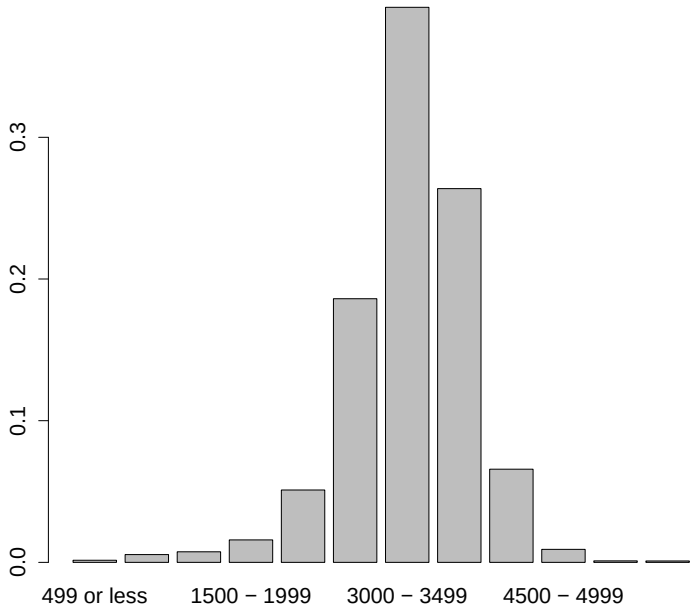


Figure 2.8: Distribution of birth weights (in grams) of babies born in the US in 2010

A^c
Complement of outcome A

complement of D represents all outcomes in our sample space that are not in D, which is denoted by $D^c = \{1, 4, 5, 6\}$. That is, D^c is the set of all possible outcomes not already included in D. Figure 2.9 shows the relationship between D, D^c, and the sample space S.

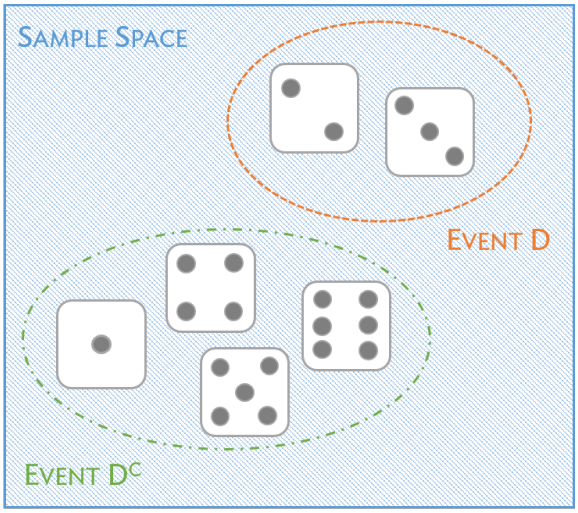


Figure 2.9: Event $D = \{2, 3\}$ and its complement, $D^c = \{1, 4, 5, 6\}$. S represents the sample space, which is the set of all possible events.

Guided Practice 2.17 (a) Compute $P(D^c) = P(\text{rolling a } 1, 4, 5, \text{ or } 6)$. (b) What is $P(D) + P(D^c)$?[8]

Guided Practice 2.18 Events $A = \{1, 2\}$ and $B = \{4, 6\}$ are shown in Figure 2.3 on page 15. (a) Write out what A^c and B^c represent. (b) Compute $P(A^c)$ and $P(B^c)$. (c) Compute $P(A) + P(A^c)$ and $P(B) + P(B^c)$.[9]

A complement of an event A is constructed to have two very important properties: (i) every possible outcome not in A is in A^c, and (ii) A and A^c are disjoint. Property (i) implies

$$P(A \text{ or } A^c) = 1 \tag{2.19}$$

[8] (a) The outcomes are disjoint and each has probability 1/6, so the total probability is 4/6 = 2/3. (b) We can also see that $P(D) = \frac{1}{6} + \frac{1}{6} = 1/3$. Since D and D^c are disjoint, $P(D) + P(D^c) = 1$.

[9] Brief solutions: (a) $A^c = \{3, 4, 5, 6\}$ and $B^c = \{1, 2, 3, 5\}$. (b) Noting that each outcome is disjoint, add the individual outcome probabilities to get $P(A^c) = 2/3$ and $P(B^c) = 2/3$. (c) A and A^c are disjoint, and the same is true of B and B^c. Therefore, $P(A) + P(A^c) = 1$ and $P(B) + P(B^c) = 1$.

That is, if the outcome is not in A, it must be represented in A^c. We use the Addition Rule for disjoint events to apply Property (ii):

$$P(A \text{ or } A^c) = P(A) + P(A^c) \tag{2.20}$$

Combining Equations (2.19) and (2.20) yields a very useful relationship between the probability of an event and its complement.

Complement

The complement of event A is denoted A^c, and A^c represents all outcomes not in A. A and A^c are mathematically related:

$$P(A) + P(A^c) = 1, \quad \text{i.e.} \quad P(A) = 1 - P(A^c) \tag{2.21}$$

In simple examples, computing A or A^c is feasible in a few steps. However, using the complement can save a lot of time as problems grow in complexity.

Guided Practice 2.22 Let A represent the event where we roll two dice and their total is less than 12. (a) What does the event A^c represent? (b) Determine $P(A^c)$ from Table 2.6 on page 19. (c) Determine $P(A)$.[10]

Guided Practice 2.23 Consider again the probabilities from Table 2.6 and rolling two dice. Find the following probabilities: (a) The sum of the dice is *not* 6. (b) The sum is at least 4. That is, determine the probability of the event $B = \{4, 5, ..., 12\}$. (c) The sum is no more than 10. That is, determine the probability of the event $D = \{2, 3, ..., 10\}$.[11]

[10] (a) The complement of A: when the total is equal to 12. (b) $P(A^c) = 1/36$. (c) Use the probability of the complement from part (b), $P(A^c) = 1/36$, and Equation (2.21): $P(\text{less than } 12) = 1 - P(12) = 1 - 1/36 = 35/36$.

[11] (a) First find $P(6) = 5/36$, then use the complement: $P(\text{not } 6) = 1 - P(6) = 31/36$.
(b) First find the complement, which requires much less effort: $P(2 \text{ or } 3) = 1/36 + 2/36 = 1/12$. Then calculate $P(B) = 1 - P(B^c) = 1 - 1/12 = 11/12$.
(c) As before, finding the complement is the more direct way to determine $P(D)$. First find $P(D^c) = P(11 \text{ or } 12) = 2/36 + 1/36 = 1/12$. Then calculate $P(D) = 1 - P(D^c) = 11/12$.

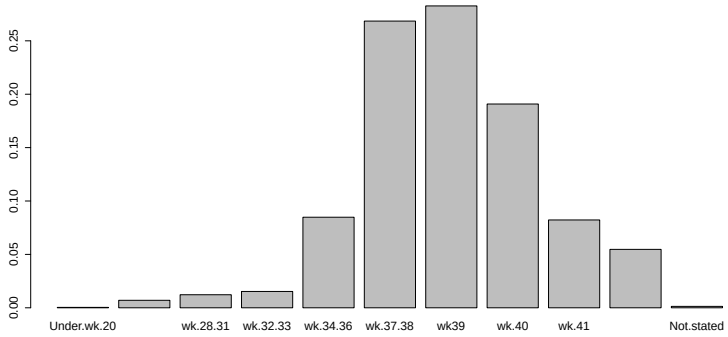


Figure 2.10: Distribution of gestational age for live births in the US in 2010, measured in weeks

Sometimes, information from a graph can be combined with using the complement of an event to calculate approximate probabilities. The gestational age of a newborn is the time between conception and birth. Because of the obvious difficulty of determining the exact date of conception, gestational ages are typically recorded in weeks. Figure 2.10 is the graphical representation of the distribution of gestational ages for the 3,999,386 babies born in 2010. Babies born between 38 and 42 weeks of gestational age is considered normal, but the term 'full term' is used for births between 39 and 40 weeks gestational age. The graph shows that approximately 30% of births occur at 39 weeks, and slightly less than 30% occur at 40 weeks, so approximately 50% of babies are considered full term. Instead of adding up the heights of the bars for gestational ages outside the full term range, using the complement of the event of a full term birth, it is clear that approximately 50% of births not considered full term.

The distribution of gestational age is shown in tabular form in Table **??**. The table shows the exact value of the proportion of babies born at 39 or 40 weeks (0.47), but when examining the important features of a distribution, approximate values are often sufficient. In some instances, the graph is all that will be available. Since small probabilities are difficult to read accurately from the graph of a distribution, they are best read from the table. Pre-term babies are those born at less than 37 weeks gestational age. Table **??** shows that the probability of this event is $0.01 + 0.01 + 0.02 + 0.08 = 0.12$. Of course, even the table shows approximate values, since the small proportion of very premature babies

born at less than 20 weeks is rounded to zero.

two problems with the example: it is too clumsy for what it accomplishes, and I have included the not stated category in the calculations of the proportions. This is negligible, but wrong.

	x
Under.wk.20	0.00
wk.20.27	0.01
wk.28.31	0.01
wk.32.33	0.02
wk.34.36	0.08
wk.37.38	0.27
wk39	0.28
wk.40	0.19
wk.41	0.08
wk.42.and.over	0.05
Not.stated	0.00

value labels should be modified

not satisfied with the way this example worked out; it should be improved or changed

2.1.7 Independence

Just as variables and observations can be independent, random phenomena can be independent, too. Two phenomena or processes are **independent** if knowing the outcome of one provides no information about the outcome of the other. For instance, flipping a coin and rolling a die are two independent processes – knowing the coin was heads does not help determine the outcome of a die roll. On the other hand, stock prices usually move up or down together, so they are not independent.

Independence was used implicitly on page 10 in the second solution to the probability that two carriers will have an affected child with cystic fibrosis. The assumption that half of the offspring who have received a mutated CF gene from the mother will receive a mutated gene from the father is essentially an independence assumption – genes are passed along from the mother and father independently.

JV: note to self, fix color of dice (and also in example...)

Example 2.5 provides a basic example of two independent processes: rolling two

dice. We want to determine the probability that both will be 1. Suppose one of the dice is red and the other white. If the outcome of the red die is a 1, it provides no information about the outcome of the white die. We first encountered this same question in Example 2.5 (page 10), where we calculated the probability using the following reasoning: $1/6^{th}$ of the time the red die is a 1, and $1/6^{th}$ of *those* times the white die will also be 1. This is illustrated in Figure 2.11. Because the rolls are independent, the probabilities of the corresponding outcomes can be multiplied to get the final answer: $(1/6)\times(1/6) = 1/36$. This can be generalized to many independent processes.

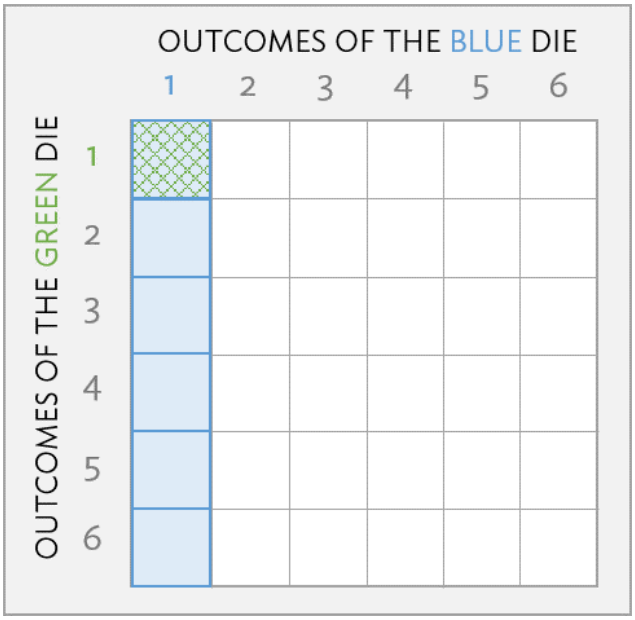


Figure 2.11: $1/6^{th}$ of the time, the first roll is a 1. Then $1/6^{th}$ of *those* times, the second roll will also be a 1.

Example 2.24 What if there was also a blue die independent of the other two? What is the probability of rolling the three dice and getting all 1s?

The same logic applies from Example 2.5. If $1/36^{th}$ of the time the white and red

dice are both 1, then $1/6^{th}$ of *those* times the blue die will also be 1, so multiply:

$$P(white = 1 \text{ and } red = 1 \text{ and } blue = 1) = P(white = 1) \times P(red = 1) \times P(blue = 1)$$
$$= (1/6) \times (1/6) \times (1/6) = 1/216$$

Example 2.24 illustrates what is called the Multiplication Rule for independent processes.

Multiplication Rule for independent processes

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$P(A \text{ and } B) = P(A) \times P(B) \tag{2.25}$$

Similarly, if there are k events A_1, ..., A_k from k independent processes, then the probability they all occur is

$$P(A_1) \times P(A_2) \times \cdots \times P(A_k)$$

In applications to biology or medicine, complicated probability problems are often solved with the simple ideas used in the dice examples.

⊙ **Guided Practice 2.26** About 9% of people are left-handed. Suppose 2 people are selected at random from the U.S. population. Because the sample size of 2 is very small relative to the population, it is reasonable to assume these two people are independent. (a) What is the probability that both are left-handed? (b) What is the probability that both are right-handed?[12]

[12] (a) The probability the first person is left-handed is 0.09, which is the same for the second person. We apply the Multiplication Rule for independent processes to determine the probability that both will be left-handed: $0.09 \times 0.09 = 0.0081$.

Guided Practice 2.27 Suppose 5 people are selected at random.[13]

(a) What is the probability that all are right-handed?

(b) What is the probability that all are left-handed?

(c) What is the probability that not all of the people are right-handed?

Suppose the variables `handedness` and `gender` are independent, i.e. knowing someone's `gender` provides no useful information about their `handedness` and vice-versa. Then we can compute whether a randomly selected person is right-handed and female[14] using the Multiplication Rule:

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

Guided Practice 2.28 Three people are selected at random.[15]

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

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

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

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

(b) It is reasonable to assume the proportion of people who are ambidextrous (both right and left handed) is nearly 0, which results in $P(\text{right-handed}) = 1 - 0.09 = 0.91$. Using the same reasoning as in part (a), the probability that both will be right-handed is $0.91 \times 0.91 = 0.8281$.

[13](a) The abbreviations RH and LH are used for right-handed and left-handed, respectively. Since each are independent, we apply the Multiplication Rule for independent processes:

$$\begin{aligned} P(\text{all five are RH}) &= P(\text{first = RH, second = RH, ..., fifth = RH}) \\ &= P(\text{first = RH}) \times P(\text{second = RH}) \times \cdots \times P(\text{fifth = RH}) \\ &= 0.91 \times 0.91 \times 0.91 \times 0.91 \times 0.91 = 0.624 \end{aligned}$$

(b) Using the same reasoning as in (a), $0.09 \times 0.09 \times 0.09 \times 0.09 \times 0.09 = 0.0000059$

(c) Use the complement, $P(\text{all five are RH})$, to answer this question:

$$P(\text{not all RH}) = 1 - P(\text{all RH}) = 1 - 0.624 = 0.376$$

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

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

Example 2.29 *Mandatory drug testing* Mandatory drug testing in the work place is common in professions such as air traffic controllers, transportation workers and government security agencies. A false positive in a drug screening test occurs when the test incorrectly indicates that a screened person is an illegal drug user. Suppose a mandatory drug test has a false-positive rate of 1.2% (i.e., has probability 0.012 of indicating that an employee is using illegal drugs even when that is not the case), and suppose a company uses the test to screen employees for drug use. Given 150 employees who are in reality drug free, what is the probability that at least one will (falsely) test positive if the outcome of one drug test has no affect on the other 149?

The solution uses independence (the assumption that the outcome of one test has no effect on the others) and the multiplication rule to calculate the probability of the complement of the event asked about.

$$\begin{aligned} P(\text{At least 1 "+"}) &= P(\text{1 or 2 or 3 ... or 150 are "+"}) \\ &= 1 - P(\text{None are "+"}) \\ &= 1 - P(\text{150 "-"}) \\ P(\text{150 are "-"}) &= P(\text{1 is "-"})^{150} \\ &= (0.988)^{150} = 0.16. \end{aligned}$$

So $P(\text{At least 1 is "+"}) = 1 - P(\text{150 are "-"}) = 0.84$.

should we be more formal here in defining events? Also, this is the example that we also solved in R, two different ways. Those solutions are candidates for the R supplement

Some people find the result surprising. Even when using a test with a small probability of a false positive, the company is more than 80% likely to incorrectly claim at least one employee is an illegal drug user.

Guided Practice 2.30 Because of the high likelihood of at least one false positive in company wide drug screening programs, an individual with a positive test is almost

always re-tested with a different screening test, one that is more expensive than the first but with a lower false positive probability. Suppose the second test has a false positive rate of 0.8%. What is the probability that an employee who is not using illegal drugs will test positive on both tests?

solution to be added if we keep the problem

Guided Practice 2.31

There are eight different common blood types, which are determined by the presence of certain antigens located on cell surfaces. Antigens are substances used by the immune system to recognize self versus non-self; if the immune system encounters antigens not normally found on the body's own cells, it will attack the foreign cells. When patients receive blood transfusions, it is critical that the antigens of transfused cells match those of the patient's, or else an immune system response will be triggered.

The ABO blood group system consists of four different blood groups, which describe whether an individual's red blood cells carry the A antigen, B antigen, both, or neither. The ABO gene has three alleles: I^A, I^B, and i. The i allele is recessive to both I^A and I^B, and does not produce antigens; thus, an individual with genotype I^Ai is blood group A and an individual with genotype I^Bi is blood group B. The I^A and I^B are codominant, such that individuals of I^AI^B genotype are AB. Individuals homozygous for the i allele are known as blood group O, with neither A nor B antigens.

a) *ABO, Independence.* Suppose that both members of a couple have Group AB blood.

   i. What is the probability that a child of this couple will have Group A blood?

   ii. What is the probability that they have two children with Group A blood?

solutions to be added if we keep the exercise

The examples in this section have used independence to solve probability problems. Sometimes the definition of independence can be used to check whether two events are independent – two events A and B are independent if they satisfy Equation (2.25).

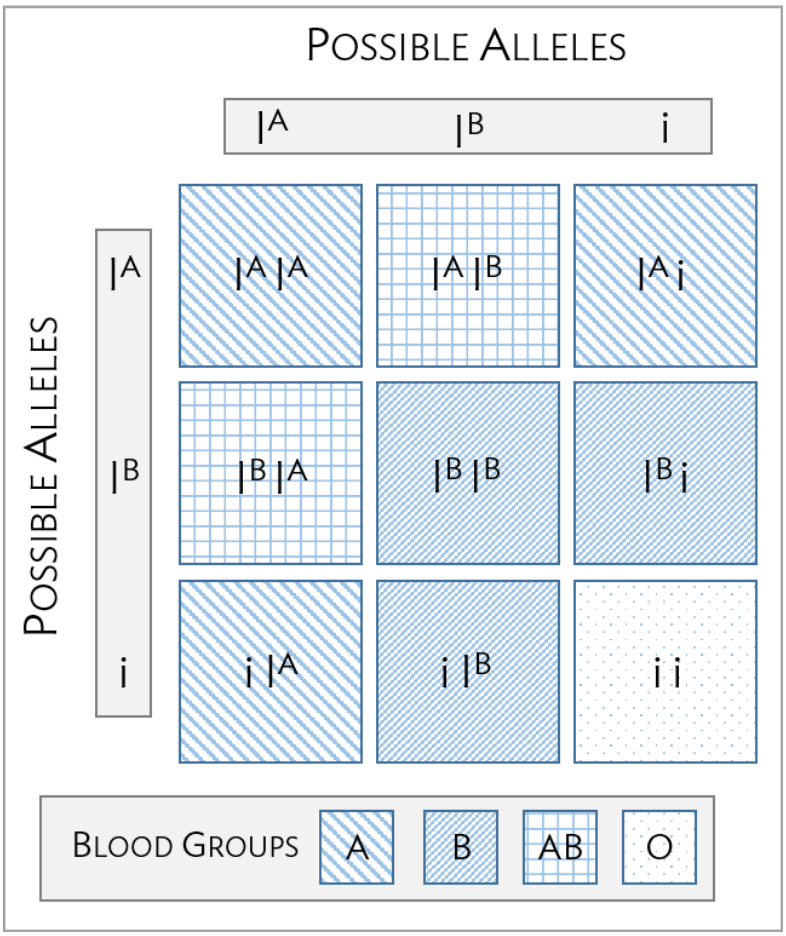


Figure 2.12: Inheritance of ABO blood groups.

Example 2.32 If we shuffle up a deck of cards and draw one, is the event that the card is a heart independent of the event that the card is an ace?

The probability the card is a heart is 1/4 and the probability that it is an ace is 1/13. The probability the card is the ace of hearts is 1/52. We check whether Equation 2.25 is satisfied:

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

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

Example 2.33 I

n the general population, about 15% of adults between 25 and 40 years of age are hypertensive. Suppose that among males of this age, hypertension occurs about 18% of the time. Is hypertension independent of sex?

solution to be filled in if we keep it. Emphasize in the solution how the wording here is more realistic than the playing card/dice examples.

2.2 Conditional probability

Precise estimates are difficult to come by, but the US CDC estimated that in 2012, approximately 29.1 million people have type 2 diabetes, or about 9.3% of the population. Twenty-one million of these cases of diabetes are diagnosed, while 8.1 million cases are undiagnosed (people living with diabetes but they and their physicians are unaware that they have the disease). A health care practitioner seeing a new patient and having no demographic or health information about the patient should expect a 9.3% chance that the patient might have diabetes, diagnosed or otherwise. But intake interviews usually include background information about patients, so that a health care practitioner knows a bit more about a new patient. Not surprisingly, the prevalence or probability of type 2 diabetes varies with age. Between the ages of 20 and 44, approximately 4% of the population have diabetes, but by age 65 and older, almost 27% of of that age group have diabetes. Knowing the age of a patient provides information about the chance of diabetes, so that age and diabetes status are not independent. While the probability of diabetes in randomly chosen member of the population is 0.093, the *conditional* probability of diabetes in a person known to be 65 or older is about 0.27.

Conditional probability is used to characterize how the probability of an outcome varies with the knowledge of another factor or condition, and is closely related to the concepts of marginal and joint probabilities.

2.2.1 Marginal and joint probabilities

Tables 2.13 and 2.14 provide additional information about the relationship between diabetes prevalence and age. [16] Table 2.13 is a contingency table like those discussed in Chapter 1, but for the entire US 2012 population; the values in the table are in thousands, to make the table more readable. The table shows in the first row, for instance, that in the entire population of approximately 313,320,000 approximately 200,000 individuals were in the age group less than 20 and suffered from type 2 diagnosis, or about 0.1%. The table also provides the information among the approximately 86,864,000 individuals less than 20 years of age, only 200,000 suffered from type 2 diabetes, approximately 0.2%. The distinction between the two statements is small but important – the first provides information about the size of the type 2 diabetes population relative to the entire population and the second about the size of the diabetes population less than 20 year old age group relative to the size of that age group.

Guided Practice 2.34

What fraction of the US population are 45 to 64 years of age and have diabetes? What fraction of the population age 45 to 64 have diabetes?

	Diabetes	No Diabetes	Sum
Less than 20 years	200	86664	86864
20 to 44 years	4300	98724	103024
45 to 64 years	13400	68526	81926
Greater than 64 years	11200	30306	41506
Sum	29100	284220	313320

Table 2.13: Contingency table showing type 2 diabetes status and age group, in thousands

The counts in Table 2.13 have been converted to proportions by dividing each value in the cells of the contingency table by the total population size, 313,320,000. The entries in this table show the proportions of the population in each of the 8 categories defined by diabetes status and age. If these proportions are interpreted as probabilities for randomly

[16] Because the CDC provides only approximate numbers for diabetes prevalence, the numbers in the table are approximations to actual population counts.

chosen individuals from the population, 0.014 in row 2 implies that the probability of selecting someone at random who has diabetes and whose age is between 20 and 44 is 0.014, or 1.4%. The entries in the 8 main table cells (excluding the values in the margins) are called **joint probabilities** since they specify the probability of two events happening at the same time – diabetes and a particular age group. In probability notation, 0.014 = P(diabetes and age 20 to 44). It is common to also write this as P(diabetes, age 20 to 44), with a comma replacing "and".

The values in the last row and column of the table are the sums of the corresponding rows or columns. Since 0.329 is the sum of the of the probabilities of the disjoint events (diabetes and age 20 to 44) and (no diabetes and age 20 to 44), it is the probability of being in the age group 20 to 44. The row and column sums are called **marginal probabilities**; they are probabilities about only one type of event, age in the case of 0.0329. The sum of the first column (0.093) is the marginal probability of a member of the population having diabetes.

	Diabetes	No Diabetes	Sum
Less than 20 years	0.001	0.277	0.277
20 to 44 years	0.014	0.315	0.329
45 to 64 years	0.043	0.219	0.261
Greater than 64 years	0.036	0.097	0.132
Sum	0.093	0.907	1.000

Table 2.14: Probability table summarizing diabetes status and age group

Marginal and joint probabilities

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

Guided Practice 2.35 What is the interpretation of the value 0.907 in the last row of the table? Of the value 1.000 in the bottom right corner?

2.2.2 Defining conditional probability

The probability that a randomly selected individual from the US has diabetes is 0.093, the sum of the first column in Table 2.14. How does that probability change if we know the individual's age is 65 or greater? Table 2.13 shows that 11,200,000 of the 41,506,000 people in that age group have diabetes, so the likelihood that someone from that age has diabetes is

$$\frac{11,200,000}{41,506,000} = 0.27,$$

or 27%. The additional information about age allows a better estimate of the probability of diabetes; the conditional probability of diabetes, given the information that an individual is older than 65, is 0.27.

Since

$$\begin{aligned}\frac{11,200,000}{41,506,000} &= \frac{11,200,000/313,320,000}{41,506,000/313,320,000} \\ &= \frac{0.036}{0.132} \\ &= 0.270,\end{aligned}$$

the calculation of conditional probability could have been done using the values in Table 2.13. The conditional probability of diabetes given age 65 or greater is simply the ratio of the proportion of the population with diabetes and age 65 or greater divided by the proportion greater than age 65. This leads to the mathematical definition of conditional probability.

Conditional probability

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

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \tag{2.36}$$

Guided Practice 2.37 (a) Write out the following statement in conditional probability notation: *"The probability a randomly selected person has diabetes, given that his or her age is between 45 and 64 "*. Notice that the condition is now based on the teenager, not the parent.

(b) Calculate the conditional probability in part (a) (c) Write out the following statement in conditional probability notation: *"The probability a randomly selected person is between 45 and 64 years old, given that the person has diabetes"*. Notice that the condition is now based on the diabetes, not the age.

(d) Calculate the probability in part (c).

2.2.3 Smallpox in Boston, 1721

I am not fond of small pox example, since it is a biased sample, but leaving it for now

The `smallpox` data set provides a sample of 6,224 individuals from the year 1721 who were exposed to smallpox in Boston.[17] Doctors at the time believed that inoculation, which involves exposing a person to the disease in a controlled form, could reduce the likelihood of death.

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

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

Table 2.15: Contingency table for the `smallpox` data set.

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

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

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

Guided Practice 2.38 Write out, in formal notation, the probability a randomly selected person who was not inoculated died from smallpox, and find this probability.[18]

Guided Practice 2.39 Determine the probability that an inoculated person died from smallpox. How does this result compare with the result of Guided Practice 2.38?[19]

Guided Practice 2.40 The people of Boston self-selected whether or not to be inoculated. (a) Is this study observational or was this an experiment? (b) Can we infer any causal connection using these data? (c) What are some potential confounding variables that might influence whether someone lived or died and also affect whether that person was inoculated?[20]

2.2.4 General multiplication rule

Section 2.1.7 introduced the Multiplication Rule for independent processes. Here we provide the **General Multiplication Rule** for events that might not be independent.

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

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

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

General Multiplication Rule

If A and B represent two outcomes or events, then

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

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

This General Multiplication Rule is simply a rearrangement of the definition for conditional probability in Equation (2.36) on page 35.

Example 2.41 Consider the `smallpox` data set. Suppose we are given only two pieces of information: 96.08% of residents were not inoculated, and 85.88% of the residents who were not inoculated ended up surviving. How could we compute the probability that a resident was not inoculated and lived?

We will compute our answer using the General Multiplication Rule and then verify it using Table 2.16. We want to determine

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \text{no})$$

and we are given that

$$P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \text{no}) = 0.8588$$
$$P(\texttt{inoculated} = \text{no}) = 0.9608$$

Among the 96.08% of people who were not inoculated, 85.88% survived:

$$P(\texttt{result} = \texttt{lived} \text{ and } \texttt{inoculated} = \text{no}) = 0.8588 \times 0.9608 = 0.8251$$

This is equivalent to the General Multiplication Rule. We can confirm this probability in Table 2.16 at the intersection of no and `lived` (with a small rounding error).

Guided Practice 2.42 Use $P(\texttt{inoculated} = \texttt{yes}) = 0.0392$ and $P(\texttt{result} = \texttt{lived} \mid \texttt{inoculated} = \texttt{yes}) = 0.9754$ to determine the probability that a person was both inoculated and lived.[21]

Guided Practice 2.43 If 97.45% of the people who were inoculated lived, what proportion of inoculated people must have died?[22]

Sum of conditional probabilities

Let A_1, ..., A_k represent all the disjoint outcomes for a variable or process. Then if B is an event, possibly for another variable or process, we have:

$$P(A_1|B) + \cdots + P(A_k|B) = 1$$

The rule for complements also holds when an event and its complement are conditioned on the same information:

$$P(A|B) = 1 - P(A^c|B)$$

Guided Practice 2.44 Based on the probabilities computed above, does it appear that inoculation is effective at reducing the risk of death from smallpox?[23]

2.2.5 Independence and conditional probability

If two events are independent, knowing the outcome of one should provide no information about the other. That intuitively clear statement can be shown mathematically.

[21] The answer is 0.0382, which can be verified using Table 2.16.

[22] There were only two possible outcomes: `lived` or `died`. This means that 100% - 97.45% = 2.55% of the people who were inoculated died.

[23] The samples are large relative to the difference in death rates for the "inoculated" and "not inoculated" groups, so it seems there is an association between `inoculated` and `outcome`. However, as noted in the solution to Guided Practice 2.40, this is an observational study and we cannot be sure if there is a causal connection. (Further research has shown that inoculation is effective at reducing death rates.)

Guided Practice 2.45 Let X and Y represent the outcomes of rolling two dice.[24]

(a) What is the probability that the first die, X, is 1?

(b) What is the probability that both X and Y are 1?

(c) Use the formula for conditional probability to compute $P(Y = 1 \mid X = 1)$.

(d) What is $P(Y = 1)$? Is this different from the answer from part (c)? Explain.

[24] Brief solutions: (a) 1/6. (b) 1/36. (c) $\frac{P(Y=1 \text{ and } X=1)}{P(X=1)} = \frac{1/36}{1/6} = 1/6$. (d) The probability is the same as in part (c): $P(Y = 1) = 1/6$. The probability that $Y = 1$ was unchanged by knowledge about X, which makes sense as X and Y are independent.

It is not difficult to show in Guided Practice 2.45(c) that the conditioning information has no influence by using the Multiplication Rule for independence events:

$$\begin{aligned} P(Y = 1 \mid X = 1) &= \frac{P(Y = 1 \text{ and } X = 1)}{P(X = 1)} \\ &= \frac{P(Y = 1) \times P(X = 1)}{P(X = 1)} \\ &= P(Y = 1) \end{aligned}$$

Guided Practice 2.46 Casinos often rely on gamblers not understanding independence. Suppose the last five outcomes on a roulette table were `black`. What is wrong with the reasoning that the next outcome is highly likely to be red?

There is a subtle but important point behind the last example. The probability of the next six outcomes being black is different than the probability that the sixth outcome is black when a gambler has seen the last five outcomes and knows that they are black. This is an example of an unconditional versus a conditional probability.

2.2.6 Tree diagrams

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. They are most useful when two or more processes occur in a sequence and each process is conditioned on its predecessors.

The `smallpox` data fit this description. We see the population as split by `inoculation`: `yes` and `no`. Following this split, survival rates were observed for each group. This structure is reflected in the **tree diagram** shown in Figure 2.17. The first branch for `inoculation` is said to be the **primary** branch while the other branches are **secondary**.

Tree diagrams are annotated with marginal and conditional probabilities, as shown in Figure 2.17. This tree diagram splits the smallpox data by `inoculation` into the `yes` and `no` groups with respective marginal probabilities 0.0392 and 0.9608. The secondary branches are conditioned on the first, so we assign conditional probabilities to these branches. For example, the top branch in Figure 2.17 is the probability that `result`

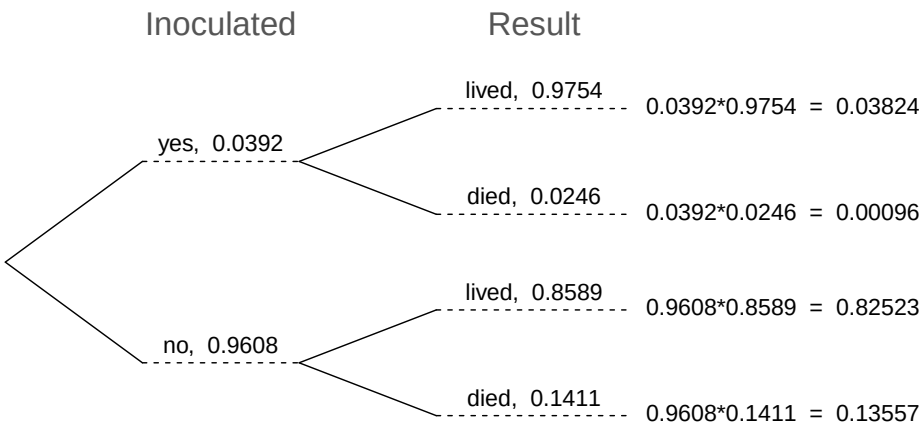


Figure 2.17: A tree diagram of the `smallpox` data set.

= `lived` conditioned on the information that `inoculated` = `yes`. We may (and usually do) construct joint probabilities at the end of each branch in our tree by multiplying the numbers we come across as we move from left to right. These joint probabilities are computed using the General Multiplication Rule:

$$\begin{aligned} &P(\texttt{inoculated} = \texttt{yes} \text{ and } \texttt{result} = \texttt{lived}) \\ &\quad = P(\texttt{inoculated} = \texttt{yes}) \times P(\texttt{result} = \texttt{lived} | \texttt{inoculated} = \texttt{yes}) \\ &\quad = 0.0392 \times 0.9754 = 0.0382 \end{aligned}$$

In addition to being a graphical representation of how to compute a probability, tree diagrams are a useful way to organize the information in a probability problem and can often reduce a seemingly difficult problem to a series of simple steps

Example 2.47

In the general population, about 1 in 28 individuals (approximately 3.6%) is an unaffected carrier of a mutation in the cystic fibrosis gene, *CFTR*, discussed in example 2.6. Most unaffected carriers are unaware that they harbor the mutation. Suppose that people with cystic fibrosis do not live long enough to reproduce with a partner. In the absence of any testing information, what is the probability that a

child of two parent will have CF?

solution not give for now, since I did not want to take the time to draw the tree. This will be a good example to also solve algebraically, to show that the tree diagram is a representation of the theorem of total prob. Useful, since it will come up with Bayes Thm. Borrow some of the wording from the solution to the grade problem in OI that I removed here. Below has a start of the solution

solution The probabilities calculated in example 2.6 were conditional probabilities given the carrier status of the parents, though that term was not used in the example. This exercise asks for an unconditional probability about the disease status of the child, with no additional information about the parents. The solution uses methods already discussed, but combined is a series of steps.

The main steps are to first enumerate all the possibilities for carrier status of the parents, calculate the probabilities of each of those probabilities, then calculate the conditional probability of an affected child, given each of the possible outcomes of the parents, and finally use a tree to calculate the probability of a child being affected.

Since the problem assumes that affected parents do not reproduce, the pair of parents must satisfy one of the events A = (neither parent is a carrier), B = (the mother is a carrier, the father i
C = (the father is a carrier, the mother is not), and D = (both parents are carriers). Since the carrier status of the mother is independent of that of the father, the probabilities of these events are

- $P(A) = (1 - 0.036)(1 - 0.036) = 0.929$
- $P(B) = P(C) = (1 - 0.036)(0.036) = 0.035$
- $P(D) = (0.036)(0.036) = 0.001$

These probabilities do not sum to one because the events where one or both parents are homozygous for the mutation, that is, are affected.

solution continues now with the conditional probabilities and the tree. We should think

about whether this example is too big of a step from the others, and perhaps add an intermediate example

2.2.7 Bayes' Theorem

This chapter began with a straightforward question – what are the chances that a woman with an abnormal (i.e., positive) mammogram has breast cancer? For a clinician, this question can be rephrased as the conditional probability that a woman has breast cancer, given that her mammogram is abnormal. This conditional probability is called the **positive predictive value** of a mammogram. The characteristics of a mammogram (and diagnostic tests in general) are given with the reverse conditional probabilities – the probability that if a woman fact has breast cancer, a mammogram will detect it. More concisely, if A = (a mammogram is positive) and B = (a woman has breast cancer), the first question is answered by finding $P(A|B)$, while the the known characteristics of a mammogram specify $P(B|A)$.

Many problems provide information about

$$P(\text{statement about variable 1}|\ \text{statement about variable 2})$$

but ask for the reverse conditional probability:

$$P(\text{statement about variable 2}|\ \text{statement about variable 1})$$

The problem arises so often in medicine, and is so important, that is useful to have several ways to approach the problem – tree diagrams, a purely algebraic approach using Bayes' Theorem, and the construction of a large, hypothetical population that allows conditional probabilities to be calculated from contingency tables. The first solution uses tree diagrams and more specific information about mammograms and breast cancer.

Example 2.48 In Canada, about 0.35% of women over 40 will develop breast cancer in any given year. A common screening test for cancer is the mammogram, but it is not perfect. In about 11% of patients with breast cancer, the test gives a **false**

negative: it indicates a woman does not have breast cancer when she does have breast cancer. Similarly, the test gives a **false positive** in 7% of patients who do not have breast cancer: it indicates these patients have breast cancer when they actually do not.[25] If a random selected woman over 40 is tested for breast cancer using a mammogram and the test is positive – that is, the test suggests the woman has cancer – what is the probability she has breast cancer?

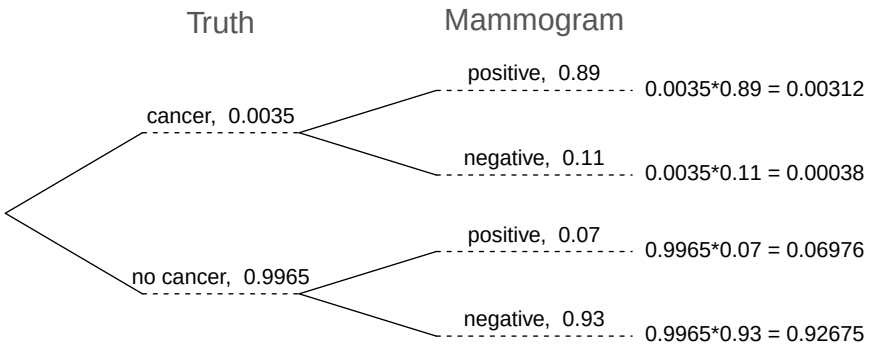


Figure 2.18: Tree diagram for Example 2.48, computing the probability a random patient who tests positive on a mammogram has breast cancer.

The problem provides enough information to compute the probability of testing positive if a woman has breast cancer ($1.00 - 0.11 = 0.89$), but not the reverse conditional probability of cancer given a positive test result. This reverse conditional probability may be broken into two pieces:

$$P(\text{has BC} \mid \text{mammogram}^+) = \frac{P(\text{has BC and mammogram}^+)}{P(\text{mammogram}^+)}$$

where "has BC" is an abbreviation for breast cancer and "mammogram$^+$" means the mammogram screening was positive. A tree diagram can be used to identify each probability and is shown in Figure 2.18. The probability the woman has breast

[25] The probabilities reported here were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

cancer and the mammogram is positive is

$$\begin{aligned} P(\text{has BC and mammogram}^+) &= P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC}) \\ &= 0.89 \times 0.0035 = 0.00312 \end{aligned}$$

The probability of a positive test result is the sum of the two corresponding scenarios:

$$\begin{aligned} P(\underline{\text{mammogram}^+}) &= P(\underline{\text{mammogram}^+} \text{ and has BC}) + P(\underline{\text{mammogram}^+} \text{ and no BC}) \\ &= P(\text{has BC})P(\text{mammogram}^+ \mid \text{has BC}) \\ &\quad + P(\text{no BC})P(\text{mammogram}^+ \mid \text{no BC}) \\ &= 0.0035 \times 0.89 + 0.9965 \times 0.07 = 0.07288 \end{aligned}$$

Then if the mammogram screening is positive for a patient, the probability the patient has breast cancer is

$$\begin{aligned} P(\text{has BC} \mid \text{mammogram}^+) &= \frac{P(\text{has BC and mammogram}^+)}{P(\text{mammogram}^+)} \\ &= \frac{0.00312}{0.07288} \approx 0.0428 \end{aligned}$$

Even with a positive mammogram, there is still only a 4% chance of breast cancer. Most people find this a surprising result. *return to this to finish why low prevalence leads to low ppv. that is the important concept for pre-meds and others*

Consider again the last equation of Example 2.48. Using the tree diagram, we can see that the numerator (the top of the fraction) is equal to the following product:

$$P(\text{has BC and mammogram}^+) = P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})$$

The denominator – the probability the screening was positive – is equal to the sum of

probabilities for each positive screening scenario:

$$P(\underline{\text{mammogram}^+}) = P(\underline{\text{mammogram}^+} \text{ and no BC}) + P(\underline{\text{mammogram}^+} \text{ and has BC})$$

In the example, each of the probabilities on the right side was broken down into a product of a conditional probability and marginal probability using the tree diagram.

$$\begin{aligned} P(\text{mammogram}^+) &= P(\text{mammogram}^+ \text{ and no BC}) + P(\text{mammogram}^+ \text{ and has BC}) \\ &= P(\text{mammogram}^+ \mid \text{no BC})P(\text{no BC}) \\ &\quad + P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC}) \end{aligned}$$

We can see an application of Bayes' Theorem by substituting the resulting probability expressions into the numerator and denominator of the original conditional probability.

$$\begin{aligned} &P(\text{has BC} \mid \text{mammogram}^+) \\ &= \frac{P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})}{P(\text{mammogram}^+ \mid \text{no BC})P(\text{no BC}) + P(\text{mammogram}^+ \mid \text{has BC})P(\text{has BC})} \end{aligned}$$

Bayes' Theorem

Consider the following conditional probability for variable 1 and variable 2:

$$P(\text{outcome } A_1 \text{ of variable 1} \mid \text{outcome } B \text{ of variable 2})$$

Bayes' Theorem states that this conditional probability can be identified as the following fraction:

$$\frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)} \tag{2.49}$$

where A_2, A_3, ..., and A_k represent all other possible outcomes of the first variable.

Bayes' Theorem is also called Bayes' rule.

Bayes' Theorem is just a generalization of what we have done using tree diagrams. The numerator identifies the probability of getting both A_1 and B. The denominator is the marginal probability of getting B. This bottom component of the fraction appears long and complicated since we have to add up probabilities from all of the different ways to get B. We always completed this step when using tree diagrams. However, we usually did it in a separate step so it didn't seem as complex.

To apply Bayes' Theorem correctly, there are two preparatory steps:

(1) First identify the marginal probabilities of each possible outcome of the first variable: $P(A_1)$, $P(A_2)$, ..., $P(A_k)$.

(2) Then identify the probability of the outcome B, conditioned on each possible scenario for the first variable: $P(B|A_1)$, $P(B|A_2)$, ..., $P(B|A_k)$.

Once each of these probabilities are identified, they can be applied directly within the formula.

2.3 Exercises

2.3.1 Defining probability

2.1 True or false. Determine if the statements below are true or false, and explain your reasoning.

(a) If a fair coin is tossed many times and the last eight tosses are all heads, then the chance that the next toss will be heads is somewhat less than 50%.

(b) Drawing a face card (jack, queen, or king) and drawing a red card from a full deck of playing cards are mutually exclusive events.

(c) Drawing a face card and drawing an ace from a full deck of playing cards are mutually exclusive events.

2.2 Roulette wheel. The game of roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball.

(a) You watch a roulette wheel spin 3 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

(b) You watch a roulette wheel spin 300 consecutive times and the ball lands on a red slot each time. What is the probability that the ball will land on a red slot on the next spin?

(c) Are you equally confident of your answers to parts (a) and (b)? Why or why not?

Photo by Håkan Dahlström (http://flic.kr/p/93fEzp) CC BY 2.0 license

2.3 Four games, one winner. Below are four versions of the same game. Your archnemisis gets to pick the version of the game, and then you get to choose how many times to flip a coin: 10 times or 100 times. Identify how many coin flips you should choose for each version of the game. It costs $1 to play each game. Explain your reasoning.

(a) If the proportion of heads is larger than 0.60, you win $1.

(b) If the proportion of heads is larger than 0.40, you win $1.

(c) If the proportion of heads is between 0.40 and 0.60, you win $1.

(d) If the proportion of heads is smaller than 0.30, you win $1.

2.4 Backgammon. Backgammon is a board game for two players in which the playing pieces are moved according to the roll of two dice. Players win by removing all of their pieces from the board, so it is usually good to roll high numbers. You are playing backgammon with a friend and you roll two 6s in your first roll and two 6s in your second roll. Your friend rolls two 3s in his first roll and again in his second row. Your friend claims that you are cheating, because rolling double 6s twice in a row is very unlikely. Using probability, show that your rolls were just as likely as his.

2.5 Coin flips. If you flip a fair coin 10 times, what is the probability of

(a) getting all tails?

(b) getting all heads?

(c) getting at least one tails?

2.6 Dice rolls. If you roll a pair of fair dice, what is the probability of

(a) getting a sum of 1?

(b) getting a sum of 5?

(c) getting a sum of 12?

2.7 Swing voters. A 2012 Pew Research survey asked 2,373 randomly sampled registered voters their political affiliation (Republican, Democrat, or Independent) and whether or not they identify as swing voters. 35% of respondents identified as Independent, 23% identified as swing voters, and 11% identified as both.[26]

(a) Are being Independent and being a swing voter disjoint, i.e. mutually exclusive?

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

(c) What percent of voters are Independent but not swing voters?

(d) What percent of voters are Independent or swing voters?

(e) What percent of voters are neither Independent nor swing voters?

(f) Is the event that someone is a swing voter independent of the event that someone is a political Independent?

2.8 Poverty and language. The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English (foreign language) at home, and 4.2% fall into both categories.[27]

(a) Are living below the poverty line and speaking a foreign language at home disjoint?

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

(c) What percent of Americans live below the poverty line and only speak English at home?

(d) What percent of Americans live below the poverty line or speak a foreign language at home?

(e) What percent of Americans live above the poverty line and only speak English at home?

(f) Is the event that someone lives below the poverty line independent of the event that the person speaks a foreign language at home?

2.9 Disjoint vs. independent. In parts (a) and (b), identify whether the events are disjoint, independent, or neither (events cannot be both disjoint and independent).

(a) You and a randomly selected student from your class both earn A's in this course.

(b) You and your class study partner both earn A's in this course.

(c) If two events can occur at the same time, must they be dependent?

2.10 Guessing on an exam. In a multiple choice exam, there are 5 questions and 4 choices for each question (a, b, c, d). Nancy has not studied for the exam at all and decides to randomly guess the answers. What is the probability that:

(a) the first question she gets right is the 5^{th} question?

(b) she gets all of the questions right?

(c) she gets at least one question right?

[26] **indepSwing**.
[27] **poorLang**.

2.11 Educational attainment of couples. The table below shows the distribution of education level attained by US residents by gender based on data collected during the 2010 American Community Survey.[28]

		Gender	
		Male	Female
Highest education attained	Less than 9th grade	0.07	0.13
	9th to 12th grade, no diploma	0.10	0.09
	HS graduate (or equivalent)	0.30	0.20
	Some college, no degree	0.22	0.24
	Associate's degree	0.06	0.08
	Bachelor's degree	0.16	0.17
	Graduate or professional degree	0.09	0.09
	Total	1.00	1.00

(a) What is the probability that a randomly chosen man has at least a Bachelor's degree?

(b) What is the probability that a randomly chosen woman has at least a Bachelor's degree?

(c) What is the probability that a man and a woman getting married both have at least a Bachelor's degree? Note any assumptions you must make to answer this question.

(d) If you made an assumption in part (c), do you think it was reasonable? If you didn't make an assumption, double check your earlier answer and then return to this part.

2.12 School absences. Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.[29]

(a) What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?

(b) What is the probability that a student chosen at random misses no more than one day?

(c) What is the probability that a student chosen at random misses at least one day?

(d) If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.

(e) If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumption you make.

(f) If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

2.13 Grade distributions. Each row in the table below is a proposed grade distribution for a class. Identify each as a valid or invalid probability distribution, and explain your reasoning.

	Grades				
	A	B	C	D	F
(a)	0.3	0.3	0.3	0.2	0.1
(b)	0	0	1	0	0
(c)	0.3	0.3	0.3	0	0
(d)	0.3	0.5	0.2	0.1	-0.1
(e)	0.2	0.4	0.2	0.1	0.1
(f)	0	-0.1	1.1	0	0

[28] **eduSex**.
[29] **Mizan:2011**.

2.14 Health coverage, frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table summarizes two variables for the respondents: health status and health coverage, which describes whether each respondent had health insurance.[30]

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health Coverage	No	459	727	854	385	99	2,524
	Yes	4,198	6,245	4,821	1,634	578	17,476
	Total	4,657	6,972	5,675	2,019	677	20,000

(a) If we draw one individual at random, what is the probability that the respondent has excellent health and doesn't have health coverage?

(b) If we draw one individual at random, what is the probability that the respondent has excellent health or doesn't have health coverage?

2.3.2 Conditional probability

2.15 Joint and conditional probabilities. P(A) = 0.3, P(B) = 0.7

(a) Can you compute P(A and B) if you only know P(A) and P(B)?

(b) Assuming that events A and B arise from independent random processes,

i. what is P(A and B)?

ii. what is P(A or B)?

iii. what is P(A|B)?

(c) If we are given that P(A and B) = 0.1, are the random variables giving rise to events A and B independent?

(d) If we are given that P(A and B) = 0.1, what is P(A|B)?

2.16 PB & J. Suppose 80% of people like peanut butter, 89% like jelly, and 78% like both. Given that a randomly sampled person likes peanut butter, what's the probability that he also likes jelly?

2.17 Global warming. A 2010 Pew Research poll asked 1,306 Americans "From what you've read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?". The table below shows the distribution of responses by party and ideology, where the counts have been replaced with relative frequencies.[31]

		Response			
		Earth is warming	Not warming	Don't Know Refuse	Total
Party and Ideology	Conservative Republican	0.11	0.20	0.02	0.33
	Mod/Lib Republican	0.06	0.06	0.01	0.13
	Mod/Cons Democrat	0.25	0.07	0.02	0.34
	Liberal Democrat	0.18	0.01	0.01	0.20
	Total	0.60	0.34	0.06	1.00

(a) Are believing that the earth is warming and being a liberal Democrat mutually exclusive?

(b) What is the probability that a randomly chosen respondent believes the earth is warming or is a liberal Democrat? **(See the next page for parts (c)-(f).)**

(c) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a liberal Democrat?

[30] **data:BRFSS2010**.

[31] **globalWarming**.

(d) What is the probability that a randomly chosen respondent believes the earth is warming given that he is a conservative Republican?

(e) Does it appear that whether or not a respondent believes the earth is warming is independent of their party and ideology? Explain your reasoning.

(f) What is the probability that a randomly chosen respondent is a moderate/liberal Republican given that he does not believe that the earth is warming?

2.18 Health coverage, relative frequencies. The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey designed to identify risk factors in the adult population and report emerging health trends. The following table displays the distribution of health status of respondents to this survey (excellent, very good, good, fair, poor) conditional on whether or not they have health insurance.

		Health Status					
		Excellent	Very good	Good	Fair	Poor	Total
Health Coverage	No	0.0230	0.0364	0.0427	0.0192	0.0050	0.1262
	Yes	0.2099	0.3123	0.2410	0.0817	0.0289	0.8738
	Total	0.2329	0.3486	0.2838	0.1009	0.0338	1.0000

(a) Are being in excellent health and having health coverage mutually exclusive?

(b) What is the probability that a randomly chosen individual has excellent health?

(c) What is the probability that a randomly chosen individual has excellent health given that he has health coverage?

(d) What is the probability that a randomly chosen individual has excellent health given that he doesn't have health coverage?

(e) Do having excellent health and having health coverage appear to be independent?

2.19 Burger preferences. A 2010 SurveyUSA poll asked 500 Los Angeles residents, "What is the best hamburger place in Southern California? Five Guys Burgers? In-N-Out Burger? Fat Burger? Tommy's Hamburgers? Umami Burger? Or somewhere else?" The distribution of responses by gender is shown below.[32]

		Gender		
		Male	Female	Total
Best hamburger place	Five Guys Burgers	5	6	11
	In-N-Out Burger	162	181	343
	Fat Burger	10	12	22
	Tommy's Hamburgers	27	27	54
	Umami Burger	5	1	6
	Other	26	20	46
	Not Sure	13	5	18
	Total	248	252	500

(a) Are being female and liking Five Guys Burgers mutually exclusive?

(b) What is the probability that a randomly chosen male likes In-N-Out the best?

(c) What is the probability that a randomly chosen female likes In-N-Out the best?

(d) What is the probability that a man and a woman who are dating both like In-N-Out the best? Note any assumption you make and evaluate whether you think that assumption is reasonable.

(e) What is the probability that a randomly chosen person likes Umami best or that person is female?

2.20 Assortative mating. Assortative mating is a nonrandom mating pattern where individuals with similar genotypes and/or phenotypes mate with one another more frequently than what would

[32] **burgers**.

be expected under a random mating pattern. Researchers studying this topic collected data on eye colors of 204 Scandinavian men and their female partners. The table below summarizes the results. For simplicity, we only include heterosexual relationships in this exercise.[33]

		Partner (female)			
		Blue	Brown	Green	Total
Self (male)	Blue	78	23	13	114
	Brown	19	23	12	54
	Green	11	9	16	36
	Total	108	55	41	204

(a) What is the probability that a randomly chosen male respondent or his partner has blue eyes?

(b) What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes?

(c) What is the probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes? What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

(d) Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning.

2.21 Drawing box plots. After an introductory statistics course, 80% of students can successfully construct box plots. Of those who can construct box plots, 86% passed, while only 65% of those students who could not construct box plots passed.

(a) Construct a tree diagram of this scenario.

(b) Calculate the probability that a student is able to construct a box plot if it is known that he passed.

2.22 Predisposition for thrombosis. A genetic test is used to determine if people have a predisposition for *thrombosis*, which is the formation of a blood clot inside a blood vessel that obstructs the flow of blood through the circulatory system. It is believed that 3% of people actually have this predisposition. The genetic test is 99% accurate if a person actually has the predisposition, meaning that the probability of a positive test result when a person actually has the predisposition is 0.99. The test is 98% accurate if a person does not have the predisposition. What is the probability that a randomly selected person who tests positive for the predisposition by the test actually has the predisposition?

2.23 HIV in Swaziland. Swaziland has the highest HIV prevalence in the world: 25.9% of this country's population is infected with HIV.[34] The ELISA test is one of the first and most accurate tests for HIV. For those who carry HIV, the ELISA test is 99.7% accurate. For those who do not carry HIV, the test is 92.6% accurate. If an individual from Swaziland has tested positive, what is the probability that he carries HIV?

2.24 Exit poll. Edison Research gathered exit poll results from several sources for the Wisconsin recall election of Scott Walker. They found that 53% of the respondents voted in favor of Scott Walker. Additionally, they estimated that of those who did vote in favor for Scott Walker, 37% had a college degree, while 44% of those who voted against Scott Walker had a college degree. Suppose we randomly sampled a person who participated in the exit poll and found that he had a college degree. What is the probability that he voted in favor of Scott Walker?[35]

2.25 It's never lupus. Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to

[33] **Laeng:2007**.
[34] **ciaFactBookHIV:2012**.
[35] **data:scott**.

a high risk of blood clotting. It is believed that 2% of the population suffer from this disease. The test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease. There is a line from the Fox television show *House* that is often used after a patient tests positive for lupus: "It's never lupus." Do you think there is truth to this statement? Use appropriate probabilities to support your answer.

2.26 Twins. About 30% of human twins are identical, and the rest are fraternal. Identical twins are necessarily the same sex – half are males and the other half are females. One-quarter of fraternal twins are both male, one-quarter both female, and one-half are mixes: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the probability that they are identical?

2.3.3 Sampling from a small population

2.27 Marbles in an urn. Imagine you have an urn containing 5 red, 3 blue, and 2 orange marbles in it.

(a) What is the probability that the first marble you draw is blue?

(b) Suppose you drew a blue marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

(c) Suppose you instead drew an orange marble in the first draw. If drawing with replacement, what is the probability of drawing a blue marble in the second draw?

(d) If drawing with replacement, what is the probability of drawing two blue marbles in a row?

(e) When drawing with replacement, are the draws independent? Explain.

2.28 Socks in a drawer. In your sock drawer you have 4 blue, 5 gray, and 3 black socks. Half asleep one morning you grab 2 socks at random and put them on. Find the probability you end up wearing

(a) 2 blue socks

(b) no gray socks

(c) at least 1 black sock

(d) a green sock

(e) matching socks

2.29 Chips in a bag. Imagine you have a bag containing 5 red, 3 blue, and 2 orange chips.

(a) Suppose you draw a chip and it is blue. If drawing without replacement, what is the probability the next is also blue?

(b) Suppose you draw a chip and it is orange, and then you draw a second chip without replacement. What is the probability this second chip is blue?

(c) If drawing without replacement, what is the probability of drawing two blue chips in a row?

(d) When drawing without replacement, are the draws independent? Explain.

2.30 Books on a bookshelf. The table below shows the distribution of books on a bookcase based on whether they are nonfiction or fiction and hardcover or paperback.

		Format		
		Hardcover	Paperback	Total
Type	Fiction	13	59	72
	Nonfiction	15	8	23
	Total	28	67	95

(a) Find the probability of drawing a hardcover book first then a paperback fiction book second when drawing without replacement.

(b) Determine the probability of drawing a fiction book first and then a hardcover book second, when drawing without replacement.

(c) Calculate the probability of the scenario in part (b), except this time complete the calculations under the scenario where the first book is placed back on the bookcase before randomly drawing the second book.

(d) The final answers to parts (b) and (c) are very similar. Explain why this is the case.

2.31 Student outfits. In a classroom with 24 students, 7 students are wearing jeans, 4 are wearing shorts, 8 are wearing skirts, and the rest are wearing leggings. If we randomly select 3 students without replacement, what is the probability that one of the selected students is wearing leggings and the other two are wearing jeans? Note that these are mutually exclusive clothing options.

2.32 The birthday problem. Suppose we pick three people at random. For each of the following questions, ignore the special case where someone might be born on February 29th, and assume that births are evenly distributed throughout the year.

(a) What is the probability that the first two people share a birthday?

(b) What is the probability that at least two people share a birthday?

2.3.4 Random variables

2.33 College smokers. At a university, 13% of students smoke.

(a) Calculate the expected number of smokers in a random sample of 100 students from this university.

(b) The university gym opens at 9 am on Saturday mornings. One Saturday morning at 8:55 am there are 27 students outside the gym waiting for it to open. Should you use the same approach from part (a) to calculate the expected number of smokers among these 27 students?

2.34 Ace of clubs wins. Consider the following card game with a well-shuffled deck of cards. If you draw a red card, you win nothing. If you get a spade, you win \$5. For any club, you win \$10 plus an extra \$20 for the ace of clubs.

(a) Create a probability model for the amount you win at this game. Also, find the expected winnings for a single game and the standard deviation of the winnings.

(b) What is the maximum amount you would be willing to pay to play this game? Explain your reasoning.

2.35 Hearts win. In a new card game, you start with a well- shuffled full deck and draw 3 cards without replacement. If you draw 3 hearts, you win \$50. If you draw 3 black cards, you win \$25. For any other draws, you win nothing.

(a) Create a probability model for the amount you win at this game, and find the expected winnings. Also compute the standard deviation of this distribution.

(b) If the game costs \$5 to play, what would be the expected value and standard deviation of the net profit (or loss)? *(Hint: profit = winnings − cost; $X - 5$)*

(c) If the game costs \$5 to play, should you play this game? Explain.

2.36 Is it worth it? Andy is always looking for ways to make money fast. Lately, he has been trying to make money by gambling. Here is the game he is considering playing: The game costs \$2 to play. He draws a card from a deck. If he gets a number card (2-10), he wins nothing. For any face card (jack, queen or king), he wins \$3. For any ace, he wins \$5, and he wins an *extra* \$20 if he draws the ace of clubs.

(a) Create a probability model and find Andy's expected profit per game.

(b) Would you recommend this game to Andy as a good way to make money? Explain.

2.37 Portfolio return. A portfolio's value increases by 18% during a financial boom and by 9% during normal times. It decreases by 12% during a recession. What is the expected return on this portfolio if each scenario is equally likely?

2.38 Baggage fees. An airline charges the following baggage fees: \$25 for the first bag and \$35 for the second. Suppose 54% of passengers have no checked luggage, 34% have one piece of checked luggage and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

(a) Build a probability model, compute the average revenue per passenger, and compute the corresponding standard deviation.

(b) About how much revenue should the airline expect for a flight of 120 passengers? With what standard deviation? Note any assumptions you make and if you think they are justified.

2.39 American roulette. The game of American roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money. Suppose you bet \$1 on red. What's the expected value and standard deviation of your winnings?

2.40 European roulette. The game of European roulette involves spinning a wheel with 37 slots: 18 red, 18 black, and 1 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. Gamblers can place bets on red or black. If the ball lands on their color, they double their money. If it lands on another color, they lose their money.

(a) Suppose you play roulette and bet \$3 on a single round. What is the expected value and standard deviation of your total winnings?

(b) Suppose you bet \$1 in three different rounds. What is the expected value and standard deviation of your total winnings?

(c) How do your answers to parts (a) and (b) compare? What does this say about the riskiness of the two games?

2.41 Cost of breakfast. Sally gets a cup of coffee and a muffin every day for breakfast from one of the many coffee shops in her neighborhood. She picks a coffee shop each morning at random and independently of previous days. The average price of a cup of coffee is $1.40 with a standard deviation of 30¢($0.30), the average price of a muffin is $2.50 with a standard deviation of 15¢, and the two prices are independent of each other.

(a) What is the mean and standard deviation of the amount she spends on breakfast daily?

(b) What is the mean and standard deviation of the amount she spends on breakfast weekly (7 days)?

2.42 Scooping ice cream. Ice cream usually comes in 1. 5 quart boxes (48 fluid ounces), and ice cream scoops hold about 2 ounces. However, there is some variability in the amount of ice cream in a box as well as the amount of ice cream scooped out. We represent the amount of ice cream in the box as X and the amount scooped out as Y. Suppose these random variables have the following means, standard deviations, and variances:

	mean	SD	variance
X	48	1	1
Y	2	0.25	0.0625

(a) An entire box of ice cream, plus 3 scoops from a second box is served at a party. How much ice cream do you expect to have been served at this party? What is the standard deviation of the amount of ice cream served?

(b) How much ice cream would you expect to be left in the box after scooping out one scoop of ice cream? That is, find the expected value of $X - Y$. What is the standard deviation of the amount left in the box?

(c) Using the context of this exercise, explain why we add variances when we subtract one random variable from another.

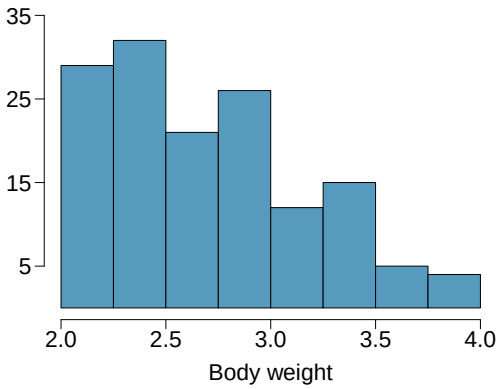
2.3.5 Continuous distributions

2.43 Cat weights. The histogram shown below represents the weights (in kg) of 47 female and 97 male cats.[36]

(a) What fraction of these cats weigh less than 2.5 kg?

(b) What fraction of these cats weigh between 2.5 and 2.75 kg?

(c) What fraction of these cats weigh between 2.75 and 3.5 kg?



[36] **cats.**

2.44 Income and gender. The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised of 59% males and 41% females.[37]

(a) Describe the distribution of total personal income.

(b) What is the probability that a randomly chosen US resident makes less than $50,000 per year?

(c) What is the probability that a randomly chosen US resident makes less than $50,000 per year and is female? Note any assumptions you make.

(d) The same data source indicates that 71.8% of females make less than $50,000 per year. Use this value to determine whether or not the assumption you made in part (c) is valid.

Income	*Total*
$1 to $9,999 or loss	2.2%
$10,000 to $14,999	4.7%
$15,000 to $24,999	15.8%
$25,000 to $34,999	18.3%
$35,000 to $49,999	21.2%
$50,000 to $64,999	13.9%
$65,000 to $74,999	5.8%
$75,000 to $99,999	8.4%
$100,000 or more	9.7%

[37] **acsIncome2005-2009**.