Contents

I Acknowledgements

II About the Manual

1 Elementary Theory of Interest
   1.1 Defining the variables
   1.2 Simple Interest
   1.3 Compound Interest
      1.3.1 Annually Compounded Interest
      1.3.2 Compounding Intervals
      1.3.3 The limit on the growth rate of an investment as a function of $n$
      1.3.4 Continuous Compounding
   1.4 The Connection between Growth Rates and Compounding
      1.4.1 Comparing Extremes: Simple and Continuous Compound Interest
   1.5 An Application of Growth Rates and Compounding
   1.6 Approximating Compound Interest by Simple interest
      1.6.1 Advanced Material: Introduction to Taylor Series
   1.7 Applications Involving the Interest Formulas
      1.7.1 Solving for the present value $P$
      1.7.2 Solving for the Annual Interest Rate $r$
      1.7.3 Solving for $t$
      1.7.4 Time to double your money and The rule of 69 and 72
      1.7.5 Effective Annual Yield
   1.8 Present value and the risk-free rate of return
   1.9 Variable interest rates
      1.9.1 Variable interest rates: discrete compounding
      1.9.2 Variable interest rates: continuous compounding
1.9.3 Continuous Rate Change-Continuous Compounding . . . . . . . 45
1.10 Variable interest rate over each compounding interval . . . . . . . . 48
  1.10.1 Geometric mean . . . . . . . . . . . . . . . . . . . . . . . . . . 49
  1.10.2 Approximating the geometric mean by the arithmetic mean . . 52
1.11 A last word on interest rates . . . . . . . . . . . . . . . . . . . . . . . 52

2 Summation notation 54
  2.1 Introduction to summation . . . . . . . . . . . . . . . . . . . . . . . . 54
  2.2 Properties of summations . . . . . . . . . . . . . . . . . . . . . . . . . 59
    2.2.1 Applications of summation properties . . . . . . . . . . . . . . . . 61
  2.3 Re-indexing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63
  2.4 Summing series . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65
    2.4.1 Applications of summing series . . . . . . . . . . . . . . . . . . . 68
  2.5 Applications of summation using spreadsheets . . . . . . . . . . . . . 70

3 Elementary set theory 73
  3.1 Introduction to sets . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
    3.1.1 Practice problems with solutions . . . . . . . . . . . . . . . . . . 78
  3.2 Intersections, unions, and complements . . . . . . . . . . . . . . . . . . 79
    3.2.1 Intersections . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79
    3.2.2 Unions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
    3.2.3 The complement of a set . . . . . . . . . . . . . . . . . . . . . . . 82
  3.3 Some basic laws . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
    3.3.1 Practice problems with basic set operations using Venn diagrams 87
    3.3.2 Solutions to practice problems . . . . . . . . . . . . . . . . . . . . 91
  3.4 The Power Set . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94
4 Counting techniques: permutations and combinations  
4.1 General multistep experiments ........................................... 95  
4.2 Permutations and combinations .......................................... 98  
   4.2.1 Permutations .......................................................... 99  
   4.2.2 Permutations of objects not all different ........................ 101  
   4.2.3 Circular permutations ............................................... 105  
   4.2.4 Combinations ....................................................... 108  
4.3 Review problems ............................................................. 110  

5 Introduction to the concept of probability  
5.1 What is probability? ......................................................... 114  
   5.1.1 The concept of relative frequencies ............................... 115  
   5.1.2 The connection between probability, mass, and Venn diagrams . 116  
5.2 Fundamental properties of probability for sample spaces ........... 117  
5.3 Finite sample spaces ....................................................... 117  
5.4 Rules for transforming words into math ................................ 120  
5.5 Some dos and don’ts when computing the probability of sets ........ 120  
5.6 Application probability problems and their solutions ............... 121  

6 Conditional probability and independence  
6.1 Introduction to conditional probability ................................. 128  
   6.1.1 The idea behind conditional probability ........................... 131  
6.2 Introduction to independent events ..................................... 134  
   6.2.1 The idea behind independence .................................... 135  
6.3 Combining independence and conditional probability ............... 136  
   6.3.1 Applications of conditional probability and independence .... 137
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Bayes’ Theorem</td>
</tr>
<tr>
<td>7.1</td>
<td>Partitioning the sample space</td>
</tr>
<tr>
<td>7.1.1</td>
<td>Tree diagram describing the law of total probability</td>
</tr>
<tr>
<td>7.2</td>
<td>Introducing Bayes’ theorem</td>
</tr>
<tr>
<td>8</td>
<td>Introduction to Random Variables</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction to finite random variables</td>
</tr>
<tr>
<td>8.2</td>
<td>Introduction to the concept of a probability mass function</td>
</tr>
<tr>
<td>8.3</td>
<td>Introduction to the expected value of a random variable</td>
</tr>
<tr>
<td>9</td>
<td>Expected value and variance</td>
</tr>
<tr>
<td>9.1</td>
<td>Introduction to the expected value of a random variable</td>
</tr>
<tr>
<td>9.2</td>
<td>Introduction to the variance and standard deviation of a random variable</td>
</tr>
<tr>
<td>9.2.1</td>
<td>A word on expected values and standard deviations for large data sets</td>
</tr>
</tbody>
</table>
10 Discrete Probability Distributions

10.1 Some basic properties of finite random variables ......... 190
10.2 Introducing the discrete random variable ................. 193
10.3 Introducing the p.m.f and c.d.f. ......................... 194
    10.3.1 Properties of the p.m.f and c.d.f. ............. 195
    10.3.2 Interpretation of the formulas ................. 197
    10.3.3 Graphing the cumulative distribution of a finite random variable . 198
    10.3.4 The range of a random variable $X$ partitions the sample space $\Omega$ . 199
10.4 Some examples of finite random variables ............... 200
    10.4.1 Computing $f_X(x)$ for the 2-dice problem .... 200
    10.4.2 Computing $F_X(x)$ for the 2-dice problem .... 201
    10.4.3 Two flips of a fair coin problem ............... 206
    10.4.4 more examples .................................. 207
10.5 Bernoulli trials ........................................ 210
10.6 The binomial distribution ................................. 211
    10.6.1 The expected value for a binomial distribution .... 213
    10.6.2 Using Excel to compute binomial distributions .... 213
10.7 Histograms in Excel ..................................... 215
11 Continuous random variables

11.1 Introduction to the concept of a continuous R.V. ........................................ 219
  11.1.1 Motivation ................................................................. 219
  11.1.2 A quick review of intervals on the real line ........................................ 221
  11.1.3 So what is a continuous random variable? ........................................ 222
11.2 Finite vs. continuous random variables ..................................................... 225
11.3 Continuous random variables ................................................................. 231
  11.3.1 The cumulative distribution and the probability density connection .... 231
  11.3.2 Computing with continuous distributions ......................................... 231
11.4 The Uniform distribution ................................................................. 235
  11.4.1 Properties of the continuous uniform random variable ....................... 236
11.5 The Exponential distribution ................................................................. 240
11.6 The Normal distribution ....................................................................... 246

12 A Simple Random Walk

A Formula summary
I Acknowledgements

Several acknowledgements are in order.

The probability component of this ebook might never have been completed were it not for a sabbatical awarded to me from Pima Community College. So I would first like to thank Pima Community College for supporting this work by offering sabbaticals to its faculty and staff! Without this sabbatical, this work would never have happened. While I’ve developed large quantities of material over winter and summer breaks on my own time, this gets old. Moreover, having a larger block of time (a summer and a semester) reduces all of the start-stop-and-restart time: it’s mid May and you have to figure out what you were on a project that you started last December - early January, so it’s a more efficient use of time. I hope the Chancellor and Board of Governors see the importance of this. And I hope the sabbatical committee stays vigilant on making sure that the sabbaticals are not turned into semester-long vacations for staff and faculty. This project has been a lot of hard work, as it should be!

Next, I would like to thank Bill Flack for proofreading this manual. I hired him to proofread my manuscript, since Bill is the best writer I know; but he went beyond the call of duty when he improved some of my figures (He’s a LaTeX picture-making master). He also suggested several key changes in the ordering of topics to make the manuscript more coherent and the topics more cohesive. Bill, what would I do without you!

Lastly, I would like to thank David Williamson and Tineke Van Zandt for all of their advice and help when I was working on my sabbatical proposal. Without a good proposal, I wouldn’t have received a sabbatical to begin with.
II About the Manual

This primer is based on a series of notes and handouts that I developed over a period of several years while teaching an experimental business math course at the University of Arizona and Pima Community College during the years of 2002, 2004-2007. Because the course was experimental, there was no textbook for the course. All students and instructors were given access to a set of PowerPoint slides that we were to follow and the students were expected to learn from. It should be no surprise to learn that this was a disaster on so many levels that it is not worth going into the details. The upshot is that my original notes for the course came about out of necessity, rather than a careful well-thought-out plan.

I have now put all of the notes I had together, covered them from MS word to LaTeX, done an entire re-write of the sections that I had, and added many new sections in order to round out the topics. Even with all of this work, this primer falls short of being a textbook on probability. Too many important topics are still missing from this text. That is, even with all of the work that I have done to date to round out this work, this text has still not shed its roots as an incomplete business math textbook.

In the future I plan to add many more topics to make this more useful to other fields. For example, material on the Poisson distribution to understand waiting times would serve both engineering students studying parts failure, and genetics students. In fact, one future goal is to add an entire chapter on population genetics. Many junior-level genetics courses serve a large population of students, many of which have no background in probability. Even with its obvious shortcoming, this primer would be an ideal introduction to probability for such students because it explains, in detail, the concept (and motivation) behind conditional probability. It also does quite a bit with Bayes’ theorem.

Look for future updates to this book. It is my hope that Pima Community College will continue to support this work.
1 Elementary Theory of Interest

Interest is the additional money that a borrower pays a lender for the use of the lender’s money. It is often expressed as a percentage of the sum borrowed. The amount of interest depends, among other things, on the amount of money borrowed, and on the time that it takes for the borrower to repay the loan.

1.1 Defining the variables

Let’s define three variables that we’ll use throughout our study of interest.

We will use $P$ to denote the principal: the amount of money borrowed, or the present value of the money invested. The interest rate will be $r$; and $r$ must be in decimal form for computation—if the interest rate is 15%, then $r = 0.15$. For time, we will use $t$; this will almost always be in years. If the time is not given in years, then it must be converted into years. For example, if $t = 15$ months, then it must be converted to years: $t = 1.25$ years.

The named, or quoted, interest rate $r$ is referred to as the nominal interest rate, or the annual interest rate. This is typically the advertised rate that banks and loan offices use: get a loan at only 8% interest! However, knowing the annual interest rate, in and of itself, holds limited meaning. To compute the actual amount of interest that you will have to pay requires knowing the compounding scheme. That is, to really predict how an investment will grow with time, you must know the pair: the interest rate $r$, and the compounding scheme. Therefore, it is useful to think of these as an order pair:

$$(r, \text{Compounding scheme}).$$

We now give a formal definition of simple and compound interest. The rest of the chapter will be devoted to giving you a deeper understanding of this definition.

Definition 1. In simple interest, only the original principal earns interest. In compound interest, the accumulated interest payments are added to the original principal and earn interest along with it.
1.2 Simple Interest

The formula for simple interest is: $I = Prt$, where $I$ represents the amount of interest earned after $t$ years. If you invest an amount $P$ at a rate of $r$ and earn simple interest for $t$ years, then at the end of that time the value of your account is $F = P + I$: principal plus interest. Combining all the terms gives the mathematical expression

$$F = P + I = P + Prt = P(1 + rt).$$

**Example 1.** Suppose that you want to invest 1000 dollars at 5% interest for 2 years. After one year you earn 5% of 1000 dollars: namely, 50 dollars. You withdraw the interest and bury it in your backyard, leaving only the original 1000 dollars to earn interest in the second year. At the end of the second year you earn another 50 dollars. At the end of 2 years, you've earned a total of $2 \times 50$ dollars in interest. Here’s a time map of your earnings

$$1000 \mapsto 1000 + .05(1000) \quad \text{(at the end of year 1)}$$

$$\mapsto 1000 + .05(1000) \quad \text{(first year earnings)} + .05(1000) \quad \text{(second year earnings)}$$

$$= 1000 + 2(.05(1000)) = 1000(1 + 2(.05)) \quad \text{(total earnings at the end of year 2)}$$

The general formula for simple interest is

$$F = P(1 + rt), \quad (1.1)$$

where $P =$ initial principal, $F =$ future value, $r =$ annual interest rate, and $t =$ number of years. Typically, $P$ and $r$ are known, and the future value is a function of time (measured in years). In our example, you invest $1000 at a simple interest rate of 5%. Then the future value as a function of time is $F(t) = 1000(1 + .05t)$. This is a linear function, since it fits the form $y = ax + b$. In this example, the domain was $t = 0, 1, 2, \ldots$. However, we can extend the formula to remain valid for all time $t \in \mathbb{R}$ by simply agreeing that for any time in between $t = 0, 1, 2, \ldots$ we will take the value of the investment to be worth the amount $F$ given by the equation (1.1).

**Note:** This assumes that interest is paid uniformly over any fractional period of time. Many accounts only pay interest in discrete payments at the end of fixed time periods. You only get interest for full periods; if you close the account partway through a period, you don’t get any interest for that fraction of a period. Suppose, for example, that a bank pays interest quarterly (with a period of 1/4 year). The value of the account is shown on the graph:
Example 2. You borrow $10,000.00 on January 1, 2000; the simple interest rate is 7%.
(a) What amount must be repaid on January 1, 2004?
(b) On that date, what is the total interest on the loan?

Solution: (a) Given \( P = 10,000 \), \( r = 0.07 \), and \( t = 4 \), we have

\[
F(4) = P(1 + r \cdot 4) = 10,000(1 + (0.07)(4)) = 12,800.
\]

(b) \( I = F - P = 12,800 - 10,000 = 2800 \).

We can also ask: “How much money should be invested today in order to have a certain amount at a given future date?”

Example 3. How much should an investor deposit in the bank now in order to have $1000 in the bank half a year from now? The simple interest rate is 8%.

Solution: Given \( F = 1000 \) (this is the future value that we want), \( r = 0.08 \), and \( t = 1/2 \), we want to find \( P \). Solving \( F = P(1 + rt) \) for \( P \) gives

\[
P = \frac{F}{1 + rt} = \frac{1000}{1 + (0.08)(1/2)} = \frac{1000}{1 + 4/100} = \frac{1000}{1.04} \approx 961.54.
\]
1.3 Compound Interest

1.3.1 Annually Compounded Interest

Again, let $P$ denote the principal (initial investment), and $r$ denote the annual interest rate. Suppose the interest is compounded annually: at the end of every year, that year’s interest is added to the principal to earn interest in future years.

After one year,

$$F = P(1 + r).$$

After two years,

$$F = [P(1 + r)] (1 + r) = P(1 + r)^2.$$

After three years,

$$F = [P(1 + r)^2] (1 + r) = P(1 + r)^3.$$

Continuing in this way, after $t$ years,

$$F = P(1 + r)^t,$$

where $t = 0, 1, 2, \ldots$. Notice that over each compounding period, the interest is determined from the simple interest formula applied to the principal over the time period in between the compoundings. Thus, the building block for the annual compound interest formula is the simple interest formula.

Although we have only derived the formula for integer values of the domain $t = 0, 1, 2, \ldots$, we can extend the formula to be valid for all values of time $t$. Thus, this is valid over any time period, including a fractional time period. How can we do this? We simply define it to be valid for all time $t$. This should not be too surprising, since the formula is valid for all $t$-values, and so long as all of the investors all agree, we can take the value of the formula to be the value of the investment. But this formula is far from random. It has the built-in structure that it compounds your money over evenly spaced time periods—in the case of annual interest, a year.
Example 4. Going back to example (1), suppose we had not taken the 50 dollars in interest out at the end of the first year. Then over the second year, those 50 dollars would have earned interest along with the 1000-dollar initial principal.

\[
1000 \mapsto 1000 + .05(1000) = 1000 + 50 \quad \text{(at the end of year 1)} \\
\mapsto (1000 + 50) + .05(1000 + 50) \\
= 1000(1 + .05)(1 + .05) = 1000(1 + .05)^2 \quad \text{(at the end of year 2)}
\]

The general formula for annually compounded interest is

\[
F = P(1 + r)^t.
\]

Typically \(P\) and \(r\) are known, and the future value is a function of time (measured in years). Paralleling the previous example, suppose you want to invest $1000 at 5% compounded annually. Then the future value as a function of time is \(F(t) = 1000(1+.05)^t\). This is a not a linear function, since it does not fits the form \(y = ax + b\). By default, it is a nonlinear function.

Example 5. (Annually Compounded Interest) Invest $1 at 10% interest, compounded annually, over 30 years.

Solution: Given: \(P = 1\), \(r = 0.1\), and \(t = 30\). Find \(F\).

\[
F = P(1 + r)^t \\
= (1.1)^{30} \\
= 17.4494
\]

The net increase was roughly 1645%! Now suppose that instead of investing $1, you had invested $10,000. What would your investment be worth? Since the relationship between \(F\) and \(P\) is linear, we can just multiply our previous result by 10,000 to get \(F = 10,000(17.4494) = 174,494\) to the nearest dollar.

### 1.3.2 Compounding Intervals

Interest can be compounded more often than once per year. New interest might be added to the old principal at the end of every quarter, or every month, or even weekly or daily.

Let \(n\) denote the number of times per year that interest is compounded (number of conversion periods in one year), \(P\) denote the principal (initial investment), and \(r\) denote the annual interest rate (assumed constant). For example if \(n = 2\), then the interest is compounded twice a year. Below is a list of common discrete compounding periods and their corresponding names.

- \(n = 1\) compounded annually
• \( n = 2 \) compounded semi-annually
• \( n = 4 \) compounded quarterly
• \( n = 12 \) compounded monthly
• \( n = 52 \) compounded weekly
• \( n = 365 \) compounded daily

Suppose the interest is compounded \( n \) times a year: at the end of every period of time \( 1/n \) years, that period’s interest is added to the principal to earn interest in future periods. Thus we have divided the year into \( n \) intervals, each with a duration of \( 1/n \) years. We then apply the simple interest formula to the principal over each subinterval (period). The interest earned over one period is \( I = Prt = Pr \cdot \frac{1}{n} = P \cdot \frac{r}{n} \).

After one period \((t = 1/n \) years\),

\[
F = P \left( 1 + r \cdot \frac{1}{n} \right) = P \left( 1 + \frac{r}{n} \right).
\]

After two periods,

\[
F = \left[ P \left( 1 + \frac{r}{n} \right) \right] \left( 1 + \frac{r}{n} \right) = P \left( 1 + \frac{r}{n} \right)^2.
\]

After three periods,

\[
F = \left[ P \left( 1 + \frac{r}{n} \right)^2 \right] \left( 1 + \frac{r}{n} \right) = P \left( 1 + \frac{r}{n} \right)^3.
\]

After \( n \) periods (one year), the principal is compounded \( n \) times. The final value is then

\[
F = P \left( 1 + \frac{r}{n} \right) \left( 1 + \frac{r}{n} \right) \ldots \left( 1 + \frac{r}{n} \right) \quad (n \text{ terms})
= P \left( 1 + \frac{r}{n} \right)^n.
\]

After two years, the final value is

\[
F = \left[ P \left( 1 + \frac{r}{n} \right)^n \right] \left( 1 + \frac{r}{n} \right)^n = P \left( 1 + \frac{r}{n} \right)^{2n}.
\]
After $t$ years, the final value is

$$F = P \left(1 + \frac{r}{n}\right)^{nt},$$

(1.2)

where $t$ need not be an integer. Again, we have derived the formula for values of $t = \frac{m}{n}$, where $m, n \in \mathbb{N}$, and we extend the formula, via definition, to be valid for all time $t$.

**WARNING:** $t$ must always be given in years.
Example 6. (Weekly Compounded Interest) Invest $10,000 at 10% interest, compounded weekly, over 1 and 30 years.

Solution: Given: $P = 10,000, r = 0.1, n = 52, \text{ and } t = 1, 30.$ Find $F$.

After one year $(t = 1)$,

$$F = 10,000 \left(1 + \frac{0.1}{52}\right)^{52 \cdot 1} = \$11,050.65$$

After thirty years $(t = 30)$,

$$F = 10,000 \left(1 + \frac{0.1}{52}\right)^{52 \cdot 30} = \$200,277.6$$

Compare this last result with the result in example 5. That example had the identical scenario, except that the interest was compounded annually instead of weekly.

In summary, the only difference between compounding $n$ times a year and once a year is that for $n > 1$, the compounding is occurring more frequently. Just as with annual interest the simple interest formula is applied to the principal over each of the $n$ compounding periods. Thus, just as in the case of annual interest, the simple interest formula is a fundamental building block for the discrete compound interest formula.

The frequency of compounding plays an important role in the growth of the principal. Given a fixed principal $P$, a fixed annual interest rate $r$, and a fixed time interval $t$, we can think of the future value $F$ in equation (1.2) as a function of $n$, the number of compounding periods per year. The rate at which your investment grows depends on how many times per year the interest is compounded. It turns out that the more compounding per year that the investment is subjected to, the greater the rate of return on the investment. But as we shall soon see, there is a limit on this rate of return.
1.3.3 The limit on the growth rate of an investment as a function of $n$

Using Calculus it can be shown that if $n, m \in \mathbb{N}$ and $m < n$, then

$$\left(1 + \frac{1}{m}\right)^m < \left(1 + \frac{1}{n}\right)^n.$$  

This inequality proves that

$$f(m) = \left(1 + \frac{1}{m}\right)^m$$

is an increasing function of $m$. Below are four graphs of $f(m)$ confirming this claim. In the top two graphs we plot $f(m)$ as a discrete function of $m$, where $m = 1, 2, 3, \ldots, 10$ (upper leftmost graph), and $m = 1, 2, \ldots, 100$ (upper rightmost graph). In the bottom two graphs, we plot $f(m)$ as a continuous function of $m$ over the intervals $[0, 10]$, $[0, 100]$ respectively. Notice that in all of the pictures, the graphs are strictly increasing.

![Graphs of $f(m)$](image)

Figure 1: Notice that as $m \to \infty$, the graphs all seem to be strictly increasing and approaching the same limit. The limit is $e \approx 2.71$.

Let $m = n/r$ and fixing $r > 0$, $P > 0$, and $t > 0$. It follows immediately that

$$F(n) = P \left(1 + \frac{r}{n}\right)^{nt}$$

is an increasing function of $n$ (the number of compounding periods). This is verified in the graph below. We have plotted the future value of $P = 1$ dollar invested at $r = .05$
over a time period of 100 years for the compounding periods $n = 1, 4, 12$. We have also included the graph of simple and continuous compound interest for purposes of comparison.

![Figure 2](image.jpg)

**Figure 2:** Various outcomes for the future value of $\$1$ at a 5% annual interest rate $r$. The simple interest is the linear graph that exhibits the least amount of growth. The graphs of the discrete compounding with $n = 1, 4$ are barely distinguishable even over a one hundred year period of growth. The difference between the $n = 12$ graph and continuous compounding graph cannot be resolved on this vertical scale over the "short" time period of 100 years!

![Figure 3](image.jpg)

**Figure 3:** A closeup of the graphs corresponding to $n = 1, 4, 12, 365$ and continuous compounding. Even in this extreme closeup, the graphs of continuous compound interest and the discrete compound interest with $n = 365$ are indistinguishable!

When graphs become so close together it is sometimes useful to plot the final values in a table for an easier comparison. Below is a table that gives the future value of $P = 1$ dollar that is invested at an annual interest rate of $r = 5\%$ for $t = 100$ years as a function of the rate of compounding.
For $n > 1$, the final value $F$ is larger than it would be for simple interest at
the same annual interest rate $r$. In fact, equation (1.2) is an increasing function of $n$. Hence,
more frequent compounding leads to increased yield. While this fact is difficult to show
mathematically, intuitively, the reason for it is simple: the sooner you start earning
interest on interest, the faster your investment grows. This increment has a limit as the
number of times of compounding (per year) increases indefinitely. For a fixed annual
interest rate $r$, as $n \to \infty$, the factor
\[
\left( 1 + \frac{r}{n} \right)^n
\]
increases to the limit $e^r$. Thus as $n$ gets very large, the compound interest formula
becomes
\[
F = Pe^{rt},
\]
where $e := 2.7182818\ldots$ is the exponential growth constant. Here we have used the fact
that
\[
\lim_{m \to \infty} \left( 1 + \frac{1}{m} \right)^m = e,
\]
where $m = n/r$. In the homework you are asked to verify this using MS Excel.

**Advanced Comment 1:** To prove this, let $x = 1/m$ and notice that
\[
\lim_{m \to \infty} \left( 1 + \frac{1}{m} \right)^m = \lim_{x \to 0^+} (1 + x)^{\frac{1}{x}},
\]
\[
\text{let } x = \frac{1}{m} \quad \Rightarrow \quad (1 + x)^{\frac{1}{x}} = \exp \left[ \frac{\ln(1+x)}{x} \right],
\]
and
\[
\frac{\ln(1+x)}{x} \to 1 \quad \text{as } x \to 0.
\]
The result now follows.
Advanced Comment 2: Notice that in both the simple and compound interest formulas, there is a linear relationship between \(F\) and \(P\). However, in the case of simple interest, the relationship between \(F\) and \(r\) and \(F\) and \(n\) is linear. In the case of compound interest, the relationship between \(F\) and \(r\) and \(F\) and \(n\) is nonlinear. This results from the fact that with compound interest, you earn interest on interest, which is not the case with simple interest. The result is that your investment under compound interest grows exponentially rather than linearly. The downside of this is that your credit card debt also grows exponentially rather than linearly.

Example 7. (Continuous Compounded Interest) Invest $10,000 at 10% interest, compounded weekly, over 1 and 30 years.

Solution: Given: \(P = 10,000\), \(r = 0.1\) and \(t = 1, 30\). Find \(F\).

After one year \((t = 1)\),

\[
F = 10,000e^{(0.1)1} = $11,051.71
\]

After thirty years \((t = 30)\),

\[
F = 10,000e^{(0.1)30} = $200,855.4
\]

Compare this last result with the result in example 6. That example had the identical scenario, except that the interest was compounded weekly instead of continuously.


1.4 The Connection between Growth Rates and Compounding

Compound interest is like a snowball rolling down a hill. The faster the snowball rolls, the more revolutions per unit of time it goes through, hence the more snow it picks up, hence the faster it grows. Similarly, the faster you overturn the interest, the more compounding periods per year, the more money you accumulate. There is limit as to how many revolutions per unit of time that a snowball can obtain without centripetal forces tearing it a part. Just as with the snowball, there is limit to the rate of growth under compounding. It is the rate corresponding to continuous compounding.

We have seen that if \( P, r, \) and \( t \) are fixed, then simple interest has the slowest growth rate. This growth rate is linear. Next, we looked at compound interest that exhibits nonlinear growth. We found that the more compounding periods per year, the faster the investment grows (it is an increasing function of \( n \)). This increased growth reaches a plateau in the limit as \( n \to \infty \). It is known as continuous compounding. We can now paint a ‘mathematical picture’ of the growth rates. Let \( P, r, \) and \( t \) be fixed. We then have the following inequality relations between the various growth rates as a function of compounding periods:

\[
P(1 + rt) < P \left(1 + \frac{r}{n}\right)^{nt} < P \left(1 + \frac{r}{m}\right)^{mt} < Pe^{rt},
\]

(1.4)

where \( n, m \in \mathbb{N} \) and \( n < m \). For example, the return on an investment of \( P \) dollars at \( r\% \) annual interest over \( t \) years will grow at various rates depending on the compounding period. The inequality between the returns that result from simple, discrete compounding with \( n = 1, 4, 12, 365 \) and continuous compounding on these investments gives a natural ordering between the various investment schemes, as shown below

\[
P(1 + rt) < P (1 + r)^t < P \left(1 + \frac{r}{4}\right)^{4t} < P \left(1 + \frac{r}{12}\right)^{12t} < P \left(1 + \frac{r}{365}\right)^{365t} < Pe^{rt}.
\]

This inequality is confirmed in table 1 for \( P = 1, \ r = .05, \) and \( t = 100 \). From these inequalities and from figures 2 and 3, it should now be clear that simple and continuous compound interest act like an envelope for discrete compounding, with the simple-interest graph being the lower bound for all \( n \) and the continuous-compound-interest graph being the upper bound for all \( n \). Thus, for fixed \( P, r, \) and \( t, \) the simple interest graph exhibits the least amount of growth. It is below all the discrete compounding graphs. The smallest discrete compounding graph is the annually discrete compounded interest corresponding to \( n = 1, \) next the semi-annually discrete compounded interest corresponding to \( n = 2, \) next the quarterly discrete compounded interest corresponding to \( n = 4, \) next the monthly \( n = 12, \) then the yearly \( n = 365, \) and finally the continuous compound interest graph that sits above all of the discrete compounding graphs.
1.4.1 Comparing Extremes: Simple and Continuous Compound Interest

Below is a graph that shows the difference between the growth rate for simple and compound interest over long time periods. We have chosen to compare the future value formula for simple interest, the slowest-growing interest scheme, to the future value formula for continuous compound interest, the fastest-growing interest scheme, to emphasize the maximum disparity in the rate of growth of an investment under the two schemes. While the difference in the future value of one dollar between the two schemes is quite striking over long times, notice that the graphs of the two functions are pretty close for about the first 20 years. This a little misleading since the principal is only one dollar. With larger principals you would see a significant difference over shorter time scales. However, this too is misleading. No matter what the value of $P$ and $r$, the two graphs would be approximately equal over some finite time interval. Moreover, if we examine the earnings ratio $F/P$ then we would find that the two graphs would be close for $rt$ small (much less than 1). It is this fact that we shall further explore in subsection 1.6.

![Graph showing comparison between simple and continuous compound interest growth]

Figure 4: A comparison between the investment of one dollar under simple interest (linear growth) and the investment of one dollar under continuous compound interest (exponential growth). The interest rate is 5%, the time is 100 years.
1.5 An Application of Growth Rates and Compounding

It is now time to put our knowledge about growth rates as a function of compounding to use. Understanding of the connection between rate of growth and the rate of compounding can give you quick-and-dirty approximate answers for multiple-choice test questions.

Suppose you’re taking a business college entrance exam. The exam consists of 20 problems with a time limit of 40 minutes. You are only given a 4-button calculator. You can only spend an average of two minutes on each problem. How can you quickly determine the correct solutions to problems involving compound interest?

By a 4-button calculator, we mean one that can only add, subtract, multiply, and divide. There are no function buttons like the square root button or the exponential button. Let us digress to discuss some examples of calculations that you can and can’t do with a 4-button calculator.

With a 4-button calculator, you can easily compute \( F = P(1 + rt) \) given \( P \), \( r \), and \( t \). To compute \( rt \), we multiply; to compute \( 1 + rt \), we add; and finally, to compute \( P(1 + rt) \), we multiply again. Thus the simple-interest formula is 4-button calculator friendly.

**Example 8.** What is the future value of $100,000 invested at a simple interest rate of 4%, after 5 years? Round your answer to whole dollars? Your choices are

(a) $100,000  
(b) $95,000  
(c) $120,000  
(d) $122,140

**Solution:** \( F = 100,000(1 + .04)5 = 120,000 \). This calculation was straightforward. The answer is (c).

It’s worth pointing out that you could compute a term of the form \( x^n \), if \( n \) is an integer. Recall the definition of the integer exponent is

\[ x^n = x \cdot x \cdots x \quad (x \text{ is multiplied by itself } n \text{ times}). \]

For example, \( 5^3 = 5 \cdot 5 \cdot 5 \).

On the other hand, you can’t calculate square roots like \( 5^{1/2} \). Remember, you don’t have a square root button. How do we multiply 5 by itself 1/2 times? What would this interpretation of a rational exponent even mean? It is worth pointing out that there are mathematical algorithms for computing the square root of a number to any desired
accuracy, but these are beyond the scope of our discussion.

Another limitation is that you cannot directly compute $e^{rt}$ on a 4-button calculator, because you can only add, subtract, multiply, and divide. Both the square root button and the exponential button on your calculator use advanced mathematical algorithms to compute $\sqrt{x}$ and $e^x$ for a given value of $x$, and chances are that you don’t know these algorithms. Moreover, you don’t have time to carry out a lengthy calculation, even if you do know how to do it. You only have two minutes per problem!

**Example 9.** What is the future value of $100,000 invested at a continuously compounded annual rate of 4%, after 5 years? Round your answer to whole dollars? Your choices are

(a) $100,000  
(b) $95,000  
(c) $120,000  
(d) $122,140  

*Solution:* Let’s start by trying a direct approach. We’re given $P$, $r$, and $t$, and we’ll try substituting them into the formula for continuous compound interest. Using only $\{+,-,\times,\div\}$, we can get as far as $F = Pe^{rt} = 100,000e^{.04 \cdot 5} = 100,000e^{.2}$. Without a key for $e$, we could try to approximate $e$ by 2.718 and substitute this value into the expression to get $100,000 \cdot (2.718)^{0.2}$. O.K., now what? How are you going to compute $(2.718)^{0.2}$ using only $\{+,-,\times,\div\}$? Notice that the exponent is 0.2, not 2. We’ve hit a dead end.

The correct approach to this problem is to first notice that we cannot determine the exact solution using a 4-button calculator. We must find some indirect way to find the solution. What qualitative properties do we know about continuous compound interest? One thing that we have learned in this section is that for a fixed annual interest rate $r$, a fixed principal $P$, and a fixed time $t$, the growth rates are increasing functions of $n$. In particular, an investment earning 4% interest, compounded continuously, grows much faster than an investment earning 4% simple interest. The future value of $P = 100,000$ invested at a simple interest rate of $r = .04$ over $t = 5$ years is a quantity that we can compute. In fact, we’ve already computed it in example 8. It was found to be $F = 120,000$. Looking at our choices (a)-(d) we see that it must be (d) since this is the only value greater than the future value under a simple interest rate.

It is worth mentioning that, without doing any calculations at all, we can immediately rule out possibilities (a) and (b). Choice (a) is the value of the principal, and choice (b) is less than the principal.
1.6 Approximating Compound Interest by Simple interest

The following subsection is advanced material. This material maybe skipped without jeopardizing your understanding of future material.

We are now going to explore the conditions under which compound interest can be approximated by simple interest. You should be very interested in this subject, since simple interest is 4-button calculator friendly.

Referring to figure 4, we see that we could approximate the future value of the $1 investment, as a function of time, under continuous compound interest by the simple interest investment for times less than 20 years. To gain a deeper understanding of this fact, let us examine the two formulas side by side. Since the principal $P is just a scaling factor, we will examine the resulting equations for the earnings ratio $F/P instead of just $F$. Moreover, the earnings ratio is a more accurate measure of the growth rate.

\[
\frac{F}{P} = e^{rt} \quad \text{(continuous compound interest)} \quad \frac{F}{P} = 1 + rt \quad \text{(simple interest)}
\]

let \( x = rt \)

\[
\to \quad \frac{F}{P} = e^{x} \quad \text{(continuous compound interest)} \quad \frac{F}{P} = 1 + x \quad \text{(simple interest)}
\]

It is clear that we can approximate continuous compound interest by simple interest, with a small relative error (defined as \( |F - P|/P \)), so long as we can approximate \( e^{x} \) by \( 1 + x \). Let us start by examining the graphs of these two functions.

![Graph](image)

Figure 5: A comparison between the graph of \( y = e^{x} \) and \( y = 1 + x \)

Notice that the graphs are approximately equal from \( x = 0 \) to about \( x = 0.2 \). In practical terms, this means that we can use the simple interest formula for \( F/P \) to approximate the continuous compound interest formula for \( F/P \) over the interval \( 0 \leq x \leq .2 \). Since
We can also approximate both discrete and continuous compound interest by the simple interest formula algebraically. Let’s start with the discrete compounding case first. If we compound the interest $n$ times in one year, then over the first interval $[0, \frac{1}{n}]$ the discrete formula $F = P(1+r/n)^nt = P(1+r/n)$, which is the simple interest formula. The results are exact. Over the first and second interval $[0, \frac{2}{n}]$, the investment has grown to

$$F = P \left(1 + \frac{r}{n}\right)^2 = P \left(1 + \frac{r}{n}\right) \left(1 + \frac{r}{n}\right) = P \left(1 + 2\frac{r}{n} + \left(\frac{r}{n}\right)^2\right) \approx P \left(1 + 2\frac{r}{n}\right),$$

where you should recognize the last expression as the simple interest formula with $t = 2/n$. The approximation is valid so long as $P(r/n)^2$ is small. Now if $r/n$ is small (much less than one), then $(r/n)^2$ is even smaller. Since most real-world $r$’s are between 1% and 18%, so $0.01 < r < 0.18$, for $n \geq 1$ we see that the ratio $r/n$ will be small, hence the approximation will be valid provided that $P$ is not so large as to off-set the smallness of $(r/n)^2$.

**Question:** What do we mean by small error? That is, can the difference of two values be small if it is greater than one?

**Answer:** Two values can be ”close” even if the error between them is much greater than one! It all depends on what we mean by close.

Keeping the initial principal in the approximation can lead to unnecessary complication. To simplify the analysis we will only focus on approximating the earnings ratio $F/P$. While this leads to certain limitations, it has the advantage of simplicity. Moreover, the error that is usually most relevant between simple interest and compound interest is the relative error, not the absolute error. If we let $F_1(n) = P(1 + r/n)^nt$ and $F_2 = P(1 + rt)$, then the relative error between these two functions is $(F_1 - F_2)/F_2$. For example, if the error in approximation is $\$1000$ is this a large error, or a small error? It really depends on the size of the investments that are being considered. If we are investing $P = 1,000,000$ dollars, then the error is small. If we are investing $P = 1,000$, then the error is large. The relative error can be expressed in terms of the relative earnings.

$$\frac{F_1 - F_2}{F_2} = \frac{F_1 - F_2}{F_2} \cdot \frac{F_2}{F_2}.$$

Thus, it suffices to work with earnings ratio $F/P$ in place of $F$. 

$x = rt$, we have $0 \leq rt \leq .2$ as the region of validity for the approximation. Thus the larger $r$ is, the smaller the time interval over which the approximation is valid. Conversely, the smaller $r$ is, the larger the time interval over which the approximation is valid.
Let us approximate the earnings ratio over the interval \([0, \frac{3}{n}]\).

\[
\frac{F}{P} = \left(1 + \frac{r}{n}\right)^3
\]

\[
= \left(1 + \frac{r}{n}\right) \left(1 + \frac{r}{n}\right) \left(1 + \frac{r}{n}\right)
\]

\[
= \left(1 + 3\frac{r}{n} + 3 \left(\frac{r}{n}\right)^2 + \left(\frac{r}{n}\right)^3\right) \approx \left(1 + r\frac{3}{n}\right),
\]

The approximation is valid so long as \(3(r/n)^2 + (r/n)^3\) is small. The smallness of this expression is related to the smallness of \(r/n\).

The next obvious question is: over what time intervals is this approximation valid? If we looked over longer and longer time intervals, then the approximation of compound interest by simple interest must break down. This fact is evident in the graph 4. To answer this question for discrete and compound interest we will need the use of a mathematical approximation tool known as a Taylor series.

It is a well-known fact in mathematics that certain functions \(f(x)\) can be expressed as an infinite series of the form

\[
f(x) = \sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots.
\]

Such a series is referred to as a Taylor series about \(x = 0\). This is a special case of a more general series known as power series. A few comments are in order:

1. The above formula may not be valid for every \(x\) in the domain of \(f\).

2. There is a formula for the \(a_n\) that depend on certain properties of the function \(f\) that are beyond the scope of this book.

3. In practice you do not sum all of the terms, of which there are infinitely many. Instead, you approximate the function by only a few terms in the series. The sum of these few terms is a polynomial. There is an amazing theorem in mathematics, known as Weierstrass’s Approximation Theorem, that says that any continuous function on a closed interval \([a, b]\) can be uniformly approximated, with arbitrary accuracy, by a polynomial. Given this piece of information, it should not be too surprising that certain continuous functions can be approximated by the first few sums of their Taylor series, since this sum is a polynomial.

Accepting these claims as true, we will now introduce two Taylor series approximations about \(x = 0\) for two functions that you are familiar with.
The Taylor series for the first function that we wish to approximate is given by

\[(1 + x)^\alpha = 1 + \alpha x + \frac{1}{2} \alpha (1 - \alpha) x^2 + \cdots.\]  

(1.5)

This series is valid for \(|x| < 1\). If we let \(x = \frac{r}{n}\) and \(\alpha = nt\), and assume that \(\frac{r}{n}\) is small, then we can apply this approximation to show that

\[(1 + \frac{r}{n})^{nt} \approx (1 + rt),\]  

(1.6)

from which it follows that

\[F = P \left(1 + \frac{r}{n}\right)^{nt} \approx P(1 + rt).\]

This says that for \(\frac{r}{n}\) small enough and for fixed time, we can approximate the discrete compound interest formula by the simple interest formula. **Warning:** this approximation assumes that \(\alpha = nt\) is fixed. It is clearly not valid for all time, since \(F = P \left(1 + \frac{r}{n}\right)^{nt}\) grows exponentially, whereas \(F = P(1 + rt)\) grows linearly.

The Taylor series for the second function that we wish to approximate is given by

\[e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{1}{2} x^2 + \frac{1}{6} x^3 + \frac{1}{24} x^4 + \frac{1}{120} x^5 + \cdots,\]  

(1.7)

where \(n! = 1 \cdot 2 \cdot 3 \cdots n\) is known as \(n\)-factorial. Unlike the first series, this series is valid for all values of \(x\). However, for large values of \(x\) you must add many terms from the series in order for the series to converge to the correct value of \(e^x\). But we are not interested in evaluating the series for large \(x\). We want to just use the first couple of terms of the series to approximate \(e^x\). This means that the truncated series will only be valid for small values of \(x\).

It is a mathematical fact that the first few terms of this series approximate \(e^x\) well for \(|x|\) much smaller than one. If we let \(x = rt\) then using (1.7) and keeping only the first two terms in the approximation we arrive:

\[e^{rt} \approx 1 + rt.\]  

(1.8)

We will find the approximations in (1.6) and (1.8) very useful in approximating the effective annual yield in section 1.7.5.

Perhaps there is no better way to build one’s intuition than to see Taylor series approximations in action. We now show two sets of graphs that compare the exact solution, in this case the graph of \(y = e^x\), to a series of approximations: \(y_1 = 1 + x\), \(y_2 = 1 + x + \frac{1}{2} x^2\), \(y_3 = 1 + x + \frac{1}{2} x^2 + \frac{1}{6} x^3\), \(y_4 = 1 + x + \frac{1}{2} x^2 + \frac{1}{6} x^3 + \frac{1}{24} x^4\), and
\[ y_5 = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \frac{1}{120}x^5. \] The first graph shows the series of approximations over the interval \([0, 1]\), and the second graph shows the series of approximations over the interval \([0, 5]\). The Taylor series about \(x = 0\) means that any truncated series, such as \(y_1\) through \(y_5\), will be most accurate for \(x\) near zero. In particular, this means that the approximations should be more accurate on the interval \([0, 1]\) than on the interval \([0, 5]\).

![Graph](image)

Figure 6: A comparison between the function \(y = e^x\) and five of its Taylor series approximations: \(y_1, y_2, y_3, y_4, y_5\). The graph on the left is a comparison between the exact and approximate functions over the interval \([0, 1]\), whereas the graph on the right is a comparison over the interval \([0, 5]\). Notice that over the interval \([0, 1]\) you cannot even see the difference between \(y_4, y_5\) and \(e^x\) since there are only 4 visible graphs, but over the longer interval \([0, 5]\) you can now see all 6 graphs and the difference becomes apparent.

### 1.7 Applications Involving the Interest Formulas

We can compute any term, or variable, in the compound interest formulas in terms of the others. We now carry out an exhaustive derivation of each possibility with examples.

#### 1.7.1 Solving for the present value \(P\)

If we know the final value \(F\), the interest rate \(r\), the amount of time \(t\), and the compounding frequency \(n\), then we can determine the principal/present value \(P\).

**Discrete compounding:** Start with equation (1.2). To solve for \(P\), divide both sides of the equation by \((1 + r/n)^{nt}\)

\[
F = P \left(1 + \frac{r}{n}\right)^{nt} \implies P = \frac{F}{\left(1 + \frac{r}{n}\right)^{nt}}
\]
or

$$P = F \left(1 + \frac{r}{n}\right)^{-nt}$$

**Continuous compounding:** Start with equation (1.3). To solve for $P$, divide both sides of the equation by $e^{rt}$

$$F = Pe^{rt}$$

$$\implies P = Fe^{-rt}$$

**Example 10.** (Solving for $P$) Given:

$F = \$1000, \quad r = 0.05, \quad n = 365, \quad t = 10$

What is the principal investment required to get $\$1000 in this case?

**Solution:**

$$P = \$1000 \left(1 + \frac{0.05}{365}\right)^{-365 \cdot 10} = \$607 \quad \text{(to the nearest dollar)}$$

### 1.7.2 Solving for the Annual Interest Rate $r$

**Discrete compounding:** Start with equation (1.2):

$$F = P \left(1 + \frac{r}{n}\right)^{nt}.$$

Next, divide by $P$, and raise both sides of the equation to the power $1/(nt)$

$$\left(\frac{F}{P}\right)^{\frac{1}{nt}} = 1 + \frac{r}{n}.$$

Lastly, solve for $r$

$$r = n \left[ \left(\frac{F}{P}\right)^{\frac{1}{nt}} - 1 \right].$$

From here forward we will sometimes use a shorthand notation. We use a right pointing arrow $\rightarrow$ to denote “next step”, and over the right arrow will be the operation that we perform on both sides of the equation. For example,

$$F = Pe^{rt} \quad \rightarrow \quad \frac{F}{P} = e^{rt}.$$
means that in this step we divide both sides of the equation by $P$.

$$\ln\left(\frac{F}{P}\right) = \ln(e^{rt}) = rt$$

means that in this step we take the logarithm of both sides of the equation. The rest should be self-explanatory. The goal of using this notation is to make the steps seem less wordy, and the presentation more clear.

**Continuous compounding:** Start with equation (1.3).

\[ F = Pe^{rt} \quad \rightarrow \quad \frac{F}{P} = e^{rt} \]

\[ \ln\left(\frac{F}{P}\right) = \ln(e^{rt}) = rt \]

\[ \frac{1}{t} \ln\left(\frac{F}{P}\right) = \frac{rt}{t} \rightarrow r = \frac{1}{t} \ln\left(\frac{F}{P}\right) \] (1.9)

Notice that if we had attempted to use the natural logarithm to solve for $r$ in the case of discrete compounding, then we would be stuck. This is because in formula (1.2), the $r$ is not in the exponent.

**Example 11.** (Doubling an investment) An investment of $P$ dollars doubled in six years under continuous compounding. Assuming the interest rate was constant over this time, what was the interest rate?

**Solution:** Set $F = 2P$. Then equation (1.3) becomes

\[ 2P = Pe^{rt} \]

\[ \rightarrow \quad 2 = e^{rt} \]

\[ \ln 2 = \ln(e^{rt}) = rt = 6r \]

\[ \frac{1}{t} \ln\left(\frac{F}{P}\right) = \frac{rt}{t} \rightarrow r = \frac{1}{6} \ln\left(\frac{2}{1}\right) = 0.1155 = 11.55\% \]

Notice that the value $P$ of the initial investment dropped out of the equation. Thus the result is valid for any initial investment $P$.

**Example 12.** Suppose $100,000 is invested at a certain rate compounded quarterly, and becomes $150,000 after 30 years. What was the interest rate?

**Solution:** Given:

\[ F = 150,000, \quad P = 100,000, \quad n = 4, \quad t = 30. \]
\[ F = P \left(1 + \frac{r}{n}\right)^{nt} \]

$150,000 = 100,000 \left(1 + \frac{r}{4}\right)^{4 \cdot 30}$

\[
\frac{3}{2} = \left(1 + \frac{r}{4}\right)^{120}
\]

Now, take the 120th root of both sides.

\[
\left(\frac{3}{2}\right)^{\frac{1}{120}} = 1 + \frac{r}{4}
\]

\[
r = 4 \left(1.5^{\frac{1}{120}} - 1\right) = 0.013538... = 1.35\%
\]

### 1.7.3 Solving for \( t \)

**Discrete compounding:** Start with equation (1.2).

\[ F = P \left(1 + \frac{r}{n}\right)^{nt} \]

Divide both sides of the equation by \( P \) and take the natural log of both sides of the equation (since the variable we wish to solve for is in the exponent). This yields

\[ \ln \left(\frac{F}{P}\right) = \ln \left(\left(1 + \frac{r}{n}\right)^{nt}\right) \]

Next, using the property of logarithms \( \ln(a^b) = b \ln(a) \), we arrive at the solution

\[
t = \frac{\ln \left(\frac{F}{P}\right)}{n \ln \left(1 + \frac{r}{n}\right)}.
\]  \hspace{1cm} (1.10)

**Example 13.** (Increasing an investment \( m \)-fold: discrete case) We can use the above formula to answer the question: How long does it take to increase an investment from \( P \) dollars into \( F = mP \) dollars at an interest rate of \( r\% \)?

**Solution:** Setting \( F = mP \) in the above formula yields

\[
t = \frac{\ln m}{n \ln \left(1 + \frac{r}{n}\right)}.
\]
Example 14. (Tripling an investment: discrete case) A sum of money is invested at an interest rate of 6%, compounded monthly. How many years does it take for it to triple?

Solution: Using the above formula, we get

\[ t = \frac{\ln 3}{n \ln \left(1 + \frac{r}{n}\right)} = \frac{\ln 3}{12 \ln(1 + 0.005)} = 18.356 \text{ years} \]

Continuous compounding: Start with equation (1.3).

\[ F = Pe^{rt} \quad \frac{F}{P} = e^{rt} \quad \ln \left(\frac{F}{P}\right) = \ln[e^{rt}] = rt \]

Thus,

\[ t = \frac{1}{r} \ln \left(\frac{F}{P}\right) . \]

(1.11)

Example 15. (Increasing an investment \(m\)-fold: continuous case) We can use this formula to answer the question: How long does it take to increase an investment from \(P\) dollars to \(F = mP\) dollars at an interest rate of \(r\)?

Solution: Setting \(F = mP\) in the preceding formula yields

\[ t = \frac{\ln (m)}{r} . \]

(1.12)

Example 16. (Tripling an investment: continuous case) A sum of money is invested at an interest rate of 6%, compounded continuously. How many years does it take for it to triple?

Solution: Using the above formula, we get

\[ t = \frac{\ln 3}{0.06} = 18.310 \]

Compare this with the result found in example 14. Notice that the time is shorter in the continuous case. Does this make sense?
1.7.4 Time to double your money and The rule of 69 and 72

All things being equal, the faster the growth rate, the faster you can double your money. Investors sometimes use “the rule of 72” to do a back-of-the-envelope calculation for the time it takes for your investment to double at a given interest rate. The rule of 72 says that the time it takes your money to double is roughly \( \frac{72}{r} \), where \( r \) is the annual interest rate.

We have now seen that the precise value of the doubling time \( t_{\text{doubling}} \) is

\[
t_{\text{doubling}} = \frac{\ln 2}{n \ln(1 + \frac{r}{n})} \quad \text{(for the discrete case)},
\]

\[
t_{\text{doubling}} = \frac{\ln 2}{r} \quad \text{(for the continuous case)},
\]

where \( \ln 2 \approx 0.6931471806 \).

**Question:** Where does the 72 come from?

**Answer:** I’m not sure!

In the case of discrete compounding, the rule 72 seems to give reasonable answers over a large range of interest rates. You can test the rule of 72 using Excel. Make one column be the exact value and the second column be the approximation \( \frac{72}{r} \). You should experiment with the parameters \( n \) and \( r \) (see rule-of-72.xls).

However, for the case of continuously compounded interest it is clear that it should be the “rule of 69”! The error between the continuous case and the approximation is

\[
\frac{(.72 - \ln 2)}{r} \approx \frac{(.72 - .69)}{r} = .03.
\]

Notice that this error is fixed and much larger than the error would be if we just used 69 instead of 72. In fact, if we compare the rule 72 to the rule of 69 over \( r \) values ranging from .01 to .20, incrementing them by a step size of \( \Delta r = .01 \) (i.e. \{.01, .02, ..., .19, .20\}) and over the values \( n = 1, 4, 12, 365 \) corresponding to yearly, quarterly, monthly, and daily, we find that the average absolute error in the rule of 72 is approximately .41, where as the error in the rule of 69 is only .17 (see rule-of-72.xls). I vote that we use the rule of 69. Moreover, the rule of 69 is almost exact for the continuous case.

**Rule of Thumb:** When it comes to choosing when to use the rule of 72 or the rule of 69 my advice is to use the rule of 69 in the case of compound interest, and discrete compounding with \( n \geq 12 \). Use the rule of 72 for all other cases.
Advanced Comment: To take this analysis further would require using mathematical analysis. The next step would require developing an approximation based on the fact that $r/n$ is a small number and so we can expand the exact solution in a Taylor series in the small parameter $r/n$ about $r/n = 0$. For $r/n$ small, it can be shown that

$$\frac{\ln m}{n \ln (1 + \frac{r}{n})} \approx \frac{\ln(m)}{r} \left(1 + \frac{r}{2n} + O\left(\frac{r}{n}\right)^2\right).$$

If $r/n = .08$ and $m = 2$ then $\ln 2 \cdot 1.04 \approx .72$. Since many interest rates are in the range $r = .08n$, this might be why the rule of 72 is used so much in practice.
1.7.5 Effective Annual Yield

Because compound interest is affected by both the nominal interest rate $r$ and the frequency of compounding $n$, it is sometimes hard to tell offhand which of two compound interest procedures is more advantageous. For example, is it better to invest at 4% compounded monthly or $4\frac{1}{4}\%$ compounded semiannually? It would be nice to have a common ground on which to compare various interest rates. The solution was to invent the effective annual yield, or annual percentage yield (APY), which we’ll call $y$. It is defined as follows:

$$ P \left( 1 + \frac{r}{n} \right)^n = F = P(1 + y). $$

Using the formula for simple interest $F = P(1 + yt)$, $y$ is the rate that would give the same future value as $F = P(1 + r/n)^{rt}$ for $t = 1$ year and with principal $P$. In words, if an investment earns compound interest, then the effective annual yield is the simple interest rate that yields the same amount at the end of one year.

Solving for $y$ in equation (1.13) gives

$$ y = \left( 1 + \frac{r}{n} \right)^n - 1. $$

The APY $y$ is the actual percentage rate earned. It is sometimes just called the yield. Since most people understand the simple interest formula $I = Prt$, the annual yield has the appeal that it is accessible to the common person. In fact, the Truth in Lending Law enacted in 1969 requires the annual yield $y$ to appear on all contracts.

For continuous compounding, we can derive the effective annual yield as follows:
First, equate the two expressions for $F$, then divide by $P$, and finally solve for the yield $y$

$$ Pe^r = F = P(1 + y) $$
$$ \frac{\div P}{\div 1} \Rightarrow e^r = 1 + y $$
$$ \frac{-1}{\times n} \Rightarrow y = e^r - 1 $$

**NOTE:** The effective annual yield is independent of the principal in both the discrete and continuous case.

Given a certain yield $y$, suppose we want to find the annual rate $r$ for the case of discrete compounding. This would require solving (1.13) for $r$.

$$ \frac{\div P}{\div 1} \Rightarrow \left( 1 + \frac{r}{n} \right)^n = 1 + y $$
$$ \frac{\left( \frac{1}{n} \right)}{\times n} \Rightarrow \left( 1 + \frac{r}{n} \right) = (1 + y)^{\frac{1}{n}} $$
$$ \frac{-1}{\times n} \Rightarrow r = n \left[ (1 + y)^{\frac{1}{n}} - 1 \right] $$
Example 17. What annual rate $r$, compounded quarterly, has a yield of 7.5%?

Solution: Given $n = 4$ and $y = .075$.

$$r = 4[(1 + .075)^{1/4} - 1] = .073.$$

Example 18. What annual rate $r$, compounded continuously, has a yield of 7.5%?

Solution: Given continuous compounding and $y = .075$.

$$r = \ln(y + 1) = \ln(1.75) = .072.$$

We end this section with the derivation of several useful formulas for approximating the yield. These formulas will then be applied to an example problem that is similar to a test question on a multiple-choice 20-question 40-minute exam where you only have a 4-button calculator, or no calculator at all.

Recall that in section 1.6, we gave formulas for approximating $(1 + x)^n$ and $e^x$ by two-term polynomials related to the simple interest formula. Referring to approximation (1.6) with $t = 1$, we have

$$\left(1 + \frac{r}{n}\right)^n \approx 1 + r \quad \xrightarrow{-1} \quad y = \left(1 + \frac{r}{n}\right)^n - 1 \approx r. \quad (1.16)$$

and referring to approximation (1.8) with $t = 1$ we have

$$e^r \approx 1 + r \quad \xrightarrow{-1} \quad y = e^r - 1 \approx r. \quad (1.17)$$

These results show that for small values of $r$ (say $r < 0.2$), $y \approx r$ regardless of whether we’re looking at the yield from discrete compounding or continuous compounding. This is a very powerful piece of information! We will now make immediate use of this valuable knowledge.

Example 19. Given the annual interest rate $r$, determine which of the following investment opportunities will have the highest effective annual yield. Your choices are

(a) 5% compounded daily
(b) 5% compounded continuously
(c) 5% compounded quarterly
(d) 10% compounded yearly
Solution: Since all or the $r$-values are much less than 20%, we can apply our approximation. If we use our approximation $y \approx r$, then in (a)-(c) $y \approx 0.05$, whereas in (d) $y \approx 0.1$. Thus, the answer is (d).

Notice that of the choices (a)-(c), (b) will have the largest yield since all of the $r$-values are the same and continuous compounding produces a larger growth rate than discrete compounding.

**Example 20.** Given the annual interest rate $r$, determine which of the following investment opportunities will have the lowest effective annual yield. Your choices are

(a) 7% Compounded daily       (b) 7% Compounded continuously

(c) 7% Compounded quarterly    (d) 6% Compounded continuously

Solution: Since all or the $r$-values are much less than 20% we can apply our approximation. If we use our approximation $y \approx r$, then in (a)-(c) $y \approx 0.07$, whereas in (d) $y \approx 0.06$. Thus, the answer is (d). Our approximation is safely accurate to 1%.

Notice that of the choices (a)-(c), (c) will have the smallest yield, since all of the $r$-values are the same and discrete quarterly compounding grows at a slower rate than continuous and daily compounding.

How do we know when it’s O.K. to use the approximation and how do we know when it’s when it’s too close to call? The answer to this question is technical. It requires comparing the next term in the series to the difference in the annual interest rates $r$ for each answer. An easier approach to this is to argue as follows: The error as a function of $r$ in approximating the yield $y$ corresponding to continuous compound interest is $E_{\text{continuous}}(r) = y_{\text{continuous}} - r = e^r - 1 - r$. The error as a function of $r$ in approximating the yield corresponding to discrete compounding is $E_{\text{discrete}}(r) = y_{\text{discrete}} - r = (1 + r/n)^n - 1 - r$. Since the continuous compound interest grows faster than discrete compounding, it follows that the error in approximating the yield from continuous compound interest $E_{\text{continuous}}(r)$ is greater than the error in approximating the yield from discrete compounding $E_{\text{discrete}}(r)$. Thus we will plot the error function corresponding to continuous compound interest and use it to decide our cutoff value for the differences in the $r$ values that are given in the multiple choice answers. Since the error corresponding to the discrete case is smaller than that for the continuous case, this analysis will hold for the discrete case as well. We are taking an approach that it is better to err on the side of caution than to produce the wrong answer. After all, we can also find a new approach to getting a wrong answer.
Using the graph of the error, we can devise a rule of thumb for this situation. So long as the \( r \) values differ by no more than 2\%, and are in the range of 0 to 20\%, you should be able to approximate the yield by the annual interest rate \( r \). If the \( r \) values in the multiple-choice answers are between 0 and .1 (10\%), then the error is less than 1\%. If the gap is tightened further, then we must refer to the graph of the error. But this is not very practical on an exam.

Using this information, we see that our approximation in example 20 is sound so long as we remember that if the \( r \) values are between 0 and .01, then the error is less than 1\%. For \(.06 \leq r \leq .07\) the approximation has an error much less than .005, or .5\%. However, in the next example the numbers are too close to call.

**Example 21.** Given the annual interest rate \( r \), determine which of the following investment opportunities will have the lowest effective annual yield. Your choices are

(a) 7\% Compounded daily  
(b) 7\% Compounded annually

(c) 7\% Compounded quarterly  
(d) 6.99\% Compounded continuously

We cannot use our method on this problem because the difference between .0700 and .0699 is .0001 and this is much less than .005. Even in this case we can throw out (a) and (c), since if all the \( r \) values are equal, then an investment with an annually compounded interest grows slower than an investment with quarterly or daily compounded interest. The answer is either (b) or (d). But without keeping many more terms in the approximation, or having the use of a scientific calculator, we cannot be sure which one is the correct answer. But do not despair! The good news is that the test makers cannot give you a question like this if you are only armed with a 4-button calculator.
1.8 Present value and the risk-free rate of return

[ADD: define risk-free rate of return]

Sometimes we want to know the amount \( P \) that we should invest today in order for our investment to be worth \( F \) at a particular time in the future. For this reason \( P \) is sometimes called the *present value* and \( F \) is called the *future value*. For example, suppose a Treasury bill will be worth 1000 dollars one year from now. The interest rate on the T-bill is 5% compounded continuously. What is a fair price for this T-bill? We’ve already seen how to answer such a question.

**Example 22.** Suppose your house burned down and you wanted to buy another house. You found a house that you’d like to buy, but the sellers will not be able to move out for 6 months. They have another potential buyer, so they demand that you pay for the house today. You estimate that the house will be worth $200,000 (the price of houses in a new neighborhood nearby). If the interest rates on Treasury bills are 5% compounded continuously, what is a fair price for the house?

**Solution:** \( P = Fe^{-rt} = 200000e^{-(.05)\cdot 6} = 200000e^{-0.25} = $195,062 \).

When you buy a US Treasury bill that matures in a few months, you’re making a risk-free investment. The amount you earn on this investment is the risk-free return. The interest rate that the T-Bill offers is known as the risk-free rate of return. By subtracting the risk-free return from the return on an investment that has the potential to lose value, you can figure out the risk premium, which is one measure of the risk of choosing an investment other than the short term T-bill.

Since the value of any T-bill is backed by the US Government (which also guarantees the worth of the dollar), it is considered a risk-free investment. The only practical danger in buying a long-term T-bill is that the guaranteed rate of return may not keep up with inflation. Over short terms like 6 months, the typical life of an option, inflation isn’t likely to change that much; so a T-bill is a good measure of what your money would be worth 6 months from now if you didn’t want to take any risks. The key word here is *risk*. You cannot make a large return on your money without incurring some level of risk. The more risk in the investment, the more money (paid in terms of interest) that you will demand. Remember, companies that want you to invest in their stocks must make a bid for your money, and they do that by paying higher interest rates. If I can make a safe investment in a fixed-index fund like the S&P 500 and earn roughly an 8% return, why would I invest in your high-risk stock if you were only offering a return of 8% as well? The answer is: I wouldn’t! You would have to offer me a much higher interest rate to entice me to buy your stock.
Many people think they make money in the stock market because they have a knack for picking good stocks. Except in rare instances, I disagree with this philosophy. The reason people make money in the stock market is because the stock market must on average pay a higher rate of return than safer investments like T-bills, bonds, and banks. If the average rate of return on an investment was low, then only a very few people (the type that like to gamble) would invest in the stock market. Most investors would leave their money in the bank.

1.9 Variable interest rates

The following subsection is advanced material. This material may be skipped without jeopardizing your understanding of future material.

So far, we’ve assumed that the simple interest rate \( r \) remains constant over the life of an investment.

That’s a valid assumption for many forms of investment. For example, most CD’s and bonds pay interest at a fixed rate.

On the other hand, many investments do not. There are CD’s whose rates change to reflect inflation. Adjustable-rate mortgages vary their rates, generally in step with the prime lending rate, the LIBOR, or some other measure of the current market rate. Money-market accounts pay at a different rate almost every month. We would like to know how to calculate and compare the yields of such investments.

1.9.1 Variable interest rates: discrete compounding

Let’s begin with a fairly simple example. We’ll assume you’ve paid $10,000 for a six-month CD paying simple interest at 5%. At the end of the six months, you let the CD renew automatically. When this happens, the bank can change the rate: in our case, the rate goes up to 6% for the second six months. What will the value of the CD be at the end of the year?

We’ll look at this in steps. At the end of the first six months, the value of the CD is

\[
F = P(1 + rt) = 10,000 \left(1 + \frac{0.05}{2}\right) = 10,250
\]
For the second six-month period, the principal $P$ is $10,250. Applying the simple-interest formula again, we get

$$F = P(1 + rt) = 10,250 \left(1 + \frac{0.06}{2}\right) = 10,000 \left(1 + \frac{0.05}{2}\right) \left(1 + \frac{0.06}{2}\right) = 10,557.50$$ \hspace{1cm} (1.18)

We can generalize this. Suppose we’ve got an investment that compounds $n$ times per year, with a different simple-interest rate in each compounding interval: in the $i$th interval, the rate is $r_i$. We start with a principal of $P$. At the end of the first interval, the simple-interest formula gives us

$$F_1 = P \left(1 + \frac{r_1}{n}\right)$$

For the second compounding interval, the principal is $F_1$, the value of the account at the beginning of the interval. The simple-interest formula gives us

$$F_2 = F_1 \left(1 + \frac{r_2}{n}\right) = P \left(1 + \frac{r_1}{n}\right) \left(1 + \frac{r_2}{n}\right)$$

For the third interval, the principal is $F_2$, and we get

$$F_3 = F_2 \left(1 + \frac{r_3}{n}\right) = P \left(1 + \frac{r_1}{n}\right) \left(1 + \frac{r_2}{n}\right) \left(1 + \frac{r_3}{n}\right)$$

We can keep doing this. For the $m$th interval, the principal is $F_{m-1}$, and the simple-interest formula gives us

$$F_m = F_{m-1} \left(1 + \frac{r_m}{n}\right) = P \left(1 + \frac{r_1}{n}\right) \left(1 + \frac{r_2}{n}\right) \cdots \left(1 + \frac{r_m}{n}\right) \hspace{1cm} (1.19)$$

Is there an easier way to do this? For example, couldn’t we just use the average of the different interest rates? Let’s look back at our first example, on page 41. There, we had two six-month periods. The simple-interest rate during the first period was 5%, and during the second period it was 6%. The average of those two is 5.5%; suppose we just used 5.5% for the whole year?

First of all, we can’t use 5.5% simple interest for the whole year. That would give us a year-end value of

$$F = 10,000(1 + 0.055) = 10,550$$

The actual value we found in Equation (1.18) was $10,557.50; so the easy way doesn’t work.
Suppose we try a slightly more complicated trick? Let’s use that average of 5.5% and compound it semi-annually. Then we’d get

\[ F = 10,000 \left( 1 + \frac{0.055}{2} \right)^2 = 10,557.56 \]

This is close, but it’s still wrong. The only correct approach to this situation is Equation (1.19).

**Example 23.** You invest $43,000 in a money-market account that compounds monthly. For the first month, the average yield is 3.72%; for the second, it’s 3.91%; for the third, it’s 4.00%; and for the fourth month, it’s 4.19%. What is the value of your account at the end of the fourth month?

**Solution:** Use Equation (1.19). Since you’re compounding monthly, \( n = 12 \).

\[
F = 43,000 \left( 1 + \frac{0.0372}{12} \right) \left( 1 + \frac{0.0391}{12} \right) \left( 1 + \frac{0.0400}{12} \right) \left( 1 + \frac{0.0419}{12} \right)
= 43,569.69
\]

1.9.2 **Variable interest rates: continuous compounding**

Now, let’s modify our first example. You’re again buying a $10,000 CD at 5% for six months, then renewing it at 6% for another six months; but this time, the interest is continuously compounded.

We’ll use the continuous-compounding formula, Equation (1.3), to get the value of the account after the first six months:

\[ F = Pe^{rt} = 10,000 e^{0.05/2} = 10,253.15 \]

Again, for the second six months, the principal is the value at the end of the first six months; and the value of the account at the end of the year is

\[ F = 10,253.15 e^{0.06/2} = 10,000 e^{0.05/2} e^{0.06/2} = 10,565.41 \]

Remember the property of exponents:

\[ e^a e^b = e^{a+b} \]

That lets us write

\[ F = 10,000 \exp \left( \frac{0.05 + 0.06}{2} \right) \]
(The notation “exp” allows us to write exponentials without tiny hard-to-read exponents. We can write \(\exp(x)\) instead of \(e^x\) to make our equation easier to read.)

Notice that in this case, we can use the average of the interest rates. Of course, that depends on the two periods being of equal length. There’s no reason why that has to be so.

Suppose we have continuously compounding interest, with a rate that changes with time. For the first \(t_1\) years, the rate is \(r_1\); for the next \(t_2\) years, it’s \(r_2\); and so on up to the last \(t_m\) years, for which it’s \(r_m\). The initial principal is \(P\).

At the end of the first \(t_1\) years, Equation (1.3) gives us the value of the account:

\[
F_1 = Pe^{r_1 t_1} = P \exp(r_1 t_1)
\]

As with discrete compounding, the principal for the next \(t_2\) years is \(F_1\), and the value at the end of \(t_1 + t_2\) years is

\[
F_2 = F_1 e^{r_2 t_2} = P e^{r_1 t_1} e^{r_2 t_2} = P \exp(r_1 t_1 + r_2 t_2)
\]

We can keep going like this until we get

\[
F_m = F_{m-1} e^{r_m t_m} = P e^{r_1 t_1} e^{r_2 t_2} \cdots e^{r_m t_m} = P \exp(r_1 t_1 + r_2 t_2 + \cdots + r_m t_m) = P \exp \left( \sum_{i=1}^{m} r_i t_i \right),
\]

where the total time interval over which the investment grows is \(t_1 + \cdots + t_m\) years.

Notice that all of these variable interest rate formulas, although messy, are defined recursively. It is for this reason that they can easily be coded into Excel, or some other programming language. By using a standard nested loop, you could input the time intervals \(t_i\) and the interest rates over the time intervals to arrive at a recursive (repeating) algorithm to compute the \(F_i\)’s.

Let’s take this analysis a little further. To have a complete understanding of what I am about to present, the reader will need to have a solid understanding of calculus. However, even those without calculus in their backgrounds should be able to glean some useful information from the following discussion.
1.9.3 Continuous Rate Change-Continuous Compounding

Suppose we want to compute the earnings of an investment of $P$ dollars over $T$ years. Furthermore suppose that the interest scheme is compounded continuously and that the interest rate changes over very small time scales that need not be of equal length. This case is more subtle than the previous cases, but it is not even close to how much more complicated things can get. For example, the analysis can get much worse if we treat the interest rate as a random variable. To solve this case we will need to take the limit as the time intervals shrink to zero. This case is addressed in calculus. It requires the use of the Riemann integral.

We begin by partitioning the time interval $[0, T]$ (time is given in years) into $n$ subintervals $[t_{i-1}, t_i]$ of length $\Delta t_i = t_i - t_{i-1}$, where $i = 1, 2, \ldots, n$, the length of the interval is denoted by $\Delta t_i$, and $t_0 = 0$ and $t_n = T$. That is, we want to cut up the original interval into little disjoint pieces.

We assume that we partition the interval into so many small partitions that the length of the largest partition, denoted by $\Delta t_{\text{max}}$, is made arbitrarily small. For example, if we take the interval $[0, 1]$ (one year) and divide the interval into $n$ subintervals of equal length, then the width of each subinterval is $\Delta t = \frac{1-0}{n} = \frac{1}{n}$. For larger and larger values of $n$ (more and more partitions of the original interval), the expression $\frac{1}{n}$ goes to zero and so the width of the subintervals goes to zero. In this case there was not a unique largest subinterval, since all of the subintervals were of equal length; however, it should be clear that if we subdivide the original interval in such a way that the maximum subinterval gets smaller and smaller, then all of the subintervals are getting smaller and smaller.

Notice that by the construction process of our subintervals, that even though the width of each subinterval gets smaller and smaller, the sum of the widths of all of the subintervals taken as a whole is the length of the total interval $T - 0$.

We will now outline the details of such a construction. Let $t_i^*$ be any time value with the property that $t_{i-1} < t_i^* < t_i$. Then as we increase the number of partitions, $t_i^*$ approaches $t_i$, and so $r(t_i^*)$ is a close approximation to the function $r(t)$ over this interval. The area under the curve $y = r(t)$ over the interval $t_{i-1} < t_i^* < t_i$ can be approximated by the rectangle with base $\Delta t_i$ and height $r(t_i^*)$. The future value is given by

$$F_n = F_{n-1} e^{r(t_i^*)\Delta t_n} = Pe^{r(t_1^*)\Delta t_1} e^{r(t_2^*)\Delta t_2} \ldots e^{r(t_n^*)\Delta t_n}$$

$$= P \exp(r(t_1^*)\Delta t_1 + r(t_2^*)\Delta t_2 + \cdots + r(t_n^*)\Delta t_n)$$

$$= P \exp\left(\sum_{i=1}^{n} r(t_i^*)\Delta t_i\right).$$
Taking the limit as the number of partitions become finer and finer (i.e. \( n \to \infty \) and \( \Delta t_{\text{max}} \to 0 \)) will yield the Riemann integral:

\[
\lim_{n \to \infty} \sum_{i=1}^{n} r(t^*_i) \Delta t_i = \int_{0}^{T} r(t) \, dt .
\]

It is important to remember that this is a process where to each new value of \( n \) we repartition the interval \([0, T]\). As the process continues the Riemann sums approach the Riemann integral. The above expression for the future value then becomes:

\[
F(T) = P \exp \left( \int_{0}^{T} r(t) \, dt \right) .
\] (1.20)

By the mean-value theorem, there exists a time \( t_* \):

\[
r(t_*) = \frac{1}{T} \int_{0}^{T} r(t) \, dt .
\]

If we define \( \bar{r} = r(t_*) \), then we may rewrite the equation above equation as

\[
F(T) = P e^{\bar{r}T} ,
\] (1.21)

where \( \bar{r} = \frac{1}{T} \int_{0}^{T} r(t) \, dt \) is the value of the interest rate averaged over the interval. Notice that this expression formally looks like the classic ”Pert” formula for continuous compound interest: \( F = P e^{rt} \).

Through a limiting process we have derived a formula for the case of continuous compounding with a variable interest rate \( r(t) \) that changes continuously. It has many similarities with the discrete case under continuous compounding, and this is no accident since it was derived as the limit of the discrete case.

We now summarize what we have done. If we define the annual interest rate \( \bar{r} \) to be the average value of the continuously changing interest rate \( r(t) \) over the interval \([0, t]\) for any \( t \in \mathbb{R} \), then

\[
\bar{r} = \frac{1}{t} \int_{0}^{t} r(s) \, ds
\]

then substituting for \( \bar{r} \) in equation (1.3) yields

\[
F = P e^{\bar{r}t} = P e^{\left( \frac{1}{t} \int_{0}^{t} \bar{r}(s) \, ds \right) t} = P e^{\int_{0}^{t} r(s) \, ds} .
\]

We can understand this formula as the limit of the discrete case. It is the end result of a process of generating a Riemann integral from a Riemann sum in the limit as \( n \to \infty \) together with a proper repartitioning of the interval \([0, t]\).
For such a situation we would probably not want to let $t$ get too large, say $t = 1$ year, since a long-time average defined this way would probably not hold much practical meaning. It would be better to use a moving average to examine long term trends.

Fortunately for you, this course does not concern itself with variable interest rates. From here forward, we will assume a constant interest rate for all time. This assumption is reasonable for Bonds, Bills and cds, but not for stocks and money market accounts.
1.10 Variable interest rate over each compounding interval

We begin by partitioning the time interval \([0, T]\) (time is given in years) into \(n\) subintervals \([t_{i-1}, t_i]\) of equal length, where \(i = 1, 2, \ldots, n\). The width of each subinterval, denoted by \(\Delta t\), is the length of the total interval \(T - 0\) divided by the number of times that we want to partition it \(n\). Thus \(\Delta t = T/n\). Then \(t_i = t_{i-1} + \Delta t\). Using \(t_1 = T/n\) and recursion, we find \(t_i = \frac{i}{n} T\) for \(i = 1, \ldots, n\). As a check, notice that \(\Delta t = t_i - t_{i-1} = \frac{T}{n}\).

Next we compute the earnings ratio over each subinterval \([t_{i-1}, t_i]\). Since the rate is constant over each subinterval, this will lead to the following results:

The future value over the \(i^{th}\) period is given by

\[
F(t_i) = F(t_{i-1}) (1 + r_i \Delta t) = F(t_{i-1}) (1 + Tr_i/n)
\]

if discrete compounding is used over the interval, and

\[
F(t_i) = F(t_{i-1}) e^{r_i \Delta t} = F(t_{i-1}) e^{Tr_i/n}
\]

if continuous compounding is used over the interval. The earnings ratios become:

\[
\frac{F(t_i)}{F(t_{i-1})} = (1 + Tr_i/n) \quad \text{(discrete compounding)}
\]

\[
\frac{F(t_i)}{F(t_{i-1})} = e^{Tr_i/n} \quad \text{(discrete compounding)}
\]

**Discrete Rate Change-Discrete Compounding:**

\[
F = P \left( 1 + T \frac{r_1}{n} \right) \left( 1 + T \frac{r_2}{n} \right) \ldots \left( 1 + T \frac{r_n}{n} \right) \quad \text{(n factors)}.
\]

(1.22)

If

\[ r = r_1 = r_2 = \cdots = r_n \]

then the interest rate is constant and for \(T = 1\) formula (1.2) is valid.

As we’ve already seen in the example above, to define the nominal interest rate \(r\) in terms of the arithmetic-average

\[
r = \frac{\sum_{i=1}^{n} r_i}{n} = \sum_{i=1}^{n} \frac{r_i}{n}
\]

is not vary useful in practice, and in general leads to an over estimate of the effective annual interest rate. The reason for this over estimate will be given shortly. Perhaps a better approach is the set the expression in 1.22 equal to an equivalent expression with
constant interest

**Discrete Rate Change-Continuous Compounding:** Suppose that over each of the \( n \) intervals that the interest is compounded continuously. The earnings ratio over the \( i^{th} \) period is \( F/P = e^{r_i \Delta t_i} = e^{Tr_i/n} \) since the duration of this time interval is \( \Delta t_i = T/n \). Then the earnings ratio at the end of the \( n \) periods becomes:

\[
\frac{F}{P} = e^{\frac{T}{n}r_1}e^{\frac{T}{n}r_2} \cdots e^{\frac{T}{n}r_n} = e^{\frac{T}{n}r_1+\frac{T}{n}r_2+\cdots+\frac{T}{n}r_n} = \exp\left(T \sum_{i=1}^{n} \frac{r_i}{n}\right).
\]

If we define the **annual arithmetic-average interest rate** as

\[
r = \frac{\sum_{i=1}^{n} r_i}{n} = \frac{\sum_{i=1}^{n} r_i}{n}.
\]

Then the resulting mathematically equivalent expression superficially resembles the standard compound interest formula

\[
\frac{F}{P} = e^{rT}.
\]

### 1.10.1 Geometric mean

Let’s start with a problem:

**Problem:** John and Mary each invested $10,000 for two years, compounded annually. John’s investment paid interest of 10% for the first year, and no interest for the second. Mary’s investment paid 5% in the first year, and 5% in the second. Which one made the better investment?

If you weren’t taking Business Math, your answer would probably be something like, “The average yield of both investments is 5%, so neither investment was better than the other.” However, as a Business Math student, you distrust this easy answer.

You’re right to be distrustful. Let’s look at the changes in both investments over the two years.

John’s investment yields 10% over the first year. At the end of the first year, the value is

\[
F = 10,000(1 + 0.10) = 11,000
\]
In the second year, his investment pays no interest at all; so the value at the end of the second year is still $11,000.

Meanwhile, Mary’s investment yields 5% in the first year; so at the end of the first year, it’s worth

\[ F = 10,000(1 + 0.05) = 10,500 \]

In the second year, it also yields 5%, so at the end of two years its value is

\[ F = 10,500(1 + 0.05) = 11,025 \]

Obviously, Mary’s investment was better than John’s, since she’s got $25 more than he does at the end of the two years.

It would be a good thing if we had some kind of “average” yield of investments like John’s, so that we could compare them with Mary’s 5% average yield. Fortunately, we have such a thing. We’re looking for a constant rate \( r \) that, over the two-year life of the investment, produces the same final value as the 10% yield in the first year and 0% in the second. In other words:

\[
(1 + r)^2 = (1 + 0.10)(1 + 0.00) = 1.10
\]

\[
1 + r = (1.10)^{1/2} = 1.0488
\]

\[
r = 4.88\%
\]

Thus John’s investment was equivalent to one that yielded 4.88% for both years. Since Mary’s investment yielded 5% for both years, hers was the better one.

An “average” like we’ve just described is called the geometric mean. We can extend the definition of it for an arbitrary number of years. Suppose we hold a certain investment for \( n \) years. In the first year, it produces a return of \( r_1 \); in the second year, of \( r_2 \); . . . ; in the \( n \)th year, it returns at a rate of \( r_n \). Then after \( n \) years, its value is

\[ F = P(1 + r_1)(1 + r_2) \cdots (1 + r_n) \]

We want to find a constant rate \( r \) that will produce the same future value over \( n \) years:

\[
(1 + r)^n = (1 + r_1)(1 + r_2) \cdots (1 + r_n)
\]

\[
1 + r = \left[ (1 + r_1)(1 + r_2) \cdots (1 + r_n) \right]^{1/n}
\]

\[
r = \left[ (1 + r_1)(1 + r_2) \cdots (1 + r_n) \right]^{1/n} - 1 \quad (1.23)
\]

This \( r \) is the geometric mean of \( r_1, r_2, \ldots, r_n \).
The “average” that you’re used to is called the *arithmetic mean*. For \( r_1, r_2, \ldots, r_n \), it’s

\[
    r_A = \frac{r_1 + r_2 + \cdots + r_n}{n}
\]

Using calculus, it can be shown that the geometric mean is always less than or equal to the arithmetic mean. In fact, careless use of the arithmetic mean can get you in serious trouble.

**Example 24.** John and Mary both invest $10,000 for two years. John chooses a high-risk investment that yields 90% in the first year, then loses 50% in the second year. Mary’s investment yields 5% in both years. Which one made the better investment?

We haven’t looked at a loss of money before, but it’s not complicated. You still use the simple-interest formula

\[
    F = P(1 + rt)
\]

but when you lose money, \( r \) is negative.

If you were naive (which you’re not), you’d say: “John’s investment yields an average of

\[
    r_A = \frac{0.90 - 0.50}{2} = 0.20 = 20\%
\]

and Mary’s yields an average of 5% per year; so John’s must be a lot better.”

As a Business Math student, you know better. Over the two years, the value of Mary’s investment becomes

\[
    F = $10,000(1 + 0.05)(1 + 0.05) = $11,025
\]

In the same time, the value of John’s investment becomes

\[
    F = $10,000(1 + 0.90)(1 - 0.50) = $9746.79
\]

Not only has John’s investment not performed as well as Mary’s; he’s actually lost money. Poor John! If only he’d known about the geometric mean:

\[
    (1 + r)^2 = (1 + 0.90)(1 - 0.50) = 0.974679
\]

\[
    1 + r = (0.974679)^{1/2} = 0.98726
\]

\[
    r = -0.0127 = -1.27\%
\]

In other words, John’s investment is equivalent to one that loses 1.27% in each of the two years.

**Example 25.** The value of an investment grows by 5% in the first year, by 10% in the second year, and by 15% in the third year. What is the equivalent constant rate \( r \) that would produce the same yield over the three years?
Solution: We are looking for the geometric mean here. Equation (1.23) gives us:

\[(1 + r)^3 = (1 + 0.05)(1 + 0.10)(1 + 0.15) = 1.32825\]
\[1 + r = (1.32825)^{1/3} = 1.0992\]
\[r = 0.992 = 9.92\%\]

Of course, you’ve noticed that this is less than the arithmetic mean of the three rates, which is 10%.

1.10.2 Approximating the geometric mean by the arithmetic mean

Under the right circumstances the geometric mean can be approximated by the arithmetic mean. This follows directly from a multi-dimensional Taylor series expansion of the geometric mean. In particular, if \(r_1, \ldots, r_n\) are all much, much less than one, denoted mathematically as \(r_i \ll 1\), for \(i = 1, \ldots, n\), then

\[(1 + r_1)(1 + r_2)\cdots(1 + r_n) \approx 1 + \sum_{i=1}^{n} r_i + \text{error terms involving the product of the } r_i'\text{s.}\]

Using the Binomial expansion formula it can be shown that

\[r = \left[ (1 + r_1)(1 + r_2)\cdots(1 + r_n) \right]^{1/n} - 1 \approx 1 + \frac{1}{n} \sum_{i=1}^{n} r_i - 1 = \frac{1}{n} \sum_{i=1}^{n} r_i.\]

Thus, in the case of small \(r_i\) it can be shown that the geometric mean can be approximated by the arithmetic mean.

1.11 A last word on interest rates

Interest may be defined as the compensation that a borrower of capital pays a lender of capital. It is sometimes expressed as a percentage of the sum borrowed. The interest rate \(r\) is sometimes referred to as the nominal interest rate, or annual interest rate. By nominal, we mean the amount or face value of a sum of money or a bond, for example, and not the purchasing power or market value. These are just decorative adjectives for the rate of interest on an investment without any adjustment for compounding or inflation. These interest rates are measured in dollars, not in material goods like food and shelter. Nominal interest is the return on an investment measured in dollars per year per dollar of investment. But dollars can become distorted units of measure under inflation. They are not absolutes like food and shelter. An interest rate that is adjusted for inflation is known as a real interest rate. Thus a real interest rate is the nominal interest rate minus the rate of inflation.
**Advanced Comment:** The following argument requires some tools from calculus. We are going to learn about many different kinds of interest: simple, discrete compounding, and continuous compounding. Associated with these are various interest rates. But the nominal interest rate $r$ is the most fundamental. All other interest rates that we will look at are defined in terms of this interest rate. This rate is intimately linked to the growth rate of the investment. In particular, if $F(t + \Delta t)$ is the future value of the investment at time $t + \Delta t$ and $F(t)$ is the future value of the investment at time $t$, then the difference $F(t + \Delta t) - F(t)$ is the price appreciation over time $\Delta t$, where $\Delta t$ is assumed to be a very small unit of time. The rate of return on the investment over the time period $[t, t + \Delta t]$ is defined as $(F(t + \Delta t) - F(t))/F(t)$. The nominal rate of interest can be determined from

$$\lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t F(t)} = \frac{F'(t)}{F(t)}.$$ 

This limit is independent of time for the case of discrete compound and continuous compound interest, and decreases as a function of time in the case of simple interest. Thus $r$ is the relative rate of increase of an investment. Notice that $F'/F$ has the dimension of one over time, and that it represents the instantaneous rate of change of the investment per unit dollar of investment. That is, the units are dollars per year per dollar of investment, which is just one over time (measured in years), hence the name annual interest rate.
2 Summation notation

2.1 Introduction to summation

The idea: Summation notation is a shorthand notation that allows us to replace a long string of similar terms being added together with a more compact expression.

Let $a_1, a_2, \ldots, a_n$ be real numbers. Consider the sum of all of them: $a_1 + a_2 + \cdots + a_n$. (We will write the sum with increasing subscripts. The dots tell you to continue the pattern until you reach the last term.) If $n$ is large, it could be cumbersome or even impossible just to write out all of the terms.

Example 1. If $n = 10^{10}$, could you write all the terms out in your lifetime? Let’s see. If you worked straight with no breaks for 100 years, which ignoring leap years is 3,153,600,000 seconds, and you were able to write one number per second, then you could only add about $3 \times 10^9$ terms. Don’t try this at home!

In summation notation, we write

$$\sum_{i=1}^{n} a_i = a_1 + \cdots + a_n.$$  

The symbol $\Sigma$ is the capital Greek letter “sigma,” which corresponds to our letter S (for “sum”). It is called the summation operator; and $i$ is the index of summation.

The index of summation tells us which of the terms $a_i$ to sum over. There’s nothing special about the letter $i$ as index of summation. It’s only a name for the variable in the expressions that have to be summed. We often use $j$, $k$, $l$, or other letters instead:

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{n} a_j = \sum_{k=1}^{n} a_k = \sum_{l=1}^{n} a_l.$$  

This dummy index of summation runs from the lower limit of summation, which in this case is 1, through the upper limit of summation, which in this case is $n$.

The index of summation doesn’t have to run from 1 through $n$. A more general way of writing summation notation would be

$$\sum_{i=n_{\text{lower}}}^{n_{\text{upper}}} a_i,$$

where $n_{\text{lower}}$ and $n_{\text{upper}}$ can be any integers, so long as $n_{\text{lower}} \leq n_{\text{upper}}$. (Remember: the integers, denoted by $\mathbb{Z}$, are the set of numbers $\mathbb{Z} = \{ \ldots, -2, -1, 0, 1, 2, \ldots \}$.)
The terms $a_i$ that are being summed are best thought of as a function of $i$ over the integers $\mathbb{Z}$. For this reason you will often see the summand expressed as $a_i = a(i)$ in order to emphasize the terms dependence on $i$. This is really no different than taking a function $y = f(x)$ defined on the real numbers $\mathbb{R}$ and restricting its domain to the integers $\mathbb{Z}$ and then replacing $x$ with $i$. We now give an example of this.

**Example 2.** Consider the continuous function

$$f(x) = \frac{1}{x^2 + 1}$$

and the discrete function

$$a_i = a(i) = \frac{1}{i^2 + 1}.$$  

Graph the continuous function on the interval $[-10, 10]$ and the discrete function over the integers from $-10$ to $10$ (i.e. $-10, -9, -8, \ldots, 7, 8, 9, 10$).

![Graphs](image)

Figure 8: The graph on the left is the continuous function $y = f(x) = 1/(1+x^2)$ over the interval $[-10, 10]$, the graph on the right is the discrete function $y = a(i) = 1/(1+i^2)$ over the integers between $-10$ and $10$. Notice that the two functions agree over the integers.

**Comment:** Notice in figure 8 that by just plotting the function over every integer we get a good idea of what the function looks like. In fact, since (by necessity) computers must truncate real numbers in order to store them in bits (0’s and 1’s) every plot of every function ever done by a computer is the graph of a function on a discrete domain that is a subset of the rational numbers $\mathbb{Q}$.

When we sum a series of the form

$$\sum_{i=n_{\text{lower}}}^{n_{\text{upper}}} a_i$$

in Excel, the value $n_{\text{lower}}$ is our start value, the value $n_{\text{upper}}$ is our stop value, $a_i = a(i)$ is our discrete function, and we increment our sum by one. On my website are many exercises for you to gain experience in doing summations.
Example 3. Evaluate the expression $\sum_{i=-5}^{3} i$.

Solution:

$$\sum_{i=-5}^{3} i = -5 - 4 - 3 - 2 - 1 + 0 + 1 + 2 + 3 = -9,$$

where $a_i = i$.

Example 4.

$$\sum_{i=95}^{100} 2i + 3 = (2 \cdot 95 + 3) + (2 \cdot 96 + 3) + (2 \cdot 97 + 3) + (2 \cdot 98 + 3) + (2 \cdot 99 + 3) + (2 \cdot 100 + 3) = 193 + 195 + 197 + 199 + 201 + 203 = 1188,$$

where $a_i = 2i + 3$.

Example 5. Evaluate the expression $\sum_{i=1}^{5} i^2$.

Solution:

$$\sum_{i=-1}^{3} i^2 = (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 = 1 + 0 + 1 + 4 + 9 = 15.$$

Example 6. Expand the sum $\sum_{i=1}^{5} \frac{1}{1 + i^2}$. Do not evaluate!

Solution:

$$\sum_{i=1}^{5} \frac{1}{1 + i^2} = \frac{1}{1 + 1^2} + \frac{1}{1 + 2^2} + \frac{1}{1 + 3^2} + \frac{1}{1 + 4^2} + \frac{1}{1 + 5^2}$$

Summation notation is commonly used in pretty much every textbook imaginable to express formulas in a compact form since any subject area that deals with probability and statistics, regularly express various formulas using summation notation (e.g., engineering, business, finance, sociology, psychology, and more practically, loan payoffs on cars or houses (amortizations) regularly express various formulas using summation notation, etc.). Thus, in order to understand these texts you will need to understand summation notation. Moreover, you may be called upon by your professor to write an Excel program for a formula given in summation notation.

We now give some examples of formulas from probability that use summation notation. For now, you don’t need to know what the formulas represent. Your job is simply to convert back and forth between summation formulas and the corresponding expanded forms.
Example 7. Expand the sum.

$$\sum_{i=1}^{n} x_i f(x_i).$$

**Solution:**

$$\sum_{i=1}^{n} x_i f(x_i) = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_n f(x_n),$$

where $a_i = x_i f(x_i)$.

Example 8. Expand the sum.

$$\sum_{i=1}^{4} (x_i - \mu)^2 f_X(x_i).$$

**Solution:**

$$\sum_{i=1}^{4} (x_i - \mu)^2 f_X(x_i) = (x_1 - \mu)^2 f_X(x_1) + (x_2 - \mu)^2 f_X(x_2) + (x_3 - \mu)^2 f_X(x_3) + (x_4 - \mu)^2 f_X(x_4),$$

where $a_i = (x_i - \mu)^2 f_X(x_i)$.

Next, we will do the reverse. Given a sum of terms, we will look for a pattern in order to write it in the more compact summation notation.

Example 9. Rewrite the sum

$$x_1 P(X = x_1) + x_2 P(X = x_2) + x_3 P(X = x_3) + x_4 P(X = x_4) + x_5 P(X = x_5)$$

using summation notation.

**Solution:**

$$\sum_{i=1}^{5} x_i P(X = x_i), \text{ where } a_i = x_i P(X = x_i).$$

Example 10. Rewrite the sum

$$1^1 + 2^2 + 3^3 + 4^4 + 5^5 + 6^6 + 7^7$$

using summation notation.

**Solution:**

$$\sum_{i=1}^{7} i^i, \text{ where } a_i = i^i.$$
Example 11. Rewrite the sum
\[ 3^2 + 4^3 + 5^4 + 6^5 + 7^6 + 8^7 + 9^8 \]
using summation notation in two different ways.

Solution:
\[ \sum_{i=3}^{9} i^{i-1} = \sum_{i=2}^{8} (i+1)^i. \]

Example 12. Expand the sum.
\[ \sum_{i=1}^{n} f(x_i) \Delta x_i. \]

Solution:
\[ \sum_{i=1}^{n} f(x_i) \Delta x_i = f(x_1) \Delta x_1 + f(x_2) \Delta x_2 + \cdots + f(x_n) \Delta x_n, \quad \text{where} \quad a_i = f(x_i) \Delta x_i. \]

Example 13. Expand the sum in the case \( n = 4 \).
\[ \sum_{i=1}^{n} \frac{e^{i/n}}{n}. \]

Solution:
\[ \sum_{i=1}^{4} \frac{e^{i/4}}{4} = \frac{1}{4} \left[ e^{1/4} + e^{2/4} + e^{3/4} + e^{4/4} \right], \quad \text{where} \quad a_i = \frac{e^{i/n}}{n}. \]

Example 14. Express the following sum using summation notation.
\[ x_1^2 P(X = x_1) + x_2^2 P(X = x_2) + x_3^2 P(X = x_3) + x_4^2 P(X = x_4) = \sum_{i=?} a_i. \]

Solution:
\[ x_1^2 P(X = x_1) + x_2^2 P(X = x_2) + x_3^2 P(X = x_3) + x_4^2 P(X = x_4) = \sum_{i=1}^{4} x_i^2 P(X = x_i). \]
2.2 Properties of summations

Most problems involving the summation of a series are not as straightforward as the ones shown above. We often need to refer to formulas that assist us in computing the more complicated summations. This is especially true when we are attempting to derive general solutions to problems that involve summation notation.

Let $a_i, b_i, r \in \mathbb{R}$ for all $i$, then

\[
\sum_{i=1}^{n} (a_i \pm b_i) = \sum_{i=1}^{n} a_i \pm \sum_{i=1}^{n} b_i. \tag{2.1a}
\]

\[
\sum_{i=1}^{n} (r \cdot a_i) = r \cdot \sum_{i=1}^{n} a_i. \tag{2.1b}
\]

\[
\sum_{i=1}^{n} r = r \cdot \sum_{i=1}^{n} 1 = rn. \tag{2.1c}
\]

Let’s look at these properties when $n = 3$:

\[
\sum_{i=1}^{3} (a_i + b_i) = (a_1 + b_1) + (a_2 + b_2) + (a_3 + b_3)
= (a_1 + a_2 + a_3) + (b_1 + b_2 + b_3) \quad \text{(rearrange terms)}
= \sum_{i=1}^{3} a_i + \sum_{j=1}^{3} b_j
\]

\[
\sum_{i=1}^{3} (r \cdot a_i) = ra_1 + ra_2 + ra_3
= r(a_1 + a_2 + a_3) \quad \text{(common factor)}
= r \sum_{k=1}^{3} a_k
\]

\[
\sum_{i=1}^{3} r = r + r + r
= r(1 + 1 + 1)
= 3r
\]

**Warning:** Don’t confuse addition with multiplication of series.

\[
\sum_{i=1}^{n} (a_i b_i) \neq \left( \sum_{i=1}^{n} a_i \right) \left( \sum_{i=1}^{n} b_i \right)
\]

To understand this, let us examine the case where $n = 2$. The case of a general $n \in \mathbb{N}$ will be similar.
Example 15.

\[ \sum_{i=1}^{2} (a_i b_i) = a_1 b_1 + a_2 b_2 \quad (2 \text{ terms}) \]  
\[ (\sum_{i=1}^{2} a_i)(\sum_{i=1}^{2} b_i) = (a_1 + a_2)(b_1 + b_2) \]
\[ = a_1 b_1 + a_2 b_2 + a_1 b_2 + a_2 b_1 \quad (4 \text{ terms}) \]  

Notice that except in the special case where \( a_1 b_2 + a_2 b_1 = 0 \), the two expressions (2.2a) and (2.2b) are not the same.

We can divide a sum into two or more shorter sums. Suppose \( 1 \leq m < n \). Then

\[ \sum_{i=1}^{n} a_i = \sum_{i=1}^{m} a_i + \sum_{i=m+1}^{n} a_i . \]

For example, if \( n = 5 \) and \( m = 3 \),

\[ \sum_{i=1}^{5} a_i = a_1 + a_2 + a_3 + a_4 + a_5 \]
\[ = (a_1 + a_2 + a_3) + (a_4 + a_5) \]
\[ = \sum_{i=1}^{3} a_i + \sum_{i=4}^{5} a_i \]

Let’s do a few examples dealing only with numbers. For both examples, we take \( n = 5 \) and \( m = 3 \) (just as above). In the first example we let \( a_i = i \), and in the second we let \( a_i = i^2 \).

Example 16.

\[ \sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5 \]
\[ = (1 + 2 + 3) + (4 + 5) \]
\[ = \sum_{i=1}^{3} i + \sum_{i=4}^{5} i \]
\[ = 15 \]
Example 17.

\[ \sum_{i=1}^{5} i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 \]
\[ = (1^2 + 2^2 + 3^2) + (4^2 + 5^2) \]
\[ = \sum_{i=1}^{3} i^2 + \sum_{i=4}^{5} i^2 \]

2.2.1 Applications of summation properties

In the following example problems you should sum the following series by hand. Do not use Excel!

Example 18. \[ \sum_{i=3}^{7} (i - 3)^2 \]

Solution: \[ \sum_{i=3}^{7} (i - 3)^2 = (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 + (6 - 3)^2 + (7 - 3)^2 \]
\[ = 0^2 + 1^2 + 2^2 + 3^2 + 4^2 = 30 \]

Example 19. \[ \sum_{i=0}^{6} (i^2 - 1) \]

Solution: \[ \sum_{i=0}^{6} (i^2 - 1) = (0^2 - 1) + (1^2 - 1) + (2^2 - 1) + (3^2 - 1) + (4^2 - 1) + (5^2 - 1) + (6^2 - 1) \]
\[ = -1 + 0 + 3 + 8 + 15 + 24 + 35 = 84 \]

Example 20. \[ \sum_{k=1}^{4} k^2 \]

Solution: \[ \sum_{k=1}^{4} k^2 = 1^2 + 2^2 + 3^2 + 4^2 = 30 \]

Example 21. \[ \sum_{k=1}^{5} \left( \frac{k}{3} \right)^2 \]

\[ \sum_{k=1}^{5} \left( \frac{k}{3} \right)^2 = \frac{1}{9} \sum_{k=1}^{5} k^2 = \frac{1}{9} \left( 1^2 + 2^2 + 3^2 + 4^2 + 5^2 \right) = \frac{55}{9} \]

Example 22. \[ \frac{2}{5(5+1)} \sum_{k=1}^{5} k \]

Solution: \[ \frac{2}{5(5+1)} \sum_{k=1}^{5} k = \frac{1}{15} (1 + 2 + 3 + 4 + 5) = 1 \]
Example 23. Evaluate $\sum_{i=-1}^{3} (i^2 - 7i + 3)$ using the properties of summation.

**Solution:** Using formulas (2.1) and evaluating the resulting sums yields:

$$\sum_{i=-1}^{3} (i^2 - 7i + 3) = \sum_{i=-1}^{3} i^2 - 7 \sum_{i=-1}^{3} i + 3 \sum_{i=-1}^{3} 1$$

$$= ((-1)^2 + 0^2 + 1^2 + 2^2 + 3^2) - 7(-1 + 0 + 1 + 2 + 3) + 3 \cdot 5$$

$$= (2 + 4 + 9) - 7(5) + 15 = 15 - 35 + 15 = -5$$

Example 24. (What can go wrong!) Evaluate the following expression, if it exists $\sum_{-2}^{3} \frac{1}{i^2} = ?$

**Solution:**

$$\sum_{-2}^{3} \frac{1}{i^2} = \frac{1}{(-2)} + \frac{1}{(-1)^2} + \frac{1}{0^2} + \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} = \infty.$$  

The sum “does not exist” since one of the terms is infinite. You must always be careful that your summand $a_i$ remains bounded for each term.
2.3 Re-indexing

We can also re-index a sum to change its lower and upper limits of summation. If \( c \) is a constant and \( j = i + c \), then \( i = j - c \); and

\[
\sum_{i=1}^{n} a_i = \sum_{j=1+c}^{n+c} a_{j-c}.
\]

Notice that when \( i = 1 \), \( j = 1 + c \) and when \( i = n \), \( j = n + c \). The indices went from \( \{1, 2, \ldots, n\} \rightarrow \{1 + c, 2 + c, \ldots, n + c\} \).

Clearly the number of indices has not changed, nor has the value of terms that they are referring to. Thus a re-indexing is nothing but a re-labelling of the terms. There were initially \( n \) terms \( \{a_1, a_2, \ldots, a_n\} \) before the change in the indices (labels); and after we rename the terms, there must still be \( n \) terms \( \{a_{1+c}, a_{2+c}, \ldots, a_{n+c}\} \). The only thing that has changed is the way we refer to them.

**Comment:** For those of you that have had calculus, this is just like a change of variables:

\[
\int_{a}^{b} f(x)dx = \int_{A}^{B} F(t)dt, \quad \text{where } x = x(t) \text{ and } F(t) = f(x(t)).
\]

**Example 25.** Re-arrange the sum

\[
\sum_{j=0}^{4} (3j - 4)
\]

so that its new index of summation \( k \) runs from 1 to 5.

**Solution:** We want \( k = 1 \) when \( j = 0 \), and \( k = 5 \) when \( j = 4 \). Thus we need to let \( k = j + 1 \). We get

\[
\sum_{j=0}^{4} (3j - 4) = \sum_{k=1}^{5} (3(k - 1) - 4) = \sum_{k=1}^{5} (3k - 7) = -4 - 1 + 2 + 5 + 8 = 10.
\]

**Example 26.** Evaluate the expression

\[
\sum_{j=1}^{10} j^2 - 5 \sum_{j=1}^{10} j
\]

by first combining the terms into a single sum. Also, change the index \( j \) to an \( i \). Does this change the value of the sum?

**Solution:**

\[
\sum_{j=1}^{10} j^2 - 5 \sum_{j=1}^{10} j = \sum_{j=1}^{10} (j^2 - 5j) = \sum_{i=1}^{10} (i^2 - 5i)
\]

We can add the first two sums because they both ran from \( j = 1 \) to \( j = 10 \). In the last expression, we substituted \( i = j \): the precise name of the index of summation doesn’t matter. The value of the sum should be independent of the index of summation.
Example 27. Begin with \( \sum_{j=0}^{4} (3j + 7) \). Re-index the summation so that the new index runs from 1 to 5.

Solution: Let \( k = j + 1 \). If \( 0 \leq j \leq 4 \), then \( 1 \leq j + 1 = k \leq 5 \).

\[
\sum_{j=0}^{4} (3j + 7) = \sum_{k=1}^{5} (3(k-1) + 7) = \sum_{k=1}^{5} (3k + 4)
\]

Example 28 (index shifting doesn’t change the sum). Consider the following sum: \( \sum_{i=1}^{n} a_i = 11 \). If we change the index of summation by letting \( j = i + 5 \) what will be the value sum?

Solution: The sum is still 11. Renaming the index of summation cannot change the value of the sum! Remember, when you shift the index you are not changing the values of the \( a_i \)'s, you are just re-labeling their index of summation. If we make a change of index by letting \( j = i + 5 \) and \( b_j \equiv a_{i+5} \). Then

\[
\sum_{i=1}^{n} a_i = \sum_{j=6}^{n+5} = 11.
\]

For example, if \( n = 2 \), then \( a_1 = b_6 \) and \( a_2 = b_7 \) and the sum becomes

\[
a_1 + a_2 = 11 = b_6 + b_7.
\]
2.4 Summing series

Sometimes it is not so easy to find a compact formula for a series of terms being added. Probably the greatest mathematician to ever live was Carl Fredrick Gauss. This next example is a problem that was posed to Gauss by his teacher when he was in grade school. It was intended to keep him busy for a long time. Not only did he solve it quickly, his solution was nothing short of brilliant!

Example 29. (Gauss’s summation formula)
Suppose we want to sum a large number of consecutive integers:

\[ S = \sum_{i=1}^{n} i. \]

If \( n \) is a large number, for example 1000, it would be a lot of work to add these up. Fortunately, there’s an easier way to get the sum.

Since the sum is finite we know the sum exists, so let’s give it a name, say \( S \) for sum. Let

\[ S = 1 + 2 + \cdots + (n-1) + n. \]  \( (2.3) \)

If we have a finite number of terms, we can add them in any order. That means we can also write the sum in the reverse order:

\[ S = n + (n-1) + \cdots + 2 + 1. \]  \( (2.4) \)

Now, here comes the stroke of genius; add equations (2.3) and (2.4) term by term:

\[
2S = [n + 1] + [(n - 1) + 2] + \cdots + [2 + (n - 1)] + [1 + n] \\
= [n + 1] + [n + 1] + \cdots + [n + 1] + [n + 1] \quad (n \text{ terms}) \\
= n(n+1).
\]

If we divide by 2, we get Gauss’s summation formula:

\[ S = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}. \]  \( (2.5) \)

For example, if \( n = 1000 \),

\[ S = \sum_{i=1}^{1000} i = \frac{(1000)(1001)}{2} = (500)(1001). \]

Example 30.
\[ \frac{1}{100} \sum_{j=0}^{100} j = \frac{1}{100} \cdot \frac{100(101)}{2} = \frac{101}{2} = 50.5. \]
Using this formula and an iterative process, we can develop the formula for any sum of the form:

\[ \sum_{i=1}^{n} i^m, \quad m \in \mathbb{N}. \]

**Example 31.** Find a formula for the sum

\[ \sum_{i=1}^{n} i^2. \]

To find \( \sum i^2 \), notice that

\[
(i - 1)^3 = i^3 + 3(-1)^1i^2 + 3(-1)^2i + (-1)^3
= i^3 - 3i^2 + 3i - 1;
\]

so \( i^3 - (i - 1)^3 = 3i^2 - 3i + 1 \)

If we sum both sides of this equation, we get

\[
\sum_{i=1}^{n} [i^3 - (i-1)^3] = 3 \sum_{i=1}^{n} i^2 - 3 \sum_{i=1}^{n} i + \sum_{i=1}^{n} 1 \quad (2.6)
\]

The expression on the left is a **telescoping sum**: if we write it out, we see that adjacent terms cancel out. To see this, let’s write it in descending order, beginning with \( i = n \) and going down to \( i = 1 \):

\[
\sum_{i=1}^{n} [i^3 - (i-1)^3] = [n^3 - (n-1)^3] + [(n-1)^3 - (n-2)^3] + [(n-2)^3 - (n-3)^3] + \ldots \\
+ [2^3 - 1^3] + [1^3 - 0^3]
\]

All the terms cancel out, except for the first and last ones: \( n^3 - 0^3 = n^3 \). Substituting this back into Equation (2.6) gives us

\[
n^3 = 3 \sum_{i=1}^{n} i^2 - 3 \sum_{i=1}^{n} i + \sum_{i=1}^{n} 1 \quad (2.7)
\]

Recalling that we have a formula for the sum of \( n \) identical constants, the rightmost sum becomes \( \sum_{i=1}^{n} 1 = 1 + 1 + \cdots + 1 = n \). Notice that there are \( n \) ones in the summation. Using this formula together with Gauss’s summation formula (2.5) in equation (2.7) gives us

\[
n^3 = 3 \sum_{i=1}^{n} i^2 - \frac{3}{2} n(n + 1) + n;
\]

and rearranging this equation yields

\[
\sum_{i=1}^{n} i^2 = \frac{n^3 + \frac{3}{2} n(n + 1) - n}{3} = \frac{2n^3 + 3n^2 + 3n - 2n}{6} = \frac{n(n+1)(2n+1)}{6}. \quad (2.8)
\]
Example 32 (Geometric series). Next, consider the sum: \( S = 1 + r + r^2 + \cdots + r^n \). Let us try a trick similar to the one used by Gauss to sum \( 1 + \cdots + n \). But instead of adding terms, we will subtract terms in order to get the series to telescope down to two terms.

Multiplying the sum \( S \) by \((-r)\) yields

\[
S = 1 + r + r^2 + \cdots + r^n \tag{2.9a}
\]
\[-rS = -r - r^2 - \cdots - r^n - r^{n+1}. \tag{2.9b}
\]

Notice that we have carefully arranged “like terms” in columns. When we add equations (2.9a) and (2.9b), all the terms cancel out except the first term in (2.9a) and the last term in (2.9b). We get:

\[
S - rS = 1 - r^{n+1} \quad \frac{\text{\divide by } (1-r)}{\implies S = \frac{1 - r^{n+1}}{1 - r}.} \tag{2.10}
\]

This is known as a geometric series.

Example 33 (Harmonic series). The series \( 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots \) is known as the harmonic series. It diverges: that is, it doesn’t have a finite sum:

\[
\sum_{n=1}^{\infty} \frac{1}{n} = \infty.
\]

Let’s look at the sum of the first \( N \) terms: \( \sum_{n=1}^{N} \frac{1}{n} \). For large \( N \), this series behaves like

\[
\sum_{n=1}^{N} \frac{1}{n} \sim \ln(N) + \gamma
\]

where \( \gamma \approx 0.577 \) is known as Euler’s constant. You can verify this by using Excel to rewrite the expression as

\[
\sum_{n=1}^{N} \frac{1}{n} - \ln(N) = \gamma
\]

and computing the left-hand side. Notice that we have cancelled out the divergent growth component of the sum by putting the log term on the left-hand side.

Example 34 (Taylor series). It is a well-known fact in mathematics that most typical functions\(^1\) that you will encounter in physics that are smooth at a point \( x_0 \) (i.e., have an infinite number of derivatives) can be expressed “locally” for \( x \) “near” \( x_0 \) by the infinite series

\[
f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!} x^n,
\]

\(^1\)there are exceptions to this, but they are beyond the scope of this discussion.
where the value of \( a_n = \frac{f^{(n)}(x_0)}{n!} \) is given by a formula involving the \( n^{th} \) derivative of the function \( f \) evaluated at \( x = x_0 \), and \( n! = 1 \cdot 2 \cdot 3 \cdots n \) is the factorial function. This is known as a Taylor series about \( x = x_0 \). In particular, we will be interested in using a Taylor series to approximate the two functions \( f(x) = e^x \) and \( f(x) = (1 + x)^\alpha \) where \( \alpha \) is a real number near \( x = 0 \) (i.e., at \( x_0 = 0 \).

For example the Taylor expansion of the exponential function about \( x = 0 \) is valid at every point on the real axis! The Taylor series is given by

\[
e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad \text{where} \quad n! = 1 \cdot 2 \cdot 3 \cdots n
\]

This expression is valid for every real number \( x \). For small values of \( x \), we can approximate the infinite series by the first \( N + 1 \) terms in the sum. That is,

\[
e^x \approx \sum_{n=0}^{N} \frac{x^n}{n!} = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots + \frac{1}{N!}x^N.
\]

This example is just meant to be a brief introduction to the concept of a Taylor series. You will see this series again in more detail in the chapter on interest theory. There you will actually apply the above approximation to the formula for compound interest.

For a more thorough discussion on sequences and series, see the text *College Algebra* by Steward, Redlin, and Watson. For a more complete discussion see your calculus text.

### 2.4.1 Applications of summing series

Use the properties of summation and the various formulas from the previous sections to sum the following series by hand. Do not use Excel!

**Example 35 (Geometric Series).** Sum the following series.

\[
\sum_{k=1}^{50} \left( \frac{k}{3} \right)^2 = ?
\]

**Solution:**

\[
\sum_{k=1}^{50} \left( \frac{k}{3} \right)^2 = \frac{1}{9} \sum_{k=1}^{50} k^2 = \frac{1}{9} \left( \frac{50(50 + 1)(2 \cdot 50 + 1)}{6} \right) = \frac{50 \cdot 51 \cdot 101}{54} = 47694.4
\]
Example 36 (Gauss summation formula). Sum the following series.

\[
\sum_{j=0}^{100} (5j + 1) = ?
\]

**Solution:** Recall the Gauss summation formula: 
\[
\sum_{i=1}^{n} i = \frac{n(n + 1)}{2}.
\]

\[
\sum_{j=0}^{100} (5j + 1) = 5 \sum_{j=0}^{100} j + \sum_{j=0}^{100} 1 = 5 \cdot \frac{100 \cdot 101}{2} + 101 = 101 \cdot 251 = 25351
\]

**Note:** In the above calculation, we used the fact that the sum from \(j = 0\) to \(j = 100\) contains 101 terms (not 100). Think about it!

Example 37 (Geometric series). Sum the following series.

\[
\sum_{k=1}^{50} \left(\frac{1}{2}\right)^k
\]

**Solution:** letting \(j = k - 1\) and using the formula for a geometric series we have

\[
\sum_{k=1}^{50} \left(\frac{1}{2}\right)^k = \frac{1}{2} \sum_{j=0}^{49} \left(\frac{1}{2}\right)^j = \frac{1}{2} \cdot \frac{1 - \left(\frac{1}{2}\right)^{50}}{1 - \frac{1}{2}} = 1 - \left(\frac{1}{2}\right)^{50}
\]

Example 38 (Harmonic series). Sum the following series.

\[
\sum_{n=1}^{1000} \frac{1}{n} = ?
\]

**Solution:** \(\sum_{n=1}^{1000} \frac{1}{n} \approx \ln(1000) + .577\)

Example 39 (Changing the Index of Summation). Let \(k = j + 1\) and compute the sum \(\sum_{j=0}^{3} (2j + 1)\).

**Solution:** \(\sum_{j=0}^{3} (2j + 1) = 2 \sum_{k=1}^{4} (2(k-1) + 1) = \sum_{k=1}^{4} (2k - 1)\)

(Here we let \(k = j + 1\). If \(0 \leq j \leq 3\), then \(1 \leq j + 1 = k \leq 4\).)

Example 40 (Renaming the Index of Summation). Consider the sum \(\sum_{j=1}^{3} a_j = 5\).

If we change the index \(j\) to \(i\), does it change the value of the sum?

**Solution:** No! The index of summation is a dummy variable. Since it is nothing but a place holder, it does not matter what name it goes by. We could be really unconventional and call it Fred. Then the sum becomes

\[
\sum_{j=1}^{3} a_j = \sum_{Fred=1}^{3} a_{Fred} = a_4 + a_2 + a_3 = 5
\]
2.5 Applications of summation using spreadsheets

We now give some problems involving evaluating sums using a spreadsheet program. On some of these problems you can use the general formulas, such as the geometric sum formula and Gauss’s summation formulas, together with the properties of sums, to find the value of the sum analytically. However, this exercise is more about learning how to harness the power of spreadsheet programs, such as Excel, than developing your mathematical prowess, so we will resist this temptation and just use spreadsheets to evaluate the sums.

For the following problems create a spreadsheet (such as an Excel spreadsheet) to evaluate the sum

\[ \sum_{i=i_{\text{initial}}}^{i_{\text{final}}} a_i. \]

In each problem you will need to identify the summand \( a_i = a(i) \), the start value \( i_{\text{initial}} \) and the stop value \( i_{\text{final}} \).

Example 41. Use Excel to verify the formulas:

\[
\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \quad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \quad \text{for } n = 100.
\]

**Solution:** Take \( a_i = i \), \( i_{\text{initial}} = 1 \), \( i_{\text{final}} = 100 \).

Example 42. Use Excel to verify the summation property

\[
\sum_{i=1}^{n} r \cdot a_i = r \left( \sum_{i=1}^{n} a_i \right)
\]

for \( n = 30 \), \( r = 7 \), and \( a_i = a(i) = 2^{-i} \).

**Solution:** It should be pointed out that this works for every \( n \), \( r \), and \( a_i \). In order to use a spreadsheet program we had to pick a specific example.

Example 43. Use a spreadsheet program to determine the sum: \( \sum_{i=1}^{10} i \)

**Solution:** Take \( a_i = i \), \( i_{\text{initial}} = 1 \), \( i_{\text{final}} = 10 \).

Answer: 55

Example 44. Use a spreadsheet program to determine the sum: \( \sum_{i=1}^{10} 3i - 5 \)

**Solution:** Take \( a_i = 3i - 5 \), \( i_{\text{initial}} = 1 \), \( i_{\text{final}} = 10 \).

Answer: 115
Example 45. Use a spreadsheet program to determine the sum: \( \sum_{i=-7}^{50} i^2 \)

**Solution:** Take \( a_i = i \), \( i_{\text{initial}} = -7 \), \( i_{\text{final}} = 50 \).

**Answer:** 43,065

Example 46. Use a spreadsheet program to determine the sum: \( \sum_{i=-3}^{100} \frac{1}{i^2 + 1} \)

**Solution:** Take \( a_i = \frac{1}{i^2 + 1} \), \( i_{\text{initial}} = -3 \), \( i_{\text{final}} = 100 \).

**Answer:** 2.866724209

Example 47. Use a spreadsheet program to determine the sum: \( \sum_{n=1}^{20} \frac{1}{n^n} \)

**Solution:** Take \( a_n = \frac{1}{n^n} \), \( n_{\text{initial}} = 1 \), \( n_{\text{final}} = 20 \). Notice that the summation index need not be \( i \).

**Answer:** 1.291285997

Example 48. Use a spreadsheet program to determine the sum: \( \sum_{n=-20}^{20} 3n - 1 \)

**Solution:** Take \( a_n = 3n - 1 \), \( n_{\text{initial}} = -20 \), \( n_{\text{final}} = 20 \). Notice that the summation index need not be \( i \).

**Answer:** -41

Example 49. Use a spreadsheet program to compare the size of the sum of the two series:

\( S_1 = \sum_{i=1}^{50} a_i \) and \( S_2 = \sum_{i=1}^{50} b_i \), where \( a_i = \frac{1}{i^2} \) and \( b_i = \frac{1}{i^4} \) respectively. Does your result agree with your intuition?

**Solution:** Take \( a_i = \frac{1}{i^2} \), \( b_i = \frac{1}{i^4} \), \( i_{\text{initial}} = 1 \), \( i_{\text{final}} = 50 \) you should find that \( S_1 = 1.625132734 \) and \( S_2 = 1.082320646 \). Thus, \( S_2 < S_1 \). We should have anticipated this result since the summands satisfy the inequality

\[ i^2 < i^4 \quad \Rightarrow \quad \frac{1}{i^4} < \frac{1}{i^2} \quad (\forall i > 1) \quad \Rightarrow \quad \sum_{i=1}^{N} \frac{1}{i^4} < \sum_{i=1}^{N} \frac{1}{i^2} , \quad \forall N \in \mathbb{N} . \]

Example 50. Use a spreadsheet program to compare the size of the sum of the two series:

\( S_1 = \sum_{i=1}^{30} a_i \) and \( S_2 = \sum_{i=1}^{30} b_i \), where \( a_i = -2i \) and \( b_i = -i^2 \) respectively. Does your result agree with your intuition?
**Solution:** Take \( a_i = -2i \), \( b_i = -i^2 \), \( i_{\text{initial}} = 1 \), \( i_{\text{final}} = 30 \) you should find that \( S_1 = 0.156517643 \) and \( S_2 = 0.386318602 \). Thus, \( S_1 < S_2 \). This is a more subtle problem than it might first appear to be. For \( i > 1 \),

\[
2i < i^2 \quad \xrightarrow{(-1)} \quad -i^2 < -2i \quad \xrightarrow{e^{i}} \quad e^{-i^2} < e^{-2i}.
\]

Thus, it follows that

\[
\sum_{i=2}^{30} e^{-i^2} < \sum_{i=2}^{30} e^{-2i}.
\]

So it should be the case that \( S_2 < S_1 \) right? Wrong! The first term corresponding to \((i = 1)\) in both sums is so much larger than the rest of the corresponding sums that the inequality is determined by comparing the first terms! In this case \( e^{-2} < e^{-1} \), so \( S_2 < S_1 \).

How could we have anticipated this? One technique is to examine the behavior of the function being summed (the same would be true if we were comparing two integrals over identical domains). A dying exponential dies off so fast with growing exponent that each term in the sum becomes negligible compared to the first term, which is of \( O(1) \). To see this let’s examine a few values of \( \exp(-x) \) for \( x = 1, 2, 3 \). The corresponding values of the function are 0.367879441, 0.018315639, and 0.00012341, which are approximately \( 1/3 \), \( 1/100 \), and \( 1/10,000 \). A similar result holds for the other terms in the sum \( S_1 \). Notice how much smaller the second and third term are than the first term. It should now be clear why the sums are dominated by their first terms.

Will such an argument as the one given above always work? No! The answer is that some times it’s simply too close to call. However, in the above case we were fortunate enough to have an exponential function, and by observing the behavior of this function over the domain of the sum we could easily see that the first term was the dominant term for each sum. In general, the behavior of two functions \( a(i) \) and \( b(i) \) whose values vary in such a way that over some values of \( i \) we have \( a_i < b_i \), and over other values of \( i \) we have \( b_i < a_i \) (i.e., the inequality is reversed), can make it very tricky to predict the outcome of which of the two sums is bigger.
3 Elementary set theory

3.1 Introduction to sets

A set is basically a number of objects grouped together and viewed as a single object or collection.

Definition 2. A set is a collection of objects. Each one of the objects is called an element or a member of the set. One way of defining a set is by writing all of its elements inside curly braces: \{\ldots\}.

Example 1. The set of counting numbers, or natural numbers is

\[ \mathbb{N} = \{1, 2, 3, \ldots\} \].

Example 2. The set of all integers is

\[ \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\} \].

Example 3. When we write out a set, we don’t repeat elements. The set of all letters in the word “Mississippi” is \{M,i,s,p\}.

To indicate that some object \(a\) is a member of a set \(A\), we write

\[ a \in A \].

This is read as “\(a\) is an element of \(A\)” To indicate that \(b\) is not a member of \(A\), we write

\[ b \notin A \].

We read this: “\(b\) is not an element of \(A\)”

We can represent ideas and properties of sets with the use of a diagram known as a Venn diagram. The concepts \(a \in A\) and \(b \notin A\) can be represented by a Venn diagram, such as the one in figure 9.

![Venn diagram](image)

Figure 9: Venn diagram of \(a \in A\) and \(b \notin A\).
The rectangular part of the diagram is the universal set \( U \), which is nothing but our “desk top” or “work space”. We’ll revisit this concept later in this section when we study the complement of a set. The circle represents the set \( A \) that we are working with, and the dots represent the elements \( a \) and \( b \).

**Example 4.** \( 1 \in \mathbb{N}, 2 \in \mathbb{N}, \) but \( 0 \notin \mathbb{N} \).

**Example 5.** Suppose \( A = \{1, 2, 4\} \). Then \( 1 \in A, 2 \in A, 4 \in A, \) but \( 3 \notin A \).

**Example 6.** Let \( \mathbb{E}^+ \) be the set of positive even numbers, and let \( \mathbb{O}^+ \) be the set of positive odd numbers\(^2\). Then \( 1 \in \mathbb{O}^+ \) but \( 1 \notin \mathbb{E}^+ \). In fact, \( \mathbb{E}^+ \) and \( \mathbb{O}^+ \) have no elements in common.

At times, it’s inconvenient or impossible to list all the elements of a set. (Try writing down the set of all birds on the planet Earth!) Instead, we define the set by specifying a property of elements of the set. This is called **set-builder notation**.

**Example 7.** Let \( S \) be the set of integers that are greater than 3 and less than or equal to 10. We could define this set by listing its members:

\[
S = \{4, 5, 6, 7, 8, 9, 10\}.
\]

Using set-builder notation, we’d write it as:

\[
S = \{x \mid x \text{ is an integer that is greater than 3 and less than or equal to 10}\}.
\]

The vertical bar is read “such that,” so if you were reading this aloud, you’d say “\( x \) such that \( x \) is an integer that is greater than three...”

A more compact way of writing this is

\[
S = \{x \mid x \in \mathbb{Z} \text{ and } 3 < x \leq 10\},
\]

which can be further abbreviated as

\[
S = \{x \in \mathbb{Z} \mid 3 < x \leq 10\}.
\]

We’d read that: “\( S \) equals the set of all \( x \) in \( \mathbb{Z} \) such that three is less than \( x \) and \( x \) is less than or equal to ten.” Mathematicians love this compact notation because it says so much with a minimum of symbols. The mind can grasp it quickly and cleanly.

Mathematicians also like to shorten sentences for clarity. For example, instead of saying “\( x \) is less than or equal to ten,” it’s easier to say “\( x \) is at most ten.” In the same way, we could say “\( x \) is at least ten” instead of “\( x \) is greater than or equal to ten.”

\(^2\)Comment on notation: Typically, when a superscript + attached to a set of real numbers, say \( A \), it denotes the subset of \( A \) consisting of all the positive numbers in \( A \), and is denoted as \( A^+ \). Similarly, the − superscript refers to the negative numbers in \( A \). For example, \( \mathbb{Z}^+ = \mathbb{N} \).
Example 8. The set of all birds can be written

\[ B = \{ b \mid b \text{ is a bird} \} . \]

If \( r \) is a robin, then \( r \in B \). On the other hand, if \( g \) is Garfield the cat, then \( g \notin B \).

Example 9. The set of all stocks in the Standard & Poor 500 can be written:

\[ S = \{ s \mid s \in \text{S&P500} \} . \]

Definition 3. If every member of a set \( A \) is also a member of a set \( B \), then we say that \( A \) is a subset of \( B \) and write

\[ A \subseteq B \, . \]

Example 10. \( \{1, 2, 3\} \subseteq \{1, 2, 3, 4, 5\} \), since all the elements in the first set are also elements of the second. Another example is the classification of numbers

\[ \mathbb{N} \subseteq \mathbb{N}_0 \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C} , \]

where \( \mathbb{N} \) is the natural numbers, \( \mathbb{N}_0 = \mathbb{N} \cup \{0\} \) is the whole numbers, \( \mathbb{Z} \) is the integers, \( \mathbb{Q} \) is the rational numbers, \( \mathbb{R} \) is the real numbers, and \( \mathbb{C} \) is the complex numbers.

Definition 4. Two sets \( A \) and \( B \) are equal if they have all the same elements. Then we write \( A = B \).

In practice we prove two sets are equal by showing:

\[ A \subseteq B \text{ and } B \subseteq A \implies A \subseteq B \subseteq A \] (see figure below).

![Venn diagram](image)

About the figure: Venn diagram depicting the situation of set \( A \) contained in set \( B \), which is again contained in set \( A \). Notice that set \( B \) is squeezed in to being equal to set \( A \).

Example 11 (the elements of a set need not be ordered). \( \{1, 2, 3\} = \{3, 1, 2\} \). The order in which we list the elements of a set doesn’t matter. Just to make life easier, we will generally list integers in increasing order.

Example 12 (textbf{expressing sets}). \( \{ x \mid x^2 = 9 \} = \{-3, 3\} \). The definitions look totally different, but the sets are the same.
It is possible for a set to have only one element: \( \{0\} \) is a perfectly good set. It’s important to remember that a one-element set is not the same thing as its element. It would be wrong to write \( 0 = \{0\} \), or to say \( 3 - \{1\} = 2 \). If you’re talking about the set, don’t say “zero”; it’s “the set consisting of zero” or “the set whose only element is zero.”

In fact, a set doesn’t have to have any elements at all!

**Definition 5.** A set with no members is called the **empty set** or the **null set**, and is written \( \emptyset \). That slash is very important: \( \emptyset \neq 0 \). We can also write the empty set as an empty pair of brackets: \( \emptyset = \{\}\). 

**Warning:** Don’t be tempted to combine these two ways of writing the empty set. \( \{\emptyset\} \) isn’t the empty set at all: it’s a set containing one element, and that element is the empty set. (Sets can be members of sets, which can get confusing very quickly: \( \emptyset \neq \{\emptyset\} \neq \{\{\emptyset\}\} \neq \{\{\{\emptyset\}\}\} \neq \ldots \)) Also, remember that \( \{0\} \) isn’t the same as the empty set: it’s not empty, because it contains 0.

If \( A \) is not a subset of \( B \), then we write \( A \nsubseteq B \). For example, \( \{1, 2, 3, 4, 5\} \nsubseteq \{1, 2, 3\} \), because 4 and 5 are members of the first set but not the second. \( \mathbb{Z} \nsubseteq \mathbb{N} \), because \(-1 \in \mathbb{Z} \) but \(-1 \notin \mathbb{N} \).

Notice that by our definition, every set is a subset of itself: \( A \subseteq A \). If \( A \subseteq B \) and \( A \neq B \), then we call \( A \) a **proper subset** of \( B \). We won’t worry too much about whether subsets are proper or not.

Also, the empty set is a subset of every set. For every set \( A \), we can write \( \emptyset \subseteq A \). This is kind of confusing, because the empty set doesn’t have any elements. So the statement is vacuously true. Strictly speaking this is an axiom and cannot but proven, but we can gain insight by using the negation of this statement. To help keep it straight, remember that the way to show \( B \nsubseteq A \) is to find an element of \( B \) that’s not a member of \( A \). Since the empty set doesn’t have any elements, it’s never true that \( \emptyset \nsubseteq A \). Then it must be true that \( \emptyset \subseteq A \). This is also a cheat, since the definition of \( B \) not being a subset of \( A \) assumes that we can always find an element in \( B \). However, if \( B \) is the empty set this is not so. Don’t let this digression make your head explode, I’ve included it to give you a taste of just how subtle set theory can be.

If \( S \) is a set, then we will use the notation \( \#(S) \) to mean the number of elements in \( S \). For example, if \( W = \{w \mid w \text{ is a day of the week}\} \), then \( \#(W) = 7 \). If \( S = \{s \mid s \text{ is a state in the United States}\} \), then \( \#(S) = 50 \).

To illustrate the relationships between sets, we use **Venn diagrams**: circles, ovals, and rectangles with shading (see examples below). The rectangular set \( U \) is known as the
universal set. The concept of a universal set will be important when we’re working with sets, since it will define our work space. We’ll discuss this in more detail soon.

![Venn diagram of $A \subseteq B$.](image)

Figure 10: Venn diagram of $A \subseteq B$. Notice that every point in the set $A$ (the gray circle) is also in the set $B$ (the oval).

![Venn diagram of $A \subsetneq B$.](image)

Figure 11: Venn diagram of $A \subsetneq B$. Here the set $B$ is the oval and the set $A$ is the circle. Only the shaded portion of $A$ is in $B$. In particular, the point $x$ is in $A$ but not in $B$. 
3.1.1 Practice problems with solutions

In this next group of examples we apply some of these ideas to the following sets of numbers and letters:

Let \( S = \{a, b, c, d, e, 1, 2, 3, 4, 5\} \),
\( A = \{1, 2, 3\} \), \( B = \{2, 4, 6\} \), \( C = \{0, 1, 2\} \), \( D = \{e, 1\} \).

**Example 13.** For the mathematical expressions 1-8, determine whether or not they are true or false.

1. \( A \subseteq S \) 5. \( 4 \in S \)
2. \( A \subseteq B \) 6. \( 4 \in A \)
3. \( C \nsubseteq S \) 7. \( 6 \notin S \)
4. \( D \subseteq S \) 8. \( \emptyset \subseteq A \)

**Solution:**

1. true 5. true
2. false 6. false
3. true 7. true
4. true 8. true

**Example 14.** Determine \#(\( S \)) and \#(\( A \))

**Solution:** \#(\( S \)) = 10 and \#(\( A \)) = 3.

**Example 15.** List all elements of the following sets:

\( S_1 = \{x \mid x \in \mathbb{Z} \text{ and } x^2 = 1\} \)
\( S_2 = \{x \mid x \in \mathbb{N} \text{ and } x^2 = 1\} \)
\( S_3 = \{x \mid x \in \mathbb{N} \text{ and } x(x - 1) = 0\} \)

**Solution:** By undoing the set notation we have

\( S_1 = \{x \mid x \in \mathbb{Z} \text{ and } x^2 = 1\} = \{-1, 1\} \)
\( S_2 = \{x \mid x \in \mathbb{N} \text{ and } x^2 = 1\} = \{1\} \), since \(-1 \notin \mathbb{N}\)
\( S_3 = \{x \mid x \in \mathbb{N} \text{ and } x(x - 1) = 0\} = \{1\} \), since \(0 \notin \mathbb{N}\)

**Example 16.** Which of the following sets are equal to \( A \), above?

\( E_1 = \{1, 3, 2\} \quad E_2 = \{2, 1\} \quad E_3 = \{3, 2, 4\} \)

**Solution:** Recall: two sets are equal if and only if they have the same number of elements. The set \( E_2 \) is missing a 3, the set \( E_3 \) is missing a 1 and has an extra element 4. The set \( E_1 \) has the same elements as \( A \). Thus, \( E_1 = A \).
3.2 Intersections, unions, and complements

3.2.1 Intersections

Definition 6. If \( A \) and \( B \) are sets, then the set of all elements that belong to both \( A \) and \( B \) is called the intersection of \( A \) and \( B \). It is written

\[ A \cap B. \]

We define it as: \( x \in A \cap B \) if and only if \( x \in A \) and \( x \in B \). The Venn diagram below shows the intersection of two sets.

![Venn diagram of intersecting sets](image)

Figure 12: Venn diagram of intersecting sets. The shaded region is the intersection of the two sets. Notice that every point in the shaded region is in both \( A \) and \( B \).

Definition 7. If two sets have no elements in common, then the sets are disjoint. In other words, \( A \) and \( B \) are disjoint if \( A \cap B = \emptyset \). In the Venn diagram below, the intersection of the two sets is empty since the sets do not overlap.

![Venn diagram of disjoint sets](image)

Figure 13: Venn diagram of sets that are disjoint (\( A \cap B = \emptyset \)). Notice that \( A \) and \( B \) have no points in common.
Example 17. Let \( A = \{1, 2, 5, 7\} \), \( B = \{2, 3, 4, 5\} \), \( C = \{8\} \). Then \( A \cap B = \{2, 5\} \); \( A \cap C = \emptyset \); \( B \cap C = \emptyset \).

If we have a collection of sets, the intersection of the collection is the set of all elements that belong to every set in the collection. If the sets in the collection are \( A_1, A_2, \ldots, A_n \), then we can write the intersection of the collection in either of two ways:

\[
A_1 \cap A_2 \cap \ldots \cap A_n = \bigcap_{i=1}^{n} A_i
\]

and we define it formally as

\[
x \in \bigcap_{i=1}^{n} A_i \quad \text{if and only if} \quad x \in A_i \text{ for every } i \in \{1, 2, \ldots, n\}.
\]

Below is a Venn diagram for the case \( n = 3 \).

![Venn diagram of three sets intersecting](image)

Figure 14: Venn diagram of three sets intersecting. The shaded region is the intersection of the three sets. Notice that every point in the shaded region is in \( A, B, \) and \( C \).

Example 18. Let \( A = \{1, 3, 5, 7\} \), \( B = \{1, 2, 3\} \), \( C = \{3, 4, 5\} \). Then \( A \cap B \cap C = \{3\} \). Don’t forget that this is a set containing 3, not the number 3 itself.

3.2.2 Unions

Definition 8. If \( A \) and \( B \) are sets, then the set of all elements that belong to either \( A \) or \( B \) is the union of \( A \) and \( B \). We write it

\[
A \cup B.
\]

We define it formally as

\[
x \in A \cup B \quad \text{if and only if} \quad x \in A \text{ or } x \in B.
\]
The Venn diagram below shows the union of two sets.

Figure 15: Venn diagram of the union of two sets. The shaded region is the union of the two sets. Notice that every point in the shaded region is in $A$, or in $B$, or in both $A$ and $B$.

**Note:** When we use the word “or,” we mean it *inclusively*. If we say “$P$ or $Q$ is true,” then we mean that $P$ is true; or $Q$ is true; or both $P$ and $Q$ are true. This is different from the *exclusive* “or,” which means “$P$ or $Q$ is true, but not both.” The exclusive form is used a lot in English (“Should I stay home and study for the math test or go out drinking at the bar?”) but in math we use the inclusive form. Thus the mathematical statement “$x \in A$ or $x \in B$” contains three cases:

- **Case 1:** $x \in A$ and $x \notin B \iff x \in A - B$
- **Case 2:** $x \notin A$ and $x \in B \iff x \in B - A$
- **Case 3:** $x \in A$ and $x \in B \iff x \in A \cap B$

To each case there corresponds a set. These sets are disjoint.

Figure 16: Venn diagram of the three cases for the union of two sets. Notice that to each case there corresponds a disjoint set: $A - B$, $A \cap B$, $B - A$. 
**Example 19.** Let $A = \{a, b, c, d, e\}$ and $B = \{d, e, f, g\}$. Then $A \cap B = \{d, e\}$ and $A \cup B = \{a, b, c, d, e, f, g\}$.

We can also take the union of a collection of more than two sets. The definition is basically the same: the union is the set of all elements that belong to one or more set in the collection. We can write it in two ways:

$$A_1 \cup A_2 \cup \ldots \cup A_n = \bigcup_{i=1}^{n} A_i$$

and we define it formally as

$$x \in \bigcup_{i=1}^{n} A_i \quad \text{if and only if} \quad x \in A_i \text{ for some } i \in \{1, 2, \ldots, n\}.$$

**Figure 17:** Venn diagram of the union of three sets: $A \cup B \cup C$. The shaded region is the union of the three sets. Notice that every point in the shaded region is in $A$, or in $B$, or in $C$, or in a combination of two or more of them.

**Example 20.** Let $A = \{1, 3, 5, 7\}$, $B = \{2, 4, 6, 8\}$, $C = \{9, 10\}$. Then $A \cap B \cap C = \emptyset$ and $A \cup B \cup C = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

### 3.2.3 The complement of a set

Given a set $A$, we may want to consider the **complement** of $A$: the set of elements not in $A$. We might try to define this new set as $\{x \mid x \notin A\}$.

But that gets us into real trouble! The trouble is that there are so many things not in $A$, most of which we’re not interested in at all.
Example 21. Suppose $A = \{1, 2, 3\}$. Then $\{x \mid x \notin A\}$ includes 4, 5, 6, ... But it also includes -5, 3.02, $\sqrt{\pi}$, and a whole bunch of other numbers. Furthermore, it includes things like “the woodpecker that pounds on my swamp cooler” and “the planet Pluto.”

One way to get rid of this problem of extraneous elements is to specify in advance that all of the elements must come from some fixed set $U$, called the universal set for the problem. How we decide to define the universal set depends on the problem. For example, we might be interested in looking at the counting numbers $\mathbb{N}$, or the integers $\mathbb{Z}$, or the real numbers $\mathbb{R}$. Universal sets are very important in the study of probability: there, the universal set is taken to be the set of all possible outcomes from an experiment.

Definition 9. If $U$ is the universal set and $A \subseteq U$, then the set of all elements in $U$ that are not members of $A$ is the complement of $A$. It is written $A^c$, and defined

$$x \in A^c \quad \text{if and only if} \quad x \in U \text{ and } x \notin A.$$ 

Below is the Venn diagram of the complement of a set.

![Venn Diagram](image)

Figure 18: Venn diagram of the complement of the set $A$. $A^c$ is everything outside of $A$ but inside of the universal set $U$.

Notice that $A$ is the complement of $A^c$, because it’s all points in $U$ that are not in $A^c$.

Example 22. Suppose $A = \{1, 2, 3\}$ and $U = \mathbb{N}$. Then $A^c = \{4, 5, 6, \ldots\}$.

The complement has a number of important properties. If $U$ is a universal set and $A \subseteq U$, then

1. $(A^c)^c = A$
2. $A \cup A^c = U$
3. $A \cap A^c = \emptyset$.

We will make much use of these identities in our study of probability.
3.3 Some basic laws

Two very important properties of complements are DeMorgan’s laws. If \( A \) and \( B \) are any two sets, then

1. \((A \cup B)^c = A^c \cap B^c\)
2. \((A \cap B)^c = A^c \cup B^c\).

We’ll prove the first of DeMorgan’s laws using Venn diagrams.

**Proof by pictures of property 1:** We must show that both sides of the equation are in fact the same set. We start by determining what part of the Venn diagram represents the set \((A \cup B)^c\). This is the complement of the union of \( A \) and \( B \), so it’s the complement of figure 15. \((A \cup B)^c\) is shaded in the picture:

![Venn Diagram of \((A \cup B)^c\)](image1)

Figure 19: Proof of DeMorgan’s Laws via Venn diagrams: the Venn diagram of \((A \cup B)^c\).

Next we look at the parts of the Venn diagram corresponding to \( A^c \) and \( B^c \).

![Venn Diagram of \(A^c\)](image2)

Figure 20: Proof of DeMorgan’s Laws via Venn diagrams: the Venn diagram of \( A^c \).
Figure 21: Proof of DeMorgan’s Laws via Venn diagrams: the Venn diagram of $B^c$.

Notice that the intersection (the overlap) of figures 20 and 21 gives the Venn diagram found in figure 19. This completes the proof by pictures.

We can also talk about subtracting sets. If we have two sets $A$ and $B$, we’ll define

$$A - B = A \cap B^c.$$

In other words,

$$x \in A - B \quad \text{if and only if} \quad x \in A \text{ and } x \notin B.$$ 

Notice that $A - B \subseteq A$, and that $(A - B) \cap B = \emptyset$. The Venn diagram is shown below.

Figure 22: Venn diagram of the set $A - B = A \cap B^c$. The shaded region is the set of all points in $A$ that are not in $B$. 
Example 23. Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{3, 5, 7, 9\}$. Then $A - B$ is the set of everything that’s in $A$ but not in $B$: $A - B = \{1, 2, 4\}$.

Lastly, there are distributive laws for sets, just as there are for real numbers.

Distributive Laws:

$$(i) \quad C \cap (A \cup B) = (C \cap A) \cup (C \cap B)$$
$$(ii) \quad C \cup (A \cap B) = (C \cup A) \cap (C \cup B)$$

Note: When computing sets involving parentheses, we use the same order of operations as we do with real numbers: any operations inside the parentheses are always performed first.

The proof of $(i)$ is straightforward and is left as an exercise. The proof of $(ii)$ follows from applying DeMorgan’s laws to expression $(i)$.

Differences between how we combine sets and how we combine numbers:

$E + F$ makes no sense for sets just as $1 \cup 2$ makes no sense for numbers. Remember

Tools for manipulating numbers: $+, -, \times, \div$

Tools for manipulating sets: $\cap, \cup, -, E^c$, etc.

Next, we look at a series of examples designed to give you some practice with applying the basic set operations to Venn diagrams.
3.3.1 Practice problems with basic set operations using Venn diagrams

For all of the examples given below in this section shade in the regions corresponding to the given sets for the diagrams shown below. You should try to answer the question yourself before reading the solution.

Figure 23: **Problem 1:** Shade the region $A \cap B$

Figure 24: **Problem 2:** Shade the region $A \cup B$

Figure 25: **Problem 3:** Shade the region $B - A$
Figure 26: **Problem 4:** Shade the region $A^c$

Figure 27: **Problem 5:** Shade the region $(B^c)^c$

Figure 28: **Problem 6:** Shade the region $(A \cap B^c)^c$
Figure 29: **Problem 7**: Shade the region $A \cap B \cap C$

Figure 30: **Problem 8**: Shade the region $A \cup B \cup C$

Figure 31: **Problem 9**: Shade the region $(A \cup B \cup C)^c$
Figure 32: **Problem 10:** Shade the region \((A \cap B) \cup C\)

Figure 33: **Problem 11:** Shade the region \((A \cup B) \cap C\)

Figure 34: **Problem 12:** Shade the region \(((A \cup B) \cap C)^c\)
3.3.2 Solutions to practice problems

Problem 1: See figure 12.

Problem 2: See figure 15.

Problem 3: Method 1: Write $B - A = B \cap A^c$ and use this form to draw the graph. Method 2: Start by examining the Venn diagram for $A - B$ in figure 22. Notice that if we relabel $A$ to be $B$ and $B$ to be $A$, then $B - A \rightarrow A - B$. Using this symmetry you should be able to reconstruct the picture for $B - A$.

Problem 4: See figure 20.

Problem 5: From the properties of complements $(B^c)^c = B$.

Problem 6: First notice that $A \cap B^c = A - B$. Thus $(A \cap B^c)^c = (A - B)^c$. To generate the Venn diagram take the complement of figure 22. The solution is

![Figure 35: Solution to Problem 6: Shade the region $(A \cap B^c)^c$](image)

Problem 7: See figure 14.

Problem 8: See figure 17.

Problem 9: The Venn diagram of $(A \cup B \cup C)^c$ is the complement of figure 17. The solution is given below.
Problem 10: To determine the Venn diagram of \((A \cap B) \cup C\) we first compute \((A \cap B)\). Then we take the union of this set with \(C\). The solution is given below.

Problem 11: To determine the Venn diagram of \((A \cup B) \cap C\) we first compute \((A \cup B)\). Then we intersect this set with \(C\). Notice that \((A \cup B) \cap C \subset C\), so you should not shade outside the set \(C\). The solution is given below.
Figure 38: Solution to Problem 10: Shade the region $(A \cup B) \cap C$

**Problem 12:** To determine the Venn diagram of $((A \cup B) \cap C)^c$ we just need to compute the complement of the set $(A \cup B) \cap C$. The Venn diagram of this set has already been determined in the solution to the preceding problem.

Figure 39: Solution to Problem 10: Shade the region $((A \cup B) \cap C)^c$
3.4 The Power Set

Definition 10. The power set of a set $A$ is the set of all subsets of $A$. It is written $\mathcal{P}(A)$.

Example 24. Let $A = \{1, 2\}$. Then
\[
\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}.
\]

This is a set containing four elements, each of which is a set. Sometimes we’ll refer to a set whose elements are themselves sets as a “collection” or a “family” of sets. This isn’t strictly necessary, but it can help reduce confusion at times.

Be careful with your notation when dealing with power sets! In our example, $\{1\} \subseteq A$, but $\{1\} \notin \mathcal{P}(A)$. In words, the set containing 1 is a subset of $A$, and a member of the power set of $A$. We can write $1 \in A$, but $1 \notin \mathcal{P}(A)$.

Example 25. Let $B = \{a, b, c\}$. Then
\[
\mathcal{P}(B) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.
\]

In Example 24, $\#(A) = 2$ and $\#(\mathcal{P}(A)) = 4 = 2^2$. In Example 25, $\#(B) = 3$ and $\#(\mathcal{P}(B)) = 8 = 2^3$.

Fact: If $A$ is a set and $\#(A) = n$, then $\#(\mathcal{P}(A)) = 2^n$.

Proof. Let $A$ be a set with $n$ elements: $A = \{a_1, a_2, \ldots, a_n\}$. Let $B$ be any subset of $A$. To describe $B$, we have to make one of two choices for every element in $A$:
\[
\begin{align*}
a_1 & \in B \quad \text{or} \quad a_1 \notin B \\
a_2 & \in B \quad \text{or} \quad a_2 \notin B \\
& \quad \vdots \\
a_n & \in B \quad \text{or} \quad a_n \notin B.
\end{align*}
\]

The total number of possibilities for $B$ is
\[
(2 \text{ choices for } a_1) \times (2 \text{ choices for } a_2) \times \cdots \times (2 \text{ choices for } a_n) = 2^n.
\]

That is, we can have $2^n$ distinct subsets. \hfill \square

Example 26. Let $S = \{a, b, c, d, e, 1, 2, 3, 4, 5\}$. Then the number of elements in the power set $\#(\mathcal{P}(S)) = 2^{\#(S)} = 2^{10}$. 
4 Counting techniques: permutations and combinations

When you study probability, one of the first things you’ll have to learn is how to calculate the number of outcomes that can result from an experiment. This sounds easy—most of us learned counting around the first grade. In fact, it can be one of the trickiest areas that you’ll encounter at the introductory math level. It’s given rise to a whole subdiscipline of mathematics, called combinatorics. Combinatorics problems can produce some very non-intuitive results, and even experienced mathematicians have been known to get in trouble by following their intuitions.

4.1 General multistep experiments

We will be looking at experiments that consist of multiple steps. We’ll start by looking at situations in which the outcome of one step doesn’t affect the possible outcomes of others. For example, if you roll a single die two times, the outcome of the first roll won’t affect the possible outcomes of the second. If you roll two dice simultaneously, the number that comes up on the first one won’t affect the number that comes up on the second.

Example 1. Suppose we roll a six-sided die two times and record the number that comes up on each roll. How many different two-roll sequences are possible?

Figure 40 on page 96 shows the possible outcomes. There are six possible outcomes for the first roll. For each of these, there are six possible outcomes for the second roll. Thus there is a total of $6 \times 6 = 36$ possible two-roll sequences.

Another way to look at this is to write the outcomes from the two rolls as ordered pairs:

$$(\text{outcome from roll 1, outcome from roll 2}) = (i, j)$$

where $i \in \{1, 2, \ldots, 6\}$ and $j \in \{1, 2, \ldots, 6\}$. Then the number of possible ordered pairs is $6 \times 6 = 36$.

Example 2. Your math class is taking a field trip, and will take box lunches. Each lunch consists of a food item and a beverage. For a food item, you can choose a ham sandwich, a peanut-butter-and-jelly sandwich, or a salad. For a beverage, you can choose cola or water. How many different lunches are possible?

Solution: One approach is to list all the possible lunches. We can write them as ordered pairs, (food, beverage):

$$(\text{ham, cola}) (\text{ham, water})$$
$$(\text{PBJ, cola}) (\text{PBJ, water})$$
$$(\text{salad, cola}) (\text{salad, water})$$

For each of the 3 food choices, there are 2 beverage choices, for a total of $3 \times 2 = 6$ different lunches.
Figure 40: Possible outcomes of rolling a six-sided die twice (see example 1). There are six possible outcomes for the first roll; for each of these, any of six outcomes is possible for the second roll.

**Example 3.** A car-rental agency offers the choice of three types of vehicles: SUV’s, minivans, and pickups. Vehicles can be chosen in any of four colors: blue, green, yellow, and red. How many different vehicle-color combinations are possible?

**Solution:** Again, we’ll list all the possibilities systematically: for each color, we’ll list all the possible vehicle types.

- blue SUV
- blue minivan
- blue pickup
- green SUV
- green minivan
- green pickup
- yellow SUV
- yellow minivan
- yellow pickup
- red SUV
- red minivan
- red pickup

For each of the 4 colors, there are 3 vehicle types, for a total of $4 \times 3 = 12$ vehicle-color combinations.

You’ve probably spotted the pattern. We’ll make it explicit in the form of a theorem:

**Theorem 1.** Suppose an experiment consists of two steps, the first of which has $n_1$ possible outcomes and the second of which has $n_2$ possible outcomes. Then the total number of possible outcomes for the experiment is the product of the numbers of individual outcomes, $n_1 n_2$.

**Proof:** Let the set of possible outcomes for the first step be $A_1 = \{a_{1,1}, a_{1,2}, \ldots, a_{1,n_1}\}$,
and let the set of possible outcomes for the second step be $A_2 = \{a_{2,1}, a_{2,2}, \ldots, a_{2,n_2}\}$. Then every outcome of the overall experiment can be written as an ordered pair $(a_{1,i}, a_{2,j})$, where $i \in \{1, 2, \ldots, n_1\}$ and $j \in \{1, 2, \ldots, n_2\}$. For each of the $n_1$ values of $i$, there are $n_2$ possible pairs, corresponding to the $n_2$ values of $j$: $(a_{1,i}, a_{2,1}), (a_{1,i}, a_{2,2}), \ldots, (a_{1,i}, a_{2,n_2})$. Thus there are a total of $n_1 \times n_2$ ordered pairs, so $n_1 \times n_2$ outcomes of the overall experiment.

The obvious next question is: what if an experiment consists of more than 2 steps?

**Example 4.** The box-lunch purveyors of Example 2, under pressure from the Wellness Committee, add a choice of four fruits: apple, orange, peach, or pear. Each lunch now consists of a food (ham sandwich, PBJ, or salad), a beverage (cola or water), and one of the four fruits. How many different lunches are now possible?

**Solution:** We can write the possible lunches as ordered triples: (food, beverage, fruit). From Example 2, we found that there were $3 \times 2 = 6$ ordered pairs (food, beverage). For each of these, we can form four ordered triples by adding one of the four fruits:

- $(\text{ham, cola, apple})$  
- $(\text{ham, cola, orange})$  
- $(\text{ham, cola, peach})$  
- $(\text{ham, cola, pear})$

- $(\text{ham, water, apple})$  
- $(\text{ham, water, orange})$  
- $(\text{ham, water, peach})$  
- $(\text{ham, water, pear})$

- $(\text{PBJ, cola, apple})$  
- $(\text{PBJ, cola, orange})$  
- $(\text{PBJ, cola, peach})$  
- $(\text{PBJ, cola, pear})$

- $(\text{PBJ, water, apple})$  
- $(\text{PBJ, water, orange})$  
- $(\text{PBJ, water, peach})$  
- $(\text{PBJ, water, pear})$

- $(\text{salad, cola, apple})$  
- $(\text{salad, cola, orange})$  
- $(\text{salad, cola, peach})$  
- $(\text{salad, cola, pear})$

- $(\text{salad, water, apple})$  
- $(\text{salad, water, orange})$  
- $(\text{salad, water, peach})$  
- $(\text{salad, water, pear})$

We get a total of $3 \times 2 \times 4 = 24$ possible lunches. This suggests that we can extend Theorem 1 to more than two steps.

**Theorem 2.** Suppose an experiment consists of $N$ steps, and each step $i$ has $n_i$ possible outcomes. Then the total number of possible outcomes for the experiment is the product $n_1 n_2 \ldots n_N = \prod_{i=1}^{N} n_i$

**Proof:** This is obviously true for $N = 1$, and Theorem 1 shows that it is true for $N = 2$. For larger values, we’ll use induction: we’ll show that if the theorem is true for $N = K$, then it is also true for $N = K + 1$.

Let the set of possible outcomes for each step $i$ be $A_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,n_i}\}$, where $i \in \{1, 2, \ldots, K + 1\}$. Then each outcome of the overall experiment can be written as an ordered $(K + 1)$-tuple $(a_{1,j_1}, a_{2,j_2}, \ldots, a_{K+1,j_{K+1}})$.

We can rewrite this $(K + 1)$-tuple as an ordered pair $((a_{1,j_1}, a_{2,j_2}, \ldots, a_{K+1,j_{K+1}}), a_{K+1,j_{K+1}})$, whose first member is an ordered $K$-tuple and whose second member is a single element from $A_{K+1}$. In other words, we can treat the experiment as consisting of two “super-steps”: the first consisting of the first $K$ steps of the experiment, and the second consisting of the $(K + 1)$st step.

If the theorem is true for $N = K$, then we know that there are $n_1 n_2 \ldots n_K$ possible outcomes of the first $K$ steps (i.e. of the first “super-step”). By Theorem 1, we
know that since the first “super-step” has $n_1 n_2 \ldots n_K$ possible outcomes and the second “super-step” has $n_{K+1}$ possible outcomes, the overall experiment has $(n_1 n_2 \ldots n_K) n_{K+1} = n_1 n_2 \ldots n_{K+1}$ possible outcomes. Hence if the theorem is true for $N = K$, it is true for $N = K + 1$.

Example 5. A state legislature consists of 30 members of the Prohibition Party, 20 members of the Anti-Masonic Party, and 10 Whigs. How many ways are there to form a committee consisting of one member from each party?

Solution: By Theorem 2, the total number of possible committees is $30 \times 20 \times 10 = 6000$.

Example 6. A six-sided die is rolled three times; on each roll, the number facing upward is recorded. How many different three-roll sequences are possible?

Solution: This experiment consists of three steps, each of which has 6 possible outcomes. By Theorem 2, the number of possible outcomes for the overall experiment is $6 \times 6 \times 6 = 6^3 = 216$.

Example 7. A multiple-choice test consists of 11 questions, each with five choices. How many different ways are there of completing the test?

Solution: We can regard this as an experiment consisting of 11 steps, each with 5 possible outcomes. By Theorem 2, the number of possible outcomes for the overall experiment is $5^{11} \approx 49,000,000$.

Example 8. You toss a coin ten times, recording whether it comes up heads or tails on each toss. How many different sequences of tosses are possible?

Solution: This is an experiment consisting of 10 steps, each with 2 possible outcomes. The number of possible outcomes for the experiment is $2^{10} = 1024$.

Example 9. A hot-dog stand offers five condiments: ketchup, mustard, relish, sauerkraut, and onion. You can eat your hot dog plain or add any set of condiments to it. How many ways are there of preparing a hot dog?

Solution: This is a problem where your intuitions might betray you. Upon first looking at it, you might think that this is a single-step experiment with five outcomes, corresponding to the five condiments. However, if you start listing possible hot-dog preparations, you’ll quickly see that there are more than five.

The correct approach is to observe that for each condiment, you’re making a yes-or-no choice: ketchup or no ketchup, mustard or no mustard, etc. Thus this is an experiment with five steps, each with two possible outcomes (yes or no). The total number of possible preparations is $2^5 = 32$.

4.2 Permutations and combinations

A very common counting and probability situation involves a process called sampling without replacement. For example, suppose you deal a hand of five cards from a standard
deck. Once you’ve dealt a card, you can’t deal that particular card again. Another example is picking a team of nine baseball players from a group. Once you’ve picked someone to be the pitcher, you can’t choose the same person to be the catcher. (We say “without replacement” because once we’ve picked a card or a person, we don’t return it to the pool from which we make our later choices.)

There are two ways of counting when we sample without replacement, called permutations and combinations. In a permutation, the order matters: which is picked first, which is picked second, and so forth. For example, if we ask about the first-, second- and third-place winners in a race, it matters who wins which place. An unordered list of the three people wouldn’t give us the information we want. In a combination, the order doesn’t matter. For example, in many card games, you’re dealt a certain number of cards, but it doesn’t matter which one came first, which one came second, and so forth: all that matters is the particular cards that are in your hand.

4.2.1 Permutations

We’ll begin with some examples of problems involving permutations. These are problems in which the order of elements of a subset matters.

**Example 10.** John, Paul, George, and Ringo are competing in a pool tournament, with a prize of $500 for the winner and a coffee mug for the runner-up. How many different ways are there for the two prizes to be awarded?

**Solution:** This is a permutation problem: it matters, at least to the two people involved, who gets the 500 bucks and who gets the mug. We can write the possibilities as ordered pairs, (winner, runner-up). Any of the four competitors can be the winner; once we’ve selected a winner, there are only three remaining choices for the runner-up (since the same person can’t be both).

\[
\begin{align*}
(\text{John, Paul}) & \quad (\text{John, George}) & \quad (\text{John, Ringo}) \\
(\text{Paul, John}) & \quad (\text{Paul, George}) & \quad (\text{Paul, Ringo}) \\
(\text{George, John}) & \quad (\text{George, Paul}) & \quad (\text{George, Ringo}) \\
(\text{Ringo, John}) & \quad (\text{Ringo, Paul}) & \quad (\text{Ringo, George})
\end{align*}
\]

Notice that, for instance, we’ve listed (John, Paul) and (Paul, John) separately, since they’re two separate and distinct permutations. The total number of permutations is

\[P(4, 2) = 4(4 - 1) = 4(3) = 12\]

We use the notation \(P(n,r)\), where \(n\) is the size of the total set and \(r\) is the size of the subset being chosen. In words, we phrase this number as “the permutations of \(n\) objects taken \(r\) at a time”. If \(r = n\), we can just phrase it as “the permutations of \(n\) objects”.

There is no single standard notation for the number of permutations. We’ve chosen \(P(n,r)\) because it’s easy to write and to read. Other notations used by respectable sources include \(\,^nP_r\), \(\,^nP_r\), and \((n)_r\). You should be prepared to deal with these different notations.
Example 11. A franchisee is opening a new store, and needs to select a day manager, evening manager, and night manager from a pool of ten applicants. How many ways are there to fill the three positions?

Solution: For the first position, the franchisee can choose any of the 10 candidates. For the second position, there are only 10 – 1 = 9 possibilities, since the same person presumably can’t fill two positions. For the third position, two applicants have been removed from the pool, so only 10 – 2 = 8 choices remain. Thus the number of ways to fill the positions is

\[ P(10, 3) = (10)(10 - 1)(10 - 2) = (10)(9)(8) = 720 \]

Example 12. A motel has ten rooms. Four people arrive one night, each asking for a room. How many ways are there of assigning people to rooms?

Solution: The first person to arrive can have any of the 10 rooms. When the second person shows up, one of the rooms is taken; so there are only 10 – 1 = 9 choices available. The third person can have any one of the 10 – 2 = 8 remaining empty rooms; and there are 10 – 3 = 7 rooms available for the fourth person. Thus the total number of ways to assign people to rooms is

\[ P(10, 4) = (10)(10 - 1)(10 - 2)(10 - 3) = (10)(9)(8)(7) = 5040 \]

Example 13. Five people arrive at the post office more or less simultaneously. There is only one clerk working, so they need to stand in line. How many different ways are there for the five people to line up?

Solution: Any one of the 5 people can be first in line. For each of these choices, there remain only 4 possibilities for the second position. This in turn leaves only 3 candidates for the third position; and then there are only 2 possibilities for the fourth position. The last position in the line must be filled by the one person remaining who hasn’t occupied the first, second, third, or fourth position. Hence the total number of arrangements is

\[ P(5, 5) = (5)(4)(3)(2)(1) = 5! = 120 \]

Definition: If you’re not familiar with the exclamation-point notation, it’s called the factorial, and it’s defined by

\[ n! = \begin{cases} n(n-1)(n-2) \cdots (3)(2)(1) & \text{if } n \in \{1, 2, 3, \ldots\} \\ 1 & \text{if } n = 0 \end{cases} \]

In words, \( n! \) is called “\( n \) factorial”.

You may wonder why 0! is defined this way; later in this section, we’ll see an example using it.

Theorem 3. The number of permutations of \( n \) objects taken \( r \) at a time, where \( r \leq n \), is

\[ P(n, r) = n(n - 1)(n - 2) \cdots (n + 1 - r) = \frac{n!}{(n - r)!} \]
Proof: We need to pick a sequence of \( r \) objects out of a set of \( n \). For the first object in the sequence, our choices are unrestricted: we can pick any of the \( n \) items in the set. Once we’ve chosen that object, we can’t choose it again; so when we pick the second object, we only have \( n - 1 \) choices. The process continues that way: when the time comes to pick the \( i \)th object, we’ve already chosen \( i - 1 \), so we only have \( n - (i - 1) = n + 1 - i \) choices remaining. Hence the total number of possible sequences is

\[
P(n, r) = \prod_{i=1}^{r} (n + 1 - i) = n(n - 1) \cdots (n + 1 - r) = \frac{n!}{(n - r)!}
\]

For the third equation in the theorem, note that

\[
\frac{n!}{(n - r)!} = \frac{n(n - 1) \cdots (n + 1 - r)(n - r)(n - r - 1) \cdots (2)(1)}{(n - r)(n - r - 1) \cdots (2)(1)}
\]

We can cancel all the factors in the denominator against all the factors less than or equal to \( n - r \) in the numerator, leaving

\[
\frac{n!}{(n - r)!} = n(n - 1) \cdots (n + 1 - r) = P(n, r)
\]

Example 14. Matthew, Mark, Luke, and John have rented a car, and are arguing about who should get which seat. How many different ways are there of putting the four people in the four seats?

Solution: This is a straightforward question of the permutations of four objects. Using the formula in Theorem 3 gives us

\[
P(4, 4) = \frac{4!}{(4 - 4)!} = \frac{4!}{0!} = 24
\]

This is one reason why mathematicians have agreed to define \( 0! = 1 \).

4.2.2 Permutations of objects not all different

So far, we’ve only looked at permutations of \( n \) distinct objects. We also need to consider situations in which some of the objects are alike.

Example 15. Four people split a box of four pastries: two identical doughnuts, a brownie, and a muffin. How many ways are there of distributing the pastries among the people?

Solution: If the pastries were all different, this would be a simple case of the permutations of four objects: \( P(4, 4) = 4! = 24 \). However, since the two doughnuts are identical, we don’t distinguish between permutations in which the same two people get different doughnuts: for example, if we write the permutations as ordered 4-tuples, we don’t distinguish between \((d_1, d_2, b, m)\) and \((d_2, d_1, b, m)\); we’d write them both as \((d, d, b, m)\).
This is the key: observe that for every permutation in which person A gets doughnut \(d_1\) and person B gets doughnut \(d_2\), there is one that is identical except that person A gets \(d_2\) and person B gets \(d_1\); and these two are not considered distinct distributions. Hence we calculate the total number of permutations as though the doughnuts were distinguishable, then divide by the number of ways of arranging the two doughnuts:

\[
\text{Number of distributions} = \frac{4!}{2} = 12
\]

We’ll list the arrangements, just so we can check that there are actually 12 of them:

- \((d, d, b, m)\)
- \((d, d, m, b)\)
- \((b, d, d, m)\)
- \((d, m, d, b)\)
- \((d, b, m, d)\)
- \((d, m, b, d)\)
- \((b, d, m, d)\)
- \((m, d, b, d)\)
- \((b, m, d, d)\)
- \((m, b, d, d)\)
- \((b, m, d, d)\)
- \((m, b, d, d)\)

**Example 16.** Alice, Betty, Connie, Dora X., Dora Y., and Dora Z. arrive more or less simultaneously at a restaurant. The greeter puts their names on a waiting list but, being new at the job, only records the first names. How many different ways are there for the waiting list to be written?

**Solution:** Consider some list of the first names, e.g. \((C, D, D, B, A, D)\). There are six full-name lists equivalent to this first-name list:

- \((C, D_X, D_Y, B, A, D_Z)\)
- \((C, D_Y, D_X, B, A, D_Z)\)
- \((C, D_Z, D_X, B, A, D_Y)\)

Notice that we got these by keeping \(A\), \(B\), and \(C\) in the same places, while changing the order of the 3 elements \(D_X\), \(D_Y\), and \(D_Z\). Thus for every first-name list, there are \(P(3, 3) = 3! = 6\) full-name lists equivalent to it. There is a total of \(P(6, 6) = 6!\) full-name lists; so if \(N\) is the number of first-name lists,

\[
N \cdot P(3, 3) = P(6, 6) \quad \Rightarrow \quad N = \frac{P(6, 6)}{P(3, 3)} = \frac{6!}{3!} = 6 \cdot 5 \cdot 4 = 120
\]

You’ve probably seen where this is going. We’ll put it in the form of a theorem.

**Theorem 4.** Given \(n\) objects, of which \(m \leq n\) are identical and the remaining \(n - m\) are all distinct, the number of ways of permuting the objects is

\[
\frac{P(n, n)}{P(m, m)} = \frac{n!}{m!}
\]

**Proof:** We’ll write the identical objects as \(a\) and the distinct objects as \(b_1\) through \(b_{n-m}\). To minimize confusion, we’ll use the word “arrangement” to describe a permutation of the objects in which we can’t distinguish one \(a\) from another. (This is not a formal
Then the number we’re looking for is the number of different arrangements. Consider any one arrangement. This will consist of an ordered list of $a$’s and $b_i$’s. We can go through this list and attach labels to the $a$’s: $a_1$ for the first one in the list, $a_2$ for the second, . . . to $a_m$ for the $m$th. Now we have an ordered list of $n$ objects, all of which are distinct: the $b_i$’s because they started out distinct; the $a_i$’s because we’ve labelled them.

If we leave the $b_i$’s in their original positions and change the order of the $a_i$’s, we obtain a new ordered list of $n$ distinct objects. If we then remove the labels from the $a_i$’s, we get an arrangement of $a$’s and $b_i$’s identical to the original arrangement.

There are $P(m, m) = m!$ such ways of changing the order of the $a_i$’s. Hence for any arrangement of identical $a$’s and distinct $b_i$’s, there are $P(m, m) = m!$ different permutations of $n$ distinct objects (the labelled $a_i$’s and the distinct $b_i$’s) that are all equivalent to the arrangement. Thus if $N$ is the number of possible arrangements,

$$N \cdot P(m, m) = P(n, n) \Rightarrow N = \frac{P(n, n)}{P(m, m)} = \frac{n!}{m!}$$

**Example 17.** At the grocery store, you buy ten items: a pepper, an onion, a cabbage, a box of crackers, and six identical cans of soup. The cashier scans them one at a time, and they appear on the receipt in the order in which they’re scanned. In how many different ways can the items be listed on the receipt?

**Solution:** There are a total of 10 items, of which 4 are distinct and 6 are identical. By Theorem 4, the number of possible receipts is

$$\frac{P(10, 10)}{P(6, 6)} = \frac{10!}{6!} = 10 \cdot 9 \cdot 8 \cdot 7 = 5040$$

There’s no reason why a set of items should include only one class of identical items, with the rest all distinct. What happens if the set contains more than one class of identical items?

**Example 18.** You are arranging five books on a shelf: three red ones and two green ones. How many different ways are there of arranging the colors?

**Solution:** Consider any one arrangement, e.g. $(R, R, G, G)$. We can give each book a label: $(R_1, R_2, R_3, G_1, G_2)$. If we change the order of the red books and then strip off the labels, we get an arrangement identical to the original one. Likewise, if we change the order of the green books, we don’t change the color arrangement. There are $3!$ different orders of the 3 red books, and $2!$ ways to order the green ones; so there are $(3!)(2!)$ permutations of the labelled books that are equivalent to the original color arrangement. There are a total of 5! permutations of the labelled books; so the number of distinct color arrangements is

$$\frac{5!}{(3!)(2!)} = \frac{120}{12} = 10$$
This number is small enough that we can check it by listing all the color arrangements:

- GGRRR
- GRGRR
- GRRGR
- GRRRG
- RGGRR
- RGRGR
- RGRRG
- RRGGR
- RRGRG
- RRRGG

**Theorem 5.** Suppose a set consists of \( n = \sum_{i=1}^{k} n_i \) elements, where for each \( i \) there are \( n_i \) elements identical to one another and different from every other element of the set. Then the number of distinct permutations of elements of the set is

\[
\frac{n!}{(n_1!)(n_2!)(n_3!)(n_4!) \cdots (n_k!)} = \frac{n!}{\prod_{i=1}^{k} n_i!}
\]

**Proof:** We won’t give the proof in detail, since the notation gets cumbersome. Conceptually, it’s a modification of the proof of Theorem 4: we show that every arrangement in which the identical elements aren’t labelled is equivalent to \( \prod_{i=1}^{k} n_i! \) different permutations of \( n \) distinct elements, in which the identical elements have been made distinct by labelling them. Since there are \( n! \) permutations of \( n \) distinct elements, the number of arrangements in which the identical elements have no labels is as given above.

Notice that we didn’t have to allow for a number of unique elements. This is because we can regard each unique element as the single member of a subset with \( n_i = 1 \). Thus if there is only one subset of identical elements with \( n_i = m > 1 \), the formula reduces to that of Theorem 4.

**Example 19.** Two teams, each consisting of five runners, engage in a race. A judge records the team to which each runner belongs, in the order in which they cross the finish line. How many different ways are there for the teams to finish?

**Solution:** We are looking at the permutations of a set of 10, consisting of two subsets of 5 identical elements. By Theorem (5), the number of such permutations is

\[
\frac{10!}{(5!)(5!)} = \frac{10 \cdot 9 \cdot 8 \cdot 7\cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 252
\]

**Example 20.** Ten people arrive at a pizza restaurant just in time to get the last ten slices of pizza: five of pepperoni, four of cheese, and one of mushroom. How many ways are there of distributing the ten slices among the ten people?

**Solution:** We want to find the number of permutations of a set of 10, consisting of three subsets: one consisting of 5 identical elements, one of 4 identical elements and one of 1 element (identical to itself, of course). By Theorem 5, the number is

\[
\frac{10!}{(5!)(4!)(1!)} = \frac{10 \cdot 9 \cdot 8 \cdot 7\cdot 6}{(4 \cdot 3 \cdot 2 \cdot 1)(1)} = 1260
\]
### 4.2.3 Circular permutations

So far, we’ve dealt with permutations in which there’s been a first element, then a second, . . . , then a last element. We can think of these as different ways of arranging the elements in a row. However, there’s another way of arranging elements in which there’s no first or last element.

**Example 21.** Suppose Curly, Larry, and Moe are riding on a small merry-go-round with three identical seats. How many different ways are there of arranging them on the merry-go-round?

Your first inclination is probably to say: \( P(3, 3) = 3! = 6 \), since Curly has a choice of three seats; that leaves Larry with two choices; and once he’s chosen, Moe has only one seat left. However, let’s see what happens when the merry-go-round starts rotating clockwise.

If we were lining the three up in a (first, middle, last) row, then (Curly, Larry, Moe), (Moe, Curly, Larry), and (Larry, Moe, Curly) would be three distinct permutations. However, on the rotating merry-go-round, where there’s no well-defined first or last position, these three permutations are all equivalent.

There are three other (first, middle, last) permutations of the three: (Curly, Moe, Larry), (Larry, Curly, Moe), and (Moe, Larry, Curly). On the merry-go-round, these three are all equivalent to one another:

Notice that there’s no way that we can rotate the merry-go-round to go from the left arrangement in the first figure (Curly, Larry, Moe) to the left arrangement in the second figure (Curly, Moe Larry). Thus there are two distinct ways of arranging the three in a circle, each corresponding to three different ways of arranging the objects in a row.
Example 22. Let’s look at another example of this kind of situation. Suppose we have four colored beads: one blue, one green, one yellow, and one red. We want to string the beads on a necklace. How many different ways are there of arranging the beads?

If we were stringing the 4 beads in a row with a beginning and an end, the answer would be \( P(4, 4) = 4! = 24 \). However, when we connect the ends of the string to form a loop, we find that some of these permutations are equivalent to one another. For example, the permutations (B,G,Y,R), (R,B,G,Y), (Y,R,B,G), and (G,Y,R,B) are all equivalent, since each one can be formed from any of the others by rotating the necklace, as shown in the figure below.

A second set of mutually equivalent permutations, shown in the figure below, is (B,R,Y,G), (G,B,R,Y), (Y,G,B,R), and (R,Y,G,B).

Notice that there is no way to get from the necklace arrangement in the first figure to the arrangement in the second figure by rotating the necklace. (We’re assuming that the necklace can’t be flipped over, only rotated.)

Every distinct necklace arrangement is called a circular permutation. In example 21, there were two circular permutations, representing the two ways of arranging people on the merry-go-round; and each circular permutation corresponded to 3 ordinary permutations of 3 items. In this example, each circular permutation corresponds to 4 ordinary permutations, since every necklace arrangement can be rotated to bring any of the 4 beads to (for example) the upper left position, then read clockwise from that upper left bead to get a distinct ordinary permutation of four objects.

This gives us an approach that we can use to calculate the number of distinct circular permutations. In this example, every ordinary permutation of the four beads corresponds to exactly one circular permutation; and every circular permutation corresponds to four ordinary permutations. Hence if \( N \) is the number of distinct circular permutations, we get

\[
4 \cdot N = P(4, 4) = 4! \quad \Rightarrow \quad N = \frac{4!}{4} = 3! = 6
\]

This number is small enough that we can check our work by listing all 6 circular permutations. We’ll try to do this in a systematic way, by showing each necklace arrangement with the blue bead in the upper left position, then rearranging the other three beads.
Obviously, we can’t get from any one of these necklaces to any other one by rotating, since rotating would remove the blue bead from the upper left position.

This way of writing the different circular permutations suggests another way that we can calculate the total number of different circular permutations. We can read each circular permutation as an ordinary permutation of the four colors, beginning with the upper left element and going clockwise. For example, the upper left circular permutation would be read as (B,G,R,Y). Then we would get six ordinary permutations, each of which would have B as its first element. The six different permutations would be produced by permuting G, R, and Y as the second, third, and fourth elements; so we’d get \( P(3, 3) = 3! = 6 \) of them.

We can generalize this to calculate the number of circular permutations of any number of distinct objects:

**Theorem 6.** The number of circular permutations of \( n \) distinct objects is \((n - 1)\)!

**Proof:** We can prove this in either of two ways. The first is to observe that every ordinary permutation of the \( n \) objects corresponds to exactly one circular permutation: we can write the ordinary permutation as an ordered \( n \)-tuple, then put the first object at the 12 o’clock position, the second object at the next position clockwise, and so forth until the last object is at the first position counterclockwise from the 12 o’clock position. Moreover, every circular permutation corresponds to exactly \( n \) ordinary permutations: given a circular permutation, we can rotate so that any one of the \( n \) items is at the 12:00 position, then write a different ordered \( n \)-tuple by beginning at the 12:00 position and reading clockwise around the circle. Thus if \( N \) is the total number of circular permutations, we get

\[
n \cdot N = P(n, n) = n! \quad \Rightarrow \quad N = \frac{n!}{n} = (n - 1)!
\]

The second approach is to choose one of the \( n \) items and declare that in writing our circular permutations, that item will always go at the 12:00 position. Then we have \( n - 1 \) items left to distribute among the \( n - 1 \) remaining positions on the circle. There are \( P(n - 1, n - 1) = (n - 1)! \) ways to distribute these items in these positions. We can’t get from one such distribution to another by rotating the circle, since that would remove our designated item from the 12:00 position; so each such distribution represents a different circular permutation.
Example 23. At a series of peace talks, representatives of six countries sit around a circular table. How many different ways are there of arranging the representatives at the table?

**Solution:** This is a case of circular permutations, so we apply Theorem 6.
\[ N = (6 - 1)! = 5! = 120 \]

Example 24. A team of eight skydivers join hands to form a ring while in free fall. How many different ways are there to arrange the ring?

**Solution:** Again, this is a question about circular permutations, since there’s no evidence of a first or last position in the ring. Use Theorem 6.
\[ N = (8 - 1)! = 7! = 5040 \]

### 4.2.4 Combinations

Now it’s time to consider the second major class of sampling-without-replacement counting: combinations. With permutations, the order of elements in a subset matters: (Jacob, Esau) is not the same permutation as (Esau, Jacob). With combinations, order doesn’t matter: both of these permutations represent the same combination \{Esau, Jacob\}.

Combinations are very common in real-world probability and counting situations. For example, in most lotteries, a set of numbers is drawn from a container; but the order in which the numbers are drawn doesn’t matter—just which numbers are drawn and which aren’t. Similarly, in many card games, a set of cards is dealt to each player, but the order in which a player gets the cards doesn’t matter: all that matters is which cards the player holds when the deal is finished.

Example 25. A very small town decides to hold a lottery. Because participation is expected to be low, people playing the lottery circle three numbers on a list running from 1 through 10. How many such combinations are possible?

**Solution:** For every possible three-number combination, there are \(P(3, 3) = 3! = 6\) permutations. For example, the combination \{1, 2, 3\} is equivalent to any of the following six permutations:

\[
(1,2,3) \quad (2,1,3) \quad (3,1,2) \\
(1,3,2) \quad (2,3,1) \quad (3,2,1)
\]

Since each combination corresponds to 3! permutations, and since each permutation of the 10 numbers taken 3 at a time corresponds to exactly one combination, we can calculate the number \(N\) of distinct combinations from

\[
N \cdot 3! = P(10, 3) \Rightarrow N = \frac{P(10, 3)}{3!} = \frac{10!}{(10-3)!3!} = \frac{10!}{7!3!} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120
\]

We will modify this argument to prove a general theorem about the number of combinations.
Theorem 7. The number of combinations consisting of \( r \) elements drawn from a set containing \( n \) elements, where \( r \leq n \), is

\[
\binom{n}{r} = \frac{n!}{r!(n-r)!}
\]  

(4.1)

Proof: Consider an \( r \)-element combination taken from a set with \( n \) elements: that is, a subset \( \{a_1, a_2, \ldots, a_r\} \). The elements of the subset can be written in \( r! \) different orders. Thus each \( r \)-element combination corresponds to \( r! \) permutations of the \( n \) elements taken \( r \) ways. Each of the \( P(n, r) \) permutations of the \( n \) elements taken \( r \) at a time corresponds to exactly one \( r \)-element combination. Hence if \( N \) is the number of distinct \( r \)-element combinations,

\[ N \cdot r! = P(n, r) \Rightarrow N = \frac{P(n, r)}{r!} = \frac{n!}{r!(n-r)!} \]

The expression on the left-hand side of equation (4.1) is the standard notation for the number of \( r \)-item combinations of \( n \) items. In words, it is usually expressed as “\( n \) choose \( r \)”. The different \( \binom{n}{r} \)'s are also known as binomial coefficients, since they appear in the binomial theorem:

\[
(a + b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}
\]

Note that \( \binom{n}{r} \) is not a fraction. Do not make the mistake of writing it with a horizontal bar, or of misreading it as \( n/r \). The parentheses are part of the notation, and should never be omitted.

Notice that \( \binom{n}{r} = \binom{n}{n-r} \). This makes sense: when we choose a subset of \( r \) elements out of a set with \( n \) elements, we are implicitly choosing a second subset with \( n-r \) elements, consisting of those elements that we don’t choose for the first subset. For example, if there are 10 people and only 8 doughnuts, then in selecting the 8 people who get doughnuts, we’re also designating the \( 10 - 8 = 2 \) people who don’t get one.

Example 26. An instructor teaching a class of 30 students asks for two volunteers to help with a demonstration. How many two-volunteer combinations are possible?

Solution: This is a straightforward application of Theorem 7:

\[
\binom{30}{2} = \frac{30!}{2!28!} = \frac{30 \cdot 29}{2 \cdot 1} = 435
\]

Example 27. An exam consists of 11 questions. To pass, you need to get at least 7 right. How many ways are there of getting exactly 7 questions right?

Solution: Apply Theorem 7:

\[
\binom{11}{7} = \frac{11!}{7!4!} = \frac{11 \cdot 10 \cdot 9 \cdot 8}{4 \cdot 3 \cdot 2 \cdot 1} = 330
\]
Example 28. A bridge hand consists of 13 cards dealt from a deck of 52. How many different bridge hands are possible?

Solution: You will probably want a calculator or equivalent for this problem:

\[
\binom{52}{13} = \frac{52!}{13!39!} \approx 6 \times 10^{11}
\]

Example 29. In the New Jersey state lottery’s “Pick-6” game, players mark six numbers on a ticket containing the numbers 1 through 49. How many possible ways of marking a ticket are there?

This is a combination problem, so you can apply Theorem 7:

\[
\binom{49}{6} = \frac{49!}{6!43!} \approx 14,000,000
\]

Example 30. In the Vermont state lottery’s “Gimme 5” game, players mark five numbers on a ticket containing the numbers 1 through 39. How many possible ways of marking a ticket are there?

The order of the numbers doesn’t matter, so this is a combination problem.

\[
\binom{39}{5} = \frac{39!}{5!34!} = 575,757
\]

4.3 Review problems

Here are some problems to test your understanding of the combinatorics topics that we’ve covered. In these problems, we won’t tell you in advance whether they involve ordered samples from multiple sets, permutations, or combinations. You’ll have to decide for yourself what kind of situation a problem involves, and which formula(s) to use to solve it.

Example 31. A high-school teacher draws up a seating chart for a class, assigning each student to a particular desk. There are 10 students in the class, and 15 desks in the room. How many different seating charts are possible?

Solution: This is a situation in which it matters which student is at which desk. We can regard each seating chart as a separate permutation of 15 desks taken 10 at a time. Then the number of possible seating charts is

\[
P(15, 10) = \frac{15!}{10!} = 15 \cdot 14 \cdot 13 \cdot 12 \cdot 11 = 360,360
\]
Example 32. Eight co-workers get together every Friday evening to play volleyball. Each week, they try to find a different way of dividing into two four-person teams. How many weeks can they play before they have to repeat a team arrangement?

Solution: In this situation, the order in which people are picked for teams presumably doesn’t matter: all that matters is who’s on one team and who’s on the other. It suffices to name the four members of one team: once we’ve specified that, we know who’s on the other team. Hence we want to know how many four-person combinations we can choose out of a set of eight.

\[
\binom{8}{4} = \frac{8!}{4!4!} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = 70
\]

Example 33. A Ferris wheel has twelve identical cars. If twelve parties show up to ride it, how many ways are there of distributing them among the cars?

Solution: This is a question of circular permutations, since there’s no first or last position on the Ferris wheel. Use Theorem 6:

Number of circular permutations = \((n-1)! = (12-1)! = 11! \approx 4 \times 10^7\)

Example 34. Three friends arrive at a bar that has 30 different beers on tap. How many different ways are there for the three people to order beers?

Solution: In this situation, one person’s ordering a particular beer doesn’t keep the next person from ordering the same kind. Hence each person has 30 choices, and the number of possible orders is

\[(30)(30)(30) = 30^3 = 27,000\]

Example 35. The National Assembly of Tyrannia consists of 50 members: 23 from the People’s Party, 22 from the Popular Party, and 5 from the Pirate Party. In how many different ways can the 50 seats in the Assembly Chamber be divided among the parties?

Solution: We can visualize this situation as one in which party labels are assigned to specific seats: 23 identical “People’s” labels, 22 identical “Popular” labels, and 5 identical “Pirate” labels. Hence this is a permutation of a set of 50 items, consisting of three subsets of identical items; it calls for Theorem 5.

\[
\frac{50!}{23!22!5!} \approx 8.7 \times 10^{18}
\]

Example 36. Sixty students have to take a particular math course, taught by two professors. Professor Pushover’s class has room for 20. Professor Cruel’s class has room for 40. How many ways are there of assigning the students to the classes?

Solution: There are two equally legitimate approaches to this problem, and they produce identical answers. The first is to look at each class assignment as a permutation of a set with 60 items: 20 identical P’s and 40 identical C’s. (Students presumably don’t care about the order in which they get assigned to a class; just the class that they’re assigned to.) The second approach is to note that by assigning 20 students to Professor
Pushover’s class, we’re implicitly assigning the remaining 40 to Professor Cruel’s. Each set of students assigned to Prof. P’s class is a combination of 20 elements taken from a set of 60. Either way, we get the formula
\[
\frac{60!}{20! \cdot 40!} \approx 4.2 \times 10^{15}
\]

**Example 37.** How many ways are there of arranging six distinct slices of pizza on a circular plate?

**Solution:** This is a question about circular permutations, since the plate has no first or last position.

Number of circular permutations = \((n - 1)! = (6 - 1)! = 5! = 120\)

**Example 38.** A World War II movie is about 12 people on a sinking ship. There is a single lifeboat, with room for 8 people. In how many different ways can the movie turn out?

**Solution:** This is a question of combinations, not permutations: presumably, the characters only care whether they get on the lifeboat, and not about the order in which they do so. We are looking at the number of 8-person combinations drawn from a total set of 12 people.

\[
\binom{12}{8} = \frac{12!}{8! \cdot 4!} = \frac{12 \cdot 11 \cdot 10 \cdot 9}{4 \cdot 3 \cdot 2 \cdot 1} = 495
\]

**Example 39.** Twelve people are taking a road trip in three vehicles: a pickup with room for two people, a hatchback with four seats, and a van with six seats. How many different ways are there of assigning people to vehicles?

**Solution:** The order in which people are assigned to cars doesn’t appear to matter; all that’s important is who’s assigned to which car. We can treat each assignment as a permutation of a set containing 2 identical P’s, 4 identical H’s, and 6 identical V’s.

\[
\frac{12!}{2! \cdot 4! \cdot 6!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7}{(2 \cdot 1)(4 \cdot 3 \cdot 2 \cdot 1)} = 13,860
\]

**Example 40.** A Thai restaurant has 15 items on the menu. When you order something, you are asked how spicy you want it, on a scale of 1 through 5. How many different kinds of orders are possible?

**Solution:** This is like a two-step experiment. The particular item you order doesn’t constrain your spiciness choice; so the total number of possible orders is \(15 \times 5 = 75\).

**Example 41.** A dentist has seven time slots available on a certain day. Four people call and make appointments for that day. In how many different ways can the patients be scheduled?

**Solution:** This is a permutation question: presumably, it matters to the patients when they’re scheduled. The number of different schedules is

\[
P(7, 4) = \frac{7!}{3!} = 7 \cdot 6 \cdot 5 \cdot 4 = 840
\]
Example 42. A group of five people arrive at a restaurant that specializes in pie. When they arrive, there are five pieces of apple pie, six of cherry, and six of pumpkin. If each of the five people can only eat one piece, how many different ways are there for them to order?

Solution: Careful! The critical thing to notice here is that the smallest number of pieces of any one kind of pie (the five of apple) is greater than or equal to the number of people in the group. This means that anyone in the group can order any kind of pie, regardless of what others have already ordered. In other words, each of the five people has three choices. Thus the total number of orders possible is

\[3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 = 3^5 = 243\]

Example 43. To play the “Powerball” lottery, one marks five different numbers on a list from 1 through 59; on a separate list, one marks a single number from 1 through 35. How many ways are there of playing?

Solution: We can regard this as a two-step experiment. In the first step, one selects a combination of 5 items from a set of 59. In the second step, one selects a single item from a set of 35. If \(N_1\) is the number of outcomes of the first step, and \(N_2\) is the number of outcomes of the second step, then the total number of outcomes is

\[N = N_1N_2 = \binom{59}{5} \times 35 = \frac{(59!)(35)}{5!54!} \approx 175,000,000\]
5 Introduction to the concept of probability

5.1 What is probability?

*Probability is the art of counting and measuring*

Probability theory is the branch of mathematics that is concerned with random (or chance) phenomena.

Any experiment that always has the same outcome under a specified set of conditions is called *deterministic*. An example of a deterministic experiment is that of dropping a ball in a vacuum and timing its fall. If we always drop the ball from the same height, we expect that it will always take the same amount of time to reach the ground\(^3\).

An experiment that does not always produce the same outcome, even under a specified set of conditions, is said to be *non-deterministic*, *random*, or *stochastic*. An example of such an experiment is flipping an unbiased coin and recording heads or tails. Even if you use the same coin and always start the flip with heads up, you not be able to predict whether you’ll get heads or tails on a given flip.

Most experiments are not purely probabilistic or purely deterministic, but lie somewhere between the two. Even in our deterministic ball-dropping experiment, we’d expect to see some random variation of the result if we measured the time to the nearest millionth or billionth of a second owing to the fact that we could never exactly reproduce the conditions under which the ball is dropped. In our non-deterministic coin-tossing experiment, if we knew enough about the forces applied to the coin, we could predict whether it would come up heads or tails. In practice, we use a combination of the deterministic and the probabilistic approach to most experiments. We can predict the results with some accuracy using the deterministic approach, but there will always be some random experimental error, which we can only deal with using probabilistic concepts.

Experience has shown that many random phenomena exhibit a statistical regularity that makes them subject to study. For example, if we toss a coin a 100 times, the number of heads in the 100 tosses will appear to be erratic. However, if we toss the coin \(n\) times and count the number \(k\) of heads, we will find that as \(n\) tends toward infinity, the ratio \(k/n\) will tend toward \(1/2\). Knowing that fact won’t help us make a prediction about the

\(^3\text{Comment: As every physics lab instructor knows, even in such a simple experiment has uncertainty and in it! The ball will never be released in quite the same way, the yard stick will add uncertainty because it will be hard to consistently measure the distance that the ball falls through, etc. In the real world every thing has a little bit of indeterministic behavior. However, this does not negate the fact that to a high degree of accuracy we can predict the time it will take a ball, starting from rest, to fall through a distance } h. \text{ In fact, the formula falling a distance } h \text{ dropped from rest is}

\[t_{\text{fall}} = \sqrt{\frac{2h}{g}}\]

and it doesn’t get anymore deterministic than this!
outcome of any particular flip of the coin, though. The most we can say is that we have a 50-50 chance of getting heads each time we flip the coin. Thus, we use this relative frequency interpretation to define the probability of getting heads on any flip of the coin. The probability is taken to be 1/2.

5.1.1 The concept of relative frequencies

Relative frequency interpretation of probabilities: If an experiment is repeated \( n \) times, where \( n \) is large, and a certain outcome occurs on \( k \) of the \( n \) tries, then we say that the outcome has probability \( k/n \).

Notice that no outcome of an experiment that is repeated \( n \) times can occur fewer than 0 times, or more than \( n \) times: \( 0 \leq k \leq n \). Thus, since the probability that a certain outcome occurs is taken to be the limit of the ratio

\[
\frac{k}{n} = \frac{\text{number of times the outcome occurs in } n \text{ experiments}}{n},
\]

we see that

\[
0 \leq \frac{k}{n} \leq 1.
\]

In other words, the probability of any outcome is

1. always greater than or equal to 0;
2. always less than or equal to 1.

Life insurance actuaries use relative frequencies to estimate the probability that you will die this year given your age, sex, location, health history, etc. This will determine how much they charge you for your life insurance policy. Engineers and physicists use relative frequency by repeating an experiment over and over to estimate the probability of a certain event, or to compare the expected outcome (i.e., the theoretical prediction) to that of the actual measured outcome.

**Definition 1.** The set of all possible outcomes of an experiment is called the sample space for the experiment. The outcomes in the sample space are called the sample points.

**Notation:** We will use the capital Greek letter omega, \( \Omega \), to denote the sample space. The sample space is the universal set, usually denoted as \( U \) in set theory. Some books use \( S \) to denote the sample space. This author believes that this is a poor choice since \( S \) is universally used throughout probability to denote success.

**Example 1.** The sample space for the experiment of tossing a coin 3 times is given by the set \( \Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\} \), where \( H \)=heads and \( T \)=tails.
Definition 2. An event is a subset of the sample space of an experiment. An event $E$ is said to occur if the outcome of the experiment is an element of $E$.

Example 2. Write down the event $E$ of getting one head and two tails after 3 tosses.

Solution: We must write down every way that we can get one head and two tails in three tosses: $E = \{HTT, THT, TTH\}$.

Definition 3. Two events $E$ and $F$ are mutually exclusive (disjoint) if $E \cap F = \emptyset$.

Definition 4. The events $E_1, \ldots, E_n$ are mutually exclusive if no two of the events can occur at the same time. In symbols, $E_i \cap E_j = \emptyset$ for $i \neq j$.

Notice that the event of getting one head and two tails is mutually exclusive from the event of getting two heads and one tail.

5.1.2 The connection between probability, mass, and Venn diagrams

We draw an analogy between sets and pieces of wood. We will associate a rectangular homogeneous piece of wood having a mass of one unit with the sample space $\Omega$. The molecules that make up the wood (the mass) are the elements of the set. Then if we cut a circle out of the piece of wood and call it $A$, the molecules that make up the wooden disc $A$ correspond to the elements (outcomes) of the set $A$. We wish to assign a probability to this set.

We associate the mass of a piece of wood with its probability. The mass of the sample space is always defined to be 1 unit, since it is the universal set. The probability mass of the wooden disc $A$ is denoted by $\mathcal{P}(A)$. The disc $A$ must weigh something, so its mass is greater than zero; and its mass is less than 1 unit, since it is not all of $\Omega$. It follows that $\mathcal{P}(A) \in (0, 1)$.

Next, we introduce the formal properties of probability. The probability-mass analogy will be used extensively when explaining certain probabilistic relations like properties 1-5 below.

\[ A^c = \Omega - A \]

4Notice that since gravity is constant, there is a one-to-one correspondence between weight and mass. It is for this reason that we will use the two interchangeably even though there is a physical difference between the two.
5.2 Fundamental properties of probability for sample spaces

Let $\Omega$ be the sample space for a given experiment. We use the notation $\mathcal{P}(E)$ to denote the probability of the event $E$. Then:

1. For any event $E$, $0 \leq \mathcal{P}(E) \leq 1$.

2. If $E$ is certain to happen, then $\mathcal{P}(E) = 1$. In particular, $\mathcal{P}(\Omega) = 1$.
   Also $\mathcal{P}(\Omega^c) = \mathcal{P}(\emptyset) = 0$.

3. If $E_1, \ldots, E_n$ are mutually exclusive events, then $\mathcal{P}(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} \mathcal{P}(E_i)$.
   In particular, for the case $n = 2$, if $E$ and $F$ are mutually exclusive events, then $\mathcal{P}(E \cup F) = \mathcal{P}(E) + \mathcal{P}(F)$.

4. For any two events $E$ and $F$, $\mathcal{P}(E \cup F) = \mathcal{P}(E) + \mathcal{P}(F) - \mathcal{P}(E \cap F)$.

5. For any event $E$, $\mathcal{P}(E^c) = 1 - \mathcal{P}(E)$.

Comments:

- The sample space $\Omega$ and events $E_i$ are sets. The probability of any event (set) is always a number between 0 and 1.

- Don’t forget De Morgan’s Laws and the distributive property for sets when computing probabilities.

- The probability measure $\mathcal{P}$ is a function from the power set of $\Omega$ to the interval $[0, 1]$. (i.e. $\mathcal{P} : \mathcal{P}(\Omega) \rightarrow [0, 1]$).

5.3 Finite sample spaces

Consider an experiment that has a sample space given by $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Notice that $\Omega$ is a finite sample space with $n$ elements (sample points). Using the definition of union we can write $\{\omega_1, \omega_2, \ldots, \omega_n\} = \{\omega_1\} \cup \{\omega_2\} \cup \cdots \cup \{\omega_n\}$. If the probability of each sample point (event) is equally likely to occur, then we can write $p = \mathcal{P}(\omega_i)$, for $i = 1, \ldots, n$. Since the events are mutually exclusive ( $\{\omega_i\} \cap \{\omega_j\} = \emptyset$, for $i \neq j$ ) it follows from properties 2 and 3 above that

\[
1 = \mathcal{P}(\Omega) \\
= \mathcal{P}(\{\omega_1, \ldots, \omega_n\}) \\
= \mathcal{P}(\{\omega_1\} \cup \cdots \cup \{\omega_n\}) \\
= \mathcal{P}(\{\omega_1\}) + \cdots + \mathcal{P}(\{\omega_n\}) \\
= p + \cdots + p \quad \text{(there are } n \text{ terms)} \\
= np
\]
Solving for \( p \), we find that \( p = 1/n \). More generally, if \( E = \{\omega_1, \omega_2, \ldots, \omega_k\} \) is any event of a sample space \( \Omega \), then \( \mathcal{P}(E) = \mathcal{P}(\{\omega_1\} \cup \{\omega_2\} \cup \cdots \cup \{\omega_k\}) = \mathcal{P}(\{\omega_1\}) + \mathcal{P}(\{\omega_2\}) + \cdots + \mathcal{P}(\{\omega_k\}) \).

**Theorem 1.** If an experiment can result in any one of \( N \) different and equally likely outcomes, and if \( K \) of these outcomes together constitute the event \( A \), then the probability of \( A \) is
\[
\mathcal{P}(A) = \frac{K}{N} = \frac{\#(A)}{\#(\Omega)}.
\]

**Proof:** Let the sample space be \( \Omega = \{\omega_1, \ldots, \omega_N\} \). Since all of the outcomes are equally likely, let
\[
p = \mathcal{P}(\omega_1) = \cdots = \mathcal{P}(\omega_N)
\]
If we write \( \{\omega_1, \ldots, \omega_N\} = \{\omega_1\} \cup \cdots \cup \{\omega_N\} \), we get
\[
1 = \mathcal{P}(\Omega)
= \mathcal{P}(\{\omega_1\} \cup \cdots \cup \{\omega_N\})
= \mathcal{P}(\{\omega_1\}) + \cdots + \mathcal{P}(\{\omega_N\}) \quad \text{(since if } i \neq j, \{\omega_i\} \cap \{\omega_j\} = \emptyset)\n= p + \cdots + p \quad \text{\(N\) terms}
= Np.
\]
Solving for \( p \) gives
\[
p = \frac{1}{N}
\]

**Note:** When events are equally likely it greatly simplifies the calculations of the probabilities because all the computations just reduce to a counting problem.

Now, let \( A = \{a_1, \ldots, a_K\} \) be any event consisting of \( K \) of these mutually exclusive outcomes. (Each \( a_i \in \{\omega_1, \cdots, \omega_N\} \), but it need not be true that \( a_i = \omega_i \).) Since we can write \( A \) as the union of the \( K \) disjoint elements, we get
\[
\mathcal{P}(A) = \mathcal{P}(\{a_1\} \cup \cdots \cup \{a_K\})
= \mathcal{P}(\{a_1\}) + \cdots + \mathcal{P}(\{a_K\}).
\]
Since \( a_1, \ldots, a_K \in \Omega \), and since the probability of each element of \( \Omega \) is \( 1/N \), we find
\[
\mathcal{P}(A) = \frac{1}{N} + \cdots + \frac{1}{N} \quad \text{\(K\) terms}
= \frac{K}{N}
\]

**Example 3.** Consider the case \( N = 3 \) and \( K = 2 \). This set is small enough that we can easily write down each subset explicitly and find its corresponding probability. For the case of three elements:
\[
\Omega = \{\omega_1, \omega_2, \omega_3\}.
\]
There are only three possible events $A$ that consist of $K = 2$ outcomes $\{\omega_i, \omega_j\}$:

- $A_1 = \{\omega_1, \omega_2\} = \{\omega_1\} \cup \{\omega_2\}$,
- $A_2 = \{\omega_1, \omega_3\} = \{\omega_1\} \cup \{\omega_3\}$,
- $A_3 = \{\omega_2, \omega_3\} = \{\omega_2\} \cup \{\omega_3\}$.

Since the three outcomes $\omega_1$, $\omega_2$, and $\omega_3$ are equally likely it follows that

$$P(\{\omega_1\}) = P(\{\omega_2\}) = P(\{\omega_3\}) = \frac{1}{3}.$$

Lastly, we compute the probabilities of each $A_i$:

- $P(A_1) = P(\{\omega_1, \omega_2\}) = P(\{\omega_1\}) + P(\{\omega_2\}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} = \frac{K}{N}$,
- $P(A_2) = P(\{\omega_1, \omega_3\}) = P(\{\omega_1\}) + P(\{\omega_3\}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} = \frac{K}{N}$,
- $P(A_3) = P(\{\omega_2, \omega_3\}) = P(\{\omega_2\}) + P(\{\omega_3\}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} = \frac{K}{N}.$
5.4 Rules for transforming words into math

1. **Not** event $A$: $A^c$.

2. Events $A$ and $B$ both occur: $A \cap B$, the intersection of $A$ and $B$.

3. Event $A$ or $B$ occurs: $A \cup B$, the union of $A$ and $B$.

4. $A \cap B \subseteq A \cup B$. This means that $A \cap B$ is generally a smaller set than $A \cup B$ (though they can be equal). This is because membership in $A \cap B$ is more restrictive than membership in $A \cup B$.

5. Exactly one of the events $A$ or $B$ occurs.
   This is the *exclusive or*. It means that $A$ or $B$ occurs, but not both. In other words, the event that $A$ and $B$ occur is not allowed. In terms of sets,
   \[
   \text{Exactly one event:}
   \]
   \[
   (A \cup B) - (A \cap B)
   \]
   \[
   = (A - B) \cup (B - A)
   \]
   \[
   \text{Note: If you see the word and connected with two events, think intersection. If you see the word or in connection with two events, think union.}
   \]

6. Neither event $A$ nor event $B$ occurs: this is the same as “$A$ does not occur and $B$ does not occur” = $A^c \cap B^c$.

7. When solving probability word problems:
   - **Step 1:** Write down the events of interest in the problem.
   - **Step 2:** Write down the given information in terms of the events.
   - **Step 3:** Write down what you’re looking for.

5.5 Some dos and don’ts when computing the probability of sets

Let $E$, $F$, and $G$ be events (sets). Below are examples of correct and common incorrect ways of combining events:

<table>
<thead>
<tr>
<th>Correct</th>
<th>Incorrect</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E \cup F = G$</td>
<td>$E + F = G$</td>
<td>(can’t add sets)</td>
</tr>
<tr>
<td>$P(E) + P(F) = 0.8$</td>
<td>$P(E) \cup P(F) = 0.8$</td>
<td>(union applies to sets, not numbers)</td>
</tr>
<tr>
<td>$P(E^C) = 1 - P(E)$</td>
<td>$E^C = 1 - E$</td>
<td>(can’t mix sets and numbers)</td>
</tr>
<tr>
<td>$P(E^C)$</td>
<td>$P(E)^C$</td>
<td>(can’t take the complement of a number)</td>
</tr>
</tbody>
</table>
5.6 Application probability problems and their solutions

Example 4. You own a restaurant that gets inspected periodically. Your restaurant is open 7 days a week. What is the probability of having an inspection on a given day? What if your restaurant was only open Monday-Friday (i.e., 5 days a week), how would the probability of an inspection change?

Solution: Take the sample space to be the days of the week:

\[ \Omega = \{ \text{Sun}, \text{Mon}, \text{Tues}, \text{Wed}, \text{Thurs}, \text{Fri}, \text{Sat} \}. \]

The probability that an inspection will happen on a particular day is equally likely for all the days. We’ll denote that probability by \( p \). Then

\[ p = \mathcal{P}(\{\text{Sun}\}) = \mathcal{P}(\{\text{Mon}\}) = \cdots \mathcal{P}(\{\text{Sat}\}) \]

Then

\[ 1 = \mathcal{P}(\Omega) = \mathcal{P}(\{\text{Sun}, \text{Mon}, \ldots, \text{Sat}\}) = \mathcal{P}(\{\text{Sun}\} \cup \{\text{Mon}\} \cup \cdots \cup \{\text{Sat}\}) = \mathcal{P}(\{\text{Sun}\}) + \cdots + \mathcal{P}(\{\text{Sat}\}) \]

(since the subsets \{Sun\},...,\{Sat\} are disjoint)

\[ = p + p + \cdots + p \quad (7 \text{ times}) \]

\[ = 7p \]

When we solve for \( p \), we get

\[ p = \frac{1}{7} \]

For the second part of the question, let \( E \) be the event that the inspection is on a weekday. Then

\[ E = \{ \text{Mon, Tues, \ldots, Fri} \} = \{ \text{Mon} \} \cup \cdots \cup \{ \text{Fri} \} \]

\[ \mathcal{P}(E) = \mathcal{P}(\{\text{Mon}\} \cup \cdots \cup \{\text{Fri}\}) = \mathcal{P}(\{\text{Mon}\}) + \cdots + \mathcal{P}(\{\text{Fri}\}) = p + p + \cdots + p = 5p \]

\[ = 5 \left( \frac{1}{7} \right) = \frac{5}{7} \]

Example 5. You are awaiting an acceptance response from your dream graduate school for physics. They have informed you that you will hear from them in January, February, or March if you are to be accepted. Write down the sample space for all of the events together with a description of the events. How many elements are in the power set?

Solution: We’ll abbreviate: \( J=\text{Jan}, F=\text{Feb}, M=\text{Mar} \). Then \( \Omega = \{J, F, M\} \). We need to list all the events of \( \Omega \), i.e. all the subsets of \( \Omega \).
Event | Description of event
---|---
$\emptyset = \{}$ & No offer in January, February, or March (out of luck)
$\{J\}$ & offer in January
$\{F\}$ & offer in February
$\{M\}$ & offer in March
$\{J,F\}$ & offer in January or February
$\{J,M\}$ & offer in January or March
$\{F,M\}$ & offer in February or March
$\{J,F,M\}$ & offer in January, February, or March

We have listed all the elements in the power set of $\Omega$. $\#(\Omega) = 3$, so the number of elements in the power set is

$$\#(\mathcal{P}(\Omega)) = 2^{\#(\Omega)} = 2^3 = 8.$$

**Example 6 (Rolling a pair of dice).** A pair of dice are rolled and the face values of the die are recorded. The outcome of each die is $\{1, 2, 3, 4, 5, 6\}$. Let $x =$ face value of the first die and $y =$ face value of the second die. Then we can express the outcome of the pair of dice as the ordered pair $(x, y)$ for $x = 1, 2, 3, 4, 5, 6$ and $y = 1, 2, 3, 4, 5, 6$. The sample space is shown below.

$$\Omega = \{(1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (1, 2), (2, 2), (3, 2), (4, 2), (5, 2), (6, 2), (1, 3), (2, 3), (3, 3), (4, 3), (5, 3), (6, 3), (1, 4), (2, 4), (3, 4), (4, 4), (5, 4), (6, 4), (1, 5), (2, 5), (3, 5), (4, 5), (5, 5), (6, 5), (1, 6), (2, 6), (3, 6), (4, 6), (5, 6), (6, 6)\}$$

Find the probability of the following events:

- $E$: the first face is odd.
- $F$: the sum of faces is greater than 8
- $G$: the sum of the faces is even

**Solution:** $\Omega = \{(i, j) | i \in \{1, \ldots, 6\}, j \in \{1, \ldots, 6\}\}$ We write out the 36 elements of $\Omega$ in the format $(i, j) \rightarrow i + j$:

$(1, 1) \rightarrow 2$ $(1, 2) \rightarrow 3$ $(1, 3) \rightarrow 4$ $(1, 4) \rightarrow 5$ $(1, 5) \rightarrow 6$ $(1, 6) \rightarrow 7$ $(2, 1) \rightarrow 3$ $(2, 2) \rightarrow 4$ $(2, 3) \rightarrow 5$ $(2, 4) \rightarrow 6$ $(2, 5) \rightarrow 7$ $(2, 6) \rightarrow 8$ $(3, 1) \rightarrow 4$ $(3, 2) \rightarrow 5$ $(3, 3) \rightarrow 6$ $(3, 4) \rightarrow 7$ $(3, 5) \rightarrow 8$ $(3, 6) \rightarrow 9$ $(4, 1) \rightarrow 5$ $(4, 2) \rightarrow 6$ $(4, 3) \rightarrow 7$ $(4, 4) \rightarrow 8$ $(4, 5) \rightarrow 9$ $(4, 6) \rightarrow 10$ $(5, 1) \rightarrow 6$ $(5, 2) \rightarrow 7$ $(5, 3) \rightarrow 8$ $(5, 4) \rightarrow 9$ $(5, 5) \rightarrow 10$ $(5, 6) \rightarrow 11$ $(6, 1) \rightarrow 7$ $(6, 2) \rightarrow 8$ $(6, 3) \rightarrow 9$ $(6, 4) \rightarrow 10$ $(6, 5) \rightarrow 11$ $(6, 6) \rightarrow 12$

Let $E$ be the event that the first face is odd:
Let $E$ be the event that the sum of faces is greater than 8:

$$E = \{(1,x),(3,x),(5,x)\mid x \in \{1, \ldots, 6\}\}$$

Let $F$ be the event that the sum of faces is greater than 8:

$$F = \{(x,y)\mid x+y > 8\}$$

The elements of $F$ are enclosed in a dashed line above.

Let $G$ be the event that the sum of the faces is even:

$$G = \{(x,y)\mid x+y \text{ is even}\}$$

$$\mathcal{P}(E) = \frac{\text{# of elements having the form } (1,x), (3,x), \text{ or } (5,x)}{\text{# elements in } \Omega} = \frac{18}{36} = \frac{1}{2}$$

$$\mathcal{P}(F) = \frac{\text{# } \{(6,3),(5,4),(6,4),(4,5),(5,5),(6,5),(3,6),(4,6),(5,6),(6,6)\}}{\text{# elements in } \Omega} = \frac{10}{36} = \frac{5}{18}$$

$$\mathcal{P}(G) = \frac{18}{36} = \frac{1}{2}$$

**Example 7.** You have just opened a liquor store in downtown East St. Louis. Records kept by the city indicate that there is a 90% chance that a business will be robbed within its first year of opening, a 70% chance that a business will fail within its first year of being open, and a 60% chance that a business will both fail and be robbed within its first year.

(a) What is the probability that your store will not be robbed in its first year?

(b) What is the probability that your store will not be robbed in its first year and will not fail in its first year?

**Solution:**

**Step 1.** Define the events:

\[
\begin{align*}
R &= \text{business is robbed in it’s first year of opening} \\
F &= \text{business fails during the first year}
\end{align*}
\]

**Step 2.** Given:

\[
\begin{align*}
\mathcal{P}(R) &= 0.9 \\
\mathcal{P}(F) &= 0.7 \\
\mathcal{P}(R \cap F) &= 0.6 \\
\mathcal{P}(R \cup F) &= \mathcal{P}(R) + \mathcal{P}(F) - \mathcal{P}(R \cap F) = 0.9 + 0.7 - 0.6 = 1
\end{align*}
\]

**Step 3.**

(a) We want the probability of the event not $R$, which is just $R^c$.

(b) We want the probability of the event ($R^c$ and $F^c$): the business does not get robbed in its first year, and it does not fail during its first year.
For part (a) we want
\[ \mathcal{P}(R^c) = 1 - \mathcal{P}(R) = 1 - 0.9 = 0.1. \]

For part (b) recall that when we use the word “and” in connection with events, we take the intersection \( \cap \). Thus we want
\[
\begin{align*}
\mathcal{P}(R^c \cap F^c) &= \mathcal{P}((R \cup F)^c) \\
&= 1 - \mathcal{P}(R \cup F) \\
&= 1 - 1 \\
&= 0
\end{align*}
\]

This says that you have no probability of success!

**Example 8.** You have just bought a new home. In the next five years, the probability that you will have to replace the heating system is 30%; the probability that you will have to replace the roof is 20%. There is a probability of 10% that you will have to replace both.

(a) What is the probability that you will not have to replace the roof, but not the heating system in the next 5 years?

(b) What is the probability that neither the heating system nor the roof will have to be replaced in the next 5 years?

**Solution:**

**Step 1.** Define the events:
\[
\begin{align*}
R &= \text{replace roof in the next 5 years} \\
H &= \text{replace heating system in the next 5 years}
\end{align*}
\]

**Step 2.** Given:
\[
\begin{align*}
\mathcal{P}(R) &= 0.2 \\
\mathcal{P}(H) &= 0.3 \\
\mathcal{P}(R \cap H) &= 0.1 \\
\mathcal{P}(R \cup F) &= \mathcal{P}(R) + \mathcal{P}(H) - \mathcal{P}(R \cap H) = 0.2 + 0.3 - 0.1 = 0.4
\end{align*}
\]

**Step 3.**

(a) We want the probability of the event \( R \) and **not** \( H \) (i.e., \( R - H \)), which is \( R \) and \( H^c = R \cap H^c = R - H \).

(b) We want the probability of the event: the roof is not replaced in the next 5 years, and the heating system is not replaced in the next 5 years, which is the event: \( R^c \) and \( H^c = R^c \cap H^c \).

For part (a) we have
\[
P(R \cap H^c) = P(R - H) = P(R) - P(R \cap H) = 0.2 - 0.1 = 1/10.
\]

Example 9. The probability that a faculty member will get a raise in any particular year is 0.68, the probability that they will get their own office is 0.41, and the probability that they will get a raise or their own office is 0.85.

(a) Find the probability that they get a raise and their own office.
(b) What is the probability that the faculty member gets neither a raise nor their own office?

Solution:

Step 1. Define the events:

\[
\begin{align*}
R &= \text{get raise} \\
R^c &= \text{don’t get raise} \\
O &= \text{get own office} \\
O^c &= \text{don’t get own office}
\end{align*}
\]

Step 2. Write down what you know. Given:

\[
\begin{align*}
P(R) &= 0.68 & P(R^c) &= 1 - P(R) = 1 - 0.68 = 0.32 \\
P(O) &= 0.41 & P(O^c) &= 1 - P(O) = 1 - 0.41 = 0.59 \\
P(R \cup O) &= 0.85
\end{align*}
\]

The last is the probability of getting a raise or one’s own office. “Or” corresponds to the union of events.

Step 3. Write down what you want.

For part (a), we want the probability of getting a raise and one’s own office. “And” corresponds to the intersection of events.

\[
P(R \cap O) = P(R) + P(O) - P(R \cup O) = 0.68 + 0.41 - 0.85 = 0.24
\]

For part (b), the first two steps are the same as in the previous example. Now, we want the probability that the faculty member gets neither a raise nor their own office, which is to say that they do not get a raise (not \(R\)) and do not get their own office (not \(O\)).

Neither \(R\) nor \(O\) \(=\) \(R^c\) and \(O^c\) \(=\) \(R^c \cap O^c\).
We want: \( P(R^c \cap O^c) \)
\[ = P((R \cup O)^c) \quad \text{(by DeMorgan’s laws)} \]
\[ = 1 - P(R \cup O) \]
\[ = 1 - 0.85 \]
\[ = 0.15 \]

**Example 10.** You have applied for an internship at Bing and at Google. The companies have told you that you probability of getting a job at Bing is 0.3 and at job at Google is 0.5. Typically, only 10% of applicants get both offers so you estimate the probability of getting both jobs is 0.1.

(a) What is the probability of getting exactly one job offer?

(b) What is the probability of getting neither of the job offers?

**Solution:**

Define the following events:

- \( B = \) Job at Bing
- \( G = \) Job at Google
- \( B^c = \) No job at Bing
- \( G^c = \) No job at Google
- \( B \cup G = \) Job at Bing or Google
- \( B \cap G = \) Job at Bing and Google

\[ P(B) = 0.3, \quad P(G) = 0.5, \quad P(B \cap G) = 0.1 \]

For part (a), we want the probability of getting exactly one of the jobs. The set \( B \cup G \) is the event of getting at least one of the jobs. It can be broken down into three parts:

- \( B - G = \) Job at Bing, no job at Google
- \( G - B = \) No job at Bing, job at Google
- \( G \cap B = \) Job at Bing and job at Google.

We need to exclude the third possibility, so we subtract out the intersection \( B \cap G \). That leaves us with two moon-shaped regions: \( B - G \) and \( G - B \).

\[ P(\text{exactly one job}) = P(B \cup G) - P(B \cap G) \]
\[ = P(B) + P(G) - 2P(B \cap G) \]
\[ = 0.3 + 0.5 - 2(0.1) = 0.6 \]

For part (b), the event “neither job” means (not \( B \)) and (not \( G \)) both occur. In other words, \( B^c \) and \( G^c \) occur. When we see “and”, we think of the intersection: \( B^c \cap G^c \)

\[ P(B^c \cap G^c) = P((B \cup G)^c) \quad \text{(by DeMorgan’s laws)} \]
\[ = 1 - P(B \cup G) \]
\[ = 1 - 0.7 \quad \text{(from the given information)} \]
\[ = 0.3 \]
**Example 11.** Your defense contracting company is planning a missile test in 3 days. The test will have to be postponed if the rain and winds get too extreme. The weather models predict that the probability of rain is 0.3, the probability of high winds is 0.4, and the probability of rain or high winds is 0.5. Your boss has asked you to calculate the following probabilities.

(a) What is the probability that it neither rains nor has high winds?

(b) What is the probability that it rains, but there are no high winds?

(c) What is the probability that there are high winds, but no rain?

**Solution:**

**Step 1:** Define the events:

- $R = \text{rain}$
- $R^c = \text{no rain}$
- $H = \text{high winds}$
- $H^c = \text{no high winds}$

**Step 2:** Write down the given information:

- $\mathcal{P}(R) = 0.3$
- $\mathcal{P}(H) = 0.4$
- $\mathcal{P}(R \cup H) = 0.5$

($R \cup H = \text{rain or high winds}$)

**Step 3:** Write down what you want.

For part (a), we want the probability that there is neither rain nor high winds.

\[
\mathcal{P}(R^c \cap H^c) = \mathcal{P}((R \cup H)^c) = 1 - \mathcal{P}(R \cup H) = 1 - 0.5 = 0.5
\]

For part (b), we write the event “$R$ and not $H$” as

\[
R \cap H^c = R - H
\]

We want $\mathcal{P}(R \cap H^c)$. To get it, we need an expression involving $R - H$. We can get one by dividing the event $R$ into two disjoint sets: $R \cap H^c = R - H$, and $R \cap H$. These two sets form a partition for the set $R$. Thus

\[
\mathcal{P}(R) = \mathcal{P}((R \cap H^c) \cup (R \cap H)) \quad \text{(since the union of the two sets is $R$)}
\]

\[
= \mathcal{P}(R \cap H^c) + \mathcal{P}(R \cap H) \quad \text{(since the two sets are disjoint)}
\]

Rearranging this gives us

\[
\mathcal{P}(R \cap H^c) = \mathcal{P}(R) - \mathcal{P}(R \cap H) \quad \text{(5.1)}
\]
To compute $\mathcal{P}(R \cap H)$, use the equation

$$\mathcal{P}(R \cup H) = \mathcal{P}(R) + \mathcal{P}(H) - \mathcal{P}(R \cap H)$$

Substituting the information from Problem 13 yields

$$0.5 = 0.3 + 0.4 - \mathcal{P}(R \cap H)$$

Solving this equation get us

$$\mathcal{P}(R \cap H) = 0.7 - 0.5 = 0.2$$

Substituting this into equation (5.1) produces

$$\mathcal{P}(R \cap H^c) = 0.3 - 0.2 = 0.1$$

For part (c), we want the probability of event $H$ and not event $R$: $\mathcal{P}(H \cap R^c) = \mathcal{P}(H - R)$. This is similar to part (b), with the roles of $H$ and $R$ reversed. Using the same result from part (b), with the roles of $H$ and $R$ reversed:

$$\mathcal{P}(H \cap R^c) = \mathcal{P}(H) - \mathcal{P}(H \cap R) = 0.4 - 0.2 = 0.2$$

## 6 Conditional probability and independence

### 6.1 Introduction to conditional probability

Suppose we roll a fair dice and it is covered up with a cup before we are able to see the outcome. We are told to choose a number \{1, \ldots, 6\} and so we choose 1. Since all outcomes are equally likely, the chance of getting a 1 in the top face is 1/6. However, if someone lifts up the cup and tells us that the face value is odd, this new information changes the probability that the face value is one.

The probability that the top face is 1 is now 1/3. Thus, if we had placed a bet of $100 that the top face was 1, we would now have increased our odds for winning. On the other hand, if we had placed our money on a face value of 2, then we would already know we lost our money. If we were given a chance to pick a new number we certainly would! And we’d pick an odd number.

The new information increased our knowledge about the possible value of the outcome. In fact, it changed the probability for the outcome of the experiment. But notice that it did not change the actual outcome since the person lifting the cup does not touch the dice. This is similar in spirit to the famous Monte-Hall problem.
We can express this mathematically as follows. Let $E_1$ be the event that the face value of the dice is 1. Let $E_{\text{odd}}$ be the event that the face value on the dice is an odd number. Then

$$P(E_1) = \frac{\# \text{ of outcomes in } E_1}{\# \text{ of possible outcomes in sample space}} = \frac{\# \{1\}}{\# \{1, 2, 3, 4, 5, 6\}} = \frac{1}{6}$$

$$P(E_1 \text{ given that } E_{\text{odd}} \text{ occurred}) = \frac{\# \text{ of outcomes in } (E_1 \cap E_{\text{odd}})}{\# \{1, 3, 5\}} = \frac{1}{3},$$

where we have introduced the following notation:

**Notation:** $\#(A)$ = number of outcomes (elements) in the set $A$.

Notice that since the sample space shrunk for the case of the conditional probability, the denominator became smaller, and hence the fraction (conditional probability) became larger. This also agrees with intuition.

**Definition:** When the probability of an event $E$ is conditioned on the fact that an event $F$ has happened, it is called a conditional probability and it is denoted by $P(E | F)$, the probability of $E$ given $F$.

If $P(F) \neq 0$, then

$$P(E | F) = \frac{P(E \cap F)}{P(F)} \quad (6.1a)$$

For any events $E$ and $F$, it follows from (6.1a) that

$$P(E \cap F) = P(F)P(E | F), \quad (6.1b)$$

$$P(E \cap F) = P(E)P(F | E). \quad (6.1c)$$

If $\Omega$ is a finite sample space and the events are all equally likely, then

$$\frac{P(E \cap F)}{P(F)} = \frac{\#(E \cap F)}{\#(F)} = \frac{\#(E \cap F)}{\#(F)}. \quad (6.2)$$

Thus, $F$ behaves as the "new" sample space $\Omega$. This is because we are told that event $F$ occurs, so our universe of all allowable outcomes has just shrunk! In other words, for $P(E | F)$ we use our new information that event $F$ has occurred to justify restricting the outcomes of the experiment to only those outcomes in the event $F \subset \Omega$. This should not be too surprising. Observe that every time we write the probability of an event $E$ as $P(E)$ that this is shorthand for $P(E | \Omega)$.

Recall that an event is a subset of the sample space $\Omega$ (universal set). It is implicitly assumed when we say an event $E$ occurred, that what we really mean is the outcomes of $E$ are composed of a subset of the collection of all the outcomes $\Omega$ of a given experiment. That is, given that all possible outcomes are contained in $\Omega$, $P(E)$ is just the probability that $E$ occurred given that $\Omega$ occurred: $P(E | \Omega)$. 

Example 1. Your defense contracting company is planning a missile test. The test will have to be postponed if the rain and wind become too extreme in the vicinity of the test site. The weather models predict that the probability of rain is 0.3, the probability of high winds is 0.4, and the probability of rain or high winds is 0.5. From this information, if you wake up on that day and find that there are high winds, what would be your new estimate for the probability of rain?

Step 1: Define the events: let

\[
\begin{align*}
\{ R &= \text{event that it rains} \\
R^c &= \text{event of no rain} \\
H &= \text{event of high winds} \\
H^c &= \text{event of no high winds}
\end{align*}
\]

Step 2: Write down the given information.

\[
\begin{align*}
\mathcal{P}(R) &= 0.3 \\
\mathcal{P}(H) &= 0.4 \\
\mathcal{P}(R \cup H) &= 0.5
\end{align*}
\]

Step 3: Write down what you want to compute. We want the probability of high winds, given that it is raining:

\[
\mathcal{P}(H \mid R) = \frac{\mathcal{P}(W \cap R)}{\mathcal{P}(R)}.
\]

In order to compute this probability we need to compute

\[
\mathcal{P}(R \cap H) = \mathcal{P}(R) + \mathcal{P}(H) - \mathcal{P}(R \cup H) = 0.3 + 0.4 - 0.5 = 0.2.
\]

Thus, the desired conditional probability is

\[
\mathcal{P}(H \mid R) = \frac{\mathcal{P}(W \cap R)}{\mathcal{P}(R)} = \frac{0.2}{0.3} = \frac{2}{3}.
\]
6.1.1 The idea behind conditional probability

Suppose you are in your instructor’s office begging for extra credit. The instructor rolls a die and puts a cup over it, then tells you to guess the number from the set

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$ 

If you guess correctly, you win extra points.

Not being able to see the die, and not having any reason to believe it will come up with one value over another, you guess 1. Then the instructor looks under the cup and tells you that the number is even.

Since 1 isn’t an even number, you’d clearly like to be able to change your guess. But before you knew that the number was even, you’d have had no reason to change.

Let $X$ be the face value of the die. Let’s look at the time line of events, in which you get additional information about $X$ from knowing that the value is even.

Before the instructor announced that $X$ was even, the sample space was

$$\Omega_{\text{before}} = \{1, 2, 3, 4, 5, 6\}.$$ 

After the instructor tells you that the face value is even, the sample space changes. The outcomes $X = 1, 3, \text{ and } 5$ are no longer possible. The new sample space is

$$\Omega_{\text{after}} = \{2, 4, 6\}.$$ 

Now, let’s look at what happens to the probability of individual outcomes like $X = 1$ as a result of the new information, and the sample space’s change from $\Omega_{\text{before}}$ to $\Omega_{\text{after}}$.

Let $E$ be an event, such as $X = 1$, whose probability we want to calculate. The probability of $E$ is usually expressed as $P(E)$, with the understanding that we mean the probability of event $E \subseteq \Omega$, where $\Omega$ is the sample space of the experiment. But in this new situation, there’s potential for confusion: does $\Omega$ mean $\Omega_{\text{before}}$ or $\Omega_{\text{after}}$?

To avoid confusion, we’ll write

$$P(E|\Omega_{\text{before}}) \quad \text{or} \quad P(E|\Omega_{\text{after}})$$

The expression $P(E|F)$ is read “the probability of $E$ given $F$” or “the probability of $E$ conditioned on $F$”. It signifies the probability of $E$ if we know that the event $F$ has definitely occurred.

In our dice example, if $E$ is the event of rolling a 1, then $E = \{1\}$ and

$$P(E|\Omega_{\text{before}}) = \frac{\#(E)}{\#(\Omega_{\text{before}})} = \frac{1}{6}.$$
We’re tempted to say that

\[ P(E|\Omega_{\text{after}}) = \frac{\#(E)}{\#(\Omega_{\text{after}})} = \frac{1}{3} \quad \text{WRONG!} \]

Why does this formula work in the first case, but not in the second? The problem is that in the second case, \( E \not\subseteq \Omega_{\text{after}} \). The only way that \( E \) can occur if we know that \( \Omega_{\text{after}} \) has occurred is if there’s an outcome \( \omega \) that’s a member of both \( E \) and \( \Omega_{\text{after}} \):

\[ \omega \in E \cap \Omega_{\text{after}}. \]

The formula we need to use is

\[ P(E|\Omega_{\text{after}}) = \frac{\#(E \cap \Omega_{\text{after}})}{\#(\Omega_{\text{after}})} = 0 \]

This formula works with \( \Omega_{\text{before}} \) as well; but since \( \Omega_{\text{before}} \) is the whole solution space, \( E \subseteq \Omega_{\text{before}} \), so \( E \cap \Omega_{\text{before}} = E \).

A Venn diagram illustrating the conditional probability of \( P(E|F) \). Notice that the single set \( F \) is filled with dots, whereas the intersection of \( E \) and \( F \) is filled with larger dots.

The Venn diagram above shows how conditional probability works. If we don’t know whether or not the event \( F \) occurs, then the probability of \( E \) is represented by the area of the circle labelled \( E \), divided by the area of the whole probability space \( \Omega \). If we do know that \( F \) took place, then the sample space shrinks down to the circle labelled \( F \). Then the conditional probability of \( E \) given \( F \) is represented by the area of the lens-shaped region \( E \cap F \), divided by the area of the new sample space \( F \). Below is a cartoon that shows how the sample space shrinks after the occurrence of event \( F \).

Figure 41: Venn diagram showing conditional probability in action. Once the event \( F \) occurs, the sample space is shrunk.
If $\Omega$ consists of a finite number of outcomes, each of which is equally likely, then we calculate probabilities by counting. If $E \subseteq \Omega$ is an event, then 

$$
P(E) = \frac{\#(E)}{\#(\Omega)}
$$

If we know that the event $F$ occurred, then the conditional probability of $E$ given $F$ is 

$$
P(E|F) = \frac{\#(E \cap F)}{\#(F)}
$$

This formula is useful, but let’s emphasize its major limitation: it is only valid if all outcomes are equally likely. There are lots of important probabilistic situations that don’t meet that rather stringent requirement. For example, the probability that someone will be involved in a car accident is much higher if that someone is a 17-year-old male with a muscle car than if it is a 46-year-old suburban mother in a Volvo station wagon.

It would be very helpful if we had a formula that we could apply even when outcomes aren’t all equally likely. Fortunately, we can use our knowledge of basic probability to devise such a formula.

We’ll begin with equation (6.3) and try to rewrite it so that the right-hand side no longer uses the counting function $\#(\ )$.

$$
P(E|F) = \frac{\#(E \cap F)}{\#(F)}
$$

Divide numerator and denominator by $\#(\Omega)$:

$$
= \frac{\#(E \cap F)}{\#(\Omega)} \cdot \frac{\#(\Omega)}{\#(F)}
= \frac{\#(E \cap F)}{\#(F)}
= \frac{P(E \cap F)}{P(F)}
$$

This equation was derived from the case in which all outcomes are equally likely. However, it’s valid even when that’s not true. We can see that by looking at the Venn diagrams above, and considering the mass interpretation of probability:

$$
P(E|F) = \frac{\text{mass of } E \cap F}{\text{mass of } F} = \frac{P(E \cap F)}{P(F)}
$$

Proving the formula for conditional probability rigorously takes a lot of higher-level math. Basically, though, it uses our counting and our probability-mass arguments and makes them very formal. We’ll skip all that, and just define it:
Definition 5. Given a sample space \( \Omega \) and events \( E \subseteq \Omega \) and \( F \subseteq \Omega \), the conditional probability of \( E \) given \( F \) is
\[
\mathcal{P}(E|F) = \frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)}
\] (6.4)

Notice that we are dividing by \( \mathcal{P}(F) \), which means we’re assuming that \( \mathcal{P}(F) \neq 0 \). We can multiply both sides of (6.4) by \( \mathcal{P}(F) \) to get
\[
\mathcal{P}(E \cap F) = \mathcal{P}(E|F) \mathcal{P}(F)
\]

If we have data concerning the real-world occurrences of the events \( F \) and \( E \cap F \), we can use the relative-frequency method to calculate \( \mathcal{P}(F) \) and \( \mathcal{P}(E \cap F) \), then use those numbers in equation (6.4) to calculate \( \mathcal{P}(E|F) \). Actuaries do this all the time: when they calculate the probability that someone of your age and sex, living in your hometown and driving your kind of car, will get in a major accident in the next year, they are using the relative-frequency method to determine probabilities. Thus your insurance rates depend on the study of probability.

6.2 Introduction to independent events

Sometimes knowing that an event \( F \) happens does not change the probability that an event \( E \) will happen. In this case, event \( E \) occurs independently of event \( F \).

Example You flip a fair coin twice. Let \( F \) be the event that you get heads on the first toss and \( E \) be the event in which you get heads on the 2nd toss. Then \( \mathcal{P}(E|F) = 1/2 \) and \( \mathcal{P}(E) = 1/2 \). In this case we say that the two events are independent since \( \mathcal{P}(E|F) = \mathcal{P}(E) \). From (6.1a) we see that if \( \mathcal{P}(F) \neq 0 \) and if \( \mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F) \), then
\[
\mathcal{P}(E|F) = \frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)} = \frac{\mathcal{P}(E)\mathcal{P}(F)}{\mathcal{P}(F)} = \mathcal{P}(E).
\] (6.5)

This insight leads to the following more general definition.

Definition: Two events \( E \) and \( F \) are independent if
\[
\mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F).
\] (6.6)

Note: Non-independent events are said to be dependent.

We can generalize the notion of independent events to any finite collection of events. \( E_1, E_2, \ldots, E_n \) are said to be independent events if, for every subset of indices
\[
\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, n\}, \text{ where } r \leq n
\]
the events \( E_{i_1}, E_{i_2}, \ldots, E_{i_r} \) are independent. In particular, if the events \( E_1, E_2 \) and \( E_3 \) are independent, then

\[
\begin{align*}
\mathcal{P}(E_1 \cap E_2 \cap E_3) &= \mathcal{P}(E_1)\mathcal{P}(E_2)\mathcal{P}(E_3) \\
\mathcal{P}(E_1 \cap E_2) &= \mathcal{P}(E_1)\mathcal{P}(E_2) \\
\mathcal{P}(E_1 \cap E_3) &= \mathcal{P}(E_1)\mathcal{P}(E_3) \\
\mathcal{P}(E_2 \cap E_3) &= \mathcal{P}(E_2)\mathcal{P}(E_3)
\end{align*}
\] (6.7a)

In general, the intersection of three events that are not necessarily independent have many possible forms (3! = 6, where the symbol ! stands for factorial). Three of the six possible combinations are shown below:

\[
\begin{align*}
\mathcal{P}(E_1 \cap E_2 \cap E_3) &= \mathcal{P}(E_1)\mathcal{P}(E_2|E_1)\mathcal{P}(E_3|E_1 \cap E_2), \\
\mathcal{P}(E_1 \cap E_2 \cap E_3) &= \mathcal{P}(E_3)\mathcal{P}(E_2|E_3)\mathcal{P}(E_1|E_2 \cap E_3), \\
\mathcal{P}(E_1 \cap E_2 \cap E_3) &= \mathcal{P}(E_2)\mathcal{P}(E_1|E_2)\mathcal{P}(E_3|E_1 \cap E_2).
\end{align*}
\] (6.8a)

6.2.1 The idea behind independence

Suppose you flip a fair coin. The probability of getting heads is 1/2. Now, suppose you first roll a die, look at the result, then flip the fair coin. What is the probability of getting heads now? It’s still 1/2—the rolling of the die would not have changed anything involving the flipping of the coin. If you knew that the die had come up with an odd number, or even if you’d known the exact value on the die, it wouldn’t have helped you in guessing whether the coin had come up heads or tails.

In such a case, the events are said to be independent. Two events \( E \) and \( F \) are independent if knowing whether or not \( E \) occurs doesn’t help you predict whether or not \( F \) occurs, and vice versa. More technically:

**Definition 6.** Two events \( E \) and \( F \) are independent if

\[
\mathcal{P}(E|F) = \mathcal{P}(E)
\]

or, equivalently, if

\[
\mathcal{P}(F|E) = \mathcal{P}(F)
\]

or, equivalently, if

\[
\mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F)
\]
If any one of these equations is true, the other two automatically follow from it. For example, suppose that the first equation is true:

$$\mathcal{P}(E|F) = \mathcal{P}(E).$$

If we substitute the above expression into the formula for conditional probability we get

$$\frac{\mathcal{P}(E \cap F)}{\mathcal{P}(F)} = \mathcal{P}(E).$$

Next, multiplying both sides by $\mathcal{P}(F)$ yields the third equation:

$$\mathcal{P}(E \cap F) = \mathcal{P}(E)\mathcal{P}(F).$$

This proves that the truth of the third equation follows from the truth of the first equation. Now, use the formula for conditional probability again:

$$\mathcal{P}(F|E) = \frac{\mathcal{P}(F \cap E)}{\mathcal{P}(E)}.\quad (6.10a)$$

Remember that intersection is commutative: $F \cap E = E \cap F$. Substitute the third equation, and cancel $\mathcal{P}(E)$:

$$\mathcal{P}(F|E) = \frac{\mathcal{P}(F)\mathcal{P}(E)}{\mathcal{P}(E)} = \mathcal{P}(F).$$

This proves the second equation.

### 6.3 Combining independence and conditional probability

The notion of independent events can be extended to conditional probabilities. The underlying reason why this works is based on the fact that $\mathcal{P}(\cdot|F)$ is a probability on the reduced sample space $F \subseteq \Omega$ in the same way that $\mathcal{P}(\cdot)$ is a probability on the sample space $\Omega$.

**Definition:** Let $E_1$, $E_2$, and $E_3$ be three events. $E_1$ and $E_2$ are independent given event $E_3$ occurred if

$$\mathcal{P}(E_1 \cap E_2|E_3) = \mathcal{P}(E_1|E_3)\mathcal{P}(E_2|E_3).\quad (6.9)$$

We shall refer to this as *conditional independence*.

Similarly, events $E_1$, $E_2$, and $E_3$ are independent given $E_4$ if

$$\mathcal{P}(E_1 \cap E_2 \cap E_3|E_4) = \mathcal{P}(E_1|E_4)\mathcal{P}(E_2|E_4)\mathcal{P}(E_3|E_4),\quad (6.10a)$$

$$\mathcal{P}(E_1 \cap E_2|E_4) = \mathcal{P}(E_1|E_4)\mathcal{P}(E_2|E_4),\quad (6.10b)$$

$$\mathcal{P}(E_1 \cap E_3|E_4) = \mathcal{P}(E_1|E_4)\mathcal{P}(E_3|E_4),\quad (6.10c)$$

$$\mathcal{P}(E_2 \cap E_3|E_4) = \mathcal{P}(E_2|E_4)\mathcal{P}(E_3|E_4).\quad (6.10d)$$
6.3.1 Applications of conditional probability and independence

Example 2 (the double-dice problem). Consider the experiment of rolling a fair die twice. We’ll use this classic example to examine a number of questions regarding conditional probabilities and independence of three sets, which we’ll define below.

To describe an outcome of the experiment, we’ll use the ordered pairs 
\[(x, y) = \text{ (face value of first roll, face value of second roll) }\].

The set of all outcomes is 
\[\Omega = \{(x, y) \mid x \in \{1, \ldots, 6\}, y \in \{1, \ldots, 6\}\}.\]

We define the following events:

- **E**: The first roll is odd \((x=1, 3, \text{ or } 5)\).
  \[E = \{(x, y) \mid x \in \{1, 3, 5\}\}.\]

- **F**: The sum of the two rolls is greater than 8.
  \[F = \{(x, y) \mid x + y > 8\}.\]

- **G**: The sum of the two rolls is even.
  \[G = \{(x, y) \mid x + y \in \{2, 4, 6, 8, 10, 12\}\}.\]

The elements of \(\Omega\) are written below, along with the sums of the two rolls for each outcome.

\[
\begin{array}{cccccccc}
(1,1) & \rightarrow & 2 & (1,2) & \rightarrow & 3 & (1,3) & \rightarrow & 4 \\
(2,1) & \rightarrow & 3 & (2,2) & \rightarrow & 4 & (2,3) & \rightarrow & 5 \\
(3,1) & \rightarrow & 4 & (3,2) & \rightarrow & 5 & (3,3) & \rightarrow & 6 \\
(4,1) & \rightarrow & 5 & (4,2) & \rightarrow & 6 & (4,3) & \rightarrow & 7 \\
(5,1) & \rightarrow & 6 & (5,2) & \rightarrow & 7 & (5,3) & \rightarrow & 8 \\
(6,1) & \rightarrow & 7 & (6,2) & \rightarrow & 8 & (6,3) & \rightarrow & 9 \\
\end{array}
\]

We’ll start by counting the elements in each set: \(E, F,\) and \(G\).

In each column, notice that three of the six elements have an odd first number in the pair: \((1, y), (3, y), (5, y)\). Thus \(#(E) = 3 + 3 + 3 + 3 + 3 + 3 = 18\): half of the 36 elements in \(\Omega\).

We have drawn dotted lines around the elements for which \(x + y > 8\). There are ten such elements; so \(#(F) = 10\).

If we go through and count, we find that for half of the elements of \(\Omega, x + y\) is even. Thus \(#(G) = 18\).

Notice that all outcomes are equally likely. For each \((x, y) \in \Omega, P(\{(x, y)\}) = \frac{1}{36}.\)
Next, we’ll need to compute \( |E \cap F| \), \( |E \cap G| \), and \( |F \cap G| \). It’s not hard to find \( E \cap F \): look for elements inside the dashed lines that have an odd number for their first component.

\[
E \cap F = \{ (x, y) \mid x \in \{1, 3, 5\} \text{ and } x + y > 8 \} \\
= \{ (5, 4), (5, 5), (5, 6), (3, 6) \}
\]

To find \( E \cap G \), we need to look at elements with an odd number for the first component (i.e. elements in the first, third, or fifth row of the table) and an even sum. There are three even sums in each of the three rows, so

\[
|E \cap G| = 3 + 3 + 3 = 9
\]

Lastly, we find \( F \cap G \) by looking at elements inside the dashed lines that have components whose sum is even. There are only four of these:

\[
F \cap G = \{ (x, y) \mid x + y \text{ is even and } x + y > 8 \} \\
= \{ (6, 4), (5, 5), (4, 6), (6, 6) \}
\]

In summary, we’ve found that

\[
\begin{align*}
|E \cap F| &= 4 & |E| &= 18 & |\Omega| &= 36 \\
|E \cap G| &= 9 & |F| &= 10 \\
|F \cap G| &= 4 & |G| &= 18
\end{align*}
\]

Since all outcomes are equally likely, we can calculate the probability of an event \( A \) as

\[
P(A) = \frac{|A|}{|\Omega|}.
\]

If we are looking at the ratio of the probabilities of two events, say \( A \) and \( B \), then we can cancel \( |\Omega| \) and substitute the ratio of the element counts of the two sets:

\[
\frac{P(A)}{P(B)} = \frac{|A|/|\Omega|}{|B|/|\Omega|} = \frac{|A|}{|B|}
\]

Using the information that we computed above we can now answer the following questions.

**Question 1.** What is \( P(E \mid F) \) and \( P(F \mid E) \)?

**Solution:**

\[
P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{|E \cap F|}{|F|} = \frac{4}{10} = \frac{2}{5}
\]

\[
P(F \mid E) = \frac{P(F \cap E)}{P(E)} = \frac{|F \cap E|}{|E|} = \frac{4}{18} = \frac{2}{9}
\]
Question 2. What is $\mathcal{P}(E \mid G)$ and $\mathcal{P}(G \mid E)$?

Solution:

$$\mathcal{P}(E \mid G) \equiv \frac{\mathcal{P}(E \cap G)}{\mathcal{P}(G)} = \frac{\#(E \cap G)}{\#(G)} = \frac{9}{18} = \frac{1}{2}$$

$$\mathcal{P}(G \mid E) \equiv \frac{\mathcal{P}(G \cap E)}{\mathcal{P}(E)} = \frac{\#(G \cap E)}{\#(E)} = \frac{9}{18} = \frac{1}{2}$$

Question 3. What is $\mathcal{P}(F \mid G)$ and $\mathcal{P}(G \mid F)$?

Solution:

$$\mathcal{P}(F \mid G) \equiv \frac{\mathcal{P}(F \cap G)}{\mathcal{P}(G)} = \frac{\#(F \cap G)}{\#(G)} = \frac{4}{18} = \frac{2}{9}$$

$$\mathcal{P}(G \mid F) \equiv \frac{\mathcal{P}(G \cap F)}{\mathcal{P}(F)} = \frac{\#(G \cap F)}{\#(F)} = \frac{4}{10} = \frac{2}{5}$$

For the next three questions, we need to compute the probabilities of the events:

Solution:

$$\mathcal{P}(E) = \frac{\#(E)}{\#(\Omega)} = \frac{18}{36} = \frac{1}{2}$$

$$\mathcal{P}(E \cap F) = \frac{\#(E \cap F)}{\#(\Omega)} = \frac{4}{36} = \frac{1}{9}$$

$$\mathcal{P}(F) = \frac{\#(F)}{\#(\Omega)} = \frac{10}{36} = \frac{5}{18}$$

$$\mathcal{P}(E \cap G) = \frac{\#(E \cap G)}{\#(\Omega)} = \frac{9}{36} = \frac{1}{4}$$

$$\mathcal{P}(G) = \frac{\#(G)}{\#(\Omega)} = \frac{18}{36} = \frac{1}{2}$$

$$\mathcal{P}(F \cap G) = \frac{\#(F \cap G)}{\#(\Omega)} = \frac{4}{36} = \frac{1}{9}$$

Question 4. Are $E$ and $F$ independent?

Solution: No.

$$\mathcal{P}(E)\mathcal{P}(F) = \frac{1}{2} \cdot \frac{5}{18} = \frac{5}{36} \neq \frac{4}{36} = \mathcal{P}(E \cap F)$$

Question 5. Are $E$ and $G$ independent?

Solution: Yes.

$$\mathcal{P}(E)\mathcal{P}(G) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = \mathcal{P}(E \cap G)$$

Question 6. Are $F$ and $G$ independent?

Solution: No.

$$\mathcal{P}(F)\mathcal{P}(G) = \frac{5}{18} \cdot \frac{1}{2} = \frac{5}{36} \neq \frac{4}{36} = \mathcal{P}(F \cap G)$$
Example 3 (Urn problem). An urn contains 9 white balls and 1 black ball all mixed up at random. You are given the falling selection algorithm: randomly select a ball from the urn, record its color, and return it back to the urn. The balls are then mixed up again and you make another selection. This process of selecting balls is repeated for a total of three times. This process is known as sampling with replacement. Answer the following questions involving this selection process.

(a) What is the probability of getting 3 white balls on the 3 draws?

(b) What is the probability of getting a black ball on the first draw, followed by two white balls on the last two draws?

(a) What is the probability of exactly one black in the three draws?

Solution:

Step 1: Define the events: let

\[ B = \text{event of choosing a black ball on any given draw} \]
\[ B^c = \text{event of not choosing a black ball on any given draw} \]

Notice that since there are only white and black balls in the urn, the event of choosing a non-black ball is the same as the event of choosing a white ball.

Since we replace the ball back into the urn that we drew before we draw again, each selection is independent because the sample space is returned to its original state; and \( \mathcal{P}(B_i) \) is the same for every value of \( i = 1, 2, 3 \). To make this easier to read, we’ll replace \( \mathcal{P}(B_i) \) and \( \mathcal{P}(B_i^c) \) with \( \mathcal{P}(B) \) and \( \mathcal{P}(B^c) \) when the subscript \( i \) is unnecessary.

Step 2: Write down the given information.

\[
\begin{align*}
\mathcal{P}(B) &= 1/10 \\
\mathcal{P}(B^c) &= 1 - \mathcal{P}(B) = 9/10
\end{align*}
\]

Step 3: Write down what you are trying to find.

For part (a), we want the probability of selecting a white ball on the first, second, and third draws.

\[
\mathcal{P}(B_1^c \cap B_2^c \cap B_3^c) = \mathcal{P}(B_1^c)\mathcal{P}(B_2^c)\mathcal{P}(B_3^c) = (\mathcal{P}(B^c))^3 = \left( \frac{9}{10} \right)^3
\]

For part (b), the event of selecting a black ball, then a white ball, then another white ball is \( B \cap B^c \cap B^c \). By independence,

\[
\mathcal{P}(B \cap B^c \cap B^c) = \mathcal{P}(B)\mathcal{P}(B^c)\mathcal{P}(B^c) = \frac{1}{10} \cdot \frac{9}{10} \cdot \frac{9}{10} = \frac{9^2}{10^3}.
\]
For part (c), Let $E$ be the event of getting exactly one black ball. Then $E$ can be written as the union of three mutually exclusive events: getting the single black ball part on the first selection; getting it on the second selection; or getting it on the third selection.

$$E = \{B \cap B^c \cap B^c\} \cup \{B^c \cap B \cap B^c\} \cup \{B^c \cap B^c \cap B\}.$$ 

$$\mathcal{P}(E) = \mathcal{P}(\{B \cap B^c \cap B^c\} \cup \{B^c \cap B \cap B^c\} \cup \{B^c \cap B^c \cap B\})$$

$$= \mathcal{P}(B \cap B^c \cap B^c) + \mathcal{P}(B^c \cap B \cap B^c) + \mathcal{P}(B^c \cap B^c \cap B)$$

(We can do this because the events are disjoint)

$$= \mathcal{P}(B)\mathcal{P}(B^c)\mathcal{P}(B^c) + \mathcal{P}(B^c)\mathcal{P}(B)\mathcal{P}(B^c) + \mathcal{P}(B^c)\mathcal{P}(B^c)\mathcal{P}(B)$$

(because the three selections are independent)

$$= 3\mathcal{P}(B)(\mathcal{P}(B^c))^2$$

$$= 3\left(\frac{1}{10}\right)\left(\frac{9}{10}\right)^2$$
7 Bayes’ Theorem

7.1 Partitioning the sample space

Before we can learn about Bayes’ theorem we must first learn how to partition the sample space. Partitioning a sample space is just a way of cutting-up/dividing-up the sample space. It’s similar to dividing up the inheritance of someone’s estate to the relatives (with the exception that the elements of a set that are being divided up don’t fight over their partition of the estate!): each person receives a set amount of the total estate, and the sum of the parts of the estate equals the total estate. Here I am also including the lawyers take in all of this!

Let $\Omega$ be the sample space for an experiment. If the events $B_1, B_2, \ldots, B_n \subseteq \Omega$ are such that the sets are pairwise disjoint and the union of the sets $B_i$, for $i = 1, \ldots, n$, is the entire sample space then, the collection of the $B_i$’s are said to be a partition of the sample space. We now give the formal definition.

**Definition:** Let $\Omega$ be the sample space for an experiment. If the events $B_1, B_2, \ldots, B_n \subseteq \Omega$ are such that

(i) $B_i \cap B_j = \emptyset$ for $i \neq j$, (the sets are pairwise disjoint), and

(ii) $\bigcup_{i=1}^{n} B_i = \Omega$, (the union of all the sets $B_i$ is the entire sample space),

then the collection of the $B_i$’s are said to be a *partition* of the sample space.

**The Idea:** Think of the set $\Omega$ as being a piece of wood. The members of the set $\Omega$ are then the molecules that compose the piece wood. We then cut the piece of wood ($\Omega$) into $n$ pieces. Each of the $n$ pieces is a subset $B_i$. The pieces are clearly disjoint, since any molecule in the piece of wood $B_i$ is not in a different piece of wood $B_j$. Moreover, if we put them back together in their original position, then we can reconstruct the original piece of wood $\Omega$ (minus any wood lost in cutting the pieces—come on folks, we have to use a little imagination here!). Below is an illustration of a partition in the case $n = 6$. 

\[
\begin{array}{ccc}
B_1 & B_3 & B_5 \\
B_2 & B_4 & B_6 \\
\end{array}
\]
Suppose that we have a partition $B_1, B_2, \ldots, B_n$ of $\Omega$. Let $A \subseteq \Omega$ be any event. Suppose that we do not know how to compute $\mathcal{P}(A)$, but if $B_i$ occurred, then we’d know how to compute $\mathcal{P}(A|B_i)$ for each $i = 1, \ldots, n$. This scenario may seem a bit contrived, but it is not. What we are doing here is a precursor to a technique known as Bayes’ Theorem, which we’ll see in the next section. It turns out that it is often advantageous to partition the sample space in order to calculate the probability of an event. An example of such a scenario is given in the concert cancelation project on my website.

We can use conditional probabilities and the members of the partition to compute the probability of $A$. For definiteness, let $n = 3$ (this is the case for the concert cancellation project). Notice that we can write

$$A = A \cap \Omega
= A \cap (B_1 \cup B_2 \cup B_3) \quad \text{(Since the $B_i$’s form a partition of the set)}
= A \cap (B_1 \cup (B_2 \cup B_3)) \quad \text{(Group $B_2 \cup B_3$ and think of them as one set)}
= (A \cap B_1) \cup (A \cap (B_2 \cup B_3)) \quad \text{(Distributive Property for Sets)}
= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \quad \text{(Applying Distributive Property Again)}$$

(7.2)

For $n=3$, several generic pictures to keep in mind are
We can formalize this as follows. Using the identical steps found in (7.2) with \( n = 3 \) replaced with a general \( n \)-value we arrive at

\[
A = \bigcup_{i=1}^{n} (A \cap B_i), \tag{7.3}
\]

where \( (A \cap B_i) \cap (A \cap B_j) = \emptyset \) for \( i \neq j \) since by construction of the partitions \( B_i \cap B_j = \emptyset \).

**Law of Total Probability:** If the sets \( B_1, B_2, \ldots, B_n \subseteq \Omega \) partition \( \Omega \) and \( A \) is any event, then

\[
P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i). \tag{7.4}
\]

**Proof:** Let \( B_1, B_2, \ldots, B_n \subseteq \Omega \) partition \( \Omega \) and \( A \subseteq \Omega \) (any event). Then

\[
A = A \cap \Omega \\
= A \cap \left( \bigcup_{i=1}^{n} B_i \right) \quad \text{(definition of a partition)} \\
= \bigcup_{i=1}^{n} (A \cap B_i) \quad \text{(distributive law)}
\]

Notice that for \( i \neq j \), \( B_i \cap B_j = \emptyset \Rightarrow (A \cap B_i) \cap (A \cap B_j) = \emptyset \). Taking the probability of this expression yields

\[
P(A) = P \left( \bigcup_{i=1}^{n} (A \cap B_i) \right) \\
= \sum_{i=1}^{n} P(A \cap B_i) \quad \text{(disjoint sets)} \\
= \sum_{i=1}^{n} P(A|B_i)P(B_i) \quad \text{(definition of conditional probability: see (6.2))}
\]

### 7.1.1 Tree diagram describing the law of total probability

An alternative way of computing the probability of an event using the total probability representing partitions is the tree diagram. Each branch of the tree splits each time we make a partition of the sample space.
The above picture shows how we compute the law of total probability from a tree diagram. The first set of branches emanating from the first node on the left denote the first partition of the sample space by $B_1$ and $B_2$. The upper branch then splits again. This time $A$ and $A^c$ are used to partition the sample space. A similar statement holds for the bottom branch.

**A closer look at the tree diagram:** Suppose $B_1$ or $B_2$, but not both, must happen. If $A$ happens, it must happen with $B_1$ or $B_2$. Thus the probability that $A$ happens is the sum of the two probabilities above:

$$\mathbb{P}(A) = \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2).$$

Below is one last picture where we make a connection between the tree diagram for the law of total probability and the set representation of the law of total probability.
Example 1. You run an evening radio show out of your local university’s radio station. A survey given by an independent group at the university found that 10% of grad students, staff, and faculty listen to your radio station and 15% of undergrad students listen to your radio station. The undergrad population makeup 75% of the total university population.

(a) Use a tree diagram to compute the fraction of your potential audience that actually listens to your radio station.

(b) Use the law of total probability to compute the fraction of your potential audience that actually listens to your radio station.

(c) Suppose you have a choice of running an ad in the local university paper that increases the number of grad students, staff, and faculty that listen to your station from 10% to 15%, or running an ad that increases the number of undergrads that listen to your station from 15% to 20%. Which plan would cause the bigger increase in the total audience that actually listens to your radio station?

Solution:

Step 1: Define the events. Let

\[ U = \text{undergrads} \quad G = \text{grad students, staff, and faculty} \]

Then if \( \Omega \) is the set of all people at the university,

\[ \Omega = U \cup G \quad \text{and} \quad U \cap G = \emptyset \]
Thus $U$ and $G$ partition the population $\Omega$. Next, let
\[ L = \text{listeners} \quad L^c = \text{non-listeners} \]

**Step 2:** Write down the given information.

- $\mathcal{P}(U) = 0.75$ (probability of randomly selecting an undergrad)
- $\mathcal{P}(G) = 0.25$ (probability of randomly selecting a grad student, staff, or faculty)
- $\mathcal{P}(L \mid U) = 0.15 \implies \mathcal{P}(L^c \mid U) = 1 - \mathcal{P}(L \mid U) = 0.85$
- $\mathcal{P}(L \mid G) = 0.10 \implies \mathcal{P}(L^c \mid G) = 1 - \mathcal{P}(L \mid G) = 0.90$

How did we get the last two lines? If 10% of grad students, staff, and faculty members are listeners, then the probability that a random grad/staff/faculty member will be a listener is 10%. In other words, given that we select a grad/staff/faculty member, the probability that the person will be a listener is 10%. We can make a similar argument about undergrads.

(a) Here is a tree diagram illustrating the situation:

- $\mathcal{P}(L \mid U) = 0.15 \quad \mathcal{P}(U \cap L) = \mathcal{P}(U)\mathcal{P}(L \mid U) = (0.75)(0.15) = 0.1125$
- $\mathcal{P}(L^c \mid U) = 0.85 \quad \mathcal{P}(U \cap L^c) = \mathcal{P}(U)\mathcal{P}(L^c \mid U) = (0.75)(0.85) = 0.6375$
- $\mathcal{P}(L \mid G) = 0.10 \quad \mathcal{P}(G \cap L) = \mathcal{P}(G)\mathcal{P}(L \mid G) = (0.25)(0.10) = 0.025$
- $\mathcal{P}(L^c \mid G) = 0.90 \quad \mathcal{P}(G \cap L^c) = \mathcal{P}(G)\mathcal{P}(L^c \mid G) = (0.25)(0.90) = 0.225$

(b) If $B_1$ and $B_2$ are a partition of $\Omega$, then:

\[ \mathcal{P}(A) = \mathcal{P}(A \mid B_1)\mathcal{P}(B_1) + \mathcal{P}(A \mid B_2)\mathcal{P}(B_2) \]

Using this formula, we get

\[ \mathcal{P}(L) = \mathcal{P}(L \mid U)\mathcal{P}(U) + \mathcal{P}(L \mid G)\mathcal{P}(G) \\
= (0.15)(0.75) + (0.10)(0.25) = 0.1375, \]

which is the same answer we get using the tree diagram.

(c) Advertising **does not** change the number of people in the population. Hence $\mathcal{P}(U)$ and $\mathcal{P}(G)$ remain the same under both plans. Advertising does change the number of listeners; so we’ll use $L_1$ and $L_2$ for the sets of listeners under plans 1 and 2.
Plan 1: \( P(L_1 \mid G) = 0.15 \) (number of grad students, staff, and faculty listeners increases)  
\( P(L_1 \mid U) = 0.15 \) (number of undergrad listeners remains the same)  

The fraction of listeners \( L_1 \) under plan 1 is:  
\[
P(L_1) = P(L_1 \mid G)P(G) + P(L_1 \mid U)P(U) 
= (0.15)(0.25) + (0.15)(0.75) 
= 0.15
\]

Plan 2: \( P(L_2 \mid G) = 0.10 \) (number of grad students, staff, and faculty listeners remains the same)  
\( P(L_2 \mid U) = 0.20 \) (number of undergrad listeners increases)  

The fraction of listeners \( L_2 \) under plan 2 is:  
\[
P(L_2) = P(L_2 \mid G)P(G) + P(L_2 \mid U)P(U) 
= (0.10)(0.25) + (0.20)(0.75) 
= 0.175
\]

Since \( P(L_1) < P(L_2) \), we’d expect a larger fraction of listeners under plan 2, since it targets a larger audience (the undergrads) and the increase in listens from each group is the same, namely a 5% increase.
Example 2. A field screening test has been developed for a certain disease found in the North American honey bee. Records show that 8% of bees have a strong form of the disease, 17% have a mild form of the disease, and the remaining population is disease-free. The field test is not completely reliable. A bee with a strong form of the disease has a 5% chance of testing negative; a bee with a mild form of the disease has a 13% chance of testing negative; and a bee with no disease has an 11% chance of testing positive.

(a) What percent of the bee population would you expect to test positive for the disease?
(b) What percent of the bee population would you expect to test negative for the disease?

Solution:

Step 1: Define the events:
\[ S = \text{set of bees with strong form} \]
\[ M = \text{set of bees with mild form} \]
\[ F = \text{set of bees who are disease-free} \]
\[ \Omega = \text{entire population of North American bees.} \]
\[ N = \text{set of bees who test negative} \]
\[ N^c = \text{set of bees who test positive} \]

Since
\[ \Omega = S \cup M \cup F \quad S \cap M = \emptyset \quad S \cap F = \emptyset \quad M \cap F = \emptyset ; \]

the collection of sets \( S, M, \) and \( F \) forms a partition of the population \( \Omega \).

Step 2: Write down the given information:
\[ P(S) = 0.08 \quad \text{(probability that a bee has the strong form)} \]
\[ P(M) = 0.17 \quad \text{(probability that a bee has the mild form)} \]
\[ P(F) = 0.75 \quad \text{(probability that a bee is disease-free)} \]
\[ P(N | S) = 0.05 \quad P(N^c | S) = 1 - P(N | S) = 0.95 \]
\[ P(N | M) = 0.13 \quad P(N^c | M) = 1 - P(N | M) = 0.87 \]
\[ P(N^c | F) = 0.11 \quad P(N | F) = 1 - P(N^c | F) = 0.89 \]

(a) What percent of bees test positive?
\[
P(N^c) = P(N^c | S)P(S) + P(N^c | M)P(M) + P(N^c | F)P(F) \\
= (.95)(.08) + (.87)(.17) + (.11)(.75) \approx 0.31 = 31\%
\]

(b) What percent of the bee population would you expect to test negative?
Since \( N \) and \( N^c \) is a partition of \( \Omega \),
\[ P(N) = 1 - P(N^c) \approx 0.69 = 69\% \]
7.2 Introducing Bayes’ theorem

Assume we want to know a conditional probability $\mathcal{P}(E|F)$, but can only estimate $\mathcal{P}(F|E)$. This happens in many applications in science, insurance, and business. In such cases, we need a method to reverse the conditional probabilities. This situation arises in the concert cancellation project on my website. We know $\mathcal{P}(R \cap G | C)$ and $\mathcal{P}(R \cap G | C^c)$ but we want $\mathcal{P}(C | R \cap G)$ and $\mathcal{P}(C^c | R \cap G)$, where $R$ is region of the country, $G$ genre, and $C$ is the event the concert gets cancelled.

We will derive Bayes’ formula for the case $n=2$. Suppose that events $B_1$ and $B_2$ partition the outcomes of an experiment and that $A$ is another event. We want to use $\mathcal{P}(B_1)$, $\mathcal{P}(B_2)$, $\mathcal{P}(A|B_1)$, and $\mathcal{P}(A|B_2)$ to compute $\mathcal{P}(B_1|A)$. Now, by the definition of conditional probability,

$$\mathcal{P}(B_1|A) = \frac{\mathcal{P}(B_1 \cap A)}{\mathcal{P}(A)}.$$ 

Also, $\mathcal{P}(B_1 \cap A) = \mathcal{P}(A|B_1)\mathcal{P}(B_1)$, so that

$$\mathcal{P}(B_1|A) = \frac{\mathcal{P}(A|B_1)\mathcal{P}(B_1)}{\mathcal{P}(A)}.$$ 

Since $B_1$ and $B_2$ form a partition of $\Omega$ and $A \subseteq \Omega$, we have by the law of total probability

$$\mathcal{P}(A) = \mathcal{P}(A|B_1)\mathcal{P}(B_1) + \mathcal{P}(A|B_2)\mathcal{P}(B_2).$$ 

Thus,

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

Similarly, by symmetry we can replace $B_1$ by $B_2$ in the above equation to arrive at

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

Such an argument is known as a symmetry argument. It works because there was nothing special about the partition set $B_1$ in comparison to $B_2$.

**Warning:** These formulas only apply when the events $B_1$ and $B_2$ form a partition. That is, if $B_1$ and $B_2$ cannot happen simultaneously but one of them must occur.

**Bayes’ Theorem:** Suppose that the events $B_1, \ldots, B_n$ partition the outcomes of an experiment and that $A$ is another event; i.e. the $B_i$’s partition $\Omega$ and $A \subseteq \Omega$. For any integer $k$ with $1 \leq k \leq n$, we have

$$\mathcal{P}(B_k|A) = \frac{\mathcal{P}(A|B_k)\mathcal{P}(B_k)}{\sum_{i=1}^{n} \mathcal{P}(A|B_i)\mathcal{P}(B_i)}. \quad (7.5)$$ 

If all outcomes are equally likely, then

$$\mathcal{P}(B_k|A) = \frac{\#(B_k \cap A)}{\#(A)} = \frac{\#(B_k \cap A)}{\sum_{i=1}^{n} \#(B_i \cap A)}.$$
Example 3. Suppose you have 4 bags labelled 1 through 4, each bag containing exactly two balls as shown in the figure below; W means white and B means black:

Assume that a bag is selected at random and that a black ball is drawn from the bag.

Question: What is the probability that the remaining ball in the bag is also black?

Notation: Let $\Omega=\{\text{all the balls in all the bags}\}$ be the sample space. Let $C_i$ denote the event of selecting bag or container $i$, and define the following events: $W$ is the event that a white ball is chosen, and $B$ is the event that a black ball is chosen.

The only way we can select two black balls is if we pick bag 3. So we want to know $P(C_3|B)$, the probability that we choose bag 3 given that the first ball selected was black. Notice that the bags partition the sample space.

Let’s do the following experiment: Select a bag, and then select a ball from the bag. The outcomes would be either a white or a black ball.

What can we compute for this experiment? If we know what bag we have selected, then we can compute the probability of selecting a black ball from that bag by using the following probabilities:

$$P(B|C_1) = \frac{1}{2} \quad (7.6a)$$
$$P(B|C_2) = \frac{1}{2} \quad (7.6b)$$
$$P(B|C_3) = 1 \quad (7.6c)$$
$$P(B|C_4) = 0 \quad (7.6d)$$
We also know the probability of picking any particular bag is 1/4. That is, $\mathcal{P}(C_i) = \frac{1}{4}$, for $i = 1, 2, 3, 4$. We can now use Bayes’ theorem (7.5):

$$
\mathcal{P}(C_3|B) = \frac{\mathcal{P}(B|C_3)\mathcal{P}(C_3)}{\mathcal{P}(B|C_1)\mathcal{P}(C_1) + \mathcal{P}(B|C_2)\mathcal{P}(C_2) + \mathcal{P}(B|C_3)\mathcal{P}(C_3) + \mathcal{P}(B|C_4)\mathcal{P}(C_4)}
$$

$$
= \frac{\mathcal{P}(B|C_3)\mathcal{P}(C_3)}{\frac{1}{4} [\mathcal{P}(B|C_1) + \mathcal{P}(B|C_2) + \mathcal{P}(B|C_3) + \mathcal{P}(B|C_4)]}
$$

$$
= \frac{\mathcal{P}(B|C_3)}{\sum_{i=1}^{n} \mathcal{P}(B|C_i)}
$$

$$
= \frac{1}{\frac{1}{2} + \frac{1}{2} + 1} = \frac{1}{2}
$$

**Question:** Why is the probability 1/2 instead of 1/3?

First of all, since we are told that we have selected a black ball we only need to consider bags 1, 2 and 3. This fact can be seen in the calculation. Notice that $\mathcal{P}(B|C_4) = 0$, so

$$
\mathcal{P}(B) = \mathcal{P}(B|C_1)\mathcal{P}(C_1) + \mathcal{P}(B|C_2)\mathcal{P}(C_2) + \mathcal{P}(B|C_3)\mathcal{P}(C_3).
$$

If we pull a black ball out of bags 1, 2 and 3, then we have three bags with 1 ball each, and there would be a black ball in only one of them, namely, in bag 3. So it would appear that we have a one in three chance of choosing a black ball. However, we have forgotten one important case! We have two choices for choosing a black ball to begin with! If we label the black balls in bag 3 as $B_1$ and $B_2$ then if we take a black ball out of one of the bags we either have $B_1$ or $B_2$ in bag 3, so there are 2 out of four possibilities for choosing another black ball in the next withdrawal.
**Example 4.** Consider two containers: Container 1 contains 3 black balls and one white ball and Container 2 contains 1 black ball and 2 white balls. Determine the following probabilities.

(a) Given that you selected a white ball, what is the probability that it came from container 1?

(b) Given that you selected a black ball, what is the probability that it came from container 1?

(c) Given that you selected a white ball, what is the probability that it came from container 2?

**Solution:**

**Step 1.** Define the events:

- $C_1 =$ selecting container 1
- $C_2 =$ selecting container 2
- $W =$ selecting a white ball
- $B =$ selecting a black ball
- $\Omega =$ set of all white and black balls

Notice that containers 1 and 2 partition $\Omega$.

**Step 2.** Write down what you are given:

$$
\begin{align*}
P(C_1) &= \frac{1}{2} & P(W \mid C_1) &= \frac{1}{4} & P(B \mid C_1) &= \frac{3}{4} \\
P(C_2) &= \frac{1}{2} & P(W \mid C_2) &= \frac{2}{3} & P(B \mid C_2) &= \frac{1}{3}
\end{align*}
$$

**Step 3.** Use Bayes’ Theorem to compute each probability.

(a) 

$$
P(C_1 \mid W) = \frac{P(W \mid C_1)P(C_1)}{P(W \mid C_1)P(C_1) + P(W \mid C_2)P(C_2)}
$$

$$
= \frac{\frac{1}{2} \cdot \frac{1}{4}}{\frac{1}{2} \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{2}}
$$

$$
= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{2}{3}} = \frac{3}{3 + 8} = \frac{3}{11} \approx 0.27
$$
(b) 

\[ P(C_1 | B) = \frac{P(B | C_1)P(C_1)}{P(B | C_1)P(C_1) + P(B | C_2)P(C_2)} \]

\[ = \frac{\frac{3}{4} \cdot \frac{1}{2}}{\frac{3}{4} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2}} \]

\[ = \frac{\frac{3}{4}}{\frac{3}{4} + \frac{1}{3}} = \frac{9}{9 + 4} = \frac{9}{13} \approx 0.69 \]

(c) 

\[ P(C_2 | W) = \frac{P(C_2 \cap W)}{P(W)} = \frac{P(W | C_2)P(C_2)}{P(W | C_1)P(C_1) + P(W | C_2)P(C_2)} \]

\[ = \frac{\left(\frac{2}{3}\right) \left(\frac{1}{2}\right)}{\left(\frac{1}{4}\right) \left(\frac{1}{2}\right) + \left(\frac{2}{3}\right) \left(\frac{1}{2}\right)} = \frac{\frac{2}{3}}{\frac{1}{4} + \frac{2}{3}} = \frac{2 \cdot 4}{3 + 2 \cdot 4} = \frac{8}{11}. \]
Example 5. Consider events $E$ and $F$: $\mathcal{P}(E) = 0.7$, $\mathcal{P}(F|E) = 0.2$, and $\mathcal{P}(F|E^c) = 0.5$. Use Bayes’ theorem to compute the following probabilities.

(a) $\mathcal{P}(E|F)$
(b) $\mathcal{P}(E^c|F)$

Solution:
Step 1. Define the events: Notice that $E$ and $E^c$ partition $\Omega$. Step 2. Write down what you know. Given:

\[
\begin{align*}
\mathcal{P}(E) &= 0.7 \\
\mathcal{P}(E^c) &= 1 - \mathcal{P}(E) = 1 - 0.7 = 0.3 \\
\mathcal{P}(F|E) &= 0.2 \\
\mathcal{P}(F|E^c) &= 0.5
\end{align*}
\]

Step 3. Use Bayes’ Theorem:

\[
\mathcal{P}(E|F) = \frac{\mathcal{P}(F|E)\mathcal{P}(E)}{\mathcal{P}(F|E)\mathcal{P}(E) + \mathcal{P}(F|E^c)\mathcal{P}(E^c)} = \frac{(0.2)(0.7)}{(0.2)(0.7) + (0.5)(0.3)} = \frac{0.14}{0.14 + 0.15} = \frac{14}{29} \approx 0.483 \approx \left(\frac{1}{2}\right)^- 
\]

\[
\mathcal{P}(E^c|F) = \frac{\mathcal{P}(F|E^c)\mathcal{P}(E^c)}{\mathcal{P}(F|E)\mathcal{P}(E) + \mathcal{P}(F|E^c)\mathcal{P}(E^c)} = \frac{(0.5)(0.3)}{(0.2)(0.7) + (0.5)(0.3)} = \frac{0.15}{0.14 + 0.15} = \frac{15}{29} \approx 0.517 \approx \left(\frac{1}{2}\right)^+
\]
Example 6. A test is available to determine if a certain circuit has a critical defect. However, the results of the test are not always reliable. Data collected over several years shows that if a circuit has a critical defect there is a 12% chance that circuit will pass the test (even though it has the defect). Similarly, there is a 15% chance that the circuit will fail the test even though the circuit has nothing wrong with it. The long-term records indicate that the defect is present in approximately 7% of all circuits made in the factory.

(a) If a circuit fails the test, what is the probability that it actually has the defect?

(b) If a circuit passes the test, what is the probability that it actually has the defect?

Solution:
The experiment consists of selecting a circuit and testing whether or not it has a certain defect.

Step 1: Let
\( \Omega = \) all circuits made at the factory,
\( D = \) event that the circuit has the defect,
\( D^c = \) event that the circuit does not have the defect,
\( P = \) event that the circuit tests positive for the defect,
\( N = \) event that the circuit tests negative for the defect.

Notice that \( D \) and \( D^c \) partition the set of all circuits into those with the defect and those without the defect \( (D \cup D^c = \Omega) \).

Step 2: We are given
\[ P(D) = .07, \]
\[ P(N|D) = .12 \text{ (probability that a circuit with the defect will test negative)} \]
\[ P(P|D^c) = .15 \text{ (probability that a circuit without the defect tests positive)} \]

From this information, we can immediately get three more facts:
(i) \( P(D^c) = 1 - P(D) = .93. \)
(ii) If 12% of the circuits with the defect test negative, then the other 88% of the circuits with the defect test positive: \( P(P|D) = .88. \)
(iii) If 15% of the circuits without the defect test positive, then 85% of the circuits without the defect test negative: \( P(N|D^c) = .85. \)

Step 3:
In Part (a), we were told that the circuit tested negative, so that event has already happened. We want \( P(D|N) \).

In Part (b), we were told that the circuit had tested positive, so that event has already happened. We want \( P(D|P) \).

We use Bayes’ theorem.

\[ P(D|N) = \frac{P(N|D)P(D)}{P(N|D)P(D) + P(N|D^c)P(D^c)} = \frac{(.12)(.07)}{(.12)(.07) + (.85)(.93)} \]

\[ P(D|P) = \frac{P(P|D)P(D)}{P(P|D)P(D) + P(P|D^c)P(D^c)} = \frac{(.88)(.07)}{(.88)(.07) + (.15)(.93)} \]
**Example 7.** A field screening test has been developed for a certain disease found in the North American honey bee. Records show that 8% of bees have a strong form of the disease, 17% have a mild form of the disease, and the remaining population is disease-free. The field test is not completely reliable. A bee with a strong form of the disease has a 5% chance of testing negative; a bee with a mild form of the disease has a 13% chance of testing negative; and a bee with no disease has an 11% chance of testing positive.

(a) If a bee tests positive for the disease, what is the probability that it actually has the strong form of the disease?

(b) If a bee tests positive for the disease, what is the probability that it actually has the mild form of the disease?

(c) If a bee tests negative for the disease, what is the probability that it actually disease free?

**Solution:**

**Step 1.** Define the events:

\[ \Omega = \text{all North American honey bees} \]
\[ S = \text{all bees with strong form of the disease} \]
\[ M = \text{all bees with mild form of the disease} \]
\[ F = \text{all bees who are free from the disease} \]

The sets \( S, M, \) and \( F \) are disjoint, and their union is \( \Omega \). Thus \( S, M, \) and \( F \) form a partition of \( \Omega \).

Also, define the events:

\[ T^+ = \text{event of testing positive} \]
\[ T^- = \text{event of testing negative} \]

The events \( T^+ \) and \( T^- \) are disjoint.

**Step 2.** Write down what you know. Given:

\[ \mathcal{P}(S) = 0.08 \]
\[ \mathcal{P}(M) = 0.17 \]
\[ \mathcal{P}(F) = 1 - 0.08 - 0.17 = 0.75 \]
\[ \mathcal{P}(T^- | S) = 0.05 \text{ so } \mathcal{P}(T^+ | S) = 1 - \mathcal{P}(T^- | S) = 0.95 \]
\[ \mathcal{P}(T^- | M) = 0.13 \text{ so } \mathcal{P}(T^+ | M) = 1 - \mathcal{P}(T^- | M) = 0.87 \]
\[ \mathcal{P}(T^+ | F) = 0.11 \text{ so } \mathcal{P}(T^- | F) = 1 - \mathcal{P}(T^+ | F) = 0.89 \]

**Step 3.** Use Bayes’ Theorem to compute the probabilities in parts (a)-(c).
part (a)

\[
\mathcal{P}(S \mid T^+) = \frac{\mathcal{P}(T^+ \mid S)\mathcal{P}(S)}{\mathcal{P}(T^+ \mid S)\mathcal{P}(S) + \mathcal{P}(T^+ \mid M)\mathcal{P}(M) + \mathcal{P}(T^+ \mid F)\mathcal{P}(F)}
\]

\[
= \frac{(0.95)(0.08)}{(0.95)(0.08) + (0.87)(0.17) + (0.11)(0.75)}
\]

\[
= 0.248
\]

part (b)

\[
\mathcal{P}(M \mid T^+) = \frac{\mathcal{P}(T^+ \mid M)\mathcal{P}(M)}{\mathcal{P}(T^+ \mid S)\mathcal{P}(S) + \mathcal{P}(T^+ \mid M)\mathcal{P}(M) + \mathcal{P}(T^+ \mid F)\mathcal{P}(F)}
\]

\[
= \frac{(0.87)(0.17)}{(0.95)(0.08) + (0.87)(0.17) + (0.11)(0.75)}
\]

\[
= 0.483
\]

part (c)

\[
\mathcal{P}(F \mid T^-) = \frac{\mathcal{P}(T^- \mid F)\mathcal{P}(F)}{\mathcal{P}(T^- \mid S)\mathcal{P}(S) + \mathcal{P}(T^- \mid M)\mathcal{P}(M) + \mathcal{P}(T^- \mid F)\mathcal{P}(F)}
\]

\[
= \frac{(0.89)(0.75)}{(0.05)(0.08) + (0.13)(0.17) + (0.89)(0.75)}
\]

\[
= 0.962
\]
8 Introduction to Random Variables

8.1 Introduction to finite random variables

In this section we will take a look at the case of a finite random variable, what it is, and how it relates to the usual functions that you saw in college algebra. We’ll start with a somewhat abstract and more general definition of a random variable and refine it as we go along.

Definition 7. A random variable $X$ is a rule that assigns a numerical value to each outcome $\omega \in \Omega$ of an experiment that has nondeterministic (random-chance) outcomes.

A formal picture of the idea behind a random variable is given below.

![Diagram of random variable](image)

The random variable $X$ assigns a numerical value to every outcome in $\Omega$. In this example, the outcome $\omega$ is given the value $x$. The set $E \subseteq \Omega$ (event) is defined by $E = \{\omega \in \Omega \mid X(\omega) = x\}$ and $\omega \in \mathcal{D}_X$ (the domain of $X$) and $x \in \mathcal{R}_X$ (the range of $X$).

Convention: We will use capital letters to denote random variables, and lower-case letters to denote the actual values that the random variables take on (the outcomes of the experiment).

At this point the idea of what a random variable is may seem very cloudy and confusing, which is why we’ll immediately focus on the special case of a finite random variable followed by a series of examples for clarification.

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be the sample space for an experiment with a finite number $n$ of outcomes.

Definition 8. A finite random variable $X$ is a function from the finite sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ to the real numbers.
In mathematical symbols, we write:

$$X : \Omega \to \mathbb{R}$$

This is just the way we’d write a function in college algebra: for example, $f : \mathbb{R} \to \mathbb{R}$.

We can simulate how the function works with a cartoon depicting the random variable as a machine that inputs an outcome $\omega \in \Omega$ from the experiment, and converts it to a real number. For example, if we input the outcome $\omega_1$, we might get an output of 2; and we’d write $X(\omega_1) = 2$.

$X$ is a function, since for every $\omega_i \in \Omega$, the value of $X(\omega_i)$ is unique. In college algebra, you learned that things like $X$ were called real-valued functions. Real-valued functions satisfy the vertical-line test, since for every $\omega_i$ in the domain $\Omega$, there is one and only one value of $X(\omega_i)$.

For our purposes, the main difference between our finite random variable $X$, which is a function, and functions from college algebra like $f(x) = x^2$, is that the domain of the college-algebra function is part of the set of all real numbers; while the domain of $X$ is a finite set $\Omega = \{\omega_1, \ldots, \omega_n\}$, whose elements need not be numbers at all. For example, if the experiment is flipping a coin, then the members of $\Omega$ might be “heads” and “tails”, which are not real numbers. If the experiment consists of undergoing a disease, then the elements of $\Omega$ might be “recovery” and “death”: again, not real numbers.

Every function, whether it’s a college-algebra function like $f(x) = x^2$ or a finite random variable like $X$, has a domain and a range. The domain is the set of all allowed input values for the function; the range is the set of all possible output values. For both of these functions, the range is a subset of the set of real numbers. In the case of $X$, since the domain is discrete and finite, the range must also be discrete and finite.
We write the range of the random variable $X$ as

$$R_X = \{x_1, x_2, \ldots, x_m\}$$

By convention, $x_1 < x_2 < \cdots < x_m$.

Since the domain $\Omega$ is finite, the range of $X$ must be finite as well: for each $\omega \in \Omega$, there can be one and only one value of $X(\omega)$ in the range. In fact, we can say:

If $\#(\Omega) = n$, then $\#(R_X) \leq n$.

To prove this, write out the elements of $\Omega$ explicitly:

$$\Omega = \{\omega_1, \ldots, \omega_n\}$$

For each value of $i$ from 1 through $n$, there is a value

$$X(\omega_i) = x_{k_i} \in \mathbb{R}$$

in the range. It is possible for a function to map two separate values in the domain to the same value in the range; for example, the function $f(x) = x^2$ maps both 2 and $-2$ in the domain to 4 in the range: $f(2) = f(-2) = 4$. This can also be true of a random variable as a function: it’s possible that

$$\omega_i \neq \omega_j \quad \text{but} \quad X(\omega_i) = X(\omega_j)$$

Thus the range can be as large as the domain, but no larger; and it could be smaller than the domain if more than one domain element are mapped to the same range element.
In summary, for any outcome of an experiment $\omega_i \in \Omega$, the random variable $X$ returns a value $X(\omega_i) = x_{ki} \in \mathbb{R}$. Since $X$ behaves like a regular function, why do we single it out and call it a random variable? The reason is that we don’t know ahead of time what the outcome of the experiment will be, so we don’t know the value of $X$ until the experiment has been conducted. This is where the randomness in $X$ lies: its domain, $\Omega$, is the set of outcomes of a random experiment. This means that an expression like $X = x_i$ is actually an event: the event

$$\{ \omega \in \Omega \mid X(\omega) = x_i \}$$

As with any event corresponding to an experiment, we can ask what the probability of the event is.

Before we wander too far from our original goal—introducing the concept of the random variable—let’s clarify things by looking at a set of examples. We’ll start with the experiment of flipping a fair coin twice and counting the number of heads.

**Example 1.** Consider the experiment of flipping a coin two times. The sample space contains four outcomes:

$$\Omega = \{HH, HT, TH, TT\}$$

To each outcome in $\Omega$, we assign the number

$$X = \text{number of heads in 2 flips}.$$ 

Notice that $X$ can only take on 3 values:

$$X = 0, 1, 2.$$ 

Here $X$ is the random variable and $0, 1, 2$ correspond to the individual outcomes from the experiment. In detail,

- $X = 0$ corresponds to the event $\{TT\} \subseteq \Omega$.
- $X = 1$ corresponds to the event $\{HT, TH\} \subseteq \Omega$.
- $X = 2$ corresponds to the event $\{HH\} \subseteq \Omega$. 

Figure 44: A visual depiction of the random variable $X : \Omega \to \mathbb{R}$ as a function from the sample space $\Omega$ to the range (the real axis). Here we see both a picture of the mapping from $\Omega$ to the real line as well as a listing of all of the values of $X$ over the domain.
Notice that $P(X = -1) = 0$, because $X \neq -1$. In fact, for any $x \notin \{0, 1, 2\}$ we have $P(X = x) = 0$.

Since $X = x$ (where $x = 0, 1, \text{ or } 2$) corresponds directly to an event in the sample space, namely the set

$$\{\omega \in \Omega \mid X(\omega) = x\} \subseteq \Omega$$

(where $\omega$ is an outcome of the experiment),

we will refer to $X = x$ as an event.

Notice that $X$ is a variable: it’s nonconstant, and can vary. Since the value that $X$ takes depends on the outcome of the experiment of flipping two coins, and that outcome is random, $X$ is known as a random variable. Below is a table designed to explicitly show the role of the outcome from the experiment, the probability of the outcome and how it relates to the random variable $X$.

<table>
<thead>
<tr>
<th>Outcome: $\omega \in \Omega$</th>
<th>Probability of outcome: $P({\omega})$</th>
<th>Value of $X$: $X(\omega) = x$</th>
<th>Event $X = x$ with ${\omega \in \Omega \mid X(\omega) = x}$</th>
<th>Probability of the event $X = x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TT$</td>
<td>$1/4$</td>
<td>-2</td>
<td>$X = -2 : {TT}$</td>
<td>$P(X = -2) = P({TT}) = 1/4$</td>
</tr>
<tr>
<td>$TH$</td>
<td>$1/4$</td>
<td>0</td>
<td>$X = 0 : {TH, HT}$</td>
<td>$P(X = 0) = P({TH, HT}) = P({TH}) + P({HT}) = 1/4 + 1/4 = 1/2$</td>
</tr>
<tr>
<td>$HT$</td>
<td>$1/4$</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$HH$</td>
<td>$1/4$</td>
<td>2</td>
<td>$X = 2 : {HH}$</td>
<td>$P(X = 2) = P({HH}) = 1/4$</td>
</tr>
</tbody>
</table>

From this information, we can compute probabilities like $P(0 \leq X \leq 1)$. Notice that $X$ can take values of 0, 1, or 2; but the value of $X = 2$ is not allowed since we require that $0 \leq X \leq 1$. Since these events are disjoint,

$$P(-2 \leq X \leq 1) = P(X = 0) + P(X = 1) = 1/4 + 1/2 = 3/4.$$

**Notation:** There is nothing special about the letter $X$. Just as we can have regular variables that are not named $x$, so also is true for random variables. Sometimes it is wise to pick a letter for a random variable that has some connection with what the random variable is representing. Notice the choice for the random variable that is used in this next example.
Next, we’ll return to our two-dice problem that we are already familiar with.

**Example 2.** Let \( S \) be the sum of the faces obtained by rolling a fair die twice. Find \( P(4 \leq S \leq 7) \). **Note:** In this case we have chosen our random variable to be \( S \), not \( X \!\!\!\!\!\!\!\!\!\!\!\!.**

**Solution:** Let’s start by writing down the sample space \( \Omega \). There are 6 outcomes for the first roll of the die, and 6 outcomes for the second roll. Since each of the 6 outcomes from the first roll (denoted \( x \in \{1, \ldots, 6\} \)) can be paired with any of the 6 outcomes from the second (denoted \( y \)), we have \( 6 \times 6 = 36 \) elements in the sample space. We can write the outcomes as ordered pairs with the form \( \omega = (x, y) \):

\[
\Omega = \begin{cases} 
(1,1) & (2,1) & (3,1) & (4,1) & (5,1) & (6,1) \\
(1,2) & (2,2) & (3,2) & (4,2) & (5,2) & (6,2) \\
(1,3) & (2,3) & (3,3) & (4,3) & (5,3) & (6,3) \\
(1,4) & (2,4) & (3,4) & (4,4) & (5,4) & (6,4) \\
(1,5) & (2,5) & (3,5) & (4,5) & (5,5) & (6,5) \\
(1,6) & (2,6) & (3,6) & (4,6) & (5,6) & (6,6) 
\end{cases}
\]

All of these outcomes are equally likely. The order is important: \((1,3)\) is not the same outcome as \((3,1)\). The values of \( S(\omega) \) are given by adding the coordinates \((S(\omega) = x+y)\):

\[
S(\omega) = \begin{pmatrix} 
2 & 3 & 4 & 5 & 6 & 7 \\
3 & 4 & 5 & 6 & 7 & 8 \\
4 & 5 & 6 & 7 & 8 & 9 \\
5 & 6 & 7 & 8 & 9 & 10 \\
6 & 7 & 8 & 9 & 10 & 11 \\
7 & 8 & 9 & 10 & 11 & 12 
\end{pmatrix}
\]

Notice that \( 2 \leq S \leq 12 \). Each value of \( S \) corresponds to a subset of \( \Omega \). To calculate \( P(S = s) \), we count the outcomes in the subset \( \{\omega \mid S = s\} \subseteq \Omega \), then divide by the 36 total outcomes in \( \Omega \):

\[
\begin{align*}
S = 2 & \rightarrow \{(1,1)\} \quad \text{(one element)} & \quad P(S = 2) = \frac{1}{36} \\
S = 3 & \rightarrow \{(1,2), (2,1)\} \quad \text{(two elements)} & \quad P(S = 3) = \frac{2}{36} \\
S = 4 & \rightarrow \{(1,3), (2,2), (3,1)\} \quad \text{(three elements)} & \quad P(S = 4) = \frac{3}{36} \\
S = 5 & \rightarrow \{(1,4), (2,3), (3,2), (4,1)\} & \quad P(S = 5) = \frac{4}{36} \\
S = 6 & \rightarrow \{(1,5), (2,4), (3,3), (4,2), (5,1)\} & \quad P(S = 6) = \frac{5}{36} \\
S = 7 & \rightarrow \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\} & \quad P(S = 7) = \frac{6}{36} \\
\vdots \\
S = 12 & \rightarrow \{(6,6)\} & \quad P(S = 12) = \frac{1}{36}
\end{align*}
\]

For example,

\[
P(4 \leq S \leq 7) = P(S = 4) + P(S = 5) + P(S = 6) + P(S = 7) = \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} = \frac{18}{36} = \frac{1}{2}.
\]
Next, we’ll look at a series of examples involving coin tosses. Random variables created from the outcomes of coin tosses will prove useful throughout the rest of this text. In particular, the section on random walks will be based on this example.

**Example 3.** Consider the experiment of flipping a coin twice. The sample space for this experiment is

\[ \Omega = \{HH, HT, TH, TT\} \]

where \( H \) indicates a head; \( T \) a tail. We can assign arbitrary \( \omega_i \) names to the outcomes:

\( \omega_1 = HH, \omega_2 = HT, \omega_3 = TH, \omega_4 = TT \).

What kind of random variables can we define on this \( \Omega \)? Some possibilities are:

\[
X = \# \text{ of heads} - \# \text{ of tails} \\
Y = \# \text{ of heads} \\
Z = |X|
\]

All of these random variables have the same domain: \( \Omega = \{HH, HT, TH, TT\} \). However, they have different ranges, and different probabilities associated with each value in the range.

Let’s look more closely at these three random variables.

**Example 4.** Compute the range and the probabilities for \( X \) in Example 3.

To determine the range, we look at the value of \( X(\omega) \) for every \( \omega \in \Omega \).

\[
X(HH) = 2 - 0 = 2 \\
X(HT) = 1 - 1 = 0 \\
X(TH) = 1 - 1 = 0 \\
X(TT) = 0 - 2 = -2
\]

Thus the range of \( X \) is \( R_X = \{-2, 0, 2\} \). The events are \( X = 2, X = 0, \) and \( X = -2 \).

To compute the probability of the event \( X = -2 \), we look at all the outcomes of the experiment that correspond to \( X = -2 \). For this experiment, there is only one such outcome: \( TT \). For each value of \( X \), we determine the set of outcomes that corresponds to it:

\[
X = -2 \quad \text{corresponds to} \quad \{TT\} \\
X = 0 \quad \text{“} \quad \{HT, TH\} \\
X = 2 \quad \text{“} \quad \{HH\}
\]

We use the fact that each outcome of the experiment is equally likely to calculate the probabilities:

\[
\mathcal{P}(X = -2) = \mathcal{P}(\{TT\}) = 1/4 \\
\mathcal{P}(X = 0) = \mathcal{P}(\{HT, TH\}) = 2/4 = 1/2 \\
\mathcal{P}(X = 2) = \mathcal{P}(\{HH\}) = 1/4
\]
**Example 5.** Find the ranges for $Y$ and $Z$ in Example 3. 

$$Y(HH) = 2 \quad Y(HT) = 1 \quad Y(TH) = 1 \quad Y(TT) = 0$$

The range of $Y$ is $\mathcal{R}_Y = \{0, 1, 2\}$.

$$Z(HH) = |2| = 2 \quad Z(HT) = |0| = 0$$

$$Z(TH) = |0| = 0 \quad Z(TT) = |-2| = 2$$

The range of $Z$ is $\mathcal{R}_Z = \{0, 2\}$.

As an additional exercise, verify that $P(Z = 0) = P(Z = 2) = 1/2$.

If we think more closely about the expression $P(X = x)$, we can see that it is a function of $x$. That allows us to write

$$f(x) = P(X = x)$$

Obviously, this is going to depend on the nature of the random variable $X$. To help us keep this in mind, we’ll write the function with a subscript $X$.

As we shall see in the following sections, the insight that the expression $P(X = x)$ for the probability function is, in fact, a function of the range (not the domain!) of the random variable $X$ is a key point to the development of the theory of probability. This subtle point can lead to much confusion on the part of the novice. Be careful!

### 8.2 Introduction to the concept of a probability mass function

**Definition 9.** The probability mass function for a random variable $X$ is the function

$$f_X(x) = P(X = x) .$$

Notice that the domain of the probability mass function is all the real numbers, not just the range of $X$. If $x$ is not in the range of $X$, then

$$f_X(x) = P(X = x) = 0 .$$
Beside the probability mass function \( f_X(x) = \mathcal{P}(X = x) \), there are a number of other probabilities that we will often need to compute:

\[
\begin{align*}
\mathcal{P}(X \text{ is at most } x) &= \mathcal{P}(X \leq x) \\
\mathcal{P}(X \text{ is at least } x) &= \mathcal{P}(X \geq x) \\
\mathcal{P}(X \text{ is less than } x) &= \mathcal{P}(X < x) \\
\mathcal{P}(X \text{ is greater than } x) &= \mathcal{P}(X > x) \\
\mathcal{P}(X \text{ is between } a \text{ and } b) &= \mathcal{P}(a < X < b) \\
&\quad\text{or} \quad \mathcal{P}(a \leq X < b) \\
&\quad\text{or} \quad \mathcal{P}(a < X \leq b) \\
&\quad\text{or} \quad \mathcal{P}(a \leq X \leq b)
\end{align*}
\]

The last four depend on the precise meaning of “between” in the context of the problem.

We’ll begin by discussing the probability that \( X \) is at most \( x \). From the expression we obtain for that, we’ll derive the expressions for the other seven cases.

**Definition 10.** Let \( X \) be a random variable with range \( \{x_1, x_2, \ldots, x_n\} \), where \( x_1 < x_2 < \cdots < x_n \). The cumulative distribution function for \( X \) is

\[
F_X(x) = \mathcal{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i).
\]

That last expression may seem intimidating; but you’ve already been computing it. To determine \( \mathcal{P}(X \leq x) \) for a finite random variable, we just add up the probability masses for each \( x_i \) that is less than or equal to \( x \).

**Example 6.** Suppose the range of \( X \) is \( \{x_1, x_2, x_3 | x_1 < x_2 < x_3\} \); and that \( x \) is a real number, with \( x_2 < x < x_3 \). Then \( F_X(x) \) is the sum of all the mass to the left of \( x \) (see figure below); so

\[
F_X(x) = f_X(x_1) + f_X(x_2).
\]

![Figure 45: The total probability mass of \( \Omega \) is distributed over 3 locations \( (x_1, x_2, x_3) \) along the \( x \)-axis.](image)
Recall that if $x$ is not in the range of $X$ (in this case, if $x \neq x_1$, $x \neq x_2$, and $x \neq x_3$), then $f_X(x) = 0$. However, unless $x < x_1$, it will be the case that $F_X \neq 0$, whether or not $x$ is in the range of $X$. This is because $F_X(x)$ is the accumulation of all mass along the number line, beginning at $-\infty$ and extending up to and including $x$.

We will return to the concept of probability mass and the cumulative distribution function in future sections. However, we must first take a digression into the definition and meaning of the expected value. We will then tie a physical meaning to the probability function and the expected value that we have only eluded to at this point in the text.

### 8.3 Introduction to the expected value of a random variable

Suppose a random variable $X$ can only assume the distinct values $x_1, x_2, \ldots, x_n$. We can compute the average (mean) value of $X$. The mean value for the random variable is given by

$$
\mathcal{E}(X) = \sum_{i=1}^{n} x_i P(X = x_i).
$$

We call this mean value the expected value of $X$. It is denoted by either $\mathcal{E}(X)$ or $\mu_X$.

If we think of the $x_i$ as positions on a massless ruler and the probabilities $P(X = x_i)$ as weights that are hanging from the ruler at the respective positions $x_1, x_2, \ldots, x_n$, then $\mathcal{E}(X)$ will be the point where the ruler weight system balances. In physics we refer to this point as the center of mass.

**Example:** For simplicity, assume $x_1 < x_2 < \cdots < x_n$. Consider the case $n = 3$ with $X = 0, 2, 5$ and $P(X = x_1) = \frac{3}{10}$, $P(X = x_2) = \frac{6}{10}$, $P(X = x_3) = \frac{1}{10}$.
The expected value is $E(X) = 0 \mathcal{P}(X = 0) + 2 \mathcal{P}(X = 2) + 5 \mathcal{P}(X = 5) = 2(3/5) + 5(1/10) = 17/10 = 1.7$. Thus, 1.7 is the position for the center of mass of the mass-rod system.

Let $X$ be a random variable whose only possible values are $x_1, x_2$ and $x_3$. Let $N$ be a large number. Suppose that after $N$ repetitions of the experiment, you have obtained the following frequencies denoted by $f$:

- $x_1$ occurs $f_1$ times,
- $x_2$ occurs $f_2$ times,
- $x_3$ occurs $f_3$ times,

where $f_1 + f_2 + f_3 = N$. If we now compute the average value of the outcomes from the experiment we have

$$
\bar{x} = \frac{(x_1 + \cdots + x_1) + (x_2 + \cdots + x_2) + (x_3 + \cdots + x_3)}{N}
= \frac{f_1 x_1 + f_2 x_2 + f_3 x_3}{N}
= \frac{f_1}{N} x_1 + \frac{f_2}{N} x_2 + \frac{f_3}{N} x_3
$$

(8.2)

For large enough $N$ we can use the frequency interpretation to approximate the following probabilities:

$$
\mathcal{P}(X = x_1) \approx \frac{f_1}{N}, \quad \mathcal{P}(X = x_2) \approx \frac{f_2}{N}, \quad \mathcal{P}(X = x_3) \approx \frac{f_3}{N}.
$$

(8.3)

For instance, the probability of heads on any particular toss of a coin can be approximated by the # of heads in $N$ tosses, provided that $N$ is large. Substituting (9.3) into (9.2) yields $\bar{x} \approx E(X)$.

**Example 7.** A standardized 5-question no-partial-credit multiple-choice exam is given to the freshman class at a certain college every year. Let $X$ be the number of correct answers of a randomly selected test. Using statistical data, the probability of $X$ is found to be

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{P}(X = x)$</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Compute the expected value of $X$. What does $E(X)$ mean in practical terms?

**Solution:**

$$
E(X) = \sum_x x \cdot \mathcal{P}(X = x) = 0(.1) + 1(.2) + 2(.3) + 3(.2) + 4(.1) + 5(.1) = 2.3
$$

This says that, on average, you will get a little less than half the problems correct. Of course, you can’t actually get 2.3 problems correct on any given exam.
Example 8. Let Range of \( X = \{x_1, x_2, \ldots, x_n\} \), \( x_1 < x_2 < \cdots < x_n \). Notice that \( x_1 = \text{minimum } X \text{ value} \) and \( x_n = \text{maximum } X \text{ value} \). Could \( \mathcal{E}(X) < x_1 \) or \( \mathcal{E}(X) > x_n \)?

Solution: No! The expected value must always lie inside the interval \([x_1, x_n]\). Think of weights balancing on a bar, as in the figure on the previous page, could the fulcrum lie on a point not on the bar? Obviously not!

For example, suppose the range of \( X = \{x_1, x_2, x_3\} = \{0, 1, 2\} \). Could \( \mathcal{E}(X) = 5 \)? No! \( 0 \leq \mathcal{E}(X) \leq 2 \).

Example 9. A coin is tossed 100,000 times. It was found that there were 70,000 heads and 30,000 tails. Is the coin fair? Let \( X \) be a random variable that is 1 if the flip is heads and 0 if the flip is tails. Approximate \( \mathcal{E}(X) \).

Solution: The coin is probably not fair. Let \( x_1 = 1 \) and \( x_2 = 0 \). The expected value is given by

\[
\mathcal{E}(X) = x_1 P(X = x_1) + x_2 P(X = x_2) \approx 1 \cdot \frac{70000}{100000} + 0 \cdot \frac{30000}{100000} = 0.7.
\]

Since we have a large number of trials, it is not likely that the coin is fair. With a fair coin and a large number of trials we’d expect about a 50% chance of heads or tails.

Example 10. Lab reports are to evaluated in according to three categories: clarity of the report \( C_1 \), following proper procedure \( C_2 \), and the quality of the analysis of the data and the conclusions drawn from the data \( C_3 \). For each category, you can receive one of three grades: \( G \) for “good,” \( F \) for “fair,” or \( P \) for “poor.” Let \( X \) equal the number of \( G \)’s your report gets. Suppose that all outcomes are equally likely: that is, in each category, you’re equally likely to get \( G \), \( F \), or \( P \).

Compute the following probabilities:

(a) \( P(X = 0) \)

(b) \( P(X = 2) \)

(c) \( P(X > 1) \)

Solution:

We’ll start by writing down the sample space. In each category, there are 3 possible grades. There are 3 categories in total. The total number of possible evaluations is

\[
(3 \text{ choices for Cat. 1}) \times (3 \text{ choices for Cat. 2}) \times (3 \text{ choices for Cat. 3}) = 3^3 = 27.
\]

We should expect to find 27 elements in the sample space:

\[
S = \{GGG, GGF, GGP, GFG, GFF, GFP, GPG, GPF, GPP, FGG, FGF, FGP, FFG, FFF, FFP, FPG, FPF, FPP, PGG, PGF, PGP, PFG, PFF, PFP, PPG, PPF, PPP\}
\]
Part (a): The event \( \{X = 0\} \) corresponds to the set of outcomes in \( S \) that include no \( G \)'s. There are eight such outcomes; so

\[
P(X = 0) = \frac{\text{# of outcomes with no } G\text{'s}}{\text{total # of outcomes}} = \frac{8}{27}.
\]

Part (b): For \( P(X = 2) \), count the outcomes with exactly two \( G \)'s. There are 6; so

\[
P(X = 2) = \frac{\text{# of outcomes with exactly 2 } G\text{'s}}{\text{total # of outcomes}} = \frac{6}{27}.
\]

Part (c): The range of \( X \) is \( \{0, 1, 2, 3\} \). \( P(X > 1) = P(X = 2) + P(X = 3) \)

Since each outcome is equally likely, we have

\[
P(X = 2) = \frac{6}{27} \quad P(X = 3) = \frac{1}{27}
\]

Notice that

\[
\{X = 2\} = \{GGF, GFG, FGG, GGP, GPG, PGG\} \\
\{X = 3\} = \{GGG\} \\
\#(\Omega) = 27
\]

Thus

\[
P(X > 1) = P(X = 2) + P(X = 3) = \frac{6}{27} + \frac{1}{27} = \frac{7}{27}
\]

Example 11. A state-run monthly lottery can sell 100,000 tickets at $2 per ticket. There are three different payouts for a winning ticket: $1,000,000 with probability 0.0000005, $100 with probability 0.008, and $10 with probability 0.01. Answer the following questions.

(a) On average, how much can the state expect to profit from the lottery per month?

(b) How much income per month could the state expect to average if it lowered the probability of winning $1,000,000 to 0.0000004?

(c) The state plans to keep the same probabilities, but raise the top prize above $1,000,000. How high could the price go and still give the state an expected value of $0.50 per ticket?

Solution:

Part (a): The expected profit from the lottery is just the expected value of the ticket times however many tickets you manage to sell. Let \( \bar{X} \) be the random variable that gives net profit to the state on a single ticket. The range of \( X \) is \( \{2 \times 10^6, 2 \times 10^2, 2 - 10, 2\} \), and the associated probabilities are \( P(X = 1,000,000) = 0.0000005 \) (top prize payout),
\[ P(X = 2 - 100) = 0.008 \text{ (second place payout), } P(X = 2 - 10) = 0.01, P(X = 2) = 1 - 0.10 - 0.008 - 0.0000005 = 0.9819995. \] Thus the expected value is

\[
E(X) = \sum x_i P(X = x_i) \\
= 2 \cdot P(X = 2) - 8 \cdot P(X = 10) \\
- 98 \cdot P(X = 100) - (1,000,000 - 2) \cdot P(X = 1,000,000) \\
= 0.60
\]

Thus, on average a ticket will bring in 60 cents. The net expected profit is the expected value from a single ticket times the number of tickets sold: \((0.60)(100,000) = \$60,000.\)

**Part (b):** Let \(\tilde{X}\) be the random variable that gives net profit to the state on a single ticket. Then

\[
\text{range of } \tilde{X} = \text{range of } X \quad \text{(from example 6)} \\
= \{2 - 10^6, 2 - 10^2, 2 - 10, 2\}
\]

\[
P(\tilde{X} = 2 - 10^6) = \frac{4}{10^7} = \frac{5 - 1}{10^7} = P(X = 2 - 10^6) - \frac{1}{10^7}
\]

\[
P(\tilde{X} = 2 - 10^2) = P(\tilde{X} = 2 - 10^2) = \frac{8}{10^3}
\]

\[
P(\tilde{X} = 2 - 10) = P(X = 2 - 10) = \frac{1}{10^2}
\]

\[
P(\tilde{X} = 2) = 1 - P(\tilde{X} = 2 - 10^6) - P(\tilde{X} = 2 - 10^2) - P(\tilde{X} = 2 - 10) \\
= 1 - P(X = 2 - 10^6) + \frac{1}{10^7} - P(X = 2 - 10^2) - P(X = 2 - 10) \\
= P(X = 2) + \frac{1}{10^7}
\]

We now have the probabilities of \(\tilde{X}\) in terms of the probabilities of \(X\).

\[
E[\tilde{X}] = (2 - 10^6)P(\tilde{X} = 2 - 10^6) + (2 - 10^2)P(\tilde{X} = 2 - 10^2) \\
+ (2 - 10)P(\tilde{X} = 2 - 10) + 2P(\tilde{X} = 2) \\
= (2 - 10^6) \left[ P(X = 2 - 10^6) - \frac{1}{10^7} \right] + (2 - 10^2)P(X = 2 - 10^2) \\
\]

\[
+ (2 - 10)P(X = 2 - 10) + 2 \left[ P(X = 2) + \frac{1}{10^7} \right] \quad \text{(substitute for } X) \\
= E[X] - \frac{1}{10^7}(2 - 10^6) + \frac{2}{10^7} \\
= E[X] + \frac{10^6 - 2 + 2}{10^7} \\
= E[X] + \frac{1}{10} \\
= 0.70
\]
Part (c): Let $T$ be the value of the top prize. All other prizes are the same, and their probabilities are the same. Let $X$ be the random variable representing the state’s net profit, as in example 6.

If a $2$ ticket wins $T$ dollars, then the state takes in $2 - T$ dollars (a loss). If a $2$ ticket wins $100$ or $10$, then $X = 2 - 10^2$ or $2 - 10$ respectively. A ticket that wins $0$ leaves the state with a profit of $X = 2$. Thus the range of $X$ is
\[
\{2 - T, 2 - 10^2, 2 - 10, 2\}
\]
and the probabilities are
\[
\mathcal{P}(X = 2 - T) = \frac{5}{10^7}
\]
\[
\mathcal{P}(X = 2 - 10^2) = \frac{8}{10^3}
\]
\[
\mathcal{P}(X = 2 - 10) = \frac{1}{10^2}
\]
\[
\mathcal{P}(X = 2) = 1 - \frac{5}{10^7} - \frac{8}{10^3} - \frac{1}{10^2} = 0.9819995
\]
We demand that $E[X] = \frac{1}{2}$ (50 cents per ticket). The formula for expected value is:
\[
E[X] = (2 - T)\mathcal{P}(X = 2 - T) + (2 - 10^2)\mathcal{P}(X = 2 - 10^2) + (2 - 10)\mathcal{P}(X = 2 - 10) + 2\mathcal{P}(X = 2)
\]
Substituting the known values, we get
\[
\frac{1}{2} = (2 - T)\frac{5}{10^7} + (2 - 10^2)\frac{8}{10^3} + (2 - 10)\frac{1}{10^2} + 2 \left( 1 - \frac{5}{10^7} - \frac{8}{10^3} - \frac{1}{10^2} \right)
\]
Multiply by $2 \cdot 10^6$ and obtain
\[
10^6 = (2 - T) \cdot 16 \cdot 10^3 + (2 - 10) \cdot 2 \cdot 10^4 + 4 \cdot 10^6 \left( 1 - \frac{5}{10^7} - \frac{8}{10^3} - \frac{1}{10^2} \right)
\]
Add $T - 10^6$ to both sides of the equation:
\[
T = (2 - 10^6) + (2 - 10^2) \cdot 16 \cdot 10^3 + (2 - 10) \cdot 2 \cdot 10^4 + 4 \cdot 10^6 \left( 1 - \frac{5}{10^7} - \frac{8}{10^3} - \frac{1}{10^2} \right)
\]
Simplifying yields
\[
T = 2 - 10^6 + 32 \cdot 10^3 - 16 \cdot 10^5 + 4 \cdot 10^4 - 2 \cdot 10^5 + 4 \cdot 10^6 - 2 - 32 \cdot 10^3 - 4 \cdot 10^4
\]
After much cancellation of terms, we get
\[
T = 3 \cdot 10^6 - 18 \cdot 10^5 = (30 - 18)10^5 = 12 \cdot 10^5 = 1.2 \cdot 10^6 = 1,200,000
\]
Example 12. Let a fair coin be tossed twice, and let \( X \) be the number of tails in the two tosses. Compute \( E(X) \).

The set of outcomes is \( S = \{TT, TH, HT, HH\} \). Thus \( X \) can take one of three values: 0, 1 or 2. Each of the four outcomes is equally likely.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Probability</th>
<th>( X )</th>
<th>( P(X) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT</td>
<td>1/4</td>
<td>2</td>
<td>( P(X = 2) = P({TT}) = 1/4 )</td>
</tr>
<tr>
<td>TH</td>
<td>1/4</td>
<td>1</td>
<td>( P(X = 1) = P({TH, HT}) = 1/2 )</td>
</tr>
<tr>
<td>HT</td>
<td>1/4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HH</td>
<td>1/4</td>
<td>0</td>
<td>( P(X = 0) = P({HH}) = 1/4 )</td>
</tr>
</tbody>
</table>

\[
E(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2) \\
= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.
\]

This is just what you’d expect.

Example 13. Your civil engineering company has two project development offers for next fall, which we’ll refer to as project 1 and 2. Unfortunately, your company is too small to take on both projects. If your company invests in Project 1 there is a 20% chance that you lose $10,000, a 50% chance that you break even, and a 30% chance that you make $50,000. If you invest in Project 2 there is a 10% chance that you lose $150,000, a 70% chance that you break even, and a 20% chance that you make $100,000. Based on the expected value of each, which investment should you make?

Let \( X_1 = \) return on investment in project 1 (\( i.e., \) profit)
Let \( X_2 = \) return on investment in project 2
Range of \( X_1 = \{-10,000, 0, 50,000\} \)
Range of \( X_2 = \{-150,000, 0, 100,000\} \)

\[
P(X_1 = -10000) = 0.2 \quad P(X_2 = -150000) = 0.1 \quad \text{(probability of loss)}
\]
\[
P(X_1 = 0) = 0.5 \quad P(X_2 = 0) = 0.7 \quad \text{(probability of break even)}
\]
\[
P(X_1 = 50000) = 0.3 \quad P(X_2 = 100000) = 0.2 \quad \text{(probability of win)}
\]

Compute the expected returns:

\[
E[X_1] = -10000 \cdot P(X_1 = -10000) + 0 \cdot P(X_1 = 0) + 50000 \cdot P(X_1 = 50000) \\
= -10000(0.2) + 50000(0.3) = -2,000 + 15,000 = 13,000
\]
\[
E[X_2] = -150000 \cdot P(X_2 = -150000) + 0 \cdot P(X_2 = 0) + 100000 \cdot P(X_1 = 100000) \\
= -150000(0.1) + 100000(0.2) = -15000 + 20000 = 5,000
\]
Since the expected return from project 1 is greater than the expected return from project 2, choose project 1. Notice that project 2 has the lure of a lot more money if things go well. Businesses that are less risk adverse (i.e., more bullish) would take the risk and go with project 2. In fact, in the real world things are not so cut and dry. The promise of large profits upon success can motivate people to work much harder than they would otherwise, thus changing the probability of success. It is presumed that the probability of success and failure come from the statistical analysis derived from a large data base.

**Example 14.** As part of a sales promotion the local bookstore will let the first 100 customers pick a gift certificate from a box of 1,000 certificates, some of which are for $1 and some of which are for $100. After each draw, the value of the certificate is noted and the piece of paper is replaced in the box (This is known as sampling with replacement). Thus, each customer has the same chance of getting the $100 prize. What is the largest number of $100 certificates that you can put into the box, if the entire promotion is to have an expected cost of at most $300?

**Solution:**

Constraint: we only have $300 to lose, and we must have 1000 certificates. Let $n$ be the number of $100 certificates in the box; let $C$ be a random variable giving the value of the certificate ($C = 1$ or 100).

$$E(C) = 1 \cdot P(C = 1) + 100 \cdot P(C = 100)$$

$$= 1 \cdot \frac{\text{# of } \$1 \text{ bills in box}}{\text{total } \# \text{ of certificates}} + 100 \cdot \frac{\text{# of } \$100 \text{ bills in box}}{\text{total } \# \text{ of certificates}}$$

$$= 1 \cdot \frac{1000 - n}{1000} + 100 \cdot \frac{n}{1000}$$

$$= 1 + \frac{99n}{1000}.$$  

Given: Expected cost = $100 \cdot E(C) = $(100 + \frac{99}{10}n).

Set: Expected cost = $300. Then

$$300 = 100 + \frac{99}{10}n$$

$$20 \cdot 10 = n$$

$$n \approx \frac{2000}{100} = 20.$$  

Of course, to make sure that we didn’t exceed $300, we’d only put in two $100 bills!

Now that we’ve had a brief introduction to random variables and expected values, we return to discrete random variables to take a closer look at their properties and how we actually use the concepts of probability mass and the cumulative distribution of probability mass to solve problems.
9 Expected value and variance

9.1 Introduction to the expected value of a random variable

Suppose a random variable $X$ can only assume the distinct values $x_1, x_2, \ldots, x_n$. We can compute the average (mean) value of $X$. The mean value for the random variable is given by

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i).$$

We call this mean value the expected value of $X$. It is denoted by either $E(X)$ or $\mu_X$.

If we think of the $x_i$ as positions on a massless ruler and the probabilities $P(X = x_i)$ as weights that are hanging from the ruler at the respective positions $x_1, x_2, \ldots, x_n$, then $E(X)$ will be the point where the ruler weight system balances. In physics we refer to this point as the center of mass.

Example: For simplicity, assume $x_1 < x_2 < \cdots < x_n$. Consider the case $n = 3$ with $X = 0, 2, 5$ and $P(X = x_1) = \frac{3}{10}$, $P(X = x_2) = \frac{6}{10}$, $P(X = x_3) = \frac{1}{10}$.

The expected value is $E(X) = 0 \cdot P(X = 0) + 2 \cdot P(X = 2) + 5 \cdot P(X = 5) = 2(3/5) + 5(1/10) = 17/10 = 1.7$. Thus, 1.7 is the position for the center of mass of the mass-rod system.

Let $X$ be a random variable whose only possible values are $x_1, x_2$ and $x_3$. Let $N$ be a large number. Suppose that after $N$ repetitions of the experiment, you have obtained the following frequencies denoted by $f$:

- $x_1$ occurs $f_1$ times,
- $x_2$ occurs $f_2$ times,
- $x_3$ occurs $f_3$ times,

where $f_1 + f_2 + f_3 = N$. If we now compute the average value of the outcomes from the experiment we have

$$
\bar{x} = \frac{(x_1 + \cdots + x_1) + (x_2 + \cdots + x_2) + (x_3 + \cdots + x_3)}{N}
= \frac{f_1 x_1 + f_2 x_2 + f_3 x_3}{N}
= \frac{f_1}{N} x_1 + \frac{f_2}{N} x_2 + \frac{f_3}{N} x_3
$$

(9.2)

For large enough $N$ we can use the frequency interpretation to approximate the following probabilities:

$$
P(X = x_1) \approx \frac{f_1}{N}, \quad P(X = x_2) \approx \frac{f_2}{N}, \quad P(X = x_3) \approx \frac{f_3}{N}.
$$

(9.3)

For instance, the probability of heads on any particular toss of a coin can be approximated by the # of heads in $N$ tosses, provided that $N$ is large. Substituting (9.3) into (9.2) yields $\bar{x} \approx \mathcal{E}(X)$.

**Example 1.** A standardized 5-question no-partial-credit multiple-choice exam is given to the freshman class at a certain college every year. Let $X$ be the number of correct answers of a randomly selected test. Using statistical data, the probability of $X$ is found to be

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{P}(X = x)$</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Compute the expected value of $X$. What does $\mathcal{E}(X)$ mean in practical terms?

**Solution:**

$$
\mathcal{E}(X) = \sum_x x \cdot \mathcal{P}(X = x) = 0(.1) + 1(.2) + 2(.3) + 3(.2) + 4(.1) + 5(.1) = 2.3
$$

This says that, on average, you will get a little less than half the problems correct. Of course, you can’t actually get 2.3 problems correct on any given exam.

**Example 2.** Let Range of $X = \{x_1, x_2, \ldots, x_n\}$, $x_1 < x_2 < \cdots < x_n$. Notice that $x_1 = \text{minimum } X \text{ value}$ and $x_n = \text{maximum } X \text{ value}$ Could $\mathcal{E}(X) < x_1$ or $\mathcal{E}(X) > x_n$?

**Solution:** No! The expected value must always lie inside the interval $[x_1, x_n]$. Think of weights balancing on a bar, as in the figure on the previous page, could the fulcrum lie on a point not on the bar? Obviously not!

For example, suppose the range of $X = \{x_1, x_2, x_3\} = \{0, 1, 2\}$. Could $\mathcal{E}(X) = 5$? No! $0 \leq \mathcal{E}(X) \leq 2$. 


Example 3. A coin is tossed 100,000 times. It was found that there were 70,000 heads and 30,000 tails. Is the coin fair? Let $X$ be a random variable that is 1 if the flip is heads and 0 if the flip is tails. Approximate $E(X)$.

Solution: The coin is probably not fair. Let $x_1 = 1$ and $x_2 = 0$. The expected value is given by

$$E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) \approx 1 \cdot \frac{70000}{100000} + 0 \cdot \frac{30000}{100000} = 0.7.$$ 

Since we have a large number of trials, it is not likely that the coin is fair. With a fair coin and a large number of trials we’d expect about a 50% chance of heads or tails.

Example 4. Lab reports are to be evaluated in accordance to three categories: clarity of the report $C_1$, following proper procedure $C_2$, and the quality of the analysis of the data and the conclusions drawn from the data $C_3$. For each category, you can receive one of three grades: $G$ for “good,” $F$ for “fair,” or $P$ for “poor.” Let $X$ equal the number of $G$’s your report gets. Suppose that all outcomes are equally likely: that is, in each category, you’re equally likely to get $G$, $F$, or $P$.

Compute the following probabilities:

(a) $P(X = 0)$
(b) $P(X = 2)$
(c) $P(X > 1)$

Solution:

We’ll start by writing down the sample space. In each category, there are 3 possible grades. There are 3 categories in total. The total number of possible evaluations is

$$(3 \text{ choices for Cat. 1}) \times (3 \text{ choices for Cat. 2}) \times (3 \text{ choices for Cat. 3}) = 3^3 = 27.$$ 

We should expect to find 27 elements in the sample space:

$$S = \{GGG, GGF, GGP, GFG, GFF, GFP, GPG, GPF, GPP, FGG, FGF, FGP, FFG, FFF, FFP, FPG, FPF, FPP, PGG, PGF, PGP, PFG, PFF, PFP, PPG, PPF, PPP\}$$

Part (a): The event $\{X = 0\}$ corresponds to the set of outcomes in $S$ that include no $G$’s. There are eight such outcomes; so

$$P(X = 0) = \frac{\# \text{ of outcomes with no } G’s}{\text{total } \# \text{ of outcomes}} = \frac{8}{27}. $$

Part (b): For $P(X = 2)$, count the outcomes with exactly two $G$’s. There are 6; so

$$P(X = 2) = \frac{\# \text{ of outcomes with exactly 2 } G’s}{\text{total } \# \text{ of outcomes}} = \frac{6}{27}. $$
Part (c): The range of \( X \) is \( \{0, 1, 2, 3\} \). \( \mathcal{P}(X > 1) = \mathcal{P}(X = 2) + \mathcal{P}(X = 3) \)

Since each outcome is equally likely, we have

\[
\begin{align*}
\mathcal{P}(X = 2) &= \frac{6}{27} \\
\mathcal{P}(X = 3) &= \frac{1}{27}
\end{align*}
\]

Notice that

\[
\begin{align*}
\{X = 2\} &= \{GGF, GFG, FGG, GGP, GPG, PGG\} \\
\{X = 3\} &= \{GGG\} \\
\#(\Omega) &= 27
\end{align*}
\]

Thus

\[
\mathcal{P}(X > 1) = \mathcal{P}(X = 2) + \mathcal{P}(X = 3) = \frac{6}{27} + \frac{1}{27} = \frac{7}{27}
\]

Example 5. A state-run monthly lottery can sell 100,000 tickets at $2 per ticket. There are three different payouts for a winning ticket: $1,000,000 with probability 0.0000005, $100 with probability 0.008, and $10 with probability 0.01. Answer the following questions.

(a) On average, how much can the state expect to profit from the lottery per month?

(b) How much income per month could the state expect to average if it lowered the probability of winning $1,000,000 to 0.0000004?

(c) The state plans to keep the same probabilities, but raise the top prize above $1,000,000. How high could the price go and still give the state an expected value of $0.50 per ticket?

Solution:

Part (a): The expected profit from the lottery is just the expected value of the ticket times however many tickets you manage to sell. Let \( X \) be the random variable that gives net profit to the state on a single ticket. The range of \( X \) is \( \{2 - 10^6, 2 - 10^2, 2 - 10, 2\} \), and the associated probabilities are \( \mathcal{P}(X = 1,000,000) = 0.0000005 \) (top prize payout), \( \mathcal{P}(X = 2 - 100) = 0.008 \) (second place payout), \( \mathcal{P}(X = 2 - 10) = 0.01 \), \( \mathcal{P}(X = 2) = 1 - 0.10 - 0.008 - 0.0000005 = 0.9819995 \). Thus the expected value is

\[
E(X) = \sum x_i \mathcal{P}(X = x_i)
\]

\[
= 2 \cdot \mathcal{P}(X = 2) - 8 \cdot \mathcal{P}(X = 10)
\]

\[
- 98 \cdot \mathcal{P}(X = 100) - (1,000,000 - 2) \cdot \mathcal{P}(X = 1,000,000)
\]

\[
= 0.60
\]

Thus, on average a ticket will bring in 60 cents. The net expected profit is the expected value from a single ticket times the number of tickets sold: \((0.60)(100,000) = $60,000\).
Part (b): Let $\tilde{X}$ be the random variable that gives net profit to the state on a single ticket. Then

$$\text{range of } \tilde{X} = \text{range of } X \quad (\text{from example 6})$$

$$= \{2 - 10^6, 2 - 10^2, 2 - 10\}$$

$$\mathcal{P}(\tilde{X} = 2 - 10^6) = \frac{4}{10^7} = \frac{5 - 1}{10^7} = \mathcal{P}(X = 2 - 10^6) - \frac{1}{10^7}$$

$$\mathcal{P}(\tilde{X} = 2 - 10^2) = \mathcal{P}(X = 2 - 10^2) = \frac{8}{10^3}$$

$$\mathcal{P}(\tilde{X} = 2 - 10) = \mathcal{P}(X = 2 - 10) = \frac{1}{10^2}$$

$$\mathcal{P}(\tilde{X} = 2) = 1 - \mathcal{P}(\tilde{X} = 2 - 10^6) - \mathcal{P}(\tilde{X} = 2 - 10^2) - \mathcal{P}(\tilde{X} = 2 - 10)$$

$$= 1 - \mathcal{P}(X = 2 - 10^6) + \frac{1}{10^7} - \mathcal{P}(X = 2 - 10^2) - \mathcal{P}(X = 2 - 10)$$

$$= \mathcal{P}(X = 2) + \frac{1}{10^7}$$

We now have the probabilities of $\tilde{X}$ in terms of the probabilities of $X$.

$$E[\tilde{X}] = (2 - 10^6)\mathcal{P}(\tilde{X} = 2 - 10^6) + (2 - 10^2)\mathcal{P}(\tilde{X} = 2 - 10^2)$$

$$+ (2 - 10)\mathcal{P}(\tilde{X} = 2 - 10) + 2\mathcal{P}(\tilde{X} = 2)$$

$$= (2 - 10^6) \left[ \mathcal{P}(X = 2 - 10^6) - \frac{1}{10^7} \right] + (2 - 10^2)\mathcal{P}(X = 2 - 10^2)$$

$$+ (2 - 10)\mathcal{P}(X = 2 - 10) + 2 \left[ \mathcal{P}(X = 2) + \frac{1}{10^7} \right] \quad \text{(substitute for } X)$$

$$= E[X] - \frac{1}{10^7}(2 - 10^6) + \frac{2}{10^7}$$

$$= E[X] + \frac{10^6 - 2 + 2}{10^7}$$

$$= E[X] + \frac{1}{10}$$

$$= 0.70$$

Part (c): Let $T$ be the value of the top prize. All other prizes are the same, and their probabilities are the same. Let $X$ be the random variable representing the state’s net profit, as in example 6.

If a $2$ ticket wins $T$ dollars, then the state takes in $2 - T$ dollars (a loss). If a $2$ ticket wins $100$ or $10$, then $X = 2 - 10^2$ or $2 - 10$ respectively. A ticket that wins $0$ leaves the state with a profit of $X = 2$. Thus the range of $X$ is

$$\{2 - T, 2 - 10^2, 2 - 10, 2\}$$
and the probabilities are

\[ P(X = 2 - T) = \frac{5}{10^7} \]
\[ P(X = 2 - 10^2) = \frac{8}{10^3} \]
\[ P(X = 2 - 10) = \frac{1}{10^2} \]
\[ P(X = 2) = 1 - \frac{5}{10^7} - \frac{8}{10^3} - \frac{1}{10^2} = 0.9819995 \]

We demand that \( E[X] = \frac{1}{2} \) (50 cents per ticket). The formula for expected value is:

\[
E[X] = (2 - T)P(X = 2 - T) + (2 - 10^2)P(X = 2 - 10^2) \\
+ (2 - 10)P(X = 2 - 10) + 2P(X = 2)
\]

Substituting the known values, we get

\[
\frac{1}{2} = (2 - T)\frac{5}{10^7} + (2 - 10^2)\frac{8}{10^3} + (2 - 10)\frac{1}{10^2} + 2\left(1 - \frac{5}{10^7} - \frac{8}{10^3} - \frac{1}{10^2}\right)
\]

Multiply by \( 2 \cdot 10^6 \) and obtain

\[
10^6 = (2 - T)5 \cdot 10^5 + (2 - 10^2)\cdot 16 \cdot 10^3 + (2 - 10) \cdot 2 \cdot 10^4 + 4 \cdot 10^6 \left(1 - \frac{5}{10^7} - \frac{8}{10^3} - \frac{1}{10^2}\right)
\]

Add \( T \cdot 10^6 \) to both sides of the equation:

\[
T = (2 - 10^6) + (2 - 10^2) \cdot 16 \cdot 10^3 + (2 - 10) \cdot 2 \cdot 10^4 + 4 \cdot 10^6 \left(1 - \frac{5}{10^7} - \frac{8}{10^3} - \frac{1}{10^2}\right)
\]

Simplifying yields

\[
T = 2 - 10^6 + 32 \cdot 10^3 - 16 \cdot 10^5 + 4 \cdot 10^4 - 2 \cdot 10^5 + 4 \cdot 10^6 - 2 - 32 \cdot 10^3 - 4 \cdot 10^4
\]

After much cancellation of terms, we get

\[
T = 3 \cdot 10^6 - 18 \cdot 10^5 \\
= (30 - 18)10^5 \\
= 12 \cdot 10^5 \\
= 1.2 \cdot 10^6 \\
= 1,200,000
\]
Example 6. Let a fair coin be tossed twice, and let $X$ be the number of tails in the two tosses. Compute $E(X)$.

The set of outcomes is $S = \{TT, TH, HT, HH\}$. Thus $X$ can take one of three values: 0, 1 or 2. Each of the four outcomes is equally likely.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Probability</th>
<th>$X$</th>
<th>$P(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TT</td>
<td>1/4</td>
<td>2</td>
<td>$P(X = 2) = P({TT}) = 1/4$</td>
</tr>
<tr>
<td>TH</td>
<td>1/4</td>
<td>1</td>
<td>$P(X = 1) = P({TH, HT}) = 1/2$</td>
</tr>
<tr>
<td>HT</td>
<td>1/4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HH</td>
<td>1/4</td>
<td>0</td>
<td>$P(X = 0) = P({HH}) = 1/4$</td>
</tr>
</tbody>
</table>

$$E(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2)$$

$$= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

This is just what you’d expect.

Example 7. Your civil engineering company has two project development offers for next fall, which we’ll refer to as project 1 and 2. Unfortunately, your company is too small to take on both projects. If your company invests in Project 1 there is a 20% chance that you lose $10,000, a 50% chance that you break even, and a 30% chance that you make $50,000. If you invest in Project 2 there is a 10% chance that you lose $150,000, a 70% chance that you break even, and a 20% chance that you make $100,000. Based on the expected value of each, which investment should you make?

Let $X_1 =$ return on investment in project 1 (i.e., profit)

Let $X_2 =$ return on investment in project 2

Range of $X_1 = \{-10,000, 0, 50,000\}$

Range of $X_2 = \{-150,000, 0, 100,000\}$

$P(X_1 = -10000) = 0.2$  \quad $P(X_2 = -150000) = 0.1$  \quad (probability of loss)

$P(X_1 = 0) = 0.5$ \quad $P(X_2 = 0) = 0.7$  \quad (probability of break even)

$P(X_1 = 50000) = 0.3$ \quad $P(X_2 = 100000) = 0.2$  \quad (probability of win)

Compute the expected returns:

$$E[X_1] = -10000 \cdot P(X_1 = -10000) + 0 \cdot P(X_1 = 0) + 50000 \cdot P(X_1 = 50000)$$

$$= -10000(0.2) + 50000(0.3) = -2000 + 15000 = 13000$$

$$E[X_2] = -150000 \cdot P(X_2 = -150000) + 0 \cdot P(X_2 = 0) + 100000 \cdot P(X_1 = 100000)$$

$$= -150000(0.1) + 100000(0.2) = -15000 + 20000 = 5000$$
Since the expected return from project 1 is greater than the expected return from project 2, choose project 1. Notice that project 2 has the lure of a lot more money if things go well. Businesses that are less risk adverse (i.e., more bullish) would take the risk and go with project 2. In fact, in the real world things are not so cut and dry. The promise of large profits upon success can motivate people to work much harder than they would otherwise, thus changing the probability of success. It is presumed that the probability of success and failure come from the statical analysis derived from a large data base.

Example 8. As part of a sales promotion the local bookstore will let the first 100 customers pick a gift certificate from a box of 1,000 certificates, some of which are for $1 and some of which are for $100. After each draw, the value of the certificate is noted and the piece of paper is replaced in the box (This is known as sampling with replacement). Thus, each customer has the same chance of getting the $100 prize. What is the largest number of $100 certificates that you can put into the box, if the entire promotion is to have an expected cost of at most $300?

Solution:

Constraint: we only have $300 to lose, and we must have 1000 certificates. Let \( n \) be the number of $100 certificates in the box; let \( C \) be a random variable giving the value of the certificate \( C = 1 \) or 100.

\[
E(C) = 1 \cdot P(C = 1) + 100 \cdot P(C = 100)
\]
\[
= 1 \cdot \frac{\# \text{ of } $1 \text{ bills in box}}{\text{total # of certificates}} + 100 \cdot \frac{\# \text{ of } $100 \text{ bills in box}}{\text{total # of certificates}}
\]
\[
= 1 \cdot \frac{1000 - n}{1000} + 100 \cdot \frac{n}{1000}
\]
\[
= 1 + \frac{99n}{1000}.
\]

Given: Expected cost = $100 \( E(C) = $(100 + \frac{99n}{10}).

Set: Expected cost = $300. Then

\[
300 = 100 + \frac{99n}{10}
\]
\[
\frac{200 \cdot 10}{99} = n
\]
\[
n \approx \frac{2000}{100} = 20.
\]

Of course, to make sure that we didn’t exceed $300, we’d only put in two $100 bills!

Now that we’ve had a brief introduction to random variables and expected values, we return to discrete random variables to take a closer look at their properties and how we actually use the concepts of probability mass and the cumulative distribution of probability mass to solve problems.
9.2 Introduction to the variance and standard deviation of a random variable

**Definition:** The *variance* of a discrete random $X$ variable is given by

$$V(X) \equiv \sum_{n=0}^{N} (x - E(X))^2 f_X(x),$$

where $E(X)$ is the expected value of the random variable $X$.

Using the fact that the expected value has the following properties:

$$\begin{align*}
E(X_1 + X_2) &= E(X_1) + E(X_2), \text{ where } X_1 \text{ and } X_2 \text{ are random variables.} \\
E(\alpha X) &= \alpha E[X] \text{ for } \alpha \in \mathbb{R}.
\end{align*}$$

$$\begin{align*}
E(E(X)) &= E(X)E(1) \text{ (since } E(X) \text{ is a constant, can factor it out)} \\
E(1) &= \sum_{i=1}^{n} 1 \cdot \mathcal{P}(X = x_i) = 1 \text{ (by definition of a probability mass function).}
\end{align*}$$

we can derive a useful formula for the variance in terms of the expected value as follows:

$$\begin{align*}
V(X) &= E[(X - E(X))^2] \quad \text{(by definition)} \\
&= E[X^2 - 2XE(X) + (E(x))^2] \quad \text{(expand argument of } E) \\
&= E[X^2] - 2(E(X))E(X) + (E(x))^2E(1) \quad \text{(by above properties)} \\
&= E[X^2] - 2(E(X))^2 + (E(x))^2 \quad (E(1) = 1) \\
&= E[X^2] - (E(X))^2.
\end{align*}$$

Thus, we have just shown that

$$V(X) = E(X^2) - (E(X))^2.$$

We’ll use this formula when computing the variance.

**Definition:** The *standard deviation* of a discrete random $X$ variable is given by

$$\sigma_X \equiv \sqrt{V(X)},$$

where $V(X)$ is the variance of the random variable $X$.

**Comment on notation:**

- The expected value $E(X)$ is often denoted by $\mu_X$, and the standard deviation is often written as

$$\sigma_X = \sqrt{\sum_{n=0}^{N} (x - \mu_X)^2 f_X(x)}.$$
It should be pointed out that in practice, the variance is used for proving theorems and the standard deviation is used for measuring how spread out a collection of data is from the mean. In fact, the standard deviation can be thought of as the “average distance” from the mean, where distance is measured by the root-mean-square metric with weight function $f_X(x)$ (the probability mass function).

In this short introduction we will not be going into more detail about the meaning of the variance and standard deviation. Instead, we will use a few straightforward examples to try to convey the idea behind the standard deviation. We’ll finish with a discussion that demonstrates a simple interpretation of the expected value and the standard deviation when it comes to analyzing experimental data.

**Example 9.** Consider the random variable $X$ with range $X = \{-3, -2, -1, 0, 1, 2, 3\}$ and probability mass function $f_X(x)$ whose value are given in the table below.

<table>
<thead>
<tr>
<th>$x$</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_X(x)$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The graph of the probability distribution of the random variable $X$ is shown below.

![Probability Distribution](image)

**Figure 46:** A graph of the probability density function $f_X(x)$.

Determine the expected value and standard deviation of $X$. Do the values agree with the picture of the distribution? Explain.

**Solution:** We start by computing the expected value of $X$:

$$E(X) = (-3)(0.1) + (-2)(0.1) + (-1)(0.2) + (0)(0.2) + (1)(0.2) + (2)(0.1) + (3)(0.1) = 0.$$
This agrees with the graph. Recall that the expected value is analogous to the center of mass. Since the graph is symmetric about the origin, it follows that the center of mass is located at the origin.

Next, we compute the variance of $X$ and the standard deviation $\sigma_X$:

\[
V(X) = E(X^2) - (E(X))^2
\]

\[
= (-3)^2(.1) + (-2)^2(.1) + (-1)^2(.2) + (0)^2(.2) + (1)^2(.1) + (2)^2(.1) + (3)^2(.1) - 0
\]

\[
= 3,
\]

\[
\sigma_X = \sqrt{V(X)} = \sqrt{3} \approx 1.7.
\]

This answer seems reasonable since the mass is more concentrated at the origin. Thus we’d expect the spread to be well less than 3, which is the largest value in the range.

**Example 10.** Consider the random variable $Y$ with range $Y = \{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ and probability mass function $f_Y(y)$ whose value are given in the table below.

<table>
<thead>
<tr>
<th>$y$</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_Y(y)$</td>
<td>0.2</td>
<td>0.15</td>
<td>0.1</td>
<td>0.05</td>
<td>0</td>
<td>0.05</td>
<td>0.1</td>
<td>0.015</td>
<td>0.2</td>
</tr>
</tbody>
</table>

The graph of the probability distribution of the random variable $Y$ is shown below.

![Graph B](image)

**Figure 47:** A graph of the probability density function $f_Y(y)$.

Determine the expected value and standard deviation of $Y$. Do the values agree with the picture of the distribution? Explain.
Solution: We start by computing the expected value of $X$:

$$E(Y) = (-4)(.2) + (-3)(.15) + (-2)(.1) + (-1)(.05) + 0(0) + 1(.05) + 2(.10) + 3(.15) + 4(.2) = 0.$$ 

This agrees with the graph. Recall that the expected value is analogous to the center of mass, which since the graph is symmetric about the origin is located at the origin.

Next, we compute the variance of $X$ and the standard deviation $\sigma_X$:

$$V(Y) = E(Y^2) - (E(Y))^2$$

$$= (-4)^2(.2) + (-3)^2(.15) + (-2)^2(.10) + (-1)^2(.05) + 0^2(0) + 1^2(.05) + 2^2(.10) + 3^2(.15) + 4^2(.2) - 0$$

$$\approx 8.8,$$

$$\sigma_Y = \sqrt{V(Y)} = \sqrt{3} \approx 3.0 > 1.7 = \sigma_X.$$ 

This answer seems reasonable since the mass is more spread out than in the previous example. Thus we’d expect $\sigma_X < \sigma_Y$.

**Example 11.** Below are two graphs of the probability distributions of two random variables $X$ (graph A) and $Y$ (graph B). Notice that both graphs are symmetric about the origin, so the expected value of each graph is zero. In fact, graph A and B are the same distributions as in the previous two examples. Without doing any computations, which graph do you expect to have a larger standard deviation? Does this agree with the findings in the previous examples?

![Figure 48: Graph 1](image1.png) ![Figure 49: Graph 2](image2.png)

9.2.1 A word on expected values and standard deviations for large data sets

When using the expected value and standard deviation to analyze experimental data the situation becomes more complicated because the probability mass function corresponding to the experiment is typically unknown. Statistical methods must be used in this case. These methods are typically based on assumption that the experiments are repeatable and the outcomes are independent of one another.
The underlying idea behind these assumptions is that although the probability distribution behind the experiment is unknown, we do know that the trials are governed by the same distribution since the experiments are independent and identically distributed (i.i.d.). This fact is enough to justify making predictions based on established statistically methods.

For example, suppose we wish to determine a certain physical quantity \( q \). To do this we devise a repeatable experiment in our local physics lab. Suppose we carefully repeat the experiment \( N \) times, where \( N \) is large. Thus, our data looks like \( \{q_1, q_2, \ldots, q_N\} \). Assuming that each of the experiments are carried out identically and that the outcomes of each experiment are independent from the other experiments, it is reasonable to use standard statistical methods on the data. The formulas for the statistical expected value \( \mu_{\text{sample}} \) and standard deviation \( \sigma_{\text{sample}} \) of the data sample are

\[
\mu_{\text{sample}} = \frac{1}{N} \sum_{i=1}^{N} q_i \quad \text{(the arithmetic average)}
\]

and

\[
\sigma_{\text{sample}} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (q_i - \mu_{\text{sample}})^2}.
\]

We will not attempt to explain all of the subtitles of the above formulas in this primer. Instead, we will look at an example from a typical freshmen physics experiment. This example will serve as a prototype for more general statistical methods.

Before we look at the example, we first give the underlying mantra repeated by every experimental physicist.

**Experimental Mantra:** A measured number without an estimate of the uncertainty is meaningless.

We will adhere to this philosophy throughout our discussion.

**Example 12.** Suppose you conduct the following experiment: using a ball-drop apparatus in your local physics lab you measure the time it takes a ball to fall one meter starting from rest. You repeat the experiment 10 times. You now have a collection of the following 10 data points\(^5\):

\[
\{0.4548, 0.4529, 0.4535, 0.4538, 0.4537, 0.4526, 0.4531, 0.4538, 0.4545\} \text{ (units of seconds)}.
\]

What is your best guess as to the “true” time and how would you quantify the uncertainty in your best guess?

\(^5\)This is real data taken from an experiment conducted at Pima Community College’s West campus!
Now we need to write down a number based on our data for the true value of the fall time, which we’ll denote by \( t_{\text{fall}} \). What shall we write down?

The answer that immediately comes to mind is the average (expected) value. This is our best guess as to the true fall time. We will denote the sample average \( \mu_{\text{sample}} \) as \( t_{\text{best}} \). We use the subscript \( \text{best} \) to remind us that this is our best guess. Taking the average of the above data we find \( t_{\text{best}} = 0.4531 \text{ s} \).

Now comes the harder question: how sure are you that the \( t_{\text{best}} \) is close to the “true” value of \( t_{\text{fall}} \)?

The answer to this question is very involved, but the upshot is that we need a way of measuring the uncertainty in our experiment (i.e., the experimental error), and we will take the standard deviation as a measure of the uncertainty. In keeping with standard convention, we typically denoted the standard deviation in the time as \( \delta t \). For the above data \( \delta t = 0.0009 \).

We have avoided showing any details of an actual experiment in order to avoid any complications that might arrive from describing an experiment. The whole point of this example is to give the reader an idea of how the standard deviation might be used in practice to report a measured number.

The end result of our best guess plus uncertainty for the fall time is

\[
t_{\text{best}} = t_{\text{best}} \pm \delta t = 0.4531 \pm 0.0009 \text{ s}.
\]

This is the proper way to report measured numbers (along with any assumptions made).
10 Discrete Probability Distributions

10.1 Some basic properties of finite random variables

Let $X$ be a finite random variable with range $\mathcal{R}_X = \{x_1, \ldots, x_n\}$, where by convention $x_1 < x_2 < \cdots < x_n$. Recall that for all real numbers $x$, we’ve defined $f_X(x) = \mathbb{P}(X = x)$. The finite random variable $X$ must also satisfy the two properties:

$$0 \leq \mathbb{P}(X = x) \leq 1 \quad \text{for all real } x$$

$$\sum_{i=1}^{n} \mathbb{P}(X = x_i) = 1$$

The first of these is because the probability of any event has to be between 0 and 1. The second is because the random variable $X$ will certainly assume one of the $n$ values in its range.

We can immediately rewrite these two properties in terms of $f_X$:

$$0 \leq f_X(x) \leq 1 \quad \text{for all real } x$$

$$\sum_{i=1}^{n} f_X(x_i) = 1$$

If $x_i \leq x < x_{i+1}$, then $\mathbb{P}(X \leq x) = \mathbb{P}(X = x_1) + \cdots + \mathbb{P}(X = x_i)$. Recall that $F_X(x) = \mathbb{P}(X \leq x)$. That gives us the property:

$$\text{If } x_i \leq x < x_{i+1} \text{ then } F_X(x) = \sum_{j=1}^{i} f_X(x_j)$$

**Proof:** Assume $x_i \leq x < x_{i+1}$. Then

$$F_X(x) = \mathbb{P}(X \leq x)$$

$$= \mathbb{P}(X = x_1) + \cdots + \mathbb{P}(X = x_i)$$

$$= f_X(x_1 + \cdots + f_X(x_i))$$

$$= \sum_{j=1}^{i} f_X(x_j)$$

If $x \geq x_n$, then from (10.2) and (10.3), we get

$$F_X(x) = \sum_{i=1}^{n} f_X(x_i) = 1$$

(10.4)
One way of thinking of probability is as mass. Imagine you have a kilogram of “probability”—think of it as a substance like clay. Divide that kilogram of material into \( n \) pieces, not necessarily of equal size, and place one piece at each of the points \( x_1, \ldots, x_n \) along the number line: the values in the range of \( X \). We get the following graphical representation of the probability distribution of \( X \):

![Graphical Representation](image)

Figure 50: One kilogram of clay is distributed along the number line at the points in the range of \( X \). The mass at the point \( x_i \) represents the probability that the finite random variable \( X \) will take the value \( x_i \).

The mass of clay at any given point \( x \) is given by the probability mass function \( f_X(x) = P(X = x) \). We can think of the cumulative distribution function, \( F_X(x) \), as the sum of all the masses stuck to the number line from \(-\infty \) up to and including \( x \). In other words, if we walk the number line from left to right, picking up every piece of clay we come to all the way up to \( x \), then \( F_X(x) \) will the the total mass in our hands when we stop at \( x \).

From this analogy, we can see that \( F_X(x) \) is a nondecreasing function of \( x \). As we continue to walk from left to right along the number line, we can only pick up more clay—we never put any down, or add a negative mass to what we’ve already gathered. We’ll now state this formally:

If \( a < b \) then \( F_X(a) \leq F_X(b) \)

Notice that if \( x < x_1 \), then in walking from \(-\infty \) to \( x \), we don’t pick up any clay at all. Hence \( F_X(x) = 0 \). Notice also that if we walk from \(-\infty \) to \( x_1 \) and stop there, then the only clay we pick up is the piece at \( x_1 \). Hence

\[
F_X(x_1) = f_X(x_1)
\]  
(10.5)

Another thing to notice: If we walk from \(-\infty \) to \( x_{i-1} \), pause for a moment, and then walk on to \( x_i \), the only additional piece of clay we pick up is the one at \( x_i \). In mathematical terms,

\[
F_X(x_i) = F_X(x_{i-1}) + f_X(x_i),
\]

which we can rearrange to read

\[
f_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \quad \text{for } i = 2, 3, \ldots, n.
\]
So far, we’ve been looking at the function $F_X(x)$: the probability that $X$ is at most $x$. Now, let’s look at the probability that $X$ is greater than $x$. The cases are complementary: that is, either $X \leq x$ or $X > x$, but not both. Because of this, we can say

$$P(X > x) = 1 - P(X \leq x) = 1 - F_X(x).$$  

(10.6)

We now summarize what we’ve learned about the behavior of the cumulative distribution for the case $n = 3$:

Figure 51: A visual of how the probability mass function works: We imagine starting from $x = -\infty$ and walking to the right picking up and gathering (in a bag) probability mass that is distributed at three points $x_1, x_2,$ and $x_3$ along the $x$-axis.

Figure 52: The corresponding jumps in the cumulative distribution function corresponding to the increases in total probability mass in our bag.
10.2 Introducing the discrete random variable

**Recall:** A random variable $X$ is a function that maps each outcome $\omega$ in the sample space $\Omega$ to a real number $x \in \mathbb{R}$, where $\omega$ is the outcome of a non-deterministic experiment. (See Definition of a random variable and the picture following it.)

**Definition 11.** The collection of all possible values that the random variable can attain is the **range** of $X$. It will be denoted $\text{Range}(X)$ or $R_X$.

**Definition 12.** A random variable $X$ is **discrete** if for any two distinct numbers $x < y \in \mathbb{R}$, the interval $(x, y)$ contains at most finitely many values in the range of $X$.

**The idea:** For a discrete random variable, every point in the range of $X$ is isolated from every other point. Suppose $X$ can only assume the values $x_1 < x_2 < \cdots$. Then we can find an interval that contains each $x_i$ but that doesn’t contain any other $x_j$ for $j \neq i$:

![Interval Diagram]

What could go wrong? Suppose that the range of $X$ consisted of the points $\{x_n = 1 - \frac{1}{n} | n = 1, 2, \ldots\}$. That is, $x_1 = 0$, $x_2 = 1/2$, $x_3 = 2/3$, \ldots. As $n$ goes to $\infty$, $x_n$ approaches 1. Here’s a picture:

![Infinite Sequence of Points Approaching 1 Diagram]

Notice that for any $n$, the interval $(x_n, 1)$ will contain an infinite number of points: $\{x_{n+1}, x_{n+2}, \ldots\}$. Thus a random variable with this kind of range would not be discrete by our definition. This is because 1 is not an isolated point. Any interval that includes 1 will contain infinitely many of the $x_n$’s.

**Advanced Discussion:** In most probability books, a discrete random variable is defined as a random variable whose range is finite or countably infinite (that is, in a one-to-one correspondence with the natural numbers $\mathbb{N} = \{1, 2, \ldots\}$). Under this definition, a random variable with a range like the one we’ve just discussed would be discrete. However, a random variable with a countably infinite range that wasn’t discrete by our definition would be weirder than anything we’d want to deal with in this class! For example, a random variable $X_{\mathbb{Q}}$ whose range is the rational numbers $\mathbb{Q}$ would not satisfy our definition of a discrete random variable since the rational numbers are dense in the real numbers (i.e. not discrete).
Definition 13. A finite random variable is a random variable that can only assume a finite number of distinct values. That is, it’s a random variable whose range is finite.

We will use $N$ (a natural number: $N \in \mathbb{N} = \{1, 2, 3, \ldots\}$) to indicate the number of elements in the range of a finite random variable $X$. For definiteness, we will write the range of a finite random variable $X$ as $\{x_1, x_2, \ldots, x_N\}$, and we will assume that $x_1 < x_2 < \cdots < x_N$. Notice that this ordering requires all the $x_i$’s to be distinct: if $i \neq j$, then $x_i \neq x_j$.

Note: All finite random variables are discrete.

10.3 Introducing the p.m.f and c.d.f.

Definition 14. The probability mass function (p.m.f.) of a finite random variable is

$$f_X(x) = \mathcal{P}(X = x).$$

Recall that $\mathcal{P}(X = x)$ is shorthand for the probability of the event $\{\omega \in \Omega \mid X(\omega) = x\}$. Thus we can think of $\mathcal{P}(X = x)$ as a function of $x$, so it would be natural to name such a function $f(x)$. To emphasize the fact that this function depends on the random variable $X$, we will include a subscript of $X$, giving us $f_X(x)$.

Because $f_X(x)$ is a function based on the probability of various events, we should expect that $f_X(x)$ will inherit some properties from probabilities. In fact, its very name suggests a connection with probability. Moreover, since $f_X(x)$ equals a probability, and probability has a connection with mass (see chapter ??), it follows that $f_X(x)$ is connected to mass, hence the name “probability mass function”.

Example 1. Let $X$ be the number of heads minus the number of tails in two tosses of a fair coin. Here $X = \#(\text{heads in two flips}) - \#(\text{tails in two flips})$. Find $f_X(x)$ and $F_X(0)$.

The sample space is the collection of outcomes from the experiment of flipping a coin twice. It is

$$\Omega = \{(TT, TH, HT, HH)\}.$$

Each outcome is equally likely. The range of $X$ consists of all possible values that $X$ can assume: $\{-2, 0, 2\}$. In our notation: $x_1 = -2, x_2 = 0, x_3 = 2$.

<table>
<thead>
<tr>
<th>$X$</th>
<th>$\mathcal{P}(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = -2$ corresponds to ${TT}$</td>
<td>$\mathcal{P}(X = -2) = 1/4$</td>
</tr>
<tr>
<td>$X = 0$  corresponds to ${TH, HT}$</td>
<td>$\mathcal{P}(X = 0) = 2/4 = 1/2$</td>
</tr>
<tr>
<td>$X = 2$  corresponds to ${HH}$</td>
<td>$\mathcal{P}(X = 2) = 1/4$</td>
</tr>
</tbody>
</table>

Using $f_X(x) = \mathcal{P}(X = x)$, and summarizing the results in a table we get:
The desired value of the cumulative distribution is

\[ F_X(0) = f_X(-2) + f_X(0). \]

Next, we introduce several properties of the p.m.f. and the c.d.f., defined below. To keep this theoretical discussion from being too stuffy, we will parallel each property of the p.m.f. and c.d.f. with an example.

### 10.3.1 Properties of the p.m.f and c.d.f.

**Property 1:** For all values of \( x \), \( 0 \leq f_X(x) \leq 1 \).

This follows because the probability of any event is always between 0 and 1. In particular, \( 0 \leq P(X = x) \leq 1 \).

**Property 2:** If the range of a finite random variable is \( \{x_1, x_2, \ldots, x_N\} \), then

\[
\sum_{all \ x} f_X(x) = \sum_{i=1}^{N} P(X = x_i) = 1
\]

This follows because the events \( \{X = x_i\} \) and \( \{X = x_j\} \) are disjoint for \( i \neq j \) (i.e. \( \{X = x_i\} \cap \{X = x_j\} = \emptyset \)), and because \( \bigcup_{i=1}^{N} \{X = x_i\} = \Omega \) (the sample space) and \( P(\Omega) = 1 \). In other words, the events \( \{X = x_i\} \) form a partition of the sample space \( \Omega \).

(see section 10.3.4)

Going back to example (1), we see that \( f_X(x) \) satisfies properties 1 and 2 since \( \sum f_X(x) = f_X(x_1) + f_X(x_2) + f_X(x_3) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1 \) and all the probabilities are positive.

**Definition 15.** The **cumulative distribution function (c.d.f.)** of a finite random variable is

\[ F_X(x) = P(X \leq x) = \sum_{t \leq x} f_X(t). \]

Let \( x_1 < x_2 < \cdots < x_N \) be the range of \( X \).
Property 3: If \( x \in [x_i, x_{i+1}) \) for some \( i \in \{1, \ldots, N\} \), then
\[
F_X(x) = \sum_{k=1}^{i} f_X(x_k) = f_X(x_1) + \cdots + f_X(x_i).
\]
From property 3, we see that

Property 4: (i) \( F_X(x_N) = \sum_{i=1}^{N} \mathbb{P}(X = x_i) = 1. \)
This is because \( x_N \) is the largest value in the range of \( X \).
(ii) If \( x < x_1 \), then \( F_X(x) = 0 \).
This is because \( x_1 \) is the smallest value in the range of \( X \).

Fact: If \( a < b \), then \( F_X(a) \leq F_X(b) \).
This shows that \( F_X(x) \) is a non-decreasing function of \( x \). We will verify this for the case \( N = 3 \).
For \( x \in [x_1, x_2) \), \( F_X(x) = f_X(x_1) \).
For \( y \in [x_2, x_3) \), \( F_X(y) = f_X(x_1) + f_X(x_2) \).
For \( z \in [x_3, \infty) \), \( F_X(z) = f_X(x_1) + f_X(x_2) + f_X(x_3) \).
Since \( f_X(x) \) can never be negative for any value of \( x \), we see that
for \( x < y < z \), \( F_X(x) \leq F_X(y) \leq F_X(z) \).

Note: If \( a, b \in [x_i, x_{i+1}) \) for some \( i \), then \( F_X(a) = F_X(b) \).

To make these concepts concrete, let’s return to example (1). First, we compute the c.d.f. for the problem:
If \( x < x_1 = -2 \), then \( F_X(x) = 0 \).
If \( x_1 = -2 \), then \( F_X(-2) = f_X(-2) = \frac{1}{4} \).
If \( -2 < x < 0 \), say \( x = -1 \), then \( F_X(-1) = \mathbb{P}(X \leq -1) = \mathbb{P}(X = -2) = f_X(-2) = \frac{1}{4} \).
If \( x_1 = 0 \), then \( F_X(0) = f_X(-2) + f_X(0) = \frac{1}{4} + \frac{1}{2} \).
If \( 0 < x < 2 \), say \( x = 1 \), then
\[
F_X(1) = \mathbb{P}(X \leq 1) = \mathbb{P}(X = -2) + \mathbb{P}(X = 0) = f_X(-2) + f_X(0) = \frac{1}{4} + \frac{1}{2}.
\]
If \( x_1 = 2 \), then \( F_X(2) = f_X(-2) + f_X(0) + f_X(2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1 \).
If \( 2 < x \), say \( x = 3 \), then
\[
F_X(3) = \mathbb{P}(X \leq 3) = \mathbb{P}(X = -2) + \mathbb{P}(X = 0) = f_X(-2) + f_X(0) = \frac{1}{4} + \frac{1}{2}.
\]

Property 5: If the range of the random variable \( X \) consists of the values \( x_1 < x_2 < \cdots < x_N \), then
\[
f_X(x_1) = F_X(x_1),
\]
and for \( i = 2, 3, \ldots, N \),
\[
f_X(x_i) = F_X(x_i) - F_X(x_{i-1}).
\]
Proof: Let $i \in \{1, \ldots, N\}$ be fixed. Then
\[
F_X(x_i) - F_X(x_{i-1}) = \sum_{k=1}^{i} f_X(x_k) - \sum_{k=1}^{i-1} f_X(x_k)
= f_X(x_i) + \sum_{k=1}^{i-1} f_X(x_k) - \sum_{k=1}^{i-1} f_X(x_k) \quad \text{(rewriting the first sum)}
= f_X(x_i).
\]

**Property 6:** $P(X > x) = 1 - F_X(x)$.
This is very straightforward: the set $\{X > x\}$ is the complement of the set $\{X \leq x\}$. Thus $\{X \leq x\}$ and $\{X > x\} = \{X \leq x\}^c$ partition the range. It follows that
\[
\Omega = \{X \leq x\} \cup \{X > x\} \quad \Rightarrow \quad 1 = P(\Omega) = P(\{X \leq x\}) + P(\{X > x\})
\]
The result follows from the definition $F_X(x) = P(\{X \leq x\}) = P(X \leq x)$.

**Property 7:** If the range of $X$ is $x_1 < x_2 < \cdots < x_N$, then
\[
P(X \geq x_i) = 1 - F_X(x_{i-1}) \quad \text{for } i = 2, \ldots, N
\]
\[
= 1 \quad \text{for } i = 1.
\]

Proof: If $i = 1$, $P(X \geq x_i) = P(X \geq x_1) = 1$, since the range of $X$ contains no values smaller than $x_1$.
Suppose $i \geq 2$. Then
\[
P(X \geq x_i) = P(X > x_i) + P(x = x_i)
= P(X > x_i) + f_X(x_i)
= 1 - F_X(x_i) + f_X(x_i) \quad \text{by (vi)}
= 1 - F_X(x_i) + F_X(x_i) - F_X(x_{i-1}) \quad \text{by (v)}
= 1 - F_X(x_{i-1})
\]

10.3.2 Interpretation of the formulas
Recall property 5, above: if the range of a finite random variable $X$ is $x_1 < x_2 < \cdots < x_N$, then
\[
f_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \quad \text{for } i = 2, 3, \ldots, N
\]
Using the definitions of $f_X$ and $F_X$, we can rewrite this as
\[
P(X = x_i) = P(X \leq x_i) - P(X \leq x_{i-1}) \quad \text{for } i = 2, 3, \ldots, N
\]
Let’s see how this version applies for the case when \( N = 5 \) and \( i = 4 \). The range of \( X \) is \( x_1 < x_2 < x_3 < x_4 < x_5 \), and \( i - 1 = 3 \).

\[
\mathcal{P}(X \leq x_4) = \mathcal{P}(X = x_1) + \mathcal{P}(X = x_2) + \mathcal{P}(X = x_3) + \mathcal{P}(X = x_4),
\]
\[
\mathcal{P}(X \leq x_3) = \mathcal{P}(X = x_1) + \mathcal{P}(X = x_2) + \mathcal{P}(X = x_3).
\]

Subtract the second equation from the first:

\[
\mathcal{P}(X \leq x_4) - \mathcal{P}(X \leq x_3) = 0 + 0 + 0 + \mathcal{P}(X = x_4).
\]

Next, we give an algorithm for computing the c.d.f.

**10.3.3 Graphing the cumulative distribution of a finite random variable**

Let \( X \) be a finite random variable with range \( x_1 < \cdots < x_N \). Let’s look at the cumulative distribution function \( F_X(x) \) for this \( X \). We now have an algorithm for generating \( F_X(x) \).

**Step 1.** Compute:

\[
F_X(x_1) = f_X(x_1),
\]
\[
F_X(x_i) = F_X(x_{i-1}) + f_X(x_i) \quad \text{for } i = 2, \ldots, N.
\]

**Step 2.** Use the fact that if \( x \in [x_i, x_{i+1}) \), i.e. \( x_i \leq x < x_{i+1} \), then \( F_X(x) = F_X(x_i) \). The graph is a nondecreasing step function:

We can make sense of this graph using our mass analogy. Imagine a mass of one kilogram of clay placed distributed at \( x_1, x_2, \) and \( x_3 \) along the \( x \)-axis. If we walk from left-to-right along the \( x \)-axis carrying a satchel (bag) and if we pick at each piece of clay when we reach the point where it is located along the \( x \)-axis, then we would gain mass as we walked from left-to-right along the \( x \)-axis. In between the discrete pieces of mass we’d expect our mass to remain constant. However, once we reach each point \( x_i \), we’d expect our mass to increase by an amount \( f_X(x_i) \), which is exactly what we see in the above graph: the jump distance of the step function at each \( x_i \) is exactly \( f_X(x_i) \).
10.3.4 The range of a random variable $X$ partitions the sample space $\Omega$

**Fact:** Suppose $X$ is a finite random variable whose range is $x_1 < x_2 < \ldots < x_N$. Then the sets $\{X = x_i\}$ for $i = 1,\ldots,N$ form a partition.

**Proof:** The set $\{X = x_i\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x_i\}$. Now, $X$ is a function that maps values in $\Omega$ to values in $\mathbb{R}$; we write this as $X : \Omega \to \mathbb{R}$. The domain of $X$ is $\Omega$. To show that $X$ partitions $\Omega$, we need to show two things:

(i) The sets $\{X = x_i\}$ includes everything in $\Omega$:

$$\bigcup_{i=1}^{N} \{\omega \in \Omega : X(\omega) = x_i\} = \Omega$$

(ii) The sets $\{X = x_i\}$ are disjoint: if $i \neq j$, then

$$\{\omega \in \Omega : X(\omega) = x_i\} \cap \{\omega \in \Omega : X(\omega) = x_j\} = \emptyset$$

To prove part (i), notice that for each $i$ the set $\{\omega \in \Omega : X(\omega) = x_i\} \subseteq \Omega$. The sets are all defined as subsets of $\Omega$. Hence

$$\bigcup_{i=1}^{N} \{\omega \in \Omega : X(\omega) = x_i\} \subseteq \Omega.$$ 

To prove equality, we still need to show that $\Omega \subseteq \bigcup_{i=1}^{N} \{\omega \in \Omega : X(\omega) = x_i\}$. Let $\omega_0 \in \Omega$. Since the domain of $X$ is all of $\Omega$ and the range of $X$ is $\{x_1,\ldots,x_N\}$, there must be some value of $j \in \{1,\ldots,N\}$ for which $X(\omega_0) = x_j$. That means

$$\omega_0 \in \{\omega \in \Omega : X(\omega) = x_j\} \subseteq \bigcup_{i=1}^{N} \{\omega \in \Omega : X(\omega) = x_i\}.$$ 

Since this is true for every $\omega \in \Omega$,

$$\Omega \subseteq \bigcup_{i=1}^{N} \{\omega \in \Omega : X(\omega) = x_i\}.$$ 

Since each set is a subset of the other, the two must be equal.

To prove part (ii), we’ll assume that it’s not true and show that this leads to false results. This is known as *proof by contradiction*. Assume that for some $i \neq j$,

$$\{\omega \in \Omega : X(\omega) = x_i\} \cap \{\omega \in \Omega : X(\omega) = x_j\} \neq \emptyset.$$ 

Then there must exist some $\omega_0 \in \Omega$ for which $X(\omega_0) = x_i$ and $X(\omega_0) = x_j$, but $x_i \neq x_j$. But that means $X(\omega_0) \neq X(\omega_0)$. Since $X$ is a function on $\Omega$, this can’t be true: functions have to take each element in their domain to one and only one element in their range. Hence no such $\omega_0$ can exist.

(Recall: A function can only have one value of $y$ for any value of $x$ in its domain. Otherwise, it would fail the vertical-line test.)
10.4 Some examples of finite random variables

Now let’s return to Example 2 on page 164. We rolled a fair die twice, and the value of the random variable \( X \) was the sum of the two rolls.

10.4.1 Computing \( f_X(x) \) for the 2-dice problem

The event \( \{X = i\} \) for \( i = 2, \ldots, 12 \) is shorthand for \( \{\omega \in \Omega \mid X(\omega) = i\} \). Let \( \tilde{x} \) and \( \tilde{y} \) be the face values of the first and second roll. Then since all outcomes are equally likely,

\[
\mathcal{P}(\{\omega \in \Omega \mid X(\omega) = i\}) = \mathcal{P}(X = i) = \frac{\# \text{ of elements } (\tilde{x}, \tilde{y}) \text{ with } \tilde{x} + \tilde{y} = i}{\#(\Omega)}.
\]

For \( i = 2 \), the only outcome \((\tilde{x}, \tilde{y})\) that will satisfy \( \tilde{x} + \tilde{y} = 2 \) is \((1, 1)\). For \( i = 3 \), there are two outcomes with \( \tilde{x} + \tilde{y} = 3 \): \((1, 2)\) and \((2, 1)\). For \( i = 4 \), there are three outcomes: \((1, 3)\), \((2, 2)\), and \((3, 1)\). The sample space \( \Omega \) contains a total of 36 outcomes (see the table on page 164), so \( \#(\Omega) = 36 \). Thus

\[
\begin{align*}
f_X(2) &= \mathcal{P}(X = 2) = \frac{1}{36}, \\
f_X(3) &= \mathcal{P}(X = 3) = \frac{2}{36}, \\
f_X(4) &= \mathcal{P}(X = 4) = \frac{3}{36}.
\end{align*}
\]

In general, we have

\[
f_X(i) = \frac{6 - |i - 7|}{36} \quad \text{for } i = 2, 3, \ldots, 12. \tag{10.7}
\]

You should verify this!

Notice that this function is symmetric about \( i = 7 \). In fact, if we generalize it to a continuous function of \( x \),

\[
f_X(x) = \frac{6 - |x - 7|}{36},
\]

then it is symmetric about the line \( x = 7 \). We would express this mathematically as \( f_X(7 + z) = f_X(7 - z) \). The results can also be expressed in a table as shown below. Notice that we can see the symmetry in the values in the table as well.

<table>
<thead>
<tr>
<th>( x )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
</table>

Figure 53 is a graph of the probability mass function \( f_X(x) \) for the two-dice example. The graph comes from Equation (10.7).
10.4.2 Computing $F_X(x)$ for the 2-dice problem

To compute $F_X(x_i)$, we will use property 5 on page 196. This states that $f_X(x_i) = F_X(x_i) - F_X(x_{i-1})$ for $i = 2, \ldots, N$. We’ll rearrange the equation:

$$F_X(x_i) = F_X(x_{i-1}) + f_X(x_i) \quad \text{for } i = 2, \ldots, N. \quad (10.8)$$

For $x < 2$, the event $X = x$ cannot occur, since $x$ is not in the range of $X$. Hence $\mathcal{P}(X = x) = \mathcal{P}(\emptyset) = 0$ (remember that $\emptyset$ is the empty set).

At $x = 2$, we have $F_X(2) = f_X(2) = \mathcal{P}(X = 2) = 1/36$. For $x \in [2, 3)$ (that is, $2 \leq x < 3$), we have $F_X(x) = \mathcal{P}(X \leq x) = \mathcal{P}(X = 2) = 1/36$.

For $i = 3, 4, \ldots, 12$, we’ll use equation (10.8) to calculate $F_X(i)$. For $x \in [i, i + 1)$, $F_X(x) = F_X(i)$. Finally, if $x \geq 12$, $F_X(x) = F_X(12) = 1$, since 12 is the largest value in the range of $X$.  

Figure 53: Two-dice example: $f_X(x)$
Let’s do the computations—

\[
F_X(2) = \frac{1}{36}
\]

\[
F_X(3) = F_X(2) + f_X(3) = \frac{1}{36} + \frac{2}{36} = \frac{3}{36} = \frac{1}{12}
\]

\[
F_X(4) = F_X(3) + f_X(4) = \frac{3}{36} + \frac{3}{36} = \frac{6}{36} = \frac{1}{6}
\]

\[
F_X(5) = F_X(4) + f_X(5) = \frac{6}{36} + \frac{4}{36} = \frac{10}{36} = \frac{5}{18}
\]

\[
F_X(6) = F_X(5) + f_X(6) = \frac{10}{36} + \frac{5}{36} = \frac{15}{36} = \frac{5}{12}
\]

\[
F_X(7) = F_X(6) + f_X(7) = \frac{15}{36} + \frac{6}{36} = \frac{21}{36} = \frac{7}{12}
\]

\[
F_X(8) = F_X(7) + f_X(8) = \frac{21}{36} + \frac{5}{36} = \frac{26}{36} = \frac{13}{18}
\]

\[
F_X(9) = F_X(8) + f_X(9) = \frac{26}{36} + \frac{4}{36} = \frac{30}{36} = \frac{5}{6}
\]

\[
F_X(10) = F_X(9) + f_X(10) = \frac{30}{36} + \frac{3}{36} = \frac{33}{36} = \frac{11}{12}
\]

\[
F_X(11) = F_X(10) + f_X(11) = \frac{33}{36} + \frac{2}{36} = \frac{35}{36}
\]

\[
F_X(12) = F_X(11) + f_X(12) = \frac{35}{36} + \frac{1}{36} = \frac{36}{36} = 1
\]

Figure 54 is the cumulative distribution function \(F_X(x)\) for the two-dice example. The data for this graph are in Equation (10.9) on page 202.
In the next example we do some computations with the cumulative distribution for the two dice problem.

Example 2. Referring to the two-dice problem above:
(a) Compute $F_X(5)$.

We just did that; but we did it by computing $F_X(2)$, then $F_X(3)$, then... A more direct way is to calculate

$$F_X(5) = \mathcal{P}(X \leq 5)$$
$$= \mathcal{P}(X = 2) + \mathcal{P}(X = 3) + \mathcal{P}(X = 4) + \mathcal{P}(X = 5)$$
$$= f_X(2) + f_X(3) + f_X(4) + f_X(5)$$
$$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{10}{36}.$$

(b) Compute $\mathcal{P}(X < 5)$ and compare it to $\mathcal{P}(X \leq 5)$.

$$\mathcal{P}(X < 5) = \mathcal{P}(X = 2) + \mathcal{P}(X = 3) + \mathcal{P}(X = 4)$$
$$= f_X(2) + f_X(3) + f_X(4)$$
$$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{6}{36}.$$
Notice that the event \( \{ X \leq 5 \} = \{ X < 5 \} \cup \{ X = 5 \} \). The two sets in the union are disjoint, so
\[
P(X \leq 5) = P(X < 5) + P(X = 5) = P(X < 5) + f_X(5) .
\]

(c) Compute \( P(4 \leq X < 9) \).
\[
P(4 \leq X < 9) = P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8)
= f_X(4) + f_X(5) + f_X(6) + f_X(7) + f_X(8)
= \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} = \frac{23}{36} .
\]

(d) Compute \( P(4 < X < 9) \) and compare it to \( P(4 \leq X < 9) \).
\[
P(4 < X < 9) = P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8)
= f_X(5) + f_X(6) + f_X(7) + f_X(8)
= \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} = \frac{20}{36} .
\]
The set \( \{ 4 \leq X < 9 \} = \{ 4 < X < 9 \} \cup \{ X = 4 \} \). The two sets in the union are disjoint, so
\[
P(4 \leq X < 9) = P(4 < X < 9) + P(X = 4) = P(4 < X < 9) + f_X(4) .
\]

(e) Compute \( F_X(4) \).
\[
F_X(4) = P(X \leq 4)
= P(X = 2) + P(X = 3) + P(X = 4)
= f_X(2) + f_X(3) + f_X(4)
= \frac{1}{36} + \frac{3}{36} + \frac{3}{36} = \frac{6}{36} = \frac{1}{6} .
\]

(f) Compute \( F_X(9) \).
\[
F_X(9) = P(X \leq 9)
= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9)
= f_X(2) + f_X(3) + f_X(4) + f_X(5) + f_X(6) + f_X(7) + f_X(8) + f_X(9)
= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} + \frac{6}{36} + \frac{5}{36} + \frac{4}{36} = \frac{30}{36} = \frac{5}{6} .
\]

(g) Compute \( P(4 < X \leq 9) \).
We could calculate this as
\[
P(4 < X \leq 9) = P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) .
\]
Since we already have $F_X(4)$ and $F_X(9)$, there’s a faster way. Rewrite equation (10.10):

$$\mathbb{P}(4 < X \leq 9) = \sum_{i=5}^{9} \mathbb{P}(X = i).$$

We can add and subtract the same quantity without changing the total, so

$$= \sum_{i=5}^{9} \mathbb{P}(X = i) + \sum_{i=2}^{4} \mathbb{P}(X = i) - \sum_{i=2}^{4} \mathbb{P}(X = i)$$

$$= \sum_{i=2}^{9} \mathbb{P}(X = i) - \sum_{i=2}^{4} \mathbb{P}(X = i)$$

$$= \mathbb{P}(X \leq 9) - \mathbb{P}(X \leq 4) = F_X(9) - F_X(4)$$

$$= \frac{30}{36} - \frac{6}{36} = \frac{24}{36}.$$

In general,

$$\mathbb{P}(x < X \leq y) = \mathbb{P}(X \leq y) - \mathbb{P}(X \leq x) = F_X(y) - F_X(x).$$

Similarly,

$$\mathbb{P}(x \leq X \leq y) = \mathbb{P}(X \leq y) - \mathbb{P}(x < X).$$

Notice we get a different formula when we include the right endpoint. In general, this is true: when dealing with the probability of an event happening over an interval, the result can vary depending on whether or not you include the endpoints!

(h) Compute the expected value of $X$. This is the average value of the random variable.

$$E(X) = \sum_{\text{all } x} x \mathbb{P}(X = x) = \sum_{\text{all } x} x f_X(x) = \sum_{i=2}^{12} i f_X(i).$$

If we write this sum out, we get

$$E(X) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36}$$

$$+ 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36}$$

$$= \frac{1}{36} (2 + 6 + 12 + 20 + 30 + 42 + 40 + 36 + 30 + 22 + 12)$$

$$= \frac{252}{36} = 7.$$

This is the value we’d expect to get, just by looking at the problem.
10.4.3 Two flips of a fair coin problem

Example 3. Let $X$ be the number of heads in two tosses of a fair coin.

(a) Find $f_X(x)$.
(b) Plot a bar graph of $f_X(x)$
(c) Find $F_X(x)$.

Solution:
The sample space is the collection of outcomes from the experiment of flipping a coin twice. It is
$$\Omega = \{(TT, TH, HT, HH)\}.$$ Each outcome is equally likely. The range of $X$ consists of all possible values that $X$ can assume: $\{0, 1, 2\}$.

| $X = 0$ corresponds to $\{TT\}$ | $P(X = 0) = 1/4$ |
| $X = 1$ corresponds to $\{TH, HT\}$ | $P(X = 1) = 2/4 = 1/2$ |
| $X = 2$ corresponds to $\{HH\}$ | $P(X = 2) = 1/4$ |

part(a): Since $f_X(x) = P(X = x)$, we get:

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_X(x)$</td>
<td>1/4</td>
<td>1/2</td>
<td>1/4</td>
</tr>
</tbody>
</table>

part(b): To plot a bar graph (probability histogram), we take the height of each rectangle to be equal to the probability that $X$ takes the value which corresponds to the midpoint of its base.
part(c):

\[ F_X(0) = f_X(0) = \frac{1}{4} \]

\[ F_X(1) = \sum_{i=0}^{1} P(X = i) \]
\[ = P(X = 0) + P(X = 1) \]
\[ = f_X(0) + f_X(1) \]
\[ = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \]

\[ F_X(2) = F_X(1) + f_X(2) \]
\[ = f_X(0) + f_X(1) + f_X(2) = 1 \]

**Note:** We have just given values for \( F_X(x) \) at the discrete points \( x = 0, 1, 2 \). If we graphed \( F_X(x) \) as a function of \( x \), it would be a discontinuous step function like the graph on page 198.

### 10.4.4 more examples

**Example 4.** Let \( X \) be a finite random variable with the following probability mass function.

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_X(x) )</td>
<td>.1</td>
<td>.3</td>
<td>.4</td>
<td>.2</td>
</tr>
</tbody>
</table>

(a) Plot a bar graph of \( f_X(x) \).

(a) Compute \( F_X(x) \) for the \( f_X(x) \)

**Solution:**

part (a): We’ll start by bar-graphing the table.
part (b): Since 1 is the smallest possible value for $X$,

$$F_X(x) = P(X \leq x) = 0 \quad \text{for } x < 1.$$  

At $x = 1$,

$$F_X(1) = f_X(1) = P(X = 1) = \frac{1}{10}.$$  

For $1 \leq x < 2$,

$$F_X(x) = F_X(1) = \frac{1}{10}.$$  

For $2 \leq x < 3$,

$$F_X(x) = f_X(1) + f_X(2) = P(X = 1) + P(X = 2) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}.$$  

For $3 \leq x < 4$,

$$F_X(x) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{1}{10} + \frac{3}{10} + \frac{4}{10} = \frac{8}{10}.$$  

For $4 \leq x$,

$$F_X(x) = \sum_{i=1}^{4} P(X = i) = 1.$$  

Here’s a graph of $F_X(x)$ versus $x$:

![Graph of $F_X(x)$ versus $x$]

Note: The hollow dots symbolize jump discontinuities.
Example 5. Let $X$ be a random variable with the following cumulative distribution function:

$$F_X(x) = \begin{cases} 
1 & \text{if } x > 8 \\
\frac{7}{10} & \text{if } 6 \leq x < 8 \\
\frac{3}{10} & \text{if } 4 \leq x < 6 \\
0 & \text{if } x < 4.
\end{cases}$$

(a) What are the possible values for $X$?
(b) Find all values for $f_X(x)$.

Solution:

part (a) The points of discontinuity in the graph of $F_X(x)$ will be the values in the range of $X$. $F_X(x)$ has discontinuities at $x = 4, 6,$ and $8$. Thus the range of $X$ is $\{4, 6, 8\}$.

part (b) We will use property (v) on page 196:

$$f_X(x_i) = F_X(x_i)$$

$$f_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \quad \text{for } i = 2, \ldots, N.$$ 

Here $x_1 = 4, x_2 = 6,$ and $x_3 = 8$. We get

$$f_X(4) = F_X(4) = \frac{3}{10},$$
$$f_X(6) = F_X(6) - F_X(4) = \frac{7}{10} - \frac{3}{10} = \frac{4}{10},$$
$$f_X(8) = F_X(8) - F_X(6) = 1 - \frac{7}{10} = \frac{3}{10}.$$ 

As a check, notice that

$$\sum_{\text{all } x} f_X(x) = f_X(4) + f_X(6) + f_X(8) = 1 = \frac{3}{10} + \frac{4}{10} + \frac{3}{10} = 1.$$
10.5 Bernoulli trials

Experiments for which there are only two possible outcomes are called **Bernoulli trials**. In statistical work, it’s traditional to label one of the outcomes “success” and the other one “failure.” Experiments with more than two possible outcomes can be converted to Bernoulli trials by grouping the outcomes into two categories. For example, if we toss a die, there are six possible outcomes: 1, 2, 3, 4, 5, 6. But if we call the outcome 1 a success, and the outcomes 2, 3, 4, 5, and 6 failures, then we can view this as a Bernoulli experiment with only two outcomes: success and failure.

In working with Bernoulli trials, we use \( \theta \) to denote the probability of success. Then the probability of failure is \( 1 - \theta \). That is, if \( P(\text{success}) = \theta \), then \( P(\text{failure}) = 1 - \theta \). If we use the event \( X = 1 \) and \( X = 0 \) to denote the number of “successes” and “failures” respectively, then \( P(X = 1) = \theta \) and \( P(X = 0) = 1 - \theta \).

The number of successes in a single Bernoulli trial must be 0 or 1.

**Definition 16.** A random variable \( X \) has a **Bernoulli distribution** if its range is \( \{0, 1\} \) and if its probability mass function is

\[
f_X(x) = \theta^x (1 - \theta)^{1-x} \quad \text{for } x = 0, 1.
\]

Thus, a **Bernoulli trial** is an experiment with only two outcomes: success and failure. In this context, “success” and “failure” are mathematical terms, not value judgements. For a particular Bernoulli trial, we might define “success” as “going bankrupt”, “catching the flu”, “being eaten by crocodiles” or something else that might not be considered especially successful in the real world.

**Example 6.** Toss a coin; define “heads” as a success.

**Example 7.** Roll a die; define “rolling a two” as a success.

**Example 8.** Launch a cruise-missile attack on a terrorist-training camp. Define “killing the leader” as a success.

Suppose that the probability of success in a single Bernoulli trial is \( \theta \). Now, suppose we repeat the trial \( n \) times. Assume that each trial is independent of all the others, and with the same probability of success \( \theta \). We can define the following finite random variable:

\[
X = \# \text{ of successes in } n \text{ trials}
\]

This is known as a **binomial random variable**. It is finite; the range of \( X \) is \( \mathcal{R}_X = \{0, 1, \ldots, n\} \), since we can have as few as zero successes in \( n \) trials, but no more than \( n \) successes. We now introduce this process formally.
### 10.6 The binomial distribution

Repeated trials play a very important role in probability and statistics, especially when the number of trials is fixed, the probability of success \( \theta \) for each trial is the same, and the trials are all independent of one another. If we do such a series of \( n \) Bernoulli trials, then the probability of \( x \) successes in \( n \) trials is

\[
\binom{n}{x} \theta^x (1 - \theta)^{n-x}
\]

where \( \binom{n}{x} = \frac{n!}{x!(n-x)!} \).

Here \( \binom{n}{x} \) is the number of ways we can have \( x \) successes in \( n \) trials; \( \theta^x \) is the probability that we will have a success on each of \( x \) trials; and \( (1 - \theta)^{n-x} \) is the probability that we will have a failure on each of \( n - x \) trials.

**Definition 17.** A random variable \( X \) has a **binomial distribution** if its probability mass function is

\[
b_X(x; n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}.
\]

**(10.11)**

**Example 9.** Suppose you flip a coin 50 times. The total number of heads is a binomial random variable, with range \( \{0, 1, \ldots, 50\} \).

**Example 10.** You randomly stop 150 students on campus and demand to see the contents of their pockets. If a student has a Missouri state quarter, it counts as a success. The total number of students carrying Missouri quarters is a binomial random variable, with range \( \{0, 1, \ldots, 150\} \).

**Example 11.** A building contains 12,000 fluorescent bulbs. Each one has an equal probability of burning out on a given day. The number of bulbs that actually burn out on a certain day is a binomial random variable, with range \( \{0, 1, \ldots, 12,000\} \).

Binomial random variables have an associated probability distribution, straightforwardly known as the **binomial distribution**. This is actually a whole family of distributions, each member of which is described by two parameters: the total number \( n \) of Bernoulli trials; and the probability of success \( \theta \) in any one Bernoulli trial.

Next, let’s use the binomial distribution given in equation (10.11) to compute some probabilities.

**Example 12.** Find the probability of getting exactly 5 heads in 12 flips of a fair coin.

**Solution:** For any one flip, we will call an outcome of heads a success and tails a failure. Since the coin is fair, \( \theta = 1/2 \). Each flip is independent. Since there is a total of 12 flips, \( n = 12 \). Plugging these numbers into Equation (10.11) gives us

\[
b_X(5; n = 12, \theta = 1/2) = \binom{12}{5} \left( \frac{1}{2} \right)^5 \left( 1 - \frac{1}{2} \right)^{12-5}
\]

\[
= \binom{12}{5} \left( \frac{1}{2} \right)^{12} \approx .19.
\]
What if we change the number of successes to be half of the total number of trials?

**Example 13.** Find the probability of getting exactly 6 heads in 12 flips of a fair coin.

**Solution:** For any one flip, we will call an outcome of heads a success and tails a failure. Since the coin is fair, \( \theta = \frac{1}{2} \). Each flip is independent. Since there is a total of 12 flips, \( n = 12 \). Plugging these numbers into Equation (10.11) gives us

\[
b_X(6; n = 12, \theta = 1/2) = \binom{12}{6} \left( \frac{1}{2} \right)^6 \left( 1 - \frac{1}{2} \right)^{12-6} = \binom{12}{6} \left( \frac{1}{2} \right)^{12} \approx .225 \neq .5.
\]

How many of you thought you were going to get 1/2?

**Example 14.** The probability that any one person with a certain tropical disease will recover is .8. Find the probability that 7 out of 10 people with the disease will recover.

**Solution:** We will assume that each person recovers independently from everyone else. If we call recovery a success, then we have \( x = 7 \), \( n = 10 \), and \( \theta = .8 \). Using Equation (10.11), we get

\[
b_X(7; n = 10, \theta = .8) = \binom{10}{7}(.8)^7(1-.8)^{10-7} \approx .20.
\]

The expected value is \( E(X) = 10 \cdot \frac{8}{10} = 8 \). This says that we expect 8 out of the 10 people to recover from the disease.

**Example 15.** If you flip a coin three times, what’s the probability of getting two heads? Define a trial to be one flip of a fair coin, and a success as the coin turning up heads on the flip. Then \( n = 3 \) (since 3 flips = 3 repeated experiments); \( x \) is the number of successes in 3 flips; and \( \theta = \frac{1}{2} \) is the probability of success on each independent identical trial. We want \( x = 2 \).

\[
b(x; n, \theta) = b \left( 2; 3, \frac{1}{2} \right) = \binom{3}{2} \left( \frac{1}{2} \right)^2 \left( \frac{1}{2} \right)^{3-2} = 3 \cdot \frac{1}{2^3} = \frac{3}{8}
\]

**Example 16.** If you roll a dice three times, what’s the probability of getting a one on two of the rolls? Define a trial as one roll of a die. Define a success as rolling the die and getting a face value of 1. Then \( \theta = \frac{1}{6} \) is the probability of success on each independent identical trial. With \( n = 3 \) and \( x = 2 \), we get

\[
b(x; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{3}{2} \left( \frac{1}{6} \right)^2 \left( 1 - \frac{1}{6} \right)^{3-2} = 3 \cdot \frac{1}{6^2} \cdot \frac{5}{6} = \frac{5}{72}
\]

**Note:** You should use repeat the above calculations using the Excel function BINOMDIST (see discussion below).
10.6.1 The expected value for a binomial distribution

The **expected value** for a binomial random variable $X$ is

$$E(X) = \mu = n\theta = \text{(number of trials)} \cdot \text{(probability of success on any one trial)}.$$  

Thus, a binomial random variable $X$ with parameters $n$ and $\theta$ has expected value

$$E[X] = n\theta.$$  

This is very intuitive: the expected total number of successes is the number of trials, times the probability of success on any one trial.

**Example 17.** Suppose you flip a fair coin 50 times. The total number of heads is a binomial random variable $X$ with parameters $n = 50$ and $\theta = 0.5$. The expected value of $X$ is: $E[X] = (50)(0.5) = 25$, which if we assume one out of every two flips (on average) results in heads, this is the “expected” outcome; hence the name.

**Example 18.** A political campaign makes automatic phone calls to 30,000 voters. The probability that any one voter will be in the shower when they call is 2%. The number of voters got out of the shower is a binomial random variable $X$, with parameters $n = 30,000$ and $\theta = 0.02$. The expected number of voters who will be called out of the shower is $E[X] = (30,000)(0.02) = 600$.

10.6.2 Using Excel to compute binomial distributions

The binomial distribution is so useful that Excel has a built-in function for computing the probability mass function $f_X(x)$ and the cumulative distribution function $F_X(x)$ for it:

$$f_X(x) = P(X = x) = \text{BINOMDIST}(x, n, \theta, \text{FALSE})$$

$$F_X(x) = P(X \leq x) = \text{BINOMDIST}(x, n, \theta, \text{TRUE})$$

**Example 19.** Suppose you flip a fair coin 5 times. Use Excel to calculate the probability that you will get no heads.

We have $n = 5$ Bernoulli trials, in which we’ll count “heads” as a success. Since the coin is fair, the probability of success is $\theta = 0.5$. We want $P(X = 0) = f_X(0)$. In Excel, we’ll type:

$$=\text{BINOMDIST}(0, 5, .5, \text{FALSE})$$
The value we’ll get is 0.03125. That’s what we’d expect: on each of the coin tosses, the probability of getting tails is $1/2$; so the probability of getting it on each of five tosses is

$$\left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.03125$$

**Example 20.** Same example—a fair coin tossed five times—but now, you want to know the probability of getting exactly 1 head.

We’ll use the same Excel function, but we’ll change the first argument from 0 to 1:

$$=\text{BINOMDIST}(1,5,.5,\text{FALSE})$$

We should get an answer of 0.15625. We can check this by hand: the probability of any one outcome of the experiment is $(1/2)^5 = 1/32$; and there are 5 possible outcomes that have exactly one head (heads on the first, tails on the rest; heads on the second, tails on the rest; etc.) Hence

$$f_X(1) = P(X = 1) = \frac{5}{32} = 0.15625$$

**Example 21.** Still the same—fair coin, tossed five times. Now, use Excel to find the probability that there will be at most one head.

We now want the cumulative distribution function instead of the probability mass function: $F_X(1) = P(X \leq 1)$. To get that, we change the last argument to **TRUE**:

$$=\text{BINOMDIST}(1,5,.5,\text{TRUE})$$

The answer we get is 0.1875. We can check this:

$$P(X \leq 1) = P(X = 0 \text{ or } X = 1)$$

$$= P(X = 0) + P(X = 1)$$

$$= 0.03125 + 0.15625 \quad \text{(from previous examples)}$$

$$= 0.1875$$

**Example 22.** A new brand of cell phone is subject to a battery defect, producing a 10% probability that the phone will explode when it is first turned on. If 700 of the phones have been sold, what is the probability that no more than 50 of them explode?

Unlike the previous examples, this would be very difficult to calculate by hand. We have $n = 700$ Bernoulli trials, with a probability of success $\theta = 0.1$. We want $F_X(50)$. In Excel, we want

$$=\text{BINOMDIST}(50,700,.1,\text{TRUE})$$
The answer we get back from Excel is approximately 0.0053, or 0.53%.

**Example 23.** 3000 people attend a concert. The probability that any one person will buy a souvenir T-shirt is 15%. What is the probability that no more than 500 T-shirts will be sold? (Assume nobody wants to buy more than one.)

Here’s a case with $n = 3000$ Bernoulli trials, with a probability of success $\theta = 0.15$. We want $F_X(500)$. This would be almost impossible to do by hand. To do it in Excel, use

$$=\text{BINOMDIST}(500,3000,0.15,\text{TRUE})$$

**Warning:** Not all versions of Excel can handle this. Older versions of Excel will return a #NUM! error. This is not because you did something wrong; there’s an intermediate number used in the calculation that is too large, and Excel chokes on it.

Excel 2003 and later versions will handle this calculation: the answer is approximately 0.0044, or 0.44%.

### 10.7 Histograms in Excel

Excel’s Data Analysis package, located in the Tools menu, includes a feature that allows you to sort data for histograms.

The histogram tool takes three inputs:

- **Input Range:** the data to be sorted. Enter the range that contains your data.
- **Bin Range:** a range containing numbers that define the bins. We’ll discuss this after this list of inputs.
- **Where the output goes:** in a range of cells on the current worksheet (in which case you specify the upper left corner of the range); in a new worksheet in the current workbook; or in a new workbook.
The concept of bins isn’t hard to understand. When you make a histogram, you sort the data according to which of several ranges each data point fall into. These ranges are called bins. For example, suppose you’ve got a database with the annual income for a number of people. You might want to sort it by income categories: under $20,000; between $20,000 and $30,000; between $30,000 and $40,000; and more than $40,000. Each of these categories is a bin.

When you specify the bins in Excel, you should list the upper ends of the bins in a column. The bin limits should be listed in increasing order. For our example of income data, the bin range should be a column of cells that looks like:

| 20,000 |
| 30,000 |
| 40,000 |

Excel will count the number of data entries for which \( X \leq 20,000 \); for which \( 20,000 < X \leq 30,000 \); for which \( 30,000 < X \leq 40,000 \); and for which \( 40,000 < X \).

In interval notation, these bins are \((-\infty, 20,000]; (20,000, 30,000]; (30,000, 40,000]; \) and \((40,000, \infty)\).

Strictly, you don’t have to specify the bin limits. If you don’t, Excel will pick limits for you. However, Excel’s judgement in such matters is not always good. Don’t be lazy: specify the bins yourself, and get it right.

In the figure below is an example of how Excel uses bin numbers to partition the real line.

![Figure 55: An example of how Excel uses the bin numbers \( \{x_1, x_2, x_3\} \) to partition the real axis.](image)

You can loosely imagine the histogram function as a mail person delivering mail to a bunch of people in an office building. The input (i.e., data to be sorted) is the incoming mail; the bin numbers are the address where the mail is going (whose box is the mail going to); and the output as the number of letters that end up in each person’s mailbox.
Example 24. The following tables give grade data for 27 students taking an exam; and bins into which to sort them using Excel.

We want the bins to represent A’s, B’s, C’s, D’s, and F’s. Remember that in Excel, if \( x_1 < x_2 \) are the limits of two consecutive bins, then a data point goes into the bin between them if and only if \( x_1 < X \leq x_2 \). We want everyone who gets at least 70 points, but less than 80, to count as a C; so we have to specify 69.5 and 79.5 points as the right-end points for the bin limits—otherwise, somebody who got exactly 70 points would go in the lower bin and be counted as a D instead of a C. The use of the 100.5 as the last right end point instead of 100 is for technical reasons with certain versions of Excel.

We’ve arranged the data in order, so that it’ll be easier for you to count how many A’s, B’s, etc. there are. In reality, the points would probably be listed in order of students’ last names and not in neat ascending order. Although bin limits have to be in ascending order for Excel’s histogram tool, data points can be in any order.

If you go through and count, you’ll find 11 F’s, 8 D’s, 2 C’s, 3 B’s, and 2 A’s. We’ll refer to these respectively as bins 1, 2, 3, 4, and 5.

If we assume that our data are representative of all students who might take this exam, we can use these numbers to estimate the probabilities of getting the various letter grades. We’ll define the sample space \( \Omega \) to be the 27 data points we have; and we’ll let \( X \) be the score of a random student selected from among them. Then—
\[
P(A) = P(90 \leq X \leq 100) = \frac{\#(\text{Bin 5})}{\#(\Omega)} = \frac{2}{27}
\]
\[
P(B) = P(80 \leq X \leq 89) = \frac{\#(\text{Bin 4})}{\#(\Omega)} = \frac{3}{27}
\]
\[
P(C) = P(70 \leq X \leq 79) = \frac{\#(\text{Bin 3})}{\#(\Omega)} = \frac{2}{27}
\]
\[
P(D) = P(60 \leq X \leq 69) = \frac{\#(\text{Bin 2})}{\#(\Omega)} = \frac{9}{27}
\]
\[
P(F) = P(0 \leq X \leq 59) = \frac{\#(\text{Bin 1})}{\#(\Omega)} = \frac{11}{27}
\]

An important skill that every one should know is how to compute an expected value from a set of data. Excel can be very useful in this way. Although we don’t pursue it at this point, below is an example of computing the expected value from a set of data.

**Example 25** (Random Sampling). The following data is a list of the starting salaries for high school teachers with masters degrees in their respective subject: \{22,900, 31,600, 28,200, 28,500, 34,000, 30,000, 25,900, 26,400, 27,100, 29,700\}. Given this data, compute the expected salary for a new high school teacher with a masters degree.

**Solution:**

Here \(S\) is a random variable giving the starting salary of a randomly selected teacher.

\[
\mu_S \approx \frac{1}{10} (31,600 + 22,900 + 28,200 + 28,500 + 34,000 + 30,000 + 25,900 + 26,400 + 27,100 + 29,700) = 28,430.
\]

Therefore, a new high school teacher should expect a starting salary of $28,430.
11 Continuous random variables

11.1 Introduction to the concept of a continuous R.V.

11.1.1 Motivation

Consider the spinner shown in the figure. The pointer is free to spin when you flick it with your finger; and it is equally likely to stop with its tip at any point on the circle.

Let $X$ be the position of the tip of the pointer when it stops. Then $X$ is a random variable that is equally likely to take any value between 0 and $2\pi$ radians. We say that the range of $X$ is $[0, 2\pi)$. This range is not discrete, but continuous: it includes every real number between 0 and $2\pi$, including fractions and irrational numbers. A random variable like this $X$ is called a continuous uniform random variable.

We want to know the probability that $X$ takes a value within a certain range: say, between $a$ and $b$. That is, we want to know $P(a \leq X \leq b)$.

To begin with, let’s look at $P(0 \leq X \leq 2\pi)$. We know that the pointer has to stop somewhere on the circle, so $X$ is always between 0 and $2\pi$. Hence $P(0 \leq X \leq 2\pi) = 1$.

Now, what is the probability that the pointer stops with its tip in the top half of the circle: $P(0 \leq X \leq \pi)$? Remember that any point on the circle is equally likely; so the probability that the pointer stops on some arc $A$ is proportional to the length of the arc. The total sample space is $\Omega = [0, \pi)$; if we let $A = [0, \pi]$, we get

$$P(A) = \frac{\text{length of } A}{\text{length of } \Omega} = \frac{\pi}{2\pi} = \frac{1}{2}$$

This accords with our intuition: the probability that the pointer stops in the top half of the circle should equal the probability that it stops in the bottom half, and those two events partition the sample space.

Compare this with the discrete case. If $A$ is some event, and if every outcome is equally likely, then the probability of $A$ is

$$P(A) = \frac{\#(A)}{\#(\Omega)}$$

In this case, we used counting to calculate the probability of $A$. In the case of the continuous uniform random variable, we use the length of the interval: since every interval contains infinitely many points, we obviously can’t count them.
In general, if $0 \leq \theta_1 \leq \theta_2 < 2\pi$, then
\[
P(\theta_1 \leq X \leq \theta_2) = \frac{\text{length of the interval } [\theta_1, \theta_2]}{\text{length of } \Omega} = \frac{\theta_2 - \theta_1}{2\pi}
\]
Let’s test this formula. First, we’ll consider the case: $\theta_1 = 0$ and $\theta_2 = 2\pi$. We get
\[
P(0 \leq X \leq 2\pi) = \frac{2\pi - 0}{2\pi} = \frac{2\pi}{2\pi} = 1
\]
This is what we expect: the pointer is certain to stop somewhere on the circle.

Now we’ll look at the probability that the pointer stops on the top half of the circle:
\[
P(0 \leq X \leq \pi) = \frac{\pi - 0}{2\pi} = \frac{\pi}{2\pi} = \frac{1}{2}
\]
This is the answer we got earlier, and it accords with our intuition about the situation.

Let’s look at a few more probabilities:
\[
P\left(0 \leq X \leq \frac{\pi}{2}\right) = \frac{\pi/2 - 0}{2\pi} = \frac{1}{4}
\]
\[
P\left(0 \leq X \leq \frac{\pi}{4}\right) = \frac{\pi/4 - 0}{2\pi} = \frac{1}{8}
\]
\[
P\left(0 \leq X \leq \frac{\pi}{6}\right) = \frac{\pi/6 - 0}{2\pi} = \frac{1}{12}
\]
\[
P\left(0 \leq X \leq \frac{\pi}{100}\right) = \frac{\pi/100 - 0}{2\pi} = \frac{1}{200}
\]
Notice that as $\theta_2 - \theta_1$ gets smaller, the probability that $X$ lies between $\theta_1$ and $\theta_2$ gets smaller as well. As $\theta_2 - \theta_1 \to 0$, we see $P(\theta_1 \leq X \leq \theta_2) \to 0$.

Now, let $\theta$ be any number in the interval $[0, 2\pi)$, and let $\Delta \theta$ be a small number. Then
\[
P(\theta \leq X \leq \theta + \Delta \theta) = \frac{\Delta \theta}{2\pi}
\]
We can make this probability arbitrarily small, just by making $\Delta \theta$ small enough. In particular, notice that
\[
P(X = \theta) = P(\theta \leq X \leq \theta) = \frac{\theta - \theta}{2\pi} = 0
\]
This is a strange result. It tells us that while $X$ has to assume some value between $0$ and $2\pi$ radians, the probability that it will assume any particular value is exactly zero!

When we think about it in a little more detail, it makes sense. Remember that every point in the range of the random variable has an equal probability of being chosen. If
there were a finite number $n$ of values, then the probability that the random variable would assume any one of the values would be $1/n$. But there are infinitely many points in the range. For any positive number $\epsilon > 0$, we can come up with a positive integer $N$ such that $1/N < \epsilon$; and there are more than $N$ points in the range, so the probability that the random variable assumes a particular value is less than $1/N$.

Our spinner gives one example of a continuous random variable. Another example is temperature. Let $T$ be the temperature at noon, measured to arbitrary accuracy. Clearly, $T$ can take on any value in a continuum. For example, there is no reason why $T$ couldn’t be a number like $66.257813990541\ldots$

The purpose of this section is to serve as a motivation for the concept of a continuous random variable. We now proceed to define a continuous R.V. more carefully.

### 11.1.2 A quick review of intervals on the real line

Before we go any further we give a quick summary of how we express intervals on the real line. This should already be familiar to you!

Intervals can take four general forms, depending on whether or not they include their endpoints. The table below shows the four forms. Note that a square bracket indicates that the endpoint is included in the interval; a curved parenthesis indicates that it’s not included. (Since $-\infty$ and $\infty$ aren’t real numbers, we always use a parenthesis for one of them.)

<table>
<thead>
<tr>
<th>Interval</th>
<th>Mathematical Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[a, b]$</td>
<td>$a \leq x \leq b$</td>
<td>$a \leq x \leq b$</td>
</tr>
<tr>
<td>$(a, b)$</td>
<td>$a &lt; x &lt; b$</td>
<td>Open interval</td>
</tr>
<tr>
<td>$[a, b)$</td>
<td>$a &lt; x \leq b$</td>
<td>$a &lt; x \leq b$</td>
</tr>
<tr>
<td>$(a, b]$</td>
<td>$a \leq x &lt; b$</td>
<td>$a \leq x &lt; b$</td>
</tr>
</tbody>
</table>

We express the above intervals in mathematical notation as follows:

Let $a$ and $b$ be real numbers with $a < b$.

$(a, b)$ denotes the set of all real numbers $x$ with $a < x < b$. Since the end points are not included in the set $(a, b)$, this is called an *open interval*.

$[a, b]$ denotes the set of all real numbers $x$ with $a \leq x \leq b$. Since both end points are included in the set $[a, b]$, this is called a *closed interval*.
Continuous R.V.s)  Copyright ©Wayne Hacker 2013. All rights reserved.  222

(a, b] denotes the set of all real numbers x with a < x ≤ b. This is a half open interval (open on the left and closed on the right)

[a, b) denotes the set of all real numbers x with a ≤ x < b. This is a half open interval (closed on the left and open on the right)

We can extend these intervals by allowing a, or b, or both to be infinite.

(a, ∞) : all real numbers x with a < x.
[a, ∞) : all real numbers x with a ≤ x.
(−∞, a) : all real numbers x with x < a.
(−∞, a] : all real numbers x with x ≤ a.
(−∞, ∞) = ℝ (the entire real number line).

11.1.3 So what is a continuous random variable?

Recall that the range of a random variable X is the collection of all possible values that the random variable can obtain. If a random variable X has a range that is an interval, then we call that random variable continuous. That is, the collection of all possible values for the random variable forms an interval. Thus, the continuous random variable is associated with experiments that have infinitely many outcomes.

Example 1. Let T = Today’s high temperature. T is a continuous random variable that can assume any value within a continuous range of numbers.

We now give a more general definition of a continuous random variable:

Informal Definition: A continuous random variable is a random variable that can assume any one in a continuum of values.

Suppose that we are concerned with the possibility that an accident will occur on a freeway which is 100 kilometers long, and that we are interested in the probability that it will occur at a given location, or perhaps on a given stretch of road. The sample space of this “experiment” consists of a continuum of points, those on the interval 0 to 100. We shall also assume that an accident is equally likely to occur anywhere along the entire stretch of road. Thus, the probability that an accident will occur on any interval of length L is L/100. Notice that we can partition the road into N non-overlapping intervals (N disjoint intervals). Since the intervals must add up to 100, we see that the length of each interval must be 100/N kilometers long. The probability that an accident will occur on any particular interval is 1/N. As we take more and more intervals (i.e., as we let N tend to infinity) we see that the probability that an accident will occur on any particular interval goes to zero. For example, if we take the intervals to be 1 centimeter long then the probability that an accident will occur on any particular interval is 10⁻⁷. In the continuous case, N = ∞, we always assign zero probability to individual points. This does not mean that the corresponding events cannot occur. Of course, if an accident
occurs, it must occur at some point along the freeway even though each point has zero probability.

Thus, for a continuous random variable, it doesn’t make sense to assign a probability to individual outcomes. Therefore, we can’t define a probability mass function for a continuous random variable. However, from the previous example it is clear that we can define a probability to arbitrarily small intervals in the range of a random variable. That is, we can speak of the probability per unit length (the one-dimensional density).

Suppose that that $a$ and $b$ are real numbers with $a < b$. Recall that $X \leq a$ is the event that $X$ assumes a value in the interval $(-\infty, a]$. Likewise, $a < X \leq b$ and $b < X$ are the events that $X$ assumes values in $(a, b]$ and $(b, \infty)$, respectively. These three events are mutually exclusive and at least one of them must happen. Thus,

$$P(X \leq a) + P(a < X \leq b) + P(b < X) = 1.$$ 

Since we are interested in the probability that $X$ takes a value in an interval, we will solve for $P(a < X \leq b)$.

$$P(X \leq a) + P(a < X \leq b) + P(b < X) = 1 \implies P(X \leq a) + P(a < X \leq b) = 1 - P(b < X) = P(X \leq b) \implies P(a < X \leq b) = P(X \leq b) - P(X \leq a) \implies P(a < X \leq b) = F_X(b) - F_X(a)$$

Since $X$ is a continuous random variable, $P(X = a) = 0$ and $P(X = b) = 0$. Thus, it makes no difference whether or not we include the end points of an interval.

**Warning:** This is not true for discrete random variables!

**Fact:** If $X$ is a continuous random variable and $a < b$, then

$$F_X(b) - F_X(a) = P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

We now make the following useful definitions.

**Definition:** A probability density function (p.d.f.) is a function, $f_X(x)$, of a continuous random variable $X$, with the property that if $a \leq b$, then $P(a \leq X \leq b)$ is the area of the region “under” the graph of $f_X(x)$ from $a$ to $b$.

The p.d.f. can be used to compute probabilities within ranges of the random variable. The total area under the graph of a p.d.f. is always equal to 1. This is analogous to the
notion that the probabilities in a p.m.f. add up to 1.

**Definition:** A *cumulative probability distribution function* (c.d.f.) is a function, defined as

\[ F_X(x) = \mathcal{P}(X \leq x) \equiv \int_{-\infty}^{x} f_X(t) \, dt. \]

This is the same definition as the one used for a discrete distribution except that the cumulative distribution was a sum in the case of a discrete distribution. The integral is the natural generalization of the sum for the case of a continuum of values that is being summed over.

**Comment:** A p.d.f. is *not* a probability mass function (a p.m.f. applies only to *discrete* random variables). The two differ in precisely the way that mass and density differ in physics. We can compare these two expressions by treating the probability as a physical quantity of mass totalling one unit (say 1 kilogram). The p.m.f. \( f_X(x) = \mathcal{P}(X = x) \) can be interpreted as a discrete probability point mass at \( X = x \), and the p.d.f. \( f_X(x) \) can be interpreted as probability per unit length at \( X = x \). After all, where do you think the names *probability mass* and *probability density* come from?

We can demonstrate this point using dimensional analysis from physics. If we let the operator \([\ ]\) denote the “take-the-fundamental-dimensions-of” operator, then applying this operator to the definition of the cumulative distribution we find

\[
[F_X(x)] = \left[ \int_{-\infty}^{x} f_X(t) \, dt \right] = [f_X(t)] \, [dt]
\]

where \([F_X]\) has fundamental units of mass \( M \), \([f_X]\) has units of density \( M/L \), and \([dt]\) has the fundamental units of length \( L \).

In summary:

- A p.d.f. is *not* a probability mass function (a p.m.f. applies only to *finite* random variables).
- The area under the graph of a p.d.f. is always equal to 1. This is analogous to the notion that the probabilities in a p.m.f. add up to 1.
- The p.d.f. can be used to compute probabilities within ranges of the random variable.
- A p.d.f. has a cumulative distribution function associated with it. The c.d.f. is defined as

\[ F_X(x) = \mathcal{P}(X \leq x) \]
11.2 Finite vs. continuous random variables

Before we write down the formal properties of a continuous random variable, let’s re-capitulate the properties of the finite discrete random variable, so that we can compare and contrast the two.

If $X$ is a discrete random variable, then its range is $\{x_1, x_2, \ldots, x_n\}$, where $x_1 < x_2 < \cdots < x_n$.

The probability mass function for $X$ is $f_X(x) = \mathcal{P}(X = x)$. If we think of probability as mass, we can think of this as expressing the mass at every point $x$. (If a point $x$ is not in the range of $X$, then there is no mass located at $x$, so $\mathcal{P}(X = x) = 0$.)

The cumulative distribution function of $X$ is $F_X(x) = \mathcal{P}(X \leq x)$. This is the total accumulation of all the probability mass along the number line from $-\infty$ up to and including $x$. Since nonzero mass is only found at the discrete points $x_1, \ldots, x_n$, we can write

$$F_X(x) = \mathcal{P}(X \leq x) = \sum_{t \leq x} f_X(t) = \sum_{k=1}^{i} f_X(x_k)$$

where $x_i \leq x < x_{i+1}$. Here, we’ve summed up the masses $f_X(x_k)$ for all the points $x_1, \ldots, x_i$, where $x_i$ is the largest value in the range of $X$ that’s less than or equal to $x$.

Now, suppose that $X$ is a continuous random variable instead. Then the range of $X$ is an interval $I$ that is a subset of the real number line. This interval $I$ could be very small, for example $I = [0, 0.00010]$; or it could be as large as the entire real number line $(-\infty, \infty)$.

The first thing to notice is the difference in the range of the two types of random variables.

**Comparing the ranges: discrete vs. continuous**

If $X$ is a finite random variable, the range is a discrete set of the form $\{x_1, \ldots, x_N\}$, with $x_1 < \cdots < x_N$.

If $X$ is a continuous random variable, then the range is an interval $[a, b]$ or $(a, b)$. The endpoints $a$ and $b$ could be real numbers; or either or both could be infinite.

The next thing we need to examine is the difference between how we represent the probability of the event $X = x$. 
Comparing the probability of the event $X = x$

For a finite random variable $X$, the probability mass function is

$$f_X(x) = \mathcal{P}(X = x).$$

If $X$ is a continuous random variable, this wouldn’t make any sense: for every $x$ in the range of $X$, $\mathcal{P}(X = x) = 0$. If we defined $f_X(x)$ the way we did for finite random variables (i.e., $f_X(x) = \mathcal{P}(X = x)$), the function would be zero everywhere. This wouldn’t give us any useful information.

Instead of defining a probability mass function for continuous random variables, we defined a probability density function, which we also denoted $f_X(x)$. First, we showed that for $a, b \in \mathbb{R}$,

$$\mathcal{P}(a \leq X \leq b) = F_X(b) - F_X(a),$$

where the cumulative distribution

$$F_X(x) = \mathcal{P}(X \leq x)$$

is defined the same way as it is for a finite random variable. We then defined the probability density function indirectly by saying that

$$\mathcal{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \text{area under the graph of } f_X(x) \text{ from } a \text{ to } b.$$

In general, for any continuous random variable $X$ with a probability density function $f_X(x)$, if $f_X(x_1) > f_X(x_2)$, then the probability of $X$ taking on a value near $x_1$ is greater than the probability of $X$ taking on a value near $x_2$.

Using the formula

$$\mathcal{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \text{area under graph of } f_X(x) \text{ from } a \text{ to } b$$

we can relate $F_X$ and $f_X$. Consider a very narrow interval $(a, b)$, where $b = a + \Delta x$ and $\Delta x$ is a very small number. The picture shows a segment of the graph of $f_X(x)$:
Since $\Delta x$ is very small, the difference between $f_X(a)$ and $f_X(b)$ is also very small. Thus the area under the graph of $f_X(x)$ from $a$ to $b$ is approximately that of the rectangle with height $f_X(a)$ and width $b - a = \Delta x$:

$$F_X(a + \Delta x) - F_X(a) \approx f_X(a) \Delta x.$$  \hspace{1cm} (11.1)

The error in this approximation is very small—in the figure, it’s the area of the small region marked “error.” (It looks like a narrow triangle, but the top is actually a very small segment of a curve.)

If we divide Equation (11.1) by $\Delta x > 0$, we get

$$\frac{F_X(a + \Delta x) - F_X(a)}{\Delta x} \approx f_X(a).$$  \hspace{1cm} (11.2)

If we take the limit as $\Delta x \to 0$, then the error that results from approximating the area under the curve as a rectangle goes to zero. The limit of the left-hand side of Equation (11.2) is the slope of the tangent line to the curve of $F_X(x)$ at $x = a$. It is the limit of the secant line shown in the picture.

![Figure 56: An example of a secant line for the curve $x = x(t)$.](image-url)

If you’ve had calculus, then the above argument should be familiar to you. It’s the fundamental theorem of calculus. Having now motivated the idea behind how we think of the relationship between the probability density function $f_X(x)$, we give a formal definition of the probability density function. It should be pointed out that we could have simply started out with the formal definition given below, but this would have been completely unmotivated.
Definition 18. Let $X$ be a continuous random variable. The probability density function $f_X(x)$ has the property that for every pair of real numbers $a$ and $b$, with $a < b$, the area under the graph of $f_X(x)$ between $a$ and $b$ is equal to the probability $\mathcal{P}(a \leq X \leq b)$. The graph below illustrates this: the shaded area between the curve and the $x$-axis, and between vertical lines extending upward from $a$ and $b$, represents $\mathcal{P}(a \leq X \leq b)$.

![Graph illustrating the definition of probability density function](image)

Figure 57: The probability $\mathcal{P}(a < X < b)$ is the area under the probability distribution $f_X(x)$.

Notice that $f_X(x) \geq 0$ for every real $x$. Notice also that the total area under the graph of $f_X(x)$ from $-\infty$ to $\infty$ must equal 1, since $X$ must assume a value between $-\infty$ and $\infty$.

Let’s consider our model of probability as mass. For a finite random variable $X$, we can think of the probability mass as contained in a finite number of beads, possibly of different masses. Each bead is located at exactly one point $x_i$ along the number line. For each of these points $x_i$, there is a positive probability that $X = x_i$, and that probability is proportional to the mass of the bead at $x_i$. For any point $x$ that’s not in the set of $x_i$’s, the probability that $X = x$ is zero.

If $X$ is a continuous random variable, we need a different kind of physical model. Since we’re in one spacial dimension and we’re dealing with the concept of probability density, the natural physical analogy is a one-dimensional wire of linear density $\lambda(x)$ that can vary with $x$.

Imagine that we take a thin metal wire of varying density$^6$ and stretch it along the $x$-axis. The density function $\lambda(x)$, if plotted on the $xy$-axes, would have the shape of the curve $y = f_X(x)$ (the probability density function).

$^6$Alternatively, we could imagine a homogeneous sheet of metal that has been cut it into the shape of the graph of $f_X(x)$ above. The bottom edge of the piece of metal (corresponding to the $x$-axis) is straight; the top edge follows the curve of the graph. The whole piece of sheet metal has a mass of 1. To find the probability that $X$ is between two values $a$ and $b$, we'd cut out the strip of metal lying between $a$ and $b$ on the $x$-axis, and weigh that piece.
Now we need a method to determine the probability mass from our probability density function. The physical analogy is we want the mass of the wire between two points: $a$ and $b$ on the $x$-axis, where $a$ can be $-\infty$ and $b$ can be $\infty$. How can we find it? In this case since we have a continuous distribution of wire along the $x$-axis, so we could cut the wire, roll it up, weigh it, and divide by the gravity $g$ (Recall: weight = mass $\times$ the acceleration due to gravity $g$) to find the mass $m$. This procedure is exactly what you do when you integrate the density function:

$$m = \int_a^b \lambda(x) \, dx.$$ 

The probability analogy is

$$P(a \leq X < b) = \int_a^b f_X(x) \, dx.$$ 

You can also see how this leads to the conclusion that for any $x$, $P(X = x)$ is zero for a continuous random variable. If you tried to cut out a piece of wire that included only $x$, that piece would have to have zero length; and a strip with zero length must have zero weight since it could even contain one molecule, atom, or even electron!

For a finite random variable, the units of $f_X(x) = P(X = x)$ were mass. For the continuous random variable, the units of $f_X(x)$ are density = mass/length. We’ll now try to show the relation between the probability density function $f_X(x)$, and the probability $P(a \leq X \leq b)$.

Suppose we divide the interval between $a$ and $b$ into $n$ pieces of equal width $\Delta x = \frac{b-a}{n}$. Each slice lies between $x_i$ and $x_{i+1}$, where $x_k = a + k\Delta x$; so we have $x_0 = a$, $x_1 = a + \Delta x$, $x_2 = a + 2\Delta x, \ldots, x_n = b$. If we make $n$ large enough (so that $\Delta x$ is very small, and the slices are very narrow), then $f_X(x)$ is nearly constant over the interval $[x_i, x_{i+1}]$. Over each such interval, we’ll approximate $f_X(x)$ by its value at the midpoint $m_{i+1}$ of the interval. That is,

$$f_X(x) \approx f_X(m_i) \quad \text{over the interval} \quad [x_{i-1}, x_i].$$

We can approximate the probability that $X$ will take a value between $x_{i-1}$ and $x_i$ by the area of the rectangle with width $\Delta x$ and height $f_X(m_i)$:

$$P(x_{i-1} \leq X \leq x_i) \approx f_X(m_i) \cdot \Delta x.$$ 

We can use this to approximate the probability that $X$ will be between $a$ and $b$, by summing up the areas of all the rectangles between $a$ and $b$ (also see the figure below):

$$P(a \leq X \leq b) = \sum_{i=1}^n P(x_{i-1} \leq X \leq x_i) \approx \sum_{i=1}^n f_X(m_i) \cdot \Delta x.$$
Notice that \( P(a \leq X \leq b) \) now has the right units. The units of \( f_X(x) \) were density, which is mass/length. The width of each rectangle \( \Delta x \) is a measure of length; so \( f_X(x) \cdot \Delta x \) has units of \[
\frac{\text{mass}}{\text{length}} \cdot \text{length} = \text{mass}.
\]

Figure 58: Use rectangles of height \( f_X(m_i) \) and width \( \Delta x \) to approximate area.

**Definition 19.** Let \( X \) be a continuous random variable. The *cumulative distribution function* for \( X \) is

\[
F_X(x) = P(X \leq x).
\]

Notice that this is identical with Definition 10 on page 167, which defined the cumulative distribution function for a finite random variable. The difference is that we can’t add up the probabilities of the points less than or equal to \( x \), since there are infinitely many of those in the case of a continuous random variable.

There is a relationship between \( f_X(x) \) and \( F_X(x) \) for continuous random variables. Remember that the probability that \( X \) takes a value between \( a \) and \( b \) is proportional to the area under the graph of \( f_X(x) \) between \( a \) and \( b \). In our wire-of-varying-density example, it’s proportional to the mass of the section of the wire between \( a \) and \( b \). \( F_X(x) \) is the probability that \( X \) takes a value less than \( x \): that is, between \(-\infty\) and \( x \). If we cut the wire at \( x \), then mass of the piece of wire to the left of the cut is \( F_X(x) \).

Don’t feel bad if you’re having a lot of trouble understanding \( f_X(x) \). We’ve tried to give you an idea of the connection between it and \( F_X(x) \), but it really requires second-semester calculus to go back and forth between the two. For that reason, we won’t do much with \( f_X(x) \). The cumulative distribution function \( F_X(x) \) is easier to understand, so we’ll use that for the most part. Since the derivation of the formulas for \( F_X(x) \) will often be beyond your current mathematical ability, we’ll generally just tell you what they are and expect you to trust us.
11.3 Continuous random variables

We now give a short introduction to the theory behind the continuous random variable. In this section we assume the reader knows calculus. If you don’t, just skim the material and don’t try to understand everything.

11.3.1 The cumulative distribution and the probability density connection

We now list a summary of the connection between the cumulative distribution and the probability density:

\[ F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt. \]

By the fundamental theorem of calculus,

\[ f_X(x) = F_X'(x). \]

The probability that \( X \) is between \( a \) and \( b \) is

\[ P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_{a}^{b} f_X(x) \, dx. \]

This is the limit of a Riemann sum; so if we divide the interval from \( a \) to \( b \) into \( N \) equal segments with width \( \Delta x \) and midpoints \( m_1, \ldots, m_N \), we can approximate

\[ P(a \leq X \leq b) \approx \sum_{i=1}^{N} f_X(m_i) \Delta x. \]

11.3.2 Computing with continuous distributions

Now, let’s find out what we can do with \( F_X(x) \). We’ll assume that we’ve already got a formula: so for every number \( x \) in the range of \( X \), we can determine the value of \( F_X(x) \). We want to use our formula to calculate probabilities like \( P(a < X) \), or \( P(X < b) \), or \( P(a \leq X \leq b) \).

The first step is to remember that \( F_X(x) = P(X \leq x) \). Second, remember that the random variable \( X \) must take some value; so for every \( x \),

\[ 1 = P(-\infty < X < \infty) = P(X \leq x) + P(X > x), \]

where we’ve use the fact that the sets: \( \{X \leq x\} \) and \( \{X > x\} \) partition the sample space.
The first equality tells us that the random variable is certain to take some real value. The second follows from the fact that $X \leq x$ and $X > x$ partition the number line: if $X$ is going to take a real value, then that value is either less than or equal to $x$, or else greater than $x$, but not both. We can see this in three different ways. First, in the form of inequalities:

$$-\infty < X < \infty = (X \leq x) \cup (X > x).$$

Second, in the form of intervals:

$$(-\infty, \infty) = (-\infty, x] \cup (x, \infty).$$

Third, on the number line:

\begin{itemize}
  \item Break $x$-axis into two disjoint intervals
  \item $(-\infty, x] \cup (x, \infty)$
\end{itemize}

We’ve seen that $\mathcal{P}(X \leq x) + \mathcal{P}(X > x) = 1$. Rearranging this equation gives us

$$\mathcal{P}(X > x) = 1 - \mathcal{P}(X \leq x) = 1 - F_X(x).$$

Sometimes this is called the \textit{survival function}:

$$S_X(x) \equiv \mathcal{P}(X > x) = 1 - F_X(x).$$

You can see where the name comes from if you think of $X$ as a random variable representing the lifetime of something—a person or a light bulb or an atomic nucleus, or anything else that might die or fail or disintegrate. Then the survival function $S_X(x)$ represents the probability that the thing is still “alive” at time $x$.

There are four other kinds of probabilities that we’d like to be able to calculate from $F_X(x)$. We’d like to be able to determine:

$$\mathcal{P}(a \leq X \leq b),$$
$$\mathcal{P}(a \leq X < b),$$
$$\mathcal{P}(a < X \leq b),$$
$$\mathcal{P}(a < X < b).$$

Luckily, if we know one of these, then we know the other three—because for a continuous random variable $X$, they’re all the same. To prove this, we need two facts. First, for any $x$, we know that $\mathcal{P}(X = x) = 0$. (This is only true for continuous random variables.)
For a discrete random variable, there has to be at least one \( x \) for which \( P(X = x) \neq 0 \).

Second, notice that
\[
a \leq X \leq b = (X = a) \cup (a < X < b) \cup (X = b)
\]
or equivalently,
\[
[a, b] = \{a\} \cup (a, b) \cup \{b\}.
\]
We’ve also shown this on the number line below.

\[
\bullet \quad \circ \quad \circ \quad \bullet
\]

\( a \quad \circ \quad X \quad \bullet \quad b \)

Since this is true, we can write
\[
P(a \leq X \leq b) = P(X = a) + P(a < X < b) + P(X = b) = 0 + P(a < X < b) + 0 = P(a < X < b).
\]

We can vary this approach slightly to show that these two are also equal to \( P(a \leq X < b) \) and to \( P(a < X \leq b) \).

We’ll work with the last of these probabilities. Starting with
\[
F_X(x) = P(X \leq x)
\]
we’ll calculate
\[
P(a < X \leq b).
\]
First, notice that if \( a < b \), the set \( X \leq b \) can be partitioned:
\[
X \leq b = (X \leq a) \cup (a < X \leq b)
\]
or equivalently,
\[
(−\infty, b] = (−\infty, a] \cup (a, b].
\]
Since the sets \((−\infty, a]\) and \((a, b]\) are disjoint, we can write
\[
P(X \leq b) = P(X \leq a) + P(a < X \leq b).
\]
Since \( F_X(x) = P(X \leq x) \), we can rewrite this
\[
F_X(b) = F_X(a) + P(a < X \leq b)
\]
and we can rearrange this as

$$
P(a < X \leq b) = F_X(b) - F_X(a).$$

Alternatively, we can show this using our graph of $f_X(x)$. Remember that $P(a < X \leq b)$ is the area under the curve of the graph between vertical lines drawn at $x = a$ and at $x = b$. In our wire-of-varying-density example, it’s the mass of the piece of wire between $a$ and $b$. From the figures below, you can see that this is the piece to the left of $b$, but to the right of $a$. If we wanted to determine the mass of this piece, we’d make a vertical cut at $b$ and keep the piece to the left; then we’d make a second cut at $a$ and throw away the piece to the left of that. The piece we’d have after our first cut at $b$ has weight $P(X \leq b) = F_X(b)$. The piece that we throw away after making the second cut at $a$ has weight $P(a \leq X) = F_X(a)$. The piece that we still have after we’ve thrown that out is

$$
P(a < X \leq b) = F_X(b) - F_X(a).$$

Figure 59: A visual of how we represent the probability $P(a < X < b)$ in terms of the cumulative distribution function $F_X(x)$.

Let’s briefly summarize our formulas for working with continuous random variables. We’ll let $X$ be our continuous random variable, and assume that $a \leq b$. Then

- $F_X(x) = P(X \leq x)$ (cumulative distribution function)
- $S_X(x) = P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)$ (survival function)
- $F_X(b) - F_X(a) = P(a < X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a \leq X \leq b)$
Our next step is to get the formulas for $F_X(x)$ for various continuous distribution functions.

### 11.4 The Uniform distribution

**Definition 20.** A random variable $X$ has a **uniform density**, and is referred to as a continuous uniform random variable, if its range is $[a, b]$ and if its probability density is

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases} \quad (11.3)$$

The midpoint $\mu_X = \frac{a+b}{2}$ is $E(X)$, the expected value of $X$.

The region under $f_X(x)$ is a rectangle of width $(b-a)$ and height $\frac{1}{b-a}$. The area is $(b-a) \cdot \frac{1}{b-a} = 1$.

If $a \leq x \leq b$, then the region under the graph of $f_X(x)$ between $a$ and $x$ is a rectangle whose base is $x-a$ and whose height is $\frac{1}{b-a}$; so its area is $\frac{x-a}{b-a}$.

The cumulative distribution function for this continuous uniform random variable is

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases} \quad (11.4)$$
By similar triangles,

\[
\frac{h(x)}{x-a} = \frac{1}{b-a} \quad \Rightarrow \quad h(x) = \frac{x-a}{b-a}.
\]

11.4.1 Properties of the continuous uniform random variable

Let \( c, d \in [a, b] \) with \( c < d \). Let \( X \) be a continuous uniform random variable on \([a,b]\). Then

\[
\mathcal{P}(c \leq X \leq d) = F_X(d) - F_X(c) = \frac{d - a}{b - a} - \frac{c - a}{b - a} = \frac{d - c}{b - a}.
\]

This is the area of the rectangle under the graph of \( f_X(x) \) between \( x = c \) and \( x = d \):

One consequence of this result is that every subinterval \([c,d] \subseteq [a,b]\) having the same length \(d - c\) will have the same probability. In fact, we could take our definition of a continuous uniform random variable to be a random variable defined over an interval \([a,b]\) such that every subinterval of \([a,b]\) having the same length has the same probability.
To see this, let \(c_1, d_1 \in [a, b]\) and \(c_2, d_2 \in [a, b]\) with \(d_1 - c_1 = L = d_2 - c_2\) and \(c_1 \neq c_2\). Then

\[
P(c_1 \leq X \leq d_1) = \frac{d_1 - c_1}{b - a} = \frac{L}{b - a} = \frac{d_2 - c_2}{b - a} = P(c_2 \leq X \leq d_2).
\] (11.6)

Thus the probability \(P(c \leq X \leq d)\) varies directly with the length of the interval. The location of the interval within \([a, b]\) is unimportant. This means that \(X\) is equally likely to assume any value in \([a, b]\).

Notice that

\[
P(-\infty < X < \infty) = P(a \leq X \leq b) = \frac{b - a}{b - a} = 1.
\]

This tells us that the total area under the curve of \(f_X(x)\) is 1. Put another way, this says that \(P(X \leq b) = F_X(b) = 1\).

**Example 2.** In certain experiments, the error made in determining the density of a substance is a continuous uniform random variable \(X\) with \(b = -a = .015\).

Find the probability that

(a) The error will be between 0 and .01;
(b) The error will be greater than .015;
(c) The error will be between .01 and .02;
(d) The size of the error will exceed .005.

**Solutions:**

(a) We want \(P(0 < X < .01)\). We’ll use Equation (11.5) with \(a = -.015\) and \(b = .015\).

\[
P(0 < X < .01) = \frac{.01 - 0}{.015 - (-.015)} = \frac{.01}{.03} = \frac{1}{3}.
\]

(b) \(P(.015 < X) = 0\). We can see this from the graph. We could also see it by writing

\[
P(.015 < X) = 1 - P(X \leq .015) = 1 - F_X(.015) = 1 - 1 = 0.
\]

(c) Be careful about plugging .02 into Equation (11.5)! We can’t just use that formula, because \(.02 > .015\). It might be safer to break the interval \((.01, .02)\) into two parts:

\[
(.01, .02) = (.01, .015) \cup (.015, .02),
\]

so

\[
P(.01 < X < .02) = P(.01 < X \leq .015) + P(.015 < X < .02).
\]

From the graph, we can see that \(P(.015 < X < .02) = 0\). Now we can use Equation (11.5) for the nonzero part:

\[
P(.01 < X \leq .015) = \frac{.015 - .01}{.03} = \frac{.005}{.03} = \frac{1}{6}.
\]

We could also have used

\[
P(.01 < X < .02) = F_X(.02) - F_X(.01)
\]
and calculated the values of $F_X$ using Equation (11.4):

$$F_X(.02) - F_X(.01) = 1 - \frac{.01 - (-.015)}{.03} = 1 - \frac{.025}{.03} = 1 - \frac{5}{6} = \frac{1}{6}.$$  

(d) When we speak of the size of an error, we refer to the absolute value. We need to compute

$$P(.005 < |X|) = P(X < -.005) + P(.005 < X).$$

In the case of continuous uniform random variables, $P(X < x) = P(X \leq x)$. Thus

$$P(X < -.005) = P(X \leq -.005) = F_X(-.005) = \frac{-.005 + .015}{.03} = \frac{1}{3}.$$  

The intervals $(-.015, -.005)$ and $(.005, .015)$ are of the same length. As we showed in Equation (11.6), this means

$$P(X < -.005) = P(.005 < X) = \frac{1}{3}.$$  

Adding the probabilities for the two intervals gives us

$$P(.005 < |X|) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}.$$  

Example 3. A moving 100 meter-long walkway uses a conveyor belt to carry people through an airport terminal. The has a tendency to break at random points along the conveyor belt. Let $X$ be the distance from the starting point where the belt breaks.

(a) Determine the probability that the belt breaks between the 40 and 80 meter mark from the starting point.

(b) Determine the probability that the belt breaks in the first 20 meters from the starting point.

Solution:

Using $a = 0$ and $b = 100$ in Equations (11.3) for $f_X(x)$ and (11.4) for $F_X(x)$ gives us

$$f_X(x) = \begin{cases} 
\frac{1}{100} & \text{if } 0 \leq x \leq 100 \\
0 & \text{if } x \notin [0, 100] 
\end{cases}$$

$$F_X(x) = \begin{cases} 
0 & \text{if } x < 0 \\
\frac{x}{100} & \text{if } 0 \leq x \leq 100 \\
1 & \text{if } x > 100 
\end{cases}$$
Part (a): We want the probability that $40 \leq X \leq 80$. We’ll use the formula for $F_X(x)$:

$$
P(40 \leq X \leq 80) = F_X(80) - F_X(40) = \frac{80 - 40}{100} = \frac{40}{100} = 0.4.
$$

Part (b): We need the probability that the belt breaks in the first 20 meters from the starting point:

$$
P(0 \leq X \leq 20) = P(X \leq 20) = F_X(20) = \frac{20}{100} = 0.5.
$$
11.5 The Exponential distribution

Definition 21. A random variable has an exponential distribution and is referred to as an exponential random variable if its probability density is

\[
f_X(x) = \begin{cases} \frac{1}{\alpha} e^{-x/\alpha} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases}
\]

where \( \alpha > 0 \) is a fixed parameter.

An exponential random variable is a continuous random variable.

The cumulative distribution for the exponential random variable is

\[
F_X(x) = \begin{cases} 1 - e^{-x/\alpha} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}
\]

Figure 61: Exponential Distribution

Figure 62: The graph of the cumulative distribution for exponential function
To see this we simply integrate the probability density function from \( t = -\infty \) to a random point \( t = x \):

\[
F_X(x) = \int_0^x f_X(t)dt
\]

\[
= \int_0^x \frac{1}{\alpha} e^{-t/\alpha} dt \quad (f_X = 0 \text{ for } x < 0)
\]

\[
= \int_0^x e^{-t/\alpha} dt
\]

\[
= e^{-t/\alpha} \bigg|_0^x
\]

The expected value (mean value) of the exponential random variable \( X \) is

\[
E(X) = \mu_X \equiv \int_{-\infty}^{\infty} x \cdot f_X(x)dx = \int_0^{\infty} \frac{x}{\alpha} e^{-x/\alpha} dx = \alpha,
\]

where we’ve used the fact that \( \int_0^{\infty} e^{-u} du = 1. \)

The exponential distribution applies to situations where there is a waiting time until an event, or to the waiting time between consecutive occurrences of a specified event.

**Definition 22.** The **Survival Function** is defined as

\[
S_X(x) = P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)
\]

For the case of the exponential distribution function the Survival Function becomes

\[
S_X(x) = P(X > x)
\]

\[
= 1 - P(X \leq x)
\]

\[
= 1 - F_X(x)
\]

\[
= 1 - (1 - e^{-x/\alpha})
\]

\[
= e^{-x/\alpha}
\]

You will find this formula useful when working with the exponential distribution.

Next, we present a series of examples. We’ll start with a mathematical example first so that the reader can see how to work with the exponential distribution function, then we’ll move to some examples involving applications of waiting times.
Example 4. Let $X$ be an exponential random variable with $\alpha = 3.5$.

(a) Compute $f_X(2)$.
(b) Compute $F_X(2)$.

(c) What is the difference between $f_X(2)$ and $F_X(2)$? In particular, how do they relate to the probability of the random variable $X$?

(d) What is the probability that $X$ is at least 4?

(e) What is the probability that $X$ is at most 4?

Solution:

Part (a): $f_X(2) = \frac{1}{3.5} e^{-2/3.5} \approx .1613$.

Part (b): $F_X(2) = 1 - e^{-2/3.5} \approx .4353$.

Part (c): $f_X(2)$ is the probability density at $X = 2$, not the probability! That is, $f_X(2) \neq P(X = 2)$. The term $F_X(2)$ is the cumulative distribution at $X = 2$. It’s how much probability mass that we’ve accumulated in walking along the $x$-axis from $-\infty$ up to $x = 2$. However, in this case since we have a continuous distribution we can think of rolling up a continuous string of clay from $-\infty$ to $x = 2$ with a density function given by $f_X$.

Part (d): The probability that $X$ takes at most 4 is

$$P(X \leq 4) = F_X(4) = 1 - e^{-4/3.5} \approx .6811.$$ 

Part (e): The probability that $X$ takes at least 4 is

$$P(4 \leq X) = P(4 < X) = 1 - P(X \leq 4) = 1 - F_X(4)$$

$$= 1 - (1 - e^{-4/3.5}) = e^{-4/3.5} \approx .3189.$$ 

Example 5. Let $X$ be an exponential random variable with probability density function $f_X(x) = \frac{1}{4} e^{-x/4}$. Determine the expected value of $X$.

Solution:

We are given $f_X(x) = \frac{1}{4} e^{-x/4}$. We want $\mathcal{E}(X)$.

For the exponential distribution, $\mathcal{E}(X) = \alpha$. By comparing the problem’s formula for $f_X(x)$ with Equation (11.7), we see that $\alpha = 4$. Thus $\mathcal{E}(X) = 4$.

Example 6. Let $T$ be an exponential random variable with an expected value $\mathcal{E} = 5$.

(a) Find a formula for the cumulative distribution of $T$.

(b) Use the result from part (a) to determine $F_T(4)$.

Solution:

We’re given that $T$ is an exponential random variable with $\mathcal{E}(T) = 5$.

Part (a): In order to find $F_T(t)$, we need to know the parameter $\alpha$. But for an exponential
random variable, \( E(T) = \alpha \); so \( \alpha = 5 \). Hence
\[
F_T(t) = 1 - e^{-t/5}.
\]

**Part (b):** We want \( \mathcal{P}(T \leq 4) \). Using our standard formula we arrive at
\[
\mathcal{P}(T \leq 4) = F_T(4) = 1 - e^{-4/5} \approx 0.55.
\]

**Example 7.** The mileage (in thousands of miles) that a car owner gets with a certain kind of radial tire is a random variable having an exponential distribution with \( \alpha = 40 \).

Find the probability that one of these tires will last

(a) at least 20,000 miles;

(b) at most 30,000 miles.

(c) What is the average mileage that the car owner should expect for these tires?

**Note:** We are ignoring ware and tear on the tires.

**Solution:**

**Part (a):** We want \( \mathcal{P}(X \geq 20) = 1 - \mathcal{P}(X \leq 20) = 1 - F_X(20) \). By Equation \((11.8)\), \( F_X(20) = 1 - e^{-20/\alpha} \), with \( \alpha = 40 \). Hence
\[
\mathcal{P}(X \geq 20) = 1 - (1 - e^{-20/40}) = e^{-1/2} \approx .6065.
\]

**Part (b):** Here we want \( \mathcal{P}(X \leq 30) \).
\[
\mathcal{P}(X \leq 30) = F_X(30) = 1 - e^{-30/40} = 1 - e^{-3/4} \approx .5276.
\]

**Part (c):** We want \( E(X) = \alpha = 40 \). Thus the expected life of a tire is 40,000 miles.

**Example 8.** A certain kind of appliance requires repairs an average of once every two years. Assuming that the times between repairs are exponentially distributed, what is the probability that such an appliance will work for at least 3 years without needing repairs?

**Solution:** Here, we are given \( E(X) = 2 \). For the exponential distribution, \( E(X) = \alpha \), so we know that \( \alpha = 2 \). What we want is
\[
\mathcal{P}(3 \leq X) = 1 - \mathcal{P}(X < 3) = 1 - \mathcal{P}(X \leq 3) = 1 - F_X(3).
\]

We can use Equation \((11.8)\): \( F_X(3) = 1 - e^{-3/2} \). Thus
\[
\mathcal{P}(3 \leq X) = 1 - F_X(3) = 1 - (1 - e^{-3/2}) = e^{-3/2} \approx .2231.
\]

Note that in Example 7, the waiting time was the time until the tire wore out. In Example 8, the waiting time was the time between repairs.
Example 9 (waiting times). Suppose that on average 30 customers per hour arrive at a drive-through window on the local fast-food joint. Let \( X \) be the random variable that gives the arrival time (in minutes) between consecutive customers.

(a) What is the probability that the time between customers is at most 1 minute and 40 seconds?

(b) What is the probability that the time between customers is at least 3 minutes and 15 seconds?

Solution:
Since we have a waiting time problem, we’ll use an exponential random variable. The first step is to find \( \alpha \) (the average arrival time). Here we have 30 arrivals per 60 minutes, which gives an average arrival time of 60 minutes/30 arrivals = 2 (i.e., \( \alpha = 2 \)).

Part (a): We want to compute the probability that the time between the arrival of consecutive bank customers is at most 1 minute 40 seconds. Since that is 5/3 minutes, we need \( P(X \leq 5/3) \).

\[
P(X \leq 5/3) = 1 - e^{-\left(\frac{5}{3}/2\right)} = 1 - e^{-5/6} \approx .5654.
\]

Part (b): Same as part (a), but since 3 minutes 15 seconds is 3\( \frac{1}{4} \) minutes, we want

\[
P(13/4 \leq X) = 1 - P(X < 13/4) = 1 - P(X \leq 13/4) = 1 - F_X(13/4)
= 1 - \left(1 - e^{-\left(\frac{13}{4} \cdot \frac{1}{2}\right)}\right) = e^{-13/8} \approx .1969.
\]

Example 10. You work for a popular guitar-player magazine. Your job is to check the claims of various guitar components. Let \( T \) be the lifetime, in hours, of a popular, and very expensive, amplifier tube. The manufacturer claims that the tubes have an average life time of 5000 hours. You have been asked by your boss to determine the following probabilities. You are to assume an exponential distribution.

(a) What is the probability that tube fails in a time less than 3000 hours?

(b) What is the probability that tube lasts for at least 5000 hours?

(c) What is the probability that tube lasts for exactly 5000 hours?

(d) Find a time \( t_0 \) such that the probability of your card lasting at most 5000 hours is 1/2.

Solution:
We’re given that \( T \) is an exponential random variable with \( \alpha = 5000 \) (the average lifetime).

Part (a): The probability that the card fails before 3000 hours is

\[
P(T < 3000) = P(T \leq 3000) = F_T(3000) = 1 - e^{-3000/5000} = 1 - e^{-3/5} \approx .4512.
\]
Part (b): The probability that the card lasts at least 5000 hours is

\[ P(5000 < T) = 1 - P(T \leq 5000) = 1 - F_T(5000) = 1 - (1 - e^{-5000/5000}) = e^{-1} \approx 0.3679. \]

Part (c): \( P(T = 5000) = 0 \). For any continuous random variable \( X \), and for any constant value \( x \), \( P(X = x) = 0 \). Your boss is obviously testing your knowledge!

Part (d): We are asked to find \( t_0 \) such that \( P(T \leq t_0) = 1/2 \). We know that

\[ P(T \leq t_0) = F_T(t_0) = 1 - e^{-t_0/5000}. \]

Equating the two expressions yields

\[ 1 - e^{-t_0/5000} = \frac{1}{2} \quad \Rightarrow \quad e^{-t_0/5000} = \frac{1}{2}. \]

Take the natural logarithm of both sides of the equation:

\[ \ln \left( e^{-t_0/5000} \right) = \frac{-t_0}{5000} \ln e = \ln \left( \frac{1}{2} \right). \]

We’ve used the property of logarithms: \( \log_b a^x = x \log_b a \). Next, we’ll use the fact that \( \ln e = 1 \), and we’ll solve for \( t_0 \):

\[ t_0 = -5000 \ln \left( \frac{1}{2} \right) = 5000 \ln(2) \approx 3466 \text{ hours}. \]
11.6 The Normal distribution

The probability density function for a random variable $X$ with a standard normal distribution (a.k.a., bell curve distribution) is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$  

![Figure 63: Standard Normal Distribution](image)

The range of the normal random variable $X$ is all of $\mathbb{R}$, i.e. $(-\infty, \infty)$. The value of $f_X(x)$ is related to the probability of the event $X = x$. For example,

$$f_X(0) = \frac{1}{\sqrt{2\pi}} > \frac{1}{\sqrt{2\pi}} e^{-2} = f_X(2),$$

which indicates that it is more probable that $X$ will land in a small interval about 0 than in a small interval about 2. This can be seen from the graph as well. Consider the area under the graph and over the intervals of equal width about $x = 0$ and $x = 2$, denoted in the figure by the dashed lines. The intervals are $[-1, 1]$ and $[1.9, 2.1]$. Clearly, the area over the interval about 0 is greater than the area over the interval about 2.

Notice that the graph together with the area under the curve were made of a thin homogeneous material, such as a sheet of plexiglass, that the it would balance on a knife edge that ran along the line of symmetry $x = 0$ (i.e., the $y$-axis). Thus, for the standard normal bell curve we expect $\mathcal{E}(X) = 0$. If we were to shift this graph to the left or to the right along the $x$-axis by an amount $\mu$, then its center of mass would be located at $x = \mu$.

Let $X$ be a random variable with expected value $\mathcal{E}(X) = \mu_X$ and standard deviation $\sigma_X$. The general Normal distribution p.d.f. is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X}} e^{-\frac{1}{2} \frac{(x-\mu_X)^2}{\sigma_X^2}}.$$
12 A Simple Random Walk

Disclaimer: This is an elementary example of something known as a random walk. Its name derives from the fact that each step (forward or backward) is determined by a random process. In general, random walks can get extremely sophisticated and have a wide range of applications in both science and business. This example is designed to be a first exposure to this topic. It is in no way a complete account of random walks, nor is it made to be accessible to individuals with limited experience in mathematics!

Consider the following experiment: Stand on a tiled floor and mark your initial position as the origin. Flip a fair coin \( N \) times. On each flip, if the coin comes up heads, then take one step forward, else take one step backwards. After \( N \) flips what is your average position w.r.t. the origin? What is your average distance from the origin?

Solution:

Define the random variable

\[
X_i = \begin{cases} 
1 & \text{if the coin turns up heads on the } i^{th} \text{ trial} \\
-1 & \text{if the coin turns up tails on the } i^{th} \text{ trial}
\end{cases}
\]

Notice that for each \( i \), the expected value of \( X_i \) is

\[
E(X_i) = (-1) \cdot P(X_i = -1) + (1) \cdot P(X_i = 1) = (-1) \cdot (1/2) + (1) \cdot (1/2) = 0.
\]

Let \( X = \sum_{i=1}^{N} X_i = (# \text{ of heads in } N \text{ flips}) - (# \text{ of tails in } N \text{ flips}) \).

If we let \( N_H = # \text{ of heads in } N \text{ flips of a fair coin} \), then the number of tails is \( N_T = N - N_H \). It follows that \( X = N_H - N_T = 2N_H - N \). Thus, \( X \) represents your position. The average value of \( X \) is the expected value of \( X \):

\[
E(X) = E \left( \sum_{i=1}^{N} X_i \right) = \sum_{i=1}^{N} E(X_i) = \sum_{i=1}^{N} 0 = 0.
\]

This result says that, on average, you would expect to be at the origin. This seems reasonable since both the probability distribution and the random variable are symmetric w.r.t. the origin. But what if we just flip the coin once? We must take a step away from the origin and so we cannot possibly be at the origin. This is common situation in discrete random variables: the average value of a discrete random value need not be in the range. In many situations the expected value does not convey a whole lot of information. We need the spread of the random variable (i.e., the variance). A more useful measure of
the spread of the random variable is the standard deviation, which is just the average distance from the mean. By distance, we mean the root-mean-squared distance.

Before we look at the variance, let’s examine the average distance from the origin. Let \( D = |X| \). Notice that \( E(|X|) = \sum |x_i|P(X = x_i) \) is the discrete \( l_1 \) weighted measure of distance, where as the root-mean-square distance \( E(X^2) = \sum x_i^2P(X = x_i) \) is the discrete \( l_2 \) weighted measure of distance.

Since the variance is always nonnegative, and since \( E(X) = 0 \), it follows that

\[
V(X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2 = E(X^2),
\]

and

\[
V(|X|) = E[(|X| - E(|X|))^2] = E(|X|^2) - (E(|X|))^2 = E(X^2) - (E(|X|))^2 > 0,
\]

where we used the fact that \( |X|^2 = X^2 \).

It follows that \( E(X^2) > (E(|X|))^2 \). I know of no elegant expression for the average distance from the origin \( E(|X|) \), so we will compute it in Excel for a few particular cases. You can find the results and the Excel file on my website. We can derive an exact expression for the average root-mean-square distance from the origin \( E(X^2) \) using the double expectation theorem for expected value: \( E(Z) = E[E(Z|Y)] \).

Let \( D_N = |X| \) be your distance from the origin after \( N \) flips of the coin. We will compute \( E(D_N^2) \). Now \( D_1 = 1 \). If you only flip the coin once, then you must be one unit away from the origin. Now let’s say that you’ve flipped the coin \( N - 1 \) times. If you are at distance \( D_{N-1} \) from the origin, then on the \( N^{th} \) flip, you will be at \( D_N = D_{N-1} \pm 1 \) depending on the outcome of \( X_N \). Thus,

\[
D_N^2 = (D_{N-1} \pm 1)^2 = D_{N-1}^2 \pm 2D_{N-1} + 1
\]

are the two possible outcomes, each with a probability of \( 1/2 \). We can now use our theorem with \( Z = D_N^2 \) and \( Y = X_N \) together with Bayes’ Theorem to arrive at

\[
E(D_N^2) = E(D_N^2|X_N = 1) \cdot P(X_N = 1) + E(D_N^2|X_N = -1) \cdot P(X_N = -1)
\]

\[
= E(D_{N-1}^2 + 2D_{N-1} + 1) \cdot \frac{1}{2} + E(D_{N-1}^2 - 2D_{N-1} + 1) \cdot \frac{1}{2}
\]

\[
= \frac{1}{2} [E(D_{N-1}^2) + 2E(D_{N-1}) + E(1) + E(D_{N-1}^2) - 2E(D_{N-1}) + E(1)]
\]

\[
= E(D_{N-1}^2) + 1,
\]

where we’ve made use of standard properties of the expected value including \( E(1) = 1 \). This is a recursion relation with \( D_1 = 1 \). The solution is easily found to be \( E(D_N^2) = N \). It follows that the standard deviation \( \sigma_{i_2} \equiv \sqrt{E(D_N^2)} \) is

\[
\sigma_{i_2} = \sqrt{N} \neq 0.
\]
We will also do a simulation that will verify this result as well (again see my website for the Excel file). This result says that even though the expected value of $X$ is zero, the standard distance from the origin after $N$ steps will be $\sqrt{N}$. This is not obvious! How can we make sense of this apparent contradiction? It turns out that what $E(X) = 0$ is actually saying is that half the time you will wonder (drift) a distance $\sqrt{N}$ from the origin in the positive direction, and half of the time you will wonder a distance $\sqrt{N}$ from the origin in the negative direction (after $N$ steps of course). The key point is this: once you start drifting away from the origin, you tend to stay drifting away from the origin. But you have a 50-50 chance of drifting to the left of the origin, or the right of the origin.
A Formula summary

Basic set theory

De Morgan’s laws

If $A$ and $B$ are any two sets, then

1. $(A \cup B)^c = A^c \cap B^c$
2. $(A \cap B)^c = A^c \cup B^c$.

Distributive laws:

\begin{align*}
(i) & \quad C \cap (A \cup B) = (C \cap A) \cup (C \cap B) \\
(ii) & \quad C \cup (A \cap B) = (C \cup A) \cap (C \cup B)
\end{align*}

Events and outcomes

Definition: The set of all possible outcomes of an experiment is called the sample space for the experiment. The outcomes in the sample space are called the sample points.

Definition: An event is a subset of the sample space of an experiment. An event $E$ is said to occur if the outcome of the experiment is an element of $E$.

Definition: Two events $E$ and $F$ are mutually exclusive (disjoint) if $E \cap F = \emptyset$.

Definition: The events $E_1, \ldots, E_n$ are mutually exclusive if no two of the events can occur at the same time.

Definition: Two events $E$ and $F$ are independent if

$$P(E \cap F) = P(E)P(F)$$

Definition: Let $\Omega$ be the sample space for an experiment. Let

$$E_1, E_2, \ldots, E_n \subseteq \Omega$$
be events. Then the collection of the $E_i$’s is a **partition** of $\Omega$ if the following two conditions are both met:

1. The sets are pairwise disjoint:
   
   For all $i \neq j$, $E_i \cap E_j = \emptyset$

2. The union of the $E_i$’s is the whole sample space:

   $$\bigcup_{i=1}^{n} E_i = E_1 \cup E_2 \cup \cdots \cup E_n = \Omega$$

In other words, the $E_i$’s are a partition if every outcome of the experiment is a member of exactly one $E_i$.

**Converting words to set notation:**

1. **Not** event $A$: $A^c$.

2. Events $A$ **and** $B$ both occur: $A \cap B$, the intersection of $A$ and $B$.

3. Event $A$ **or** $B$ occurs: $A \cup B$, the union of $A$ and $B$.

4. $A \cap B \subseteq A \cup B$. This means that $A \cap B$ is generally a smaller set than $A \cup B$ (though they can be equal). This is because membership in $A \cap B$ is more restrictive than membership in $A \cup B$.

5. **Exactly one** of the events $A$ or $B$ occurs.
   This is the **exclusive or**. It means that $A$ or $B$ occurs, but not both. In other words, the event that $A$ and $B$ occur is not allowed. In terms of sets,

   $$(A \cup B) - (A \cap B) = (A - B) \cup (B - A)$$

**Note:** If you see the word *and* connected with two events, think intersection. If you see the word *or* in connection with two events, think union (Also remember,”$\cup$ for Union”).

6. **Neither** event $A$ **nor** event $B$ occurs: this is the same as “$A$ does not occur and $B$ does not occur” = $A^c \cap B^c$.

7. When solving probability word problems:

   **Step 1:** Write down the events of interest in the problem.

   **Step 2:** Write down the given information in terms of the events.

   **Step 3:** Write down what you’re looking for.

   **Step 4:** Solve
Fundamental properties of probability for sample spaces

Let $\Omega$ be the sample space for an experiment. Then:

1. If $E$ is any event, then $0 \leq P(E) \leq 1$.
2. If $E$ is certain to happen, then $P(E) = 1$. In particular, $P(\Omega) = 1$.
3. $P(E^c) = 1 - P(E)$. In particular, $P(\Omega^c) = P(\emptyset) = 0$.
4. If $E$ and $F$ are two events, then
   $$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$
5. If $E$ and $F$ are mutually exclusive (disjoint) events, then
   $$P(E \cup F) = P(E) + P(F)$$
   If $E_1, E_2, \ldots, E_n$ are mutually exclusive (disjoint) events, then
   $$P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)$$
6. $P(E^c \cap F^c) = P((E \cup F)^c) = 1 - P(E \cup F)$
7. $P(E^c \cup F^c) = P((E \cap F)^c) = 1 - P(E \cap F)$
8. If $P(F) \neq 0$, then
   $$P(E|F) = \frac{P(E \cap F)}{P(F)}$$
   It follows that
   $$P(E \cap F) = P(F)P(E|F) = P(E)P(F|E)$$
9. Suppose $E$ is an event; and $F_1, \ldots, F_n$ is a partition of the sample space $\Omega$. Then
   $$P(E) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

Bayes’s Theorem

Let $E$ be an event; and let $F_1, \ldots, F_n$ be a partition of the sample space $\Omega$ for some experiment. Then for any $k \in \{1, \ldots, n\}$,
$$P(F_k | E) = \frac{P(E \mid F_k)P(F_k)}{\sum_{i=1}^{n} P(E \mid F_i)P(F_i)}$$
If all outcomes are equally likely, then
$$P(F_k | E) = \frac{\#(E \cap F_k)}{\sum_{i=1}^{n} \#(E \cap F_i)}$$
Finite random variables

Range of $X$ is $\Omega = \{x_1, x_2, \ldots, x_N\}$ with $x_1 < x_2 < \cdots < x_N$.

The probability mass function (p.m.f.) is defined as
\[ f_X(x) = P(X = x) . \]

The cumulative distribution function (c.d.f.) is defined as
\[ F_X(x) = P(X \leq x) . \]

Some useful facts:
\textbf{Fact 1:} If $x < x_1$, then $F_X(x) = 0$.
\textbf{Fact 2:} If $x \in [x_i, x_{i+1})$ for some $i = 1, \ldots, N$, then
\[ F_X(x) = \sum_{k=1}^{i} f_X(x_k) . \]
\textbf{Fact 3:} If $x \geq x_N$, then $F_X(x) = 1$.
\textbf{Fact 4:} $F_X(x_i) = F_X(x_{i-1}) + f_X(x_i)$, for $i = 2, \ldots, N$
and $F_X(x_1) = f_X(x_1)$.
\textbf{Fact 5:}
\[ E(X) = \sum_{n=1}^{N} x_n f_X(x_n) \]
\textbf{Fact 6:} Suppose all outcomes are equally likely. Then for every $i$,
\[ f_X(x_i) = P(X = x_i) = \frac{1}{N} \]
If $E$ is an event, and $\#(E)$ is the number of outcomes in the set $E$, then
\[ P(E) = \frac{\#(E)}{\#(\Omega)} = \frac{\#(E)}{N} \]
If $E$ and $F$ are both events, then
\[ P(E \mid F) = \frac{P(E \cap F)}{P(F)} = \frac{\#(E \cap F)}{\#(F)} = \frac{\#(E \cap F)}{\#(F)} \]
\textbf{Fact 7:} If the expected value of $X$ is $\mu_X = E(X)$, then the formula for the variance of $X$ is
\[ V(X) = \sum_{i=1}^{N} (x_i - \mu_X)^2 f_X(x_i) = E(X^2) - (E(X))^2 \]
\textbf{Fact 8:} The formula for the standard deviation of $X$ is
\[ \sigma_X = \sqrt{V(X)} \]
Continuous random variables

The range of $X$ is an interval.

**Definition:** The cumulative distribution function (c.d.f.) is defined as

$$F_X(x) = P(X \leq x).$$

This is the same definition as in the discrete case.

**Fact 1:** For each $x$ in the range of $X$: $P(X = x) = 0.$

**Fact 2:** $P(a \leq X \leq b) = F_X(b) - F_X(a)$

**Definition:** If $X$ is a continuous random variable, the **probability density function** (p.d.f.) is a function $f_X$ with the property that if $a \leq b$, then $P(a \leq X \leq b)$ is the area of the region “under” the graph of $f_X(x)$ from $a$ to $b$.

**Fact 3:** The formula for the expected value of $X$ is

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x f_X(x) \, dx$$

**Fact 4:** If the expected value of $X$ is $\mu_X$, then the formula for the variance of $X$ is

$$V(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) \, dx = E(X^2) - (E(X))^2$$

**Fact 5:** The formula for the standard deviation of $X$ is

$$\sigma_X = \sqrt{V(X)}$$

**Fact 6:** Suppose $X$ is a random variable with the normal distribution, with mean $\mu_X$ and standard deviation $\sigma_X$. Then the formula to convert the random variable $X$ to the standard normal random variable $Z$ is

$$Z = \frac{X - \mu_X}{\sigma_X}$$
Compound Interest (Business math only)

Let $F$ denote future value, $P$ denote principal, or initial investment, $r$ denote annual interest rate, $n$ denote the rate of compounding per year, and $t$ denote the time (in years).

The formula for discrete compounding is

$$F = P \left(1 + \frac{r}{n}\right)^{nt}.$$

The formula for continuous compounding is

$$F = Pe^{rt}.$$

Formulas for effective annual yield.

Discrete case:

$$y = \left(1 + \frac{r}{n}\right)^n - 1.$$

Continuous case:

$$y = e^r - 1.$$