LIBERAL ARTS
MATHEMATICS

MATH-100

ZERO TEXTBOOK COST

An Internationalized Version

Liberal Arts Mathematics

An Internationalized Version

An Open Educational Resources Publication by College of the Canyons

Authored and compiled by Saburo Matsumoto

Editors: Trudi Radtke & Alex Gavilan

Cover by Kyra Karatsu
Cover image by Manik Roy on Unsplash

Version 2.7.3

2025

Acknowledgements

We would like to extend appreciation to the following people and organizations for allowing this textbook to be created:

California Community Colleges Chancellor's Office
Santa Clarita Community College District
College of the Canyons Distance Learning Office

We would also like to acknowledge the following faculty in the Mathematics Department at College of the Canyons for reviewing parts of this textbook:

Tigran Alikhanyan, Eli Barbullushi, Kahroliné Di Passero, Collette Gibson, Joe Gerda (emeritus), Brandon Hilst, Kathy Kubo.

*Unless otherwise noted, the content in this textbook is licensed under CC BY 4.0

Table of Contents

Chapter 1: Mathematics and Problem Solving **13**

1.1 Fallacies in Common Language *14*

1.2 Statements and Logic *24*

1.3 Truth Tables *36*

1.4 Analyzing Arguments *58*

1.5 Problem Solving *74*

For Further Explorations… *87*

Chapter 2: Mathematics and Numbers **89**

2.1 Percentages *90*

2.2 Deceptive and Misleading Numbers *99*

2.3 Rates, Proportions, and Conversions *109*

2.4 Estimating the Answer *122*

2.5 Really Big and Really Small Numbers *127*

For Further Explorations… *133*

Chapter 3: Mathematics and Finance **136**

3.1 Simple and Compound Interest *137*

3.2 Continuous Compounding *150*

3.3 Savings and Annuities *155*

3.4 Loans and Mortgages *161*

3.5 National Deficit and Debt *170*

For Further Explorations… *175*

Chapter 4: Probability and Odds **178**

4.1 Basic Concepts *179*

4.2 Conditional Probability *190*

4.3 Counting *203*

4.4 The Law of Averages and Expected Value *223*

For Further Explorations… *234*

Chapter 5: Data and Statistics **236**

5.1 Basics of Statistics *237*

5.2 Types of Statistical Studies *254*

5.3 Describing Data *266*

5.4 Numerical Measures of Center and Variation 279
5.5 Normal Distribution 299
For Further Explorations… 308

Chapter 6: Growth and Decay **310**
6.1 Growth Models 311
6.2 Logarithmic Scales in Natural Sciences 329
6.3 The pH Scale 337
For Further Explorations… 340

Chapter 7: Mathematics and the Arts **343**
7.1 Perspective in Art 344
7.2 The Golden Ratio and the Fibonacci Sequence 349
7.3 Musical Scales 354
7.4 Fractals 361
For Further Explorations… 378

Chapter 8: Mathematics and Politics **380**
8.1 Apportionment 381
8.2 Gerrymandering 402
8.3 Theory of Voting 410
For Further Explorations… 440

Chapter 9: Selected Topics **443**
9.1 Four-Color Theorem 444
9.2 How Big Is Infinity? Or Is It "Infinities"? 447
9.3 Seven Bridges of Königsberg 455
9.4 Russell's Paradox 459
9.5 Non-Euclidean Geometry 461
9.6 Mathematical Problem Solving Illustrated 468

Chapter 10: International Topics Supplement **475**
10.1 Project 1: The Pythagorean Theorem 476
10.2 Project 2: Musical Scales 481
10.3 Project 3: Symbols and Symmetries 485
10.4 Project 4: Currencies and Denominations 494
10.5 Project 5: The Language of Numbers 497
10.6 Project 6: License Plates 503

10.7 Project 7: Units of Measurement *507*
10.8 Project 8: Word Problems *511*
10.9 Project 9: Numerology *516*
10.10 Project 10: Health Indices *520*
10.11 Project 11: Research on π *525*
10.12 Project 12: History of Algebra *531*
10.13 Project 13: Population Growth *537*
10.14 Project 14: The Golden Ratio *542*
10.15 Project 15: Ethnomathematics *549*
10.16 Project 16: Mathematics in Literature, Cinema, and Theater *554*
10.17 Project 17: Games and Mathematics *558*
10.18 Project 18: Gambling and Mathematics *563*
10.19 Project 19: Fields Medal and Its Winners *568*
10.20 Project 20: Fibonacci, Leonardo of Pisa *572*

Appendix A - Try It Now Answers **576**

Appendix B: Solutions to Exercises **591**

References **631**

Introduction

What Is Liberal Arts Mathematics?

Until about the last decade of the 20th century, almost every college and university in the United States required College Algebra or a similar algebra-based mathematics course for a degree, regardless of the student's major. Then in the 1990s, institutions of higher learning began offering a new mathematics course more applicable and appropriate for students not majoring in STEM (Science, Technology, Engineering, and Mathematics) fields. This big push was in part due to what was regarded as unnecessary skills and procedures in algebra, which seemed rather irrelevant to many students' academic goals. With increasing significance of data and information, many began to see the need for a course in which quantitative understanding and reasoning were emphasized. Their idea was that, while STEM students were justifiably required to take calculus (and courses leading to calculus, such as trigonometry and intermediate algebra), others majoring in liberal studies, humanities, fine arts, hospitality, and other non-STEM fields could benefit more from a different type of mathematics course.

Soon textbooks began to be published as these new courses started to appear in colleges and universities across the country. These courses went by many names, covering a wide variety of topics. It turns out that the algebra requirement was a much bigger issue than originally thought, and this new approach quickly gained momentum. Because different schools saw different needs for different students, several non-algebra options were developed to satisfy the general-education requirement in mathematics.

Perhaps the most obvious alternative was an introductory statistics course. Most institutions already had such a course, and it became, and still is, a very popular substitute for College Algebra. While statistics education has seen many changes in recent years, most introductory statistics courses include topics such as data collection, bias, graphs and charts, sampling distributions, probability, confidence intervals, and hypothesis testing.

Other schools slightly modified their existing algebra course with a focus on modeling. This approach is basically to offer topics common in College Algebra (polynomial, exponential, radical, rational, and logarithmic functions) and use them to model real-world phenomena often studied in natural and social sciences. Such a course typically has titles like "College Algebra for Liberal Arts."

Another popular approach was the development of a course in finite mathematics. Yet other colleges allowed students to take symbolic logic or history of mathematics to satisfy the degree requirements. Many schools also began teaching new courses with names such as Quantitative Reasoning, Mathematics for Liberal Arts, and Thinking Mathematically. While the course titles vary, the central idea was the same: to deliver content that would help non-STEM students see real-life applications of mathematics and appreciate its usefulness and versatility. With these objectives, such courses typically cover topics including voting methods, compound

interest, probability, statistics, mathematics in fine arts, exponential growth, and major achievements in the history of mathematics. For consistency, we use the title “Liberal Arts Mathematics” to refer to courses containing these topics.

In other words, if you are a student not majoring in a STEM discipline, and if your degree program does not require Statistics, then Liberal Arts Mathematics is a course specifically designed for you!

Today, there are many textbooks available on the market as well as open educational resources (OERs) for Liberal Arts Mathematics. In fact, a typical book contains far more topics than can be covered in a usual, one-semester, three-unit course. This book is no exception. Are there topics more important than others? Which topics should be included/emphasized? How should those topics be taught? How deeply should they be covered? How much mastery should you, as students, demonstrate? These are some of the questions that your college and your instructor would consider as they prepare for your Liberal Arts Mathematics course.

Some Myths about Liberal Arts Mathematics

The main purpose of a Liberal Arts Mathematics course is to help you develop skills for problem solving and critical thinking. Too many students grow up thinking that mathematics is all about memorizing formulas and procedures. On the contrary, problem solving means defining/identifying the problem, figuring out a strategy, perhaps working on simpler cases first, recognizing patterns, carrying out trial-and-error, carefully and logically finding answers, and reflecting on the entire process. Whether the topic involves fine arts, finance, politics, economics, risks, logic, or simple percentages, these problem-solving skills are the goals we want you to achieve. To this end, we present some commonly believed myths about Liberal Arts Mathematics.

“Topics and questions must be practical and useful in the real world.”

We definitely want you to be able to solve real-world problems, and that is, as mentioned above, one of our main objectives. However, that does not mean that every topic and every question must be real and practical. It may sound contradictory, but the truth is that mathematics is all about *abstract* thinking. Abstraction and imagination are what empower us to use mathematics to solve real-life problems in so many different areas. Throughout this book, you will see the “same math” applied in completely different fields. For instance, exponential growth is used to describe not only populations but also compound interest and musical scales. “Abstracting” certain underlying patterns is a great problem-solving tool, and it is important to build this skill. Some exercises may be quite artificial (such as determining validity of an argument with a bunch of made-up words), fictional (such as a toxic substance taking over the whole world), and silly (such as asking if there are people with the same number of hairs), but they are useful in developing this essential abstracting skill. If you are an athlete, you may lift weights and do many silly exercises which you will never do in actual games; these physical exercises are meant to develop your basic muscles, agility, and

coordination. Similarly, some exercises in this book may not appear very applicable, but they are designed to build and stretch those thinking/mental muscles you have.

"Topics must be easy enough for non-STEM students."

Contrary to popular opinion, this course is not, should not, and must not be a watered-down course for mathematically challenged students to pass just to get their degree. It may be true that some students enrolled in Liberal Arts Mathematics already feel anxious about mathematics. It is possible that some of you did poorly in your high school math classes. But time and again, I have seen students who appreciate the challenge of rigorous mathematical thinking. Your earlier math grades may not be an accurate indication of your mathematical prowess. Perhaps you are a good mathematics student after all—just not the type who can memorize and follow algebraic procedures for good grades. In fact, I have had Liberal Arts Mathematics students discover some complex topics fascinating (such as the Fibonacci Sequence, the Golden Ratio, the Four-Color Theorem, and fractals) and study them further in their projects. Others, maybe for the first time, see that their area of interest (such as animation, art, hospitality, social justice, and politics) has a lot to do with mathematics and give deep and profound insight in class discussions.

Many students will find a few topics in this book quite challenging: an indeterminant form that leads to the definition of *e* (Section 3.2), the formula for the normal distribution function (Section 5.5), fractal dimensions (Section 7.4), geometric series that gives us annuity and loan formulas (Sections 3.4 and 3.5), and the like. But the idea is that, while you may not fully understand these topics, just being exposed to these concepts could help you appreciate the usefulness and unity of mathematics. It is not necessary for you to master these, but hopefully you will remember that these elegant concepts do exist in mathematics, and perhaps you may even run into these ideas when they are alluded to or presented in films, books, and other resources.

"Many formulas and procedures are essential and must be memorized."

Perhaps this is one of the biggest misconceptions about mathematics, especially in American schools. As mentioned above, the power of mathematical thinking lies in its abstraction. Problem solving requires recognition and discovery of an underlying relation or pattern. New problems may require thinking "outside the box" with creativity as well as logic (as illustrated in the story of Archimedes in the bathtub). For this reason, whenever possible, this book attempts to explain the logical thinking process behind the formulas and procedures. Where there is a complicated formula (like Bayes' Theorem, permutations, and combinations), the book adds intuitive ways to carry out the calculations.

"Math is a great tool but nothing more."

No one would deny that mathematics is a tool so powerful that it can be applied in all kinds of fields. In this book, we see that it is a tool used in business, finance, sociology, music, art,

politics, and more. However, nothing could be further from the truth if you think of mathematics *only* as a technical tool for problem solving. In fact, one of the key objectives of this book is to show mathematics as a part of humanities; like art, literature, and music, it can and should enhance our human experience and shed light on what it means to be human. Unfortunately, this is not how mathematics is traditionally taught in schools today. Throughout history, however, mathematics was not necessarily regarded as a natural science but rather a part of liberal arts (so the term "Liberal Arts Mathematics" could be redundant). At the entrance to his "Academy," Plato put up a sign "Let no one ignorant of geometry enter." In fact, the word "mathematics" itself comes from a Greek word meaning "things/lessons to be learned," and a student was referred to as a *mathetes*, a "learner" or a "disciple." Mathematics was taught not as a tool, but as a way to seek truth and think logically; it was an aspect of philosophy.

We all ask questions and seek explanations. It is in our nature to discover patterns and relations. We feel joy and excitement when we finally figure out how something works, shouting things like "Oh, I get it!" That sense of pleasure and accomplishment is perhaps a reward for enduring a long thinking process. This could happen when you solve a crossword puzzle or when a mechanic finally finds out why a car is not running smoothly. Curiosity is part of being human, and so is a desire to understand things. That is, in essence, a mathematical pursuit.

In a book appropriately titled *Mathematics for Human Flourishing,* the mathematician Francis Su writes that "math skills are valuable, but they cannot serve as goals. The skills society needs from math may change, but the virtues needed from math will not."[1]

"Pure mathematics is, in its way, the poetry of logical ideas."
—Albert Einstein

Some Highlights in This Edition

Based on the objectives and principles stated above, I have included some new content in this edition of the book. Exercise sets (now appearing at the end of each section, not each chapter) do contain challenging problems, which your instructor may assign for extra credit or optional homework. At the end of each chapter, you will find "**For Further Explorations…"** topics, where you are encouraged to pursue more advanced topics related to that chapter. These could provide ideas for your projects as well. The text itself, in many sections, has been updated for more in-depth exposition. I have also added inspiring quotations where appropriate. At the end of Chapter 9, I have added a new section in which I walk through a simple but intriguing mathematical problem to demonstrate what it means to "do mathematics" and to be engaged in problem solving. I highly recommend reading it (Sec. 9.6).

[1] Francis Su, *Mathematics for Human Flourishing* (New Haven: Yale University Press, 2020), p. 11. He also states, on p. 10, that the "proper practice of mathematics cultivates virtues that help people flourish."

Finally, the new coronavirus and the society's response have given us many timely and crucial lessons; these learning opportunities are not to be missed. The COVID-19 pandemic has taught us about cause and effect, various fallacies, risks and expected values, statistical placebo-controlled experiments, and exponential growth, just to name a few. In addition, recent news stories on racism, social justice, discrimination, the election, and the 2020 Census also provide valuable resources related to the content of this book. Many of these current events are included in this edition. It is my hope that you learn to think—logically and rationally—as you see and hear what is taking place in our world today.

Sab Matsumoto
Department of Mathematics
College of the Canyons
Santa Clarita, California

Topics Related to Current or Recent Events

- COVID-19 vaccines, side effects, and health protocols (Sec. 1.1)
- California Proposition 16 (Sec. 1.2)
- California's Anti-Publish Housing Law (Sec. 1.2)
- One Trillion Trees Initiative (Sec. 2.5)
- Federal deficit (Sec. 3.5)
- Risk and COVID-19 vaccines (Ch. 4 Explorations)
- Exponential growth of COVID-19 (Ch. 6 Explorations)
- The 2020 Census (Sec. 8.1)
- Ranked-choice voting in New York (Sec. 8.3)
- Huntington-Hill method after the 2020 Census (Ch. 8 Explorations)

Challenging or Mind-Boggling Mathematical Results

- Infinite ping pong balls in a bin (Sec. 1.5 & 9.2 Exercises)
- Five Points on the Sphere (Sec. 1.5)
- Hilbert's Hotel (Sec. 9.2 Exercises)
- The "Hair Problem" (Sec. 2.5 Exercises)
- Some big numbers (Ch. 2 Explorations)
- Indeterminate form and *e* (Sec. 3.2)
- The Birthday Problem (Sec. 4.3)
- The Hat-Check Problem (Sec. 4.4 Exercises)
- The Monty Hall Problem (Sec. 4.4 Exercises)
- The Doubling Bet Problem (Sec. 4.4 Exercises)
- Calabresi's Fable (Ch. 4 Explorations)
- Doubling growth (Sec. 6.1 Examples & Exercises)
- Cantor's Disappearing Table (Sec. 7.4 Exercises)
- Four-Color Theorem (Sec. 9.1)
- Infinities (Sec. 9.2)
- Seven Bridges of Königsberg (Sec. 9.3)
- Russell's Paradox (Sec. 9.4)
- Non-Euclidean geometry (Sec. 9.5)
- Riddle of 100 Lightbulbs (Sec. 9.6)

Chapter 1: Mathematics and Problem Solving

"Fallacies do not cease to be fallacies because they become fashions."—Gilbert K. Chesterton

We are surrounded by information—information in various forms such as numbers, charts, graphs, figures, sentences, songs, speeches, and conversations. Much of the information comes through media including television, magazines, and online communications. Many of the things we hear and read are true while many are not. How are we to determine what is true and what is not? What is truth anyway? Are there ways to detect false information? What are we to watch out for?

One of the primary goals of this book is to answer some of these questions through logic and mathematics. This is particularly important in the Internet-dominated, media-centered, and information-driven age in which we find ourselves living today. With new technologies, it is becoming much easier for businesses and organizations to send out deceptive messages and misleading advertisements, and they can target specific individuals based on their online activities. Throughout this book you will find data, numbers, and statements that can be presented honestly or dishonestly in a deceptive manner. They may be numbers, percentages, odds, statistics, or written sentences.

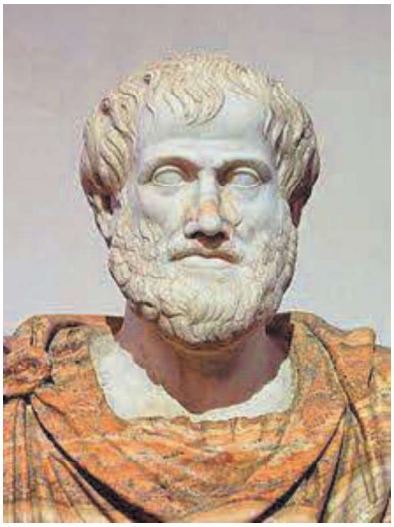


Figure 1.1: Aristotle[2]

In this first chapter we study logic, both correct use of logic and logical fallacies. Before we get into the formal side of logic, however, we first look at some types of fallacies presented as English sentences. In other words, we will see some bad examples of logical arguments presented in common language. Then, in the following sections we will look carefully into some formal components of logic like propositions, truth tables, and logical equivalences.

[2] Image is in the public domain

1.1 Fallacies in Common Language[i]

An argument can be wrong when the assumptions are not true, when they are not sufficient to guarantee the conclusion, or when there are invalid chains in logic. There are a number of other ways in which arguments can be faulty, and we will study these using logical structures later. But first, let us get familiar with some faulty arguments in common language. A faulty argument is called a fallacy.

***Ad Hominem* (Personal Attack)**

An *ad hominem* argument attacks the person making the argument, ignoring the argument itself.

In this fallacy, the character flaw mentioned should be unrelated to the argument.

"X is stated by someone with a character flaw (which is not related to X). Therefore, X is false." The "not related" part is explained in each of the examples below.

Example 1

"The doctor says I have a rare type of cancer, but I don't believe him because he is an atheist."

This statement claims that the doctor's diagnosis is not trustworthy because of his religious belief (or lack thereof), which is unrelated to his abilities as a doctor. This is an example of an *ad hominem* (or personal attack) fallacy.

If the statement were "The doctor says I have a rare type of cancer, but I don't believe him because he is an animal doctor," then the "flaw" (being an animal doctor as opposed to a human doctor or a cancer specialist) is related to the claim, i.e., the "flaw" may be a valid reason for doubting his diagnosis. In that case, it would not be a fallacy.

Example 2

"Jane says that whales aren't fish, but she is only in second grade, so she can't be right."

Here the argument is attacking Jane, not the validity of her claim, so this is an *ad hominem* argument. Note that her age (or grade) is not directly related to whales and fish.

If the statement had been, "Jane says that getting old is no fun, but she is only in second grade, so I can't believe her," then this is not an *ad hominem* argument since her age is related to the "getting old" part (i.e., she does not yet have credentials to speak about being old).

On the other hand, consider this statement: "Jane says that whales aren't fish, but everyone knows that they're really mammals. She's so stupid." This certainly is not very nice, but it is not *ad hominem* since a valid counter-argument is made along with the personal insult. Even though *ad hominem* is often called "personal attack" (as a fallacy type), it is different from a personal insult.

Appeal to Ignorance

An appeal to ignorance assumes something is true because it has not been proven false

"X is not proved false. Therefore, X is true."

Example 3

"Nobody has proven that photo isn't of Bigfoot, so it must be Bigfoot."

"I know there are aliens in space because science has not yet eliminated the possibility."

Certainly, these examples illustrate the faulty argument that the lack of evidence against a claim is not a proof for the claim itself.

Appeal to Authority

An appeal to authority argument is an attempt to use the authority of a person to prove a claim. While often authority can provide strength to an argument, problems can occur when the person's opinion is not shared by other experts, or when the authority is irrelevant to the claim.

"X is stated by an expert in the field. Therefore, X is true."

Example 4

"A diet high in bacon can be healthy; my personal doctor said so."

Here, an appeal to the authority of a doctor is used for the argument. This generally would provide strength to the argument, except that the opinion that eating a diet high in saturated fat runs counter to general medical opinion. The opinion of one (or a few) doctors does not necessarily make a claim true. Opinions of doctors vary on this topic. More supporting evidence would be needed to justify this claim.

Example 5

"Jennifer Hudson lost weight with Weight Watchers, so their program must work." Here, there is an appeal to the authority of a celebrity. While her experience does provide evidence, it provides no more than any other person's experience would.

Appeal to Consequence

An appeal to consequence concludes that a statement is true or false based on whether the consequences are desirable or not.

"I would like X to be true. Therefore, X is true."

Example 6

"Humans will travel faster than light because faster-than-light travel would be beneficial for space travel."

"Global warming must be fake because many dire consequences will happen if the earth is actually getting warmer; businesses will suffer, with huge negative impact on the economy."

Appeal to consequence is much more prevalent than we may realize. All of us are emotional beings, and we tend to believe what we want to believe and reject evidence we would rather not believe. We have more to say on this subject at the end of this chapter in "For Further Explorations…"

False Dilemma (Limited Choice)

A false dilemma argument falsely frames an argument as an "either or" choice, without allowing for additional options.

"Either X or Y is true (and only these two are possible). X is not true. Therefore, Y must be true." Here, the fallacy occurs in the assumption that only two options exist. This is a very common type of fallacy—the two limited choices are often the two extreme positions, with nothing in the middle.

Example 7

"Either those lights in the sky were an airplane or aliens. There are no airplanes scheduled for tonight, so it must be aliens."

This argument ignores the possibility that the lights could be something other than an airplane or aliens.

As mentioned above, while many people may not recognize the faulty premise, this is very common. During a virus epidemic, one may say, "You are *against* shutting down all businesses? Don't you care about the epidemic?" This is a false dilemma; the implicit assumption is that there are only two choices—to not care about the epidemic or to shut down all businesses. There may be a way to care about the epidemic without shutting down all businesses.

Similar attacks are quite prevalent from all political angles. One could, for instance, say, "You are against universal health care. Do you want people to die because they do not have access to health care?" Another could also ask, "You support universal health care? You must be a socialist." It is easy to see that false dilemma could be an obstacle in proper and productive arguments and a democratic process. One can also say that false dilemma is often used as a "weapon" in political attacks.

Circular Reasoning

Circular reasoning is an argument that relies on the conclusion being true for the reason to be true.

"X is true because Y is true; Y is true because X is true."

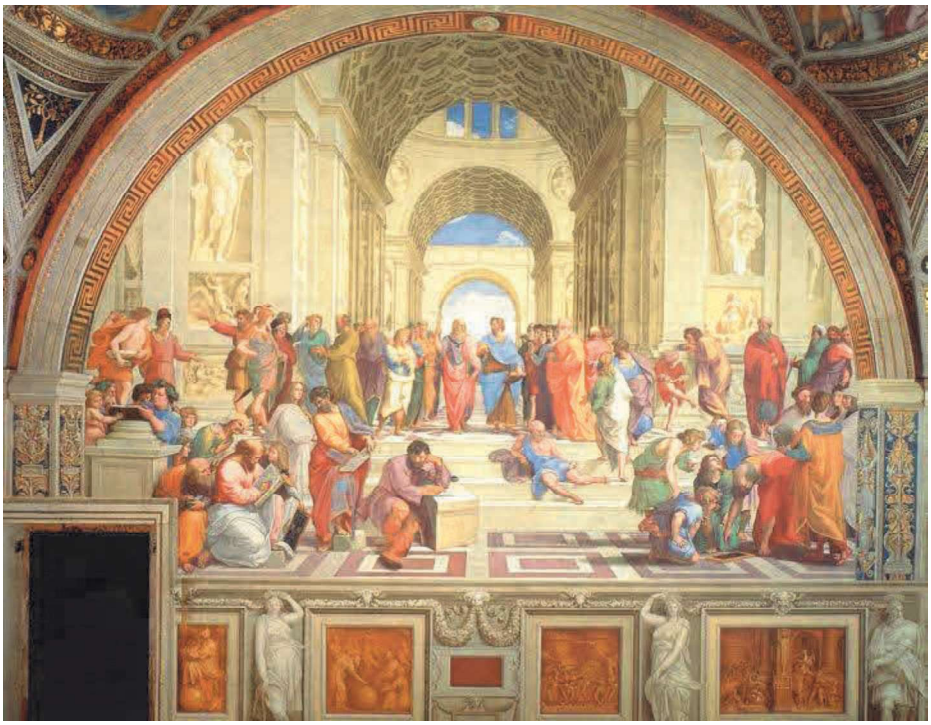


Figure 1.2: The School of Athens[3]

Example 8

"America is the best country to live in because it is better than any other place." While America may be the greatest country to live in, this claim is "circular." According to this statement, the reason America is the best country for living is because it is better than any other place. Why is it better than any other place? Because it is the best country to live in. If you want to support this statement, you need some specific or general evidence.

[3] Image is in the public domain

Example 9

Tom said, "I shouldn't have gotten a C in that class because I am an A student!"
In this argument, the student is claiming that because he is an A student, he should not have gotten a C. But because he got a C, he is not an A student.

The fundamental problem in circular reasoning is an important one to remember: *one cannot prove a conclusion based on something that assumes that conclusion*. Imagine someone saying this: "I know I am going north because my back is facing south. How do I know my back is facing south? Well, because I am facing north." While he *may* be going north, such an argument is clearly not valid.

***Post Hoc* (False Cause)**

A *post hoc* argument claims that because two things happened sequentially, the first must have caused the second.

"X happened before Y happened. Therefore, X caused Y."

Example 10

"Today I wore a red shirt, and my football team won! I need to wear a red shirt every time they play to make sure they keep winning."

"I ate some potato salad for lunch, and I felt sick all afternoon. That potato salad must have been bad."

The first example about a red shirt and football is so obvious that it sounds more like a joke. The second example is more realistic and reasonable. However, whereas potato salad may indeed have been the cause of sickness, this, in and of itself, cannot guarantee causation. It could have been something else this person ate or drank (or some other factor).

A word of warning—many headlines in the media, intentionally or unintentionally, report that something had happened right before something else happened, possibly suggesting a link between those two events. Fully aware of the *post hoc* fallacy, the editor may be trying to manipulate the reader's mind. So watch out. *Just the fact two events happened sequentially does not mean the first caused the second*. Here is an example of one such headline from January 22, 2021: "Home Run King Hank Aaron Dies of 'Undisclosed Cause' 18 Days After receiving Moderna Vaccine."[4] While the headline does not explicitly link his death with the vaccine, it may be a way to mislead the reader's mind.

[4] CHDT. (2021). Home Run King Hank Aaron Dies of 'Undisclosed Cause' 18 Days after Receiving Moderna Vaccine. *The Defender.* https://childrenshealthdefense.org/defender/hank-aaron-dies-days-after-receiving-moderna-vaccine/?fbclid=IwAR3W2wtp3h8SSuRBIwK2-GYifYcLk6LPzDbje_aPkiDGUxN-IVT8vVH85a4.

Sometimes a *post hoc* argument can be debunked with a careful study or analysis. For example, during a 10-week period from December, 2020, to March, 2021, it was reported that about 150 people suffered sudden hearing loss following their COVID-19 vaccine.[5] So vaccine could cause hearing loss, right? When a team of doctors from Johns Hopkins University investigated the cases further, they discovered that the rate of people suffering sudden hearing loss after the vaccination was significantly lower than before the pandemic; they concluded that there was no association.

Straw Man

A straw man argument involves misrepresenting the argument in a less favorable way to make it easier to attack.

"X is similar to Y, and Y is false (or easier to attack). Therefore, X is false."

This fallacy takes someone's argument or position, exaggerates it or distorts it in such a way that it would look ridiculous, and then attacks that ridiculous idea. It is like making an image (straw man) of someone and attacking the image because an image is easier to attack than a real human being.

Example 11

"Senator Jones has proposed reducing military funding by 10%. Apparently, he wants to leave us defenseless against attacks by terrorists."

Here the person arguing has represented a 10% funding cut as equivalent to leaving us defenseless, making it easier to attack Senator Jones' position. This may remind you of the "false dilemma" fallacy, and you are right. Many fallacies belong to more than one category; these types are not mutually exclusive.

Correlation-Causation

Correlation-causation is similar to *post hoc*, but without the requirement of sequence. This fallacy assumes that because two things are related to each other (or two quantities are associated with each other), one must have caused the other. Often there is a third variable not considered.

"X and Y seem to be correlated with each other. Therefore, X causes Y." Remember this wise saying, "**Correlation (or association) does not imply causation**." We will see this statement repeated a few times throughout this book. "Correlation" means as one changes, so does the other, in some related or predictable way. Look at the examples below.

[5] Kaplan, K. (May 2021). Covid 19 Vaccines don't cause sudden hearing loss, study finds. *Los Angeles Times.* https://www.latimes.com/science/story/2021-05-20/covid-19-vaccines-dont-cause-sudden-hearing-loss-study-finds

Example 12

(1) "Months with high ice cream sales also have a high rate of deaths by drowning. Therefore, ice cream must be causing people to drown."

(2) "During the last few weeks, the average speed of vehicles on our freeways increased, but the number of fatal accidents went down. Thus, high speeds are causing fewer traffic-related deaths; in other words, speed saves lives."

The first argument is (falsely) implying a causal relation when really both are more likely dependent on the weather; that ice cream and drowning are both more likely during warm summer months. This is an important point to remember. When you see a correlation, often there is a "common underlying" factor that are causing both of these quantities to change in the way that leads to that correlation. Try to find the underlying cause (if any).

The second argument occurred during a statewide shutdown during a pandemic, which caused general traffic volume to dip (most people stayed home), causing motorists to speed (because fewer cars were there) and the number of traffic fatalities to go down (because fewer cars were there). So again, there was a common underlying cause. Higher speeds definitely did not cause fewer fatalities.

This particular type of fallacy is very, very common and could result in serious consequences, so we will add some explanations (see more examples in Section 5.2). We can observe many quantities are associated with other quantities. For example, one can observe that people in countries with high wine consumption tend to live longer (association—a positive correlation). But this fact, by itself, does not prove that drinking wine makes one live longer. It may, and it may not. Studies have shown that dog owners tend to stay healthier, and music conductors tend to live longer. But again, one cannot conclude that having a dog makes one healthier or music (or conducting) keeps one alive longer. As seen in our examples, there are often **common underlying causes** that could explain the association or correlation. For instance, the truth may be (may not be) that dog owners end up exercising more, which contributes to better health. Perhaps there is something in the lifestyle or schedule of music conductors that allows them to live longer. The point is this: **an observational study can show association but cannot establish causation**. As we will see in Section 5.2, an experimental statistical study is about the only way in which we can establish causation. In general, however, determining causation is a very difficult task.

During the COVID-19 pandemic, this problem hit home in a realistic way. "What is causing the widespread infection?" This was an extremely relevant question. People (experts and others alike) have observed that mask-wearing seems to reduce the spread, and so do social-distancing, shutting down businesses, and banning indoor gatherings. Did masks slow down the rate of infection? Probably, but again, observation by itself cannot definitively establish

causation. Is it possible that those wearing a mask may have been cautious people who stayed home more and washed their hands more frequently? Without additional experiments, it is difficult to answer these questions.

We will have more to say about this in Chapter 5 when we discuss statistical studies. In particular, the correlation-causation fallacy is closely related to factors known as **confounding variables** and **lurking variables**.

It may be difficult to identify one particular fallacy type for an argument. As we stated earlier, this list is not mutually exclusive, meaning that some claims can belong to multiple types of fallacy. Nor is our list exhaustive; there are many other types of fallacies.[6] As you read or hear people making various statements or claims, be sure to exercise your critical thinking, with these fallacy types in mind.
Also, remember that just the fact that a fallacy is in an argument does not mean that the conclusion is wrong. See the examples below.

Example 13

(1) "Whales are mammals because my big brother told me so."

(2) "That medicine really works because my pain went away last time I had a huge headache."

(3) "I know there are aliens in space because science has not yet eliminated the possibility."

Note that (1) contains an appeal to authority ("my big brother"). Clearly, not everything one's big brother says is correct; however, that does not mean that *everything* he says is false. In particular, whales are indeed mammals. (2) can be considered a case of "false cause" (*post hoc*). Just because taking of a medicine preceded the alleviation of pain does not *necessarily* mean the medicine really works, but it may actually be effective as a pain killer. We just do not know. (3) is an actual example used as an appeal to ignorance. Just because science has not yet completely rejected the existence of aliens does not mean that aliens *do* exist. But they may still exist. Again, we just do not know, or we do not have sufficient evidence to say yes or no.

Try It Now 1

Identify the logical fallacy in each of the arguments.

a. Only an untrustworthy person would run for office. The fact that politicians are untrustworthy is proof of this.

[6] For example, some books have a fallacy type called **diversion** (also known as a "**red herring**"), a way to distract, mislead, and/or divert the attention from the real issue in question. There is also one called **appeal to popularity** ("Everyone says that, so it must be right.") There are many other fallacy types.

b. Since the 1950s, both the atmospheric carbon dioxide level and obesity levels have increased sharply. Hence, atmospheric carbon dioxide causes obesity.
c. The oven was working fine until you started using it, so you must have broken it.
d. You can't give me a D in the class because I can't afford to retake it.
e. The senator wants to increase support for food stamps. He wants to take the taxpayers' hard-earned money and give it away to lazy people. This isn't fair, so we shouldn't do it.
f. Victoria dislikes coffee! No worries, I got some tea.

Exercises 1.1

For each of the following statements, name the type of logical fallacy being used.

1. If you don't want to drive from Boston to New York, then you will have to take the train.

2. Every time I go to Dodger Stadium, the Dodgers win. I should go there for every game.

3. Legendary ex-New England Patriots quarterback Tom Brady likes his footballs slightly under-inflated. That team has a history of bending or breaking the rules, so Brady must have told the equipment manager to make sure that the footballs were under-inflated.

4. What you are doing is clearly illegal because it's against the law.

5. The county supervisor voted against the new education tax. He must not believe in education.

6. "Apples a day keeps doctors away." No one has said apples are bad, so this old saying must be true.

7. Wine has to be good for your health because… I mean, can you imagine a life without wine?

8. Studies show that slightly overweight senior citizens live longer than underweight ones. The extra weight must make people live longer.
9. Whenever our smoke detector beeps, my kids eat cereal for dinner. The loud beeping sound must make them want to eat cereal for some reason.

10. There is a scientist who works at a really good university, and he says there is no strong evidence for climate change, especially global warming. Some politicians also question climate change. That's why I don't really believe it.

11. My neighbor cheats on his tax returns. I don't believe anything he says about health.

12. Suppose a boy wants to eat ice cream. He goes to his mother and asks, "Mom, can I eat ice cream?" She says, "As long as Dad says it's OK, you can." Then the boy goes to his father and asks, "Dad, can I eat ice cream?" He replies, "Sure, as long as Mom says it's OK." The boy then eats ice cream.

13. A: "Don't fight over small things. Just let them go."
    B: "What exactly are 'small things'? How do I know what is small and what isn't?"
    A: "Well, small things are things you really don't want to fight over."

Figure 1.3[7]

[7] Image on PxHere

1.2 Statements and Logic

Logic is basically the study of valid reasoning. When searching the internet, we use Boolean logic – terms like "and" and "or" – to help us find specific web pages that fit in the sets we are interested in. After exploring this form of logic, we will look at logical arguments and how we can determine the validity of a claim.

Boolean Logic

We can often classify items as belonging to sets (for our purposes, a "set" is simply a collection of objects).[8] If you went to the library to search for a book and they asked you to express your search using unions, intersections, and complements of sets, that would feel a little strange. Instead, we typically using words like "and," "or," and "not" to connect our keywords together to form a search. These words, which form the basis of Boolean logic, are directly related to our set operations. (Boolean logic was developed by the 19th-century English mathematician George Boole.)

Boolean Logic

Boolean logic combines multiple statements that are either true or false into an expression that is either true or false.
In connection to sets, a search is true if the element is part (a member, an element) of the set.
In connection to sets, a search is true if the element is part of the set.

Suppose M is the set of all mystery books and C the set of all comedy books. If we search for "mystery," we are looking for all the books that are elements of the set M; the search is true for books that are in the set.

When we search for "mystery and comedy," we are looking for all the books that are elements of both sets, in the **intersection**. If we were to search for "mystery or comedy," we are looking for a book that is a mystery, a comedy, or both, which is the **union** of the two sets. If we searched for "not comedy," we are looking for any book in the library that is not a comedy, the **complement** of the set C.

Connection to Set Operations

Expression	Element
A and *B*	elements in the intersection $A \cap B$
A or *B*	elements in the union $A \cup B$
not *A*	elements in the complement $\sim A$, A^c, or $\bar{A}$

[8] It may be surprising, but the notion of a set, arguably the most fundamental concept in all of mathematics, is extremely difficult to define. In set theory, the word "set" is left undefined.

Notice here that the expression "A *or* B" contains everything in A and everything in B and everything in both. This use of "or" is called *inclusive*, as opposed to *exclusive*. This is a difference between the Boolean logic use of the word and common everyday use. When restaurant servers ask, "Would you like soup or salad?" they usually mean an exclusive choice – one option or the other, but not both. In Boolean logic, however, the *or* is not exclusive – more like being asked at a restaurant "Would you like fries *or* a drink with that?" Answering "both, please" is an acceptable answer.

Example 1

Suppose we are searching a library database for Mexican universities. Express a reasonable search using Boolean logic.

[Solution]

We could start with the search "Mexico and university," but it would be likely to find results for the U.S. state New Mexico. To account for this, we could revise our search to read

Mexico *and* university *not* "New Mexico."

In most internet search engines, it is not necessary to include the word ***and***; the search engine assumes that if you provide two keywords you are looking for both. In Google's search, the keyword *or* has be capitalized as ***OR***, and a **negative sign** in front of a word is used to indicate *not*. Quotes around a phrase indicate that the entire phrase should be looked for. The search from the previous example on Google could be written:

Mexico university **–**"New Mexico"

Example 2

Describe the numbers that meet the condition:

even and greater than 0 and less than 10.

[Solution]

The numbers that satisfy all three requirements are {2, 4, 6, 8}.

Sometimes statements made in English can be ambiguous. For this reason, Boolean logic uses parentheses to show precedent, just like in the algebraic order of operations.

Example 3

Describe the numbers that meet the condition:

odd number and greater than 0 and less than 20 and (multiple of 3 or multiple of 5).

[Solution]

The first three conditions limit us to the set {1, 3, 5, 7, 9, 11, 13, 15, 17, 19}.

The last grouped conditions tell us to find elements of this set that are also either a multiple of 3 or a multiple of 5. This leaves us with the set {3, 5, 9, 15}.

Notice that we would have gotten a very different result if we had written

(odd number and greater than 0 and less than 20 and multiple of 3) or multiple of 5.

The first grouped set of conditions would give {3, 9, 15}. When combined with the last condition, though, this set expands without bounds:

{3, 5, 9, 10, 15, 20, 25, 30, 35, 40, 45, …}

Example 4

The English phrase "Go to the store and buy me eggs and bagels or cereal" is ambiguous; it is not clear whether the person is asking for eggs always along with either bagels or cereal or the person is asking for either the combination of eggs and bagels, or just cereal. How can these meanings be distinguished?

[Solution]

Using parentheses could clarify the intent:

"Eggs and (bagels or cereal)" means Option 1: Eggs and bagels. Option 2: Eggs and cereal

"(Eggs and bagels) or cereal" means Option 1: Eggs and bagels. Option 2: Cereal

Be aware that when a string of conditions is written without grouping symbols, it is often interpreted from the left to right, resulting in the latter interpretation.

Statements and Negation

So far we have been discussing Boolean algebra as it applies to sets and their operations like intersection and union. But now, we apply similar ideas, not to sets, but to sentences. As mathematical objects, a sentence (or a claim) is called a **proposition** or a **statement**. In this context, a (mathematical, logical) **statement** is a declarative sentence (thus has a subject and a verb/predicate) that is either true (T) or false (F). In the next section, we will study propositions and their operations more formally, in terms of truth tables, but for now we will focus on ways to create new statements from one or more statements; these operations include negation, conjunction, disjunction, and conditional statements. The logical study of statements (propositions) is sometimes referred to as "propositional calculus."

Statements (Propositions)

A statement is a complete declarative sentence which is either true (T) or false (F). The "T" or "F" is called the **truth value** of the statement.

Example 5

Determine if each of the following is a statement (proposition); if so, find its truth value.

- "George Washington"
- "Is today January 1?"
- "George Washington served as the President."
- Thomas Jefferson grew up in England."
- "$3 < 5$"
- "A rat is a fish."
- $x - 3 = 7$

[Solution]

"George Washington" is not a statement because it is not a complete sentence.

"Is today January 1?" is not a statement because it is not a declarative sentence.

"George Washington served as the President." This is a declarative sentence with a subject (George Washington) and a verb (served). Hence, this is a statement (proposition). It happens to be a true statement.

"Thomas Jefferson grew up in England." This is a declarative sentence with a subject (Thomas Jefferson) and a verb (grew). So this is a statement (proposition). Its truth value is false (F) since he grew up in Virginia, but it is still a statement.

"3 < 5" is also a statement because it has a subject (the number 3) and a verb/predicate (is less than). Its truth value is true (T).

"A rat is a fish" is a statement though its truth value is false (F).

"$x - 3 = 7$" is a statement, with a subject $x - 3$ and a verb/predicate (is equal to), but its truth value depends on the value of x.

Now that we have defined what a statement is, we begin with the **negation** of a statement.

Negation of a Statement

The negation of a statement A, denoted $\sim\mathrm{A}$, is the statement that always has the truth value opposite of the truth value of A.

If A is a simple affirmative statement, its negation $\sim\mathrm{A}$ can be obtained simply by inserting the word "not." If it is already a negative sentence, its negation is usually obtained by deleting the word "not" since double negatives cancel each other.

Example 6

Negate each of the following statements.

A: Joe Biden is a Democrat.
B: Ronald Reagan was not a Democrat.
C: Lionel Messi is a baseball player.

[Solution]

$\sim A$: Joe Biden is not a Democrat.
Note that the negation ~A is *not* "Joe Biden is a Republican" because it is possible that someone is neither a Republican nor a Democrat.
$\sim B$: Ronald Reagan was a Democrat.
$\sim C$: Lionel Messi is not a baseball player.

Of course, the truth value of C is false in reality (since Messi is a soccer player). But ~C is not "Messi is a soccer player." Although this last statement is true in reality, in formal logic, you must distinguish between what is real and what is logically possible. The negation of a statement must have the opposite truth value, regardless of what is really true. Suppose, for instance, Messi were neither a baseball player nor a soccer player. Then, C would be false, and so is the statement "Messi is a soccer player." This is why the statement "Messi is a soccer player" cannot be $\sim C$; both of these statements would be false under that assumption.

Things can quickly get more complicated when there are multiple words with negative connotation. Unfortunately, these complex sentences are getting more common in our society. Here is one such example from the real world.

Example 7

The 2020 California Proposition 16 sought to *repeal* Proposition 209 Affirmative Action Amendment (1996), which *prohibits* state and local governments from "discriminating against, or granting preferential treatment to, individuals or groups on the basis of race, sex, color, ethnicity, or national origin." (What does it mean to *repeal* something that *prohibits* something?) Suppose a political candidate says, "I do not support Proposition 16." Is the candidate for affirmative action? Do you support the candidate?

[Solution]

Of course, this proposition involves affirmative action, which is, generally speaking, favoring individuals belonging to certain groups known to have been discriminated against in the past. Is Proposition 16 *for* or *against* affirmative action? *For* or *against* discrimination?

As stated above, Proposition 209 generally prohibits discrimination by governments. So, for instance, a state university cannot consider race or ethnicity in its admission decision. Proposition 16 asked California voters to repeal Proposition 209. This politician does not support Proposition 16. Would you vote for the politician? What would that mean in terms of affirmative action? It is complicated, isn't it?
Let us summarize: The politician does not support repealing the discrimination-prohibiting proposition (Proposition 209). Note this sentence contains several negative terms ("not," "repeal," and "prohibiting").

Here is a way to think about this. Let us denote by "A" the position supporting affirmative action. Proposition 209 prohibits any favoring based on race or ethnicity, so Proposition 209 is against A, or "not A." The 2020 Proposition 16 sought to repeal this. So Prop. 16 is not "not A," which is "A." (Yes, Proposition 16 is *for* affirmative action.) This politician does not support that, so he/she is "not A." In other words, the politician does not support affirmative action. Therefore, if you strongly favor affirmative action, then you probably will not agree with this politician's position. (Remember, Proposition 209 is against affirmative action, and this politician wants to keep that proposition.)

Another good (and very confusing) example is a 2024 news story about *pulling from a ballot a bill that will repeal the "Anti-Public Housing" Law*.[9] (Can you spot a triple negative?) Negation gets even more complicated when quantifiers and conditionals are involved. We will see these later.

[9] See the L.A. Times Article California lawmakers are giving up a bid to repeal a nearly 75-year-old anti-public-housing measure - Los Angeles Times (June 6, 2024).

Conditional Statements

Boolean logic is commonly used in spreadsheet applications like Excel to do conditional calculations. When an action or outcome depends on the value of a statement, it forms a **conditional**.

Conditionals

A conditional is a compound statement of the form "If A, then B" (or "If A then B, else C").

(For consistency and accuracy, it is recommended that you always start with the "if" clause.)

Formally, in the statement "If A, then B," A is called the **antecedent** (or the **hypothesis**) while B is called the **consequent** (or the **conclusion**).

Example 8

In common language, an example of a conditional statement would be "If it is raining, then we will go to the mall." Identify the antecedent and the consequent.

[Solution]

The statement "it is raining" is the condition (and thus the antecedent). If the condition is true, then we will follow the first course of action and go to the mall. So "we will go to the mall" is the consequent. Note that this statement does *not* preclude the possibility that we could still go to the mall *even if* it does not rain.

Example 9

As mentioned earlier, conditional statements are commonly used in spreadsheet applications like Excel or Google Sheets. In Excel, the formula "=IF(X, Y, Z)" means "if X is true, Y is entered in the cell; otherwise, Z is entered in the cell." What does the following expression do?

=IF(A1<2000, A1+1, A1*2)

[Solution]

Notice that after the IF, there are three parts. The first part is the condition, and the second two are calculations. Excel will look at the value in cell A1 and compare it to 2000. If that condition is true (i.e., if the number in cell A1 is smaller than 2000), then the first calculation is used, and 1 is added to the value of A1 and the result is stored/entered. If the condition is false, then the second calculation is used, and A1 is multiplied by 2 and the result is stored/entered.

In other words, this statement is equivalent to saying "If the value of A1 is less than 2000, then add 1 to the value in A1. Otherwise, multiply A1 by 2." (In most spreadsheet programs, the symbol * means multiplication).

As we did earlier, we can create more complex conditions by using the operators *and, or*, and *not* to join simpler conditions together.

Example 10

A parent might say to their daughter, "If you clean your room and take out the garbage, then you can have ice cream." Under what conditions can the daughter have ice cream?

[Solution]

Here, there are two conditions:

1) The child cleaning her room.
2) The child taking out the garbage.

Since these conditions were joined with *and*, the combined conditional will be true only if both conditions are true; if either chore is not completed, then the parent's condition is not met.

By the way, the child may still get ice cream even if the two chores are not completed. The promise here applies only if the conditions are met. If the conditions are not met, no conclusions can be drawn.

Notice that if the parent had said, "If you clean your room *or* take out the garbage, then you can have ice cream," then the child would need to complete only one chore to meet the condition.

Quantified Statements

Not every statement is simple like "George Washington served as the President" and "$3 < 5$." Sometimes certain items are quantified as we will see now. There are two types of "**quantifiers**" in logic. Words that describe every element or no element of a set, such as "all," "every," and "none" are called **universal** quantifiers. In contrast, words or phrases such as "some," "one," and "at least one" are called **existential** quantifiers because they describe the existence of at least one element in a set.

Quantifiers

A ***universal quantifier*** indicates that everything in a set (or none) shares a characteristic.

An ***existential quantifier*** indicates that a set contains at least one element with a certain characteristic.

We make these statements often. "Everyone is passing this class" and "No one can live for 200 years" are examples of statements with a universal quantifier. "Someone left the cake out in the rain," "There was something in the box," and "Some students passed that course" are examples of statements with an existential quantifier. Note that the last statement can also be stated as "There exists at least one student who passed that course" to be more precise. Something interesting happens when we negate – or state the opposite of – a quantified statement.

Example 11

Suppose your friend says, "Everybody cheats on their taxes." What is the minimum amount of evidence you would need to prove your friend wrong?

[Solution]

To show that it is not true that everybody cheats on their taxes, all you need is one person who does not cheat on their taxes. It would be perfectly fine to produce more people who do not cheat, but one counterexample is all you need. In other words, the negation is an existential statement.

It is important to note that you do not need to show that absolutely nobody cheats on their taxes.

Example 12

Suppose your friend says, "Some of these six cartons of milk is leaking." What is the minimum amount of evidence you would need to prove your friend wrong?

[Solution]

In this case, you would need to check all six cartons and show that none of them is leaking. You cannot disprove your friend's statement by checking only one of the cartons.

These examples illustrate an important fact; when we negate a statement with a universal quantifier, we get a statement with an existential quantifier, and vice-versa.

Negating a Quantified Statement

- The negation of "all A are B" is "at least one A is not B."
- The negation of "no A are B" is "at least one A is B."
- The negation of "at least one A is B" is "no A are B."
- The negation of "at least one A is not B" is "all A are B."

Example 13

"Somebody brought a flashlight." Write the negation of this statement.

[Solution]

The negation is "Nobody brought a flashlight."

Example 14

"There are no prime numbers that are even." Write the negation of this statement.

[Solution]

The negation is "At least one prime number is even" or "Some prime numbers are even" or "There exists at least one prime number that is even." (Since 2 is a prime number, the negation is a true statement, not the original one.)

Try It Now 2

Write the negation of each. (Hint: Focus on the structure. Do not worry if the statements are true.)

a. **"All Icelandic children learn English in school."**
b. **"No French children speak English."**
c. **"Some Brazilian children speak Japanese."**

Exercises 1.2

1. List the set of integers that satisfy the following statement:
   A positive multiple of 5 and not a multiple of 2

2. List the set of integers that satisfy the following statement:
   Greater than 12 and less than or equal to 18

3. List the set of integers that satisfy the following statement:
   Positive even number less than 10 or odd number between 12 and 20

4. You qualify for a special discount if you are either

- a full-time college student in California or
- at least 18 and your income is less than $20,000 a year.

For each person below, determine if the person qualifies for this discount. If more information is needed, indicate that. Be precise.

(a) A 17-year-old full-time student with no job attending a California community college
(b) A 28-year-old man earning $50,000 a year
(c) A 60-year-old grandmother who does not work and does not go to school
(d) A boy in first grade
(e) A mother making $18,000 a year and not enrolled in any college
(f) A 22-year-old full-time student at Arizona State University with no income
(g) An 18-year-old woman earning exactly $20,000 a year while attending UCLA part-time

5. For each of the following, determine if it is a statement (proposition). If not, explain why not. If it is a statement, find its truth value.

(a) Dallas is in Texas.
(b) Dallas, Texas
(c) May the force be with you.
(d) Please answer all the questions.
(e) 4 is an odd number.
(f) Thomas Jefferson was the third President of the United States.
(g) Thomas Jefferson
(h) $2 + 3 = 5$

6. Write the negation: The earth is flat.

7. Write the negation: I stopped smoking.

8. Write the negation: The sun is not red.

9. Write the negation: Everyone failed the quiz today.

10. Write the negation: Someone in the car needs to use the restroom.

11. How can you prove this statement wrong?

"Everyone who ate at that restaurant got sick."

12. How can you prove this statement wrong?

"There was someone who ate at that restaurant and got sick."

13. How can you prove this statement wrong?

"There is no baseball player who can excel at both pitching and hitting. Everyone must choose one or the other."

14. How can you prove this statement wrong?

"Every student must have an ID number before registering for classes."

Figure 1.4[10]

[10] Image by on PIxabay

1.3 Truth Tables

Just as you can add, subtract, multiply, and divide numbers (these are called **binary operations**), you can also define certain operations on statements (propositions). Sometimes two complicated statements mean the same thing; we call those statements "**logically equivalent**" to each other. One of the tools to check for logical equivalence is truth tables. In this section, we learn how to carry out certain logical operations and construct truth tables.

Basic Logic Symbols

Before we focus on truth tables, we are going to introduce some symbols that are used for *and, or,* and *not.* These are quite similar to the symbols we used before for the set operations (intersection, union, and complement, respectively). Here, we assume A and B are statements (propositions).

Logic Symbols

- The symbol ⋀ is used for *and*: "A and B" is notated A ⋀ B.
- The symbol ⋁ is used for *or*: "A or B" is notated A ⋁ B.
- The symbol ~ is used for *not:* "not A" is notated ~A.

You can remember all of these symbols by relating them to the earlier symbols learned for sets. When we are working with sets, we use the rounded version of the symbols; when we are working with statements, we use the pointy version. The negation (complement) symbol is the same as before (in some logic books, you may also see $\neg A$ or $\bar{A}$ for the negation of A).

Example 1

Translate each statement into symbolic notation. Let P represent "I like Pepsi" and let C represent "I like Coke".

(1) I like Pepsi or I like Coke.
(2) I like Pepsi and I like Coke.
(3) I do not like Pepsi.
(4) It is not the case that I like Pepsi or Coke.
(5) I like Pepsi and I do not like Coke.

[Solution]

(1) P ⋁ C
(2) P ⋀ C
(3) ~P
(4) ~(P ⋁ C)
(5) P ⋀ ~C

As you can see, we can use parentheses to organize more complicated statements.

Try It Now 3

Translate "We have carrots or we will not make soup" into symbols. Let C represent "we have carrots" and let S represent "we will make soup."

Basic Truth Tables

Because complex Boolean statements can get tricky to think about, we can create a truth table to keep track of what truth values for the simple statements make the complex statement true and false.

Truth Table

A truth table is a table showing what the resulting truth value of a complex statement is for all the possible truth values for the simple statements. In a truth table, a "T" is used for true, and an "F" for false.

As with any table (rectangular chart), a horizontal set of entries is called a **row** whereas each vertical set of entries is called a **column.** In a truth table, the first few columns are reserved for the simple, individual statements; the last column is for the complex sentence you want to analyze. There should be enough rows to account for all possibilities.

Example 2

Suppose you are picking out a new couch, and your significant other says, "Get a sectional or something with a chaise." Create a truth table for this.

[Solution]

As stated above, when creating any truth table, we need to list all the possible truth value combinations for A and B. Notice below how the first column contains 2 Ts followed by 2 Fs, and the second column alternates T, F, T, F. This pattern ensures that all 4 combinations are considered. In other words, whenever two statements are involved (say *A* and *B*), we usually begin with these columns:

A	***B***	
T	T	
T	F	
F	T	
F	F	

This particular question involves a complex statement made of two simpler conditions: "(It) is a sectional" and "(It) has a chaise." For simplicity, let us use *S* to designate "(It) is a sectional" and *C* to designate "(It) has a chaise."

A truth table for this situation would look like this:

S	***C***	***S* or *C***
T	T	T
T	F	T
F	T	T
F	F	F

In the first row, if S is true and C is also true, then the complex statement "S or C" is true. This would be a sectional that also has a chaise, which meets our desire. (Remember that *or* in logic is not exclusive; if the couch has both features, it meets the condition.)

Note that since each of the two statements can be (independently) T or F, there are four (2×2) possibilities altogether. If a truth table involves three independent statements, it should have eight $(2 \times 2 \times 2)$ rows.

In the previous example about the couch, the truth table was really just summarizing what we already know about how the "or statement" works. The truth tables for the basic *and, or,* and *not* statements are shown below. These are called **conjunction**, **disjunction**, and **negation**, respectively.

Conjunction (*and*)

A	B	$A \wedge B$
T	**T**	**T**
T	F	F
F	T	F
F	F	F

Disjunction (*or*)

A	B	$A \vee B$
T	**T**	**T**
T	F	T
F	T	T
F	F	F

Negation (*not*)

A	***~A***
T	F
F	T

As you can see above (highlighted rows), a conjunction $A \wedge B$ is true only when both *A* and *B* are true; a disjunction $A \vee B$ is false only when both *A* and *B* are false.

Truth tables really become useful when we analyze more complex Boolean statements.

Example 3

Create a truth table for the statement "A or not B."

[Solution]

After creating the first two columns with those initial values listing the four possibilities as before, we create a third column for the expression $\sim B$. Now we will temporarily ignore the column for A and write the truth values for $\sim B$. Its truth value is the opposite of the truth value for B in each row. (Throughout this section, we color truth tables so that the yellow, italicized column is based on the green, bolded columns.)

A	**B**	*~B*
T	**T**	*F*
T	**F**	*T*
F	**T**	*F*
F	**F**	*T*

Next we can find the truth values of $A \vee \sim B$, using the first and third columns. Since the connector is "or," the resulting truth value is T as long as at least one of those columns is T.

A	B	~B	$A \vee \sim B$
T	T	F	*T*
T	F	T	*T*
F	T	F	*F*
F	F	T	*T*

This completes the truth table.

The truth table shows that A ∨ ~B is true in three cases and false in one case. If you are wondering what the point of this is, here is one way to interpret/apply this table. Suppose it is the last day of the baseball season and two teams, who are not playing each other, are competing for the final playoff spot. Anaheim (The Los Angeles Angels of Anaheim) will make the playoffs if it wins its game or if Boston (The Boston Red Sox) does not win its game. (Anaheim owns the tie-breaker; if both teams win, or if both teams lose, then Anaheim gets the playoff spot.) If "A" is the statement "Anaheim wins its game" and "B" is that "Boston wins its game," then A ∨ ~B represents the situation "Anaheim wins its game or Boston does not win its game." The truth table shows us the different scenarios related to Anaheim making the playoffs. In the first row, both teams win, so it is true that Anaheim makes the playoffs. In the second row, Anaheim wins and Boston does not, so it is true that Anaheim makes the playoffs. In the third row, Anaheim does not win its game and Boston wins its game, so it is false that Anaheim makes the playoffs. In the fourth row, neither teams win, so it is true that Anaheim makes the playoffs.

Try It Now 4

Create a truth table for this statement: $(\sim A) \wedge B$.

Example 4

Create a truth table for the statement A ∧ ~(B ∨ C).

[Solution]

It helps to work from the inside out when creating a truth table. It is also a good idea to create columns in the table for intermediate operations. We start by listing all the possible truth value combinations for A, B, and C. Notice how the first column contains

4 Ts followed by 4 Fs, the second column contains 2 Ts, 2 Fs, then repeats, and the last column alternates T, F, T, F... This pattern ensures that all 8 combinations are considered. After creating columns with those initial values, we create a fourth column for the innermost expression, B ∨ C. Now we will temporarily ignore the column for A and focus on B and C, writing the truth values for B ∨ C.

A	**B**	**C**	*B ∨ C*
T	**T**	**T**	*T*
T	**T**	**F**	*T*
T	**F**	**T**	*T*
T	**F**	**F**	*F*
F	**T**	**T**	*T*
F	**T**	**F**	*T*
F	**F**	**T**	*T*
F	**F**	**F**	*F*

Next, we can find the negation of B ⋁ C, working off the B ⋁ C column we just created. (Ignore the first three columns and simply negate the values in the B ⋁ C column.)

A	B	C	**B ⋁ C**	*~ (B ⋁ C)*
T	T	T	**T**	*F*
T	T	F	**T**	*F*
T	F	T	**T**	*F*
T	F	F	**F**	*T*
F	T	T	**T**	*F*
F	T	F	**T**	*F*
F	F	T	**T**	*F*
F	F	F	**F**	*T*

Finally, we find the values of A ∧ ~ (*B* ⋁ *C*). (Ignore the second, third, and fourth columns.)

A	B	C	B ⋁ C	**~ (B ⋁ C)**	*A ⋀ ~ (B ⋁ C)*
T	T	T	T	**F**	*F*
T	T	F	T	**F**	*F*
T	F	T	T	**F**	*F*
T	F	F	F	**T**	*T*
F	T	T	T	**F**	*F*
F	T	F	T	**F**	*F*
F	F	T	T	**F**	*F*
F	F	F	F	**T**	*F*

We have finally completed the truth table.

It turns out that this complex expression is true in only one case: when A is true, B is false, and C is false. To illustrate this situation, suppose that Anaheim will make the playoffs if:

(1) Anaheim wins, and (2) neither Boston nor Cleveland wins. Assuming that "T" means a win, "TFF" is the only scenario in which Anaheim will make the playoffs.

Try It Now 5

Create a truth table for this statement: $(\sim A \wedge B) \vee \sim B$.

Conditional

Earlier, we discussed conditional statements, in which we take an action based on the value of the condition (antecedent). We are now going to look at another version of a conditional, sometimes called an implication, which states that the second part must logically follow from the first.

Conditional

A conditional is a compound statement in which a statement p, called the **antecedent**, implies a statement q, called the **consequent**.
A conditional is written as $p \rightarrow q$ and is translated as "if p, then q."

In mathematics and logic, there are many ways to say "if p, then q." One can say that "p implies q." It is also equivalent to saying "p is sufficient for q" (because p is a sufficient condition for q to be true) or "q is necessary for p" (because q is necessarily true if p is true).

Example 5

The English statement "If it is raining, then there are clouds in the sky" is a conditional statement. It makes sense because if the antecedent "it is raining" is true, then the consequent "there are clouds in the sky" must also be true.

Notice that the statement tells us nothing of what to expect if it is not raining; there might be clouds in the sky, or there might not. If the antecedent is false, then the consequent becomes irrelevant, and no conclusions can be made.

Example 6

Suppose you order a team jersey online on Tuesday and want to receive it by Friday so you can wear it to Saturday's game. The website says that if you pay for expedited shipping, they guarantee that you receive the jersey by Friday. In what situation would the guarantee be violated?

There are four possible outcomes:

1) You pay for expedited shipping and receive the jersey by Friday.
2) You pay for expedited shipping and do not receive the jersey by Friday.
3) You do not pay for expedited shipping and receive the jersey by Friday.
4) You do not pay for expedited shipping and do not receive the jersey by Friday.

Only one of these outcomes would violate their guarantee: the second outcome, in which you pay for expedited shipping but do not receive the jersey by Friday. The first outcome is exactly what was promised, so there is no problem with that. The third outcome is not a violation of the guarantee because the website never said what would happen if you did not pay for expedited shipping; maybe the jersey would arrive by Friday whether you paid for expedited shipping or not. The fourth outcome is fine also

because, again, the website did not make any promises about when the jersey would arrive if you did not pay for expedited shipping.

It may seem strange that the third outcome in the previous example, in which the first part is false but the second part is true, is not a problem. Remember, though, that if the antecedent is false, we cannot make any judgment about the consequent. The website never said that paying for expedited shipping was the only way to receive the jersey by Friday. (In other words, paying for expedited shipping is *sufficient*, but *not necessary*, for delivery by Friday.)

Example 7

A friend tells you, "If you upload that picture to Facebook, you will lose your job." Under what conditions can you say that your friend was wrong?

There are four possible outcomes:

1) You upload the picture and lose your job.
2) You upload the picture and do not lose your job.
3) You do not upload the picture and lose your job.
4) You do not upload the picture and do not lose your job.

There is only one possible case in which you can say your friend was wrong: the second outcome, in which you upload the picture but still keep your job. In the last two cases, your friend cannot be called wrong because your friend never said anything about what would happen if you did not upload the picture. Your friend never said that you were guaranteed to keep your job. You might lose your job for many other reasons.

Therefore, in logic, we have the following truth values for conditionals. *This is very important! a conditional is false only when the antecedent is true AND the consequent is false* (the second row highlighted below). All other cases are true.

Truth Values for Conditional

A conditional $p \rightarrow q$ is **false** ONLY when the antecedent (p) is true AND the consequent (q) is false (the second row highlighted below). All other cases are **true**.

Truth table for the conditional

p	*q*	*p → q*
T	T	T
T	*F*	*F*
F	T	T
F	F	T

Simply said, this shows that the only time the proposition "If p, then q" is false is when p is true and q is false. Notice that the last two rows, where p is false, the conditional is true. In other words, the sentence "If the moon is made of green cheese, then $2 + 2 = 5$" is considered a true statement (even though the conclusion (the consequent) is not. The point is that if you start with a false assumption (antecedent), you could get away with making *any* conclusion.

A professor once said, "If I get super-hungry in my office, I eat dog food." After everyone stared at him in disbelief, he added, "It is true. I never get super-hungry in my office." Indeed, logically speaking, it was a true statement because the antecedent was never true.

Example 8

Construct a truth table for the statement (m ⋀ ~p) → r.

[Solution]

We start by constructing a truth table with 8 rows to cover all possible scenarios. Next, we can focus on the antecedent, m ⋀ ~p. Again, note that you construct the yellow column based on the green column(s) at each step.

m	p	r
T	T	T
T	T	F
T	F	T
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

m	**p**	r	*~p*
T	**T**	T	*F*
T	**T**	F	*F*
T	**F**	T	*T*
T	**F**	F	*T*
F	**T**	T	*F*
F	**T**	F	*F*
F	**F**	T	*T*
F	**F**	F	*T*

m	p	r	**~p**	*m ⋀ ~p*
T	T	T	**F**	*F*
T	T	F	**F**	*F*
T	F	T	**T**	*T*
T	F	F	**T**	*T*
F	T	T	**F**	*F*
F	T	F	**F**	*F*
F	F	T	**T**	*F*
F	F	F	**T**	*F*

Now we can create a column for the conditional. Because it can be confusing to keep track of all the truth values, we suggest that you copy the column for r to the right of the column for $m \wedge \sim p$. This makes it a lot easier to read the conditional from left to right.

m	p	r	~p	**m ∧~ p**	**r**	*(m ∧~ p) → r*
T	T	T	F	**F**	**T**	*T*
T	T	F	F	**F**	**F**	*T*
T	F	T	T	**T**	**T**	*T*
T	F	F	T	**T**	**F**	*F*
F	T	T	F	**F**	**T**	*T*
F	T	F	F	**F**	**F**	*T*
F	F	T	T	**F**	**T**	*T*
F	F	F	T	**F**	**F**	*T*

When m is true, p is false, and r is false—the fourth row of the table—then the antecedent, which is "$m \wedge \sim p$," is true but the consequent, which is "r," is false, resulting in an F. All other cases are true.

If you want a real-life situation that could be modeled by $(m \wedge \sim p) \rightarrow$ r, consider this: let m be "we order meatballs," p be "we order pasta," and r be "Rob is happy." The statement $(m \wedge \sim p) \rightarrow r$ is translated "If we order meatballs and do not order pasta, then Rob is happy." If m is true (we order meatballs), p is false (we do not order pasta), and r is false (Rob is not happy), then the statement is false because we satisfied the antecedent but Rob did not satisfy the consequent.

Converse, Inverse, and Contrapositive

For any conditional, there are three related statements: the converse, the inverse, and the contrapositive.

Statements Related to a Conditional

The original conditional is	"if p, then q"	$p \rightarrow q$
The converse is	"if q, then p"	$q \rightarrow p$
The inverse is	"if not p, then not q"	$\sim p \rightarrow \sim q$
The contrapositive is	"if not q, then not p"	$\sim q \rightarrow \sim p$

Example 9

Consider again the conditional "If it is raining, then there are clouds in the sky." It seems reasonable to assume that this is true. Write the converse, the inverse, and the contrapositive of this conditional.

[Solution]

The converse would be "If there are clouds in the sky, then it is raining." This is not always true.

The inverse would be "If it is not raining, then there are not clouds in the sky." Likewise, this is not always true (for the same reason that the converse is not true—since you can have clouds without rain).

The contrapositive would be "If there are not clouds in the sky, then it is not raining." This statement is true and is equivalent to the original conditional.

Here is another way to remember these. We begin with an original conditional, "if p, then q." Recall that p is called the antecedent while q is called the consequent. If you switch p and q, then you get the converse ("if q, then p"). If, instead, you negate each part without switching their position, then you get the inverse ("if not p, then not q").

Now, if you take the converse and then negate each part, you get the contrapositive ("if not q, then not p"). This means that the inverse of the converse is the contrapositive.

Also, if you take the inverse and then switch the parts, you also get the contrapositive ("if not q, then not p"). This means that the converse of the inverse is also the contrapositive.

Confusing? Maybe. But it may be worth noting that the "*inverse of the converse is the contrapositive, and so is the converse of the inverse*." In other words, to write the contrapositive, you switch and negate (or you negate and switch—the order does not matter).

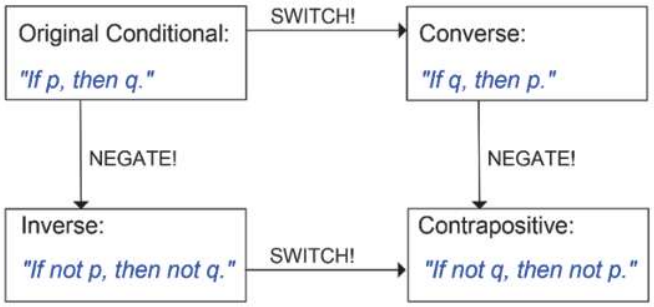


Figure 1.1

Of course, the converse of the converse is (like negatives) the original conditional itself, and the same goes for the inverse of the inverse.

Here is the interesting fact. Switching gives you something *not* logically equivalent. Negating each part also gives you something *not* logically equivalent. However, doing both of these things (or doing one of these twice) will give you something **logically equivalent** to the original sentence. It is sort of like adding a negative sign (or multiplying by -1). Doing twice would *undo* what was done. Cool.

Example 10

Write the other three statements for the original conditional: "If you are 2 months old, you are a baby."

[Solution]

- Converse (switch the two parts): "If you are a baby, then you are 2 months old."
- Inverse (negate each part): "If you are not 2 months old, you are not a baby."
- Contrapositive (do both): "If you are not a baby, then you are not 2 months old."

One warning—in logic, always start with the "if" (conditional clause). In English, you can switch these clauses to mean the same thing. For instance, "If you are 2 months old, you are a baby" means the same thing as "You are a baby if you are 2 months old." However, in formal logic, for consistency (and to avoid confusion), start every implication/conditional with the "if" clause. In other words, the antecedent should come first (as the very word "antecedent" suggests).

Looking at truth tables, we can see that the original conditional and the contrapositive are logically equivalent and that the converse and inverse are logically equivalent.

		Conditional	**Converse**	**Inverse**	*Contrapositive*
p	*q*	$p \rightarrow q$	$\mathbf{q \rightarrow p}$	$\mathbf{\sim p \rightarrow \sim q}$	$\sim q \rightarrow \sim p$
T	T	*T*	**T**	**T**	*T*
T	F	*F*	**T**	**T**	*F*
F	T	*T*	**F**	**F**	*T*
F	F	*T*	**T**	**T**	*T*

Equivalent

Logical Equivalence

A conditional statement and its **contrapositive** are logically equivalent.
The **converse** and **inverse** of a conditional statement are logically equivalent.

In other words, the original statement and the contrapositive must agree with each other; they must both be true, or they must both be false. Similarly, the converse and the inverse must agree with each other; they must both be true, or they must both be false.

Be aware that symbolic logic cannot represent the English language perfectly. For example, we may need to change the verb tense to show that one thing occurred before another.

Example 11

Suppose this statement is true: “If I eat this giant cookie, then I will feel sick.” Which of the following statements must also be true?
(1) If I feel sick, then I ate that giant cookie.
(2) If I do not eat this giant cookie, then I will not feel sick.
(3) If I do not feel sick, then I did not eat that giant cookie.

[Solution]

You can see that (1) is the converse, which is not necessarily true. I could feel sick for some other reason, such as drinking sour milk. (2) is the inverse, which is not necessarily true, either. Again, I could feel sick for some other reason; avoiding the cookie does not guarantee that I will not feel sick. (3) is the contrapositive, which must be true if the original statement is true. I had eaten the cookie, I would have felt sick; since I do not feel sick, I must not have eaten the cookie.

Notice again that the original statement and the contrapositive have the same truth value (both are true), and the converse and the inverse have the same truth value (both are false).

Try It Now 6

“If you microwave salmon in the staff kitchen, then I will be mad at you.” If this statement is true, which of the following statements must also be true?

a. **If you don’t microwave salmon in the staff kitchen, then I will not be mad at you.**
b. **If I am not mad at you, then you did not microwave salmon in the staff kitchen.**
c. **If I am mad at you, then you microwaved salmon in the staff kitchen.**

Negation of a Conditional

Consider the statement “If you park here, then you will get a ticket.” What set of conditions would prove this statement false?

- You did not park here, and you did not get a ticket.
- You did not park here, and you got a ticket anyway.
- You parked here, and you did not get a ticket.
- You parked here, and you got a ticket.

The first two statements are irrelevant because we did not know what would happen if you parked somewhere else. The third statement, however, contradicts the conditional statement “If you park here, then you will get a ticket” because you parked here but did not get a ticket. The last statement, of course, is exactly what the original conditional said would happen.

This example demonstrates a general rule. The negation of a conditional can be written as an “and” statement, i.e., a conjunction: “It is not the case that if you park here, then you will get a ticket” is equivalent to “You park here and you do not get a ticket.”

The Negation of a Conditional

The negation of a conditional statement is logically equivalent to a conjunction of the antecedent and the negation of the consequent.

$$\sim (p \to q) \text{ is equivalent to } p \wedge \sim q.$$

Note that the negation of a conditional does **not** have the word "if." It is **not** a conditional; rather, it is a conjunction. *Be sure to use the word "and" in the negation.*

Example 12

Which of the following statements is equivalent to the negation of "If you don't grease the pan, then the food will stick to it"?

(1) I didn't grease the pan and the food didn't stick to it.
(2) I didn't grease the pan and the food stuck to it.
(3) I greased the pan and the food didn't stick to it.

[Solution]

The first statement is the correct answer. To disprove the original claim that not greasing the pan will cause the food to stick, all you have to do is to *not* grease the pan and have the food *not* stick.

Note that (2) is exactly what the original statement said. (3) is like the inverse of the original conditional. If you greased the pan, you could not disprove the original statement no matter what the result is.

Try It Now 7

Consider the statement "If I go swimming, I will get wet." Which of the following statements is the negation of this statement?

a. I did not go swimming, and I got wet.
b. I did go swimming, and I did not get wet.
c. I did not go swimming, and I did not get wet.
d. I did go swimming, and I got wet.

Biconditional

In everyday life, we often have a stronger meaning in mind when we use a conditional statement. Consider "If you submit your hours today, then you will be paid next Friday." What the payroll person really means is "If you submit your hours today, then you will be paid next Friday, and if you don't submit your hours today, then you won't be paid next Friday." Such a conditional statement actually includes "*if p, then q*" and its inverse "*if not p, then not q.*" A more compact way to express this statement is "You will be paid next Friday *if and only if* you submit your hours today." A statement of this form is called a **biconditional** because the implication goes in both directions.

Biconditional

A biconditional is a conditional statement in which the antecedent and consequent are interchangeable.
A biconditional is written as $p \leftrightarrow q$ and is translated as "p if and only if q." (A biconditional is often denoted as "p *iff* q.")

Because a biconditional statement $p \leftrightarrow q$ is equivalent to $(p \rightarrow q) \wedge (q \rightarrow q)$, we may think of it as a conditional statement combined with its converse: "*if* p*, then* q" and "*if* q*, then* p." The double-headed arrow shows that the conditional statement goes from left to right and from right to left. A biconditional is considered true as long as the antecedent and the consequent have the same truth value; that is, they are either both true or both false.

Truth Table for the Biconditional

p	*q*	*p* ↔ *q*
T	*T*	*T*
T	F	F
F	T	F
F	*F*	*T*

Notice that the fourth row, where both components are false, is true; if you do not submit your hours today and you do not get paid next Friday, the person from payroll told you the truth.

Example 13

Suppose this statement is true: "The garbage truck comes down my street if and only if it is Thursday morning." Which of the following statements could be true?

(1) It is noon on Thursday, and the garbage truck did not come down my street this morning.
(2) It is Monday, and the garbage truck is coming down my street.
(3) It is Wednesday at 11:59PM, and the garbage truck did not come down my street today.

[Solution]

Statement (1) cannot be true. This is like the second row of the truth table; it is true that I just experienced Thursday morning, but it is false that the garbage truck came. Statement (2) cannot be true, either. This is like the third row of the truth table; it is false that it is Thursday morning, but it is true that the garbage truck came. Statement (3) could be true, and in fact, it must be true. This is like the fourth row of the truth table; it is false that it is Thursday morning, but it is also false that the garbage truck

came. The biconditional clearly implies that the garbage truck cannot and must not come on a Wednesday.

Try It Now 8

Suppose this statement is true: "I wear my running shoes if and only if I am exercising." Determine whether each of the following statements must be true or false.

a. I am exercising, and I am not wearing my running shoes.
b. I am wearing my running shoes, and I am not exercising.
c. I am not exercising, and I am not wearing my running shoes.
d. I am wearing my running shoes, and I am exercising.

Example 14

Create a truth table for the statement (A ∨ B) ↔ ~C.

[Solution]

Whenever we have three component statements, we start by listing all the possible truth value combinations for A, B, and C. After creating those three columns and eight rows, we can create a fourth column for the antecedent, A ∨ B. Now we will temporarily ignore the column for C and focus on A and B, writing the truth values for A ∨ B.

A	B	C
T	T	T
T	T	F
T	F	T
T	F	F
F	T	T
F	T	F
F	F	T
F	F	F

A	**B**	C	*A ∨ B*
T	**T**	T	*T*
T	**T**	F	*T*
T	**F**	T	*T*
T	**F**	F	*T*
F	**T**	T	*T*
F	**T**	F	*T*
F	**F**	T	*F*
F	**F**	F	*F*

Next, we can create a column for the negation of C. (Ignore the A ∨ B column and simply negate the values in the C column.)

A	B	**C**	A ∨ B	*~C*
T	T	**T**	T	*F*
T	T	**F**	T	*T*
T	F	**T**	T	*F*

A	B	**C**	A ∨ B	*~C*
T	F	**F**	T	*T*
F	T	**T**	T	*F*
F	T	**F**	T	*T*
F	F	**T**	F	*F*
F	F	**F**	F	*T*

Finally, we find the truth values of (A ∨ B) ↔ ~C. Remember, a biconditional is true when the truth value of the two parts match and is false when the truth values do not match.

A	B	C	**A ∨ B**	**~C**	*(A ∨ B) ↔ ~C*
T	T	T	**T**	**F**	*F*
T	T	F	**T**	**T**	*T*
T	F	T	**T**	**F**	*F*
T	F	F	**T**	**T**	*T*
F	T	T	**T**	**F**	*F*
F	T	F	**T**	**T**	*T*
F	F	T	**F**	**F**	*T*
F	F	F	**F**	**T**	*F*

To illustrate this situation, suppose your boss needs you to do either project A or project B (or both, if you have the time). If you do one of the projects, you will not get a crummy review (C is for crummy), but if you do not complete either project, then you will get a crummy review. This is a biconditional. So (A ∨ B) ↔ ~C means "You will not get a crummy review if and only if you do project A or project B." Looking at a few of the rows of the truth table, we can see how this works out. In the first row, A, B, and C are all true: you did both projects and got a crummy review, which is not what your boss told you would happen! That is why the final result of the first row is false. In the fourth row, A is true, B is false, and C is false: you did project A and did not get a crummy review. This is what your boss said would happen, so the final result of this row is true. And in the eighth row, A, B, and C are all false: you did not do either project and did not get a crummy review. This is not what your boss said would happen, so the final result of this row is false. (Even though you may be happy that your boss did not follow through on the threat, the truth table shows that your boss lied about what would happen.)

Sometimes the entire column for a statement may be true. This happens if the statement is always true. For instance, "if p, then p" and "p or not p" are always true. Such a statement is called a tautology. We will have more to say about tautologies later.

De Morgan's Laws

A contemporary of Boole's, Augustus De Morgan formalized two rules of logic that had previously been known informally. They allow us to rewrite the negation of a conjunction as a disjunction (and vice-versa).

For example, suppose you want to schedule a meeting with two colleagues Adam and Britney, and you need both of them to be available. Let "A" be the statement "Adam is available," and let "B" be the statement "Britney is available." What situation would make it impossible to have the meeting? The question is to describe $\sim (A \wedge B)$, meaning "When is it not true that both are available?" The answer is when either Adam is unavailable **or** Britney is unavailable (or both): $\sim A \vee \sim B$. In other words, the negation of a conjunction is the disjunction of the two negations. The negation of a disjunction is similar.

De Morgan's Laws

1. The negation of a conjunction is equivalent to the disjunction of the negation of the statements making up the conjunction. To negate an "and" statement, negate each part and change the "and" to "or."

   $$\sim (p \wedge q) \text{ is equivalent to } (\sim p) \vee (\sim q).$$

2. The negation of a disjunction is equivalent to the conjunction of the negation of the statements making up the disjunction. To negate an "or" statement, negate each part and change the "or" to "and."

   $$\sim (p \vee q) \text{ is equivalent to } (\sim p) \wedge (\sim q).$$

Example 15

For Valentine's Day, you got your partner neither flowers nor candy. Which of the following statements is logically equivalent to this statement?

(1) Either you did not get your partner flowers or you did not get your partner candy.
(2) You did not get your partner flowers, and you did not get your partner candy.
(3) You got your partner flowers, or you got your partner candy.

[Solution]

Statement (1) does not go far enough; it leaves open the possibility that you got your partner one of the two things. Statement (2) is the correct answer since it is equivalent to the original; $\sim(f \vee c)$ is equivalent to $\sim f \wedge \sim c$. Statement (3) says that you got your partner something, but we know that you did not.

Try It Now 9

To serve as the President of the United States, a person must have been born as a U.S. citizen and must be at least 35 years old. Based on these two eligibility requirements, what minimum set of conditions would disqualify someone from serving as President?

Exercises 1.3

1. Translate each statement from symbolic notation into English sentences. Let A represent "Elvis is alive" and let G represent "Elvis gained weight."
   (a) $A \vee G$
   (b) $\sim (A \wedge G)$
   (c) $G \rightarrow \sim A$
   (d) $A \leftrightarrow \sim G$
   (e) $A \wedge \sim G$
   (f) $\sim (A \vee G)$

2. Create a truth table for each statement below.
   (a) $A \wedge \sim B$
   (b) $\sim (\sim A \vee B)$
   (c) $(A \wedge B) \rightarrow C$
   (d) $(A \vee B) \rightarrow \sim C$

3. Complete the truth table for $(A \vee B) \wedge \sim (A \wedge B)$.

A	B	$A \vee B$	$A \wedge B$	$\sim (A \wedge B)$	$(A \vee B) \wedge \sim (A \wedge B)$
T	T				
T	F				
F	T				
F	F				

4. We have been studying the *inclusive* or, which allows both A and B to be true. The *exclusive or* does not allow both to be true; it translates to "either A or B, but not both." For each situation, decide whether the "or" is most likely exclusive or inclusive.
   (a) An entrée at a restaurant includes soup or a salad.
   (b) You should bring an umbrella or a raincoat with you.
   (c) We can keep driving on I-5 or get on I-405 at the next exit.
   (d) Use Gate 1 if you are at least 35 years old, or Gate 2 if you are younger.
   (e) You should save this document on your computer or a flash drive.
   (f) I am not sure if my pregnant wife is going to have a boy or a girl.

5. Consider the statement “If you are under age 17, then you cannot watch this movie.”
   (a) Write the converse.
   (b) Write the inverse.
   (c) Write the contrapositive.

6. Consider the statement “If you have a house in Beverly Hills, you are rich.”
   (a) Write the converse.
   (b) Write the inverse.
   (c) Write the contrapositive.

7. Assume that the statement “If you cheat, you will get a discipline referral” is true. Which of the following statements must also be true?
   (a) If you don’t cheat, then you won’t get a discipline referral.
   (b) If you don’t get a discipline referral, then you did not cheat.
   (c) If you get a discipline referral, then you cheated.

8. Write the negation: If Luke faces Vader, then Obi-Wan cannot interfere.

9. Write the negation: If you look both ways before crossing the street, then you will not get hit by a car.

10. Write the negation: If you weren’t talking, then you wouldn’t have missed the instructions.

11. Write the negation: If you score a goal now, we will win.

12. Assume that the biconditional statement “You will play in the game if and only if you attend all practices this week” is true. Which of the following situations could NOT happen?
   (a) You attended all practices this week and didn’t play in the game.
   (b) You didn’t attend all practices this week and played in the game.
   (c) You didn’t attend all practices this week and didn’t play in the game.

13. Use De Morgan’s Laws to rewrite the disjunction as a conjunction: It is not true that Tina likes Sprite or 7-Up.

14. Use De Morgan’s Laws to rewrite the disjunction as a conjunction: It is not true that the father or the mother of that child will be required to testify.

15. Use De Morgan’s Laws to rewrite the conjunction as a disjunction: It is not the case that both the House and the Senate passed the bill.

16. Use De Morgan's Laws to rewrite the conjunction as a disjunction: It is not true that I need both a car and a house.

Figure 1.5[11]

[11] Image by Tony Hoffarth is licensed under CC BY-NC-ND 2.0

1.4 Analyzing Arguments

A logical argument is a process of reaching a conclusion based on a set of statements called premises. There are two general types of arguments: inductive and deductive arguments. You may have heard of these in your English, speech, or philosophy classes in the context of persuasive communication.

Types of Arguments

Argument Types

An **inductive** argument uses a collection of specific examples as its premises and uses them to propose a general conclusion.
A **deductive** argument uses a collection of general statements as its premises and uses them to propose a specific conclusion.

In short, "inductive" goes from specific to general while "deductive" goes from general to specific.

Example 1

Consider the following argument: "When I went to the store last week, I forgot my purse. When I went today, I forgot my purse. I always forget my purse when I go to the store." This is an inductive argument.

The premises:

- I forgot my purse last week.
- I forgot my purse today.

The conclusion:

- I always forget my purse.

Notice that the premises are specific situations while the conclusion is a general statement. In this case, this is a fairly weak argument since it is based on only two instances.

Example 2

Consider the following argument: "Every day for the past year, a plane flies over my house at 2:00 P.M. So a plane will fly over my house today at 2:00 P.M." This is a stronger inductive argument since it is based on a larger set of evidence. While it is not necessarily true (the airline may have cancelled its afternoon flight today), it is probably a safe bet.

Evaluating Inductive Arguments

An inductive argument is never able to prove the conclusion true, but it can provide either weak or strong evidence to suggest that it may be true.

Many scientific theories, such as the Big Bang theory, can never be proven completely (partly because no one was there to observe it). Instead, there are inductive arguments that support it with a wide variety of evidence. Usually in science, an idea is considered a hypothesis until it has been well tested, at which point it graduates to being considered a theory. Common scientific theories, like Newton's theory of universal gravity, have all stood up to years of testing and evidence though sometimes they need to be adjusted based on new evidence, such as when Einstein proposed the theory of general relativity.

A deductive argument is more clearly valid or not, which makes it easier to evaluate.

Evaluating Deductive Arguments

A deductive argument is considered **valid** if the conclusion follows logically from the premises. Validity depends on the structure of the argument itself, not the truth values or meaning of the premises.

If a deductive argument is valid and, in addition, the premises are all true, the conclusion will necessarily be true. In this case, we say the argument is **sound**.

Here is an important point; a deductive argument can be "**valid**" without the conclusion being correct. As stated above, validity has to do with the structure of the argument (explained fully below). **Validity** means that *the conclusion is a logical consequence of the premises*. In other words, if the premises are all true in a valid deductive argument, then logically the conclusion must be true (it is then a **sound** argument). Therefore, a sound argument always has a correct conclusion. Be sure to distinguish "valid" and "sound" when analyzing deductive arguments. We now look more deeply into deductive arguments.

Evaluating Deductive Arguments with Venn Diagrams

We can interpret a deductive argument visually with a Venn diagram. This can make it easier to determine whether the argument is valid or invalid.

Example 3

Consider the deductive argument "All cats are mammals. A tiger is a cat. So a tiger is a mammal." Is this argument valid?

The premises:

- All cats are mammals.
- A tiger is a cat.

The conclusion:

- A tiger is a mammal.

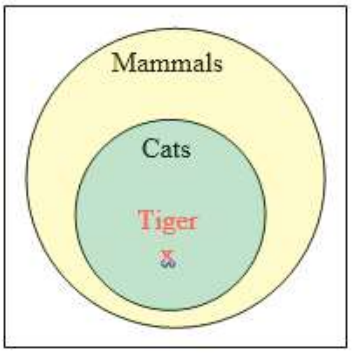


Figure 1.2[12]

Both premises happen to be true. But remember, that fact has nothing to do with the validity of the argument! To see that the premises must logically lead to the conclusion, we can use a **Venn diagram**. From the first premise, we draw the set of cats (the inner circle) as a subset of the set of mammals (the outer circle). From the second premise, we are told that a tiger is contained within the set of cats (the inner circle). From that, we can see in the Venn diagram that the tiger must also be inside the set of mammals (since everything inside the inner circle must also be in the outer circle), so the conclusion is valid.

Note that the sentence "All A are B" is represented by two circles, where A is the inner circle and B the outer. This is a Venn diagram describing the subset relation; "A is a subset of B" when every member of A also belongs to B. You can see that premises with "all," "every," and "each" (positive universal quantifiers) can always be represented this way, and this is a key to evaluating the validity of a deductive argument.

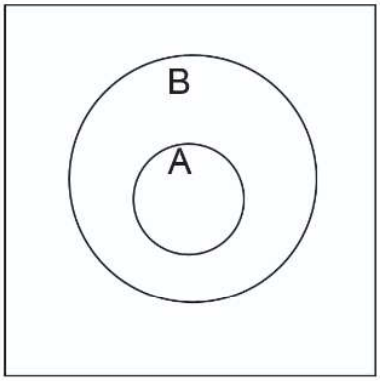


Figure 1.3

[12] Logic by David Lippman is licensed under CC BY-SA 3.0 US.

Analyzing Arguments with Venn Diagrams

1) Draw a Venn diagram based on the premises of the argument.
2) The argument is **valid** if the diagram cannot be drawn to make the conclusion false.
3) The argument is **invalid** if there is a way to draw the diagram that makes the conclusion false.

In other words, If the premises are insufficient to determine the location of an element in the conclusion, then the argument is invalid.

Try It Now 10

Determine the validity of this argument:
Premise: **All cats are scared of vacuum cleaners.**
Premise: **Max is a cat.**
Conclusion: **Max is scared of vacuum cleaners.**

Example 4

Premises:

- All firefighters know CPR.
- Jill knows CPR.

Conclusion:

- Jill is a firefighter.

From the first premise, we know that all firefighters lie inside the set of those who know CPR. (Firefighters are a subset of people who know CPR.) Whether this is true or not does not matter in determining logical validity here (although it matters greatly in saving lives). From the second premise, we know that Jill is a member of that larger set, but we do not have enough information to know whether she is also a member of the smaller subset that is firefighters.

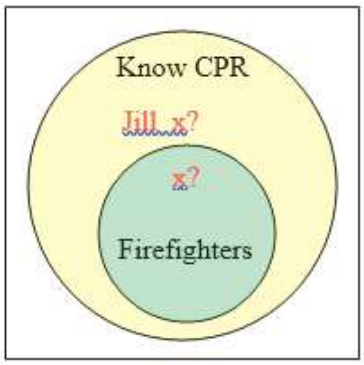


Figure 1.4[13]

[13] Logic by David Lippman is licensed under CC BY-SA 3.0 US.

Since the conclusion does not necessarily follow from the premises, this is an invalid argument. It is possible that Jill is a firefighter, but the structure of the argument does **not** allow us to conclude that she definitely is. In fact, Jill may, in real life, be a firefighter, but this argument is still invalid (and thus cannot be sound).

It is important to note that whether or not Jill is actually a firefighter is irrelevant in evaluating the validity of the argument; we are concerned with whether the premises are enough to prove the conclusion. Remember, validity has to do with the logical structure, not the truthfulness of the premises or the conclusion.

Try It Now 11

Determine the validity of this argument:
Premise: **All bicycles have two wheels.**
Premise: **This Harley-Davidson has two wheels.**
Conclusion: **This Harley-Davidson is a bicycle.**

Now, if you have a *negative* universal quantifier ("no," "none," etc.), then the Venn diagram would look different. Instead of one circle inside of the other, you would draw two *separate* circles. The sentence "No A are B" then is represented as two circles with nothing in common. We call two sets **disjoint** or **mutually exclusive** if their intersection is empty. If x is in A, then it cannot be in B. If x is in B, then it cannot be in A. However, if x is *not* in A, we cannot determine if it is in B or not. Use the diagram below to see if these sentences make sense. With this, try the following example.

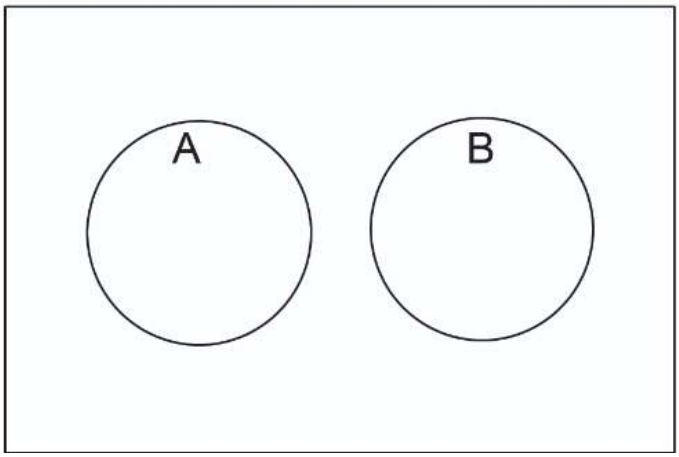


Figure 1.5

Try It Now 12

Determine the validity of this argument:
Premise: **No cows are purple.**
Premise: **Fido is not a cow.**
Conclusion: **Fido is not purple.**

In addition to these categorical style premises of the form "all ___", "some ____", and "no ____", it is also common to see premises that are conditionals.

Example 5

Premises:

- If you live in Seattle, you live in Washington (the state).
- Marcus does not live in Seattle.

Conclusion:

- Marcus does not live in Washington.

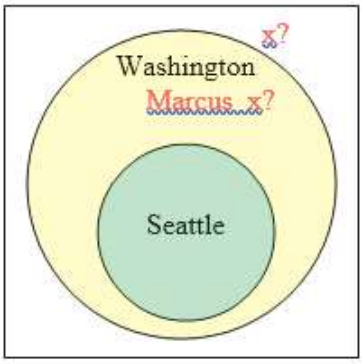


Figure 1.6[14]

From the first premise, we know that the set of people who live in Seattle is inside the set of those who live in Washington. From the second premise, we know that Marcus does not lie in the Seattle set, but we have insufficient information to know whether Marcus lives in Washington or not. In other words, this is an invalid argument.
If Marcus actually lives in New York, the conclusion would be true (and so would the premises), but that does not make the argument valid.

You may have noticed that the same diagram can illustrate "All A are B" and "If A, then B." That is because, logically speaking, these two sentences are equivalent. Compare the statements "All cats are mammals" and "If something is a cat, then it is a mammal." In other words, every conditional statement (implication) describes a subset relation and can be represented by two nested circles.

[14] Logic by David Lippman is licensed under CC BY-SA 3.0 US.

Try It Now 13

Determine the validity of this argument:
Premise: **If you cheat on the final, you will be expelled.**
Premise: **You did not cheat on the final.**
Conclusion: **You will not be expelled.**

Forms of Valid Arguments

Arguments can actually be analyzed using truth tables, but this can be a lot of work. What one would do is to represent each premise symbolically, create an appropriate conditional, and examine the truth values of the complex statement which has the conclusion as its consequent. The argument is then valid if (and only if) the entire complex statement is always true (an entire column of T's).

Recall that a statement that is always true is called a tautology. A deductive argument is valid if and only if it is a tautology. If there is any chance (any row in the truth table) that it could be false, then the argument is not valid.

Rather than making a truth table for every argument, we may be able to recognize certain common forms of arguments that are valid. If we can determine that an argument fits one of the common forms, we can immediately state whether it is valid or invalid.

The Law of Detachment (*Modus Ponens)*

The Law of Detachment applies when a conditional and its antecedent are given as premises, and the consequent is the conclusion. The general form is:

Premise: $p \rightarrow q$
Premise: p
Conclusion: q

The Latin name, *modus ponens*, translates to "mode that affirms."

The first premise states that p is a sufficient condition for q to be true, and the second premise states that p is true. It is then easy to see that q has to be true. Another way to see this is the Venn diagram. The p circle is inside of the q (outer) circle. If something is in p (the inner circle), it has to be in q (the outer circle). Naturally, this is a valid argument.

Example 6

Analyze the following deductive argument.

Premises:

- If you bought bread, then you went to the store.
- You bought bread.

Conclusion:

- You went to the store.

[Solution]

In symbolic form,
Premise: $b \rightarrow s$
Premise: b
Conclusion: s

This argument has the structure described by the Law of Detachment. (The second premise and the conclusion are simply the two parts of the first premise detached from each other.) Instead of making a truth table, we can say that this argument is valid by the Law of Detachment.

By the way, you may be saying (correctly), "Yes, but that is not true! You can buy bread without going to the store. You could have someone else go to the store, or you can order bread and have it delivered." You are absolutely right. You can thus argue that this argument is false (based on your claim that the first premise is false). However, note that this argument is absolutely valid based on the structure. If the premises are correct and the argument is valid, then the conclusion **must** be true. If a premise is false, however, then even a valid argument could produce a false conclusion. It is very important, therefore, to remember the difference between validity and soundness.

Now we move on to the second way to recognize a valid argument—based on contrapositive.

The Law of Contraposition (*Modus Tollens)*
The Law of Contraposition applies when a conditional and the negation of its consequent are given as premises, and the negation of its antecedent is the conclusion. The general form is:

Premise: $p \rightarrow q$
Premise: $\sim q$
Conclusion: $\sim p$

The Latin name, *modus tollens*, translates to "mode that denies".

Notice that the second premise and the conclusion look like the contrapositive of the first premise, $\sim q \rightarrow \sim p$, but they have been detached. You can think of the Law of Contraposition as a combination of the Law of Detachment and the fact that the contrapositive is logically equivalent to the original statement.

Example 7

Premises:

- If I drop my phone into the swimming pool, my phone will be ruined.
- My phone is not ruined.

Conclusion:

- I did not drop my phone into the swimming pool.

Let "*d*" be "I drop the phone in the pool" and "*r*" be "The phone is ruined." Then we can represent the argument this way:

Premise $d \rightarrow r$
Premise $\sim r$
Conclusion: $\sim d$

The form of this argument matches what we need to invoke the Law of Contraposition, so it is a valid argument.

Try It Now 14

Is this argument valid?
Premise: If you just brushed your teeth, then your toothbrush is wet.
Premise: Your toothbrush is not wet.
Conclusion: You did not just brush your teeth.

Some arguments involve more than two propositions as illustrated by the following.

The Transitive Property (Hypothetical Syllogism)

The **Transitive Property** has as its premises a series of conditionals, where the consequent of one is the antecedent of the next. The conclusion is a conditional with the same antecedent as the first premise and the same consequent as the final premise. The general form is:

Premise: $p \rightarrow q$
Premise: $q \rightarrow r$
Conclusion: $p \rightarrow r$

This describes a chain reaction; if the first thing happens, then the second thing happens, and if the second thing happens, then the third thing happens. Therefore, we can say that if the first thing happens, then the third thing will happen. We can simply say that the first event leads to the final event. We could even have more than two premises; as long as they form a chain reaction, the transitive property will give us a valid argument.

Example 8

Analyze the following argument.

Premises:

- If a soccer player commits a reckless foul, she will receive a yellow card.
- If Hayley, a soccer player, receives a yellow card, she will be suspended for the next match.

Conclusion:

- If Hayley commits a reckless foul, she will be suspended for the next match.

If we let "*r*" be "committing a reckless foul," "*y*" be "receiving a yellow card," and "*s*" be "being suspended for the next match," then the argument looks like this:

Premise $r \rightarrow y$
Premise $y \rightarrow s$
Conclusion: $r \rightarrow s$

This argument has the exact structure required to use the Transitive Property, so it is a valid argument.

Try It Now 15

Is this argument valid?
Premise: **If you arrive after 2 pm, you will miss the train.**
Premise: **If you miss the train, you will miss the job interview.**
Premise: **If you miss the job interview, you will not get this job.**
Premise: **If you do not get this job, you cannot pay your rent.**
Premise: **If you cannot pay your rent, you may be evicted.**
Conclusion: **If you arrive after 2 pm, you may be evicted.**

Disjunctive Syllogism

In a disjunctive syllogism, the premises consist of an *or* statement and the negation of one of the options. The conclusion is the other option. The general form is:

Premise: $p \vee q$
Premise: $\sim p$
Conclusion: q

The order of the two parts of the disjunction is not important. In other words, we could have the premises $p \vee q$ and $\sim q$, and the conclusion p. This is essentially saying the following: "Either p or q must be true, and we know one is not. Therefore, the other must be true."

Example 9

Premises:

- I will either drive or take the train.
- I will not drive.

Conclusion:
- I will take the train.

If we let "*d*" be "I will drive" and "*t*" be "I will take the train," then the symbolic representation of the argument is:

Premise $d \vee t$
Premise $\sim d$
Conclusion: t

This argument is valid because it has the form of a disjunctive syllogism. I have only two choices, and one of them is not going to happen, so the other one must happen.

Try It Now 16

Is this argument valid?
Premise: Alison was required to write a 10-page paper or give a 5-minute speech.
Premise: Alison did not give a 5-minute speech.
Conclusion: Alison had to write a 10-page paper.

Keep in mind that, when you are determining the validity of an argument, you need to keep your personal opinion out of it. Your job is to pretend that the premises are true and then determine whether they force you to accept the conclusion. You may attack the premises in a court of law or a political discussion, of course, but here we are focusing on the structure of the arguments, not the truth of what they actually say.

In fact, you can check the validity of an argument even if you have no idea what the premises are saying as long as you can identify the logical structure of the argument. The exercises include some of these arguments.

Forms of Logical Fallacies

We have just looked at four forms of valid arguments; there are two common forms that represent invalid arguments, which are also called fallacies. Detecting these logical fallacies is a very important skill as these are quite prevalent in politics, advertisements, and perhaps our daily conversations.

The Fallacy of the Converse

The fallacy of the converse arises when a conditional and its consequent are given as premises and the antecedent is the conclusion. The general form is:

Premise:	$p \rightarrow q$
Premise:	q
Conclusion:	p

Notice that the second premise and the conclusion look like the converse of the first premise, $q \rightarrow p$, but they have been detached. The fallacy of the converse incorrectly tries to assert that the converse of a statement is equivalent to that statement. This is also called the fallacy of **affirming the consequent**.

Example 10

Analyze the following argument.
Premises:

- If I drink coffee after noon, then I have a hard time falling asleep that night.
- I had a hard time falling asleep last night.

Conclusion:

- I drank coffee after noon yesterday.

If we let "*c*" be "I drink coffee after noon" and "*h*" be "I have a hard time falling asleep," then our argument looks like this:

Premise	$c \rightarrow h$
Premise	h
Conclusion:	c

This argument uses converse reasoning, so it is an invalid argument. There could be plenty of other reasons why I could not fall asleep. I could have been worried about money, or my neighbors might have been setting off fireworks.

Try It Now 17

Is this argument valid?

Premise:	**If you pull that fire alarm, you will get in big trouble.**
Premise:	**You got in big trouble.**
Conclusion:	**You must have pulled the fire alarm.**

The Fallacy of the Inverse

The fallacy of the inverse occurs when a conditional and the negation of its antecedent are given as premises and the negation of the consequent is the conclusion. The general form is:

Premise:	$p \rightarrow q$
Premise:	$\sim p$
Conclusion:	$\sim q$

Again, notice that the second premise and the conclusion look like the inverse of the first premise, ~p → ~q, but they have been detached. The fallacy of the inverse incorrectly tries to assert that the inverse of a statement is equivalent to that statement. This is also called the fallacy of **denying the antecedent**.

Example 11

Premises:

- If you listen to vinyl records, you are old.
- Sky does not listen to vinyl records.

Conclusion:

- Sky is not old.

If we let "*v*" be "listen to vinyl records and "*o*" be "is old," then this is the argument:

Premise	$v \rightarrow o$
Premise	$\sim$v
Conclusion:	$\sim o$

This argument is invalid because it uses inverse reasoning. The first premise does not imply that all old people listen to vinyl records; there could be some old people who do not, and Sky may just be one of them.

Try It Now 18

Is this argument valid?

Premise:	**If a player trips an opponent, He will be given a penalty.**
Premise:	**Diego did not trip an opponent.**
Conclusion:	**Diego will not be given a penalty.**

Of course, arguments are not limited to these six basic forms; some arguments have more premises or premises that need to be rearranged before you can see what is really happening. There are plenty of other forms of arguments that are invalid. If an argument does not seem to fit the pattern of any of these common forms, though, you may want to use a Venn diagram or a truth table instead.

Lewis Carroll, author of *Alice's Adventures in Wonderland*, was a mathematics and logic teacher and wrote two books on logic. In those books, he proposed premises as a puzzle, to be connected using syllogisms. The following example is one such puzzle.

Example 12

Solve the puzzle. In other words, find a logical conclusion from these premises.

All babies are illogical.

Nobody is despised who can manage a crocodile.

Illogical persons are despised.

Let "*b*" be "is a baby," "*d*" be "is despised," "*I*" be "is illogical," and "*m*" be "can manage a crocodile."

Then we can write the premises as:

$b \rightarrow i$

$m \rightarrow \sim d$ (If you can manage a crocodile, you are not despised.)

$i \rightarrow d$

Using the Transitive Property with the first and third premises, we can conclude that b → d; that all babies are despised. Using the contrapositive of the second premise, d → ~m, we can then use the Transitive Property with b → d to conclude that b → ~m; that babies cannot manage a crocodile. While it is silly, this is a logical conclusion from the given premises.

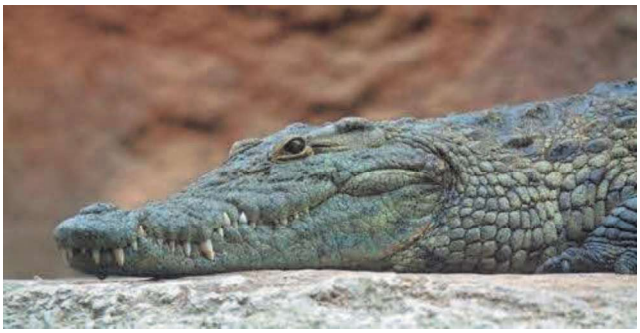


Figure 1.7[15]

Of course, the contrapositive of this conclusion is equivalently true: "if you can manage a crocodile, then you are not a baby.

[15] Image by Leigh Bedford is licensed under CC BY 2.0

Example 13

Premise: If I work hard, I'll get a raise.
Premise: If I get a raise, I'll buy a boat.
Conclusion: If I don't buy a boat, I must not have worked hard.

If we let "*h*" be "working hard," "*r*" be "getting a raise," and "*b*" be "buying a boat," then we can represent our argument symbolically:

Premise: $h \rightarrow r$

Premise: $r \rightarrow b$

Conclusion: $\sim b \rightarrow \sim h$

Using the Transitive Property with the two premises, we can conclude that h → b; if I work hard, then I will buy a boat. We also know that the conditional statement h → b is equivalent to its contrapositive ~b → ~h. Therefore, the conclusion is indeed a logical syllogism derived from the premises.

Try It Now 19

Is this argument valid?
Premise: If I go to the party, I will be really tired tomorrow.
Premise: If I go to the party, I will get to see friends.
Conclusion: If I don't see friends, I will not be tired tomorrow.

Exercises 1.4

1. Determine whether each of the following is an inductive or deductive argument.
   (a) We tried that new medicine on 100 patients, and all of them were cured. So the new medicine works.
   (b) Every student has an ID number. Sandra is a student, so she has an ID number.
   (c) Every angle of a rectangle is 90 degrees. A soccer field (pitch) is rectangular, so every corner is 90 degrees.
   (d) Whenever someone throws things up in the air, it seems to come down. Thus, what goes up must come down.
   (e) Every time it rains, my grass grows fast. Rain speeds up the growth of grass.
   (f) All the people I know in Europe seem to live for a long time. The Europeans must have a long life span.

2. Analyze the validity of the argument. If valid, is it sound, too?

   Everyone who gets a degree in science will get a good job. I got a science degree. Therefore, I will get a good job.

3. Analyze the validity of the argument. If valid, is it sound, too?

   If someone turns off the switch, the lights will be off. The lights are off. Therefore, someone must have turned off the switch.

4. Suppose the statement: "If today is Dec. 25, then the store is closed."
   (a) "Today is Dec. 25. Thus, the store is closed." Which property was used? Is this valid?
   (b) "The store is not closed. So today is not Dec. 25." Which property was used? Is this valid?
   (c) "The store is closed. Therefore, today must be Dec. 25." Which property was used? Is this valid?
   (d) "Today is not Dec. 25. Thus, the store is open." Which property was used? Is this valid?

5. Is this a valid argument? Explain. "Every student brought a pencil or a pen. Marcie brought a pencil. Therefore, Marcie did not bring a pen."

6. Is this a valid argument? Explain. "Jamie must scrub the toilets or hose down the garbage cans. Jamie refuses to scrub the toilets. Therefore, Jamie will hose down the garbage cans."

For the following questions, use a **Venn diagram** or a truth table to determine the validity.

7. Every person on this reality show is self-absorbed. Laura is not self-absorbed. Therefore, Laura cannot be on this reality show.

8. If you are a triathlete, then you have outstanding endurance. LeBron James is not a triathlete. Therefore, LeBron does not have outstanding endurance.

9. If a creature is a chimpanzee, then it is a primate. If a creature is a primate, then it is a mammal. Bobo is a mammal. Therefore, Bobo is a chimpanzee.

10. Every cripsee is a domwow. Mekep is not a domwow. Therefore, Mekep is not a cripsee. (This sentence has a lot of made-up words, but it is still possible to check for the validity of the argument. This is a good practice for abstract thinking.)

11. Whoever touches a kipoc will be bopied. I did not touch any kipoc. Therefore, I will not be bopied. (Again, you do not need to know the meaning of each word to do this exercise.)

1.5 Problem Solving[ii]

Unlike exercises, there is never a simple recipe for solving a problem. You can get better and better at solving problems, both by building up your background knowledge and by simply practicing. As you solve more problems (and learn how other people solved them), you learn strategies and techniques that can be useful. However, no single strategy works every time.

Pólya's How to Solve It

George Pólya was a great champion in the field of teaching effective problem-solving skills. A professor at Stanford University (among other universities), he wrote many mathematical papers along with three books, most famously, *How to Solve It* in 1945. This book gives a four-step method for solving mathematical problems:

1. Understand the problem.
2. Devise a plan.
3. Carry out the plan.
4. Review and extend.

The first step is the understanding and the identification of the problem. After all, if you do not know what the problem is, you will not be able to solve it. The second is the planning stage; this may require creativity, thinking outside the box, and views from different angles. Then, you execute that plan or plans in the third stage. Here, you may have to revise the plan (if necessary). Be flexible and patient. Then, once you find the solution, take time to reflect, and then ask more questions to possibly extend your results.

This is easy to state in general, but how do you actually do these steps? Steps 1 and 2 may be particularly mysterious! How do you "make a plan?" That is where you need some tools in your toolbox, and some experience to draw upon.

Much has been written since 1945 to explain these steps in more detail, but the truth is that they are more *art* than *science*. This is where mathematics becomes a creative endeavor (and where it becomes so much fun). We will articulate some useful problem-solving strategies, but no such list will ever be complete. This is really just a start to help you on your way. The best way to become a skilled problem solver is to learn the background material well and then to solve a lot of problems!

Strategy 1: Guess and Test (Trial and Error)

Make a guess and test to see if it satisfies the demands of the problem. If it does not, alter the guess appropriately and check again. Keep doing this until you find a solution that works.

Example 1

Mr. Jones has chickens and cows, a total of 25 of those animals, on his farm. How many of each does he have if the total number of feet is 76? (Here, "feet" does not represent any length.)

Step 1: Understanding the problem.
We are given in the problem that there are 25 chickens and cows.
All together there are 76 feet.
Chickens have 2 feet and cows have 4 feet.
We are trying to determine how many cows and how many chickens Mr. Jones has on his farm.

Step 2: Devise a plan.
We are going to use "trial-and-error" along with making a table, which helps you look for a pattern. Here is the procedure: create a table reflecting the data in the problem. If done in an orderly way, such a table will often reveal patterns and relationships that suggest how the problem can be solved.

Step 3: Carry out the plan.
Below, on each row, we ensure that the number of chickens and the number of cows add up to the given total of 25. For instance, my first guess may be 20 chickens and 5 cows.

Chickens	Cows	Number of chicken feet	Number of cow feet	Total number of feet
20	5	40	20	60
21	4	42	16	58

Notice we are going in the wrong direction! The total number of feet is decreasing!

Chickens	Cows	Number of chicken feet	Number of cow feet	Total number of feet
19	6	38	24	62

Better! The total number of feet is increasing!

Chickens	Cows	Number of chicken feet	Number of cow feet	Total number of feet
15	10	30	40	70
12	13	24	52	76

We got it!

Step 4: Review and extend.
It appears that Mr. Jones has 12 chickens and 13 cows. There are 12 + 13 = 25 heads, and 24 + 52 = 76 feet, indeed. It is always good to check your answer just to be sure. We have found a solution to this problem. I could use this strategy when there are a limited number of possible answers. It was relatively easy to guess and adjust our answer.

But now that you have finished the problem, stop and reflect for a moment. What made you stop after moving from 20 chickens to 21 chickens? Because the total number of feet went down (from 60 to 58). Why did that happen? You may have noticed that when a cow (with four feet each) is replaced by a chicken (with two each), then the number of feet clearly goes down by 2. So when the total number of feet was 60 at the beginning, you could have guessed that you have to increase the number of cows to get more feet. In fact, each time you replace a chicken with a cow, you gain 2 feet, and $76 - 60 = 16$ (we needed 16 more feet than the initial guess), so you could have directly guessed that you needed to replace 8 of those 20 chickens with cows. We started with 20 chickens and 5 cows, so this adjustment would make it 12 chickens and 13 cows, which is the correct answer.

While this type of reflection may occur after you have found your solution, recognizing such a pattern is helpful in seeing the problem much more clearly and in solving similar problems. You could even generalize the pattern and come up with a formula. This is the fun and elegant part of problem solving. After all, we humans all get great joy out of "getting it" (experiencing the "Aha!" moment or "Eureka!" moment[16]).

Remember, there is no formula or standard procedure here. Try and adjust. You will soon discover a few strategies along the way (what not to do, how to do it better, etc.). Do not forget to reflect on the problem and your solution after solving it.

[16] "Eureka" is a Greek word meaning "I have found/discovered." The use of this word in English is based on a famous story of Archimedes (and his discovery involving buoyancy and a bathtub) some 2200 years ago.

Try It Now 20

Mrs. Lee has chickens and goats in the barnyard—the combined total of 250 animals. Together, there are 640 feet. How many of each animal does she have?

Strategy 2: Think Outside the Box

Carl Friedrich Gauss (1777-1855) is one of the greatest mathematicians of all times, making lasting contributions in algebra, geometry, probability theory, number theory, and even astronomy. When he was in elementary school, his classroom teacher got upset because the class was not behaving well, and the entire class was ordered to sit down and add up all the numbers (meaning integers) from 1 to 100. Boys and girls in the class reluctantly began: 1 + 2 = 3, 1 + 2 + 3 = 6, and so on. No one finished it—except for the little Carl, who kept thinking and thinking, without jotting down anything. In a few minutes, Carl wrote "5050" and told his teacher. The answer was, of course, correct.

This story may or may not be true (the authenticity of this legendary tale is somewhat controversial), but it is reported that Gauss's thought process was as follows. Everyone started to add from the beginning to the end (but never reached the answer as the problem gets more and more time-consuming). Instead, Gauss thought of adding the first and the last: 1 + 100 = 101. Then, the second and the second-from-the last: 2 + 99 = 101. Then, the third and the third-from-the-last: 3 + 98 = 101. Clearly, each sum is 101. How many 101s are there? Since there are 100 numbers in total, there must be 50 pairs (the last one being 50 + 51 = 101). So Gauss could simply multiply 101 by 50 to get 5050.

This example illustrates the importance (and cleverness) of thinking outside the box. Try to look at a problem from various points of view. Try something others have not thought about.

Many people mistakenly think that mathematicians are good at arithmetic, sort of like super computers. However, the truth is that many professional mathematicians admit that they are not good at calculations. There are well-known stories about groups of mathematicians going out to dinner and not being able to split the bill correctly or calculate appropriate tips. This is because fundamentally mathematics is not about calculations but about recognizing clever patterns, looking at things from different perspectives, formulating conjectures, and proving statements. (Statements proved true are called theorems.) Gauss, like many other famous mathematicians, had creative problem-solving skills.

Example 2

Suppose you have a list of numbers: 2, 5, 8, ... Find the 200^{th} number.

[Solution]

You immediately notice that you can get the next number by adding 3. So you *could* try to add 3 repeatedly until you find the 200^{th} number. But is there a better way to do this? Let us figure out the pattern. The common difference is 3 as noted already. The

first number is 3 – 1. The second is 6 – 1, and that is $2 \times 3 - 1$. The third is 9 – 1, and that is $3 \times 3 - 1$. Then the 10th number must be $10 \times 3 - 1 = 29$. Is that right? Let us check directly: 2, 5, 8, 11, 14, 17, 20, 23, 26, 29, … The 10th number is indeed 29. So this pattern should work. In other words, to find the *n*-th number, you multiply *n* by 3 and then subtract 1. The 100th number must then be $100 \times 3 - 1 = 299$, and the 200th number must be $200 \times 3 - 1 = 599$. The answer is 599.

Reflect on our thinking process for a moment. The "thinking-outside-the-box" part of this question could have been reached if you thought of adding 1 to every number in the sequence. Adding 1 would have produced a new sequence, 3, 6, 9, …, a rather familiar list (positive multiples of 3). This new sequence could have easily given us the method we used above to find the answer.

Try It Now 21

Here is another list of numbers: 7, 11, 15, 19, … Find the 300th number in this sequence.

Example 3

Here is an example where the question sounds super-hard but the answer is astonishingly (almost magically) simple.

Pick five cities (or five locations) anywhere on the earth. Assuming that the earth is a perfect sphere, describe a way to find a hemisphere (exactly half of the sphere, including the boundary circle) that contains at least four of the five locations chosen. Note: This is possible regardless of the configuration of the five locations chosen.

[Solution]

You may want to start with some examples, like picking five U.S. cities or five European cities. But these are easy since all of those cities are on the Northern Hemisphere (with the Equator as its boundary). It seems much more difficult if the five locations are spread apart, all around the globe.

You probably realized that an abstract form of the same question involves a simple sphere and five random points on the sphere. Here, "thinking outside the box (or the sphere in this case)" may be to start with a fewer number of points. What if there were only two points? Then, simply connect those two points and extend the line to a "great circle" (an "equator-like" length-minimizing circle) through the two points. Such a circle exists for any pair of points (there are infinitely many if the two points are exactly on the opposite sides of the sphere). Now, this circle defines two hemispheres, and each of those hemispheres would, by definition, contain the two points on its boundary. Where are the other three points? Well, one of those hemispheres must contain at least two (maybe all three), and that is the hemisphere with at least four of the points picked at the beginning.

This question, titled "Five Points on the Sphere," has been known for a while now; however, without knowing this trick, even professional mathematicians could have a hard time coming up with this simple answer. This is also an application of the so-called **Pigeonhole Principle**, which suggests, in this case, that since there are only two hemispheres (two pigeonholes) and three points (three pigeons) left, at least one of the hemispheres must contain more than one point.

Strategy 3: Work Backwards

Some problems can be "traced back" to find the answer. If you are given the result of a mathematical operation (like addition or multiplication), for instance, you can "undo" the operation to get back to the previous stage of the calculation. Repeating this process, you may be able to determine the starting point.

Go Deeper! *Working Backwards*

1. Click https://www.youtube.com/watch?v=v34NqCbAA1c to watch a TED Talk that explains the benefits of working backwards.

Example 4

Karen is thinking of a number. If you double it and subtract 7, you obtain 11. What is Karen's number?

[Solution]

We start with 11 and work backwards. The opposite of subtraction is addition. We will add 7 to 11. We are now at 18. We now know that doubling her original number resulted in 18. The opposite of doubling something is dividing by 2. So the answer is $18 \div 2 = 9$.

It is always good to check your answer whenever possible. If Karen started with 9, doubling it would give her 18. Subtracting 7 from 18 gives 11. Yes, we have the right answer.

Try It Now 22

You asked Stephanos how old he is. His answer was this: "If you multiply my age by 93, add 6, and divide it by 3, you will get 436." How old is Stephanos? Solve this problem by working backwards.

This "Try It Now" question is similar to many despised exercises found in algebra books. You may have seen this type of problems in the past and wondered, "What kind of crazy person is this dude?" The purpose of this question is not to give you the impression that algebra is useless or to simply frustrate you. It is critical thinking and problem solving that we are after. Practice! By using your analytic skills and applying them to these questions, you will have many

"tools" available in your virtual "toolbox" so that you may be able to, like a master craftsman, pull out and use appropriate tools when solving problems in completely different areas.

Strategy 4: Look for a Pattern

In some previous examples, we saw lists of numbers. These numerical lists turn out to be very important in mathematics, and we have a formal name for these. Because these can describe growth, decay, and other changes, we will be using these throughout this book.

Sequence

A **sequence** is an ordered list of numbers. It may be finite or infinite. It may or may not have a pattern. Each number on the list is called a **term** of the sequence.

Applications of sequences in this book include compound interest (Chapter 3), factorials (Chapter 4), linear and exponential growth (Chapter 6), and musical scales (Chapter 7). It may surprise you that sequences played a very significant role in laying a solid foundation of calculus and are widely used to provide solutions in many application problems.

Here, we use sequences primarily to describe or express patterns. When you see a sequence, ask yourself as you search for a pattern. Are the numbers growing steadily larger? Steadily smaller? Increasing faster and faster? How is each term related to the previous term?

Here are some examples. Some are more obvious than others. See if you can name the next few terms.

(1) $-1, 1, -1, 1, -1, \ldots$ Clearly, 1 and -1 are alternating in this sequence.
(2) $3, 1, -1, -3, -5, -7, \ldots$ The numbers are going down by 2 each time.
(3) $1, 10, 100, 1000, \ldots$ The terms are increasing fast, but the pattern is clear, right?

Example 5

For each sequence below, see the pattern and find the next two terms.
(1) $1, 4, 9, 16, \ldots$
(2) $1, 2, 4, 8, 16, 32, \ldots$
(3) $2, 3, 5, 9, 17, 33, \ldots$
(4) $1, 1, 2, 3, 5, 8, 13, \ldots$

[Solution]

For (1), note that each number is the square of a whole number. Another way to write these numbers is $1^2, 2^2, 3^2, 4^2, \ldots$ So the next two terms are $5^2 = 25$ and $6^2 = 36$.

For (2), you see that the numbers are increasing faster and faster. This is typical when the growth is "exponential" (we will study this in later chapters). If you recognized that the terms are doubling each time, it is easy to see that the next two terms are 64 and 128.

(3) may be difficult at first. These are not very nice numbers like squares or powers of 2. But we do see that the terms are growing faster and faster just as in (2). Stare at them for a while. Think of different possibilities. After a while, you may notice that all these numbers are just one larger than those numbers in (2). See the correspondence? It turns out that the terms of (3) are simply one more than the corresponding terms in the line above. Since the next two terms of (2) were 64 and 128, the next two terms of (3) are 65 and 129.

The last one (4) is actually one of the most famous sequences in the history of mathematics. It is called the Fibonacci Sequence. If you are interested, you can read more about it in Chapter 7. The pattern may be quite hard to discover first, but it is quite simple; after starting with 1 and 1, each term is just the sum of the two previous terms ($1 + 1 = 2,\ 1 + 2 = 3,\ 2 + 3 = 5,\ 3 + 5 = 8$, and so on). So the next term is $8 + 13 = 21$, and the one after that is $13 + 21 = 34$.

Strategy 5: List Examples and Observe

Sometimes it may be best to list some examples (start with easy or simple ones), make observations, and come up with your own hypothesis, which may simply be an educated guess first. In mathematics, a hypothesis is often referred to as a **conjecture**. A conjecture becomes a **theorem** (a word derived from the Greek word meaning "observe" or "behold") when it is proved. A proof is basically a deductive argument showing that the conclusion logically follows from a given set of assumptions. Some proofs are short and simple while others could take up hundreds of pages (and possibly centuries).

Example 6

A number is called a "perfect square" if it is the square of a whole number (like those in the first sequence in Example 5 above). Can perfect squares end in a 2 or a 3?

[Solution]

Do perfect squares end with certain digits? Let us list all the squares of the numbers 1 to 20 and observe.

1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, and 400. Now look at the number in the ones digits. Notice they are 0, 1, 4, 5, 6, or 9. None of the perfect squares seem to end in a 2, 3, 7, or 8. This list suggests that perfect squares cannot end in a 2, 3, 7 or 8. In fact, you can try the next hundred squares, and you will not see a counterexample (a counterexample is an example that disproves our

conjecture). But note that this is just an observation. To answer this question, one would need a mathematically rigorous proof.

This example illustrates something very important. Unlike in science, *in mathematics it is not enough to provide examples*—even thousands of examples (and no counterexamples). A mathematical statement requires a definitive, deductive proof. For this particular pattern seen here, there is, in fact, such a proof. One can even prove that the last digit of this sequence has a very clear pattern: 1, 4, 9, 6, 5, 6, 9, 4, 1, 0, and these digits repeat every 10th number in the sequence. (Can you figure out why?)

In fact, it is also true that any perfect square, when divided by 4, will have only 0 or 1 as the remainder, never 2 or 3. (Can you see why?)

Example 7

How many different amounts of money can you have in your pocket if you have only three coins consisting of only dimes and quarters?

[Solution]

Since the total number of coins is quite small here, your list is not very long. You can list every single possibility now.

Quarters	Dimes	Amount
0	3	30 cents
1	2	45 cents
2	1	60 cents
3	0	75 cents

The answer is clearly 4 amounts, but reflect and think about the problem again. You could have guessed the answer without doing all the calculations. How? You could have simply reasoned that the number of quarters must be 0, 1, 2, or 3, and that number would uniquely determine the number of dimes, and each pair will always give you different amounts since you are replacing a dime with a quarter (15 cents more) each time. Then you can easily see why these amounts increased by 15 cents each time as well.

Try It Now 23

How many ways can you make change for 23 cents using only pennies, nickels, and dimes? (Hint: Make an organized list. Your list is probably going to be quite a bit longer than the list for Example 7.)

Strategy 6: Use the Process of Elimination

This strategy can be used when there is only one possible solution. This is a widely used method in which one starts with a large number of possibilities and crosses out or "eliminates" those that cannot be the answer. The process is repeated until only one is left.

Example 8

What is a number that satisfies all the following requirements? Are there more than one?

(1) The number is odd.
(2) It is more than 1 but less than 100.
(3) It is greater than 20.
(4) It is less than 5 times 7.
(5) The sum of the digits is 7.

[Solution]

From the first two sentences, we know that the number is an odd number between 1 and 100. The next two sentences narrow down the choices to a number between 20 and 35 (but not 35). So far, the possibilities are the following:

21, 23, 25, 27, 29, 31, 33

Then, (5) gives us a powerful clue; we can eliminate 21, 23, 27, 29, 31, and 33. In fact, it is clear that 25 is the only one satisfying (5) because $2 + 5 = 7$ and no other number on this list has this attribute. By using the elimination strategy, we have found our answer.

Exercises 1.5

Apply any problem-solving strategies and/or your critical-thinking skills to solve various types of problems below. There is no single formula or procedure to follow. Be flexible and consider all possibilities.

1. There are 13 postage stamps on the table. Some are 20-cent stamps while others are 45-cent stamps. The total postage value of these stamps is $4.10.
   (a) If they were all 20-cent stamps, would 13 of them add up to $4.10?
   (b) If there were five 20-cent stamps, would these stamps add up to $4.10?
   (c) If there were ten 20-cent stamps, would these stamps add up to $4.10?

   OK, you probably got the idea now.

   (d) How many 20-cent stamps are there?
   (e) Can you think of another way to solve the problem?

(f) What would you say to a friend of yours who tries to help you out by writing a system of two equations to solve this problem?

2. Find the next two terms of each of the following sequences and explain why.
   (a) 9, 7, 5, 3, …
   (b) 2, 22, 222, 2222, …
   (c) 3, 6, 12, 24, …
   (d) 0, 1, 3, 6, 10, 15, …
   (e) 5, 5, 7, 5, 5, 7, 5, 5, …
   (f) 1, 2, 6, 24, 120, …

3. A man bought an old car for $2000. He fixed it up and sold it for $2,500. But he missed it, so he brought it back for $3,200. Later, he sold it for $4,000. How much did he make in these transactions?
   (a) Is it $4,000 - $2,000 = $2,000?
   (b) He made $500 on the first sale and then $800 on the second. But he lost $700 in between when he re-purchased it. So is it $800 - $700 = $100?
   (c) Is it $500 + $800 = $1,300?

4. See the pattern below. Can you make a conjecture? Is it true? Can you prove it?
   - $1 = 1\ (= 1^2)$.
   - $1 + 3 = 4\ (= 2^2)$.
   - $1 + 3 + 5 = 9\ (= 3^2)$.
   - $1 + 3 + 5 + 7 = 16\ (= 4^2)$.

5. A heart surgeon is about to perform a medical procedure on a boy. The surgeon told the nurses, “I want you to know that this is my son. I am operating on my own child today.” Everyone there clearly knew that the boy was the surgeon’s son. But the surgeon was not the boy’s father, and everyone clearly knew that, also. How can this be true?

6. Sometimes a pre-conceived notion could give us bias. Consider the following description about “H.D.”

   “H.D. sat on a wall. H.D. had a great fall. All the president’s horses and all the president’s men couldn’t put H.D. together again.”

   (a) What do you think this “H.D.” is? How do you know that?
   (b) Now, suppose you grew up never hearing any nursery rhymes and all you are given about “H.D.” is the statements above. What *must* be true about this “H.D.”?

7. You have decided to work out at the gym at least twice a week, but never on two consecutive days. If you are to keep the same schedule every week, list all possible days (of the week) you can exercise at the gym.

8. A man must be married to have a mother-in-law but can be single and have a brother-in-law. Explain why.

9. Your sprinklers are set to water your plants at 6 am every morning during summer, when the daylight saving time is in effect. Once the time goes back to regular time (in November, say from PDT to PST), what time do your sprinklers start? (Hint: it's either 5 am or 7 am.) What is the best way to explain this to your friend?

10. You are to visit a friend who lives 300 miles away. You drive to his house early in the morning, averaging 60 mph, but you return in the afternoon, when the freeway is jammed, at an average speed of 40 mph. Is the overall average speed 50 mph? Why or why not?

11. You have 10 identical pairs of black socks and 9 identical parts of white socks in your drawer, except each pair is not "paired up," i.e., each sock is in the drawer, separated from all others. The room is completely dark, and you cannot see which socks you are taking out. How many socks do you have to pull out if you want to be sure that you get…
    (a) A pair of white socks?
    (b) A pair of black socks?
    (c) Any matching pair?

12. A 49-year-old widower George has a 25-year-old son, Mike. George marries a young woman Gloria whose mother Michelle is a widow. Mike marries Michelle, and they have a baby girl together.
    (a) How is that baby related to George? (The baby is his son's daughter.)
    (b) How else is that baby related to George? (The baby is his wife's mother's daughter.)
    (c) Does that baby have a half-brother? If so, under what conditions?
    (d) Who is that baby's half-sister?

13. A person can be $1/2$ Chinese and $1/2$ Italian (by having a Chinese mother and an Italian father, for example). Are the following cases possible? If so, how? If not, why not? What are the criteria that would make any combination possible?
    (a) $1/2$ Japanese, $1/4$ Russian, $1/4$ Irish
    (b) $3/8$ Scottish, $3/8$ Vietnamese, $1/4$ Spanish
    (c) $3/8$ Mexican, $1/4$ Turkish, and $1/2$ Norwegian
    (d) $1/6$ French, $1/2$ Canadian, $1/3$ Brazilian
    (e) $1/2$ Armenian, $1/4$ German, $1/2$ Swedish

14. John says, “I don’t have any brothers, sisters, half-brothers, or half-sisters. See that tall woman? Her father is my mother’s child.” Who is the tall woman?

15. Suppose you have infinitely many ping pong balls, each numbered with a natural number, like 1, 2, 3, 4, … In Stage 1, Balls 1 through 100 are thrown into a bin, and Ball 1 is removed from the bin. In Stage 2, Balls 101 through 200 are thrown into the bin, and Ball 2 is removed. Continue this process indefinitely.[17]
    (a) After Stage 2, how many balls are in the bin? Which ones?
    (b) How about after Stage 9? How many balls are in the bin? Which ones?
    (c) After Stage n, how many balls are in the bin? Which ones?

    This problem has an interesting, fascinating, and perplexing sequel. See the exercise set for Section 9.2 if you are interested.

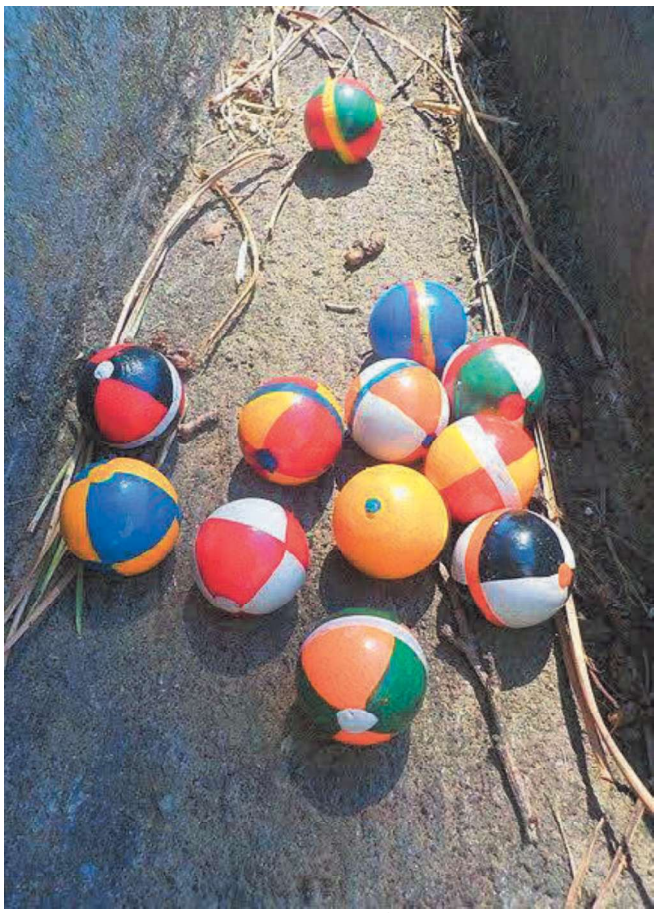


Figure 1.6[18]

[17] If you are worried that this would take forever, imaging Stage 1 taking 1 minute, Stage 2 taking half a minute, Stage 3 taking a quarter minute, and so on, each stage taking half the time as the previous stage. Then, this “whole” process will take no more than 2 minutes.

[18] Image by Albarubescens is licensed under CC BY-SA 4.0

For Further Explorations…

Fallacies (Section 1.1) are everywhere. In fact, although the first section in this book lists many types of fallacies, you may be able to add more to the list. Being aware of these different types will help identify fallacies all around us and hopefully prevent you from believing false conclusions, misinformation, and misleading claims.

However, the danger is real, and finding truth turns out to be much more difficult in general—in fact, it is so difficult that even experts could be tricked and fall prey to all types of fallacies. In his 2021 book *The Data Detective*[iii], author and economist Tim Harford lists several real-life examples of "Appeal to Consequence." Remember that this is the fallacy type "X is true because I want it to be true." While this may appear obvious and easy to avoid, examples abound; our emotions often lead to conscious or unconscious bias. One can also refer to this as "wishful thinking" or "motivated reasoning." Here are some of the examples from Harford's book.[19]

- Abraham Bredius, an art scholar specializing in works of the 17th century master artist Johannes Vermeer, had a theory that some important religious works by Vermeer were missing. For years he was hoping to discover those missing art pieces by Vermeer. When a piece called *Christ at Emmaus* was rediscovered, Bredius called it the "Dutch master's finest work." The picture was placed in a famous museum as an authentic work by Vermeer. Later, a former artist and an art dealer named Han van Meegeren, charged with treason, admitted that he had painted *Christ at Emmaus* himself, fully aware that Bredius wanted there to be a missing religious work by Vermeer.

- A psychologist showed an article to a group of people. The article showed evidence that caffeine could increase the risk of developing breast cysts. Coffee-drinkers among the female subjects, who obviously did not want the conclusion to be true, did not find the article or its evidence convincing while many other people did. This is the other side of the fallacy; "X must not be true because I do not want it to be true."

- Science has overwhelmingly showed that HIV (human immune-deficiency virus) causes AIDS. However, there are many, even today, who refuse to believe that AIDS is caused by this particular virus. In fact, there are those who flatly deny the existence of HIV. A large number of people around the world reject this truth because they do not want it to be true.

- In March, 2020, just when COVID-19 was quickly making its way across the globe, a group of U.K. scientists published an epidemic model (a "tip of the iceberg" model) suggesting that this new, then little-known, virus is not as dangerous as it was first believed, implying that the problem would not last for a long time or spread to a large scale. Many people who saw

[19] All these examples are from the first chapter "Search Your Feelings" (pp. 20—46) of Tim Harford's book *The Data Detective*.

this model happily believed this news because it was what they were looking for. Unfortunately, they could not have been more wrong as we now know.

- A 2011 experiment by an Oxford University economist divided up his subjects into two fictional groups: “farmers,” who would benefit if the wheat price is high, and “bakers,” who would benefit if it is low. The subjects were asked to predict the price of wheat and would be rewarded for accuracy. It turns out that the predicted wheat prices were high among the “farmers” and low among the “bakers.”

Consider some of your beliefs, particularly those with which your friends and family may disagree. Perhaps they are about the types of food and drinks you consume. Maybe they involve your favorite hobbies, sports, or other activities. If you enjoy watching football, do you believe football is a safe sport even when the news tells you otherwise? If you enjoy red meat, what is your reaction when you read that red meat is generally unhealthy? If you ride a motorcycle, what do you think of statistics on motorcycle accidents?

As humans, none of us is immune to bias based on our own feelings. That is why, when you see data or statistical reports, it would be wise to stop, ask yourself how you feel (and why you feel that way), and not rush to conclusions.

Figure 1.7[20]

[20] Image by Yin_Min_Tun is licensed under CC BY-SA 4.0

Chapter 2: Mathematics and Numbers

"You don't have to be a mathematician to have a feel for numbers."—John F. Nash, Jr.

Problem solving may require identifying an appropriate procedure and determining what information you will need (and will not need) to answer the question. Frequently you may have to recognize a certain pattern yourself, come up with your own procedure, and/or find some creative way to resolve the issue as you saw in the last section.

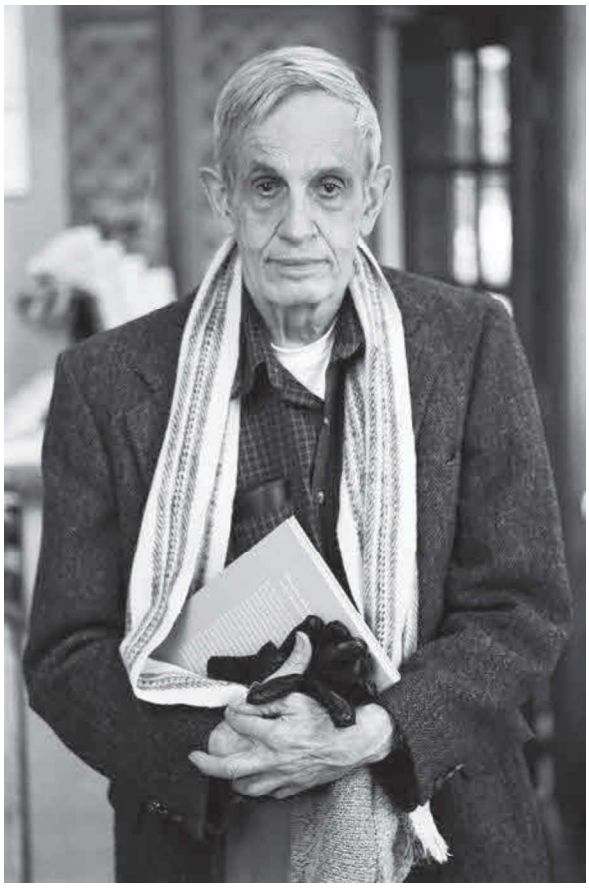


Figure 2.1: John F. Nash, Jr.[21]

In this chapter, we will review several basic but powerful ideas like percentages, rates, proportions, conversions, and estimations. We will discuss deceptive and misleading use of numbers, focus on the problem-solving process, and explore how to use these ideas to solve problems where we do not have perfect information. You can have a "feel" for numbers without being a professional mathematician, as quoted above by the Nobel laureate featured in the 1998 movie *A Beautiful Mind.*

[21] Image by Peter Badge is licensed under CC BY-SA 3.0

2.1 Percentages[iv]

One of the most useful and common ways to express a proportion or a rate is percentages. They are found literally everywhere—from store sales, advertisements, sports, and hobbies, to some of the most advanced research studies and global issues.

Percentages as Ratios

Percent comes from the Latin expression *per centum*, which literally means "per 100," or "parts per hundred." So it is a **ratio** with the base number 100. When we write 40%, this is equivalent to the fraction $\frac{40}{100}$ or the decimal 0.40. Notice that 80 out of 200 and 2 out of 5 are also 40%, since

$$\frac{80}{200} = \frac{2}{5} = \frac{40}{100} = 0.40 = 40\%.$$

Now, we all know that 2 hits out of 5 at-bats in baseball is significantly different from 80 hits out of 200 at-bats. The former could be just luck while the latter requires a significant level of skill and consistency. However, as a percentage, both of these situations result in the same batting average. It is important to know that a percentage may or may not give us the whole story (in this situation, the small sample space—only 5 at-bats—should be pointed out).

Example 1

243 people out of 400 state that they like dogs. What percent is this?

[Solution]

$\frac{243}{400} = 0.6075 = \frac{60.75}{100}$.

This is 60.75%.

Notice that the percent can be found from the equivalent decimal by **moving the decimal point two places to the right**. Conversely, if you are given a percentage, move the decimal point two places to the left to find the equivalent decimal number:

- From decimal to percent (to write answer appropriately)
  - $0.825 = 82.5\%$
  - $0.009 = 0.9\%$
  - $1.5 = 150\%$

- From percent to decimal (for calculation input use)
  - $73\% = 0.73$
  - $0.3\% = 0.003$ (3 out of 1000)
  - $250\% = 2.5$ (2 and half times as much)

This type of conversion becomes quite important as you gain understanding of calculations involving percentages and as you carry out various operations on your calculator.

Example 2

Write each as a percent.

(1) $\frac{1}{4}$ (2) 0.02 (3) 2.35

[Solution]

(1) $\frac{1}{4} = 0.25 = 25\%$ (2) $0.02 = 2\%$ (3) $2.35 = 235\%$

Every percentage expresses a proportion of some specific part (like the number of hits) out of a whole (like the number of at-bats). When we say "2% of the students," for instance, we are describing a small portion (a **part**) of all the students (a **whole**, or a **base**).[22] In other words, every percentage involves a part and a whole. Because of how fractions work, the part is placed on the top (the numerator) and the whole (base) is placed on the bottom (the denominator).

The English language is quite convenient because we refer to the portion in question as "part" **of the whole**. The preposition "of" is almost always followed by the whole (the denominator). Keep this in mind if you get confused about which number is which. Further, because the "whole" is what the percentage is *based on*, for consistency, we will use the term "base" to refer to the "whole."

Percent

If we have a *part* that is some *percent* of a *whole*, then

Percent= $\frac{\text{part}}{\text{whole}}$, or equivalently, part = percent × whole, or

Percent= $\frac{\text{part}}{\text{base}}$ or equivalently, part = percent × base

To do the calculations, we write the percent as a decimal.

You have seen math books where they define percentage $P\%$ with A as the part and B as the base as follows:

$$\frac{P}{100} = \frac{A}{B} \quad \text{or} \quad P = \frac{A}{B} \times 100.$$

This is correct if you write the percentage as "$P\%$." **However, for many applications**, **it is better to write percentages as decimal numbers**. (This is especially true in financial

[22] Normally speaking, we think of a "part" smaller than the "whole." However, we often use percentages exceeding 100%, in which case the term "whole" could be misleading. For this reason, we mostly use the term "base" (sometimes "reference") for the denominator of percentages after these introductory remarks.

calculations in Chapter 3.) In other words, 40% should be written as 0.40, and 0.3% should be written as 0.003 instead. See the example below.

Example 3

The sales tax in a town is 9.4%. How much tax will you pay on a $140 purchase?

[Solution]

Here, $140 is the base, and we want to find 9.4% *of* $140. We start by writing the percent as a decimal by moving the decimal point two places to the left (which is equivalent to dividing by 100). We can then compute:

$$0.094 \times 140 = 13.16$$

So the answer is $13.16 in tax.

Example 4

In the news, you hear "tuition is expected to increase by 7% next year." If tuition this year was $1200 per quarter, what will it be next year?

[Solution]

We calculate 7% of $1200 first: $0.07 \times 1200 = 84$ (only the *increase*; not full tuition). We then add the increase to the tuition: $\$1200 + \$84 = \$1284$ (expected tuition). *Quicker Method (Recommended)*: The tuition next year will be the current tuition plus an additional 7%, so it will be ***107% of this year's tuition***. In other words, instead of finding the increase and then adding it to the current tuition, we can find the answer more directly as follows:

$$1.07 \times 1200 = 1284.$$

So the tuition will be $1284 per quarter. The decimal number 1.07 represents 107%.

Since this point is relevant in many future sections, let me add an explanation. Obviously, 100% plus 7% is 107%. So it makes sense to multiply the base by 1.07 to get your answer here. Another way to think about this is what is known as the **Distributive Property** in algebra to write the sum of $1200 and the amount $84 of increase this way:

$$\begin{aligned} 1200 + (0.07 \times 1200) &= (1 \times 1200) + (0.07 \times 1200) \\ &= (1 + 0.07) \times 1200 \\ &= 1.07 \times 1200 \\ &= 1284. \end{aligned}$$

So, in general, you see the following general method, which is useful in more complicated calculations, such as those in Chapter 3 (compound interest).

Quantity After a Percent Increase/Decrease

When the base number *B* is increased by a given percentage, the total will be expressed as $(\mathbf{1}+\boldsymbol{r})\times\boldsymbol{B}$**,** where *r* is the decimal representation of that percentage.

Similarly, if the base number *B* is reduced by a given percentage, the result is expressed as $(\mathbf{1}-\boldsymbol{r})\times\boldsymbol{B}$, where *r* is the decimal representation of that percentage.

Try It Now 1

A TV set originally priced at $800 is on sale for 30% off. There is then a 9.2% sales tax. Find the price after including the discount and sales tax. (The sales tax is based on the sale price.)

Finding the Base

Sometimes you are given a percentage and the part represented by that percentage, but not the base. It requires division to find the base.

Example 5

6900 people of a small town got sick with the flu last year, and that represents 23% of the population. What is the population?

[Solution]

To compute the population, which is the base (that is, the whole population), we can rephrase this situation as follows: "23% **of what number** is 6900?" Remember, use 0.23 for "23%." In other words,

$$0.23\times\text{Population}=6900$$
$$(\text{percent}\times\text{base}=\text{part})$$

In order to find the population, we divide 6900 by 0.23.

$$6900\div 0.23=30{,}000.$$

So the population of this town is 30,000. You can easily check the answer and see that this is correct since $0.23\times 30{,}000=6900$.

Example 6

The price of an international airline ticket suddenly increased by 25% today, and now it is $1000. How much was the original price before the price jump?

[Solution]

Observe that the percentage given was **not 25% of the new price**, but rather it is **25% of the original price**, which we do not know. Hence, the correct way to think about this problem is this: "A 25% increase on the original price means that the new price, $1000,

is 125% of the old price." In other words, the question is "125% of what is $1000?" Note that *the base is the unknown here*. As we did in the previous example, we can write it this way:

$$1.25 \times \text{Original Price} = 1000$$
$$(\text{percent} \times \text{base} = \text{part})$$

Dividing 1000 by 1.25, we find that the original price is $800. This makes sense because 25% of $800 is indeed $200, and $200 more than $800 is $1000.

It is tempting to give the following answer, which is *one of the most common errors in percentage calculations*: "25% of $1000 is $250, so subtract this from $1000. The original price must have been $750." One can easily check and see this is incorrect because 25% of $750 is

$$0.25 \times 750 = 187.50,$$

and the sum of this number and $750 is only $937.50, not $1000.

This is tricky, right? It turns out there are many common errors involving percentages. In the next section, we will study these in particular. However, before moving to that section, we will introduce the concept of "absolute" versus "relative" change.

Absolute and Relative Change

Percentages are frequently used to describe changes with respect to time. We have already seen an example where tuition increases by a certain percentage. There are primarily two ways to express changes: **absolute change (numerical change)** and **relative change (percentage change)**.

Absolute Change and Relative Change

Given two values (starting quantity and ending quantity),
Absolute change = $\text{ending quantity} - \text{starting quantity}$
Relative change= $\frac{\text{absolute change}}{\text{starting quantity}}$ expressed as a percentage
Absolute change has the same units as the original quantity. Relative change gives a percent change. The base of the percentage change is always the starting quantity.

In mathematics, the term *absolute value* exists, and that number is never negative; however, the word *absolute* here is used in contrast to the word *relative*, and an absolute change can be positive, 0, or negative. When the quantity increases, the absolute and the relative changes are positive; when the quantity decreases, these changes are negative. For instance, if a population goes down by 7%, we say that the relative change is -7% (and the quantity of the decrease is 7% of the *starting* population).

Example 7

The value of a car dropped from $7400 to $6800 over the last year. What percent decrease is this?

[Solution]

To compute the percent change, we first need to find the dollar value change:

$$6800 - 7400 = -600.$$

The absolute (numerical) change is $-\$600$.

Since we always compute the decrease based on (relative to) the starting value, we compute the percentage of this absolute change out of $\$7400$:

$$\frac{\text{absolute change}}{\text{staring quantity}} = \frac{-600}{7400} \approx -0.081 = -8.1\%.$$

This is the relative change. We can say that the decrease was about 8%. (The symbol $\approx$ denotes "approximately." We will simply use reasonable approximations in this book instead of significant digits in a precise way as in scientific calculations.)

As seen in this example, when a change is expressed, the base of the percentage is always the former (starting) quantity because that is the first quantity with respect to time.

The base of a percent is very important. For example, it was argued that marijuana was a "gateway" drug, with some claiming that 80% of marijuana smokers went on to use harder drugs like cocaine. The problem? This was not true. The true claim is that 80% of harder drug users first smoked marijuana. The difference is one of base: 80% of marijuana smokers using hard drugs, versus 80% of hard drug users having smoked marijuana. These numbers are not equivalent. As it turns out, only one in 2,400 marijuana users actually go on to use harder drugs.[23]

Absolute and Relative Difference

When comparing two numbers (not necessarily a quantity change), either of the numbers can be used as the base, so one must be very specific when presenting percentages. Look for the key terms like "**of**," "**more than**," and "**less than**." The quantity that follows these *markers* serve as the base number and thus should be used as the denominator in percentage calculations. As with changes, when comparing two quantities, we have an **absolute (numerical) difference** and a **relative (percentage) difference.**

[23] Lies, Damned Lies, and Statistics. (n.d.). *tvtropes.* http://tvtropes.org/pmwiki/pmwiki.php/Main/LiesDamnedLiesAndStatistics is licensed under CC BY-NC-SA 3.0

Example 8

Suppose a rental car company charges $75 a day for a full-size sedan and $215 a day for a convertible sports car. Compare the prices of the two rental cars.

[Solution]

The absolute difference is 215 – 75 = 140. Using this value, we could say "The sports car costs $140 a day more than the sedan." However, that number may not mean much by itself; the relative difference may be more meaningful. There are **two different relative differences** we could calculate, depending on which car we use as the base:

Using the sedan as the base, $\frac{140}{75} \approx 1.867$. This tells us that the sports car costs 186.7% **more than** the sedan.

Using the sports car as the base, $\frac{140}{215} \approx 0.651$. This tells us that the sedan costs 65.1% **less than** the sports car.

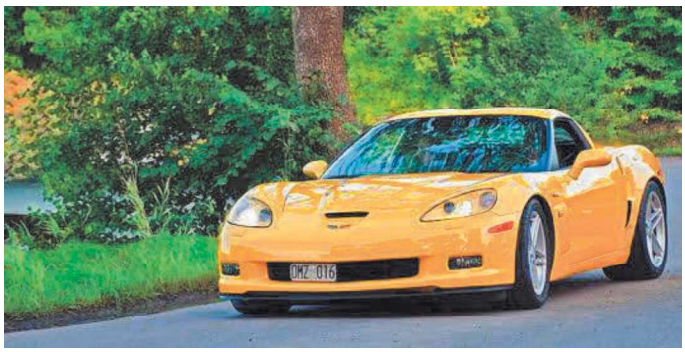


Figure 2.2[24]

Notice both of these are showing relative (percent) differences. We could also calculate the daily rate of the sports car relative to that of the sedan: $\frac{215}{75} = 2.867$, which tells us that the price of renting the sports car is 286.7% **of** that of the sedan (exactly 100% more than the first relative difference we calculated). Likewise, we could calculate the daily rate of the sedan relative to that of the sports car: $\frac{75}{215} = 0.349$, which tells us that the daily rate of the sedan is 34.9% **of** that of the sports car (exactly the difference between 100% and the second relative difference we calculated).

Notice that, in the above example, we used phrases like "less than," "more than," and "of." The first two signal relative differences (in which the absolute difference appears as the numerator) while the preposition "of" has a different meaning. Be sure to distinguish these two. But whichever you use, the number that immediately follows these phrases is the base of the percentage. We will have more on "of" versus "than" in the next section.

[24] Image by Staffan Andersson is licensed under CC BY-SA 4.0

Try It Now 2

The U.S. federal debt at the end of 2016 was $19.57 trillion and grew to $20.24 trillion by the end of 2017. At the end of 2019 it was $22.72 trillion and grew to $26.95 trillion by the end of 2020.[25] Calculate the absolute and relative increase for 2016-2017 and 2019-2020. Which year saw a larger increase in federal debt?

Because percentages largely depend on what you decide as the base (reference), always pay close attention when you read statements involving percentages.

Example 9

A local paper article on high school graduation rates reported the following: "The number of schools graduating 60% or fewer students in four years decreased by 17 during that time period. The number of students attending schools with such low graduation rates was cut in half."

(1) Is the phrase "decrease by 17" useful in this story?

(2) Considering the last sentence, can we conclude that the number of schools graduating 60% or fewer students in four years was originally 34 (because 17 is half of 34)?

[Solution]

(1) This number is hard to evaluate since we have no basis for judging whether this is a large or small change. If the number of these schools dropped from 20 to 3, that would be a very significant change; however, if the number dropped from 217 to 200, that would be less of an improvement.

(2) The last sentence states that only half the *students* now attend these schools. This provides a relative change which helps put the first sentence in perspective. We can speculate that the number of such schools may have been around 34. However, note that 17 was a number of schools whereas the "half" refers to a number of students, not schools. You can assume that most schools are similar in size, but we do not know that. Further, it is possible that students simply moved schools rather than the schools improving, so that estimate might not be fully accurate.

[25] Duffin, E. (2020). Public Debt of the United States 1990-2020. *Statista.* https://www.statista.com/statistics/187867/public-debt-of-the-united-states-since-1990/

Exercises 2.1

1. Patrick left an $8 tip on a $50 restaurant bill. What percent tip is that?

2. Out of 230 racers who started the marathon, 212 completed the race, 14 gave up, and 4 were disqualified before reaching the goal. What percentage did not complete the marathon?

3. Suppose that every employee pays 4.2% of their gross wages towards social security (FICA tax). How much will someone earning $45,000 a year pay towards social security?

4. Ireland has a 23% VAT (value-added tax, similar to a sales tax). How much will the VAT be on a purchase of a €250 item (an item that costs 250 euros)?

5. A project on Kickstarter.com was aiming to raise $15,000 for a precision coffee press. They ended up with 714 supporters, raising 557% of their goal. How much did they raise?

6. Another project on Kickstarter for an iPad stylus raised 1,250% of their goal, raising a total of $312,500 from 7,511 supporters. What was their original goal?

7. The population of a town increased from 3,250 in 2018 to 4,300 in 2020. Find the absolute and relative (percent) increase.

8. The number of CDs sold in one year was 114 million, down from 147 million the previous year. Find the absolute and relative (percent) decrease.

9. A company wants to decrease their energy use by at least 15%.
   (a) If their electric bill is currently $2,200 a month, what will their bill be if they are successful?
   (b) If their next bill is $1,700 a month, were they successful? Why or why not?

10. A store is hoping an advertising campaign will increase their number of customers by at least 30%. They currently have about 80 customers a day.
    (a) How many customers will they have if their campaign is successful?
    (b) If they increase to 120 customers a day, were they successful? Why or why not?

2.2 Deceptive and Misleading Numbers

Although mathematics covers a whole lot more than numbers, it is still true that arithmetic, including the basic operations on numbers such as integers, fractions, decimals, and percentages, remains one of the most practical and fundamental applications of mathematics that you encounter every day. Many decisions are made, both in large and small scales, based on numbers, figures, statistics, and all sorts of data. Based on various numbers, companies may decide where to build a distribution center or a shopping mall, and government agencies may decide to implement certain policies. Different political parties may present different numbers to push their own agenda or to silence their opponents. Advertisers know that figures can be very effective in persuading consumers. In fact, many of us may not even be aware of how we make our shopping decisions based on numbers presented to us in a variety of forms. This is especially true in the Internet era in which we live since advertisements are targeted based on our online activities and digital footprints.

It is precisely because of this fact that percentages and other figures are frequently manipulated. We now look at how certain numbers can be presented to us in misleading or deceptive ways to give certain impressions or to change our mind.

Absolute and Relative Difference

We have already seen that percentages can be calculated in different ways to produce different numbers to describe the same change. Let us recall how we compare two quantities, both in absolute and relative senses. (This is a review of the previous section.)

Absolute and Relative Difference

When comparing two quantities,

- **Absolute Difference**: numerical difference between the two numbers
- **Relative Difference**: percentage difference based on one of the numbers

$$\text{Relative Difference} = \frac{\text{Absolute Difference}}{\text{Base Number}}$$

When describing a change (from an old number to a new number),

- **Absolute Change**: numerical difference between the new and the old numbers
- **Relative Change**: percentage difference based on the old (original) number

$$\text{Relative Change} = \frac{\text{Absolute Change}}{\text{Original Number}}$$

Absolute change has the same units as the original quantity. Relative change gives a percent change. The base of the percentage change is always the starting quantity.

For instance, suppose the annual tuition is $50,000 at College A and $20,000 at College B. In terms of absolute difference, you can state, truthfully, that the tuition is $30,000 higher at College A than at College B. In terms of relative difference, however, each of the following statements is true:

1. The tuition at College B is 60% cheaper **than at College A** since
$$(50{,}000 - 20{,}000) \div 50{,}000 = 30{,}000 \div 50{,}000 = 0.60.$$

2. The tuition at College A is 150% more **than at College B** since
$$(50{,}000 - 20{,}000) \div 20{,}000 = 30{,}000 \div 20{,}000 = 1.50.$$

In addition, without referring to the absolute difference, the following statements are also correct:

3. The tuition at College B is 40% **of the tuition at College A** since
$$20{,}000 \div 50{,}000 = 0.40.$$

4. The tuition at College A is 250% **of the tuition at College B** since
$$50{,}000 \div 20{,}000 = 2.50.$$

Here, the first and the third statements use the tuition at College A as the base while the second and the fourth use the tuition at College B as the base. Note the words like "of," "more than," and "cheaper than." The amount that follows "of" or "than" is the base, the reference number.

Now, 60, 150, 40 and 250 are very different numbers, and yet, each of these percentages can be used to represent the tuition difference between these two colleges. If you are trying to persuade someone to go to College B based on the tuition, you could say, "Hey, College A charges you 250% as much as College B." And *that* sounds dramatic.

Similar calculations can be carried out in comparing many things including prices, index numbers, crime rates, populations, and averages.

Example 1

According to Wikipedia.org, the 2018 population estimate for California was 39,557,000 while the population estimate for Michigan was 9,996,000.

(1) The population of Michigan is what percent less than the population of California?
(2) The population of California is what percent more than the population of Michigan?
(3) The population of Michigan is what percent of the population of California?
(4) The population of California is what percent of the population of Michigan?

[Solution]

Note that the term “less than the population of California” is used in (1). That means we need to divide the absolute difference by the population of California.

$$(39{,}557{,}000 - 9{,}996{,}000) \div 39{,}557{,}000 = 0.7473.$$

Therefore, the population of Michigan is about 74.7% less than the population of California.

Similarly, for (2), the term “more than the population of Michigan” is used. So we divide that same absolute difference by the population of Michigan:

$$(39{,}557{,}000 - 9{,}996{,}000) \div 9{,}996{,}000 = 2.9573.$$

Therefore, California has about 295.7% more people than Michigan does.

Now, for (3), neither “more than” nor “less than” appear in the question. This time, the absolute difference is not needed:

$$9{,}996{,}000 \div 39{,}557{,}000 = 0.2527.$$

Therefore, the population of Michigan is about 25.3% of that of California.

Similarly, for (4), we do not need the absolute difference. This time, the population of Michigan is the base.

$$39{,}557{,}000 \div 9{,}996{,}000 = 3.957.$$

Therefore, the population of California is about 396% of that of Michigan.

Of course, they all imply the same fact: California has about 4 times as many people as Michigan.

Try It Now 3

1. **Make the same comparison as above (finding four percentage figures) between**
   a. **Virginia (8,518,000) and Mississippi (2,987,000)**
   b. **Michigan (9,996,000) and Wyoming (578,000)**

2. **As of 2018, California (39,557,000) is the most populous state in the union while Wyoming (578,000) is the least populous. The population of Wyoming is about what percent of that of California? How many times is the population of California as compared to that of Wyoming?**

A Percentage Change Followed by Another

When percentage change is followed by another percentage change, the base for percentage calculations changes, so one must be careful ***not* to simply add the two percentages**. This is not a difficult concept, but many people are unaware of this. Here are a few examples.

Example 2

The number of students at a city college increased 20% one year, and the number again increased by 30% the following year. Does this mean the two-year increase was 50%?

[Solution]

These problems can be simplified if you start with a fixed number. For example, let's assume that the college had 1000 students two years ago. Then, in one year, with a 20% increase, the college had 1200 students. But then, in the second year, the college added 30% of 1200, which is 360. With the addition of 360 students over 1200, the college has 1560 students after two years. That represents an increase of 560 over 1000, which amounts to a 56% (not 50%) increase.

By the way, it turns out that the original number is irrelevant. Regardless of the number you start with, a 20% increase followed by a 30% increase would *always* result in a total increase by 56% since $(1 + 0.2) \times (1 + 0.3) = 1.56$.

Example 3

The value of a certain stock jumped up by 50% one day, only to lose 50% the following day. Does it mean that the stock is back to the original value after the two-day period?

[Solution]

Again, let us consider a specific number to start with. Suppose the stock was valued at $100 two days ago. After jumping up by 50%, the stock was valued at $150 yesterday. Then, after a 50% loss (and 50% of $150 is $75), the stock is now only worth $75, meaning that it lost 25% of the value in the two-day period.

Why do these things happen? It is because the base for the percentage has changed when the value changed. Again, remember that percentages cannot be simply added in these cases.

Example 4

Suppose that a stock drops in value by 60% one week but then increases in value the next week by 75%. Is the value higher or lower than where it started?

[Solution]

To answer this question, suppose the value started at \$100. After one week, the value dropped by 60%:

$$100 - (0.06 \times 100) = 100 - 60 = 40.$$

So the value is \$40 after the 60% drop. Then the next week, notice that base of the percent has changed to the new value, \$40. Computing the 75% increase gives

$$40 + (0.75 \times 40) = 40 + 30 = 70.$$

In the end, the stock is valued at \$70, still \$30 lower (30% lower), than it started.

Finding the Base

As seen in the previous section, confusion in the base can happen in the simplest of cases, as illustrated by the following examples in which **the base is the unknown**. These examples illustrate one of the most common errors when percentage increases and decreases are mentioned.

Example 5

(1) A local theme park just increased the parking fee by 25%, and **now it is \$20**. How much was the parking fee before?
(2) A dining room set just went on sale and is **now available at \$1400**, which is 30% off the regular price. What is the regular price?

[Solution]

(1) Many people make the error in assuming that \$20 is the base for 25%. However, careful reading suggests that this is not correct. The new parking fee is \$20 while we do not yet know what the fee was before. In other words, the problem does *not* state the base explicitly. But we *do* know that whatever it was, adding 25% of that to the old parking fee results in the new fee of \$20. Now, the old parking fee is of course 100% of itself, so \$20 represents 125% of the old parking fee. Hence, if PF represents the old parking fee, we have

$$(1 + 0.25) \times PF = 1.25 \times PF = 20.$$
$$PF = 20 \div 1.25 = 16.$$

Therefore, the parking fee used to be \$16. If you make the common mistake of finding 25% of \$20, that would be \$5, and deducting \$5 from \$20 would give you \$15. But this is clearly incorrect because adding 25% to \$15 would result in \$18.75 (as shown below), not \$20.

$$(1 + 0.25) \times 15 = 18.75.$$

(2) The calculation is similar except we subtract (not add) 30% (which is 0.30) from 1. Whatever the regular price of the dining room set was, a 30% discount means that you are still paying 70% of the original price ($1 - 0.30 = 0.70$). Denoting the regular price by RP, we have

$$(1 - 0.3) \times RP = 0.7 \times RP = 1400.$$
$$RP = 1400 \div 0.7 = 2000.$$

So the original price of the dining room set was \$2000. This makes sense because 30% of \$2000 is \$600, which is the discount you are getting when you pay \$1400 rather than \$2000 for it.

You may have noticed that *you divide whenever you need to find the base*. This is a simple tip you may want to remember: to calculate the base, you always divide the "part" by the percentage (written as a decimal number).

Percentages of Percentages

Suppose the mayor has worked very hard to fight poverty in your city and, as a result, the poverty rate has gone down from 16% to 12% over the last four years. How much of a drop does that represent?

Of course, you can say that the poverty rate has gone down by 4% (16% – 12%), right? Well, that would be a little ambiguous, and here is why. Generally, when we use percentages to express a relative change, we use the earlier number (in this case, 16%) as the base. Therefore, the relative change in this case is

$$(12 - 16) \div 16 = -0.25,$$

representing a 25% drop in the poverty *rate*.

So what are we saying? Is 4% equal to 25%? Of course not! A more accurate way of describing this is as follows: "The 4% drop represents 25% of the base, which is 16%" (because 4 is 25% of 16). Yes, we are dealing with a percentage of a percentage. These numbers are particularly misleading and quite confusing. For that reason, when we use the *absolute* difference in two percentages (such as the 4% drop from 16% to 12%), we refer to that as "**points** (or percentage points)." In other words, here we can say that the poverty rate has dropped by 4

points (as opposed to the more confusing "4 percent"). -25% then refers to the *relative* difference of the two poverty rates, which changed from 16% to 12%.

Example 6

(1) The approval rate of a senator has gone up from 44% to 49%. How many points has the rate increased by? What percentage has the approval rate increased by?
(2) The fatality rate of a certain disease has gone down from 22% to 10%. How many points has the rate gone down by? What percentage drop does this represent?

[Solution]

(1) The rate has gone up by 5 *points*. However, one can also say that the rate has gone up by about 11.4% since

$$(49 - 44) \div 44 = 5 \div 44 = 0.1136$$

(2) The rate has dropped by 12 *points*. Also, the rate has gone down by about 54.5% (a little more than half) because

$$(10 - 22) \div 22 = -12 \div 22 = -0.5454 \ldots$$

This number is negative, representing a decrease, not an increase.

We conclude this section with a caution against averaging percentages.

Example 7

A basketball player scores on 40% of two-point field goal attempts and on 30% of three-point field goal attempts during a season. Can you find the player's overall field goal percentage?

[Solution]

It is very tempting to average these values and claim the overall average is 35%, but this is likely not correct, since most players make many more two-point attempts than three-point attempts. We do not actually have enough information to answer the question. Suppose the player attempted 200 two-point field goals and 100 three-point field goals. Since 40% of 200 is 80 and 30% of 100 is 30, this player made $80 + 30 = 110$ shots out of 300, for a $\frac{110}{300} = 0.367$ = 36.7% overall field goal percentage. But again, the problem does not give us sufficient information to answer it.

Exercises 2.2

1. You invested some money in a start-up company, and now your investment has gained 60% and is worth $6,720. How much was your original investment?

2. A college football ticket can be purchased at $28 after a 20% student discount is applied. How much is the ticket without the student discount?

3. The unemployment rate in a certain small town has changed from 4% to 7% during the last ten years. The rate went up by how many percentage points? By what percent?

4. Suppose that the survival rate of a certain form of cancer increased from 33% to 59%. Express this change in terms of points and as a percentage.

5. An article reports "attendance dropped 6% this year, to 3008." What was the attendance before the drop?

6. An article reports "sales have grown by 30% this year, to $273 million." What were sales before the growth?

7. Are these two claims equivalent, in conflict, or not comparable because they are talking about different things?
   (a) "16.3% of Americans are without health insurance."[26]
   (b) "Only 55.9% of adults receive employer-provided health insurance."[27]

8. Are these two claims equivalent, in conflict, or not comparable because they are talking about different things? **Hint**: Profit margin expresses what percent of the selling price is the profit. For instance, if the store buys an item for $18 and sells it for $20, the profit is $2, and that is 10% of the price ($\frac{2}{20} = 0.10$), so the profit margin is 10%.
   (a) "We mark up the wholesale price by 25% to come up with the selling (retail) price."
   (b) "The store has a 20% profit margin."

9. Are these two claims equivalent, in conflict, or not comparable because they are talking about different things?
   (a) "Every year since 1950, the number of American children gunned down has doubled."
   (b) "The number of child gunshot deaths has doubled from 1950 to 1994."

[26] Smith E., & Stark C. (2012). *By the numbers: Health insurance*. CNN. http://www.cnn.com/2012/06/27/politics/btn-health-care/index.html
[27] Mak, T. (2012). *Poll: Fewer insured through work*. Politico. http://www.politico.com/news/stories/0712/78134.html

10. A high school currently has a 30% dropout rate. They have been tasked to decrease that rate by 20%. Find the equivalent percentage point drop.

11. A politician's support grew from 42% by 3 percentage points to 45%. What percent (relative) change is this?

12. Suppose you hear someone say, "The number of homicides in my school district has quadrupled over the last two years." What does this mean? What possible scenarios can you think of? Is there anything else you would want to know regarding this claim?

13. Is it possible for the price of a concert ticket to increase by 150%? To decrease by 150%? Why or why not?

14. If your favorite shoes are on sale with a 30% discount, and if you get an additional employee discount of 10%, can you get the shoes at 40% off? Why or why not?

15. How would you react if your boss tells you the following? "I will have to give you a 50% pay cut now, but I will give you a permanent 60% pay raise next week." (Start with some fixed pay now and see what happens. Then explain why this is counter-intuitive.)

16. Alfred University had 47,456 students in 2020 while Beltran University had 77,966 students in 2020. Complete the following statements:
    (a) Beltran's enrollment was ___% larger than Alfred's.
    (b) Alfred's enrollment was ___% smaller than Beltran's.
    (c) Alfred's enrollment was ___% of Beltran's.
    (d) Beltran's enrollment was ___% of Alfred's.

17. In the 2012 Olympics, Usain Bolt ran the 100m dash in 9.63 seconds. Jim Hines won the 1968 Olympic gold with a time of 9.95 seconds.
    (a) Bolt's time was ___% shorter than Hines'.
    (b) Hines' time was ___% longer than Bolt's.
    (c) Hines' time was ___% of Bolt's.
    (d) Bolt's time was ___% of Hines'.

18. A store has clearance items that have been marked down by 60%. They are having a sale, advertising an additional 30% off clearance items. What percent of the original price do you end up paying?

19. Which is better: having a stock that goes up 30% on Monday then drops 30% on Tuesday, or a stock that drops 30% on Monday and goes up 30% on Tuesday? In each case, what is the net percent gain or loss?

20. Suppose a baseball player had a batting average of .290 in the first half and .270 in the second half of this season. Was his season average .280? Explain why or why not.

Figure 2.3[28]

[28] Image by Dave Herholz is licensed under CC BY-SA 2.0

2.3 Rates, Proportions, and Conversions

Although percentages are a very common way to express a ratio (a part out of a whole), the most fundamental concept of a ratio comes from fractions. In fact, every fraction where the numerator and the denominator are both whole numbers is called a *rational number* since it refers to a ratio. As we have already seen, every percentage represents a ratio out of 100 ("per cent"), which can also be expressed as a decimal number. If, for instance, you cut a small pizza into 4 slices and ate one, you ate a quarter of the pizza, which can be written as $\frac{1}{4}$ or 25% or 0.25. Having studied percentages, we now move to a few practical topics closely associated with percentages: rates, proportions, and conversions.

Rates and Proportions

If you wanted to power your city using wind power, how many windmills would you need to install? Questions like these can be answered using rates and proportions.

Rates

A rate is the ratio (fraction) of two quantities. A **unit rate** is a rate with a denominator of one.

Example 1

Your car can drive 300 miles on a tank of 15 gallons. How many miles does your car run per gallon? Express this as a rate.

[Solution]

The fraction is the rate, $\frac{300 \text{ miles}}{15 \text{ gallons}}$, where we can carry out the division (simplify the fraction) to find a unit rate: $\frac{20 \text{ miles}}{1 \text{ gallon}}$. We could then write this as $20 \frac{\text{miles}}{\text{gallon}}$, or just 20 miles per gallon.

Proportion Equation

A proportion equation is an equation showing the equivalence of two rates or ratios.

A key to solving a proportion equation is "**cross multiplication**." When two ratios are equal to each other, two products obtained by cross multiplication are equal: $\frac{a}{b} = \frac{c}{d}$ if and only if $a \times d = b \times c$ (by the definition of fractions). You may remember from algebra that we often omit the multiplication "times" symbol when using letters. So this will be written as follows from now on:

Equality of Two Fractions (Proportions): Cross Multiplication

$$\frac{a}{b} = \frac{c}{d} \text{ if and only if } ad = bc.$$

It is now necessary to use a little bit of algebra; in particular, we often represent an unknown quantity with a letter like x. We then do the same thing to both sides of the equation to isolate x.

Example 2

Solve the proportion $\frac{5}{3} = \frac{x}{6}$ for the unknown value x.

[Solution]

This proportion is asking us to find a fraction with denominator 6 that is equivalent to the fraction $\frac{5}{3}$. We can solve this by cross multiplication, giving

$$3x = 5 \times 6 = 30.$$

Now, we divide both sides of the equation by 3 to get $x = 10$.

Example 3

A map scale indicates that ½ inch on the map corresponds with 3 real miles. How many miles apart are two cities that are $2\frac{1}{4}$ inches apart on the map?

[Solution]

We can set up a proportion by setting equal two $\frac{\text{map inches}}{\text{real miles}}$ rates, and introducing a variable x to represent the unknown quantity – the real mile distance between the cities.

$$\frac{\frac{1}{2}\text{ map inch}}{3\text{ miles}} = \frac{2\frac{1}{4}\text{ map inches}}{x\text{ miles}}.$$

Converting the fractions to decimals ($\frac{1}{2} = 0.5$ and $2\frac{1}{4} = 2.25$), we can rewrite this as

$$\frac{0.5}{3} = \frac{2.25}{x}.$$

Cross-multiplying, we obtain

$$0.5x = 3 \times 2.25 = 6.75.$$

When we divide both sides by 0.5, we get the answer, $x = 13.5$ (real miles).

Alternatively, we can use those fractions to solve the question, but it is a bit tricky since the mixed number $2\frac{1}{4}$ must first be converted to an improper fraction. Observing that

$$2\frac{1}{4} = 2 + \frac{1}{4} = \frac{8}{4} + \frac{1}{4} = \frac{9}{4},$$

we can rewrite our original equation as

$$\frac{1/2}{3} = \frac{9/4}{x},$$

from which we set up the equation

$$\frac{1}{2}x = \frac{9}{4} \times 3 = \frac{27}{4}.$$

Multiply both sides by 2, we get our final answer:

$$x = \frac{27}{4} \times 2 = \frac{27}{2} = 13\frac{1}{2} \text{(miles)}.$$

Many proportion problems can also be solved using **dimensional analysis**, the process of multiplying a quantity by rates to change the units. This process is also used in conversion problems. Keeping track of units will also help you interpret different numbers in your calculations and make sense of what you are dealing with.

In terms of notation, we will use the *times* symbol "$a \times b$" and the *dot* symbol "$a \cdot b$" (as well as the *algebraic* expression "ab") to refer to multiplication (the product of two numbers) throughout this book.

Example 4

Your car can drive 300 miles on a tank of 15 gallons. How far can it drive on 40 gallons?

[Solution]

We could certainly answer this question using a proportion: $\frac{300 \text{ miles}}{15 \text{ gallons}} = \frac{x \text{ miles}}{40 \text{ gallons}}$

However, we earlier found that 300 miles on 15 gallons gives a rate of 20 miles per gallon. (Notice that the unit on the top is "miles" and the unit on the bottom is "gallons.") If we multiply the given 40-gallon quantity by this rate, the *gallons* unit "cancels" and we are left with a number of miles:

$$40 \text{ gallons} \times \frac{20 \text{ miles}}{\text{gallon}} = \frac{40 \text{ gallons}}{1} \cdot \frac{20 \text{ miles}}{\text{gallon}} = 800 \text{ miles}.$$

If, instead, we were asked "how many gallons are needed to drive 50 miles?" we invert the 20 mile-per-gallon rate so that the *miles* unit cancels and we are left with gallons:

$$50 \text{ miles} \cdot \frac{1 \text{ gallon}}{20 \text{ miles}} = \frac{50 \text{ miles}}{1} \cdot \frac{1 \text{ gallon}}{20 \text{ miles}} = \frac{50 \text{ gallons}}{20} = 2.5 \text{ gallons}.$$

In this equation, we see the fraction $\frac{1}{20}$. The units are "gallons per mile" as opposed to "miles per gallon." This fraction essentially tells us that the car uses 1/20 of a gallon of fuel for every mile it runs.

Unit Conversions

In the United States, we do not use the metric system as most of the world does. The system used in the U.S. is similar to the U.K.-based Imperial System, but these two systems of measurement are also slightly different. This is somewhat unfortunate since these differences make the U.S. the only country which uses what is known as the United States Customary Units."

Here are some unit conversions for reference. Your instructor may or may not require you to memorize these although it is definitely convenient and useful if you know some of the basic ones. For everyday use, you may want to download a free app for unit conversions. The list below includes both the U.S. units and the metric units of measurement for convenience.

Length

- 1 foot (ft) = 12 inches (in)
- 1 mile = 5,280 feet
- 1000 millimeters (mm) = 1 meter (m)
- 1000 meters (m) = 1 kilometer (km)
- 1 yard (yd) = 3 feet (ft)
- 100 centimeters (cm) = 1 meter
- 2.54 centimeters (cm) = 1 inch

Weight and Mass

- 1 pound (lb) = 16 ounces (oz)
- 1000 milligrams (mg) = 1 gram (g)
- 1 kilogram = 2.2 pounds (on earth)
- 1 ton = 2000 pounds
- 1000 grams = 1kilogram (kg)

Capacity (Volume)

- 1 cup = 8 fluid ounces (fl oz)[29]
- 1 quart = 2 pints = 4 cups
- 1000 milliliters (ml) = 1 liter (L)
- 1 pint = 2 cups
- 1 gallon = 4 quarts = 16 cups

Example 5

A bicycle is traveling at 15 miles per hour. How many feet will it cover in 20 seconds?

[Solution]

To answer this question, we need to convert 15 miles per hour. If we know the speed of the bicycle in feet per second, this question would be simpler. Since we do not, we will need to do additional unit conversions. Here, we will need to know that 5280 ft = 1 mile. We might start by converting the speed into feet per second.

$$15\frac{\text{miles}}{\text{hr}} = \frac{15 \text{ miles}}{1 \text{ hr}} \cdot \frac{5280 \text{ ft}}{1 \text{ mile}} = 79{,}200\frac{\text{ft}}{\text{hr}} = \frac{79{,}200 \text{ ft}}{\text{hr}} \cdot \frac{1 \text{ hr}}{60 \text{ min}} \cdot \frac{1 \text{ min}}{60 \text{ sec}} = 22\frac{\text{ft}}{\text{sec}}.$$

$$22\frac{\text{ft}}{\text{sec}} \cdot 20 \text{ sec} = 440 \text{ ft}.$$

You may have noticed that conversions require multiplication, but always multiplication by a form of 1 (since we do not want to change the quantity). In the above example, $\frac{5280 \text{ ft}}{1 \text{ mile}}$, $\frac{1 \text{ hr}}{60 \text{ min}}$, and $\frac{1 \text{ min}}{60 \text{ sec}}$ all represent 1 since the top and the bottom of each fraction are the same quantity.

Try It Now 4

A 1000-foot spool of bare 12-gauge copper wire weighs 19.8 pounds. How much will 18 inches of the wire weigh, in ounces?

[29] While "ounce" is a unit of weight, "fluid ounce" is a unit of capacity (volume) for liquids, not weight. 1 fluid ounce ≈ 1 ounce (weight) for water.

Notice that with the miles-per-gallon example, if we double the miles driven, we double the gas used. Likewise, with the map distance example, if the map distance doubles, the real-life distance doubles. This is a key feature of proportional relationships and one we must confirm before assuming two things are related proportionally.

Example 6

Suppose that you are tiling the floor of a 10-ft-by-10-ft room and find that 100 tiles will be needed. How many tiles will be needed to tile the floor of a 20-ft-by-20-ft room?

[Solution]

In this case, while the width of the room has doubled, the area has quadrupled. Since the number of tiles needed corresponds with the area of the floor, not the width, 400 tiles will be needed. We could find this using a proportion based on the areas of the rooms:

$$\frac{100 \text{ tiles}}{100 \text{ ft}^2} = \frac{x \text{ tiles}}{400 \text{ ft}^2}.$$

In this case, the answer was pretty obvious.

Other quantities just do not scale proportionally at all.

Example 7

Suppose a small company spends $1000 on an advertising campaign and gains 100 new customers from it. How many new customers should they expect if they spend $10,000?

[Solution]

While it is tempting to say that they will gain 1000 new customers, it is likely that additional advertising will be less effective than the initial advertising. For example, if the company is a hot tub store, there are likely only a fixed number of people interested in buying a hot tub, so there might not even be 1000 people in the town who would be potential customers.

Geometric Ratios

Next, we consider another important aspect of mathematics application: geometry. Geometric shapes, as well as area and volumes, can often be important in problem solving.

Example 8

You want to find out how tall a tree is but do not have any way to climb it. Describe a method for determining the height.

[Solution]

There are several approaches we could take. We will use one based on triangles, which requires that it is a sunny day. Suppose the tree is casting a shadow, say 15 ft long. I can then measure my own shadow. Suppose I am 6 ft tall and cast a 2-ft shadow. Since the triangle formed by the tree and its shadow has the same angles as the triangle formed by me and my shadow, these triangles are called **similar triangles,** and their sides scale proportionally. In other words, the ratio of height to width will be the same in both triangles. Using this, we can find the height of the tree, which we will denote by *h*:

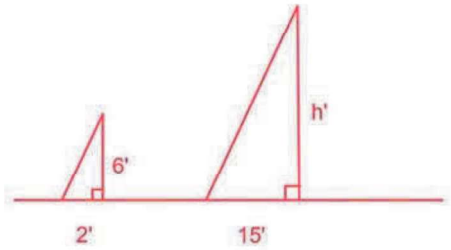


Figure 2.4

$$\frac{6}{2} = \frac{h}{15}$$

By cross multiplication, we get $2h = 6 \times 15 = 90$. Thus, $h = 45$ (feet).

It may be helpful to recall some formulas for areas and volumes of a few basic shapes.

[Areas]

Rectangle
Area = $L \times W$
Perimeter = $2L + 2W$

Circle
Area = $\pi\, r^2$
circumference = $2\, \pi\, r$

Figure 2.5

[Volumes]

Rectangular Box

Volume = $L \times W \times H$

Circular Cylinder

Volume = $\pi r^2 h$

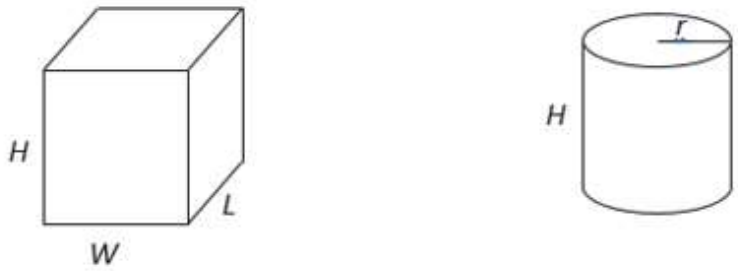


Figure 2.6

Example 9

If a 12-inch-diameter pizza requires 10 ounces of dough, how much dough is needed for a 16-inch pizza?

[Solution]

To answer this question, we need to consider how the weight of the dough will scale. The weight will be based on the volume of the dough. Assuming that both pizzas will be about the same thickness, the weight will scale with the area of the pizza. We can find the area of each pizza using the formula for area of a circle, πr^2:

A 12" pizza has radius 6 inches, so the area will be $6 \times 6 \times \pi = 36\pi \approx 113$ square inches.

A 16" pizza has radius 8 inches, so the area will be $8 \times 8 \times \pi = 64\pi \approx 201$ square inches.

We can now set up a proportion to find the weight of the dough for a 16" pizza:

$$\frac{10 \text{ ounces}}{113 \text{ in}^2} = \frac{x \text{ ounces}}{201 \text{ in}^2}$$

Cross-multiplying and dividing both sides of the equation by 113, we get

$$x = 201 \times 10 \div 113 \approx 17.8.$$

So you will need about 17.8 ounces of dough for a 16" pizza. It is interesting to note that while the diameter is $\frac{16}{12}$ = 1.33 times as large, the dough required, which scales with area, is 1.33^2 = 1.78 times as much.

Try It Now 5

A website says that you will need 48 fifty-pound bags of sand to fill a sandbox that measure 8 ft by 8 ft by 1 ft. How many bags would you need for a sandbox 6 ft by 4 ft by 1 ft?

Example 10

A company makes regular and jumbo marshmallows. The regular marshmallow has 25 calories. How many calories will the jumbo marshmallow have?

[Solution]

We would expect the calories to scale with volume. Since the marshmallows have cylindrical shapes, we can use that formula to find the volume. From the grid in the image, we can estimate the radius and height of each marshmallow.

Figure 2.7 Photo courtesy Christopher Danielson[30]

The regular marshmallow appears to have a diameter of about 3.5 units, giving a radius of 1.75 units, and a height of about 3.5 units. The volume is about $\pi(1.75)^2\ (3.5) = 33.7\ \text{units}^3$.

The jumbo marshmallow appears to have a diameter of about 5.5 units, giving a radius of 2.75 units, and a height of about 5 units. The volume is about $\pi(2.75)^2\ (5) = 118.8\ \text{units}^3$.

We could now set up a proportion. The regular marshmallow has 25 calories for 33.7 cubic units of volume.

$$\frac{25\ \text{cal}}{33.7\ \text{units}^3} = \frac{x\ \text{cal}}{118.8\ \text{units}^3}.$$

[30] Problem Solving by David Lippman is licensed under CC BY-SA 3.0 US.

Cross-multiplying and dividing both sides of the equation, we get

$$x = 25 \times 118.8 \div 33.7 \approx 88.1.$$

The jumbo marshmallow should have about 88.1 calories. Note that while the diameter and height are about 1.5 times as large for the jumbo marshmallow, the volume and calories are about $1.5^3 = 3.375$ times as large.

As these examples show, the area of a flat shape (a 2-dimensional figure, e.g., circles, rectangles, triangles) grows proportionally to the *square* of each side or radius while the volume of a 3-dimensional object (e.g., boxes, balls, marshmallows) grows proportionally to the *cube* of each side or radius.

For example, if you have a square floor, doubling each side length (multiplying by 2) will result in the area which is $2^2 = 4$ times as large. Tripling each side will give you the area $3^2 = 9$ times as large. If there is a ball, doubling its radius (multiplying by 2) will result in the volume which is $2^3 = 8$ times as large. Tripling its radius will result in a ball whose volume is $3^3 = 27$ times as large as the original ball.

In general, the following holds for shapes/objects on a plane or in space *if all the dimensions increase by a certain percentage* (without changing the shape). This could surprise you because a paper cup which looks slightly bigger than another paper cup could actually hold much more drink than it may appear to you. See examples below.

Relative Changes of Area and Volume

If every dimension of a shape on a plane becomes $(1 + r)$ times as large, its **area** becomes $(1 + r)^2$ times as large.

If every dimension of an object in space becomes $(1 + r)$ times as large, its **volume** becomes $(1 + r)^3$ times as large.

Example 10

Suppose that the radius and the height of a medium cup is 30% more than those of a small cup. Compare the volume/capacity of the medium cup compared to the small cup. By what percentage does the volume increase? Also, compare the area of the bottom side of each cup.

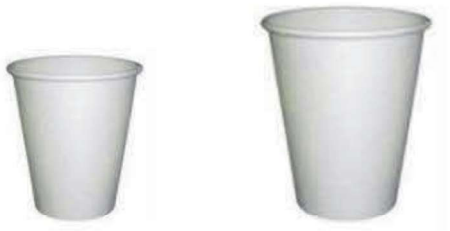


Figure 2.8

[Solution]

"30% more" means $r = 0.3$, so the volume is $(1 + 0.3)^3 = 1.3^3 = 2.197$. In other words, the medium cup holds more than twice as much drink as the small cup. For the area, $(1 + 0.3)^2 = 1.3^2 = 1.69$, so the area of the bottom of the cup increased by 69%.

Example 11

In shotput, a typical shot (a metal ball) for men weighs 16 pounds with a diameter of about 5 inches. How much would the weight be if a solid ball is made of the same material with same density but with a diameter of 10 inches? How about 15 inches?

[Solution]

10 inches would double the original radius, so the weight would be $2^3 = 8$ times as much, meaning it would weigh $16 \times 8 = 128$ pounds (incredible, right?). 15 inches would be tripling the radius, resulting in a shot with $3^3 = 27$ times as much weight, which is $16 \times 27 = 432$ pounds. Good luck throwing (or even lifting) it.

Exercises 2.3

1. Find a unit rate: You bought 10 pounds of potatoes for $4.

2. Find a unit rate: Joel ran 1500 meters in 4 minutes, 45 seconds.

3. Solve: $\frac{2}{5} = \frac{6}{x}$

4. Solve: $\frac{n}{5} = \frac{16}{20}$

5. Convert 8 feet to inches.

6. Convert 6 kilograms to grams.

7. A wire costs $2 per meter. How much will 3 kilometers of wire cost?

8. Sugar contains 15 calories per teaspoon. A "cup" in the U.S. is equivalent to 16 tablespoons, and 1 tablespoon is 3 teaspoons. How many calories are in 1 cup of sugar?

9. A car is driving at 100 kilometers per hour. How far (how many meters) does it travel in 2 seconds?

10. A chain weighs 10 pounds per foot. How many ounces will 4 inches weigh?

11. A crepe recipe calls for 2 eggs, 1 cup of flour, and 1 cup of milk. How much flour would you need if you use 5 eggs?

12. A crown molding 8 feet long and 4 inches wide costs $14. How much will it cost to buy 40 feet of crown molding if the width is to be 4 inches everywhere?

13. Four wind turbines can supply enough electricity to power 3000 homes. How many turbines would be required to power 55,000 homes?

14. A highway had a landslide, where 3,000 cubic yards of material fell on the road, requiring 200 dump truck loads to clear. On another highway, a slide left 40,000 cubic yards on the road. How many dump truck loads would be needed to clear this slide?

15. The sound of lightning (thunder) travels at the speed of sound, which is about 750 miles per hour. Assume that light travels so fast that you see a flash of lightning instantly.
    (a) If you see a flash of lightning and then hear the thunder 4 seconds later, how far away is the lightning?
    (b) Now generalize that result. Suppose it takes n seconds to hear the thunder after a flash of lightning. How far away is the lightning, in terms of n?

16. Sound travels about 750 miles per hour. If you stand in a parking lot near a building and sound a horn, you will hear an echo.
    (a) Suppose it takes half a second to hear the echo. How far away is the building[31]?

[31] Kraft, N. (2012) *Tuba Echo*. Vimeo http://vimeo.com/40377128

(b) Now generalize that result. Suppose it takes n seconds to hear the echo. How far away is the building, in terms of n?

17. Which of these glasses contains more liquid? What information would you need to answer this question?

Figure 2.9[32]

18. I was at the store and saw two sizes of avocados being sold. The regular size (on the left in the photo below) sold for $0.88 each, while the jumbo ones sold for $1.68 each. Which is the better deal? (You may need to physically measure their sizes by holding a ruler on this page.)

Figure 2.10[33]

[32] Problem Solving by David Lippman is licensed under CC BY-SA 3.0 US
[33] Problem Solving by David Lippman is licensed under CC BY-SA 3.0 US

2.4 Estimating the Answer

We now bring together the mathematical tools we have reviewed and use them to approach more complex problems. In many problems, it is tempting to take the given information, plug it into whatever formulas you have handy, and hope that the result is what you were supposed to find. Chances are that this approach has served you well in other math classes.

This approach may not work well with real life problems. Instead, problem solving is best approached by starting at the end: identifying exactly what you are looking for. From there, you then work backwards, asking "what information and procedures will I need to find this?" Very few interesting questions can be answered in one mathematical step; often you will need to chain together a solution pathway, a series of steps that will allow you to answer the question.

Here is another problem-solving process, similar to Polya's method presented in Chapter 1. This time, the emphasis is on determining what you need to know and **estimating** the answer.

1. Identify the question you are trying to answer.
2. Work backwards, identifying the information you will need and the relationships you will use to answer that question.
3. Continue working backwards, creating a solution pathway.
4. If you are missing necessary information, look it up or estimate it. If you have unnecessary information, ignore it.
5. Solve the problem, following your solution pathway.

In most problems we work on, we will be approximating a solution because we do not have perfect information. We will begin with a few examples where we will be able to **approximate** the solution using basic knowledge from our lives.

Example 1

Estimate how many times your heart beats in a year.

[Solution]

This question is asking for the rate of heart beats per year. Since a year is a long time to measure heart beats for, if we knew the rate of heart beats per minute, we could scale that quantity up to a year. So the information we need to answer this question is heart beats per minute. This is something you can easily measure by counting your pulse while watching a clock for a minute.

Suppose you count 80 beats in a minute. Since there are 60 minutes in an hour, you can mentally calculate $80 \times 60 = 4800$ and then estimate that the heart beats *about* 5000 times an hour. Multiplying this by 24, we get 120,000 times a day, or maybe about 100,000 times a day. Since there are 365 days in a year, we can come up with an

estimate of 36.5 million beats per year. But since we rounded 120,000 down to 100,000, perhaps the real answer may be more like 40 million beats a year. This type of rough estimate may be good enough in many situations.

If more precision is required, we can convert 80 beats per minute to beats per year as follows:

$$\frac{80 \text{ beats}}{1 \text{ minute}} \times \frac{60 \text{ minutes}}{1 \text{ hour}} \times \frac{24 \text{ hours}}{1 \text{ day}} \times \frac{365 \text{ days}}{1 \text{ year}} = 42,048,000 \text{ (beats per year)}.$$

Since this is a rough estimate, you could round the answer to "about 42 million heart beats per year."

Example 2

How thick is a single sheet of paper? How much does it weigh?

[Solution]

While you might have a sheet of paper handy, trying to measure it would be tricky. Instead, we might imagine a stack of paper and then scale the thickness and weight to a single sheet. If you have ever bought paper for a printer or copier, you probably bought a ream, which contains 500 sheets. We could estimate that a ream of paper is about 2 inches thick and weighs about 5 pounds. Scaling these down,

$$\frac{2 \text{ inches}}{\text{ream}} \cdot \frac{1 \text{ ream}}{500 \text{ pages}} = 0.004,$$

So a single sheet of paper is about 0.004 inch thick. Similarly,

$$\frac{5 \text{ pounds}}{\text{ream}} \cdot \frac{1 \text{ ream}}{500 \text{ pages}} = 0.01,$$

So a single sheet of paper weighs about 0.01 pounds, or 0.16 ounces.

Example 3

A recipe for zucchini muffins states that it yields 12 muffins, with 250 calories per muffin. You instead decide to make mini-muffins, and the recipe yields 20 muffins. If you eat 4, how many calories will you consume?

[Solution]

There are several possible solution pathways to answer this question. We will explore two.

First, a rough estimate. The recipe is for 12 muffins, but you just made 20 mini-muffins. If each mini-muffin is exactly half the original size, you would have ended up with 24 muffins, so each mini-muffin is a little more than half of the original muffin. Eating 4 of those mini-muffins then would be like eating a little more than 2 of the original muffins. Since each of those has 250 calories, your guess is that 4 mini-muffins would have a little more than 500 calories. Perhaps you would estimate about 600 calories.

Now, if you need more precision, let us determine the number of calories in each mini-muffin. To find the calories in each mini-muffin, we could first find the total calories for the entire recipe and then divide it by the number of mini-muffins produced. To find the total calories for the recipe we could multiply the calories per standard muffin by the number of muffins. Notice that this produces a multi-step solution pathway. It is often easier to solve a problem in small steps rather than trying to find a way to jump directly from the given information to the solution.

We can now execute our plan:

$$12 \text{ muffins} \cdot \frac{250 \text{ calories}}{\text{muffin}} = 3000 \text{ calories for the whole recipe.}$$

$$\frac{3000 \text{ calories}}{20 \text{ mini}-\text{muffins}} = 150 \text{ calories per mini-muffin.}$$

$$4 \text{ mini muffins} \cdot \frac{150 \text{ calories}}{\text{mini}-\text{muffin}} = 600 \text{ calories, which is the precise answer.}$$

Note that the estimate we came up with first was actually the correct answer.

Example 4

You need to replace the boards on your deck and want to estimate how much the materials will cost. Your deck is rectangular, measuring 16 ft by 24 ft. A visit to the local home store suggests that an 8-ft-by-4-inch cedar deck board costs about $7.50. Find an estimated cost.

[Solution]

There are two approaches we could take to this problem: 1) estimate the number of boards we will need and find the total cost, or 2) estimate the area of the deck and find the approximate cost per square foot for deck boards. We will do both and compare.

1) Note that 4 inches is 1/3 of a foot, so you will need 3 of these cedar boards to cover an area 8’ by 1’. So 6 of these boards would cover an area 16’ by 1’. Your deck is 16’ by 24’. So multiply 6 by 24. This may be hard to do mentally. Well, 6 times 25 is 150, so you need about 150 boards (an over-estimate) for your deck. Each board is $7.50, so 100 boards would cost $750. Adding 50 more would mean another $400

or so, but remember, this is an over-estimate. So you can estimate the total cost to be a little above $1000.

2) For this solution pathway, we will be able to answer the question if we know the cost per square foot for deck boards and the square footage of the deck. To find the cost per square foot for deck boards, we could compute the area of a single board and divide it into the cost for that board. We can compute the square footage of the deck using geometric formulas.

Since your deck is 16 ft by 24 ft, the total area is the product of these numbers, 384 ft^2.

The area of each deck board, doing the necessary conversion from inches to feet, is

$$8 \text{ feet} \times (4 \text{ inches} \cdot \frac{1 \text{ foot}}{12 \text{ inches}}) \approx 2.667 \text{ ft}^2.$$

The cost per square foot is then

$$\frac{\$ 7.50}{2.667 \text{ ft}^2} \approx \$2.8125 \text{ per ft}^2.$$

This will allow us to estimate the material cost for the whole 384 ft^2 deck:

$$\$2.8125 \times 384 \approx \$1080.$$

This is very consistent with the estimate obtained by the first approach. Of course, this cost estimate assumes that there is no waste, which is rarely the case. It is common to add at least 10% to the cost estimate to account for waste. You may want to say that it will be around $1200.

You may have noticed that you could have also solved the problem by setting up the proportion:

$$\frac{\$ x}{384 \text{ ft}^2} = \frac{\$7.50}{2.667 \text{ ft}^2}$$

where each side is representing the cost per square foot.

By the way, it may be more practical to get an even rougher estimate. This may be more realistic when thinking of starting this project. Once you figure out that the cost per square foot is about $2.81 (as above), you could just think of it as $3 (you over-estimate it since there will be some wasted pieces). Roughly the deck is just less than 400 square feet, so multiplying it by $3, you can "guestimate" that the project will cost you a little less than $1200. This is a reasonable estimate, and such a "ballpark" figure could be quite useful in the beginning.

Exercises 2.4

1. It takes an air pump 5 minutes to fill a twin-size air mattress (39 by 75 by 8.75 inches). How long will it take to fill a queen-size mattress (60 by 80 by 8.75 inches)? First, estimate the answer. Then, find the answer by setting up a proportion equation.

2. It takes your garden hose 20 seconds to fill a 2-gallon watering can. How long will it take to fill each? (1 cubic foot is about 7.5 gallons.)
   (a) An inflatable pool 3 feet wide, 8 feet long, and 1 foot deep?
   (b) A circular inflatable pool 12 feet in diameter and 3 feet deep?

3. You want to put a 2-inch-thick layer of topsoil for a new garden 20 feet wide and 30 feet long. The dirt store sells by the cubic yards. How many cubic yards will you need to order? First, estimate the answer. Then, calculate the precise answer.

4. Your chocolate milk mix says to use 4 scoops of mix for 2 cups of milk. After pouring in the milk, you start adding the mix but get distracted and accidentally put in 5 scoops of mix. How can you adjust the mix?

5. A recipe for sabayon calls for 2 egg yolks, 3 tablespoons of sugar, and ¼ cup of white wine. After cracking the eggs, you start measuring the sugar but accidentally put in 4 tablespoons of sugar. How can you compensate? Estimate first, and then calculate the precise answer.

6. The store is selling lemons at 2 for $1. Each lemon yields about 2 tablespoons of juice. How much will it cost to buy enough lemons to make a 9-inch lemon pie requiring ½ cup of lemon juice?

7. You read online that a brick patio 15 ft by 20 ft would cost about $2,275 to have professionally installed. Estimate the cost of having a brick patio 18 ft by 22 ft installed. Then, find the precise answer.

2.5 Really Big and Really Small Numbers

In January, 2020, the United States announced that it will join the "One Trillion Trees Initiative," an international project by the World Economic Forum. This initiative is an effort to grow, restore, and conserve one trillion trees across the globe by 2030.

But what does it look like? Can you imagine one trillion trees? It is not one *million* or one *billion*. It is one thousand million. Exactly how big *is* this number? How can we understand such a large number? One of the common ways to understand very large (and very small) numbers is to make comparisons with something we *could* imagine or understand. For instance, for the One Trillion Trees Initiative, we could figure out how many trees each human being on the earth will have if the trees are allocated to the inhabitants of this planet. We could also calculate how many years it would take to plant so many trees if one tree can be planted every second. We will return to this problem later.

First, however, it may be worth noting that there is a variety of ways in which large (and small) numbers are written, read, and denoted around the world. There is no global agreement, unfortunately. For instance, "one billion" used to mean different numbers in the U.S. and the U.K.[34]

Note that in both of these English-speaking countries, a comma separates every three digits to make it easier to read large numbers, but in some European countries, a comma may be used as the decimal point (sometimes referred to as the "decimal comma"). You may be shocked to see the distance of a marathon race "42.195 km" written as "42,195 km," for instance (that looks like over 26,200 miles). The number π may be written as "3,14159…" (Very confusing, right?)

Also, in many Asian languages, large numbers have names based on 10,000s (every four digits) rather than 1,000s. For instance, in Chinese, 10,000 is referred to as a "wan," 100,000 as "10 wan," 1 million as "100 wan," and 10 million as "1000 wan." Only when you multiply that by 10, which is 100 million, do you get a new numerical unit, "yi" (100 million is one "yi"). Large and small numbers can be quite confusing.

[34] "One billion" in America refers to 10^9, which is 1,000,000,000; however, in the U.K., until fairly recently (like the mid-1970s) this word "(one) billion" used to mean 10^{12}, which is 1,000,000,000,000, the number we would call one trillion in America.

The following table describes the standardized and scientific names and Latin prefixes for very large and very small numbers. Understanding your properties of exponents will help you understand this. (Note: we use metric prefixes, e.g., 1000 meters = 1 kilometer).

Big Numbers

Number	Name	Metric Prefix
$10^1 = 10$	Ten	deka-
$10^2 = 100$	Hundred	hecto-
$10^3 = 1000$	Thousand	kilo-
$10^4 = 10{,}000$	Ten Thousand	myria-
$10^6 = 1{,}000{,}000$	One Million	mega-
$10^9 = 1{,}000{,}000{,}000$	One Billion	giga-
10^{12}	One Trillion	tera-
10^{15}	One Quadrillion	peta-
10^{18}	One Quintillion	exa-

Figure 2.8[35]

By the way, 10^{100}, a one followed by one hundred 0's, is actually named "googol." Note that this spelling is different from the name of a large tech company.

Small Numbers

Number	Name	Metric Prefix
$10^{-1} = 0.1$	One Tenth	deci-
$10^{-2} = 0.01$	One Hundredth	centi-
$10^{-3} = 0.001$	One Thousandth	milli-
$10^{-6} = 0.000001$	One Millionth	micro-
$10^{-9} = 0.000000001$	One Billionth	nano-
10^{-12}	One Trillionth	pico-
10^{-15}	One Quadrillionth	femto-
10^{-18}	One Quintillionth	atto-

Figure 2.9[36]

Sometimes when working with rates, proportions, and percentages, the process can be made more challenging by the magnitude of the numbers involved. Sometimes large numbers are just difficult to comprehend.

[35] Really Big and Really Small Numbers. (n.d.). *Varsity Tutors*. https://www.varsitytutors.com/hotmath/hotmath_help/topics/big-and-small-numbers

[36] Really Big and Really Small Numbers. (n.d.). *Varsity Tutors*. https://www.varsitytutors.com/hotmath/hotmath_help/topics/big-and-small-numbers

The tables above list a lot of numbers and prefixes. For everyday use, it may be helpful if you know how many 0's there are in some large numbers you may encounter in real life, in the media, and in your professional fields. While there are bigger numbers (as shown above), most real-life situations (e.g., federal budgets) can be described by these.

- A million: 1,000,000 (There are six 0's. It's a thousand times a thousand.)
- A billion: 1,000,000,000 (There are nine 0's. It's a thousand times a million.)
- A trillion: 1,000,000,000,000 (There are twelve 0's. It's a thousand times a billion.)

When dealing with these numbers, which are difficult for most of us to imagine, we often use comparisons so we can grasp just how big these numbers are. Another helpful hint: when dividing a large number by another large number, the same number of 0's can be "cancelled" on the top and the bottom of a fraction.

Example 1

What is 100 billion divided by 5 million?

[Solution]

We are asked to find $\frac{100,000,000,000}{5,000,000}$. While you *can* enter these two numbers on your calculator to find the answer directly, it is possible to cancel six 0's from the numerator and the denominator, so the answer is the same as the quotient $100{,}000 \div 5$, which is 20,000.

Example 2

The 2010 U.S. military budget was $683.7 billion.

(1) Compare this number to the 1.4 million Walmart employees nationwide.
(2) If each of the 330 million people in America splits the military budget equally, how much would it be per person? (How much is every American paying for the military?)
(3) If you were to count one $100 bill every second, how long will it take to count the 2010 U.S. military budget?
(4) Which comparison do you think would make most sense?

[Solution]

Here we have a very large number, $683,700,000,000.

(1) When you divide 683.7 billion by 1.4 million, you can simplify the operation by "cancelling millions." Since 683.7 billion is 683,700 million, this is equivalent to

$$683{,}700 \div 1.4 \approx 488{,}357.$$

Hence, if the amount of the 2010 U.S. military budget were used to pay the 1.4 million Walmart employees in the U.S., each would earn over $488,000.

(2) Again, cancelling millions, we can perform the following division:

$$683{,}700 \div 330 \approx 2072.$$

The 2010 U.S. military budget was over $2,000 per person.

(3) Converting seconds to years, we see the following:

$$\frac{\$683.7 \text{ billion}}{\$100 \text{ / sec}} \times \frac{1 \text{ min}}{60 \text{ sec}} \times \frac{1 \text{ hr}}{60 \text{ min}} \times \frac{1 \text{ day}}{24 \text{ hr}} \times \frac{1 \text{ yr}}{365 \text{ days}} \approx 216.80 \text{ (yrs)}.$$

If you were to put $683.7 billion in $100 bills and count out one per second, it would take you over 216 years to finish counting it.

(4) Which one makes most sense is a subjective question, and there is no one correct answer here. However, counting $100 bills every second sounds a little artificial, and some may question why one particular retail chain was chosen for our first comparison. It may be argued that "costing over $2000 per person" may be the easiest one to related to.

Example 3

The One Trillion Trees Initiative of 2020 was mentioned at the beginning of this section. If one tree can be planted every second, around the clock, every year, how many years would it take for this goal to be achieved? Is it possible to reach the goal by 2030?

[Solution]

$$\frac{10^{12} \text{ trees}}{1 \text{ tree / sec}} \times \frac{1 \text{ min}}{60 \text{ sec}} \times \frac{1 \text{ hr}}{60 \text{ min}} \times \frac{1 \text{ day}}{24 \text{ hr}} \times \frac{1 \text{ yr}}{365 \text{ days}} \times \approx 31{,}709.8 \text{ (yrs)}.$$

This shows that the goal cannot be achieved by 2030. In fact, it will take over 31,700 years to reach one trillion trees. Similar calculations show that, if the goal is to be reached, then over 3,170 trees need to be planted around the world every second, every day for the next 10 years. (We'd better hurry.)

In this chapter, we have seen many different aspects of numbers—percentages, deceptive numbers, proportions, conversions, estimates, small and large numbers. Some numbers are hard to understand, and some can be quite confusing. For these reasons, numbers and figures are often used to present information that appear quite objective and truthful, yet they could easily be manipulated to provide deception, exaggeration, and misinformation. How do you guard yourself against such abuse by means of numbers? Just as in Chapter 1, the answer is to

exercise critical-thinking skills and get familiar with these ways in which some people can try to form opinions for you. In this chapter, we have presented various ways in which this could happen.

There are many other ways wherein numbers can deceive. Some of those will be discussed in later chapters on finance, statistics, growth, and politics. What has been presented in this chapter is just a small portion of number manipulations. The message is clear: we must all be numerically literate and quantitatively intelligent to make wise decisions in this data-driven world.

Exercises 2.5

1. The Deepwater Horizon oil spill (in 2010) resulted in 4.9 million barrels of oil spilling into the Gulf of Mexico. Each barrel of oil can be processed into about 19 gallons of gasoline. How many cars could this have fueled for a year? Assume an average car gets 20 miles to the gallon and drives about 12,000 miles in a year.

2. A box of Jell-O costs $0.50 and makes 2 cups. How much would it cost to fill a swimming pool 4 feet deep, 8 feet wide, and 12 feet long with Jell-O? (1 cubic foot is about 7.5 gallons).

3. The sun is about 93 million miles from the earth while the moon is about 240,000 miles away (on the average). What percentages can you use to compare these two distances?

4. The population of the U.S. is about 332,200,000, covering a land area of 3,797,000 square miles. The population of India is about 1,388,700,000, covering a land area of 1,269,000 square miles. Compare the population densities of the two countries.

5. The GDP (Gross Domestic Product) of China was $14.34 trillion in 2019, and the GDP of Sweden was $531 billion. The population of China was about 1.40 billion while the population of Sweden was about 10.2 million. Compare the GDP per capita (GDP per person) of the two countries.

6. In 2020, Twitter was reporting about 500 million tweets per day. Each tweet can consist of up to 280 characters. Create a comparison to help understand the amount tweeted in a year, assuming that 20 drops of water make up about 1 milliliter and that an Olympic-size pool contains 2.5 million liters of water.

7. The photo-sharing site Flickr had 6 billion images in August 2011. Assuming each picture is about 2 megabytes in size, compare the amount of data at Flickr to the data stored on an 8GB flash drive. How many flash drives are needed to store all the images?
8. Define a “millionaire.” Does it depend on what currency is used?

9. You are concerned that your friend is taking too much ibuprofen for her headache, but she says, “I never take any more than 1.3 grams of the stuff. Some people take like 900 mg.” What would you say to her?

10. Which is harder? Walking 1,000,000 inches or walking 20 miles? Does one “sound” harder than the other?

11. [The “Hair Problem”] Here is a bizarre but enlightening question that illustrates the power of reasoning and counting: are there two people in the world with exactly the same number of hairs? Follow the steps below to determine the answer.
    (a) An average person has about 1600 hairs per square inch on the head. No one has more than 3000 per square inch. An average person has about 2500 square inches of skin. No one has more than 4000 square inches of skin. Based on these numbers, calculate an upper bound for the number of hairs a person can have.
    (b) Describe a person who would have that many hairs (the answer to Part (a) of this question) if such a person existed. What must be true about that person?
    (c) Using an online resource, find out how many people are in the world.
    (d) Based on these answers, can you determine if there are at least two people in the world with exactly the same number of hairs? Explain your reasoning. (Imagine there are a bunch of rooms labeled 0, 1, 2, …, all the way to that upper bound you found in Part (a). Then, imagine every person in the world entering the room whose label is the number of hairs that the person has. Would there be any room with more than one person? This is an application of what is known as the **Pigeonhole Principle** in discrete/finite mathematics.)

For Further Explorations…

Some Big Numbers Indeed[37]

Do you ever shuffle your playlist while listening to music? Have you ever wondered how many different orders are possible when you randomize the songs on your playlist of, say, 20 songs? Many people like to use the "random play" feature because they want to enjoy a variety of songs instead of the first few songs every time they open their audio app. So how likely is it that you and your friend "shuffle-play" and listen to the entire playlist (20 songs) in exactly the same order? Not likely, right? (And you will be absolutely right as explained below.)

Figure 2.11[38]

We will formally learn how to answer these questions in Section 4.3 (this particular type of problem is called *permutations*), but the number of different orders in which 20 songs can be played turns out to be a rather big number:

$$20! = 20 \times 19 \times 18 \times \ldots \times 3 \times 2 \times 1 \approx 2.4320 \times 10^{18}.$$

What is this number? It is 2.432 quintillion, or, if you prefer, 2.432 billion billion.

Cool.

Sure, but it is hard to grasp just how big this number is, isn't it? Let us make some comparisons.

Astronomers tell us that our universe is about 13.77 billion years old (give or take a few billion). If you multiply 13.77 billion by 365, and by 24, and then by 3600, we see that the age

[37] Some ideas of this essay were presented at a conference in Pomona, CA, on March 4, 2017, by a Harvey Mudd College Professor Francis Su, then President of the Mathematical Association of America and the author of the book *Mathematics for Human Flourishing* (2020).

[38] Photo by Hans-B. Sickler is licensed under CC BY 2.0

of the universe is about 4.3425×10^{17} seconds (again, give or take a few). This is the same as 434.25×10^{15}; in other words, according to science, the universe is about 434 quadrillion seconds old. But note that when you divide this number by $20!$ (the number of different permutations of those 20 songs), the quotient is about 0.178, or 17.8%.

In other words, if your audio player had been picking a random permutation (ordering) of those 20 songs on the playlist every second, 24/7, all year long, every year, for the last 13.77 billion years since the beginning of the universe, it would have chosen less than 20% of those possible permutations. It would then take another 63.38 billion years (theoretically) to play every permutation! (The exclamation mark here is used intentionally.)

So, going back to the original problem of how likely it is for you and your friend to get exactly the same permutation of 20 songs—almost impossible. The chances are 1 in 2.432 quintillion. In fact, from this perspective, you can say that every time you shuffle-play that playlist, you are "making history."

Figure 2.12: Spiral Galaxy NGC4414[39]

Similarly, a standard deck of playing cards (without the two jokers) consists of 52 cards. If you shuffle a deck of cards thoroughly, how many possible outcomes are there? In other words, in how many orders can the 52 cards be stacked?

This is a question on similar to the first one, and the number of possible permutations is this number:

$$52! = 52 \times 51 \times 50 \times \ldots \times 3 \times 2 \times 1 \approx 8.0658 \times 10^{67}.$$

[39] Photo by NASA & ESA is in the public domain.

This is a much bigger number. Just as a reminder, 10^{67} is the number beginning with a 1 and having 67 zeros after that. This is beyond any number we stated in Section 2.5, and we do not even have a name for such a big number. So, once again, let us try to make some comparisons.

As calculated above, the universe is about 4.3425×10^{17} seconds old. Today, there are almost 8 billion people on our planet. According to a 2016 report, scientists believe there are 2 trillion galaxies.[40] If every one of the 8 billion people had been shuffling a whole deck, every second, around the clock, since the beginning of the universe, and if this had been happening in each of the 2 trillion galaxies (I know—this is getting ridiculous), there would have been

$$(4.3425 \times 10^{17}) \times (8 \times 10^{9}) \times (2 \times 10^{12}) \approx 6.96 \times 10^{39}$$

permutations (orderings) of the 52 cards, assuming that no ordering is repeated. But, dividing $52!$ by this number, we see that the total number of possible permutations is 11.5×10^{27} times as big! In other words, $52!$ Is more than 11 billion billion billion times as big as the number of card orderings that have occurred in 2 trillion galaxies, each with 8 billion people shuffling cards every second since the beginning of time.

Now, *these* are big numbers indeed.

No wonder these numbers are called "astronomical."

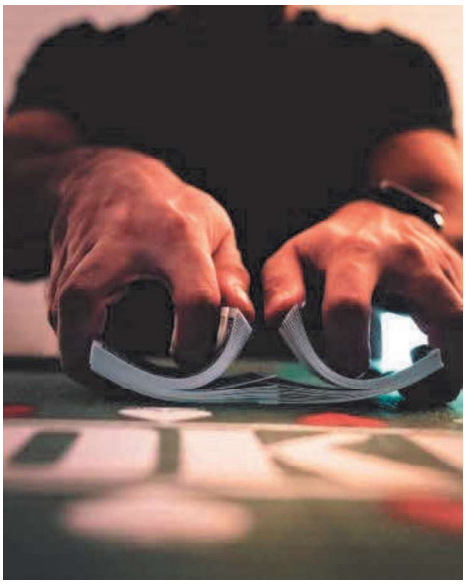


Figure 2.13 [41]

[40] Fountain, H. (2016). *Two Trillion Galaxies, at the Very Least.* The New York Times. https://www.nytimes.com/2016/10/18/science/two-trillion-galaxies-at-the-very-least.html. (However, this number was reduced to 200 billion five years later.)

[41] Photo by Amol Tyagi on Unsplash

Chapter 3: Mathematics and Finance

"Blessed are the young, for they shall inherit the national debt."—
Herbert Hoover, the 31st President of the United States

Figure 3.1: A portrait of President Herbert Hoover[42]

We have to work with money every day. Balancing your checkbook or calculating your monthly expenditures on espresso requires only arithmetic, but we need more complex mathematics when we start saving, plan for retirement, or need a loan.

In this chapter, we study mathematics of finance. It involves interest, whether you are saving money or borrowing money. The most fundamental idea of this chapter is the **time value of money**; because of the potential earning capacity of money, i.e., earning interest, $100 today is worth more than $100 in the future. Yes, money makes money. Keep this in mind: **money is time-sensitive.** The entire field of mathematics of finance depends on this one basic principle. This is why we use terms like *present* values as opposed to *future* values.

We will then conclude the chapter with a brief section on spending and borrowing by governments, which involves national deficits and debts. It is interesting that the above quotation by President Hoover is from almost 100 years ago.

[42] Photo by the Library of Congress on Unsplash

3.1 Simple and Compound Interest [v]

There are basically two types of interest: **simple** and **compound**. We will spend most of our time on compound interest because of its significance. The money deposited or borrowed at the beginning is called the **principal** (note the spelling), denoted P. This is the initial amount that earns interest. The amount of **interest** earned is expressed by I, which depends on the **interest rate** r (written as a decimal) and often on **time** t. The balance in the future is obviously the sum of the principal and the interest, and this is called the (future) **amount**, denoted by A.

In Sections 3.1 and 3.2, we only consider situations where money is deposited **once** at the beginning. In later sections we study plans where a series of regular deposits or payments are made. Let us begin with simple interest.

Simple Interest

The principal P_0 could be a starting investment or the starting amount of a loan. Interest, in its most simple form, is calculated as a percent of the principal. For example, if you borrowed \$100 from a friend and agree to repay it with 5% interest, then the amount of interest you would pay is 5% of 100: $\$100 \times 0.05 = \$5$. The total amount you would repay would be \$105, the original principal plus the interest. Note that, as before, you can calculate this total as 105% of the principal:

$$A = 100 + (100 \times 0.05) = 100 \times 1.05 = 105.$$

Simple One-Time Interest

$$I = Pr$$
$$A = P + I = P + Pr = P(1 + r)$$

I is the interest; A is the end (future) amount (principal plus interest).
P is the principal (starting amount); r is the interest rate (in decimal, e.g., $5\% = 0.05$).

Example 1

A friend asks you if he could borrow \$300 and agrees to repay it in 30 days with 3% interest. How much interest will you earn?

[Solution]

The principal is $P = \$300$, and the rate is $r = 0.03$, so $I = 300 \times 0.03 = 9$. You will earn \$9 interest.

One-time simple interest is only common for extremely short-term loans. For longer term loans, it is common for interest to be paid on a daily, monthly, quarterly, or annual basis. In that case, interest would be earned regularly. For example, bonds are essentially a loan made to the bond issuer (a company or government) by you, the bond holder. In return for the loan,

the issuer agrees to pay interest, often annually. Bonds have a maturity date, at which time the issuer pays back the original bond value.

Example 2

Suppose your city is building a new park and issues bonds to raise the money to build it. You obtain a $1,000 bond that pays 5% interest annually that matures in 5 years. How much interest will you earn?

Each year, you would earn 5% interest: $\$1000 \times 0.05 = \$50$ in interest. So over the course of five years, you would earn a total of $250 in interest. When the bond matures, you would receive back the $1,000 you originally paid, leaving you with a total of $1,250. ■

We can generalize this idea of simple interest over time t. Since the amount of simple interest stays the same each time, we multiply the one-time interest amount $I = Pr$ by time t.

Simple Interest over Time

$$\boldsymbol{I = Prt}$$
$$\boldsymbol{A = P + I = P + Prt = P(1 + rt)}$$

I is the interest; A is the end (future) amount (principal plus interest).
P is the principal (starting amount), r is the interest rate in decimal, and t is time.

In the above example, we assumed that interest was paid **annually**; frequently, however, interest is calculated and paid (or added) more often, such as **monthly** or **quarterly**. So the units of time (year, months, etc.) in the above equation should match the time period for the interest rate.

To avoid confusion, interest rates are almost always given as an **annual percentage rate** (**APR**), the total rate for the year. If the interest is paid in shorter time increments, the APR is divided by the number of interest periods in a year. For example, a 6% APR paid **monthly** would be divided into twelve 0.5% payments: $0.06 \div 12 = 0.005$. A 4% annual rate paid **quarterly** would be divided into four 1% payments: $0.04 \div 4 = 0.01$.

Example 3

Treasury notes (T-notes) are bonds issued by the federal government to cover its expenses. Suppose you obtain a $1000 T-note with a 4% annual rate, paid semi-annually (twice a year), with a maturity in 4 years. How much interest will you earn?

[Solution]

Since interest is being paid semi-annually, the 4% interest will be divided into two 2% payments (2% of $1000 is $20). Note that now there are eight $20 payments. We have P= $1000, and

$I = \$1000 \times 0.02 \times 8 = \$160$. You will earn $160 interest total over the four years.

Note that if this were paid annually (once a year), you will get four payments of $40, so the total amount of interest is the same. This is one of the characteristics of simple interest.

Compound Interest

With simple interest, we were assuming that the amount of interest stays the same each period. In a standard bank account, however, any interest we earn is automatically added to our balance, and we earn interest on that interest in future years. This reinvestment of interest is called compounding.

Virtually everything in the world of finance, including personal savings, loans, and mortgages, is calculated using compound interest. Because the balance changes every interest period (referred to as a "**compounding period**"), *the amount of compound interest changes every time as well*; this is the most obvious difference between compound interest and simple interest studied above.

Example 4

Suppose that we deposit $1000 in an investment account earning 12% interest compounded annually. How will our money grow?

[Solution]

Note that the principal is $1000. After one year, 12% of $1000 is added to the account, making the balance $1120 (since 12% of $1000 is $120). If it were simple interest, the interest stays the same, at $120 *every year*. However, with compound interest, the amount of interest for the second year is based on the new balance of $1120. Because 12% of $1120 is $134.40, we will have

$$\$1120 + \$134.40 = \$1254.40.$$

Then, for the third year, we add 12% of $1254.40 (which is $150.53 after rounding) to this amount, resulting in

$$\$1254.40 + \$150.53 = \$1404.93.$$

Note that the amount of interest itself is increasing every year: $120, $134.40, $150.53, etc. Also, you may have noticed that the balance at the end of each year is 112% of the previous year's balance (since the new balance is always 12% *more than* the previous balance). This pattern gives us a general formula (which will be more fully develop below):

Principal: $P = \$1000$
After 1 year: $A = \$1000 \times 1.12$
After 2 years: $A = \$1000 \times 1.12 \times 1.12 = \$1000 \times (1.12)^2$
After 3 years: $A = \$1000 \times 1.12 \times 1.12 \times 1.12 = \$1000 \times (1.12)^3$
So after t years, the amount in the account will be $A = \$1000 \times (1.12)^t$.

In other words, we find "1 + (the rate per period)," which is 1.12 in this example, and raise it to the number of compounding periods. This factor is to be multiplied by the principal. We will generalize this concept below.

Compound Interest (for **Annual Compounding** only):

$$A = P(1 + r)^t$$

A is the balance (future amount) in the account after t years. P is the principal.

r is the annual interest rate (APR) in decimal form.

It is very important to remember that *this formula applies only when the compounding occurs annually* (which is not very often). Think of this formula as a starting point, a stepping stone so that you can understand the next, more general, formula.

As stated above, it is very common for financial institutions to pay interest more than once a year. The frequency is described by adverbs like "**annually**" (once a year), "**semiannually**" (every 6 months), "**quarterly**" (every 3 months), "**monthly**" (every month), and "**daily**" (every day). When dealing with these non-annual periods, you must exercise some care.

(1) Divide r(APR) by the number of interest periods per year (4 if quarterly; 12 if monthly), and use this number, which is essentially the "interest rate per period."
(2) Adjust t (time) as the total number of interest periods. For instance, "10 years" would be $10 \times 4 = 40$ periods under "quarterly" compounding but $10 \times 12 = 120$ periods under "monthly" compounding. Therefore, the exponent in the above formula should represent the "total number of interest periods."

Let us see some examples.

Example 5

Suppose that we deposit $1000 as before, but the account earns 12% interest *compounded monthly* instead of annually. How would the money grow now?

[Solution]

As explained above, "12% compounding monthly" technically means 1% every month because $12\% \div 12 = 1\% = 0.01$. So the balance increases by 1% (0.01 in decimal) every month, or (better yet) it becomes $1 + 0.01 = 1.01$ times the previous balance. Let us first calculate what happens, step by step, month by month, to the initial principal of $1000.

In the first month, $P = \$1000$ earns 1% interest: $\$1000 \times 0.01 = \$10$. This interest is added to the principal, so the balance after one month is $\$1000 + \$10 = \$1010$. (Or you can calculate 101% of $1000 more directly as $\$1000 \times 1.01 = \$1010$.)

In the second month, 1% of this new balance, $\$1010 \times 0.01 = \$10.10$, is added to the previous balance, so after two months, the balance is $\$1010 + \$10.10 = \$1020.10$. (Or you can calculate this more directly as $\$1000 \times (1.01)^2 = \$1020.10$.)

In the third month, 1% of this new balance, $\$1020.10 \times 0.01 = \$10.20$ (after rounding to the nearest penny), is added to the previous balance, so after three months, the balance is $\$1020.10 + \$10.20 = \$1030.30$. (Or more directly as $\$1000 \times (1.01)^3 = \$1030.30$.)

Calculating out a few more months, we get the following table:

Month	Starting balance	Interest earned	Ending Balance
1	1000.00	10.00	1010.00
2	1010.00	10.10	1020.10
3	1020.10	10.20	1030.30
4	1030.30	10.30	1040.60
5	1040.60	10.41	1051.01
6	1051.01	10.51	1061.52
7	1061.52	10.62	1072.14
8	1072.14	10.72	1082.86
9	1082.86	10.83	1093.69
10	1093.69	10.94	1104.62
11	1104.62	11.05	1115.67
12	1115.67	11.16	1126.83

In the last two examples, we compared annual growth and monthly growth at the same APR. A bit of warning first: in the world of finance, even though this rate, 12%, is referred to as the "**annual percentage rate** (APR)," technically is it not the percentage rate that the money really earns every year. (You may be in disbelief and say, "Seriously?" Yes, I *am* saying that the "annual percentage rate" does not mean the actual percentage rate for the year.) We will explain this below. The convention used in business and finance is, as described under simple interest, that *we always find the rate per period* and use that rate instead of this APR. As illustrated above, "12% compounding monthly" means, more accurately, that this account earns 1% interest every month. But traditionally, since everyone in the industry seems to prefer using *annual* percentage rates, we refer to this as "12% compounded monthly." Note that this definitely does NOT mean that the money is increasing by 12% every month (that would be a scheme "way too good to be true").

In short, therefore, we always calculate using the "rate per period," which is $\frac{r}{k}$ where r is the annual percentage rate and k is the number of compounding periods per year. It means that $k = 12$ if the compounding occurs monthly. Obviously, k =1 if the compounding occurs annually (once a year).

Now, look at the previous example ("12% monthly"). Note that after 12 months, i.e., one year, the balance turns out to be $1126.83, which is $126.83 more than the original principal of $1000. Since this amounts to precisely 12.683% of $1000, the *true* annual percentage earned in the year is 12.683%, not the APR, which is 12%. Thus, the term "annual percentage rate" *r* is a bit misleading because the true percentage earned in a year is always slightly more than *r*. This *true* annual rate is referred to as the **annual percentage yield** and is abbreviated "**APY**." It is also called the "effective rate of return."

Let us compare this balance with the other example, where the compounding was 12% *annually*. The balance after 1 year with annual compounding was $1,120.00 (precisely 12% more than $1000) whereas it was $1,126.83 with monthly compounding. Remember that the APR was 12% in both of these examples. This shows that **the frequency matters**. It can be shown that more frequently the compounding occurs, the more interest the money earns.

At this point, you may say that the difference is very slim. After all, the difference between these two balances is $6.83 after 12 long months, not quite enough for two café lattes. Well, as you will see below, the difference becomes enormous after a long period of time.

General Formula for Compound Interest

Let us find an equation to represent this growth; if A_m represents the future amount (balance) after m months, then we can derive our formula as follows. We will use the "12% monthly" example. Recall that each month, the balance is 1.01 times (101% of) the previous balance. This is a type of exponential growth (to be studied in a later chapter).

$P = 1000$ (the principal, the original deposit)

$A_1 = 1000 \times 1.01 = 1010$
$A_2 = 1000 \times 1.01 \times 1.01 = 1000 \times (1.01)^2 = 1020.10$
$A_3 = 1000 \times 1.01 \times 1.01 \times 1.01 = 1000 \times (1.01)^3 = 1030.30$

Observing a pattern, we could conclude:

A_m=$1000 \times (1.01)^m$.

Notice that we found 1.01 by dividing the APR $r = 0.12$ by $k = 12$ and adding 1 to that quotient (since we were compounding 12 times per year).

Generalizing our result, we reach the following formula.

$$A_m = P\left(1 + \frac{r}{k}\right)^m$$

where r is the APR, k is the number of compounding periods per year, and m is the number of compounding periods (months in this example).

While this formula works fine, it is more common to use a formula that involves the number of years rather than the number of compounding periods. If t is the number of *years*, then $k \times t = kt$ is the total number of compounding periods (written as m above). Making this change gives us the standard formula for compound interest.

General Formula for Compound Interest

$$A_t = P\left(1 + \frac{r}{k}\right)^{kt}$$

A_t is the amount (balance) in the account after t years. P is the principal.

r is the annual interest rate (APR); k is the number of compounding periods in one year.

k= 1 means annual compounding. Other common frequencies used are k = 4 (quarterly), k = 12 (monthly), and k = 365 (daily). Theoretically, compounding could occur more frequently, like every hour, every minute, every second, and even continuously. This will be explored in the next section.

Example 6

In the previous examples, we calculated the balance after 1 year when the principal was $1000 and the APR was 12% (annually in the first example and monthly in the second). Now, compare the balance after 10 years and 20 years. Also, compare these balances with the balance if the account gives simple interest at 12%.

[Solution]

Note that $P = \$1000$ and $r = 0.12$ in each case.

(1) First, let us consider the balance at 10 years (t $= 10$).

- 12% annual compounding:

$$A_{10} = \$1000 \times \left(1 + \frac{0.12}{1}\right)^{1\times 10} = \$3{,}105.85.$$

- 12% monthly compounding:

$$A_{10} = \$1000 \times \left(1 + \frac{0.12}{12}\right)^{12\times 10} = \$3{,}300.39.$$

- 12% simple interest:

$$A_{10} = \$1000 \times (1 + (0.12 \times 10)) = \$2{,}200.00.$$

(2) Let us now consider the balance after 20 years ($t = 20$).

- 12% annual compounding:

$$A_{20} = \$1000 \times \left(1 + \frac{0.12}{1}\right)^{1\times 20} = \$9{,}646.29.$$

- 12% monthly compounding:

$$A_{20} = \$1000 \times \left(1 + \frac{0.12}{12}\right)^{12\times 20} = \$10{,}892.55.$$

- 12% simple interest:

$$A_{20} = \$1000 \times (1 + (0.12 \times 20)) = \$3{,}400.00.$$

Now, compare and contrast these figures again. You can easily see that, after many years, there are significant differences between annual compounding and monthly compounding and even a more drastic difference between these and simple interest. This example shows the power of compound interest. This power works *for you* if you are saving money (but works *against you* if you are borrowing).

Evaluating Exponents on the Calculator

When we need to calculate something like 5^3, it is easy enough to just multiply $5 \times 5 \times 5 = 125$. But when we need to calculate something like $(1.01)^{240}$, it would be very tedious to multiply 1.01 by itself 240 times! Use a scientific calculator.

All scientific calculators have a button for exponents. It is typically labeled like

^ , x^y, or y^x.

To evaluate $(1.01)^{240}$ we would type 1.01 ^ 240, or 1.01 y^x 240. Try it out – you should get something around 10.89255365.

Example 7

A certificate of deposit (CD) is a savings instrument that many banks offer. It usually gives a higher interest rate, but you cannot access your investment for a specified length of time. Suppose you deposit $3000 in a CD paying 6% interest, compounded monthly. How much will you have in the account after 20 years?

[Solution]

In this example, P=$3000, r=0.06, k=12, and $t = 20$, so

$$A_{20} = \$3000\left(1 + \frac{0.06}{12}\right)^{12\times 20} = \$3000 \times 3.310204\ldots = \$9930.61$$

after rounding to the nearest penny.

Figure 3.2[43]

[43] Image by Micheile on Unsplash

Let us compare the amount of money earned from compounding against the amount you would earn from simple interest.

Years	Simple Interest ($15 per month*	6% compounded monthly (0.5% each month)
5	$3900	$4046.55
10	$4800	$5458.19
15	$5700	$7362.28
20	$6600	$9930.61
25	$7500	$13,394.91
30	$8400	$18,067.73
35	$9300	$24,370.65

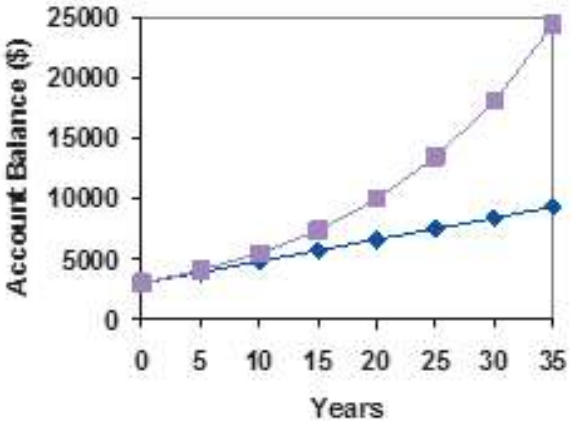


Figure 3.2[44]

Again, you can see that over a long period of time, compounding makes a huge difference in the account balance. You may want to keep this in mind as a main difference between linear growth and exponential growth. We will see this in Ch. 6 again.

[44] Compound Interest by David Lippman is licensed under CC BY-SA 3.0 US.

Using Your Calculator

In many cases, you can avoid rounding completely by how you enter things in your calculator. For example, suppose you need to calculate the following:

$$A_{30} = 1000\left(1 + \frac{0.05}{12}\right)^{12\times 30}$$

We can quickly calculate $12 \times 30 = 360$, giving us $A_{30} = 1000\left(1 + \frac{0.05}{12}\right)^{360}$.

Now we can use the calculator as follows:

Type this	Calculator shows
0.05 ÷ 12 =	0.00416666666667
+ 1 =	1.00416666666667
y^x 360 =	4.46774431400613
× 1000 =	4467.74431400613

(This would be the balance after 30 years if $1000 is deposited with 5% compounding monthly. Be sure to find the quantity inside the parentheses first.)

Example 8

You know that you will need $40,000 for your child's education in 18 years. If your account earns 4% compounded quarterly, how much would you need to deposit now to reach your goal?

[Solution]

In this example, we are looking for P, given r=0.04, k=4 (since the compounding is quarterly), $t = 18$, and A_{18}=$40,000. In this case, we have to set up the equation and solve for P.

$$40{,}000 = P\left(1 + \frac{0.04}{4}\right)^{4\times 18} = P \times 2.047099.$$

Dividing both sides of the equation by 2.047099, we obtain P= 19,539.84. So you would need to deposit $19,539.84 now to have $40,000 in 18 years.

Using Your Calculator with Parentheses

Some advanced calculators allow you to type in the entire expression to be evaluated with parentheses. If you have a calculator like that, you will just need to enter the following to find the answer to this last example.

$$40000 \div (1 + 0.04 \div 4)\ x^y\ (4 \times 18) =$$

These smart calculators know the correct order of operations.

Exercises 3.1

1. A friend lends you $200 for a week, which you agree to repay with 5% one-time interest. How much will you have to repay?

2. Suppose you obtain a $3,000 T-note with a 3% annual rate with maturity in 5 years. How much interest will you earn?

3. A treasury bill (T-bill) is a type of bond that is sold at a discount over the face value. Suppose you buy a 13-week T-bill with a face value of $10,000 for $9,800. This means that in 13 weeks, you can redeem it for the face value, earning you $200. What annual interest rate do you earn? (Hint: 13 weeks is a quarter of a year.)

4. You deposit $300 in an account earning 5% interest compounded annually. How much will you have in the account in 10 years?

5. You deposit $1000 in an account earning 7% interest compounded annually. What will be balance be in 20 years?

6. You deposit $1000 in an account earning 6% interest compounded monthly. When does the amount double? Do this by trial-and-error. (Try a few exponents and estimate.)

7. You deposit $2000 in an account earning 3% interest compounded monthly.
   (a) How much will you have in the account in 20 years?
   (b) How much interest will you earn?

8. You deposit $10,000 in an account earning 4% interest compounded quarterly.
   (a) How much will you have in the account in 25 years?
   (b) How much interest will you earn?

9. How much would you need to deposit now in order to have $6,000 in 8 years? Assume the account earns 6% interest compounded monthly.

10. How much would you need to deposit now in order to have $20,000 in the account in 4 years? Assume the account earns 5% interest compounded quarterly.

11. You have $2,000 to invest and want it to grow to $3,000 in two years. What interest rate (compounded annually) would you need to find to make this possible? Use trial-and-error (unless you are familiar with how to solve radical equations).

12. You have $5,000 to invest and want it to grow to $20,000 in ten years. What interest rate (compounded monthly) would you need to find to make this possible? Use trial-and-error (unless you are familiar with how to solve radical equations).

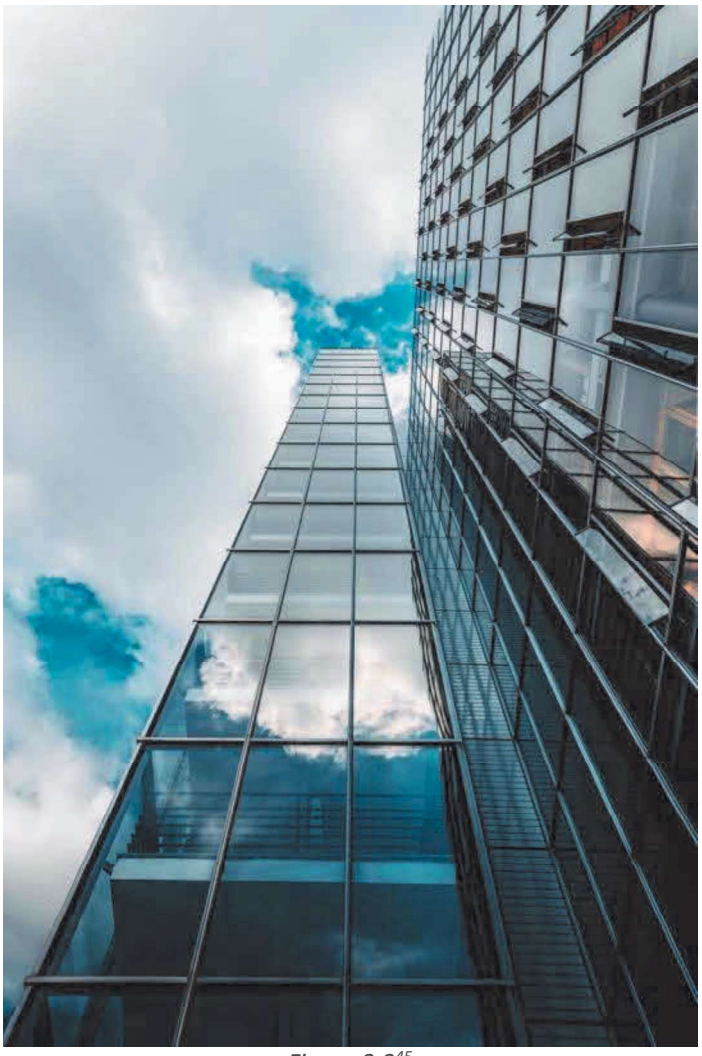


Figure 3.3[45]

[45] Image by Christian Wiediger on Unsplash

3.2 Continuous Compounding[vi]

We continue the discussion that was started in the previous section. So far, we have only studied an investment where you start with some money (the **principal**), which is a one-time deposit, and watch it grow with a fixed interest **rate** (APR) and **frequency** (defined by **compounding period**). Compound interest, by definition, is interest calculated on the principal amount together with accumulated interest.

More and More Frequent Compounding

We have already seen that the higher the frequency is (i.e., the shorter each compounding period is), the larger the final balance is. Even with the same APR, monthly compounding ($k = 12$) gives you more interest than quarterly ($k = 4$) or annual (k=1) compounding.

What happens as you increase this number (k), the number of times interest is compounded each year?

Your earnings will get bigger, but by how much? Will your earnings increase without bound?

Equivalently, what happens as you shorten each compounding period? In other words, what happens as you calculate compound interest every day, every hour, every minute, every second, every hundredth of a second? (You will be very tired.)

The amount of money produced under these situations does *not* increase without bound! (Wishful thinking, but it does not happen.) Instead, the irrational number e emerges to describe the limiting amount earned. The resulting formula is called the **Continuous Compounding Formula** and is the subject of this section.

In the last section, we learned that compound interest is a type of exponential growth and that the general formula for the (future) amount, A_t (after t years) is

$$A_t = P\left(1 + \frac{r}{k}\right)^{kt}$$

where P is the principal (amount of initial deposit), r is the APR, k is the number of compounding periods per year, and t is the number of years.

An Indeterminant Form

Now, we want to explore what happens as the compounding becomes more and more frequent, every second, fraction of a second, and eventually to "continuous compounding." This means that our value k goes to infinity.

As k gets larger, $\frac{r}{k}$ gets smaller, approaching 0. That is, as $k \to \infty$, $\frac{r}{k} \to 0$.

Now, as $\frac{r}{k} \to 0$, clearly $1 + \frac{r}{k} \to 1$, which is the quantity inside the parentheses.

On the other hand (or the outside the parentheses), as k gets larger, kt gets bigger. That is, as $k \to \infty,\ kt \to \infty$.

So, as compounding gets more frequent (letting $k \to \infty$), here is what happens to our formula:

$$A_t = P\left(1 + \frac{r}{k}\right)^{kt} \text{ becomes more like } A_t = "P\ (1)^{\infty}."$$

This last expression is put in quotation marks because this is not a proper mathematical expression. The form 1^{∞} is called an **indeterminant form** and is not well-defined. Of course, we do not really mean to take 1 to the infinite power. We cannot.

It turns out that there are lots of things that make the '1^{∞}' form both interesting and difficult to analyze. In particular, all of the following are true:

- 1 to any finite power equals 1. For example, $1^{1,000,000,000} = 1$.
- When a number a bit less than 1 is raised to higher and higher powers, it approaches zero.
- For example, $\left(\frac{99}{100}\right)^n$ approaches 0 as n approaches infinity. (This is an exponential function with base between zero and one and describes "exponential decay.")
- When a number a bit more than 1 is raised to higher and higher powers, it approaches infinity.

  For example, $\left(\frac{101}{100}\right)^n$ approaches ∞ as n approaches infinity. (This is an exponential function with base greater than one and describes "exponential growth.")

Depending on precisely *how* the base approaches 1 and *how* the exponent approaches infinity, the form 1^{∞} can approach different numbers (thus the term "indeterminant form").

An **indeterminate form** occurs when you cannot figure out what is happening without further analysis.

That is, each occurrence of must be investigated separately. So what happens in our case? As you saw above, the number inside, the base number, approaches 1 but remains slightly bigger than 1. If you take any number bigger than 1 to large powers, it will get infinitely bigger (the last bullet point above). So this expression approaches infinity, right? Not so fast (literally). Because 1 raised to any power is always 1 (the first bullet point).

Here is a way to look at this dilemma; some people call it the "Peter Pan" dilemma. The base number (the number inside the parentheses), which is $(1+\frac{r}{k})$, gets closer and closer to 1, wanting to stay "small," like 1. But the exponent, kt, wants to stretch the indeterminant form bigger and bigger, to infinity. So we have an internal force wishing to stay small like a child while the external exponent is trying to make it grow up to be an adult. This "tug of war" is the Peter Pan dilemma.

"Who wins?" You may ask. Well, neither does. But they come to a beautiful compromise. As k goes to infinity, mathematicians have figured out *exactly* what happens to this particular indeterminant form.

For us, this investigation involves the following definition of the irrational number e.

Definition of e

The real number e is defined as the limit of the quantity $\left(1+\frac{1}{n}\right)^n$ as n approaches infinity. That is,

$$\left(1+\frac{1}{n}\right)^n \to e \quad as \;\; n \to \infty.$$

This number has been extensively studied. It is known that, like π, this number is irrational and thus is not a terminating decimal or a repeating decimal.

$$e \approx 2.71828182845904523 \ldots$$

This number e is arguably the most important real number and plays an absolutely crucial role in calculus, engineering, probability, statistics, and sciences. You can call it a "superstar" among all the real numbers. In fact, there is an entire book written on this one number, entitled *e: The Story of a Number*.

Let us go back to continuous compounding. Based on this definition, one can prove, using some tools of mathematics, that as $k \to \infty$,

$$\left(1+\frac{r}{k}\right)^{kt} \to e^{rt}.$$

So this was the answer to our indeterminant form. Equipped with this elegant truth, we can now see that our formula

$$A_t = P\left(1+\frac{r}{k}\right)^{kt} \text{ becomes } A_t = Pe^{rt}.$$

Thus, we have the following formula.

Continuous Compounding Formula

If the principal P is invested in an account earning interest at rate r compounding continuously, then the balance A_t after t years is

$$A_t = Pe^{rt}.$$

This formula is affectionately known as the "Pert" formula or the "Shampoo" formula. Roughly, continuous compounding describes interest being added every moment. When using this formula, be sure to calculate the part e^{rt} before multiplying it by P.

Example 1

Suppose that \$1000 is invested at 3% APR. Compare the amount after ten years for compounding monthly, daily, and continuously.

[Solution]

Compounded monthly: $A_{10} = \$1000\left(1 + \frac{0.03}{12}\right)^{12\times 10} = \$1{,}349.35.$

Compounded daily: $A_{10} = \$1000\left(1 + \frac{0.03}{365}\right)^{365\times 10} = \$1{,}349.84.$

Compounded continuously: $A_{10} = Pe^{rt} = \$1000\, e^{0.03\times 10} = \$1349.86.$

Not much difference! You will not get rich if your bank decides to compound continuously!

Example 2

If your initial deposit of \$100 earns an incredibly high APR of 15%, how much would you (or your descendent) end up with in 80 years if it is compounded *continuously*? Compare your result with *annual* compounding at the same APR.

[Solution]

Continuously: $A_{80} = Pe^{rt} = \$100\, e^{0.15\times 80} = \$16{,}275{,}479.14$ (over \$16 million).

Annually: $A_{80} = \$100\,(1 + 0.15)^{1\times 80} = \$7{,}175{,}087.94$ (about \$7 million).

You can see that when the APR is high over a very long period of time, it makes a big difference if the compounding is continuous or less frequent.

Compounding, as you will see in Chapter 6, is considered a type of **exponential growth** (growth by a fixed percentage), which can describe many phenomena seen in natural and social sciences. Of course, nature (such as growth in zoology, global temperature in oceanography, and population growth in sociology) do not decide to compound monthly or

quarterly; they almost always grow *continuously*. It is because of this fact, along with a powerful calculus-related property of the number *e*, that exponential growth based on *e* (sometimes referred to as **the natural exponential function**) plays a central role in many areas of science.

Exercises 3.2

1. Suppose you just discovered that your uncle, who died 100 years ago, had left you $2000 as a gift so you can attend college.
   (a) If your uncle invested that money at 4% annually for 100 years, how much would you have today?
   (b) Repeat Part (a), using 4% monthly instead.
   (c) Repeat Part (a), using 4% continuously instead. Compare your answers.

2. If $15,000 is invested in an account earning 3% interest continuously, how much is the investment worth after 20 years?

3. If $200 is invested in an account earning 7% interest continuously, how much is the investment worth after 40 years?

4. You want to invest some money so that in 36 years, you will have $100,000. Supposing that you have an account earning interest at 5.2% continuously, how much do you have to invest today?

5. You have just found a great investment opportunity, assuring you that the money will be compounded continuously at 9%. How much money do you have to set aside now so that in 10 years you will have $1,000,000?

6. Shockingly, you have just discovered a cryptocurrency whose value grows at an "annual percentage rate" of 69.315% compounded continuously. (This is not a random number. This rate was carefully crafted, specifically for this exercise.) In the text, you learned that an "annual percentage rate" does not actually mean the rate for annual yield. Pick your favorite amount of deposit, pretend that you invested that amount in this cryptocurrency, and see what happens to the value after just one year. How about in two years? How about in three years? Explain what is happening.

3.3 Savings and Annuities

So far in this chapter, we have looked at how a one-time deposit grows with compound interest. For most of us, however, we are not able to put a large sum of money in the bank today. Instead, we save for the future by setting aside a small amount of money from each paycheck, little by little, into a savings account or a retirement account. This idea is called a savings annuity. Most retirement plans like 401k plans or IRA plans are examples of savings annuities.

Annuities

An annuity can be described as a series of equal deposits made on a regular basis to save money for the future. Recall that in the two previous sections we looked at how the principal (deposited only one time) grows with respect to time. In an annuity, we make a series of regular, equal deposits. Let us use d to denote the amount of each deposit.

So how much will we have in a savings annuity after t years? It will be easiest to see what happens by working with an example first rather than working in general.

Suppose we will deposit $1000 at the end of each year into an account paying 6% annually. How much money will we have at the end of ten years?

Note that r=0.06, k=12 since the compounding is monthly, and we are given that d=$1000 each year. Assuming that we start with an empty account, we can begin with the first deposit, made at the end of Year 1. Then, another deposit of d=$1000 is made at the end of Year 2, another at the end of Year 3, and so on, until the last deposit is made at the end of Year 10. You have made 10 deposits of $1000 each, so the amount will be $10,000, right? No! Remember the "**time value of money**" mentioned at the beginning of this chapter? Money is **time-sensitive**. Compound interest is added to your account. In fact, because each deposit stays in this account for different durations, they will earn different amounts of interest. Let us look more precisely.

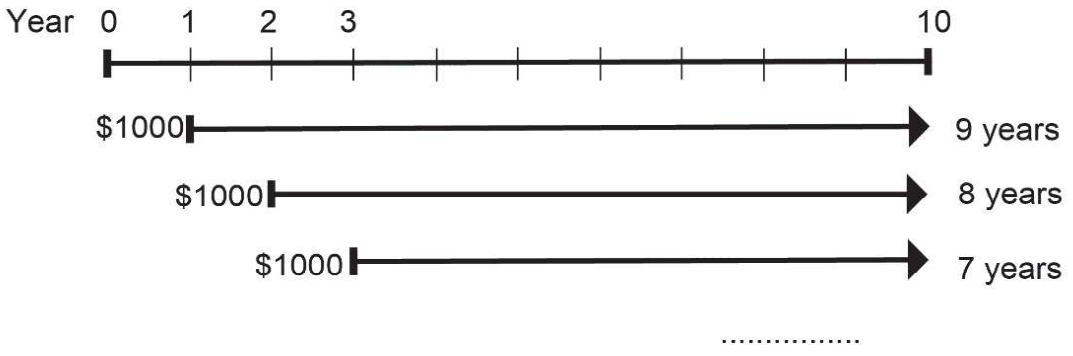


Figure 3.4

Here is how to calculate your balance at the end of the 10-year period, having just made your last deposit of $1000 for that year.

The first deposit of $1000, made at the end of Year 1, will have sat in the account for 9 years at APR 6% compounded annually. Thus, this first deposit is now worth $\$1000 \times (1 + 0.06)^9 = \$1689.48$.

The second deposit of $1000, made at the end of Year 2, will have sat in the account for 8 years, so the second deposit is now worth $\$1000 \times (1 + 0.06)^8 = \$1593.85$.

The third deposit, similarly, will earn interest for 7 years, so it is now $\$1000 \times (1 + 0.06)^7 = \$1503.63$.

The same applies to all the deposits, with the last deposit of $1000 made at the end of Year 10, which does not earn any interest since it was just deposited.

During this 10-year period, 10 deposits (each in the amount of $1000) were made, and the balance at the end of the 10 years will be the sum of all of these amounts:

$$A = 1000 \times (1.06)^9 + 1000 \times (1.06)^8 + 1000 \times (1.06)^7 + \cdots + 1000 \times (1.06)^1 + 1000.$$

Now, if you simply put all this money under the mattress (an old way of hiding money at home), you will have $10,000, obviously ($\$1000 \times 10$). But with a high interest rate of 6% annually, this amount turns out to be $13,180.79. The difference

$$\$13{,}180.79 - \$10{,}000 = \$3{,}180.79$$

is the total interest earned during the 10-year period.

You may be asking, “Now, how did you calculate this amount, the sum of ten numbers?” One way is, of course, calculate each term as above and add them: 1689.48 + 1593.85 + 1503.63 + … But this seems like a lot of work and is not practical when you have, like in a retirement account, hundreds of monthly deposits.

Fortunately, there is a mathematical tool available for us for calculating this type of sum; it is called a **geometric series**. This series can be used when the amount of each deposit is the same, the base (the number in the parentheses, i.e., $1 + \frac{r}{k}$) is the same, and the exponents are sequential. Here is the formula:

Annuity Formula

When deposits of \$ d are made on a regular basis into an account with an annual percentage rate r (written as a decimal) compounded k times a year, the sum S_t saved after t years is

$$S_t = d \times \frac{\left(1+\frac{r}{k}\right)^{kt} - 1}{\left(\frac{r}{k}\right)}$$

Here, we assume that k is consistent with the frequency of deposits (e.g., if a deposit is made every quarter, then $k = 4$, and the compounding is quarterly). You may notice that $\frac{r}{k}$ appears twice in this formula; it is the rate per period.

It is rather complicated, isn't it? It is not a main goal of this course that you learn this type of formula. Memorization is not central to mathematical thinking as many people are led to believe. In fact, it is unlikely that many people, even in the world of finance, actually use this formula on their calculators. Most of these calculations are performed by computer software. Your instructor may or may not make you learn this. However, some exercises will require this (and other) formulas to be used.

So when do you use this? Remember, annuities assume that you put money in the account on a regular schedule (every month, year, quarter, etc.) and let it sit there earning interest. Therefore, use this formula when **equal and regular deposits are made to save money for the future**. Each deposit (payment) is "**moved forward**" on the timeline to be added at a point in the future (along with the interest earned on each deposit).

Example 1

A traditional individual retirement account (IRA) is a special type of retirement account in which the money you invest is exempt from income taxes until you withdraw it. If you deposit \$100 each month into an IRA earning 6% interest, how much will you have in the account after 20 years?

[Solution]

In this example, $d = \$100$, $r = 0.06$, $k = 12$ (since deposits are made monthly), and $t = 20$.

Putting all these numbers into the equation, we get

$$S_{20} = 100 \times \frac{\left(1+\frac{0.06}{12}\right)^{12\times 20} - 1}{\left(\frac{0.06}{12}\right)} = 100 \times \frac{(1.005)^{240} - 1}{0.005} = 100 \times 462.040895\ldots$$

This gives us the answer; the balance after 20 years is $S_{20} = \$46{,}204.09$.
Notice that you deposited into the account a total of $24,000 ($100 a month for 240 months); this is referred to as your **total payments**. The difference between what you end up with and how much you put in is the **interest earned**. In this case the total interest earned is $\$46{,}204.09 - \$24{,}000 = \$22{,}204.09$.

This is why you would want to set up a retirement account or some savings account long before you plan to retire. Let the power of compound interest work for you.

Total Payments and Interest Earned

When deposits of \$ d are made on a regular basis into an account with an annual percentage rate r (written as a decimal) compounded k times a year, and S_t is the balance after t years,

- $d \times kt$ (each payment multiplied by the number of payments) is your total payments.
- $S_t - (d \times kt)$ is the interest earned (the difference between your balance and how much you put into the account).

Now, sometimes the question is not how much you will end up with, but how much you would have to set aside each month for a certain target amount in the future.

Example 2

You want to have $200,000 in your account when you retire in 30 years. Your retirement account earns 8% interest. How much do you need to deposit each month to meet your retirement goal?

[Solution]

In this example, we are looking for d, given that $r = 0.08$, $k = 12$ (assumed monthly since you are making monthly deposits), $t = 30$, and S_{30}= $200,000.

In this case, we will set up the equation and solve for d.

$$200{,}000 = d \times \frac{\left(1 + \frac{0.08}{12}\right)^{12 \times 30} - 1}{\frac{0.08}{12}}.$$

The complex fraction on the right-hand side may be hard to enter into a calculator. If your calculator can store a number into its memory, you may want to do so. If your calculator comes with parentheses, then you can enter something like this:

$$((1 + 0.08 / 12) \boxed{x^y} (12*30) - 1) / (0.08 / 12) =$$

to get approximately 1490.36. (The x^y key may be substituted by the caret ^ key.) So we have

$$200{,}000 = 1490.36\, d.$$

Dividing both sides of the equation by this number, we get

$$d = \frac{200{,}000}{1490.36} = 134.20.$$

So you would need to deposit $134.20 each month to have $200,000 in 30 years if your account earns 8% interest.

If you are following along with your calculator, you may get a slightly different answer, but your amount should be a little more than $134.00. The difference is due to rounding errors, and you should not be concerned about it. The answer shown here, $134.20, is obtained with a scientific calculator using one long sequence of operations without rounding at any stage. If you rounded 0.08/12 to 1.00667, for instance, you will get an answer more like $134.09. In the real world, more precise computing is used to minimize rounding errors. For the purposes of this course, however, these errors are negligible. Just make sure you do not round in early stages of your calculations, and try to use a reasonable number of digits. Typically, the more digits you use, the more accurate the final answer will be.

Try It Now 1

A more conservative investment account pays 3% interest. If you deposit $150 a month into this account, how much will you have after 10 years? How much is from interest?

Exercises 3.3

1. You deposit $200 each month into an account earning 3% interest compounded monthly.
   (a) How much will you have in the account in 30 years?
   (b) How much total money will you put into the account?
   (c) How much total interest will you earn?

2. You deposit $1000 each year into an account earning 8% compounded annually.
   (a) How much will you have in the account in 10 years?
   (b) How much total money will you put into the account?
   (c) How much total interest will you earn?

3. You deposit $300 every quarter (every 3 months) into an account earning 8% compounded quarterly.
   (a) How much will you have in the account in 20 years?
   (b) How much total money will you put into the account?
   (c) How much total interest will you earn?

4. As a child, you began setting aside $10 each month into an account earning 6% compounded monthly. Your brother did the same, except he kept all his money under his bed.
   (a) How much will you have in the account in 50 years?
   (b) How much total money will you put into the account? (This is also the amount your brother will have after 50 years.)
   (c) How much total interest will you earn?

5. Jose has determined he needs to have $800,000 for retirement in 30 years. His stock-based investment account earns 7.2% interest.
   (a) How much would he need to deposit in the account each month?
   (b) How much total money will he put into the account?
   (c) How much total interest will he earn?

6. You have determined you will need to have $32,000 saved up for your dream car in 6 years. Your savings account earns 4% interest compounded quarterly.
   (a) If you can set aside money every quarter, how much would you need to deposit into your savings account every quarter?
   (b) How much total money will you put into the account?
   (c) How much total interest will you earn?

7. You wish to have $3000 in 2 years to buy a fancy new stereo system. How much should you deposit each month into an account paying 8% compounded monthly?

Figure 3.5[46]

[46] Image on PxHere

3.4 Loans and Mortgages

In the first three sections of this chapter, we learned how to save. Money was moved "*forward*" (to the right on the time line) to a point in the future. For this reason, the amount you will end up with, P_t, is referred to as a "**future value**." We learned an important (and positive) side of finance: **compound interest works FOR you if you are saving money**. Unfortunately, there is another side of finance, and that is about borrowing, the type of loans in which you get the money now and make a series of regular payments in the future. Because all the payments are in the future, in order to find the sum of their present values, we need to move money "*backwards*" (to the left on the time line) to the present. We will then see the other (ugly) side of finance: **compound interest works AGAINST you if you are borrowing money**. Study this section carefully. It may just save you thousands of dollars in your life.

Loans and Mortgages

In this section, you will learn about conventional loans (also called amortized loans or installment loans). Examples include auto loans and home mortgages. The key for understanding how loans work is, once again, to look at the time-dependent value of money. In the previous section, on annuities, we learned that $1000 invested today is worth more in the future because of compound interest. In fact, because the interest "compounds," it grows exponentially (something we will study in Chapter 6). Yes, money "grows."

In contract, when you borrow money, because your payments are all in the future, the **present value** of those future payments must add up to the amount being borrowed today. Money "shrinks" when you move it backwards, to the left on the time line. In other words, $1000 that you will pay back in a year is not worth $1000 today; it is less. This is easy to see—if you want to borrow $1000 *today*, then clearly you will have to pay more than $1000 a year from now to settle the debt, right?

A song, popular in the 1940s, called "A Dollar Ain't a Dollar Any More" illustrates this time-sensitive nature of money vividly.

Let us see how this works by looking at an example. This illustration is (intentionally) very similar to the one presented in the previous section. This helps you see and contrast borrowing with saving.

Suppose you need to borrow some money today. How much can you borrow if you promise to pay back $1000 at the end of every year for ten years, with an APR of 6% compounded annually?

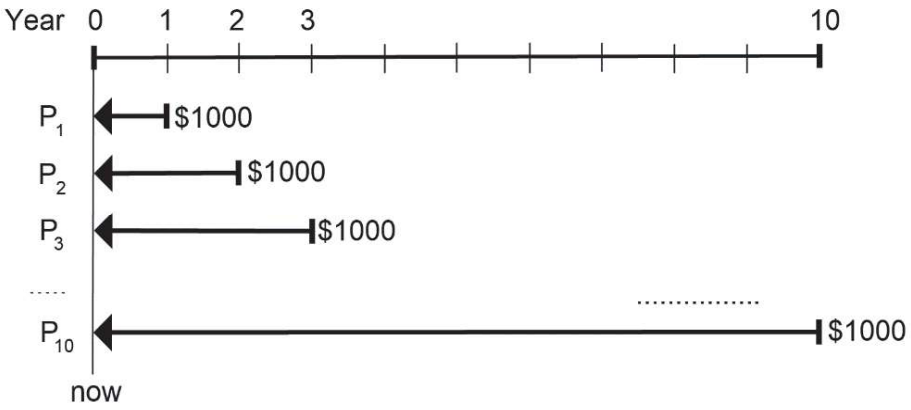


Figure 3.6

The figure above illustrates this loan with ten future payments. Let us denote by P_1 the present value of the first \$1000 payment, to be made one year from today. In other words, this is how much that first payment is worth today. Let P_2 be the present value of the second \$1000 payment, and so on. P_{10} would then represent how much the last payment, to be made 10 years later, is worth today. The amount you can borrow today is definitely not \$10,000. The amount you can borrow today, denoted P (for "**principal**" or "**present value**"), which is the "loan amount," is the sum of all these present values, i.e.,

$$P = P_1 + P_2 + \cdots + P_{10}$$

So what is P_1? Well, we have a way to figure this out—using the first formula we learned in this chapter. P_1, after the 6% annual interest is added to it for one year, becomes \$1000. So we have

$$\$1000 = P_1 \times (1 + 0.06)^1$$

So $P_1 = \frac{1000}{(1.06)^1}$ but we can re-write this using a negative exponent as $P_1 = 1000 \times (1.06)^{-1}$. This turns out to be \$943.40. In other words, if we deposit \$943.40 today, with 6% annual interest, it will be exactly \$1000 in one year. This means that the present value of my first \$1000 payment is \$943.40. Similarly, since

$$\$1000 = P_2 \times (1 + 0.06)^2,$$

we can write $P_2 = 1000 \times (1.06)^{-2}$. This turns out to be \$890.00.

We can do this for all ten payments, so the loan amount, the sum of the present values of all these future payments, is

$$P = 1000 \times (1.06)^{-1} + 1000 \times (1.06)^{-2} + \cdots + 1000 \times (1.06)^{-10}.$$

You can carry out each calculation, which shows that

$$P = 943.40 + 890.00 + \cdots + 558.39 = \$7360.09.$$

In other words, while you will be paying back a total of $10,000 over the next ten years, today you can only borrow $7,360.09. The difference, $2,639.91, is the interest you pay, which is a major part of what is known as "financial charge."

Negative Exponents

Negative exponents are not something we will spend a lot of time on, but you will see them in some formulas. In general, x^{-N} simply means $\frac{1}{x^N}$ in mathematics (assuming that $x \neq 0$). We will not study these algebraically, though, as they are not within the scope of this textbook.

Of course, it is annoying to have to calculate each term and add them together, especially if you have hundreds of payments, as in a home mortgage (a typical mortgage term is for 30 years, i.e., 360 payments). So is there not a way to calculate this? The answer is "Yes!" The same mathematical idea that gave us the annuity formula, i.e., **geometric series**, gives us a long (but convenient) formula to calculate the loan amount, whether it is a car loan or a home mortgage. Essentially all your debts, including student loans, credit card debts, and loans at retail stores, are based on this formula.

Loan Formula

When payments of $\$ d$ are made on a regular basis over t years for a loan with an annual percentage rate r (written as a decimal) compounded k times a year, the amount of the loan P (for "**present value**") is

$$P = d \times \frac{1 - \left(1 + \frac{r}{k}\right)^{-kt}}{\left(\frac{r}{k}\right)}$$

The duration of the loan, t years, is called the **term** of the loan.

Here, as before, we assume that k is consistent with the frequency of payments (e.g., if the payments are made monthly, then $k = 12$). Also, notice the similarity between the Annuity Formula (in the previous section) and the Loan Formula here. The subtraction is reversed in the numerator, and the exponent is negative, but everything else looks pretty much the same. Like the Annuity Formula, your instructor may or may not make you learn this. However, some exercises will require this (and other) formulas to be used.

By the way, this loan amount P will be significantly less than the total sum of all your payments. Remember, it is not the sum of all future payments; rather, it is the **sum of the present values of all future** payments. The difference between these values is the amount of **total interest paid** over the term of the loan. See the example below.

Example 1

You can afford $200 per month as a car payment. If you can get an auto loan at 3% interest for 60 months (5 years), how expensive of a car can you afford? In other words, what loan amount can you pay off with $200 per month?

[Solution]

In this example, you see that $d = \$200$, $r = 0.03$, $k = 12$ (since you are making monthly payments), and $t = 5$.

We are looking for P, the amount of the loan.

$$P = 200 \times \frac{1 - \left(1 + \frac{0.03}{12}\right)^{-12\times 5}}{\left(\frac{0.03}{12}\right)} = 200 \times \frac{1 - (1.0025)^{-60}}{0.0025} = 200 \times \frac{1 - 0.8608691}{0.0025}.$$

(Learn how to enter negative numbers on your calculator. Some calculators require you to enter the negative sign or a specific key after the number while some do not.)
This comes out to be $200 \times 55.65235 \ldots = 11{,}130.47$, so you can afford to borrow $11,130.47 today. You may have to give up buying your dream car for now or come up with a sizable down payment.

You will pay a total of $12,000 (because you pay $200 per month for 60 months) to the loan company; this amount is known as your **total payments**. The difference between the amount you pay and the amount of the loan is $\$12{,}000 - \$11{,}130.47 = \$869.53$. This is **the total interest paid**. This is the cost of borrowing money. ■

Total Payments and Total Interest Paid

When payments of $\$ d$ are made on a regular basis to pay for a loan with an annual percentage rate r (written as a decimal) compounded k times a year for t years, and P is the amount of the loan,

- $d \times kt$ (each payment multiplied by the number of payments) is your **total payments**.
- $(d \times kt) - P$ is the **total interest paid** (the difference between how much you pay back and the amount of the loan) over the term of the loan.

So far we have learned how to compute the amount of loan based on payments. However, a more realistic question is this: how much do I have to pay back every month if I buy this car (or

house or whatever)? So it is common to be given the amount of loan P, along with the APR and the terms, and then to have to compute the amount of each payment. Here is an example.

Example 2

You want to take out a $140,000 mortgage (home loan). The interest rate on the loan is 6%, and the loan is for 30 years. How much will your monthly payments be? What are the total payments over the 30-year period? How much interest is paid over the life of this loan?

[Solution]

In this example, we are looking for d. We are given that $r = 0.06$, $k = 12$, $t = 30$, and the loan amount, $P = 140{,}000$.

In this case, we will have to set up the equation and solve for d.

$$140{,}000 = d \times \frac{1 - \left(1 + \frac{0.06}{12}\right)^{-12 \times 30}}{\left(\frac{0.06}{12}\right)} = d \times \frac{0.833958}{0.005}.$$

This gives us

$$140{,}000 = 166.7916 \times d.$$

Dividing both sides of the equation by 166.7916, we get $d = 839.37$. So you will be paying $839.37 every month for the next 30 years (not to mention your property tax, maintenance costs, and other related expenses).

Note that you will be making 360 payments, so your total payments will be

$$\$839.37 \times 360 = \$302{,}173.20.$$

Wait. What?

You are only borrowing $140,000. In other words, you are paying a total of

$$\$302{,}173.20 - \$140{,}000 = \$162{,}173.20$$

in total interest over the life of the loan. Yes, over the next 30 years, you will be spending more money on the interest than on the house you are about to buy. This is the reality!

Remember, again, the power of compound interest works FOR you when saving but AGAINST you when borrowing.

Of course, the amount of interest you pay depends very heavily on the interest rate and the terms, so be wide. Even a fraction of a percentage point in APR may make a huge difference in total interest paid.

Try It Now 2

Ciara bought $3,000 of new furniture on credit. Because her credit score is not very good, the store is charging her a fairly high interest rate on the loan of 16%. If she agreed to pay off the furniture over 2 years, how much will she have to pay each month? How much total interest will she be paying during that time?

Remaining Loan Balance

With loans, it is often desirable to determine what the remaining loan balance will be after some number of years. For example, if you purchase a home and plan to sell it in five years, you might want to know how much of the loan balance you will have paid off and how much you have to pay from the sale.

To determine the remaining loan balance after some number of years, we first need to know the amount of each payment (if we do not already know it). Remember that only a portion of your loan payments go towards the loan balance; a big portion is going to go towards interest. For example, if your payments were $1,000 a month, after a year you will *not* have paid off $12,000 of the loan balance.

To determine the remaining loan balance, we can think "how much loan will these loan payments be able to pay off in the remaining time on the loan?"

Example 3

Suppose you bought a house 20 years ago. After making a big down payment and paying some painful fees, the original amount of the 30-year mortgage was $166,791.61. Your APR was 6%, so you have been making a payment of $1000 every month (this amount was calculated using our formula above—you can verify that). Now that 20 years have passed (only ten more years to go in your mortgage), you want to know how much you still owe (the remaining balance). What is the current balance of your mortgage, given that you have only $10 \times 12 = 120$ more payments to make?

[Solution]

This may sound complicated, but (as mentioned above) the current balance is simply the sum of the present values of all future payments. This is exactly how we derived our formula earlier. So the question is equivalent to this: what is the present value of the next 120 payments of $1000 each, to be paid over the next 10 years at an APR of

6% monthly? In other words, we are looking for P when $d = \$1{,}000$, $r = 0.06$, $k = 12$, and $t = 10$.

$$P = 1000 \times \frac{1 - \left(1 + \frac{0.06}{12}\right)^{-12\times 10}}{\frac{0.06}{12}} = 1000 \times \frac{1 - (1.005)^{-120}}{0.005} = 1000 \times \frac{0.450367}{0.005}.$$

This gives us $P = 90{,}073.45$. The loan balance with 10 years to go is $90,073.45.

Now, reflect on this for a moment. How much was the original loan? Just over $166,000. You have already made 240 out of the 360 monthly payments. You have already made 2/3 of all the payments. How is it possible that you still owe more than $90,000? Well, this is the reality (again). The power of compounding interest works AGAINST you when you borrow.

Example 4

A couple purchases a home with a $180,000 mortgage at 4% for 30 years with monthly payments. What will the remaining balance on their mortgage be after 5 years? (The answer may be depressing.)

[Solution]

First, we will calculate their monthly payments.

We are looking for d with $r = 0.04$, $k = 12$, $t = 30$, and $P = 180{,}000$.

$$180{,}000 = d \times \frac{1 - \left(1 + \frac{0.04}{12}\right)^{-12\times 30}}{\left(\frac{0.04}{12}\right)} = 209.46124 \times d.$$

Dividing both sides by 209.46124, we get $d = 859.35$.

Now that we know the monthly payments, we can determine the remaining balance. We want the remaining balance after 5 years, with 25 years remaining on the loan, so we calculate the loan balance that will be paid off with the monthly payments over those 25 years.

Note that $d = \$859.35$, $t = 25$, and everything else remains the same.

$$P = 859.35 \times \frac{1 - \left(1 + \frac{0.04}{12}\right)^{-12\times 25}}{\left(\frac{0.04}{12}\right)} = 859.35 \times 189.45248 = 162{,}805.99.$$

The loan balance after 5 years, with 25 years remaining on the loan, will be $162,805.99.

Over that 5 years, the couple has paid off $\$180{,}000 - \$162{,}805.99 = \$17{,}194.01$. Remember, they have paid $859.35 every month for 5 years (60 payments), for a total of $\$859.35 \times 60 = \$51{,}561$. The difference, $\$51{,}561 - \$17{,}194.01 = \$34{,}366.99$, is the total interest paid thus far.

The amount of the initial loan they have actually paid off (how much of their house they own) is called the **equity**. In their case, they own just about $17K of their $180K home.

Which Equation to Use?

When presented with a finance problem (on an exam or in real life), you are usually not told what type of problem it is or which equation to use. Here are some hints on deciding which equation to use based on the wording of the problem.

If you are trying to find an amount or a balance in the future, you are moving the money "forward," as in saving money. Find out if you are considering a one-time deposit or a series of equal, regular deposits. Either case, you need to find the rate r (always use a decimal form), the time t (in years), and the number of compounding periods per year (k). For a one-time deposit, use the formula in Section 3.1. If the interest is compounding continuously, you do not have k, and you use the formula in Section 3.2. For a series of equal, regular deposits, use the Annuity Formula in Section 3.3. If you are dealing with a loan with a series of equal, regular payments, use the Loan Formula in this section.

Remember, an important part of answering any kind of question, money or otherwise, is first to correctly identify what the question is really asking, and then to determine what approach will best allow you to solve the problem.

Try It Now 3

For each of the following scenarios, determine if it is a compound interest problem with one deposit, a savings annuity problem, or a loan problem. Then solve each problem.

a. **Dylan wants to buy a new car. Rather than take out a loan, he decides to save $200 a month in an account earning 3% interest compounded monthly. How much will he have saved up after 3 years?**
b. **Jeanette is managing investments for a non-profit company. They want to invest some money now in an account earning 5% interest compounded annually with the goal to have $30,000 in the account in 6 years. How much should Jeanette deposit into the account now?**
c. **Brittany is going to finance new office equipment at a 2% rate over a 4-year term. If she can afford monthly payments of $100, how much new equipment can she buy?**
d. **How much would you need to save every month in an account earning 4% interest to have $5,000 saved up in two years?**

Exercises 3.4

1. You can afford to make payments of $1700 per month. You have found a 30-year loan at 3% interest.
   (a) How big of a loan can you afford?
   (b) How much total money will you pay the loan company?
   (c) How much of that money is interest?

2. Marie can afford a $250-per-month car payment. She has found a 5-year loan at 7% interest.
   (a) How expensive of a car can she afford?
   (b) How much total money will she pay?
   (c) How much of that money is interest?

3. You want to buy a $25,000 car. The car dealership is offering a 2% interest rate for 48 months (4 years). What will your monthly payments be?

4. You decide to finance a $12,000 car at 3% compounded monthly for 4 years. What will your monthly payments be? How much interest will you pay over the life of the loan?

5. You want to buy a $200,000 home. You plan to pay 10% as a down payment and take out a 30-year loan for the rest.
   (a) How much is the loan amount going to be?
   (b) What will your monthly payments be if the interest rate is 5%?
   (c) What will your monthly payments be if the interest rate is 6%?

6. Lynn bought a $300,000 house, paying 10% down, and financing the rest at 6% interest for 30 years.
   (a) Find her monthly payments.
   (b) How much interest will she pay over the life of the loan?

7. Vahid bought a car for $24,000 three years ago. The loan had a 5-year term at 3% interest rate, making monthly payments. How much does he still owe on the car? (You must first find her monthly payments.)

8. Mina bought a house 15 years ago, taking out a $420,000 mortgage at 6% for 30 years, making monthly payments. How much does she still owe on the mortgage? (You must first find her monthly payments.) Note that half of the 30-year term is over. Does it mean she has paid off half of the mortgage?

3.5 National Deficit and Debt[vii]

Throughout this chapter, we have talked about mathematics of finance, mostly involving compound interest. It is at the heart of **personal finance**, and knowing these fundamental concepts can make a big difference, possibly saving you a significant amount of money when purchasing a home.

However, this mathematics of finance is not just for individuals. After all, groups, organizations, companies, municipalities, and governments also spend money, invest money, and go into debts. In fact, federal budgets and sovereign debts are big topics in business and economics. Before we conclude this chapter, we briefly discuss these big financial/economic issues, including debts and deficits.

Deficit vs. Debt

If you earned $3000 last month and spent $3500, you have *overspent*, and the difference is called the **deficit,** which may be the amount you owe on your credit card for the month. You may have already had a $400 balance on your credit card bill during the previous month. Now, you need to pay $900 and possibly additional fees and charges. This amount is called a **debt**. Simply put, a deficit has to do with the negative amount over a certain period of time (like a month or a year) while a debt has to do with the accumulation of money owed.

Similarly, a government deficit is the difference between the amount the government spends and the amount the government collects (mostly through taxes) over a certain period of time, typically a year (called a **fiscal year**). If the revenues exceed the expenditures, the amount of difference is called a **surplus**. Not surprisingly, the government often spends much more than it collects, resulting in large deficits year after year. The total money owed by the government, i.e., the accumulation of these deficits, is referred to as the **national debt**. (Do not think this is someone else's problem. You and I will be paying this off.)

For instance, year 2020 was a terrible year with the new coronavirus as well as many issues related to social injustice and the controversial presidential election. The Congressional Budget Office projected, in April 2020, the deficit for Fiscal year 2020 will be at least $3.7 trillion (that's $3,700,000,000,000). Because of deficits from previous years, the United States federal debt, as of the end of 2020, was approximately $27 trillion.

How big is $27 trillion, exactly? If you could pay this amount by putting down a $100 bill every second, 24 hours a day, 7 days a week, 365 days a year, it would take you over 8,560 years to pay it back. (Check this! It sounds like a question from Chapter 2.)

Let us take a look at federal deficits for the last 90 or so years.

Figure 3.1 shows the pattern of annual federal budget deficits and surpluses, back to 1930, as a share of **GDP** ("**Gross Domestic Product**," value of all goods and services made in a year). When the line is above the horizontal axis, the budget is in surplus. When the line is below the

horizontal axis, a budget deficit occurred. Clearly, the biggest deficits as a share of GDP during this time were incurred to finance World War II. Deficits were also large during the 1930s, the 1980s, the early 1990s, and during the 2008-2009 recession.

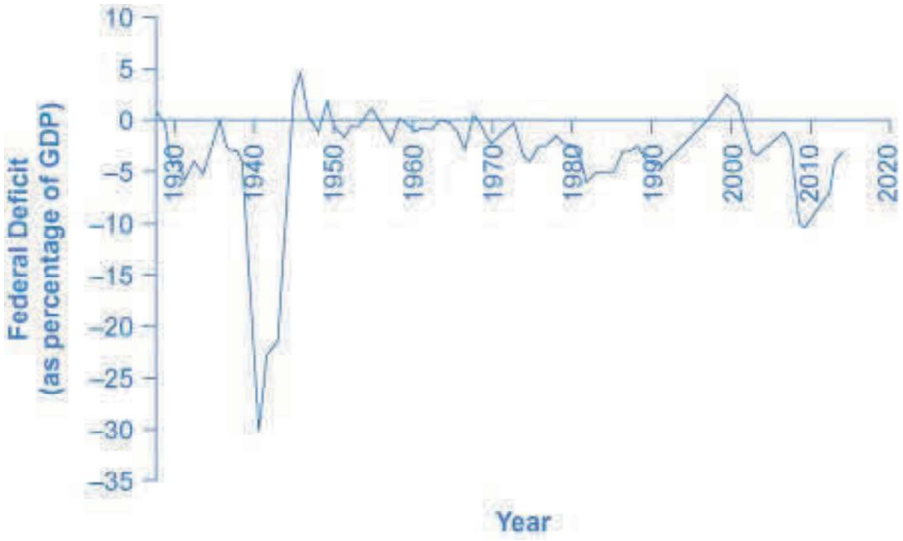


Figure 3.7: Pattern of Federal Budget Deficits and Surpluses, 1929–2014. The federal government has run budget deficits for decades. The budget was briefly in surplus in the late 1990s.[47]

Debt/GDP Ratio

Another useful way to view the budget deficit is through the prism of accumulated debt rather than annual deficits. As introduced above, the **national debt** refers to the total amount that the government has borrowed over time. Figure 2 shows the ratio **debt/GDP** since 1940. Until the 1970s, the debt/GDP ratio revealed a fairly clear pattern of federal borrowing. The government ran up large deficits and raised the debt/GDP ratio in World War II, but from the 1950s to the 1970s the government ran either surpluses or relatively small deficits, and so the debt/GDP ratio drifted down. Large deficits in the 1980s and early 1990s caused the ratio to rise sharply. When budget surpluses arrived from 1998 to 2001, the debt/GDP ratio declined substantially. The budget deficits starting in 2002 then tugged the debt/GDP ratio higher—with a big jump when the recession took hold in 2008–2009, when the subprime mortgage crisis caused the market to crash.

[47] U.S. Office of Management and Budget and Federal Reserve Bank of St. Louis, Federal Surplus or Deficit [-] as Percent of Gross Domestic Product [FYFSGDA188S], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/FYFSGDA188S

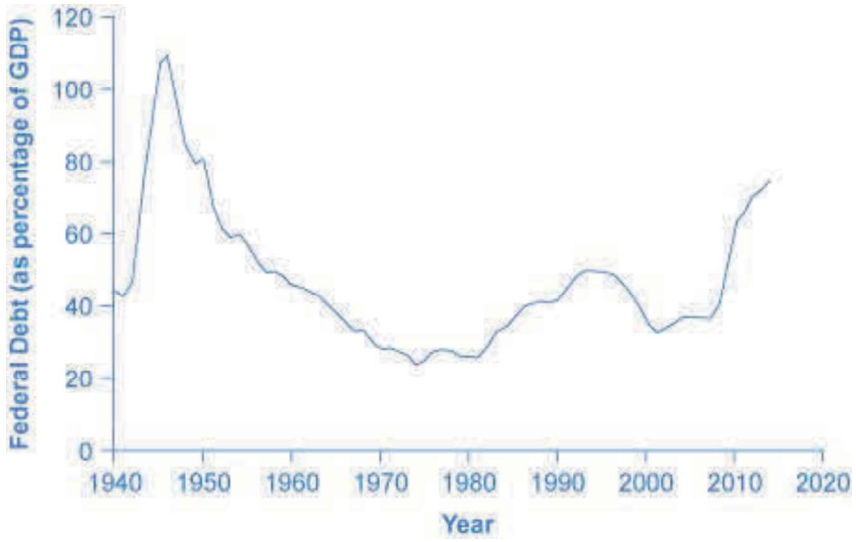


Figure 3.8: Federal Debt as a Percentage of GDP, 1942–2014. Federal debt is the sum of annual budget deficits and surpluses. Annual deficits do not always mean that the debt/GDP ratio is rising.[48]

The Path from Deficits to Surpluses to Deficits

Why did the budget deficits suddenly turn to surpluses from 1998 to 2001, and why did the surpluses return to deficits in 2002? Why did the deficit become so large after 2007? Figure 3 suggests some answers. The graph combines the earlier information on total federal spending and taxes in a single graph but focuses on the federal budget since 1990.

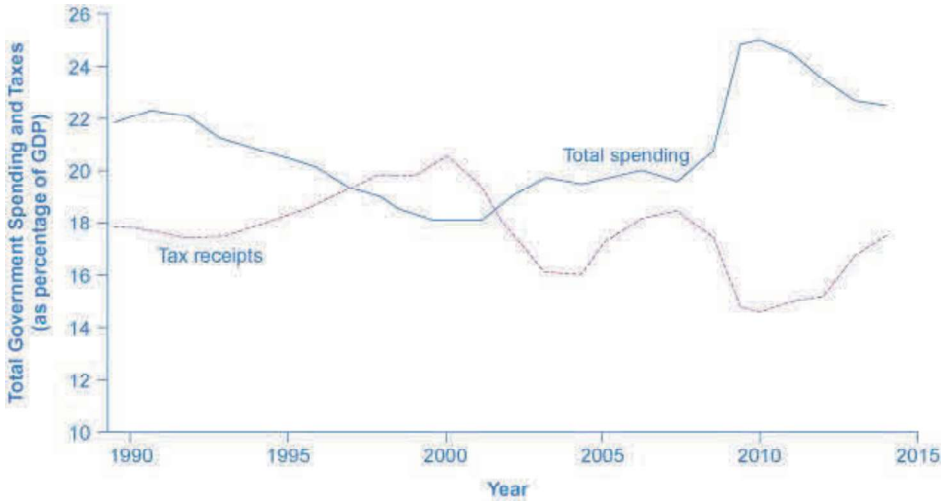


Figure 3.9: Total Government Spending and Taxes as a Share of GDP, 1990–2014. When government spending exceeds taxes, the gap is the budget deficit. When taxes exceed spending, the gap is a budget surplus.[49]

[48] Image from Economic Report of the President, Table B-20, (http://www.gpo.gov/fdsys/pkg/ERP-2015/content-detail.html) is in the public domain.

[49] Image from Economic Report of the President, Tables B-21 and B-1, (http://www.gpo.gov/fdsys/pkg/ERP-2015/content-detail.html) is in the public domain.

Government spending as a share of GDP declined steadily through the 1990s. The biggest single reason was the end of the Cold War, when defense spending declined from 5.2% of GDP in 1990 to 3.0% in 2000, but interest payments by the federal government also fell by about 1.0% of GDP. However, federal tax collections increased substantially in the later 1990s, jumping from 18.1% of GDP in 1994 to 20.8% in 2000. Powerful economic growth in the late 1990s fueled the boom in taxes. Personal income taxes rise as income goes up; payroll taxes rise as jobs and payrolls go up; corporate income taxes rise as profits go up. At the same time, government spending on transfer payments such as unemployment benefits, foods stamps, and welfare declined with more people working.

This sharp increase in tax revenues and decrease in expenditures on transfer payments were largely unexpected even by experienced budget analysts, so budget surpluses came as a surprise. However, in the early 2000s, many of these factors started running in reverse. Tax revenues sagged, due largely to the recession that started in March 2001 (triggered by the collapse of the "dotcom" bubble), which reduced revenues. Congress enacted a series of tax cuts starting in 2001. In addition, government spending swelled due to increases in defense, healthcare, education, Social Security, and support programs for those who were hurt by the recession and the slow growth that followed. Deficits returned. When the severe recession hit in late 2007, spending climbed and tax collections fell to historically unusual levels, resulting in enormous deficits.

Longer-term U.S. budget forecasts, a decade or more into the future, predict enormous deficits. The higher deficits during the 2008-2009 recession have repercussions, and the demographics will be challenging. The primary reason is the "baby boom"—the exceptionally high birthrates that began in 1946, right after World War II, and lasted for about two decades. Starting in 2010, the front edge of the baby boom generation began to reach age 65, and in the next two decades, the proportion of Americans over the age of 65 will increase substantially. The current level of the payroll taxes that support Social Security and Medicare will fall well short of the projected expenses of these programs.

Then, 2020 happened. The results of the COVID-19 pandemic include government spending on programs to rescue businesses, stimulus checks, vaccine development and distribution, health and safety protocols, and many other areas. Yes, the forecast is for large budget deficits. A decision to collect more revenue to support these programs or to decrease benefit levels may alter this long-term forecast, but there is no doubt that the country is facing major challenges.

Exercises 3.5

1. Briefly describe the main difference between a “deficit” and a “debt.”

2. Give a simple definition of “federal deficit.”

3. In FY 2017 (the fiscal year 2017 is from October 1, 2017 to September 30, 2018), the U.S. Federal Government had a total revenue of about $3.33 trillion and total expenditures of $4.11 trillion. The GDP for that year was $20.23 trillion.
   (a) How much was the federal deficit for FY 2017?
   (b) The national debt as of the end of FY 2016 was $20.245 trillion. Using (a), calculate the national debt as of the end of FY 2017.
   (c) Find the debt/GDP ratio at the end of FY 2017.
   (d) Find the deficit/GDP ratio for FY 2017.

4. In FY 1987, the national deficit was $155 billion. By what percentage did the deficit change between 1987 and 2017 (30 years)? Use your answers from the previous exercise.

Figure 3.10[50]

[50] Image by Chris Yunker is licensed under CC BY 2.0

For Further Explorations…

Comparing Mortgage Payments, Total Interests, and Savings

Many people do not realize the true power of compound interest, particularly when the term is long, like with 30-year mortgages. When a long period of time is involved, small differences may result in a lot of saving (or a lot of interest). Consider the following situations involving a $500,000 mortgage. Most of the calculations below are covered in Section 3.4, and you should be able to verify the numbers.

1) You called two mortgage houses to get a 30-year home loan for $500,000. The first lender says your APR is 3.0% while the second says it will be 4.0%. How much will your monthly payments be at each rate? How much would you be saving at 3.0% compared to 4.0%?
2)

   With APR = 3.0%, you need to solve the following:

$$500{,}000 = d \times \frac{1 - \left(1 + \frac{0.03}{12}\right)^{-12\times 30}}{\left(\frac{0.03}{12}\right)} = 237.18938 \times d.$$

   This gives the amount of each monthly payment: $d = \$2108.02$.

   With APR = 4.0%, the equation is similar:

$$500{,}000 = d \times \frac{1 - \left(1 + \frac{0.04}{12}\right)^{-12\times 30}}{\left(\frac{0.04}{12}\right)} = 209.46124 \times d.$$

   Here, $d = \$2387.08$. The difference is $279.06 per month. Multiplying this with 360 payments, you see that the total saving (for the next 30 years) will be $\$279.06 \times 360 = \$100{,}461.60$. Did you see that? Yes, a difference of 1% could save you more than $100,000.

3) You waited for a few days to call the first mortgage house. You were then told that you are no longer eligible for 3.0% but you can qualify for 3.25%. You think, "Well, it's only a quarter of one percent. It cannot be that bad." What would your monthly payments be, and how much would you have saved if you could get the 3.0% rate promised first?

   With APR = 3.25%, the equation is this:

$$500{,}000 = d \times \frac{1 - \left(1 + \frac{0.0325}{12}\right)^{-12\times 30}}{\left(\frac{0.0325}{12}\right)} = 229.77608 \times d.$$

Here, $d = \$2176.03$. The difference between this and the amount with an APR of 3.0% is

$\$2176.03 - \$2108.02 = \$68.01$ (per month) only. However, multiplying this with 360 payments, you see that the total saving (for the next 30 years) will be $\$68.01 \times 360 = \$24{,}483.60$. Did you see this? Yes, a quarter-point difference could save you enough money to buy a car! This is why you would want to pay very close attention to the interest rate.

4) With any 30-year monthly mortgage, the rate 5.30404% signifies something special. Find your total interest payment at this rate and see why this is significant.
This time, the equation is as follows:

$$500{,}000 = d \times \frac{1 - \left(1 + \frac{0.0530404}{12}\right)^{-12\times 30}}{\left(\frac{0.0530404}{12}\right)} = 180 \times d.$$

This gives us $d = \$2777.78$, and the total payments will be $\$2777.78 \times 360 = \$1000{,}000.80$, almost exactly a million dollars. Since the loan was for $500,000, this APR, 5.30404%, is the rate that results in the total interest being equal to the amount financed. In other words, if your APR in a 30-year loan is higher than 5.30404%, then you will be paying more money for the interest than for the house (or whatever you are buying).

You can also see that at an APR of 9.3968%, you will be paying twice as much for the interest as for the house itself.

5) Many of today's mortgages have no pre-payment penalty, meaning that you can pay extra whenever you can, and that will save you a lot of money in the long run. Calculate your monthly payments if the APR is 3.0%. Then, find out how much you can save in the 30-year period *if you decide to pay $100 extra every month*. (This question requires the use of an amortization table, not covered in this chapter.)

This is a bit tricky since the balance is actually calculated every month, and you will end up shortening the term of your mortgage. But it is worth looking at it. In Part 1) above, we determined that each monthly payment will be $2108.02 at an APR of 3.0%. "$100 extra every month" means you will be paying $2208.02 every month. Detailed calculations (using an amortization table) show that the balance at the end of the 334th payment (27 years 10 months into the mortgage) will be $904.67, which will accrue an

interest of \$2.26 for one month, and the 335th and the last payment will be \$906.93. Your total payments will then be

$$\$2208.02 \times 334 + \$906.93 = \$738{,}385.61.$$

If you are making the minimum payment of \$2108.02 each month, the total payments will be

$$\$2108.02 \times 360 = \$758{,}887.20.$$

The difference is \$20,501.59, which is how much you will be saving over the 30-year period simply by paying \$100 extra every month.

(I am not sure how much you have paid to take this Liberal Arts Mathematics course, but these scenarios have just showed you how to save tens of thousands of dollars. Now, *that* is a great return on your investment!)

Figure 3.11[51]

[51] Image on PxHere

Chapter 4: Probability and Odds

"The biggest risk is not taking risk… In a world that is changing really quickly, the only strategy that is guaranteed to fail is not taking risks"—Mark Zuckerberg

We live in an uncertain world. We do not know what will happen. While some things may be easier to predict than others, we must face the reality that we do not know the future. However, there is a way to quantify the degree of uncertainty.

The probability of a specified event is the chance or likelihood that it will occur. There are several ways of viewing probability. One would be experimental in nature, where we repeatedly conduct an experiment. This is obviously an inductive approach as we are using specific instances and data to draw general conclusions.

Suppose we flipped a coin over and over and over again and it came up heads about half of the time; we would expect that in the future whenever we flipped the coin, it would turn up heads about half of the time. When a weather reporter says "there is a 10% chance of rain tomorrow," she is basing that on prior evidence; that out of all days with similar weather patterns, it has rained on 1 out of 10 of those days.

Another view would be subjective in nature, an educated guess. If someone asked you the probability that the Los Angeles Dodgers would win their next baseball game, it would be impossible to conduct an experiment where the same two teams played each other repeatedly, each time with the same starting lineup and starting pitchers, each starting at the same time of day on the same field under the precisely the same conditions. Since there are so many variables to consider, someone familiar with baseball and with the two teams involved might make an educated guess that there is a 75% chance the Dodgers will win the game; that is, if the same two teams were to play each other repeatedly under identical conditions, the Dodgers would win about three out of every four games. But this is just a guess, with no way to verify its accuracy, and depending upon how educated the educated guesser is, a subjective probability may not be worth very much.

We will return to the experimental and subjective probabilities from time to time, but in this course we will mostly be concerned with theoretical probability.

4.1 Basic Concepts[viii]

If you roll a die, pick a card from deck of playing cards, or randomly select someone and observe the person's hair color, we are executing an experiment or a trial. In probability, we look at the likelihood of different results. We begin with some important terminology.

Events and Outcomes

Events and Outcomes

The result of an experiment is called an **outcome**. An outcome is an individual result that cannot be broken down further.

The **sample space** is the **set** of all possible outcomes.

An **event** is any particular outcome or a collection of outcomes.

It should be emphasized that an **outcome** is a "bottom-level" result (cannot be split any further), that every trial must result in one and only one (unique) outcome, and that the outcomes are mutually exclusive (two outcomes cannot happen at the same time). List all of the possible outcomes, and you then have the **sample space**. One, two, or more (even all) outcomes together could be considered an **event**. In other words, an event is any subset of the sample space. Here is an example.

Example 1

Describe the sample space when a standard 6-sided die is rolled. What are some events?

[Solution]

The outcomes are 1, 2, 3, 4, 5, and 6. The sample space is the **set** {1,2,3,4,5,6}. Some examples of events are as follows:

- "We roll a 3" (consisting of just one outcome).
- "We roll a number bigger than 4" (consisting of the last two outcomes).
- "We roll an even number" (consisting of the three outcomes 2, 4, and 6).
- "We roll a number other than a 2" (consisting of five outcomes).

Probability is a **mathematical theory**. While it is a very practical theory and is applied in many fields, as a formal theory, it comes with some basic assumptions. We use $P(E)$ to represent the probability of an event E. First, it may be obvious, but the probability of an event is never negative, i.e., $P(E) \geq 0$ for all events *E*. Second, the probability of an event that is absolutely certain to occur is 1. For instance, if we roll a die, the probability that we get either an odd or an even number is 1 (which is, of course, 100%). Third, the probability of mutually exclusive

events is the sum of the two probabilities. In other words, if *A* and *B* are two events that can never occur simultaneously, then

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

Easy to see, right? It turns out that based on these three simple (and somewhat obvious) assumptions, we can build a rich mathematical theory of probability. Let us begin with a simple formula.

Basic Probability

Basic Probability

Given that all outcomes are equally likely, we can compute the probability of an event E using this formula:

$$P(E) = \frac{\text{Number of outcomes corresponding to the event } E}{\text{Total number of equally likely outcomes}}$$

Example 2

If we roll a 6-sided die, calculate

(1) the probability of rolling a 1.
(2) the probability of rolling a number bigger than 4.

[Solution]

Recall that the sample space is {1,2,3,4,5,6}. For (1), only one outcome corresponds to the event "rolling a 1," so the answer is $\frac{1}{6}$. For (2), the event consists of two outcomes (5 and 6), so the answer is $\frac{2}{6}$ or $\frac{1}{3}$.

Probabilities are essentially fractions and can be reduced to lower terms or can be converted to percentages.

Example 3

You have a bag with 20 cherries: 14 sweet and 6 sour ones. If you pick a cherry at random, what is the probability that it will be sweet?

[Solution]

There are 20 cherries that could be picked. Of these 20 possible outcomes, 14 are favorable (sweet), so the probability that the cherry will be sweet is $\frac{14}{20} = \frac{7}{10}$ (or 70%).

There is one potential complication to this example, however. It must be assumed that the probability of picking any of the cherries is the same as the probability of picking any other. This would not be true if (let us imagine) the sweet cherries are smaller than the sour ones. (The sour cherries would come to hand more readily when you sampled from the bag.) Let us keep in mind, therefore, that when we assess probabilities in terms of the ratio of favorable to all potential cases, we rely heavily on the assumption of equal probability for all outcomes.

Try It Now 1

At some random moment, you look at your digital clock and note the minutes reading.

a. **What is the probability of the minutes reading :15?**
b. **What is the probability the minutes reading less than :15?**

A standard deck of 52 playing cards consists of four suits (hearts, spades, diamonds, and clubs). Spades and clubs are black while hearts and diamonds are red. Each suit contains 13 cards, each of a different rank: an ace (which in many games functions as both a low card and a high card), cards numbered 2 through 10, a jack, a queen and a king.

Example 4

Compute the probability of randomly drawing one card from a deck and getting an ace.

[Solution]

There are 52 cards in the deck and 4 aces so $P(ace) = \frac{4}{52} = \frac{1}{13} \approx 0.0769$. There is a 7.69% chance that a randomly selected card will be an ace.

Certain and Impossible Events

An impossible event has a probability of 0.

A sure (certain) event has a probability of 1.

The probability of any event must be $0 \leq P(E) \leq 1$.

If you compute a probability and get an answer that is negative or greater than 1, you have made a mistake and should check your work. You cannot get an answer like 5 (even if you are really sure).

Complementary Events

Now let us examine the probability that an event does **not** happen. As in the previous section, consider the situation of rolling a six-sided die and first compute the probability of rolling a 6: the answer is P(6) =1/6. Now consider the probability that we do *not* roll a 6; there are 5 outcomes that are not a 6, so the answer is P(not a 6) = $\frac{5}{6}$. Notice that

$$P(6) + P(\text{not } 6) = \frac{1}{6} + \frac{5}{6} = \frac{6}{6} = 1.$$

This is not a coincidence. Consider a generic situation with n possible outcomes and an event E that corresponds to m of these outcomes. Then the remaining $n - m$ outcomes correspond to E not happening. Since "event E does not happen" is also an event (it is the complement of E and thus, after all, a subset of the sample space), let us denote "not E" by the symbol $\sim E$.

$$P(\sim E) = \frac{n-m}{n} = \frac{n}{n} - \frac{m}{n} = 1 - \frac{m}{n} = 1 - P(E).$$

Complement of an Event

The **complement** of an event is the event "E does not happen."
The notation $\sim E$ is used for the complement of event E.
We can compute the probability of the complement using $P(\sim E) = 1 - P(E)$.
Notice also that $P(E) = 1 - P(\sim E)$, also.

This is pretty obvious, right? If the probability of something happening is 40%, then the probability of that thing *not* happening must be 60%. But this idea of complementary events can be quite useful as we will see later. This idea, as well as the symbol ($\sim E$), should remind you of the negation of a statement in Chapter 1. This equation basically says that an event must either "happen" or "not happen" just as a statement must be either "True" or "False."

Example 5

If you pull a random card from a deck of playing cards, what is the probability it is not a heart?

[Solution]

There are 13 hearts in the deck, so $P(\text{heart}) = \frac{13}{52} = \frac{1}{4}$.

The probability of not drawing a heart is the complement:

$$P(\sim\text{heart}) = 1 - P(\text{heart}) = 1 - \frac{1}{4} = \frac{3}{4}.$$

Independent Events

When there are two or more events, one of them could influence the other and affect the probability. Imagine, for instance, a family with one boy and two girls. They are to choose two of the children to go for a walk. If the first one picked is a girl, then the probability that the second one to be picked is also a girl is 50% (because one boy and one girl are left); however, if

the first one picked is the boy, then the probability that the second one to be picked is a girl is 100% (because only the two girls are left). So the first outcome clearly affects the probability for the second choice.

On the other hand, many events do not influence each other. These are called **independent** events as you see below.

Example 6

Suppose we flipped a coin and then rolled a die. What is the probability of landing heads up on the coin and getting a 6 on the die?

[Solution]

We could list all possible outcomes, where "H" denotes heads and "T" denotes tails:

$$\{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}.$$

Notice there are $2 \times 6 = 12$ total outcomes. Out of these, only $H6$, is the desired outcome, so the probability is $\frac{1}{12}$. Note that the coin and the die do not in any way affect each other.

This example was looking at two **independent** events.

Independent Events

Events A and B are **independent events** if the probability of Event B occurring is the same whether or not Event A occurs.

In the previous example, the probability of getting a 6 on the die is $\frac{1}{6}$ regardless of whether or not the coin landed heads up. Therefore, these two events are independent.

Example 7

Are these events independent?

(1) A fair coin is tossed twice. The two events are (a) the first toss lands heads up and (b) the second toss lands heads up.
(2) The two events are (a) "It will rain tomorrow in San Francisco" and (b) "It will rain tomorrow in Oakland."
(3) You draw a card from a deck and then draw another card *without* replacing the first. The two events are (a) the first card is red and (b) the second card is red.

[Solution]

(1) The probability that the second coin toss lands heads up is 1/2 regardless of whether the first toss comes up heads or tails, so these events are independent.
(2) These events are not independent because it is more likely that it will rain in Oakland on days it rains in San Francisco since the two cities are right next to each other.
(3) The probability of the second card being red depends on whether the first card is red or not, so these events are *not* independent. Remember, if you draw one red card, now the stack has one fewer red card, affecting the probability on the second draw.

When two events are independent, the probability of both occurring is the product of the probabilities of the individual events.

$P(A \text{ and } B)$ for Independent Events

If events A and B are independent, then the probability of both A and B occurring is

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

If you look back at the coin and die examples from earlier, you can see that the number of outcomes of the first event multiplied by the number of outcomes in the second event is the total number of possible outcomes in the combined event. We will learn how to count the total number of outcomes in a later section.

Example 8

In your drawer you have 10 pairs of socks, and 6 of those pairs are white. You also have 7 shirts, and 3 of them are white. If you randomly reach in and pull out a pair of socks and a shirt, what is the probability that both are white?

[Solution]

The probability of choosing a white pair of socks is $\frac{6}{10}$. The probability of choosing a white shirt is $\frac{3}{7}$. Note that these are independent events because which socks you pick does not influence which shirt you pick. Therefore, the probability of both being white is $\frac{6}{10} \times \frac{3}{7} = \frac{18}{70} = \frac{9}{35}$.

Try It Now 2

A card is pulled from a deck of cards and noted. The card is then replaced, the deck is shuffled, and a second card is removed and noted. What is the probability that both cards are aces?

The previous examples looked at the probability of both events occurring. This is called the **conjunction** (or **intersection**) of two events. Now we will look at the probability of either (one or both) event occurring. This is called the **disjunction** (or **union**).

Example 9

Suppose we flipped a coin and rolled a die. What is the probability that the coin lands heads up *or* you get a 6 on the die?

[Solution]

Here, there are still 12 possible outcomes:

$$\{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}.$$

By simply counting, we can see that 7 of the outcomes have heads on the coin *or* a 6 on the die *or* both – we use "*or*" inclusively here, meaning "heads, a 6, or both" (these 7 outcomes are H1, H2, H3, H4, H5, H6, T6), so the probability is $\frac{7}{12}$. How could we have found this from the individual probabilities?

As we would expect, $\frac{1}{2}$ of these outcomes have heads, and $\frac{1}{6}$ of these outcomes have a 6 on the die. If we add these, $\frac{1}{2}+\frac{1}{6}=\frac{6}{12}+\frac{2}{12}=\frac{8}{12}$,not the correct probability. Looking at the outcomes, we can see why. The outcome H6 was counted twice since it contains both "heads" and a "6"; the probability of both "heads" *and* a "6" is $\frac{1}{12}$. If we subtract out this double count, we would get the correct probability: $\frac{8}{12}-\frac{1}{12}=\frac{7}{12}$.

$P(A$ or $B)$

The probability of either A or B (or both) occurring is

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

Example 10

Suppose we draw one card from a standard deck. What is the probability that we get a queen or a king?

[Solution]

There are 4 queens and 4 kings in the deck. Note that there are no cards that are both a queen and a king, so $P(\text{king or queen}) = 0$. So we have

$$P(\text{king or queen}) = P(\text{king}) + P(\text{queen}) - P(\text{king and queen}) = \frac{4}{52} + \frac{4}{52} - 0 = \frac{8}{52} = \frac{2}{13}.$$

In this example, the events were **mutually exclusive**, so $P(A \text{ or } B) = P(A) + P(B)$.

Example 11

Suppose we draw one card from a standard deck. What is the probability that we get a red card or a king?

[Solution]

Half the cards are red, so $P(\text{red}) = \frac{26}{52}$. There are four kings, so $P(\text{king}) = \frac{4}{52}$. But there are two red kings, so $P(\text{red and king}) = \frac{2}{52}$. We can then find the answer:

$$P(\text{red or king}) = P(\text{red}) + P(\text{king}) - P(\text{red and king}) = \frac{26}{52} + \frac{4}{52} - \frac{2}{52} = \frac{28}{52} = \frac{7}{13}.$$

Example 12

The table below (called a **two-way table**) shows how many drivers (out of 665) received a speeding ticket last year as well as the color of their car. Find the two probabilities below and determine if these events (having a red car and getting a ticket) are independent.

(1) has a red car *and* got a speeding ticket.
(2) has a red car *or* got a speeding ticket.

	Speeding ticket	**No speeding ticket**	**Total**
Red car	15	135	150
Not red car	45	470	515
Total	60	605	665

[Solution]

(1) We can see that 15 people of the 665 drivers had a red car *and* got a speeding ticket, so the probability is $\frac{15}{665} \approx 0.02256$, or approximately 2.256%. Note that the probability of a driver having a red car is $\frac{150}{665}$ and the probability of a driver getting a ticket is $\frac{60}{665}$. The product of these is $\frac{150}{665} \times \frac{60}{665} \approx 0.02035$, or approximately 2.035%. These probabilities are not equal to each other, showing something quite important. **These two events are not independent** (that is, the probability of

getting a ticket is affected in one way or another by the vehicle color). If they were independent, the "and" probability would be exactly the product of the two probabilities.

(2) We could answer this question by simply adding up the numbers: there are 15 people with red cars and speeding tickets, 135 with red cars but no ticket, and 45 with a ticket but no red car. The sum is 195. So the probability is $\frac{195}{665} \approx 0.2932$, or about 29.32%. We also could have found this probability by complementary event. The *opposite* of "having a red car or getting a ticket" is "neither a red car nor a speeding ticket," which describes the 470 drivers on the bottom right box. Thus, the answer is $1 - \frac{470}{665} = \frac{195}{665}$.

Exercises 4.1

1. A ball is drawn randomly from a jar that contains 6 red balls, 2 white balls, and 5 yellow balls. Find the probability of the given event.
   (a) a red ball is drawn
   (b) a white ball is drawn

2. Suppose you write each letter of the alphabet on a different slip of paper and put the slips into a hat. What is the probability of drawing one slip of paper from the hat at random and getting a vowel (a, e, i, o, u)?

3. A group of people were asked in a survey if they had run a red light last year. 150 responded "yes" while 185 responded "no." Find the probability that a randomly chosen person in this survey ran a red light last year.

4. In a survey, 205 people indicated they prefer cats, 160 indicated they prefer dogs, and 40 indicated they do not enjoy either pet. Find the probability that a randomly chosen person in this survey prefers cats.

5. What is the probability of tossing a die and getting an odd number?

6. What is the probability of tossing a die and getting a 7?

7. A test was given to a group of students. The grades and gender are summarized below. If one student is chosen at random, what is the probability that the student is female? Also, what is the probability that the student did *not* earn a C?

	A	B	C	Total
Male	8	18	13	39
Female	10	4	12	26
Total	18	22	25	65

8. The table below shows the number of credit cards owned by individuals. If one person is chosen at random, what is the probability that the person has no credit cards? Also, what is the probability that the person is male?

	Zero	One	Two or more	Total
Male	9	5	19	33
Female	18	10	20	48
Total	27	15	39	81

9. What is the probability of tossing a die and getting a number other than 3?

10. What is the probability of tossing a die and getting a number less than 3?

11. If you pick one card at random from a standard deck of cards, what is the probability it will be a red king?

12. If you pick one card at random from a standard deck of cards, what is the probability it will be a diamond?

13. What is the probability of rolling a 12-sided die (called a *dodecahedral* die) and getting a number other than 8?
14. If you pick one card at random from a standard deck of cards, what is the probability it is not the ace of spades?

15. If you pick one card at random from a standard deck of cards, what is the probability it is not an ace?

16. If you pick one card at random from a standard deck of cards, what is the probability it is not a spade?

17. If you pick one card at random from a standard deck of cards, what is the probability it is neither an ace nor a spade?

18. Suppose that 18% of people own dogs in the world. If you pick two people at random, what is the probability that they both own a dog? (What do you have to assume here?)

19. Suppose $P(A) \neq 0$ and $P(B) \neq 0$ (in other words, A and B are events that *could* happen). If these two events are mutually exclusive, are they always independent, sometimes independent, or never independent? Explain why.

Figure 4.1[52]

[52] Image by Becky Lai is licensed under CC BY-NC-ND 2.0

4.2 Conditional Probability

We often need to find the probability of an event **given** that another event has occurred. As we will see below, this type of probability has a significant role to play on issues like social justice, discrimination, equity, and disproportionate impact in society.

Conditional Probability

The probability that the event A occurs, **given** that event B has occurred, is called the conditional probability of A **given** B. It is denoted $\boldsymbol{P(A|B)}$**.**

Let us think about how conditional probabilities are related to other probabilities. The example in the previous section (Example 12) involving red cars and speeding tickets is a good starting point. Here is that **two-way table** again. Let T be the event that the driver got a ticket, and let R denote the event that the driver was in a red car.

	Speeding ticket (T)	**No speeding ticket ($\sim T$)**	**Total**
Red car (R)	15	135	150
Not red car ($\sim R$)	45	470	515
Total	60	605	665

What is the probability that a driver got a speeding ticket *given* that the driver had a red car? Note that this additional information—the fact the driver had a red car—is assumed ("given"), so we need only to look at the top row. Out of the 150 drivers with red cars, 15 of them got a ticket. Thus, the probability that a driver got a speeding ticket *given* that the driver had a red car is

$$P(T|R) = \frac{15}{150} = \frac{1}{10} = 10\%.$$

(Similarly, one can easily see that the probability that a driver got a ticket *given* that the car was not red is, looking at the numbers in the second row, $P(T|{\sim}R) = \frac{45}{515} = \frac{9}{103} \approx 8.738\%$.)

Now, let us look at conditional probabilities mathematically. How is $P(T|R)$ related to $P(T)$ and $P(R)$? Note that 15 is the number of drivers who drove a red car *and* got a ticket (conjunction). This is the number associated with the event "R and T." 150 is the number associated with the event R. One can see that $P(R \text{ and } T) = \frac{15}{665}$ while $P(R) = \frac{150}{665}$. Now, remember that "T given R" means that we assume that R is already true (for sure) and then we look for the probability that T is also true (T occurs). So the conditional probability $P(T|R)$ is essentially the probability that both R *and* T occur after knowing for sure that R has

occurred. In other words, you are 100% sure that R occurs, and that is your "universe" (assumption, the denominator, the base). Thus, we have

$$P(T|R) = \frac{P(R \text{ and } T)}{P(R)}.$$

In this example, this number, as we saw already, is 10%. You can also use this formula to see this:

$$P(T|R) = \frac{P(R \text{ and } T)}{P(R)} = \frac{{}^{15}/_{665}}{{}^{150}/_{665}} = \frac{15}{150} = \frac{1}{10} = 10\%.$$

When a table is given as in this case, note that both the numerator and the denominator of this calculation are fractions, each with the same denominator (665 in this example). Hence, you can ignore that total number and simply use those numbers appearing in the table, as we did the first time we got the answer $P(T|R) = 10\%$.

More generally, suppose we have two events, A and B. The probability $P(A|B)$ is the likelihood of A occurring, having known with certainty that B has occurred. For this to happen, both A and B must occur, but we seek to find the likelihood of this happening out of all the situation where B has already occurred. So $P(B)$ becomes the "universe," or "sample space," or the *base* (the denominator in a percentage). This leads us to the formula for conditional probability:

Conditional Probability $P(A \mid B)$:
The **probability of A given B is**

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

(where we assume $P(B) \neq 0$).
If events A and B are independent, then the probability of both A and B occurring is

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

This is equivalent to

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

This is the general formula for the probability of a conjunctive ("and") event. If A and B happen to be **independent**, then $P(A)$ is not influenced by whether B has occurred or not, so $P(A) = P(A|B)$, and this gives us the earlier formula $P\ (A \text{ and } B) = P(A) \times P(B)$.

But remember, if you are given a table of numbers as above, you do not need to use these formulas. Instead, you can simply look at one row or one column of the table to figure out conditional probabilities.

Example 1

Find the probability that a die rolled shows a 3, given that a flipped coin lands heads.

[Solution]

These are independent events, so the probability of the die rolling a 3 is $P(3) = \frac{1}{6}$ regardless of the result of the coin flip. Or, using the formula,

$$P(3|H) = \frac{P(3 \text{ and } H)}{P(H)} = \frac{1/12}{1/2} = \frac{1}{6}$$

Example 2

What is the probability that two cards drawn at random from a deck of playing cards will both be aces?

[Solution]

It might seem that you could use the formula for the probability of two independent events and simply multiply $\frac{4}{52} \cdot \frac{4}{52} = \frac{1}{169}$. This would be incorrect, however, because the two events are not independent. Here is a way to think about this more carefully.

Let $A1$ be the event that the first card drawn is an ace, and let $A2$ be the event that the second card is an ace. Clearly, $P(A1) = \frac{4}{52} = \frac{1}{13}$. Now, if the first card drawn is an ace, what is the probability that the second card drawn is an ace? It is *not* $\frac{1}{13}$. Why not? After an ace is drawn on the first draw, there are 3 aces out of 51 total cards left. So the conditional probability of drawing an ace after one ace has been already drawn is

$$P(A2|A1) = \frac{3}{51} = \frac{1}{17}.$$

Thus, the probability of both cards being aces is

$$P(A1 \text{ and } A2) = P(A1) \times P(A2|A1) = \frac{1}{13} \times \frac{1}{17} = \frac{1}{221}$$

Example 3

Using the same data as before (shown again below), compare and contrast the probabilities that a randomly chosen driver (1) got a speeding ticket *given* that the car was red; (2) got a speeding ticket *given* that the car was *not* red; (3) had a red car *given* that the driver got a speeding ticket; and (4) had a red car *given* that the driver did *not* get a ticket.

	Speeding ticket (T)	No speeding ticket ($\sim T$)	Total
Red car (R)	15	135	150
Not red car ($\sim R$)	45	470	515
Total	60	605	665

[Solution]

We already answered (1) and (2) at the beginning of this section. For (1),

$$P(T|R) = \frac{15}{150} = \frac{1}{10} = 10\%.$$

Similarly, for (2),

$$P(T|\sim R) = \frac{45}{515} \approx 8.738\%.$$

Now, for (3) the situation is quite different. Here, because we know that the driver got a speeding ticket, we are only considering the 60 people in the first column of the table. Of those, 15 had a red car, so

$$P(R|T) = \frac{15}{60} = \frac{1}{4} = 25\%.$$

Similarly, for (4), we are looking only at the 605 drivers in the second column. Of those, 135 had a red car. Therefore,

$$P(R|\sim T) = \frac{135}{605} \approx 22.314\%$$

What do all these probabilities mean? How would you interpret, compare, and contrast them? First, we discovered that $P(T|R) = 10\%$. This means that if you drive a red car, you have a 10% chance of getting a speeding ticket. Since $P(T|\sim R) \approx 8.7\%$, if you do not drive a red car, you have less chance of getting a speeding ticket. This is the type of data used in discrimination lawsuits and equity dialogues. Also, insurance companies and actuaries carefully look at these data to determine your insurance rates.

We also found out that $P(R|T) = 25\%$, suggesting that if you got a speeding ticket, there is a 25% chance you were in a red car. Since $P(R|\sim T) \approx 22\%$, a driver who did not get a speeding ticket has a lower chance of having been driving a red car.

Note that all these are different conditional probabilities, and $P(T|R) \neq P(R|T)$. The left-hand side is the probability that a driver of a red car gets a speeding ticket whereas the right-hand side is the probability that a driver who got a speeding ticket was in a red car. Pay close attention when you use conditional probabilities so you do not mix up these different numbers.

Example 4

If you pull 2 cards out of a deck, what is the probability that both are spades?

[Solution]

The probability that the first card is a spade is $\frac{13}{52} = \frac{1}{4}$. The probability that the second card is a spade, given the first was a spade, is $\frac{12}{51}$ since there is one less spade in the deck and one less total cards. Note $\frac{12}{51} = \frac{4}{17}$. The probability that both cards are spades is $\frac{1}{4} \times \frac{4}{17} = \frac{1}{17} \approx 0.0588$, which is less than 6%.

Try It Now 3

In your drawer you have 10 pairs of socks, and 6 of those pairs are white. If you reach in and randomly grab two pairs of socks, what is the probability that both are white?

One very important application of conditional probabilities is **false positives** and **false negatives** in various kinds of tests. Since no test is completely flawless, these errors may result, sometimes causing confusion and dire consequences.

Example 5

A home pregnancy test was given to women, and then pregnancy was verified through more accurate blood tests. The following two-way table shows the home pregnancy test results. Find

(1) the probability that a woman was not really pregnant even though (given that) she tested **positive** on the home pregnancy test.
(2) the probability that a woman tested **positive** on the home pregnancy test even though (given that) she really was *not* pregnant.

	Positive (POS)	**Negative (NEG)**	**Total**
Pregnant (PG)	70	4	74
Not Pregnant (~PG)	5	14	19
Total	75	18	93

[Solution]

These two questions sound very similar, right? Are they asking the same thing? No, there is a subtle difference. Pay close attention when dealing with conditional probabilities.

We first introduce some notation. We will let "PG" to denote "she is pregnant," "~PG" for "not pregnant," "POS" for "testing positive," and "NEG" for "testing negative" on the home pregnancy test.

In (1), POS is given. The problem is asking for $P(\sim PG|POS)$, the probability that a woman is $\sim PG$ given that she tested positive (POS). Since we are given that the test result was positive, we are only considering the 75 women in the first column. Out of this group, 5 were not actually pregnant. Hence, $P(\sim PG|POS) = \frac{5}{75} \approx 0.0667$.

b) is asking for $P(POS|\sim PG)$ because the fact that the woman is not pregnant is given and we are asked for the probability that she tested positive. Since we are considering only those who are not pregnant, we focus our attention on the second row (the 19 women), and we see that $P(POS|\sim PG) = \frac{5}{19} \approx 0.263$.

In this example, five women *tested* **positive** even though they were in fact not pregnant. This is what is called a false positive. A **false positive** is a positive test result when the test should have shown a negative result. (The positive result of the test was a "false alarm.")

On the other hand, note that four women *tested* **negative** even though they were indeed pregnant. This is what is called a **false negative**. A false negative is a negative test result when the test should have shown a positive result. (The negative result of the test was "false.") In this case, this could be problematic because a false negative allows the pregnant woman to assume that she is not pregnant and she may continue a lifestyle that a pregnant woman may need to avoid.

Let us now concentrate on more complex conditional probability problems. Misunderstanding of these probabilities can lead to serious consequences in public health, policy, and other areas of our society.

Example 6

Suppose a certain rare virus has an incidence rate of 2% (that is, it afflicts 2% of the population). A test has been devised to detect this virus. The test does not produce false negatives (that is, anyone who has the disease will test positive for it), but the false positive rate is 5% (that is, 5% of people who do not have the virus will test positive). Suppose a randomly selected person takes the test and tests positive. What is the probability that this person actually has the virus?

[Solution]

There are two ways to approach the solution to this problem. One involves an important result in probability theory called **Bayes' Theorem**, which we will not use here (but will be stated later). We use an alternative and much more **intuitive approach** called "hypothetical 1000."

Let us break down the information in the problem. The incidence rate of 2% means 2 in every 100 people has the virus. No false negative means everyone who has the disease will test positive. *If a person tests negative, the person cannot have the virus.* However, **its inverse may not be true**; *if you test positive, you may or may not have the virus*. In fact, the 5% false positive rate means out of every 100 people who do not have the virus, unfortunately 5 will test positive.

The key here is to create a hypothetical situation and apply the numbers as set out above. First, suppose we randomly (and hypothetically) select 1000 people and give the test. How many of them actually have the virus? Only 2%, which is 20 people. The other 980 do not have the virus. We will now create a table and enter these numbers.

	Tested Positive (P)	**Tested Negative ($\sim P$)**	**Total**
Virus (V)			20
No Virus ($\sim V$)			980
Total			1000

Now, all 20 with the virus will test positive, so the top row is 20 for the first column and 0 for the second. Out of the 980 without the virus, 5% of those will test positive. $980 \times 0.05 = 49$, so we enter this number under "Tested Positive" in the second row. The rest of the 980 people, i.e., $980 - 49 = 931$, will test negative. Enter this number for the "No Virus, Tested Negative" box.

After adding each column, we have the completed table completed.

	Tested Positive (P)	**Tested Negative ($\sim P$)**	**Total**
Virus (V)	20	0	20
No Virus ($\sim V$)	49	931	980
Total	69	931	1000

Now, the original question was asking for the probability of someone having the virus given that the person tested positive. If P denotes "tested positive" and V denotes "(actually) having the virus," what we are looking for is $P(V|P)$. There are 69 people tested positive, but only 20 of them had the virus, so the answer is $\frac{20}{69} \approx 0.289855 \ldots \approx 29.0\%$.

Now, stop and take a look at this number. Try to interpret it. You see that a vast majority (71.0%) of those who tested positive are actually virus-free *even though the test itself boasts a 5% error level*. The reason this happens is that the percentage of people with the virus was low, at 2%. When a high percentage of people are not sick (or with a virus), more false positives will occur, distorting these conditional probabilities.

If you just tested positive on this test, you need not panic. In fact, an informed doctor or healthcare provider may kindly tell you, "Your test came back positive, but don't worry about this yet. There is more than an 70% chance that you don't actually have the virus." Then another test is probably recommended for you.

Unfortunately, not many people, including some health professionals, know how these probabilities work. A question similar to this was posed to doctors and medical students at the Harvard Medical School, and the results were published in a 1978 *New England Journal of Medicine* article.[53] Only about 18% of the participants got the right answer.

The significance of this finding and similar results from other studies in the intervening years lies in the possibly catastrophic consequences it might have for patient care. If doctors (erroneously) think the chances that a positive test result nearly guarantees that a patient has a disease, they might begin an unnecessary and possibly harmful treatment regimen on a healthy patient. Or worse, as in the early days of the AIDS crisis when being HIV-positive was often equated with a death sentence, the patient might take a drastic action.

As we have seen in this hypothetical example, the most responsible course of action for treating a patient who tests positive may be to comfort the patient and order further, more reliable, tests to verify the diagnosis.

One of the reasons that the doctors and medical students in the study did so poorly is that such problems, when presented in the types of statistics courses that medical students often take, are solved by Bayes' Theorem, which is stated as follows (you do *not* need to learn this):

[53] Casscells, W., Schoenberger, A., and Graboys T. (1978). Interpretation by Physicians of Clinical Laboratory Results. *New England Journal of Medicine*, (299), 999—1001. https://www.nejm.org/doi/full/10.1056/NEJM197811022991808

Bayes' Theorem

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

The problem is that we are much more likely to remember the way we solved the problem with a table, with a fictional group of 1000 people, than to remember Bayes' Theorem. We highly recommend that you use the table method, which has gained so much popularity in recent years that it has its own name: **"hypothetical 1000 table**."

Example 7

A certain disease has an incidence rate of 4%. If the false negative rate is 5% and the false positive rate is 10%, find the probability that a person who tests positive actually has the disease.

[Solution]

Imagine 1000 people who are tested. 4% of them, or 40 people, have the disease. The other 960 does not. 5% of those 40 with the disease, which is 2 patients, test negative. The other 38 test positive. Out of the 960 without the disease, 10% of them, or 96 of them, test positive (false positive). Others, $960 - 96 = 864$, test negative. Adding each column, we get the following table.

	Tested Positive (P)	**Tested Negative ($\sim P$)**	**Total**
Disease (D)	38	2	40
No Disease ($\sim D$)	96	864	960
Total	134	866	1000

The probability that someone who tests positive actually has the disease is

$$P(D|P) = \frac{38}{134} \approx 0.28358,$$

which is a little less than 30%.

Try It Now 4

A certain disease has an incidence rate of 0.5%. If there are no false negatives and if the false positive rate is 4%, find the probability that a person who tests positive actually has the disease.

Exercises 4.2

1. Suppose a jar contains 17 red marbles and 32 blue marbles. If you reach in the jar and pull out 2 marbles at random, find the probability that both are red.

2. Suppose you write each letter of the alphabet on a different slip of paper and put the slips into a hat. If you pull out two slips at random, find the probability that both are vowels (a, e, i, o, u).

3. Bert and Ernie each have a well-shuffled standard deck of 52 cards. They each draw one card from their own deck. Compute the probability that:
   (a) Bert and Ernie both draw an ace.
   (b) Bert draws an ace but Ernie does not.
   (c) neither Bert nor Ernie draws an ace.
   (d) Bert and Ernie both draw a heart.
   (e) Bert gets a card that is not a jack and Ernie draws a card that is not a heart.

4. Bert has a well-shuffled standard deck of 52 cards, from which he draws one card. Ernie has a 6-sided die, which he rolls at the same time Bert draws a card. Compute the probability that:
   (a) Bert gets a jack and Ernie rolls a 5.
   (b) Bert gets a heart and Ernie rolls a number less than 6.
   (c) Bert gets a face card (jack, queen, or king) and Ernie rolls an even number.
   (d) Bert gets a red card and Ernie rolls a 2.

5. Find the probability of drawing a king from a deck of cards and then drawing a queen.

6. Find the probability of drawing two spades from a deck of cards.

7. A math class consists of 25 students: 14 females and 11 males. Two students are selected at random to participate in a probability experiment. Find the probability that
   (a) a male is selected and then a female.
   (b) a female is selected and then a male.
   (c) two males are selected.
   (d) two females are selected.
   (e) no males are selected.

8. A math class consists of 25 students, 14 females and 11 males. Three students are selected at random to participate in a probability experiment. Find the probability that
   (a) a male is selected and then two females.
   (b) a female is selected and then two males.
   (c) two females are selected then one male.
   (d) three males are selected.
   (e) three females are selected.

9. In a science class, the grades of a test and gender are summarized below. If one student is chosen at random, find the probability that
   (a) the student is female and earned an A.
   (b) the student is female or earned a B.
   (c) the student is male.
   (d) the student is female given that the student earned an A.
   (e) the student got an A given that it is a male student.

	A	B	C	Total
Male	8	18	13	39
Female	10	4	12	26
Total	18	22	25	65

10. The table below shows the number of credit cards owned by individuals. If one person is chosen at random, find the probability that
    (a) the person is male and has two or more credit cards.
    (b) the person is female or has no credit cards.
    (c) the person is female.
    (d) the person is male given that the person has two or more credit cards.
    (e) the person has no credit cards given that he is male.

	Zero	One	Two or more	Total
Male	9	5	19	33
Female	18	10	20	48
Total	27	15	39	81

11. Find the probability of flipping a coin and getting heads, given that the previous flip was tails.

12. Find the probability of rolling a “1” on a fair die, given that the last 3 rolls were all 1’s.

13. Suppose a math class contains 25 students: 14 females (three of whom speak French) and 11 males (two of whom speak French). Find the probability that a randomly selected student speaks French, given that the student is female.

14. Suppose a math class contains 25 students: 14 females (three of whom speak Spanish) and 11 males (two of whom speak Spanish). Find the probability that a randomly selected student is male, given that the student speaks Spanish.

15. A certain virus infects one in every 400 people. A test used to detect the virus in a person comes out positive 90% of the time if the person has the virus and 10% of the time if the person does not have the virus. Let V be the event "the person is infected" and P be the event "the person tests positive."
    (a) Find the probability that a person has the virus given that the person has tested positive, i.e. find $P(V|P)$.
    (b) Find the probability that a person does not have the virus given that they test negative, i.e. find $P(\sim V|\sim P)$.

16. A certain virus infects one in every 2000 people. A test used to detect the virus in a person comes out positive 95% of the time if the person has the virus and 5% of the time if the person does not have the virus. Let V be the event "the person is infected" and P be the event "the person tests positive."
    (a) Find the probability that a person has the virus given that the person has tested positive, i.e. find $P(V|P)$.
    (b) Find the probability that a person does not have the virus given that they test negative, i.e. find $P(\sim V|\sim P)$.

17. A certain disease has an incidence rate of 0.3%. If the false negative rate is 6% and the false positive rate is 4%, find the probability that a person who tests positive actually has the disease.

18. A certain disease has an incidence rate of 0.1%. If the false negative rate is 8% and the false positive rate is 3%, find the probability that a person who tests positive actually has the disease.

19. A certain group of symptom-free women between the ages of 40 and 50 are randomly selected to participate in mammography screening. The incidence rate of breast cancer among such women is 0.8%. The false negative rate for the mammogram is 10%. The false positive rate is 7%. If the mammogram results for a particular woman are positive (indicating that she has breast cancer), what is the probability that she actually has breast cancer?

20. About 0.01% of men with no known risk behavior are infected with HIV. The false negative rate for the standard HIV test 0.01% and the false positive rate is also

0.01%. If a randomly selected man with no known risk behavior tests positive for HIV, what is the probability that he is actually infected with HIV?

Figure 4.2[54]

[54] Image by Marco Verch Professional Photographer is licensed under CC BY 2.0

4.3 Counting

Counting? You already know how to count or you would not be taking a college-level math class, right? Well yes, but what we will really be investigating here is ways of counting efficiently. For some probability questions, we will need to count some very large numbers, like the number of possible lottery tickets. One way to do this would be to write down every possible set of numbers that might show up on a lottery ticket, but you would not want to do this. You will soon see why.

Basic Counting

We will start with some reasonable counting problems to develop the ideas that we will soon need.

Example 1

Suppose at a particular restaurant you have three choices for an appetizer (soup, salad, or breadsticks) and five choices for a main course (hamburger, sandwich, quiche, fajita, or pizza). If you are allowed to choose exactly one item from each category for your meal, how many different meal options do you have?

[Solution]

It is possible that you systematically list all the possibilities here: soup & hamburger, soup & sandwich, …, breadsticks and pizza. The key word here is "systematically." For example, here you could list all five choices involving soup first, then five with salad, and then five with breadsticks. Alternatively, you could construct a 3-by-5 table and list all possibilities. But as a more general approach, let us draw a **tree diagram**:

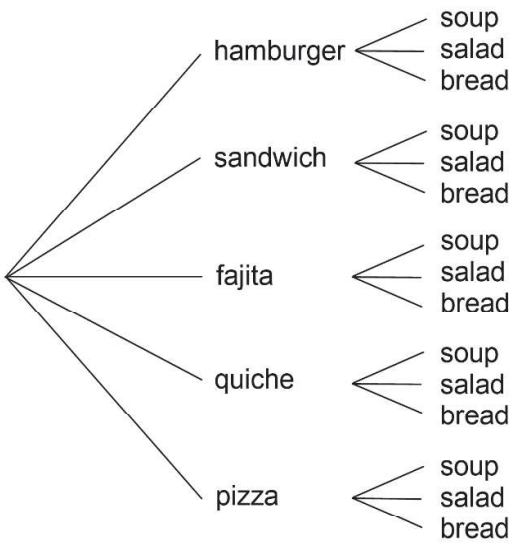


Figure 4.3[55]

[55] Probability by David Lippman is licensed under CC BY-SA 3.0 US.

This is called a "tree" diagram because at each stage we branch out, like the branches on a tree. In this case, we first drew five branches (one for each main course) and then for each of those branches we drew three more branches (one for each appetizer). We count the number of branches at the final level and get 15.

If we wanted, we could instead draw three branches at the first stage for the three appetizers and then five branches (one for each main course) branching out of each of those three branches.

Now we know how to count possibilities using lists and tree diagrams. In both cases, we ended up with the answer 15 because there were 3 choices in the first stage (appetizers) and 5 in the second (main course). We generalize this technique as our first counting principle:

The Multiplication Rule (Basic Counting Rule)

If we are to choose one item from each of two separate categories where there are m items in the first category and n items in the second category, then the total number of available choices is $m \times n$.

Example 2

There are 21 novels and 18 volumes of poetry on a reading list for a college English course. How many different ways can a student select one novel and one volume of poetry to read during the semester?

[Solution]

There are 21 choices from the first category and 18 for the second, so there are $21 \times 18 = 378$ possibilities.

The Multiplication Rule can be extended when there are more than two categories as we see in the following examples.

Example 3

Suppose at a particular restaurant you have three choices for an appetizer (soup, salad, or breadsticks), five choices for a main course (hamburger, sandwich, quiche, fajita, or pasta) and two choices for dessert (pie or ice cream). If you are allowed to choose exactly one item from each category for your meal, how many different meal options do you have?

[Solution]

There are 3 choices for an appetizer, 5 for the main course, and 2 for dessert, so there are $3 \times 5 \times 2 = 30$ possibilities.

Example 4

A quiz consists of 4 true-or-false questions. In how many ways can a student answer the quiz?

[Solution]

There are 4 questions. Each question has 2 possible answers (true or false), so the quiz may be answered in $2 \times 2 \times 2 \times 2 = 2^4 = 16$ different ways.

This last example actually illustrates a particular type of situations. If the choices are identical in each stage or category, then the product is simply the number of choices in each stage, raised to the appropriate exponent. For instance, if you roll a die 5 times, the total number of outcomes will be $6^5 = 7776$. Because it is still the Multiplication Rule, we will not state it as a separate rule.

Try It Now 5

Suppose at a particular restaurant you have eight choices for an appetizer, eleven choices for a main course, and five choices for dessert. If you are allowed to choose exactly one item from each category for your meal, how many different meal options do you have?

Permutations

In counting problems, two aspects are quite important as you will see below. One is whether or not repetition is allowed. In other words, ask yourself, "Can I pick the same item repeatedly?" The other is whether or not the order in which the choices are made is relevant. Let us start with an example that illustrates one of these problem types.

Example 5

How many different ways can the letters W, X, Y, and Z be arranged to form a four-letter code?

[Solution]

First, recognize that this problem is different from the meal examples seen earlier. Instead of choosing one item from each of several different categories, we are repeatedly choosing items from the *same* category (the category is the four letters W, X, Y, and Z). Also, each time we choose an item we *do not replace* it, so there is one fewer choice at the next stage. We have 4 choices for the first letter (say we choose W), then 3 choices for the second (X, Y, and Z). Say we choose Y; then 2 choices are left for the third letter (X and Z). Say we choose Z. Now, only one choice is left at the last stage (X). Therefore, there are $4 \times 3 \times 2 \times 1 = 24$ possible codes consisting of these four letters.

Note that in this example, *once a letter is used, you cannot repeat that same letter* (because you are arranging these four letters to create a code). In other words, **no repetition is allowed**. Further, the order makes a difference. After all, the code "XYZW" is different from "ZYWX." In other words, **the order is relevant**. These two attributes will be further explored below.

It is easy to see, as in this example, that if we had started with n items, the number of different ways to arrange all of these n items would have been $n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$. It turns out that this expression shows up quite often in mathematics. It is called the factorial and is denoted $n!$ (yes, with an exclamation mark).

n Factorial

$$\boldsymbol{n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1}$$

For example, in the above example, we saw that $4! = 24$. You can easily check that $3! = 3 \times 2 \times 1 = 6$, $2! = 2 \times 1 = 2$, and $1! = 1$. Surprisingly, for reasons far more sophisticated than the scope of this textbook allows, it is defined that $0! = 1$.

You may be wondering why the exclamation mark (!) is used as the factorial symbol. The reason is that these numbers are often surprisingly big. For instance, similar to the example above, $10!$ would be the number of codes one can create using ten letters of the alphabet, or, equivalently, the number of ways in which you can line up 10 people. This number is surprisingly large: $10! = 3{,}628{,}800$. Yes, there are over 3.6 million ways to line up 10 people. Shocking! (and this is why the exclamation mark is used.)

Example 6

In how many ways can five different door prizes be distributed among five people? [Solution]

There are 5 choices of prize for the first person, 4 choices for the second, and so on. So the number of ways in which the prizes can be distributed is $5 \times 4 \times 3 \times 2 \times 1 = 5! = 120$ ways.

In the last two examples, we were arranging all available items, and the answer involved a factorial expression. Now we will consider some slightly different examples.

Example 7

A charity benefit is attended by 25 people, and 3 gift certificates are given away as door prizes: one gift certificate for $100, another for $25, and the third for $10. Assuming that no person receives more than one prize, how many different ways can the three gift certificates be awarded?

[Solution]

We see that there are 25 choices for the person who receives the $100 certificate, 24 remaining choices for the $25 certificate, and 23 choices for the $10 certificate. Therefore, by the Multiplication Rule there are $25 \times 24 \times 23 = 13{,}800$ ways in which the prizes can be awarded.

Example 8

Eight sprinters have made it to the Olympic Finals in the 100-meter race. In how many different ways can the gold, silver, and bronze medals be awarded?

[Solution]

There are 8 choices for the gold medal winner, 7 remaining choices for the silver, and 6 for the bronze, so there are $8 \times 7 \times 6 = 336$ ways the three medals can be awarded to the 8 runners.

Note that in all of these examples, the prizes, the certificates, and the medals were awarded **without replacement (repetition)**; that is, once we have chosen a winner of the first prize or the medal, that winner is not eligible for the other prizes. Thus, at each succeeding stage of the solution there is one fewer choice (8, then 7, then 6 in the Olympic example). Contrast this with the situation of a true-or-false test, where there are two possible answers (T or F) for each question on the test (the same answer can obviously be repeated).

Note also that **the order of selection was important** in each example. For the certificate example, being chosen first means that you receive substantially more money; in the Olympics example, coming in first means that you get the gold medal instead of the silver or bronze. In each case, if we had chosen the same three people in a different order, the outcome would have been different. Contrast this with the situation where we might draw three names out of a hat and each receives a $10 gift certificate. In this case the order of selection is not important since each of the three people receives the same prize.

Recall that we found the answer as follows when choosing the three certificate winners out of 25 guests in the charity example:

$$25 \times 24 \times 23$$

and as follows when choosing the three top runners out of eight in the Olympics example:

$$8 \times 7 \times 6$$

In other words, in each case, we multiple three numbers, beginning with the number of possible choices and going down by 1 each time. The number “three” corresponds with the number of people being chosen.

We can generalize the situation in the few previous examples to any problem **without replacement (repetition)** where the **order of selection is important**. If we are arranging r items in order out of n possibilities, the number of possible arrangements will be given by

$$n \times (n-1) \times (n-2) \times \ldots \times (n-r+1).$$

If you do not see why $(n-r+1)$ is the right number to use for the last factor, just think back to the certificate example, where we calculated $25 \times 24 \times 23$. In this case $n = 25$ and $r = 3$, so $n - r + 1 = 25 - 3 + 1 = 23$, which is exactly the right number for the final factor.

We can now present a special definition to this situation where we are choosing (arranging) r items out of n possibilities **without replacement** and where the **order is important**. Each of these arrangements is called a **permutation**. We write:

Permutations

$${}_nP_r = n \cdot (n-1) \cdot (n-2) \cdots (n-r+1)$$

We say that there are ${}_nP_r$ **permutations of** size r selected from among n choices without replacement when order matters.

It turns out that we can express this result more simply using factorials.

$${}_nP_r = \frac{n!}{(n-r)!}$$

The last formula may be a little confusing. Here is a way to make sense. You can write the number of ways in which eight Olympic runners can win the three medals (the number of permutations of size 3 selected from 8), as

$$8 \times 7 \times 6 = 336$$

as before, but note that I cannot use the factorial notation here because the numbers do not go down all the way to 1. However, notice this product is equivalent to

$$\frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1} = \frac{8!}{5!} = 8 \times 7 \times 6 = 336.$$

This is because all the numbers from 5 down appear on both the top and the bottom, cancelling each other. The reason that the denominator is 5! is that $8 - 3 = 5$. This example then illustrates a way to express the number of permutations with two factorials (in a fraction): ${}_nP_r = \frac{n!}{(n-r)!}$ as above.

For our purposes, however, *do not depend on these formulas* (unless the numbers are very large)! It is more important to be able to think and figure out ways to count accurately than to plug numbers into formulas.

For permutations problems, where r items are chosen without replacement from n items and the order matters, start with $n \times (n-1) \times \ldots$ until you have multiplied r numbers together. This is basically the Multiplication Rule.

Example 9

I have 9 paintings and have room to display only four of them at a time on my wall in a certain order. In how many different ways could I do this?

[Solution]

Since we are choosing 4 paintings out of 9 without replacement where the order of selection is important there are

$$_9P_4 = 9 \times 8 \times 7 \times 6 = 3{,}024 \text{ permutations.}$$

Example 10

In how many ways can a four-person executive committee (president, vice-president, secretary, treasurer) be selected from a 16-member board of directors of a non-profit organization?

[Solution]

We want to choose 4 people out of 16 without replacement (i.e., no one person holds two or more positions) and where the order of selection is important. So the answer is

$$_{16}P_4 = 16 \times 15 \times 14 \times 13 = 43{,}680 \text{ ways.}$$

Try It Now 6

How many 5-character passwords can be made using the letters A through Z

a. **If letters can be repeated?**
b. **If no repetition is allowed?**

Combinations

Above we considered the situation where we chose r items out of n possibilities without replacement and where the order of selection was important. We now consider a similar situation in which the order of selection is not important.

Example 11

A charity benefit is attended by 25 people at which three $50 gift certificates are given away as door prizes. Assuming no person receives more than one prize, in how many different ways can the gift certificates be awarded?

[Solution]

Using the Multiplication Rule, there are 25 choices for the first person, 24 remaining choices for the second person, and 23 for the third, so there would be $25 \times 24 \times 23 = 13{,}800$ ways to choose three people if the order mattered. Suppose for a moment that Abe is chosen first, Bea second, and Cindy third; this is one of the 13,800 possible outcomes. Another way to award the prizes would be to choose Abe first, Cindy second, and Bea third; this is another of the 13,800 possible outcomes. But either way Abe, Bea, and Cindy each get $50, so the order in which we select them really does not matter. In how many different orders can Abe, Bea, and Cindy be listed? It turns out there are 6.

ABC ACB BAC BCA CAB CBA

How can we be sure that we have counted them all? We are really just choosing 3 people out of 3, so there are $3 \times 2 \times 1 = 6$ ways to do this. We did not need to list them all; we can just use permutations.

So, out of the 13,800 ways to select 3 people out of 25, six of them involve Abe, Bea, and Cindy. The same argument works for any other group of three people (say Abe, Bea, and David or Frank, Gloria, and Haley) so each three-person group is counted *six times*. Thus the number 13,800 is six times too big. Therefore, we divide the number by 6 to get our answer:

$$13800 \div 6 = 2300$$

There are 2300 different ways in which the three identical certificates are given to 25 people. Note that this answer can also be expressed as follows:

$$\frac{{}_{25}P_3}{{}_3P_3}$$

We can generalize the situation in this example to any problem of choosing a collection of items **without replacement** where the **order of selection is not important**. If we are choosing r items out of n possibilities, the number of possible choices will be given by $\frac{{}_nP_r}{{}_rP_r}$, and we could use this formula for computation. However, this type or problems arises so frequently that we give a special definition: each group chosen is called a **combination**. We write:

Combinations

$_nC_r = \frac{_nP_r}{_rP_r}$

We say that there are $_nC_r$ **combinations** of size r selected from among n choices *without replacement* where *order does **not** matter.*

It turns out that we can also express this result in terms of factorials:

$$_nC_r = \frac{n!}{(n-r)!\,r!}$$

Note that this last formula is similar to the permutations formula. In fact, if you remove the r! part from the denominator, you will get the permutations formula. This is precisely because $_nC_r$ is $\frac{_nP_r}{r!}$. But once again, it is not important that you memorize and plug in numbers into this formula. It is more important to keep in mind the fact that the number of combinations is obtained from the number of permutations along with the process of correcting the overcount (duplicate count) since several permutations represent one combination (because the order does not matter with combinations). This "correction" is the division by $r!$, which is the number of ways in which you can arrange the r items chosen.

Example 12

A group of four students is to be chosen from a 35-member class to represent the class on the student council. How many ways can this be done?

[Solution]

We are choosing 4 people out of 35 without replacement. Note that these four students will not be holding different positions or ranks, so the order of selection is not important. So we look for the number of combinations. There are $_{35}C_4 = \frac{35\times34\times33\times32}{4\times3\times2\times1} = 52{,}360$ combinations.

This example may suggest another reason for *not* using the formula involving three factorials. Here, we were simply able to calculate the answer by multiplying 35, 34, 33, and 32 and then dividing the product by 4!. However, if you try to use the formula on your calculator, the numerator is 35!, which may appear very strange on your calculator: "1.0333…e+40." This is a way for your calculator to express very large numbers. In particular, 35! is 1.0333×10^{40}. Some calculators may not even be able to handle such a big number.[56] Remember why we use the exclamation mark to denote factorials? These numbers grow very quickly. (Of course, most

[56] Many cell phones are now equipped with highly capable scientific calculators, but some of them will display "error" when you try to calculate 104! or bigger while some other models will calculate 19,500! and more.

people do not use the factorial notation in daily living. Next time you see a coupon saying, "Save \$10!" do not assume that you can save $\$10! = \$3.26$ million.)

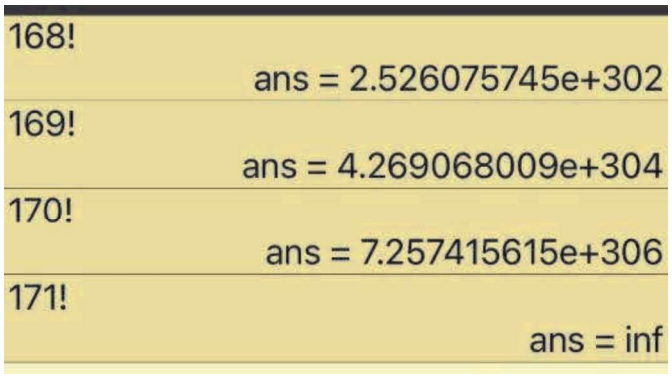


Figure 4.4: A screenshot of a calculator app on a cell phone (How far can yours go?)

Try It Now 7

Suppose a government committee consists of 13 members, and one of its subcommittees is to have 3 members. How many different 3-member subcommittees may be chosen from among the 13 committee members?

In the Try It Now problem above we assumed that the members of the subcommittee were chosen without regard to party affiliation. In reality, however, many choices come with additional constraints such as political affiliation. Let us consider a similar problem again, in a slightly more complicated form:

Example 13

Suppose a government committee consists of 29 members: 15 Republicans and 14 Democrats. A subcommittee is to have 19 members: 10 Republicans and 9 Democrats. How many different ways can the members of the subcommittee be chosen from among the 29 committee members?

[Solution]

In this case we need to choose 10 of the 15 Republicans and 9 of the 14 Democrats. There are ${}_{15}C_{10} = 3003$ ways to choose the 10 Republicans and ${}_{14}C_{9} = 2002$ ways to choose the 9 Democrats. But now what? How do we finish the problem?

Suppose we listed all of the possible 10-member Republican groups on 3003 slips of red paper and all of the possible 9-member Democratic groups on 2002 slips of blue paper. How many ways can we choose one red slip and one blue slip? This is a job for the Multiplication Rule! We are simply making one choice from the first category and one choice from the second category, just like in the restaurant menu problems from earlier.

Therefore, there must be $3003 \times 2002 = 6{,}012{,}006$ possible ways of selecting the members of the subcommittee.

Sometimes it may be difficult to determine whether or not repetition is allowed and whether or not the order is relevant. In some cases clarifying questions may be needed. For example, suppose that you and your friend are at an ice cream store which has 31 flavors and your friend asks you, "In how many ways can you choose 3 flavors for a triple cone?" You may then have to say, "Well, that depends. Do you want to specify the order (so that a vanilla-chocolate-strawberry triple cone is to be distinguished from a chocolate-strawberry-vanilla cone)? Can you order the same flavor more than once (so you can order a vanilla-vanilla-chocolate triple cone)?"

Here is an example with both permutations and combinations.

Example 14

Suppose there are seven audio books you are interested in listening to. In how many ways can you

(1) Listen to 2 out of the 7 audio books if the order is important?
(2) Listen to 5 out of the 7 audio books if the order is important?
(3) Listen to 2 out of the 7 audio books if the order is *not* important?
(4) Listen to 5 out of the 7 audio books if the order is *not* important?

[Solution]

You should immediately notice that (1) and (2) involve permutations (since the order is important) whereas (3) and (4) involve combinations. For (1), the answer is obtained either by the Multiplication Rule or by the permutations formula: $7 \times 6 = 42$. The answer is 42 (which sounds like the answer to the "ultimate question of life, the universe, and everything"). Similarly, for (2), the answer is $7 \times 6 \times 5 \times 4 \times 3 = 2520$. Now, for (3), the answer is $\frac{7 \times 6}{2 \times 1} = 21$. Note that this is much smaller than (1) or (2). But look at (4). The answer is $\frac{7 \times 6 \times 5 \times 4 \times 3}{5 \times 4 \times 3 \times 2 \times 1} = 21$ again. You should notice that $5 \times 4 \times 3$ appear on both the numerator and the denominator, cancelling each other.

Were you expecting a larger number because (2) was much bigger than (1)? Is it an accident that (3) and (4) have the same answer?

Think about it for a second. These last two answers are suggesting that there are as many ways to choose 2 out of 7 items as to choose 5 out of 7. (You can easily verify that, in a similar manner, there are as many ways to choose 3 out of 7 as to choose 4 out of 7.) You may, at this point, conjecture the following: **There are as many combinations of choosing r out of n as choosing $(n - r)$ out of n.**

Is this correct? Indeed it is! Consider it this way, for example. You have 7 friends you want to invite to a birthday party, but you can only choose 2 of them. There are, as seen above, 21 ways of doing so. However, each of those "combinations" is equivalent to choosing the 5 (out of 7) whom you are *not* inviting. Similarly, if you can choose only 3 out of 7, that means you are selecting the 4 (out of 7) whom you are not going to invite. In other words, choosing r out of n items is essentially equivalent to choosing $n - r$ out of n items. This may sound philosophical, but it is true indeed; **choosing whom to include is equivalent to choosing whom to exclude**. Mathematically speaking, we can now summarize it as follows. (This is the conjecture stated above.)

$$ {}_nC_{n-r} = {}_nC_r \qquad (0 \le r \le n) $$

Probability Using Permutations and Combinations

We can use permutations and combinations to solve more complex probability questions.

Example 15

A 4-digit PIN number is selected. What is the probability that there are no repeated digits?

[Solution]

There are 10 possible values for each digit of the PIN (namely: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9), so there are $10 \times 10 \times 10 \times 10 = 10^4 = 10{,}000$ total possible PIN numbers.

To have no repeated digits, all four digits would have to be different, i.e., selecting without replacement. The order matters, of course. We could either compute $10 \times 9 \times 8 \times 7$ or notice that this is the same as ${}_{10}P_4$ = 5040.

The probability of no repeated digits is the number of 4-digit PIN numbers with no repeated digits divided by the total number of 4-digit PIN numbers. This probability is

$$ 5040 \div 10{,}000 = 0.504 = 50.4\%. $$

Example 16

In a certain state's lottery, 48 balls numbered 1 through 48 are placed in a machine, and 6 of them are drawn at random. If the six numbers drawn match the numbers that a player had chosen, the player wins $1,000,000. In this lottery, the order in which the numbers are drawn does not matter. Find the probability that you win the million-dollar prize if you purchase a single lottery ticket.

[Solution]

In order to compute the probability, we need to count the total number of ways in which six numbers can be drawn and the number of ways in which the six numbers on the player's ticket could match the six numbers drawn from the machine. Since there is no stipulation that the numbers be in any particular order, the number of possible outcomes of the lottery drawing is

$$_{48}C_6 = \frac{48\times47\times46\times45\times44\times43}{6\times5\times4\times3\times2\times1} = 12{,}271{,}512.$$

Of these possible outcomes, only one combination would match all six numbers on the player's ticket, so the probability of winning the grand prize is

$$\frac{_6C_6}{_{48}C_6} = \frac{1}{12271512} \approx 0.0000000815$$ (Less than 1 in 12 million)

Example 17

In the state lottery from the previous example, if five of the six numbers drawn match the numbers that a player has chosen, the player wins a second prize of $1,000. Find the probability that you win the second prize if you purchase a single lottery ticket.

[Solution]

As above, the number of possible outcomes of the lottery drawing is $_{48}C_6$ = 12,271,512. In order to win the second prize, five of the six numbers on the ticket must match five of the six winning numbers; in other words, we must have chosen five of the six winning numbers and one of the 42 losing numbers. The number of ways to choose 5 out of the 6 winning numbers is given by $_6C_5$ = 6, and the number of ways to choose 1 out of the 42 losing numbers is given by $_{42}C_1$ = 42. Thus the number of favorable outcomes is then given by the Multiplication Rule:

$$_6C_5 \cdot {}_{42}C_1 = 6 \times 42 = 252.$$

So the probability of winning the second prize is

$$\frac{(_6C_5)(_{42}C_1)}{_{48}C_6} = \frac{252}{12271512} \approx 0.0000205$$ (About 205 in 10 million).

Try It Now 8

A multiple-choice question on an economics quiz contains 10 questions with five possible answers each. Compute the probability of randomly guessing the answers and getting exactly 9 questions correct.

Example 18

Find the probability of getting exactly one ace when randomly drawing five cards from a deck.

[Solution]

We use combinations to compute the possible number of 5-card hands, ${}_{52}C_5$. This number will go in the denominator of our probability formula since it is the number of possible outcomes.

$$ {}_{52}C_5 = \frac{52\times51\times50\times49\times48}{5\times4\times3\times2\times1} = 2{,}598{,}960. $$

For the numerator, we need the number of ways to draw exactly one ace and four other cards (none of them aces) from the deck. Since there are four aces and we want exactly one of them, there will be ${}_4C_1$ ways to select one ace. Since there are 48 non-aces and we want 4 of those, there are ${}_{48}C_4$ ways to select the four non-aces. Now we use the Multiplication Rule to calculate that there will be ${}_4C_1 \times {}_{48}C_4$ ways to choose one ace and four non-aces.

$$ {}_4C_1 \times {}_{48}C_4 = \frac{4}{1} \times \frac{48\times47\times46\times45}{4\times3\times2\times1} = 778{,}320. $$

Putting this all together, we have

$$ P(\text{one ace}) = \frac{778{,}320}{2{,}598{,}960} \approx 0.229. $$

We see that the probability of getting exactly one ace is just about 30%.

Example 19

Find the probability of getting exactly two aces when randomly drawing five cards from a deck.

[Solution]

The solution is similar to the previous example, except now we are choosing 2 aces out of 4 and 3 non-aces out of 48. The denominator remains the same:

$$ {}_4C_2 \times {}_{48}C_3 = \frac{4\times3}{2\times1} \times \frac{48\times47\times46}{3\times2\times1} = 103{,}776. $$

$$ P(\text{two aces}) = \frac{103{,}776}{2{,}598{,}960} \approx 0.0399. $$

The probability of getting exactly two aces is about 4%.

It is useful to note that these card problems are remarkably similar to the lottery problems discussed earlier.

Try It Now 9

Find the probability of getting three aces and two kings when randomly drawing five cards from a deck.

Birthday Problem

We now take a pause to consider a famous problem in probability theory: the "Birthday Problem."

> Suppose you have a room full of 30 people. What is the probability that there is at least one shared birthday, i.e., at least two people having the same birthday (month and day only)?

Take a guess at the answer. Was your guess fairly low, like around 10%? That seems to be the intuitive answer (30/365, perhaps?). Let's see if we should listen to our intuition. This appears to be a tough problem. Let's start with a simpler version instead.

Example 20

> Suppose three people are in a room. What is the probability that there is at least one shared birthday among these three people?
>
> [Solution]
>
> There are a lot of ways in which there is at least one shared birthday. For instance, if A, B, and C are the three people in the room, A and B could have the same birthday, and that birthday could be any one of the 365 days of the year. The same can be said of A and C as well as of B and C. But then what if they all have the same birthday? You can see that this gets complicated very quickly.
>
> Fortunately there is an easier way. We ask ourselves, "What is the *alternative* to having at least one shared birthday?" In asking this question, we are finding the probability of the **complementary event**. The *alternative* is that there are **no** shared birthdays; the negation of "**at least one**" is having **none.**
>
> $$\mathrm{P}(\text{at least one}) = 1 - P(\text{none}).$$
>
> We will start, then, by finding the probability that there is no shared birthday. Let's imagine that you are one of these three people. Your birthday can be anything without conflict, so there are 365 choices out of 365 for your birthday. What is the probability

that the second person does *not* share your birthday? There are 365 days in the year (we ignore leap years) and removing your birthday from contention, there are 364 choices that will guarantee that you do not share a birthday with this person, so the probability that the second person does not share your birthday is 364/365. Now we move to the third person. What is the probability that this third person does not share the birthday with either you or the second person? There are 363 days that will not duplicate your birthday or the second person's, so the probability that the third person does not share a birthday with the first two is 363/365.

We want the second person not to share a birthday with you *and* the third person not to share a birthday with the first two people, so we use the Multiplication Rule:

$$P\ (\text{no shared birthday}) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \approx 0.9918.$$

Then subtract this from 1 to get

$$P\ (\text{shared birthday}) = 1 - P(\text{no shared birthday}) \approx 1 - \ 0.9918 = 0.0082.$$

This is a pretty small number (less than 1%), so maybe it makes sense that the answer to our original problem will be small. Let's make our group a bit bigger.

Example 21

Suppose five people are in a room. What is the probability that there is at least one shared birthday among these five people?

[Solution]

Continuing the pattern of the previous example, the answer should be

$$P\ (\text{shared birthday}) = 1 - \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \frac{362}{365} \times \frac{361}{365} \approx 0.0271.$$

Note that we could rewrite this more compactly as

$$P(\text{shared birthday}) = 1 - \frac{{}_{365}P_5}{365^5} \approx 0.0271$$

which makes it a bit easier to type into a calculator or computer, and this suggests a nice formula as we continue to expand the population of our group.

Example 22 (the Original "Birthday Problem")

Suppose 30 people are in a room. What is the probability that there is at least one shared birthday among these 30 people?
[Solution]

Here we can calculate

$$P(\text{shared birthday}) = 1 - \frac{{}_{365}P_{30}}{365^{30}} \approx 0.706$$

This gives us the surprising result that when there are 30 people in a room, there is more than a 70% chance that at least two people share the same birthday!

This is one of many results in probability theory that is counter-intuitive; that is, it goes against our gut instincts. If you still do not believe the mathematics, you can carry out a simulation. Just so you do not have to go around rounding up groups of 30 people, someone has kindly developed an online applet so that you can conduct a computer simulation. Go to https://mste.illinois.edu/activity/birthday/.

Set the class size to 30. If you keep track of the number of times that there is a repeated birthday, you should get at least one duplicate birthday about 7 out of every 10 times you run the simulation. The applet also keeps track of this experimental probability.

Try It Now 10

Suppose 10 people are in a room. What is the probability that there is at least one shared birthday among these 10 people?

Exercises 4.3

1. One of the challenges in this section is to identify when the order is important and when repetition is allowed. For each situation below, determine if the order is important and if repetition is allowed. Explain why.
   (a) Creating a 4-digit PIN for your debit card
   (b) Choosing 4 people out of 10 candidates to serve on a town committee
   (c) Making a playlist of 4 different songs out of 10 you have downloaded
   (d) Selecting a 4-digit lock combination where the digit "0" cannot be used.

2. For each situation below, determine if the order is important and if repetition is allowed. Explain why.
   (a) Choosing 3 out of 6 children to go on a bus
   (b) Assigning 7 seats to a small class of 7 students
   (c) Picking 2 out of 5 books at the store to give to a friend
   (d) Making a license plate where the first three characters are alphabet letters and the last three are digits
   (e) Coming up with a 9-hitter starting lineup for a baseball team
   (f) Selecting a 7-digit phone number

3. A boy owns 8 pairs of pants, 3 shirts, 2 ties, and 2 jackets. How many different outfits can he wear to school if he must wear one of each item?

4. At a restaurant you can choose from 3 appetizers, 5 entrees, and 2 desserts. How many different three-course meals can you have?

5. How many three-letter codes can be made from 4 letters *A, B, C,* and *D* if repetition of letters is allowed? What if repetition is not allowed?

6. How many four-letter codes can be made from 6 letters *A, B, C, D, E,* and *F* if repetition of letters is allowed? What if repetition is not allowed?

7. All of the license plates in a particular state feature three letters followed by three digits (e.g. ABC 123). How many different license plate numbers are available to the state's Department of Motor Vehicles?

8. A password must be exactly eight characters long. How many passwords are possible if only the 26 letters of the alphabet are allowed?

9. A pianist plans to play 4 pieces at a recital. In how many ways can she arrange these pieces in the program?

10. In how many ways can first, second, and third prizes be awarded in a contest with 210 contestants?

11. Seven Olympic sprinters are eligible to compete in the 4 x 100 m relay race for the U.S. Olympic team. How many four-person relay teams can be selected from among the seven athletes?

12. A computer user has downloaded 25 songs and wants to create a playlist with ten songs. If the order of the songs is important to her, how many different playlists could be made from the 25 songs available to her?

13. In how many ways can 4 pizza toppings be chosen from 12 available toppings? (Assume they must be 4 different toppings.)

14. At a baby shower 17 guests are in attendance and 5 of them are randomly selected to receive a door prize. If all 5 prizes are identical, in how many ways can the prizes be awarded?

15. In the 6/50 lottery game, a player picks six numbers from 1 to 50. How many different choices does the player have if order does not matter?

16. In a daily lottery game, a player picks three different numbers from 0 to 9. How many different choices does the player have if order does not matter?

17. A jury pool consists of 27 people. In how many different ways can 11 people be chosen to serve on a jury and one additional person be chosen to serve as the jury foreman?

18. Suppose that the United States Senate Committee on Commerce, Science, and Transportation consists of 23 members: 12 Republicans and 11 Democrats. Assume also that the Surface Transportation and Merchant Marine Subcommittee consists of 8 Republicans and 7 Democrats. In how many ways can members of the subcommittee be chosen from the committee?

19. You are at a pizza joint that feature 15 toppings. You are interested in buying a 2-topping pizza. How many choices for the 2 toppings do you have in each situation below?
    (a) They must be two different toppings, and you must specify the order.
    (b) They must be two different toppings, but the order of those two is not important. (After all, a pizza with ham and extra cheese is the same as one with extra cheese and ham.)
    (c) The two toppings can be the same (they will just give you twice as much), and you must specify the order.
    (d) The two toppings can be the same, and the order is irrelevant.

20. You own 16 CDs. You want to randomly arrange 5 of them in a CD rack. What is the probability that the rack ends up in alphabetical order?

21. A jury pool consists of 27 people: 14 men and 13 women. Find the probability that a randomly selected jury of 12 people is all male.

22. In a lottery game, a player picks six numbers from 1 to 48. If 5 of the 6 numbers match those drawn, they player wins second prize. What is the probability of winning this prize?

23. In a lottery game, a player picks six numbers from 1 to 48. If 4 of the 6 numbers match those drawn, they player wins third prize. What is the probability of winning this prize?

24. Find the probability that a 5-card poker hand is dealt to you that contains all hearts.

25. Find the probability that a 5-card poker hand is dealt to you that contains four aces.

Figure 4.5[57]

[57] Image by Poker Photos is licensed under CC BY 2.0

4.4 The Law of Averages and Expected Value

Expected value is perhaps the most useful probability concept we will discuss. It has many applications, from insurance policies, corporate strategies, all the way down to our everyday decision-making. But before we can truly grasp this important concept, we need to understand the Law of Averages; this is the bridge between the mathematical theory of probability and events in the real world:

The Law of Averages

The Law of Averages (The Law of Large Numbers)

When an independent trial is repeated many times, the actual probability that some event occurs approaches its theoretical (mathematical) probability.

Here is a practical way to illustrate this principle. Suppose the "trial" referred to in this statement is a simple coin toss. Assuming that the coin is fair, the theoretical/mathematical probability of the coin landing heads up is ½, or 50%. So what happens if you toss the coin 10 times? You may get 5 heads, but you may also get 7 heads, which is 70% of the 10 tosses. Now, you would not be too surprised because it is only 7 out of 10 times. Some would say that the "sample size is too small." And that is absolutely right. However, suppose you toss this coin 1000 times and got 700 heads. Now, that would be surprising. In fact, at that point, you would probably suspect that the coin is not fair.

Here is another example. Imagine an NBA player, known to make about 90% of free throws. If, in a particular game, he misses the first 2 out of 3 free throws, you may be disappointed but not shocked. "It happens" sometimes. However, if the same player misses 20 out of 30, then everyone would be concerned because it is *very unlikely* that he misses 2/3 of free throws as the number gets larger and larger.

This is the Law of Averages, or the Law of Large Numbers. In fact, using the theory of probability (what is known as the **binomial probability distribution**), one can show that the probability of getting 6 or more heads out of 10 coin tosses is about 38% (not all that unusual) whereas the probability of getting 60 or more heads out of 100 coin tosses is only about 2.8% (very rare).

So the point is this: *the more times you repeat something, the probability that some event happens gets closer and closer to the probability you calculate mathematically*.

Example 1

Suppose you are at a roulette table, betting on even numbers. You have noticed that the last 5 spins have all resulted on odd numbers. Which of the following is correct?

A. “Bet on even. Things have to balance out. Eventually you will have to get an even number.”
B. “Bet on odd. Clearly, there is a streak going on.”
C. “Odd or even, it doesn’t matter. The chances of you winning are the same as before.”

[Solution]

The fundamental thing to remember here is the **independence** of events (here, the spins). The roulette does not have memory to remember history. Both A and B may sound convincing, but these are examples of what is known as **the Gambler’s Fallacy**. Although A claims (correctly) that things will balance out, that is true only when considering larger and larger numbers of spins. It is not so unusual to get 5 odd numbers in a row, but it is extremely unlikely to get 500 odd numbers in a row. But that fact still does not affect the probability for the next spin. The “streak” which B speaks of is also a fact, but that again does not, in any way, influence the next spin. C is correct.

Expected Value

The **Expected value** of a game (or an event) is the average gain (or loss) one can expect to make every time the game is played (or the event is held), based on the **Law of Averages**. A positive expected value means that it is to your advantage to play the game (or to hold the event). If it is negative, that means that it is more likely that you end up losing money. There is a formula to calculate the expected value (see below), but a better way to approach the problem, from the perspective of problem solving, is to consider playing the game (or holding the event) hundreds of times and keeping track of how much you will and lose. Keep in mind that the expected value does *not* tell you how much you will win or lose on your next game or wager, or not even the next ten. It tells you, on the average, after playing the game or betting *many, many times*, how much you expect to gain or lose each time.

Let us look at some examples.

Example 2

Suppose you play a game in which a fair die is rolled; you win $7 if the outcome is a 6 but you lose $2 otherwise. What is the expected value for this game, i.e., how much, on the average, do you expect to win (or lose) each time you play this game?

[Solution]

Instead of thinking of playing this game once and wonder what would happen, consider playing this game 600 times (or 6000 times) and see what you can expect. The probability of getting a 6 on the die is 1/6, so out of the 600 times you play, you expect to win 100 times and lose 500 times. For each win, you gain $7, so your winnings will

be $700. On the other hand, each time you lose, you pay $2, so you will lose $1000. So, after playing it 600 times, your net gain is $\$700 - \$1000 = -\$300$, which is a loss of $300. So the average per game is a loss of $\$300 \div 600 = 0.50$; the expected value is -50 cents, or a 50-cent loss.

Now, this case I chose 600 times because the probability of winning each game was 1/6. In order to avoid fractions and messy calculations, you can choose whatever number you want; here, I wanted a large number that is a nice multiple of 6.

Example 3

In the casino game roulette, a wheel with 38 spaces (18 red, 18 black, and 2 green) is spun. In one possible bet, you bet $1 on a single number. If that number is spun on the wheel, then you receive $36 (their original $1 + $35). Otherwise, you lose your dollar. On average, how much money should you expect to win or lose if you play this game repeatedly?

[Solution]

This time, the probability that you win a given game is 1/38. So suppose you play this game 380 times (38 games would work, too). Out of 380 times, you expect to win 10 times. Each time you win, you receive $35, so your total winnings will be $350. Note that it may appear that you are winning $36 each time you win, but one of those dollars is your bet, so you are simply getting it back, not winning it. On the other hand, you expect to lose $1 each of the other 370 times. That is a loss of $370. The net gain is then $\$350 - \$370 = -\$20$, a loss of $20. To calculate the expected value, you divide this by 380, and the expected value is $-20 \div 380 = -0.05263\ldots$, which is about a loss of 5.3 cents per game.

Notice that no one ever loses exactly 5.3 cents; most people (in fact, about 37 out of every 38) lose $1 and a very few people (about 1 person out of every 38) gain $35 (the $36 they win minus the $1 they spent to play the game).

There is another way to compute expected value without imagining what would happen if we play every possible space. There are 38 possible outcomes when the wheel spins, so the probability of winning is $\frac{1}{38}$. The complement, the probability of losing, is $\frac{37}{38}$. Summarizing these along with the values, we get this table:

Outcome	Probability of Outcome
$35	$\frac{1}{38}$
-$1	$\frac{37}{38}$

Notice that if we multiply each outcome by its corresponding probability and *add them together*, we get

$$35 \times \frac{1}{38} + (-1) \times \frac{37}{38} \approx 0.9211 - 0.9737 \approx -0.0526.$$

This is the expected value we computed above. Here, we multiplied each outcome by its probability before adding the products. It is an example of what is known as a **weighted sum** (the probabilities are the "weights").

Expected Value

Expected Value is the average gain or loss of an event if the procedure is repeated many times.

We can compute the expected value by multiplying each outcome by the probability of that outcome and then adding up the products (a weighted sum).

The official definition uses this weighted sum approach; however, in general, it is far easier to calculate expected values (and makes more sense) if you follow the scenario of playing the game many times, as described in the examples above.

Try It Now 11

You purchase a raffle ticket to help out a charity. The raffle ticket costs $5. The charity is selling 2000 tickets. One of them will be drawn and the person holding the ticket will be given a prize worth $4000. Compute the expected value for this raffle.

Example 4

In a certain state's lottery, 48 balls numbered 1 through 48 are placed in a machine and 6 of them are drawn at random. If the six numbers drawn match the numbers that you had chosen, you win $1,000,000. If you match 5 numbers, then you win $1,000. It costs $1 to buy a ticket. Find the expected value.

[Solution]

Earlier, we calculated the number of all possible ways the 6 numbers can be picked out of 48, that is, the number of combinations of size 6 taken from 48 items:

$$_{48}C_6 = 12{,}271{,}512.$$

So how many tickets should we consider "buying" (theoretically) this time? In previous examples, we always picked a large number that is a multiple of the total number of outcomes. Here, the probability of winning the million-dollar prize is 1 out of this number, larger than 12 million. Why not, then, do we pretend to buy this many? In

other words, let us suppose we buy 12,271,512 lottery tickets, each time choosing a different combination (thereby picking every possible combination). We will definitely, certainly, undoubtedly *win* the grand prize! But at what cost? Let us calculate the expected value.

Out of the 12,271,512 tickets you purchase, exactly one of them will win $1,000,000. Earlier, we also calculated how many combinations will win the $1000 prize (which you will win by choosing 5 of the 6 numbers chosen and 1 out of the 42 not chosen randomly):

$$ {}_6C_5 \times {}_{42}C_1 = 252. $$

Each of these 252 combinations will win $1000, so our total winnings will be

$$ 1{,}000{,}000 + (252 \times 1000) = 1{,}252{,}000. $$

Congratulations! You have just won $1,252,000. This sounds great, except that we purchased each ticket for $1, so we spent $12,271,512. The difference indicates our net loss:

$$ 1{,}252{,}000 - 12{,}271{,}512 = -11{,}019{,}512. $$

We have lost more than $11 million. Dividing this number by the number of tickets purchased, we get the expected value per ticket:

$$ -11{,}019{,}512 \div 12{,}271{,}512 \approx -0.898. $$

The expected value is about -$0.898. In other words, for each ticket we buy for a dollar, we should expect to lose about 90 cents.

Of course, it is possible to use the method of weighted sum (probabilities and outcomes) to come up with the same answer. However, this will get messy as you see below. Remember, when you win the million-dollar prize, you still spent one dollar, so the "outcome" then would be $999,999. Similarly, if you win a $1000 prize, your outcome will be $999. With these values, you could calculate the expected value as follows:

$\frac{{}_6C_6}{{}_{48}C_6} = \frac{1}{12271512} \approx 0.0000000815$ for the probability of picking all 6 numbers.

$\frac{({}_6C_5)({}_{42}C_1)}{{}_{48}C_6} = \frac{252}{12271512} \approx 0.0000205$ for picking 5 numbers and missing one.

Our probabilities and outcome values are:

Outcome	Probability of outcome
\$999,999	$\frac{1}{12271512}$
\$999	$\frac{252}{12271512}$
$-\$1$	$1-\frac{253}{12271512}=\frac{12271259}{12271512}$

The expected value, then is:

$$(\$999{,}999)\cdot\frac{1}{12271512}+(\$999)\cdot\frac{252}{12271512}+(-\$1)\cdot\frac{12271259}{12271512}\approx-\$0.898$$

Again, on average, one can expect to lose about 90 cents on a lottery ticket. Of course, most players will lose \$1.

In general, if the expected value of a game is negative, it is not a good idea to play the game since on average you will lose money. It would be better to play a game with a positive expected value (good luck trying to find one) although keep in mind that even if the average winnings are positive, it could be the case that most people lose money and one very fortunate individual wins a great deal of money. If the expected value of a game is 0, we call it a fair game since neither side has an advantage.

Not surprisingly, the expected value for casino games is **always negative for the player** and positive for the casino. It must be positive, or they would go out of business. (There is a reason that those gigantic casinos in Las Vegas stay in business.) It is often said, correctly, that the "only way to make money at a casino is to own one." Players just need to keep in mind that when they play a game repeatedly, they should expect to lose money because their expected value is negative. That is fine so long as they enjoy playing the game and think it is worth the cost. But it would be wrong to expect to come out ahead.

Try It Now 12

A friend offers to play a game in which you roll 3 standard dice. If all the dice roll different values, you give him \$1. If any two dice match values, you get \$2. What is the expected value of this game? Would you play?

Expected value also has applications outside of gambling. Expected value is very common in making insurance decisions.

Example 5

A 40-year-old man in the U.S. has a 0.242% risk of dying during the next year. An insurance company charges $275 for a life-insurance policy that pays a $100,000 death benefit. What is the expected value for the person buying the insurance?

[Solution]

The probability given is 0.242%, which represents 0.242 men out of 100, or 242 men out of 100,000. In order to avoid using decimals, let us assume 100,000 men purchase this policy. Out of those, 242 men will (unfortunately) die, receiving $100,000 each as the promised benefit. Their families will be collecting a total of $\$100{,}000 \times 242 = \$24{,}200{,}000$ altogether. While the insurance company must pay out this amount, over $24 million, they also got $275 from each of the 100,000 men, the revenue totaling $27,500,000. The difference

$$24{,}200{,}000 - 27{,}500{,}000 = -3{,}300{,}000$$

is the total cash flow from the group of 100,000 men who purchased this policy to the insurance company. In other words, this group of 100,000 men "lost" $3.3 million altogether. Dividing this amount by 100,000, we get $-\$33$. The expected value of each policy purchased is $-\$33$, meaning that, on the average, each of those men lost $33. Said differently, each policy sold made the insurance company $33 more.
If you prefer to use the weighted sum method, here is how:

Outcome	Probability of outcome
$100,000-$275=$99,725	0.00242
$275	1-0.00242=0.99758

The expected value is

$$99{,}725 \times 0.00242 + (-275) \times 0.99758 = -33.$$

Not surprisingly, the expected value is negative; the insurance company can only afford to offer policies if they, on average, make money on each policy. They can afford to pay out the occasional benefit because they offer enough policies that those benefit payouts are balanced by the rest of the insured people.

For people buying the insurance, there is a negative expected value, but there is a security that comes from insurance that is worth that cost.

Example 6

Here is one additional application of expected value—in sports. In baseball, a hitter can collect a base hit, which can be a single (1 TB), a double (2 TB), a triple (3 TB), or a home run (4 TB), where "TB" stands for "total bases." Imagine a (good) hitter with 100 at-bats with 20 singles, 5 doubles, 2 triples, and 5 home runs. What is the expected value (for this hitter) of the number of total bases gained per at-bat?

[Solution]

This player's total bases (TB) gained would be

$$(20 \times 1) + (5 \times 2) + (2 \times 3) + (5 \times 4) = 56.$$

He has earned 56 total bases out of 100 at-bats, so his "expected value for TB per at-bat" will be 0.560. In baseball statistics, this number is actually quite important; it is called the "slugging percentage" (SLG). The career leaders in SLG in Major League Baseball are Babe Ruth (.6897), Ted Williams (.6338), and Lou Gehrig (.6324).

Odds, Risks, and Probability[ix]

The **odds** of an event ("odds," always plural) occurring are the probability (e.g**. risk**) that this event will occur divided by the **probability** that the event will not occur.

$$\text{Odds of event} = \frac{P}{1-P}$$

This probability measure is popular in the world of gambling. If we compute the number of people putting money on one horse winning and the number of people putting money on the horse not winning (i.e. putting money on other horses) we can compute the odds of winning. For example, suppose that among 3100 persons gambling on horses, 100 persons put money on Horse A to win and 3000 do not (they bet on other horses). The odds of winning are then 1/30:

$$\text{Odds} = \frac{100/3100}{3000/3100} = \frac{100}{3000} = \frac{1}{30}.$$

In fact in gambling the *odds of not winning* are preferred and expressed as a ratio $X/1$. In our example, the odds are 30/1, or in words "thirty to one." This means that for every dollar that you bet, you will receive 30 if you win.

While odds are not the same as expected value, we often hear both of these terms as they relate to chance and probability. One way to remember the difference is this: odds are a ratio while expected value is a weighted sum as explained above. Odds are very similar to probability, but be aware of the difference. For instance, if you only have 10% chance of

winning a game, the odds against you are 9 to 1 while the probability of your not winning would be 9/10 (which, of course, is 90%).

Whether expressed as odds or probability, the concept of expected value is practically affecting every decision we make, every single day. You decide whether to wear a short-sleeve shirt or a heavy sweater based on the temperature outside. Basically you are asking, "What if I wear my shirt and get too cold?" or "How likely is it that it will get too hot if I wear this sweater?" You would probably drive faster if the negative consequences (speeding ticket, accident) are unlikely whereas you would drive at the speed limit if there is a police car right next to you and the probability of being caught is high. College education costs money, but hopefully the probability of getting gainfully employed in the future is much higher if you have a degree. These decisions are all based on some forms of expected value. We are constantly weighing the pros and cons along with their probabilities; that is exactly what expected value is.

Exercises 4.4

1. A bag contains 3 gold marbles, 6 silver marbles, and 28 black marbles. Someone offers to play the following game. You randomly select one marble from the bag. If it is gold, you win $3. If it is silver, you win $2. If it is black, you lose $1. What is your expected value of this game?

2. A friend devises a game that is played by rolling a single six-sided die once. If you roll a 6, he pays you $3; if you roll a 5, he pays you nothing; if you roll a number less than 5, you pay him $1. Compute the expected value for this game. Should you play this game?

3. In a lottery game, a player picks 6 numbers from 1 to 23. If the player matches all 6 numbers, the player wins $30,000 (and loses $1 for playing the game). Otherwise, the player loses $1. Find the expected value of this game.

4. A game is played by picking two cards from a deck. If they are the same value, then you win $5; otherwise you lose $1. What is the expected value of this game?

5. A company estimates that 0.7% of their products will fail after the original warranty period but within 2 years of the purchase, with a replacement cost of $350. If the company offers a 2-year extended warranty for $48, what is the company's expected value of each warranty sold?

6. An insurance company estimates the probability of an earthquake in the next year to be 0.13%. The average estimated damage caused by an earthquake is $60,000. If the company offers earthquake insurance for $100, what is their expected value of the policy?

7. Three people put their names in a hat, then each draws a name as part of a randomized gift exchange. What is the probability that no one draws their own name? What about with four people? What about with five? (This is a classic mathematics problem known as **the Hat-Check Problem**. It is fascinating that the solution is related to our old friend e, the irrational number that appeared in the continuous compounding formula).

8. *The Monty Hall problem* is named for the host of the game show *Let's make a Deal*. In this game, there are three doors, behind one of which there was a prize. The contestant was asked to choose one of the doors. Monty Hall would then open one of the other doors to show there was no prize there. The contestant was then asked if he/she wanted to stay with the original door or switch to the other unopened door. Is it better to stay or switch, or does it matter? (You may want to do a quick research on this famous problem.)

9. When three coins are tossed, the probability of getting three tails is 1/8. Suppose you get $6 if you get three tails and lose $2 otherwise. Calculate the expected value.

10. A die is tossed in a game. Suppose you win $10 if you get a 6 and lose $2 otherwise. Calculate the expected value of this game.

11. You agree to play this game in which the dealer rolls a die. You lose $2 if you get a 1, 2, or 3. You lose $5 if you get a 4 or 5. But you win $20 if you get a 6. Would you play the game? Answer this by computing the expected value per game.

12. Repeat the previous problem, except now that you have to pay $1 to play the game. Would you keep playing?

13. A 50-50 raffle is a way to raise money (often used by non-profit organizations). Tickets are sold at an event for a certain price, and the total proceeds are divided evenly (50-50) between one lucky winner and the host organization. Suppose a ticket costs $1, and 700 tickets were sold. Find the expected value per ticket purchased. In other words, for each ticket you purchase, how much do you expect to lose/donate?

14. Repeat the previous problem, except now each ticket costs $2, and 1000 tickets are sold.

15. Here is an interesting game presented and discussed by psychologists, mathematicians, philosophers, and people in business, risk management, and gambling. You start with $1, and a die is rolled. You "win" if the outcome is *not* a 6, in which case you double your amount. You "lose" if the outcome is a 6, in which you lose everything you have accumulated. At any point of this game, you can

continue or quit. Once you lose your money, there is no point in continuing since twice 0 is still 0.

For example, on the first roll, say a 3 comes up. Now you have \$2. You decide to continue, and the second outcome is a 5. Now you have \$4. You decide to continue, and the third outcome is a 4. Now you will have \$8. You decide to continue, and the four outcome is a 3. You will now have \$16. You decide to continue, and the fifth outcome is a 6. You lose everything (your \$16) and go down to \$0.

(a) If no 6 comes up the first 5 rolls, how much would you have?
(b) Suppose no 6 comes up the first 10 rolls. How much would you have?
(c) If the first 6 comes up on the n^{th} roll, how much would you lose then?
(d) There is about a 93.5% chance that at least one 6 comes up during the first 15 rolls. If you have played this game 15 times, therefore, you have about a 93.5% chance that you have nothing, but you still have a 6.5% chance ($\left(\frac{5}{6}\right)^{15} \approx 0.0649$) that you have money. How much would you have if no 6 comes up 15 times in a row?
(e) After playing this game n times, how much do you expect to have?
(f) When would you stop playing this game? Put yourself in this situation and answer according to your personality. Then, find the "correct" answer using mathematics.
(g) Instead of sequentially rolling one die n times, imagine rolling n dice simultaneously, once. Would any of your answers change? Why or why not?

16. Comment on how each of the following everyday experiences involves the idea of expected value. Can you think of more? (There are many, many more.)
    (a) Many drivers regularly exceed the speed limit. However, this practice is not so blatant if the driver sees a police car nearby.
    (b) When traveling to Seattle, many people take an umbrella even though it is inconvenient to do so.
    (c) Most people do not mind spending a lot of money to have their car brakes fixed.
    (d) When booking an expensive travel ticket, some people buy an insurance policy so that the money is fully refunded if they cannot go on the trip.
    (e) People suddenly become interested in getting a life insurance policy when they become parents.
    (f) If you are in a big hurry to park your car in downtown, you may just park for a few minutes with an expired parking meter, but you dare not park in a space marked for disabled person parking.
    (g) Many shoppers choose to buy products at a big warehouse-type store with a generous return policy even if the prices are similar at other stores.

For Further Explorations…

Risk and Irrational Fears

Guido Calabresi is a senior federal judge and a former dean and professor of the Yale Law School. He is famous for what has come to be known as "Calabresi's Fable" on irrational fears and risk-related behaviors.[58] His story goes like this:

> *Suppose someone comes to tell you about an exciting, innovative, and life-changing invention that saves you a huge amount of time every day, allows you to go to many places you have never been to before, and enables you to enjoy being with your loved ones and having a much higher quality of life overall. The only catch is that about 40,000 Americans will unfortunately die every year because of this new product. Would you support this invention?*

Most people will say, "Of course not. Life is too precious to sacrifice for convenience." Then Calabresi would ask, "What is the difference between this and the automobile?" His point is that we fear so many things and worry about relatively low risks while we have no problem driving every day; many of our fears and concerns, he says, are irrational.

Indeed, there are about 30,000 to 50,000 car-related fatalities every year in America, but we do not have people trying to ban the use of automobiles. We accept the cost of traffic fatalities because it is just another aspect of our normal, daily life now. In a world without cars, we might be able to save tens of thousands of lives, but it is a society almost impossible for us to imagine.

This fable has been "resurrected" widely in recent times due to the COVID-19 pandemic, along with fears of vaccine side effects such as blood clot cases. You can read more about Calabresi's Fable online and consider its far-reaching implications. In light of this fable, what do you think of the following data and figures? Discuss these with your classmates, friends, and others.

- COVID-19 vaccines were rolled out in early 2021, but vaccinated people could still get the disease. Out of the first 87 million people fully vaccinated in the U.S., it was reported[59] about 7157 of them still caught COVID-19. Out of those, 88 died (though 13% of these deaths were not due to COVID-19). Should vaccination be stopped?

[58] A version of this fable was contained in an article by David Leonhardt in the morning newsletter of *The New York Times* on April 19, 2021: https://www.nytimes.com/2021/04/19/briefing/european-soccer-league-mars-helicopter-navalny.html.

[59] Dolan, M. (May 6, 2021). Scientists scramble to see why, in rare cases, even the vaccinated can get COVID-19. *The Los Angeles Times*. https://www.latimes.com/california/story/2021-05-03/in-rare-cases-even-the-vaccinated-can-get-covid-19-and-die. Similar numbers also appeared in a report by CNN on April 15, 2021. https://www.cnn.com/2021/04/14/health/breakthrough-infections-covid-vaccines-cdc/index.html

- Airplane crashes are always tragic. In 2020, large commercial plane crashes killed 299 people worldwide.[60] Should airplane travels be banned? (Someone jokingly suggested that most plane crashes can be avoided if the planes just travel slowly on the road and not take off.)

- The COVID-19 vaccine by Johnson & Johnson had been given to 7 million Americans in the spring of 2021 when it was discovered that six people (all women) suffered blood clots. The U.S. and some European governments paused the use of this vaccine out of an abundance of caution. Do you support or oppose this action?

- About 5,000 motorcyclists die every year in America, making motorcyclists 27 times more likely to die than passengers in cars (per vehicle miles traveled).[61] With such a high risk, why are motorcycles permitted on the road? Should anything be done?

- What other risks (perhaps high risks) do we accept in general besides driving vehicles? Think of commonplace items around your house or town, such as backyard pools, various sports, food, outdoor activities, etc.

Figure 4.6[62]

[60] Reuters. (2021). Plane crash deaths rose in 2020 despite pandemic. *The Guardian*. https://www.theguardian.com/world/2021/jan/02/plane-crash-deaths-rose-in-2020-despite-pandemic

[61] Facts + Statistics (n.d.). *Insurance Information Institute*. https://www.iii.org/fact-statistic/facts-statistics-motorcycle-crashes.

[62] Image by AngMoKio is licensed under CC BY-SA 3.0

Chapter 5: Data and Statistics

"Statistics are used much like a drunk uses a lamp post: for support, not illumination."—Vin Scully

We are experiencing unprecedented times in terms of data, information, and current events. In recent years, terms like "fake news," "alternative facts," and "conspiracy theories" have emerged, making it hard to discern what is true and what is not. As a matter of fact, one of the main goals of a course like this is to help you acquire tools to determine what is true and to detect fallacies and deception, which are so prevalent in the society in which we live.

Are facts relative? Do they depend on what media or news channels you watch? Are you relying solely on the information you get "fed" based on algorithms created specifically for you by social media companies using AI (artificial intelligence)? Is there any objective reality? Does truth even exist?

"There's a gullible side to the American people," said the filmmaker Michael Moore. Proverbs 14:15 in the Old Testament says, "A fool will believe anything; smart people watch their step."[63] Indeed, we do not want to believe *everything* we watch in the news or read on the Internet. So how do we discern what is true from what is not?

These are perhaps questions in epistemology (philosophy of how we gain knowledge), so we will set them aside for now. What we *can* do is to collect data carefully and objectively, describe it, analyze it, and draw some conclusions with some level of certainty. The science of collecting/analyzing data from samples and drawing conclusions is called **statistics**. The first part is referred to as "descriptive statistics" whereas the second part is called "inferential statistics." Statistics is an area of applied mathematics which has become extremely crucial in the modern world in virtually every discipline and field, including business, economics, politics, education, psychology, sociology, and other sciences.

If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, statistics is one of the most important things that you can study. By mastering the use of statistics, you can use it for support *and* illumination, despite the above quotation by a legendary baseball broadcaster.

[63] The Good News Translation of the Bible

5.1 Basics of Statistics[x]

Here are some claims that you may have heard on several occasions. (We are not saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentyne.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Native Americans are significantly more likely to be hit crossing the streets than are people of other ethnicities.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- It is very unlikely there will ever be another baseball player with a batting average over .400.
- There is a 70% chance that in a room full of 30 people at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

Before going further, however, two distinct meanings of the word "statistics" should be clarified. First, "**statistics**" could mean the academic discipline, a field of study, involving the collection, presentation, interpretation, and use of data. When used in this sense, the word is considered singular though it looks plural (same with "physics" and "mathematics"). Second, in a narrower sense, a number computed from data in a survey or study is referred to as a "**statistic**." For instance, the average height of a sample of 100 people would be a (one) statistic. This is a singular word (of course, two or more of these numbers would be referred to as "statistics"). Be aware of the difference when you encounter this word.

Data are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life. This chapter will help you learn statistical essentials. It will make you into an intelligent consumer of statistical claims.

You can take the first step right away. To be an intelligent consumer of statistics, your first reflex must be to question the statistics that you encounter. The British Prime Minister Benjamin Disraeli (1804—1881) famously said, "There are three kinds of lies -- lies, damned lies, and statistics." This quote reminds us why it is so important to understand statistics. No longer will you blindly accept numbers or findings; instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

We have already put the emphasis on defending ourselves against misleading claims such as fallacies (Ch. 1), deceptive percentages (Ch. 2), rates of interest (Ch. 3), and surprising

probabilities (Ch. 4). Now we do the same statistics. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. When a research team is testing a new treatment for a disease, statistics allows them to conclude, based on a relatively small experiment, that there is good evidence their drug is effective. Statistics allowed prosecutors in the 1950's and 60's to demonstrate racial bias in jury panels. Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases.

Populations and Samples

Before we begin gathering and analyzing data, we need to characterize the population we are studying. If we want to study the amount of money spent on textbooks by a typical first-year college student, our population might be all first-year students at our college. Or it might be

- All first-year community college students in our state;
- All first-year students at public colleges and universities in the state;
- All first-year students at all colleges and universities in the state; or
- All first-year students at all colleges and universities in the entire United States.

Population

The population of a study is the group which the collected data is intended to describe. This is the entire set, the universal set, about which you want to make a claim or an inference.

In this context, the population is not a number (it is the group itself, not how many people are in the group). Sometimes the intended population is called the **target population**. If we design our study poorly, the collected data might not actually be representative of the intended population. It is, therefore, essential that the target population be identified before you begin collecting any data. For instance, in the above example on the textbook cost, we might get different answers to our question as we vary the population we are studying. First-year students at a large university might take slightly more diverse courses than those at your college, and some of these courses may require less popular textbooks that cost more. On the other hand, a large university bookstore might have a larger pool of used textbooks, reducing the cost of these books to the students. Whichever the case (it is likely that some combination of these and other factors are in play), the data we gather from *your* college alone will probably not be the same as that from a large university. Particularly when conveying our results to others, we want to be clear about the population we are describing with our data.

Example 1

A newspaper website contains a poll asking people their opinion on a recent news article. What is the population?

[Solution]

While the target (intended) population may have been all people, the real population of the survey is readers of the website.

If we were able to gather data on every member of our population, say the average (we will define "average" more carefully in a subsequent section) amount of money spent on textbooks by each first-year student at your college during a recent academic year, the resulting number would be called a **parameter**, and that is the *true value* we are seeking.

Parameter

A parameter is a value (average, percentage, etc.) calculated using *data* from every member of a population.

We seldom see parameters, however, since surveying an entire population is usually very time-consuming and expensive, if not impossible (unless the population is very small or we already have the data collected).

Census

A census is a survey of an entire population. *Every member* of the population is to be included in a census.

You are probably most familiar with one common type of census: the official government census that attempts to count the population of the U.S. every ten years.[64] This demonstrates one additional problem with a census: the difficulty in finding and getting participation from everyone in a large population. This can bias, or skew, the results.

There are occasionally times when a census is appropriate, usually when the population is fairly small. For example, if the manager of Starbucks wanted to know the average number of hours her employees worked last week, she should be able to pull up payroll records or ask each employee directly.

Since surveying an entire population is often impractical, we usually select a **sample** to study.

Sample

A sample is a smaller subset of the entire population, ideally one that is fairly representative of the whole population.

[64] as required by Article 1, Section 2 of the U.S. Constitution. The Census generates the apportionment population for each state, which is used to calculate the number of congressional representatives of each state. See Section 8.1 for apportionment methods.

Note that the term "sample" refers to the *set* from which data is collected. If you are one of the people in the sample, you yourself are *not* "a sample." You are a *member* of the sample. Of course, since a sample is a subset of the population, you are also a member of the population. We will discuss sampling methods in greater detail later. For now, let us assume that samples are chosen in a careful and appropriate manner. If we survey a sample, say 100 first-year students at your college, and find the average amount of money spent by these students on textbooks, the resulting number is called a **statistic**.

Statistic

A statistic is a value (average, percentage, etc.) calculated using the data from a sample.

Example 2

A researcher wanted to know how citizens of Los Angeles felt about a voter initiative. To study this, she goes to a large shopping mall and randomly selects 500 shoppers and asks them their opinion. 60% indicate they are supportive of the initiative. What are the sample and the population? Is the 60% value a parameter or a statistic?

[Solution]

The sample is the 500 shoppers questioned. The population is less clear. While the intended population of this survey was all the citizens of Los Angeles, the effective population was mall shoppers. There is no reason to assume that the 500 shoppers questioned would be representative of all the citizens of Los Angeles.

The 60% value was based on the sample, so it is a statistic. (Note that this is singular).

Here is a good way to summarize the main concepts so far. The **population** is the entire set of people (or products, or whatever) in which we are interested, of which we want to make some inference. What we are trying to determine is some value (average, approval rate, etc.) about this population. That *real value*, the true number, is called a **parameter**. But often it is impossible to collect data from every member of the population. In other words, *the parameter is often impossible to find*. So we study and collect data from a **sample**, which is a *subset* of (usually much smaller than) the population, and we then carry out some calculation from the data collected (such as the average value or a percentage). That number is called a **statistic**. Thus, a statistic belongs to the sample. However, we use that statistic to make an inference about the population. For this reason, the statistic is often called an **estimator** of the parameter, the true number which we may never be able to find.

For example, there is no way that anyone can ever find the true average height of all Americans since it is impossible to accurately measure the height of every single American, ideally all at the same time. So the *parameter* "the average height of all Americans" is impossible to know. However, by carefully selecting a representative *sample*, measuring the height of those Americans in the sample, and computing their average, we can come up with a

statistic, the sample mean, which is then used to *estimate* the parameter. This is statistics at its fundamental level.

Example 3

For each of the studies described below, identify the population, the sample, the parameter, and the statistic.

(1) A university has 52,000 students, and the administration surveys the age of 500 randomly chosen students to estimate the average age of all the students there.
(2) A factory has just manufactured 500,000 parts. The quality engineer has picked 1000 of them and discovered that 3.7% of the parts in the sample were defective.
(3) A consumer group has measured the volume of a popular soda in 50 two-liter bottles randomly purchased all over America. The average volume was 2.03 liters.
(4) 1000 registered voters in a Midwest town was asked if they would support a political candidate, and only 29% said they would.

[Solution]

In each case, try to determine who/what was actually studied (sample), the data from that subgroup (statistic), what that group is supposed to represent (population), and the truth number or ratio you are looking for (parameter) for that group. Note that the *population is not a number*. Rather, it is the group about which the researchers are trying to make an inference.

(1) The 52,000 students at the university is the population, and the 500 students chosen is the sample. The true average age of the 52,000 students is the parameter, and the average age of the 500 students in the sample is the statistic.
(2) The population is all the parts just manufactured, i.e., 500,000 parts. The 1000 of those parts picked by the engineer is the sample, and 3.7% is the statistic. The parameter would be the true defective ratio out of all the parts manufactured, which is estimated by the statistic.
(3) The population appears to be all the two-liter bottles sold in America. The sample is the 50 bottles measured. 2.03 liters is the statistic (called a **sample mean**), which is used to estimate the parameter, which is the real average volume in a two-liter bottle of this soda sold in America.
(4) The population is all the registered voters of this town (the number is not mentioned in the problem). The 1000 voters who were asked this question make up the sample, and the value 29% is the statistic. The parameter is the true percentage of the registered voters who support this candidate.

One more thing should be mentioned before going further; it has to do with the word **data.** Technically speaking, the word “data” was the plural form of the singular noun “datum” (the Latin word meaning *something given*). Because of this etymology, a few English speakers and writers still insist that “data” should be treated as a plural word; they write, “The data are…”

or “The data show…” However, most people now use “data” as a singular noun as well. So you can say, “The data is…” or “The data shows…,” which may sound more natural. In this textbook, we will use both because there is no plural form of “data” if it is considered a singular noun. (No one should say, “datas.”)

Try It Now 1

A college reports that the average age of their students is 28 years old. Is this a statistic or a parameter? (What else do you need to know to answer this question?)

Categorizing Data

Once we have gathered data, we might wish to classify it. Roughly speaking, data can be classified as categorical data or quantitative data.

Quantitative and Categorical Data

Quantitative data are data that are numerical in nature and with which we can perform meaningful arithmetic calculations (such as finding a difference or the average value).

Categorical (qualitative) data are pieces of information classifying the objects into descriptive or conceptual categories. These are typically non-numerical.

Example 4

Which type of data are you collecting by asking each of the following questions? Quantitative or categorical (qualitative)? Why?

(1) “What is your favorite movie of all time?”
(2) “How many movies have you see in a movie theater in the last 12 months?”
(3) “What is your ZIP code?”
(4) “On a scale of 1 to 10, how would you rate the movie you just saw?”

[Solution]

(1) The responses would look like *Soul, Return of the Jedi,* and *Citizen Kane.* We might count the number of people who give each answer, but the answers themselves do not have any numerical values. We cannot perform computations with an answer like "*Finding Nemo*." This would be categorical data.
(2) This data would be quantitative since you could carry out meaningful numerical calculations such as finding the average and say, “The sample average was 3.7 movies watched in a theater last year.”
(3) ZIP codes are numbers, but we cannot do any meaningful mathematical calculations with them. It does not make sense to say that 98036 is "twice" 49018 — that would be like saying that Lynnwood, WA is "twice" Battle Creek, MI, which does not make sense at all). Therefore, ZIP codes are really categorical data, not quantitative.

(4) Note that a movie that rates an 8 is not necessarily twice as good as a movie that rates a 4, whatever that means. If two people see the movie and one of them thinks it is a 1 and the other thinks it is a 9, it does not necessarily make sense to say that their average rating was a 5. Scales do not work that way. However, it *does* make sense to discuss the average rating and other statistics in this case. For instance, a movie whose average rating is a 1.8 is clearly different from one whose average is a 9.2. Also, it would mean something if the maximum rating of another movie was a 4. So, this data is quantitative.

Try It Now 2

Classify each measurement as categorical (qualitative) or quantitative.

a. **Eye color of a group of people**
b. **Daily high temperature of a city over two weeks**
c. **Annual income**
d. **Occupation of people in a neighborhood**

Sampling Methods

As we mentioned in a previous section, the first thing we should do before conducting a survey is to identify the population that we want to study. Suppose we are hired by a politician to determine the amount of support she has among the electorate should she decide to run for another term. What population should we study? Every person in the district? Not every person is eligible to vote, and regardless of how strongly someone likes or dislikes the candidate, they do not have much to do with her being re-elected if they are not able to vote.

What about eligible voters in the district? That might be better, but if someone is eligible to vote but does not register by the deadline, that person does not have any say in the election either. What about registered voters? Many people are registered but choose not to vote. What about "likely voters"?

This is the criteria used in much political polling, but it is sometimes difficult to define a "likely voter." Is it someone who voted in the last election? In the last general election? In the last presidential election? Should we consider someone who just turned 18 a "likely voter”? They were not eligible to vote in the past, so how do we judge the likelihood that they will vote in the next election?

In November 1998, former professional wrestler Jesse "The Body" Ventura was elected Governor of Minnesota. Up until right before the election, most polls showed he had little chance of winning. There were several contributing factors to the polls not reflecting the actual intent of the electorate:

- Ventura was running on a third-party ticket, and most polling methods are better suited to a two-candidate race.

- Many respondents to polls may have been embarrassed to tell pollsters that they were planning to vote for a professional wrestler.
- The mere fact that the polls showed Ventura had little chance of winning might have prompted some people to vote for him in protest to send a message to the major-party candidates. (Something similar may have happened in the "Brexit" Referendum in the U.K. in 2016.)

But one of the major contributing factors was that Ventura recruited a substantial amount of support from young people, particularly college students, who had never voted before and who registered specifically to vote in the gubernatorial election. The polls did not deem these young people likely voters (since in most cases young people have a lower rate of voter registration and a turnout rate for elections), so the polling samples were subject to **sampling bias**; they omitted a portion of the electorate that was weighted in favor of the winning candidate.

Sampling Bias (Selection Bias)

A sampling bias is the bias caused when the sample chosen does not fairly represent the population being studied. The sampling method, therefore, must be carefully considered to avoid this type of bias.

Even identifying the population can be a difficult job, but once we have identified the population, how do we choose an appropriate sample? Remember, although we would prefer to survey all members of the population, this is usually impractical unless the population is very small. This is why we choose a sample. There are many ways to sample a population, but there is one goal we need to keep in mind; we would like the sample to be **representative of the population**. This is much easier said than done.

Returning to our hypothetical job as a political pollster, we would not anticipate very accurate results if we drew our sample from among the customers at a Starbucks, nor would we expect that a sample drawn entirely from the membership list of the local Elks Club would provide a useful picture of district-wide support for our candidate.

One way to ensure that the sample has a reasonable chance of mirroring the population is to employ **randomness**.

Random Sampling

A random sample is one in which each member of the population has an equal probability of being chosen.

If we could somehow identify all likely voters in the voting district, put each of their names on a piece of paper, toss the slips into a (very large) hat and draw 1000 slips out of the hat, we

could have a random sample. In practice, computers are better suited for this sort of endeavor than millions of slips of paper and extremely large headgear.

It is always possible, however, that even a random sample might end up not being totally representative of the population. If we repeatedly take samples of 1000 people from among the population of likely voters in the district, some of these samples might tend to have a slightly higher percentage of Democrats (or Republicans) than does the general population. Some samples might include more older people whereas some samples might include more younger people. In most cases, however, this **sampling variability** is not significant if the sample is large enough.

Sampling Variability

The natural variation of samples is called sampling variability. This is unavoidable and expected in random sampling. (In most cases it is not an issue.)

Here is another way to understand sampling variability. If everyone in your class flips 10 coins (or one coin 10 times) and counts the number of heads, many students may get 4, 5, or 6 while very few will get 0 or 10. In this case, the samples of size 10 are showing natural variation. However, as the sample size gets larger, the sampling variability becomes less significant (this is closely related to the Law of Large Numbers stated in Section 4.4).
To help ensure that the sample is representative, pollsters might instead use a **stratified sample**.

Stratified Sampling

In stratified sampling, a population is divided into a number of **subgroups** (**strata**). A random sample is then taken from each subgroup (**stratum**) with sample sizes *proportional* to the size of the subgroup in the population.

Example 5

Suppose in a particular state that previous data indicated that the electorate was comprised of 39% Democrats, 37% Republicans, and 24% independents. How would a pollster choose a stratified sample of 1000 voters?

[Solution]

In a sample of 1000 people, the data given shows that we would expect to have about 390 Democrats, 370 Republicans, and 240 independents. To accomplish this, the pollster could use computer-generated random numbers to select 390 people from among those voters known to be Democrats, 370 from those known to be Republicans, and 240 from those with no party affiliation.

Stratified sampling can be used to select a sample with people in desired age groups, a specified mix ratio of males and females, various ethnicities, races, and religious affiliations. If a *specific number* of people in each group (stratum) is to be included in the sample, the number is called "**quota**." One can collect data until the desired quota is met for each group.

Another sampling method is cluster sampling, in which the population is divided into groups, and one or more groups are randomly selected to be in the sample.

Cluster Sampling

In cluster sampling, the population is divided into subgroups (**clusters**), and a set of subgroups is selected to be in the sample.

If a college wanted to survey its students, since students are already divided into classes, the researchers could randomly select 10 classes and give the survey to all the students in those classes. This would be cluster sampling.

Other sampling methods include **systematic sampling**.

Systematic Sampling

In systematic sampling, every n^{th} member of the population is selected to be in the sample.

To select a sample using systematic sampling, a pollster could call every 100th name in the phone book, for example.

Systematic sampling is not as random as a random sampling. If your name is Albert Aardvark and your sister Alexis Aardvark is right after you in the phone book, there is little chance you could both end up in the sample), but systematic sampling *can* yield acceptable samples. Perhaps the worst types of sampling methods are convenience sampling and voluntary-response sampling.

Convenience Sampling and **Voluntary-Response Sampling**

Convenience sampling is the method of choosing a sample based on what is convenient or easy. (Sometimes there may not be any other options.)

Voluntary-response sampling is the method of choosing a sample based on who responds to a voluntary, optional survey (as in "How Is Our Service?" at a restaurant).

Suppose you stand on a campus sidewalk and interview the first 100 people who agree to speak to you. This is convenience sampling. You can see why this sample may not be representative of the whole student body of your school since you were at just one campus

location at a certain time of the day on a certain day. You may have talked to only those people taking a Tuesday morning class in a science building.

Imagine a website, say MSNBC or Fox News, with a survey asking readers to give their opinion on a tax proposal. This is a self-selected sample and a voluntary-response sample. Often voluntary-response samples are skewed towards people who have a particularly strong opinion about the subject of the survey or who just have way too much time on their hands and enjoy taking surveys.

Try It Now 3

In each case below, indicate what sampling method was used.

a. **Every 4th person in the class was selected.**
b. **A sample was selected to contain 25 men and 35 women.**
c. **Viewers of a new show are asked to vote on the show's website.**
d. **A website randomly selects 50 of their customers to send a satisfaction survey to.**
e. **To survey voters in a town, a polling company randomly selects 10 city blocks and interviews everyone who lives on those blocks.**

How to Mess Things Up Before You Start

There are several ways that a study can be ruined before you even start collecting data. The first we have already explored – sampling or selection bias, which occurs when the sample is not representative of the population. One example of this is voluntary-response bias, which is bias introduced by collecting data from only those who volunteer to participate. This is not the only potential source of bias. The following is a list of bias types; however, this is neither comprehensive nor mutually exclusive. There are more types not listed, and sometimes more than one may apply.

Types of Bias

- **Sampling bias** – when the sample is not representative of the population
- **Voluntary-response bias** – the sampling bias that often occurs when the sample consists entirely of volunteers (This is also called "**participation bias**.")
- **Self-interest study** – bias that can occur when the researchers have an interest in the outcome
- **Response bias** – when the responder gives inaccurate responses for any reason
- **Perceived lack of anonymity** – when the responder fears giving an honest answer might negatively affect them
- **Loaded questions** – when the question wording influences the responses significantly
- **Non-response bias** – when people refusing to participate in the study can influence the validity of the outcome

Clearly, the sample should not be too small—it should be big enough to represent the population. However, even a small sample could actually be quite useful if the sampling is done correctly. Most opinion polls and quality-control studies are carried out with a sample consisting of a tiny fraction of the population (e.g., 3 million would be a huge sample size, but it is still less than 1% out of the U.S. population of 331 million).[65]

Conversely, a large sample size is not automatically good; it *can* be quite biased and misleading. A classic example is the 1936 opinion poll carried out by the *Literary Digest*, which predicted that Republican Alf Landon would win over Democratic candidate Franklin D. Roosevelt with a large margin. They polled 10 million people and got 2.4 million responses, one of the largest samples studied ever. Unfortunately, however, most of those polled were those who read this magazine, who owned cars or telephones (back in the 1930s). While the sample was very large, it was extremely biased.

Example 6

Consider a recent study which found that chewing gum may raise math grades in teenagers. This study was conducted by the Wrigley Science Institute, a branch of Wrigley, a chewing gum company. Identify possible types of bias involved.

[Solution]

Obviously, this is an example of a **self-interest study**, one in which the researches have a vested interest in the outcome of the study. While this does not necessarily mean that the study was biased, it certainly suggests that we should subject the study to extra scrutiny.

Many examples similar to the above can be given. For instance, a group of scientists have concluded that dehydration can "drastically impair" athletes' endurance performance, according to a 2019 article published by the Gatorade Sports Science Institute.[66] Again, the fact that this research was carried out by an organization associated with a sports drink does not automatically invalidate the conclusion, but you would want to ensure that there is no **conflict of interest** involved.

Example 7

A survey asks people "when was the last time you visited your doctor?" Identify possible types of bias involved.

[Solution]

[65] Typical sample sizes are 600 or 1200 for opinion polls and smaller for controlled studies.
[66] Gonzalez-Alonso, J. (2019). New Ideas About Hydration and its Impact on The Athlete's Brain, Heart and Muscles. *GSSI*. https://www.gssiweb.org/sports-science-exchange/article/new-ideas-about-hydration-and-its-impact-on-the-athlete-s-brain-heart-and-muscles#articleTopic_2

This might suffer from **response bias** since many people might not remember exactly when they last saw a doctor and give inaccurate responses. Also, many people may not necessarily want you to know when you last saw a doctor.

Sources of response bias may be innocent, such as bad memory, or as intentional as pressuring by the pollster. Consider, for example, how many voting initiative petitions people sign without even reading them.

Example 8

Identify any potential types of bias involved in each of the following situations.

(1) A survey asks participants a question about their interactions with members of other races.

(2) An employer puts out a survey asking their employees if they have a drug abuse problem and need treatment help.

(3) A survey asks, "Do you support funding research of alternative energy sources to reduce our reliance on high-polluting fossil fuels?"

(4) A telephone poll asks the question, "Do you often have time to relax and read a book?"

[Solution]

(1) Here, a **perceived lack of anonymity** could influence the outcome. The respondent might not want to be perceived as racist even if they are and give an untruthful answer. It could also be classified under **response bias.**

(2) Here, answering truthfully might have consequences; responses might not be accurate if the employees do not feel their responses are anonymous or fear retribution from their employer. Clearly, this is an example of a **perceived lack of anonymity**.

(3) This is an example of a **loaded (or leading) question** – questions whose wording leads the respondent towards an answer.

(4) In recent years, phone surveys are becoming quite unreliable because of **non-response bias**, which is clearly reflected in this example. (In this specific case, ironically those who refused to answer the question may have been just too busy to relax and read a book.) Many people simply do not answer calls from pollsters any more. It is unlikely that the results will be representative of the entire population. When people refuse to participate, we can no longer have a meaningful sample.

One more thing should be mentioned in regard to bias. Sometimes the **order** in which questions are asked in a survey or an interview could dramatically change the results. A psychology researcher provides an example:

> "My favorite finding is this: we did a study where we asked students, 'How satisfied are you with your life? How often do you have a date?' The two answers were not statistically related - you would conclude that there is no relationship between dating frequency and life satisfaction. But when we reversed the order and asked, 'How often do you have a date? How satisfied are you with your life?' the statistical relationship was a strong one. You would now conclude that there is nothing as important in a student's life as dating frequency."

One way to remove bias based on the question order is for the researcher to provide two or more versions of the survey in which the questions are arranged in different order.

Try It Now 4

In each situation below, identify a potential source of bias.

a. **A survey asks how many sexual partners a person has had in the last year.**
b. **A radio station asks listeners to phone in their choice in a daily poll.**
c. **A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score.**
d. **High school students are asked if they have consumed alcohol in the last two weeks.**
e. **The Beef Council releases a study stating that consuming red meat poses little cardiovascular risk.**
f. **A poll asks, "Do you support a new transportation tax, or would you prefer to see our public transportation system fall apart?"**

Exercises 5.1

1. A political scientist surveys 28 of the current 106 members in a state's assembly. 14 of the 28 members surveyed said they were supporting a new education bill, 12 said there were not supporting the bill, and 2 were undecided.
   (a) What is the population of this survey? (Remember, the answer is *not* a number.)
   (b) What is the size of the population?
   (c) What is the size of the sample?
   (d) Give the sample statistic for the proportion of assembly members surveyed who said they were supporting the education bill.
   (e) Based on this sample, how many of the representatives can we expect to support the education bill?

2. A city has 9500 registered voters. There are two candidates for city council in an upcoming election: Brown and Feliz. The day before the election, a telephone poll

of 350 randomly selected registered voters was conducted. 112 said they would vote for Brown, 207 said they would vote for Feliz, and 31 were undecided.
   (a) What is the population of this survey?
   (b) What is the size of the population?
   (c) What is the size of the sample?
   (d) Give the sample statistic for the proportion of voters surveyed who said they would vote for Brown.
   (e) Based on this sample, how many of the 9500 voters can we expect to vote for Brown?

3. Identify the most relevant source of bias in this situation: a survey asks the following: “Should the mall prohibit loud and annoying rock music in clothing stores catering to teenagers?”

4. Identify the most relevant source of bias in this situation: to determine opinions on voter support for a downtown renovation project, a surveyor randomly questions people working in downtown businesses.

5. Identify the most relevant source of bias in this situation: a survey asks people to report their actual income and the income they reported on their IRS tax form.

6. Identify the most relevant source of bias in this situation: a survey randomly calls people from the phone book and asks them to answer a long series of questions.

7. Identify the most relevant source of bias in this situation: a survey asks the following: “Should the death penalty be permitted if innocent people might die?”

8. Identify the most relevant source of bias in this situation: a study seeks to investigate whether a new pain medication is safe to market to the public. They test by randomly selecting 300 men from a set of volunteers.

9. Identify the most relevant source of bias in this situation: a study by a milk manufacturer investigates the benefits of drinking milk daily.

10. In a study, you ask the subjects their age in years. Is this data categorical or quantitative?

11. In a study, you ask the subjects their gender. Is this data categorical or quantitative?

12. Classify each type of data below into categorical or quantitative.
    (a) Favorite food
    (b) Nationality
    (c) Household income
    (d) GPA

13. Classify each type of data below into categorical or quantitative.
    (a) Phone number
    (b) Batting average
    (c) LDL (the "bad" cholesterol level)
    (d) PIN for one's ATM card

14. In a research project, participants were asked to provide their weight. Is this data categorical or quantitative? Explain why

15. In a study, the sample is chosen by separating all cars by size and then selecting 10 of each size grouping. What is the sampling method?

16. In a study, the sample is chosen by writing everyone's name on a playing card, shuffling the deck, then choosing the top 20 cards. What is the sampling method?

17. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths.

    (a) What is the population of this study?

    (b) Would you expect the data to be identical? Why or why not?

    (c) The first researcher collected her data by randomly selecting 40 states and then selecting 1 person from each of those states. What sampling method is that?

    (d) The second researcher collected his data by choosing 40 patients he knew. What sampling method did that researcher use? What concerns would you have about this data set, based upon the data collection method?

18. In 1992, Roper Starch Worldwide Inc. (a survey company) interviewed almost 1000 adults for the American Jewish Committee concerning the Holocaust. The study (originally) concluded that "1 in 5 Americans doubted that the Holocaust had happened."

(a) The question on the survey was this: “Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?” What do you think of this question?

(b) About 22% of the people responded to the above question, indicating that it was possible for the Holocaust to have never happened. Another 12% indicated they were not certain it was possible. Based on these results, do you agree with the conclusion of this study?

(c) Later, Roper gave a new poll in which the question was re-phased to this: “Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?” Do you think this was phrased better than the original question?

(d) In the new survey, only 1% responded that it was possible that the Holocaust had never happened. 91% of the people were certain it had happened. What do you think was the reason for such a big gap in the results? (This question was based on a 1994 article by the *New York Times*, digitized recently. You can read more details at https://www.nytimes.com/1994/07/08/us/poll-on-doubt-of-holocaust-is-corrected.html).

Figure 5.1: Anne Frank, a victim of the Holocaust[67]

[67] Image is in the public domain

5.2 Types of Statistical Studies

So far, we have primarily discussed observational studies – studies in which conclusions would be drawn from observations of a sample or the population. In some cases these observations might be unsolicited, such as studying the percentage of cars that turn right at a red light even when there is a "no turn on red" sign. In other cases the observations are solicited, like in a survey or a poll.

Experiments

In contrast to observational studies, it is common to use experiments when exploring how subjects react to an outside influence. In an experiment, some kind of treatment is applied to the subjects, and the results are measured and recorded.

Types of Statistical Studies

1. An **observational study** is a study based on observations or measurements without any influence on the subjects.

2. An **experiment** is a study in which the effects of a treatment given to one group are measured and compared to a group not given the same treatment.

This definition highlights the two important attributes that characterize an **experiment**. One is suggested by the term "treatment"; in an experiment, *something is done to a group of subjects* (participants), such as administering a drug or controlling their daily routine. The second is suggested by the term "compared"; in an experiment, *the subjects are classified into at least two groups* so that the results can be compared and contrasted.

Here are some examples of experiments:

- A pharmaceutical company tests a new medicine for treating Alzheimer's disease by administering the drug to 50 elderly patients with recent diagnoses. The results are then compared to patients who did not get the new medicine. The treatment here is the new drug.
- A gym tests out a new weight loss program by enlisting 30 volunteers to try out the program. The results are then compared to those who did not participate. The treatment here is the new program.
- You test a new kitchen cleaner by buying a bottle and cleaning your kitchen. You then determine if the new cleaner is better than the one you have been using. The new cleaner is the treatment.
- A psychology researcher explores the effect of music on temperament by measuring people's temperament while they listen to different types of music. The various types of music is the treatment.

Try It Now 5

Is each scenario describing an observational study or an experiment?

a. **The weights of 30 randomly selected people are measured.**
b. **Subjects are asked to do 20 jumping jacks, and then their heart rates are measured and are compared to their heart rates of those who did not do jumping jacks.**
c. **Twenty coffee drinkers and twenty tea drinkers are given a concentration test.**

Here is something extremely important. When conducting experiments, it is essential to **isolate** the treatment being tested.

Example 1

Suppose a middle school finds that their students are not scoring well on the state's standardized math test. They decide to run an experiment to see if an alternate curriculum would improve scores. To run the test, they hire a math specialist to come in and teach a class using the new curriculum. To their delight, they see an improvement in test scores. Can they conclude that the new curriculum improved the test scores?

[Solution]

The answer is "maybe." The experiment did not give a conclusive answer. The difficulty with this scenario is that it is not clear whether the curriculum is responsible for the improvement or the improvement is due to a math specialist teaching the class.

This is called **confounding** (or **confounding variables**)– when it is not clear which factor or factors caused the observed effect. Confounding is the downfall of many experiments.[68]

Confounding

Confounding occurs when there are two or more potential variables that could have caused the outcome and it is not possible to determine which actually caused the result.

Example 2

Identify possible confounding variables in each case below.

(1) A drug company study about a weight loss pill might report that people lost an average of 8 pounds while using their new drug. However, in the fine print you find a statement saying that participants were also encouraged to diet and exercise.

[68] In addition to confounding variables, there are also **lurking variables**. While "confounding" usually refers to two variables whose effects are hard to distinguish, "lurking" refers to a variable not explicitly considered in a study.

(2) Researchers conduct an experiment to determine whether students will perform better on an arithmetic test if they listen to music during the test. They first give the student a test without music and then give a similar test while the student listens to music.

[Solution]

(1) The confounding variables here are diet and exercise. It is not clear in this case whether the weight loss is due to the pill, to diet and exercise, or a combination of both.

(2) Possible confounding variables here are the student's mental state after taking the test once and the fatigue (among other things). The student just might perform better on the second test, regardless of the music, simply because it was the second test and the student was warmed up. It is also possible that the first test exhausted the student (especially if it was long).

There are a number of measures that can be introduced to help reduce the likelihood of confounding. The primary measure is to use a control group.

Control Group

In experiments, the participants are divided into two or more groups, typically a **control group** and a **treatment group**. The treatment group receives the treatment being tested; the control group does not receive the treatment.

Ideally, the groups are otherwise as similar as possible, isolating the treatment as the only potential source of difference between the groups. For this reason, the method of dividing groups is important. Researchers attempt to ensure that the groups have similar characteristics (similar mix of male/female subjects, age groups, etc.), but it is nearly impossible to control for every characteristic. Random assignment is very commonly used to avoid bias or confounding. When the groups are randomly assigned in an experiment, we call it a **randomized controlled trial (RCT)**; this is considered a "gold standard" when it comes to clinical trials.

Example 3

Identify the treatment group and the control group in each of the experiments below.

(1) To determine if a two-day prep course would help high school students improve their scores on an AP test, a group of students was randomly assigned into two subgroups; one group was given the prep course while the other was not.

(2) A company testing a new plant food grows two crops of plants in adjacent fields. One group receives the new plant food while the other does not.

[Solution]

The answers are pretty clear in both of these. For (1), the group given the prep course is the treatment group; the group that did not get the prep course is the control group. For (2), the crop of plants given the plant food is the treatment group; the crop without the plant food is the control group.

Placebo Effect

Now, in experiment (1) above, we may not have eliminated confounding altogether. A group of students received a two-day prep course for an AP test. The fact that they participated in a new prep program, not the content of the program itself, may actually give them extra confidence and improve their score. Such a psychological effect cannot be ignored. For example, suppose a medicine study is testing a new headache pill by giving the treatment group the pill and the control group nothing. If the treatment group showed improvement, we would not know whether it was due to the medicine in the pill or a response to have taken any pill (the idea "I should be getting better because I am taking some pill"). This is called a **placebo effect**.

Placebo Effect

The placebo effect refers to the effectiveness of a treatment influenced by the patient's perception of how effective they think the treatment should be, so an improvement might be seen even if the treatment itself is ineffectual.

For this reason, the control group is often given a "fake pill," called a **placebo**, which appears identical to the real treatment but without the active ingredient(s). A placebo could be a sugar pill, pure water, or a saline injection (as used in new vaccine trials).

For example, a study found that when doing painful dental tooth extractions, patients who were told they were receiving a strong painkiller (while actually receiving a saltwater injection) found as much pain relief as patients receiving a dose of morphine. In this case, the saltwater was the placebo. Many other **placebo effects** are well-documented. There is no question that some placebos work; it might even over-perform the real medicine being tested.

A placebo also plays a crucial role in identifying false "side effects." Some people may suffer adverse effects because they think they are taking an experimental medicine. If both the treatment group and the placebo group are experiencing various "side effects" at similar rates, they are not considered to have been caused by the treatment in question.

Placebo and Placebo-Controlled Experiments

A **placebo** is a harmless, inactive treatment given to control for a psychological effect, known as the "placebo effect."

An experiment that gives the control group a placebo is called a **placebo-controlled experiment**.

During the 2020 COVID-19 pandemic, several companies were developing vaccines for the new coronavirus in an effort promoted by "Operation Warp Speed." Every one of those firms used randomized and placebo-controlled experiments in their clinical trials because that is the only accepted method of testing new medication products such as vaccines. "Randomized" means that the treatment group and the control group were formed so that, except for the vaccine treatment, everything else should be as identical as possible to eliminate any confounding variables. It is extremely important to isolate reasons for improvement (and any possible side effects); a placebo is an essential part of such a process.

Example 4

What can/should be used as a placebo, if any, under each circumstance below?

(1) In a study for a new medicine that is dispensed in a pill form
(2) In a study on the effect of alcohol on memory
(3) In a study of a frozen meal diet plan
(4) In a study to test the effectiveness of a new plant food

[Solution]

(1) Since the new medicine is a pill, a sugar pill can be used as a placebo.
(2) The treatment group could be given regular beer while the control group could get a non-alcoholic beer as a placebo. But this could be a problem as some (possibly many) people could actually tell the difference between a non-alcoholic beer and a regular beer.
(3) Standard frozen meals (not for any diet) stripped of their original packaging would serve as a placebo here. They should look like the meals given under the frozen food diet plan.
(4) We can safely assume that plants will never experience a placebo effect, which is psychological in nature. When the subjects are non-human, placebos are not necessary.

In some cases, it is more appropriate to compare to a conventional treatment than to a placebo. For example, in a cancer research study, it may not be ethical to deny any treatment or to give a placebo to the control group. In this case, the currently acceptable medicine would be given to the control group, which is (in this case) called a **comparison group**. In our AP test

prep example, the non-treatment group would most likely be encouraged to study on their own, rather than be asked to not study at all, to provide a meaningful comparison.
When using a placebo, it would defeat the purpose if the participant knew they were receiving the placebo. This leads to the idea of **blindness** in experiments.

Blind Studies

A **blind study** is one in which the participant does not know whether or not they are receiving the treatment or a placebo.

A **double-blind study** is one in which even those interacting with the participants, such as the doctors who assess the patients, do not know who is in the treatment group and who is in the control group.

A blind study that is not double-blind is sometimes called a **single-blind study**.

Note that a placebo-controlled experiment is, by necessity, at least single-blind. Double-blind tests are costly but can provide more security and validity than single-blind tests as illustrated below.

Example 5

What can be used as a placebo in each case of the following cases (if a placebo is possible)? If a placebo is used, should it be a single-blind test or a double-blind test? Why?

(1) A pharmaceutical company wants to test a new pill to reduce the LDL (bad cholesterol) level.
(2) A medical group wants to see if one particular yoga exercise reduces the anxiety level of corporate managers.
(3) Education experts want to study if piano lessons improve the children's intelligence.
(4) A drug company wants to show their new anti-depressant is effective in treating depression.
(5) A researcher wants to examine if the music of Mozart makes better wines during the fermentation process.

[Solution]

(1) A sugar pill, identical to the real pill being tested, would serve as a placebo. In this case, a single-blind test is sufficient since the evaluator (the doctor, the assessor) has an objective measure, i.e., the LDL level. Whether the evaluator knows a particular patient is taking the placebo or the real medicine is irrelevant because of this objective (not subjective) data.
(2) Because they are testing a certain (specific) exercise, a placebo should be used—perhaps some other yoga exercise similar to the one being tested. Here is an

example where a **double-blind study** is needed. If the evaluator knows that a particular patient is in the treatment group (and thus the anxiety level of the patient *should* be reduced), the evaluator may give more positive assessment to that patient while the anxiety level of a patient known to be in the control group may be diagnosed as "steady" or "worse." Here, the evaluator's assessment is subjective, not like the LDL level before. This is when a double-blind test is appropriate.

(3) Ask yourself, "What could be a placebo for piano lessons?" A fake piano lesson? It may be difficult to find an appropriate "fake pill." A placebo is probably not possible here.

(4) This is similar to (2) above. Since the level of depression is very subjective, a double-blind study should be conducted with a placebo, which can be a sugar pill.

(5) During fermentation, yeasts transform sugars in the grape juice into alcohol. Neither the yeasts nor the juice suffers from placebo effects; they do not feel pressed to produce better wine because of Mozart. They are not fooled. Therefore, no placebo is necessary. A placebo would be difficult anyway (would you play Bach or Beethoven instead?) Since no placebo is used, this is not a blind study.

Occasionally, it may be necessary to conduct trials where placebos are impossible or unnecessary but the evaluator must be kept blind. For instance, in (5) above, the subjects (wine being made) do not have to be blinded since there is no placebo. However, those who evaluate the quality of the finished products (wine made while Mozart is being played and traditional wine without Mozart) may have their judgment affected by knowing which wine is which. In situations like this, the evaluators/assessors may be asked to "**blind-test**" the results.

A similar accommodation was made for an orchestra tryout, where some suspected that older, male musicians had been far more likely to be hired to play in the orchestra. The tryout was conducted as each musician played behind a curtain so that they can be judged for the quality of their music alone.

It should be noted, however, that not every experiment needs a control group. For example, if a researcher is testing whether a new fabric can withstand fire, she simply needs to torch multiple samples of the fabric – there is no need for a control group.

Case-Control Studies

Before ending this section, one more type of study should be mentioned: **case-control studies**. This type of research contains elements in both observational studies and experiments.

Case-Control Studies

A case-control study is a statistical study in which two **existing groups** with different outcomes are identified and their association with certain possible causes is studied.

It is an observational study in the sense that researchers *do not influence* the patients/subjects (no medicine or placebo is given); it is also an experiment in the sense that *two groups are compared*.

The way in which the two groups are identified in a case-control study is not random as in an experiment. These are **naturally occurring groups** as the following examples show. These examples also suggest why sometimes a case-control study is the only option (where an experiment may be impossible or unethical).

Example 6

Consider why a case-control study is appropriate in each of the following cases.

(1) A group of doctors want to study if smoking causes lung cancer.
(2) Researchers want to study if wearing seat belts actually reduces the risk of fatal accidents.
(3) A company wants to know if alcohol consumption is harmful to pregnant women.

[Solution]

The answer may be obvious in each case. (1) is actually a classic case where a case-control study succeeded in establish a clear association. No researcher can ethically tell half the subjects *to smoke* and tell the other half not to smoke. However, some people choose to smoke while others do not; therefore, the cases group and the control group are naturally and voluntarily identified. Researchers then studied the rate of lung cancer among the smokers and compare that to the rate among the non-smokers.

For (2), again, it is not ethical to ask half the drivers to wear seat belts and the other half not to wear seat belts. But drivers choose for themselves which group they belong to. Researchers can then compare the rate of traffic-related deaths of the two voluntarily defined groups.

Similarly, for (3), researchers cannot make certain pregnant women drink alcohol. But retrospectively, they can study the rate of pregnancy-related problems (miscarriages, birth defects, etc.) between the two groups.

Notice that case-control studies are often necessary for **ethical reasons**. You cannot make certain people consume alcohol or drugs; you cannot endanger people's lives for the sake of research. For this reason, certain ethical standards are widely held among statisticians and researchers. In the past, some placebo-controlled studies were conducted in blatantly unethical ways. One of the most infamous examples is the **Tuskegee Syphilis Study**, which began in 1932 and lasted for some 40 years in Alabama. 600 African-American males were promised free health care and medication while given just placebo treatments. The study led to 128 deaths.

It should be mentioned that, technically speaking, **neither observational studies nor case-control studies can actually prove causation**; all they can do is to establish an association. But even strong association does not automatically imply causation. In the above example of smoking and lung cancer (a 1950 article[69] by Richard Doll and Bradford Hill, for example), researchers were able to show "statistically significant" association between smoking and lung cancer using a case-control study, but further analysis was required to prove causation. Here is something very important to remember. If you learn nothing else from this book or the course, learn this. It was one of the fallacy types studied in Chapter 1:

Association vs. Causation
Association (or correlation) does not establish causation. Therefore, an observational study can never prove causal relationship.

Example 7

A group of sociologists carried out a study on very wealthy people and others. This is a case-control study because these groups naturally existed already prior to the comparison. (Researchers did not make one group wealthy and the other group poor.)

The groups are those whose personal assets exceed $100 million (Group A, the case group) and everyone else (Group B, the control group). The research discovered that the percentage of those practicing meditation therapy was significantly higher in Group A than in Group B. They concluded that there is strong association between meditation therapy and wealth. Is that correct?

[Solution]

Read their claim carefully again. You may have read it with bias or a pre-conceived notion. Note that the researchers did *not* claim that practicing meditation therapy will help you get rich, nor did they explicitly state that meditation therapy **causes** more wealth. They simply made a statement about the **association**, or (positive) **correlation**. Assuming that their data indeed showed some correlation, their conclusion is actually correct.

The problem is that, when most of us read this, we frequently jump to *our* conclusion because, well, we easily fall for the **fallacy "correlation implies causation."** It is very important that you distinguish the two. Yes, there seems to be association here, but meditation therapy is not necessarily a cause for their riches. In fact, the opposite may be true (though you cannot prove it either); their wealth may be enabling them to practice meditation therapy, or their wealth may drive them to seek meditation therapy. So do not jump on the bandwagon and look for your local meditation therapy center, hoping to get rich.

[69] Doll, R., & Hill, B. (1950). Smoking and Carcinoma of the Lung. *British Journal of Medicine*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2038856/pdf/brmedj03566-0003.pdf

In fact, there is also (very) strong association between Group A and those who own private jets. (The percentage of those who own private jets is significantly higher in Group A than in Group B.) Clearly, you would not think that buying a private jet will improve your chance of becoming wealthy, right?

I hope you will remember, and recognize, this type of fallacy; this is very prevalent in the media.

Example 8

A case-control study in 2020 showed that there was a close link between wearing a mask and reduced risk of COVID-19 infection. Does it prove that wearing a mask reduces the risk of infection?

[Solution]

As much as one may hope this is true, no case-control study can prove causation. This study may prove strong association, but it does not prove that wearing a mask *causes* reduction in the infection risk. This is *not* to say that the causal relationship does not exist. (It should be noted that health experts overwhelmingly support the view that wearing a mask *does* indeed reduce the risk. Scientific evidence seems abundant—but not because of observational or case-control studies.)

Various websites and online videos suggest strange but real associations in all types of quantities; these serve as excellent examples to show that *correlation does not imply causation* (and that in many such cases the association may be purely **coincidental**). For example, a YouTube video[70] explains how ice cream used to be blamed for polio because of the seasonal correlation between consumption of ice cream and the number of polio cases. A website called "Spurious Correlations" list many closely associated quantities, including one between mozzarella cheese consumption and the number of civil engineering doctorates awarded.[71]

Let us conclude this section with a silly but illuminating example.

Example 9

There is a strong positive correlation in Europe between the number of babies born in a country and the number of storks there; nations with more storks give birth to more (human) babies. Does this prove the old saying that storks deliver babies?[72]

[Solution]

[70] https://www.youtube.com/watch?v=lbODqslc4Tg
[71] https://www.tylervigen.com/spurious-correlations
[72] This example is used in the introduction of Tim Harford's book *The Data Detective*.

Of course, right?

No. An easy explanation here is that larger countries simply have more birds *and* more people. This is an example of a **common underlying cause** (the sheer size of each country). Claiming that storks deliver babies is just as baseless as claiming that babies bring storks. Whenever there is an association, consider some common underlying cause that could explain the association.

Exercises 5.2

1. Does this describe an observational study or an experiment? The temperature on randomly selected days throughout the year was measured.

2. Does this describe an observational study or an experiment? A group of students are told to listen to music while taking a test and their results are compared to a group not listening to music.

3. A team of researchers is testing the effectiveness of a new vaccine. They randomly divide the subjects into two groups. Group 1 receives new vaccine, and Group 2 receives the existing vaccine. The patients in the study do not know which group they are in.
   (a) Which is the treatment group?
   (b) Which is the control group (if there is one)?
   (c) Is this study single-blind, double-blind, or neither?
   (d) Is this a placebo-controlled experiment?

4. For the clinical trials of a weight-loss drug containing *garcinia cambogia* the subjects were randomly divided into two groups. The first received an inert pill along with an exercise and diet plan, while the second received the test medicine along with the same exercise and diet plan. The patients do not know which group they are in, nor do the fitness and nutrition advisors.
   (a) Which is the treatment group?
   (b) Which is the control group (if there is one)?
   (c) Is this study single-blind, double-blind, or neither?
   (d) Is this an observational study, a case-control study, or a placebo-controlled experiment?

5. A teacher wishes to know whether the males in his class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes.
   (a) Is this a sampling or a census?
   (b) Is this an observational study or an experiment?
   (c) Are there any possible sources of bias in this study?

6. Several farmers believe that playing Barry Manilow songs to their peas will increase their yield. Describe a controlled experiment the farmers could use to test their theory. Should they use a placebo? If so, what? If not, why not?

7. A sports psychologist believes that people are more likely to be extroverted as adults if they played team sports as children. There are at least two possible studies to test this theory. Design one as a case-control study and the other as an experiment. Which is more practical?

8. A food scientist is studying if eating a certain South American plant reduces one's blood pressure. She identifies 400 subjects to participate in the study and randomly divides them into two groups.
   (a) Should this be an observational study, a case-control study, or an experiment?
   (b) What should the treatment group get?
   (c) What type of placebo, if any, should be used?
   (d) Should this be a single-blind test, a double-blind test, or neither?

9. A group of epidemiologists is studying the effectiveness of social-distancing during a virus pandemic. They have identified 1000 people who constantly stay at least 6 feet away from everyone; they have also identified 1000 people who do not practice social-distancing. They then compare the infection rate of each group.
   (a) Explain why this is a case-control study.
   (b) Explain why the use of a placebo is not possible in this case.
   (c) Explain why the researchers cannot conduct an experiment here.

10. See Exercise 9 first. Suppose that they have discovered that the social-distancing group had a significantly lower rate of infection compared to the other group. Explain why this does *not* establish that social-distancing is the cause of the lower rate. (Are there any confounding variables?)

11. Find a newspaper or magazine article (print or online) which describes the results of a recent study (the results of a poll are not sufficient). Give a summary of the study's findings and analyze whether the article provided enough information to determine the validity of the conclusions. If not, produce a list of things that are missing from the article that would help you determine the validity of the study. Look for the things discussed in the text: population, sample, randomness, bias, placebos, and experiments.

12. Do some research on the "Tuskegee Syphilis Study." Explain why it was inappropriate to conduct a placebo-controlled study in that case. Could they have conducted a case-control study instead?

5.3 Describing Data[xi]

Once we have collected data from surveys or experiments, we need to summarize and present the data in a way that will be meaningful to the reader or the listener. We will begin with graphical presentations of data and then explore numerical summaries of data.

Presenting Categorical Data Graphically

Categorical, or qualitative, data are pieces of information that allow us to classify the objects under investigation into various categories. We often start working with categorical data by summarizing the data into a **frequency table**.

Frequency Table

A frequency table is a table with two columns. One column lists the categories, and another for the frequencies with which the items in the categories occur (how many items fit into each category).

Example 1

An insurance company determines vehicle insurance premiums based on known risk factors. If a person is considered a higher risk, their premiums will be higher. One potential factor is the color of your car. The insurance company believes that people with some color cars are more likely to get in accidents. To research this, they examine police reports for recent total-loss collisions. The data is summarized in the frequency table below.

Color	Frequency
Blue	25
Green	52
Red	41
White	36
Black	39
Grey	23

Sometimes we need an even more intuitive way of displaying data. This is where charts and graphs are helpful. There are many, many ways of displaying data graphically, but we will concentrate on one very useful type of graph called a bar graph. Here, we will work with bar graphs that display categorical data; later, we will be devoted to bar graphs that display quantitative data.

Bar Graph

A bar graph is a graph that displays a bar for each category with the length of each bar indicating the frequency of that category.

Example 2

Using our car data from above, note the highest frequency is 52, so our vertical axis needs to go from 0 to at least 52 (we use 0 to 55 here). We can also put a hash mark every 5 units:

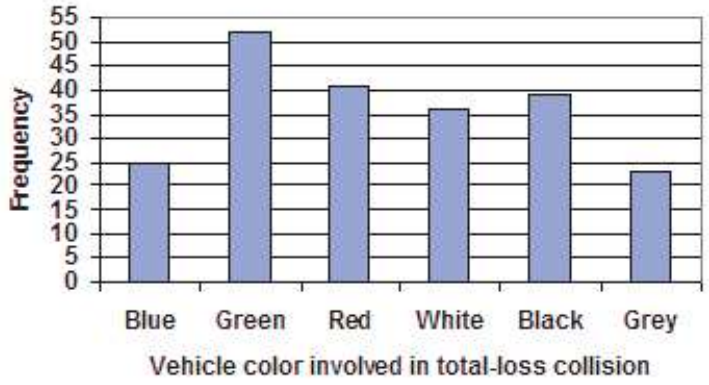


Figure 5.2[73]

Notice that the height of each bar is determined by the frequency of the corresponding color. The horizontal gridlines are a nice touch but not necessary. Instead of gridlines, we might also list the frequencies at the top of each bar, like this:

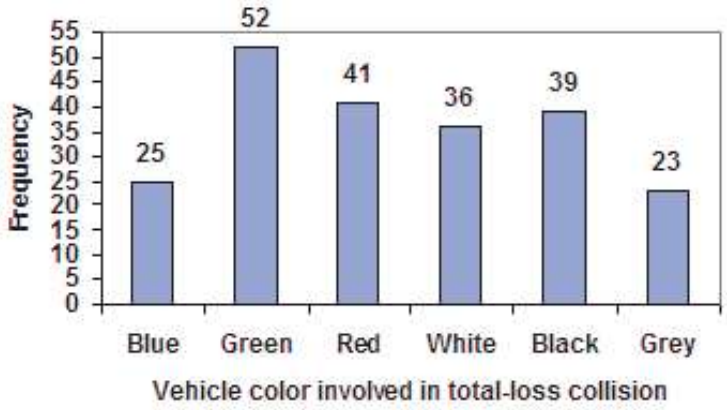


Figure 5.3[74]

[73] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.
[74] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

In this case, our chart might benefit from being re-ordered **from largest to smallest frequency values**. This arrangement can make it easier to compare similar values in the chart even without gridlines.

Example 3

Re-ordering our bars from the tallest to the shortest, we get:

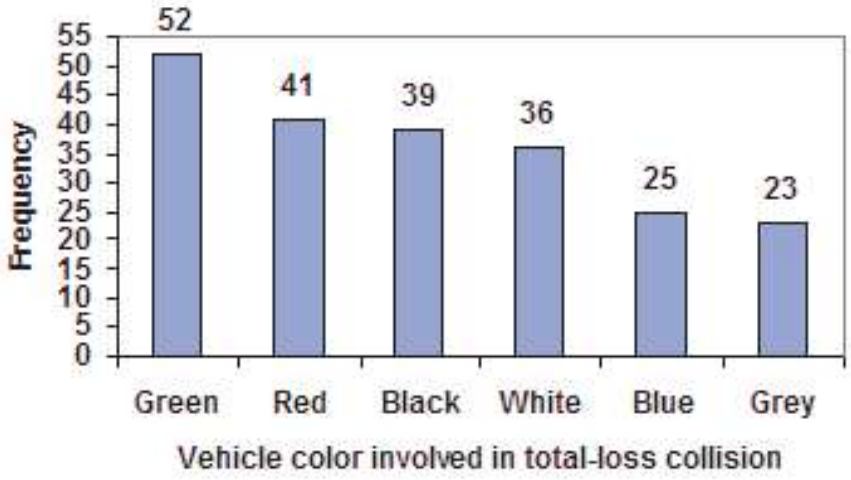


Figure 5.4[75]

A bar graph ordered in this way is sometimes referred to as a **Pareto chart**, but typically a Pareto chart also has a line graph showing the **cumulative total** above the bars.

To show relative (percentage) sizes, it is common to use a pie chart.

Pie Chart

A pie chart is a circle with wedges cut of varying sizes marked out like slices of pie. The relative sizes (angles) of the wedges correspond to the relative frequencies of the categories.

[75] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

Example 4

For our vehicle color data, a pie chart might look like this:

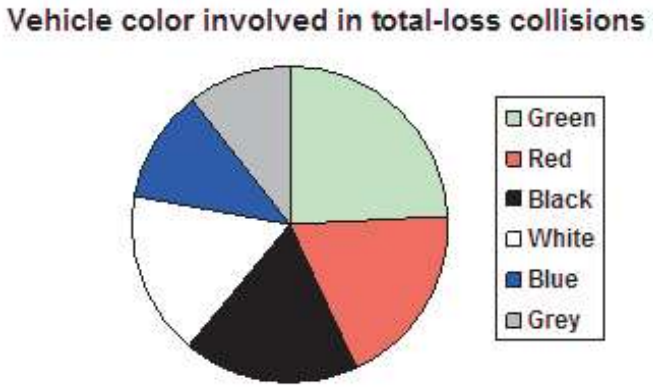


Figure 5.5[76]

Pie charts can often benefit from including frequencies or relative frequencies (percentages) in the chart next to the pie slices. Having the category names next to the pie slices also makes the chart clearer.

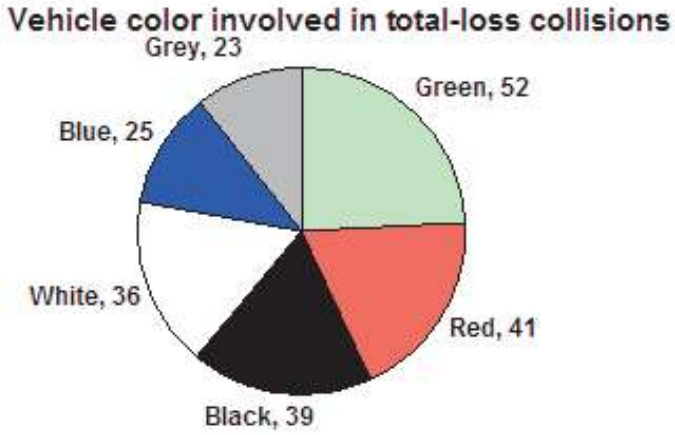


Figure 5.6[77]

Pie charts look nice but are harder to draw by hand than bar charts. Common apps and programs like Microsoft Excel are able to create bar graphs, pie charts, and other graph types.

Try It Now 6

Create a bar graph and a pie chart to illustrate the grades on a history exam below.

A: 12 B: 19 C: 14 D: 4 F: 5

[76] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.
[77] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

Fancy graphs and images can be helpful, but be careful since they can also be misleading. Instead of plain bars, we could use meaning images, called **pictograms.**

Pictogram

A pictogram is a statistical graphic in which the size of the picture is intended to represent the frequencies or size of the values being represented.

Example 5

A labor union might produce the graph below to show the difference between the average manager salary and the average worker salary.

Looking at the picture, it would be reasonable to guess that the manager salaries is 4 times as large as the worker salaries – the *area* of the bag looks about 4 times as large. However, the manager salaries are in fact only twice as large as worker salaries, as reflected in the picture by making the manager bag twice as *tall*.

Figure 5.7[78]

Remember from Chapter 2 that, given the same shape, when a dimension (such as the height of these bags) doubles, the area (the circular images above) quadruples, and the volume (3-dimensional) becomes 8 times as large. This type of distortion could be deceptive and sometimes intentionally used to exaggerate quantities.

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the vertical axis, representing the least number of cases that could have occurred in a category. Normally, this number should be **zero**.

[78] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

Example 6

Compare the two graphs below showing support for same-sex marriage rights from a poll taken in 2008. The vertical scale on the first graph (from 0 to 100%) shows the true difference in percentages; the second graph makes it look like three times as many people oppose marriage rights as support it.

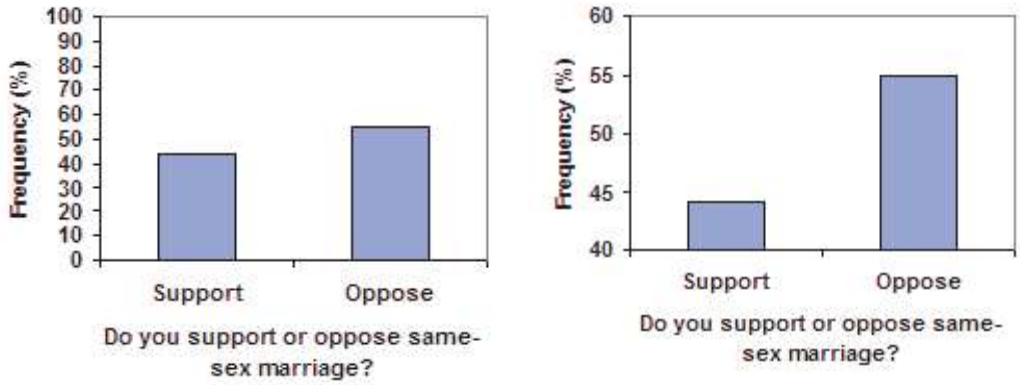


Figure 5.8

Try It Now 7

A poll was taken asking people if they agreed with the positions of the 4 candidates for a county office. Does the pie chart present a good representation of this data? Explain what might be happening here.

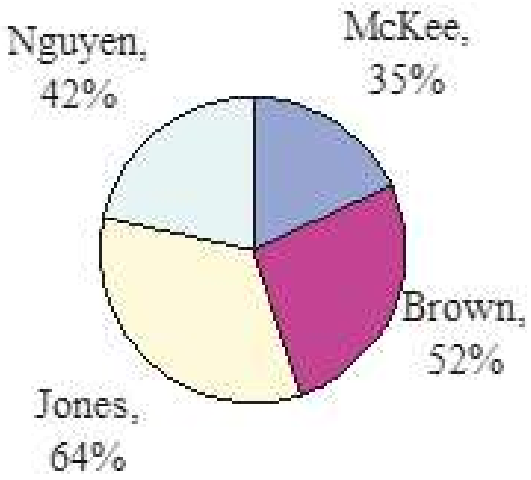


Figure 5.9[79]

Presenting Quantitative Data Graphically

Quantitative, or numerical, data can also be summarized into frequency tables. However, if we have a large number of widely varying data values, creating a frequency table that lists every

[79] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

possible value as a category would lead to an exceptionally long frequency table and may not reveal any patterns. For this reason, it is common with quantitative data to group data into **class intervals**. (Here, assume that all data values are whole numbers, i.e., no decimals or fractions.)

Class Intervals

Class intervals are groupings of the data. In general, we define class intervals so that:

- **Each interval is equal in size** (for example, if the first class contains values from 120-129, the second class should include values from 130-139).
- We have a **reasonable number of classes**, somewhere between 5 and 20, typically (depending upon the number of data values we are working with).

Sometimes each class is referred to as a **bin** or a **bucket**.

Once we define class intervals and the corresponding frequency table for the given set of quantitative data, we can construct a graph similar to a bar chart presented earlier. However, since the data values are numerical, it is better to treat the horizontal axis as a **continuous number line**. Such a graph is called a **histogram**.

Histogram

A histogram is a way to represent numerical data, similar to a bar graph, except the horizontal axis is a number line. Each class (bin) is represented by an interval on that line, and the height of the rectangle over that interval represents the frequency of the data value in that class.

Example 7

Suppose that we have collected weights from 100 male subjects as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of 263-121 = 142. We could create 7 intervals with a width of around 20, or we can have 14 intervals with a width of around 10, or somewhere in between. We often have to experiment with a few possibilities to find something that represents the data well. Let us try using an interval width of 15. We could start at 121, or at 120 since it is a nice round number.

Interval	Frequency
120-134	4
135-149	14
150-164	16
165-179	28
180-194	12
195-209	8
210-224	7
225-239	6
240-254	2
255-269	3

A histogram of this data would look like this (note that the horizontal axis is a number line):

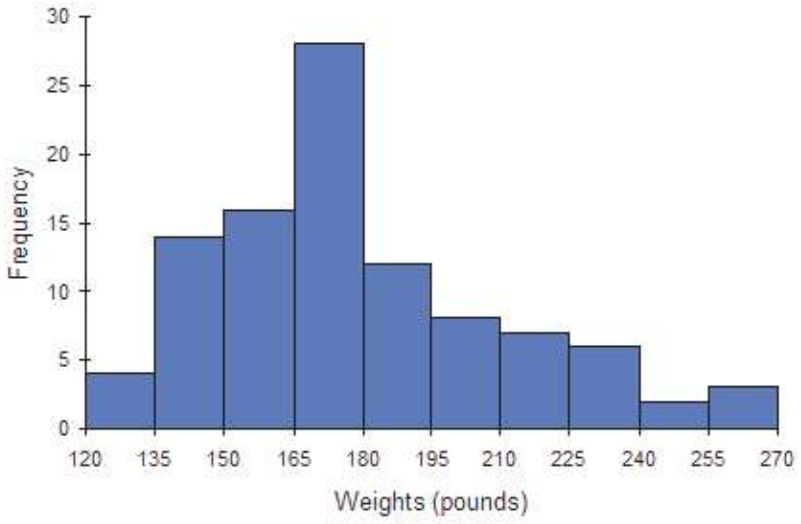


Figure 5.10[80]

In many software packages, you can create a graph similar to a histogram by putting the class intervals as the labels on a bar chart.

[80] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

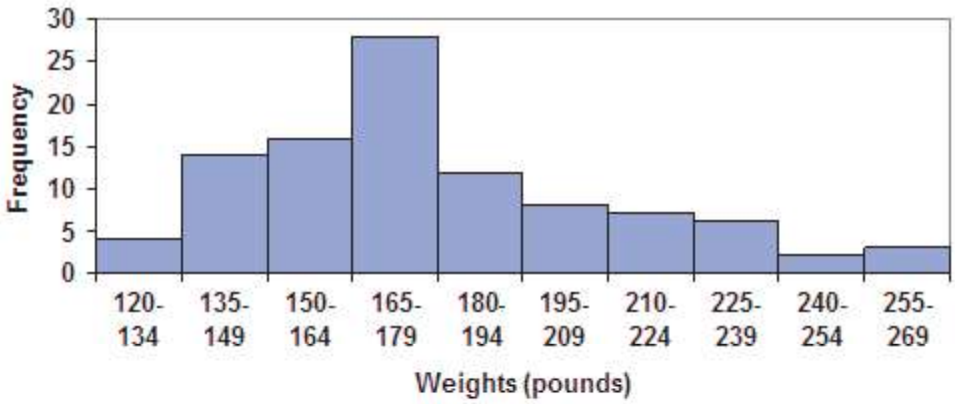


Figure 5.11[81]

Other graph types such as pie charts are possible for quantitative data. The usefulness of different graph types will vary depending upon the number of intervals and the type of data being represented. For example, a pie chart of our weight data is difficult to read because of the quantity of intervals we used.

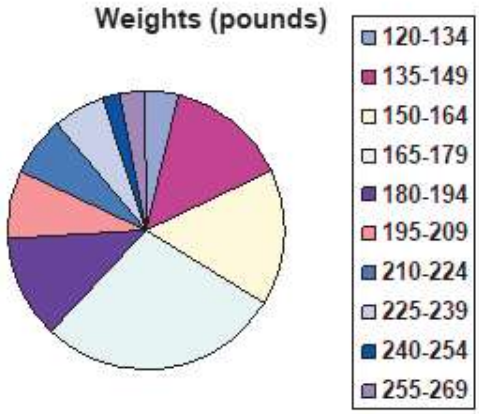


Figure 5.12[82]

Try It Now 8

The total cost of textbooks for the term was collected from 36 students. Create a histogram for this data. (Suggestion: Group them into class intervals of size 55, that is, 140 to 194, 195 to 249, etc.)

$140	$160	$160	$165	$180	$220	$235	$240	$250	$260	$280	$285
$285	$285	$290	$300	$300	$305	$310	$310	$315	$315	$320	$320
$330	$340	$345	$350	$355	$360	$360	$380	$395	$420	$460	$460

[81] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.
[82] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

When collecting data to compare two groups, it is desirable to create a graph that compares quantities.

Example 8

The data below came from a task in which the goal is to move a computer mouse to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial.

Interval (milliseconds)	Frequency small target	Frequency large target
300-399	0	0
400-499	1	5
500-599	3	10
600-699	6	5
700-799	5	0
800-899	4	0
900-999	0	0
1000-1099	1	0
1100-1199	0	0

One option to represent this data would be a comparative histogram or bar chart in which bars for the small target group and large target group are placed next to each other.

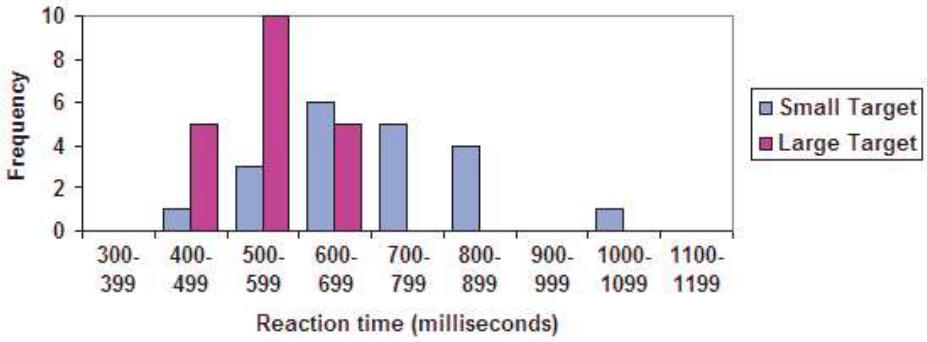


Figure 5.13[83]

[83] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

With the advancement in technology, we are now seeing incredibly sophisticated representations of data. For instance, with the use of various colors, animation, and video, one can visually describe the relationship between wealth and health of various countries over a 200-year span.[84] Software programs such as Excel can easily and quickly produce attractive graphs and charts. You must be careful, however, and remember that visual aids can be created and presented in inaccurate and/or deceptive ways. Because they are powerful tools, they could lead to gross misunderstanding or wrong conclusions. Sometimes this happens because the data itself is skewed (this is what Tim Harford, the author of *The Data Detective*, is describing when he writes, "the visualization is so good but the data are so bad").[85] But even with good data, graphs can be constructed to focus on one aspect of the data or a trend, and infographics could be drawn to provoke certain emotions and stir up our feelings like fear, anger, and confusion.

There are many, many other types of graphs and visual data-representations (both categorial and quantitative), but we will not cover them in this textbook. It is generally a good idea to keep them simple so that the target audience can easily understand the point. Do not make them too fancy or over-complicated; those graphs can be distracting instead of informative.

Exercises 5.3

1. The table below shows scores on a recent history test.
   (a) Complete the frequency table for the test scores.
   (b) Construct a graph of the data.
   (c) Construct a pie chart of the data.

   80, 50, 50, 90, 70, 70, 100, 60, 70, 80, 70, 50
   90, 100, 80, 70, 30, 80, 80, 70, 100, 60, 60, 50

2. A group of adults where asked how many cars they had in their household.
   (a) Complete the frequency table for the car number data.
   (b) Construct a graph of the data.
   (c) Construct a pie chart of the data.

   1, 4, 2, 2, 1, 2, 3, 3, 1, 4, 2, 2,
   1, 2, 1, 3, 2, 2, 1, 2, 1, 1, 1, 2

3. A group of adults were asked how many children they have in their families. The bar graph below shows the number of adults who indicated their number of children.
   (a) How many adults where questioned?

[84] You can see a brief BBC video at https://www.youtube.com/watch?v=jbkSRLYSojo.
[85] Chapter 9 "Remember That Misinformation Can Be Beautiful, too" (pp. 213—237) of *The Data Detective* has many other examples where visual aids are used as manipulative tools. He adds, "Any vivid graphic has the potential to go viral, whether true, false, or a bit of both" (p. 225).

(b) What percentage of the adults questioned had 0 children?

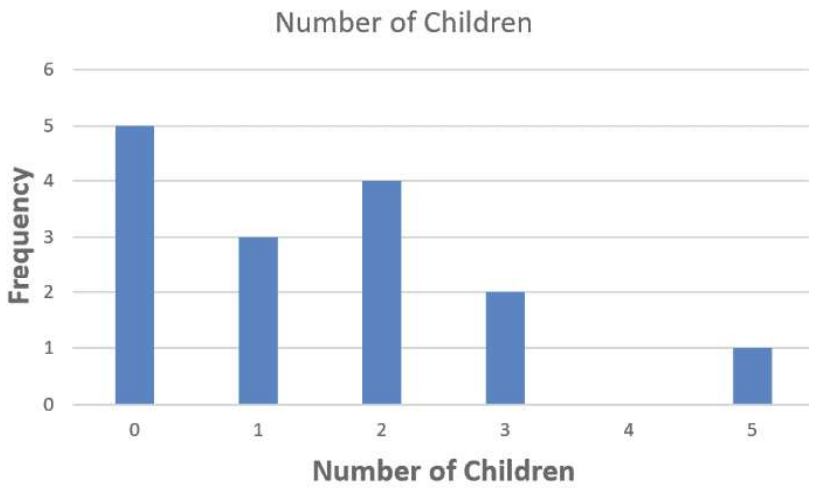


Figure 5.14[86]

4. Jasmine was interested in how many days it would take an order from an online retailer to arrive at her door. The graph below shows the data she collected.
   (a) How many orders did she place?
   (b) What percentage of the orders arrived in one day?

Figure 5.15[87]

[86] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.
[87] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

5. The bar graph below shows the percentage of students who received each letter grade on their last English paper. The class contains 20 students. How many students earned an A on their paper?

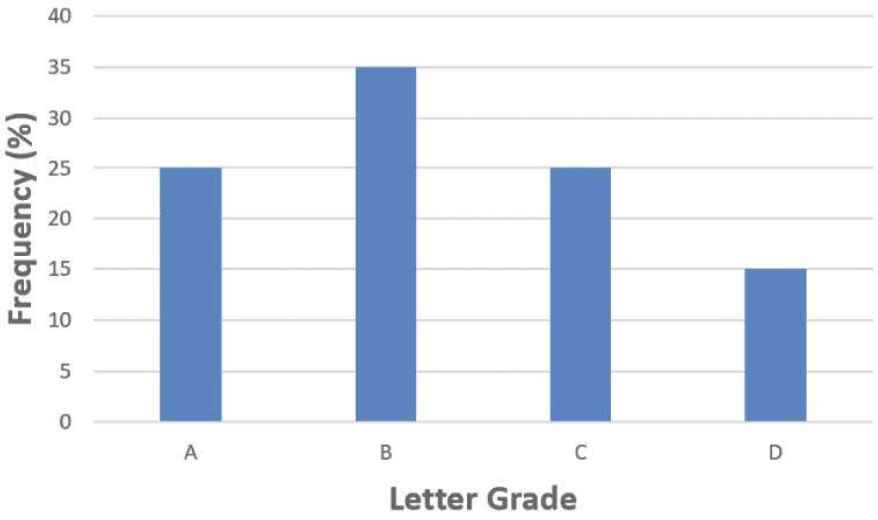


Figure 5.16[88]

6. Kori categorized her spending for this month into four categories: food, rent, fun, and other. The percentage she spent in each category is pictured here. If she spent a total of $2600 this month, how much did she spend on rent?

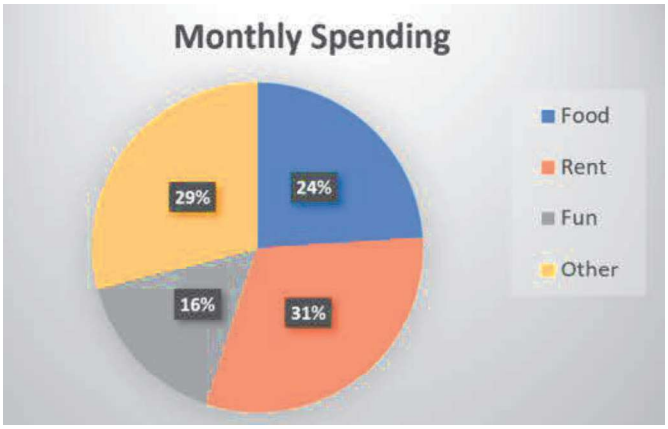


Figure 5.17[89]

[88] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.
[89] Describing Data, by David Lippman, is licensed under CC BY-SA 3.0 US.

5.4 Numerical Measures of Center and Variation

It is often desirable to use a few numbers to summarize a distribution of data. One important aspect of a distribution is where its center is located. Measures of central tendency are discussed first. A second aspect of a distribution is how spread out it is. In other words, how much the data in the distribution vary from one another. The second part of this section describes measures of variation.

Measures of Central Tendency

Let us begin by trying to find the most "typical" value of a data set. Note that we just used the word "typical" although in many cases you might think of using the word "average." We need to be careful with the word "average" as it could mean different things to different people. One of the most common uses of the word "average" is what mathematicians and statisticians call the **arithmetic mean** (or just "mean" for short). "Arithmetic mean" sounds rather fancy, this is what most people think of when they use the word "average."

Mean

The mean of a set of data is the sum of the data values divided by the number of values.

Example 1

Marci's exam scores for her last chemistry class were 79, 86, 82, and 94. Find the mean.

[Solution]

The mean of these values is

$$\frac{79 + 86 + 82 + 94}{4} = 85.25.$$

Example 2

The number of touchdown (TD) passes thrown by each of the 32 teams in the National Football League in the 2019 season are shown below. Find the mean.

28, 20, 18, 21, 16, 22, 33, 20, 24, 30, 22, 30, 22, 27, 29, 30,
24, 19, 28, 26, 17, 25, 22, 22, 37, 18, 36, 31, 18, 27, 29, 26.

[Solution]

Adding these values, we get 797 total TD passes. Dividing it by 32, the number of data values, we get $797 \div 32 \approx 24.9$ TD passes.

It would be correct for us to report that "the mean number of touchdown passes in the NFL in the 2019 season was 24.9 passes per team," but it is not uncommon to see the more casual word "average" used in place of "mean."

Try It Now 9

The prices of a jar of peanut butter at 5 stores were $3.29, $3.59, $3.79, $3.75, and $3.99. Find the mean price.

Example 3

One hundred families in an organization reported their annual household income (to the nearest $5000). The results are summarized in a frequency table below. Find the mean.

Income (thousands of dollars)	Frequency
15	6
20	8
25	11
30	17
35	19
40	20
45	12
50	7

[Solution]

Calculating the mean by hand could get tricky if we try to type in all 100 values:

$$\frac{\overbrace{15+\cdots 15}^{\text{6 terms}}+\overbrace{20+\cdots+20}^{\text{8 terms}}+\overbrace{25+\cdots+25}^{\text{11 terms}}+\cdots}{100}$$

We could calculate this more easily by noticing that adding 15 to itself six times is the same as 90. Using this simplification, we get

$$\frac{15\cdot 6+20\cdot 8+25\cdot 11+30\cdot 17+35\cdot 19+40\cdot 20+45\cdot 12+50\cdot 7}{100}=\frac{3390}{100}$$
$$=33.9.$$

Since the income in this table is expressed in "thousands of dollars," we multiply this mean by $1000 to find our answer. The mean household income is $33,900.

Example 4

Extending off the last example, suppose a new family joins this organization, and they have a household income of $5 million ($5000 thousand). Now, find the mean.

[Solution]

Note that all we need to do is to add 5000 to our numerator and 1 to our denominator. Our new mean is as follows:

$$\frac{3390 + 5000 \cdot 1}{100 + 1} = \frac{8390}{101} \approx 83.07.$$

While $83,070 is the correct mean household income, it no longer represents a "typical" value. This illustrates an **outlier**.

Outlier

An outlier is an extreme value, much bigger or much smaller than most other values in a given data set. An outlier can affect the mean significantly.

The mean of a set of data is the sum of the data values divided by the number of values.

In situations like this, where one value is much bigger or smaller than most other value in the data, there is a better measure of center than the mean.

Median

The median of a set of data is the value **in the middle** when the data is in order.

To find the median, begin by listing the data **in order** from smallest to largest or largest to smallest. If the number of data values, N, is **odd**, then the median is the middle data value. (This can be found by rounding N/2 up to the next whole number.) If the number of data values is **even**, there is no one middle value, so we find the mean of the two middle values (values N/2 and N/2 + 1).

Example 5

Returning to the football touchdown data, we would start by listing the data in order:

37, 36, 33, 31, 30, 30, 30, 29, 29, 28, 28, 27, 27, 26, 26, 25,
24, 24, 22, 22, 22, 22, 22, 21, 20, 20, 19, 18, 18, 18, 17, 16.

(If you are wondering, the Ravens had 37, and the Broncos had 16.)

[Solution]

Since there are 32 data values, an even number, the median will be the mean of the two middle numbers, the 16th and the 17th data values ($32 \div 2 = 16$). The 16th value is 25 while the 17th is 24, so the median number is

$$\frac{24 + 25}{2} = 24.5.$$

Notice that for this data, the median is fairly close to the mean we calculated earlier, 24.9.

Example 6

Find the median of quiz scores: 5, 10, 8, 6, 4, 8, 2, 5, 7.

[Solution]

We start by listing the data **in order**: 2, 4, 5, 5, 6, 7, 8, 8, 10. Since there are 9 data values, an odd number, there is one middle number, the 5th value (the next whole number after $9 \div 2 = 4.5$ is 5). So we find the median is 6. Note that the value 6 is the 5th from the beginning and also the 5th from the end of this ordered list.

Try It Now 10

The prices of a jar of peanut butter at 5 stores were $3.29, $3.99, $3.79, $3.75, and $3.59. Find the median price. Do not forget to re-arrange the values if necessary.

Example 7

Let us return now to our original household income data. Find the median value.

Income (thousands of dollars)	Frequency
15	6
20	8
25	11
30	17
35	19
40	20
45	12
50	7

[Solution]

Here we have 100 data values. (If we did not already know that, we could find it by adding the frequencies.) Since 100 is an even number, we need to find the mean of the middle two data values, the 50^{th} and the 51^{st} data values. To find these, we start counting up from the bottom:

- There are 6 data values of $15, so values 1 to 6 are $15,000;
- The next 8 data values are $20, so values 7 to (6+8) =14 are $20,000;
- The next 11 data values are $25, so values 15 to (14+11) =25 are $25,000;
- The next 17 data values are $30, so values 26 to (25+17) =42 are $30,000;
- The next 19 data values are $35, so values 43 to (42+19) =61 are $35,000.

From this we can tell that values 50 and 51 will be $35,000, so the mean of these two values is $35,000. The median income in this organization is $35,000.

Example 8

Suppose that we add in that new family with a $5 million household income. What would the new median be?

[Solution]

There will be 101 data values, and the 51^{st} value will be the median. As we discovered in the last example, the 51st value is $35,000. Notice that the new family did not affect the median at all in this case. **The median is not swayed as much by outliers as the mean is.**

This is the main reason the median, not the mean, is used to represent certain data, such as the average household income or home price.

In addition to the mean and the median, there is one other common measurement of the "typical" value of a data set: the **mode.**

Mode

The mode is the value in the data set that occurs most frequently.

The mode is fairly useless with quantitative data like weights or heights where there are a large number of possible values. The mode is most commonly used for categorical data, for which median and mean cannot be computed.

Example 9

In our vehicle color survey, we collected the following data. Find the mode.

Color	Frequency
Blue	3
Green	5
Red	4
White	3
Black	2
Grey	3

[Solution]

For this data, “Green” is the mode since it is the data value that occurred the most frequently.

It is possible for a data set to have more than one mode (if several categories have the same frequency).

Try It Now 11

Reviewers were asked to rate a product on a scale of 1 to 5. Find the mean, the median, and the mode ratings.

Rating	Frequency
1	4
2	8
3	7
4	3
5	1

Measures of Variation

Consider these three sets of scores from a recent 10-point quiz.

- Section A: 5, 5, 5, 5, 5, 5, 5, 5, 5, 5
- Section B: 0, 0, 0, 0, 0, 10, 10, 10, 10, 10
- Section C: 4, 4, 4, 5, 5, 5, 5, 6, 6, 6

All three of these data sets have a mean of 5 and median of 5, yet the sets of scores are clearly quite different. In Section A, everyone had the same score; in Section B half the class got no points and the other half got a perfect score. Section C was not as consistent as Section A, but not as widely varied as Section B.

In addition to the mean and median, which are measures of the "typical" or "middle" value, we also need a measure of how "spread out" or varied each data set is. There are several ways to measure this "spread" of the data. The first is the simplest and is called the **range.**

Range

The range is the difference between the maximum value and the minimum value of the dataset.

Example 10

Using the quiz scores from above, find the range in each section.

[Solution]

- For Section A, the range is 0 since both maximum and minimum are 5 and 5 – 5 = 0.
- For Section B, the range is 10 since 10 – 0 = 10.
- For Section C, the range is 2 since 6 – 4 = 2.

In this example, the range seems to be revealing how spread out the data is. However, suppose we add a fourth section, Section D, with scores 0, 5, 5, 5, 5, 5, 5, 5, 5, 10. This section also has a mean and median of 5. The range is 10, yet this data set is quite different from Section B. To better illuminate the differences, we will have to turn to a more sophisticated measure of variation called the **standard deviation**, which is far more important in the formal field of statistics.

Standard Deviation

The standard deviation is a measure of variability based on how far each data value deviates, or is different, from the mean. (It is approximately the average distance a typical value is from the mean.) A few important characteristics:

- Standard deviation is never negative. Standard deviation is zero only if all the data values are equal.
- Standard deviation will get larger as the data values spreads out more.
- Standard deviation has the same unit as the original data.
- Standard deviation, like the mean, is highly influenced by outliers.

Standard deviation is rather complicated. Let me explain below, step by step, how to calculate it.

Suppose there are n values in a set of data. The difference between each value and the mean is called the **deviation**. A deviation can be positive, negative, or 0 (it is 0 only if the particular value is equal to the mean). **Square** each deviation and find their sum. When this sum is divided by $n - 1$, that quotient is called the **variance** of the data set. The **standard deviation** is the positive square root of the variance.

This sounds super-complicated, right? Yes, and because of the complexity, standard deviations are almost always calculated by computer. However, it is worth learning how these calculations are carried out by hand (for a small data set) so you can have a "feel" for how this works and what this measure represents.

Example 11

Suppose that the quiz scores from yet another section, Section E, were as follows:

$$\{2, 3, 4, 4, 5, 6, 7, 9, 10, 10\}.$$

Find the standard deviation of this data set.

[Solution]

There are ten values, and their sum is 60. Hence, the **mean** is $60 \div 10 = 6$. Next, find the **deviation** for each value as shown, and then **square each deviation**. We can construct a table as shown below. (You may think that, since the deviations express how far each data value is from the mean, we should just find the mean of the deviations to measure the variation. The problem is that the deviations, by definition, will always add up to 0, so the mean of the deviations will always be 0 and, therefore, cannot be used to express the variation.)

Data value	Deviation: data value - mean	Deviation squared
2	$2 - 6 = -4$	$(-4)^2 = 16$
3	$3 - 6 = -3$	$(-3)^2 = 9$
4	$4 - 6 = -2$	$(-2)^2 = 4$
4	$4 - 6 = -2$	$(-2)^2 = 4$
5	$5 - 6 = -1$	$(-1)^2 = 1$
6	$6 - 6 = 0$	$(0)^2 = 0$
7	$7 - 6 = 1$	$(1)^2 = 1$
9	$9 - 6 = 3$	$(3)^2 = 9$
10	$10 - 6 = 4$	$(4)^2 = 16$
10	$10 - 6 = 4$	$(4)^2 = 16$

Now, add up the squared deviations to get $16 + 9 + 4 + 4 + 1 + 0 + 1 + 9 + 16 + 16 = 76$. You would think that we want to divide this by the number of scores, n (which is 10 here), to find the mean of the deviations squared. But if the data set represents a sample (as it almost always does), we divide the sum by $n - 1$(which is 9 here) instead. This may be the trickiest point (see below for further explanations). The quotient is called the **variance** (or **sample variance**); here, it is $76 \div 9 = 8.444$. The standard deviation, which is the positive square root of the variance, is $\sqrt{8.444} \approx 2.91$. The standard deviation for the quiz scores in Section E is, therefore, 2.91.

More precisely, what we just calculated is referred to as the **sample standard deviation** because we used $n - 1$ in the denominator for the variance. If, instead, we used n, we would get what is called the **population standard deviation**. We almost always use **sample standard deviation**, assuming that our data set is just one of the many samples taken from a large pool of possible data sets. For example, here, our data set (as well as the other data sets for Sections A, B, C, and D above) is simply one sample of how 10 students in one section can score on one quiz. Where a data set represents the entire population (remember the difference between a sample and a population?), one can use the **popular standard deviation** to describe the population distribution (but this is rare).

Why use n for the population and $n-1$ for the sample? The difference is quite technical and requires advanced mathematics for a full explanation, but the idea is this: dividing by $n-1$ results in a slightly larger variance than dividing by n. When we have only a sample (a subset of the population), the chances are that the particular sample does not represent the entire distribution of the population, which is generally more widespread. Hence, the sample variance is slightly smaller than the population variance, and the same goes for their square roots, the standard deviations. So dividing by $n-1$ is a way to compensate for the "bias" that occurs by using only a little part of the population instead of the whole population. It turns out that theoretical/mathematical statistics can actually prove that $n-1$ is precisely what we need for an **unbiased estimator** for the parameter, the true standard deviation for the population.

If you are using spreadsheet software such as Excel, there are functions to calculate both of these standard deviations: "*stdev.p*" for population standard deviation and "*stdev.s*" for sample standard deviation. But for the purposes of this course, we will use **sample standard deviation** only, so always use $n-1$.

As you can see, it takes multiple steps to calculate the standard deviation of a data set. Here is a brief summary.

To Compute Standard Deviation (for a data set with *n* values):

1. Find the deviation of each data value from the mean.
2. Square each deviation.
3. Add all the squared deviations.
4. Divide by $n-1$ to find the **variance**.
5. Find the square root of the variance to calculate the **standard deviation**.

Example 12

Find the standard deviation for Section B above, with the scores as follows:

0, 0, 0, 0, 0, 10, 10, 10, 10, 10

[Solution]

We first see that the total sum of the 10 values (scores) is 50, so the mean is 5. Using an abbreviated table can help our calculations for the standard deviation:

Data value	Deviation: data value - mean	Deviation squared
0	$0-5=-5$	$(-5)^2=25$
10	$10-5=5$	$(5)^2=25$

Now, of course, each of these values occurs five times, so what we have on the last column is really ten occurrences of 25. So the sum of the deviation squared is $25 \times 10 = 250$. We then divide this sum by 9 and take the square root:

$$\sqrt{\frac{250}{9}} \approx 5.27.$$

Notice that the standard deviation of this data set is much larger than that of section E since the data in this set is more spread out. (Each score is actually 5 away from the mean.)

For comparison, the standard deviations of all five sections are as follows. You can easily see that the standard deviation is larger when the data values are more spread apart.

Section	Standard Deviation
Section A: 5, 5, 5, 5, 5, 5, 5, 5, 5, 5	Standard deviation: 0
Section B: 0, 0, 0, 0, 0, 10, 10, 10, 10, 10	Standard deviation: 5.27
Section C: 4, 4, 4, 5, 5, 5, 5, 6, 6, 6	Standard deviation: 0.82
Section D: 0, 5, 5, 5, 5, 5, 5, 5, 5, 10	Standard deviation: 2.36
Section E: 2, 3, 4, 4, 5, 6, 7, 9, 10, 10	Standard deviation: 2.91

Try It Now 12

The prices of a jar of peanut butter at 5 stores were: $3.29, $3.59, $3.79, $3.75, and $3.99. Find the standard deviation of the prices.

By the way, in finding the standard deviation, you *squared* the differences (called **deviations**), added them up, and later took a square root. If this reminded you of the Pythagorean Theorem or the Distance Formula from your algebra class, that is no accident; theoretically the idea of standard deviation is related to distance in space.

Quartiles of a Data Set

We now present a simple yet convenient way to express distributions; it is called **quartiles**. Where standard deviation is a measure of variation based on the mean, quartiles are based on the median.

Quartiles

Quartiles are values that divide the data in **quarters**.
The first quartile (**Q1**) is the value such that 25% of the data values are below it; the third quartile (**Q3**) is the value such that 75% of the data values are below it. The second quartile (**Q2**) is the same as the median, the value such that 50% of the data values are below it.

These divide the data into quarters; 25% of the data is between the minimum and Q1, 25% is between Q1 and the median Q2, 25% is between the median Q2 and Q3, and 25% is between Q3 and the maximum value.

While quartiles are not a 1-number summary of variation like standard deviation, the quartiles are used with the minimum and the maximum values to form a **5-number summary** of the data.

Five-Number Summary

The five-number summary takes this form: Minimum, Q1, Median (Q2), Q3, Maximum

To find the first quartile, we need to find the data value such that 25% of the data is below it. But remember that this is the same as the "median" of the first (lower) half of the data set. So an easy way to find the first quartile is to split the original data set in half and find the median of the lower half. The third quartile is similar; find the median of the second (higher) half of the data set. (Warning: The precise definitions and method of finding quartiles vary slightly from book to book. You may see slightly different explanations in other resources.)

Example 13

Suppose that the heights (in inches) of nine athletes, sorted from shortest to tallest are as follows. Find the five-number summary.

59, 60, 62, 64, 66, 67, 69, 70, 72

[Solution]

Note the median is 66. The lower half, {59, 60, 62, 64}, has the median of 61 (the mean of the two middle ones), and the higher half, {67, 69, 70, 72}, has the median of 69.5. So Q1 is 61, Q2 is 66, and Q3 is 69.5. The five-number summary includes the minimum and the maximum numbers as well, so it will be "59, 61, 66, 69.5, 72."

Example 14

Suppose that the heights (in inches) of eight musicians, sorted from shortest to tallest are as follows. Find the five-number summary.

57, 60, 63, 64, 66, 67, 68, 69

[Solution]

Note the median in this case is 65 (the mean of the two numbers in the middle). Then the lower half has the median 61.5 (the mean of 60 and 63). The higher half has the median 67.5. So, Q1 is 61.5, Q2 is 65, and Q3 is 67.5. The five-number summary is "57, 61.5, 65, 67.5, 69."

Try It Now 13

The total cost of textbooks for the term was collected from 36 students. Find the 5-number summary of this data.

$140	**$160**	**$160**	**$165**	**$180**	**$220**	**$235**	**$240**	**$250**	**$260**	**$280**	**$285**
$285	**$285**	**$290**	**$300**	**$300**	**$305**	**$310**	**$310**	**$315**	**$315**	**$320**	**$320**
$330	**$340**	**$345**	**$350**	**$355**	**$360**	**$360**	**$380**	**$395**	**$420**	**$460**	**$460**

Of course, with a relatively small data set, finding a five-number summary is a bit meaningless since the summary contains almost as many values as the original data. However, for **visualizing data**, there is a very nice graphical representation of a 5-number summary called a **box plot**, or **box and whisker graph**.

Box Plot

A box plot is a graphical representation of a 5-number summary, drawn as follows:

A number line is first drawn. A box is drawn from the first quartile to the third quartile, and a line is drawn through the box at the median. "Whiskers" are extended out to the minimum and maximum values.

Example 15

Draw the box plot for a data set whose five-number summary is 59, 62, 66, 69, 72.

[Solution]

Given a data set, normally we need to find the median, Q1, and Q3 first. However, in this case, since the five-number summary is given to us, we can go directly to drawing the box plot.

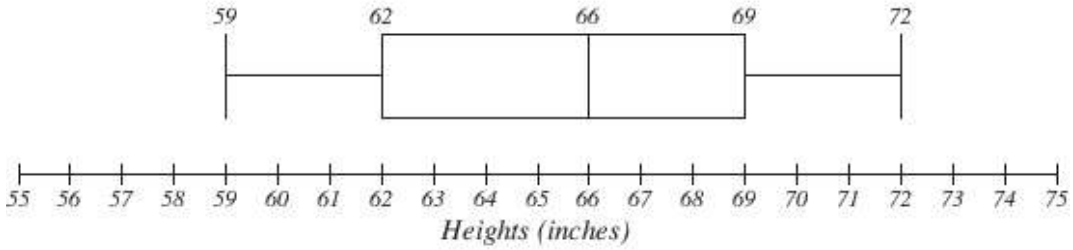


Figure 5.18[90]

Example 16

Draw the box plot for a data set listing the household incomes of the members of an organization. The five-number summary is 15, 27.5, 35, 40, 50 (in thousands of dollars).

[Solution]

Again, we do not need to find the quartiles since they are given to us. Below is the box plot for this data set. You can easily see that the income distribution is not symmetric.

[90] Describing Data by David Lippman, is licensed under CC BY-SA 3.0 US

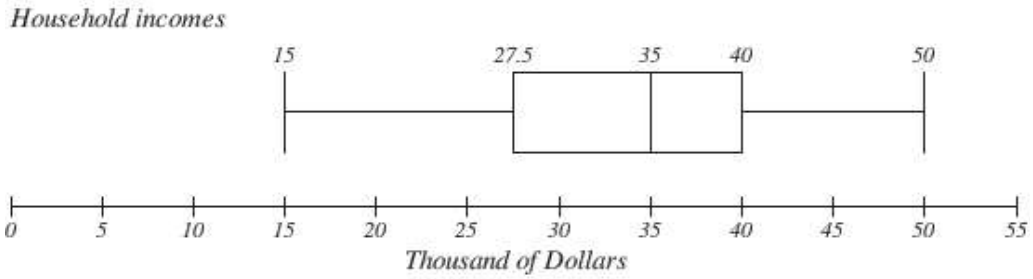


Figure 5.19[91]

Try It Now 14

Create a boxplot based on the textbook price data from Try It Now 8 (or Try It Now 13).

Box plots are particularly useful for comparing data from two populations. The following examples show the power and convenience of box plots. I hope you agree that these visuals tell us a lot about the data.

Example 17

The box plots of service times for two fast-food restaurants are shown below. What do you know about these two fast-food restaurants? Which one would you prefer to go to if they serve the same food at the same price?

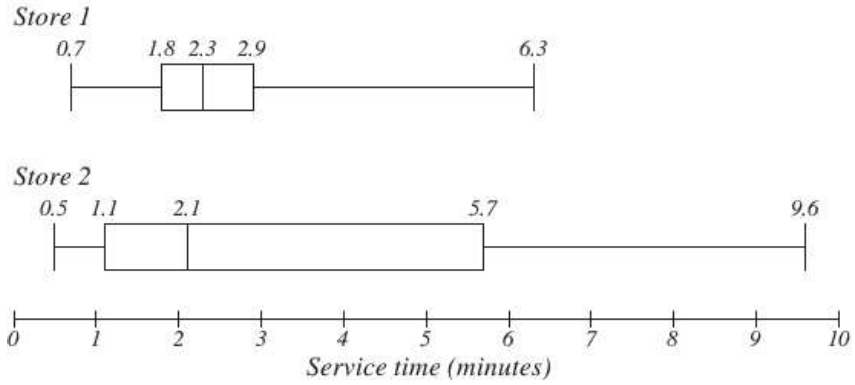


Figure 5.19[92]

[Solution]

While Store 2 had a slightly shorter median service time (2.1 minutes vs. 2.3 minutes), Store 2 is less consistent, with a wider spread of the data.

At Store 1, about 75% of customers were served within 2.9 minutes while at Store 2, about 75% of customers were served within 5.7 minutes. At Store 1, everyone in the

[91] Describing Data by David Lippman is licensed under CC BY-SA 3.0 US
[92] Describing Data by David Lippman is licensed under CC BY-SA 3.0 US

sample was served within 6.3 minutes while at Store 2, some people had to wait for almost 10 minutes.

Which store should you go to in a hurry? That depends upon your opinions about luck. A quarter of the customers at Store 2 had to wait between 5.7 and 9.6 minutes, but again, a quarter of the customers at the same store were served within 1.1 minutes.

Example 18

The boxplot below is based on the birth weights of infants with severe idiopathic respiratory distress syndrome (SIRDS). The boxplot is separated to show the birth weights of infants who survived and those that did not. Compare the two groups and make some observations.

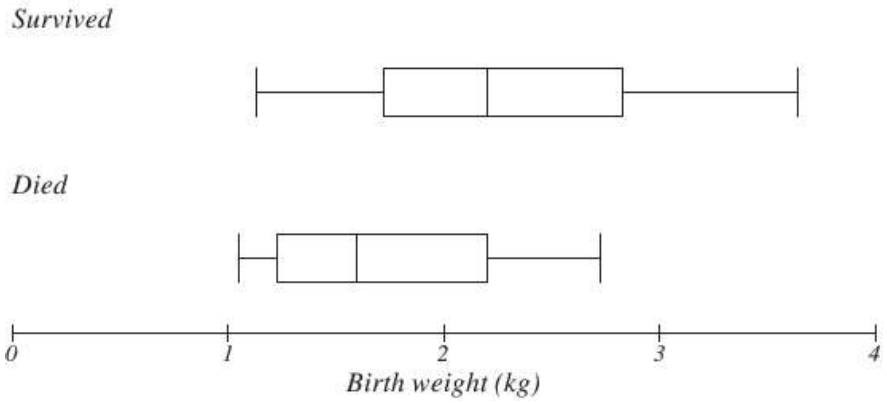


Figure 5.20[93]

[Solution]

The boxplots reveal that the birth weights of the infants that died appear to be, overall, smaller than the weights of infants that survived. In fact, we can see that the median birth weight of infants that survived is the same as the third quartile of the infants that died.

Similarly, we can see that the first quartile of the survivors is larger than the median weight of those that died, meaning that over 75% of the survivors had a birth weight larger than the median birth weight of those that did not survive.

Looking at the maximum value for those that died and the third quartile of the survivors, we can see that over 25% of the survivors had birth weights higher than the heaviest infant that died.

Note, however, this is *not* to say that less weight is the cause of death for infants with SIRDS. Remember, association—no matter how strong it may be—cannot ever

[93] Describing Data by David Lippman is licensed under CC BY-SA 3.0 US.

establish causation in and of itself. In this case, the association may be explained by some **common underlying cause**, which makes the babies smaller and the syndrome more severe. Further studies are needed.

Exercises 5.4

1. Students in a class reported the number of minutes they spend commuting from home to school as follows: 17, 50, 33, 52, 10, 4, 21, 18, 15.
   (a) Find the mean.
   (b) Find the median.
   (c) Find the range.

2. Members of a club reported the number of vacations they have taken in the last 5 years as follows: 5, 12, 10, 7, 2, 0, 9, 10, 4, 10.
   (a) Find the mean.
   (b) Find the median.
   (c) Find the mode.
   (d) Find the range.

3. A group of diners were asked how much they would pay for a meal. Their responses were as follows: $7.50, $8.25, $9.00, $8.00, $7.25, $7.50, $8.00, $7.00.
   (a) Find the mean.
   (b) Find the median.
   (c) Find the range.

4. You recorded the time (in seconds) it took for 7 participants to solve a puzzle. Their times were as follows: 15.2, 18.8, 19.7, 20.2, 21.8, 22.1, 29.4.
   (a) Find the mean to two decimal places.
   (b) Find the median.
   (c) Find the range.
   (d) Write the five-number summary for this data.
   (e) Find the standard deviation of this data.

5. The following is the frequency chart of a five-star review on an online retailer website for a certain book you are interested in. Find the mean, the median, and the mode rating.

Rating	Frequency
0	32
1	55
2	36
3	89
4	102
5	46

6. In a geography class, students were asked how many overseas trips they have taken in their life, and the following is the frequency chart summarizing the responses. Find the mean, the median, the mode, and the range.

# of Overseas Trips	Frequency
0	23
1	31
2	18
3	7
4	3
7	1
10	2

7. Which data set below has the bigger mean? Which has the bigger standard deviation?

$$A = \{20, 30, 40, 50, 60\} \text{ or } B = \{30, 40, 50, 60, 70\}$$

8. For the data set below, find the mean, the median, the range, and the standard deviation.

$$\{42, 42, 42, 42, 42, 42, 42\}$$

9. Create a data set in which the first quartile (Q1) is 10 and the third quartile (Q3) is 20.

10. The box plot below shows salaries for actuaries (business professionals who analyze and manage risks, who often work with underwriters in insurance companies) and CPAs. Andrew makes the median salary for an actuary. Kelsey makes the first-quartile salary for a CPA. Who makes more money? How much more?

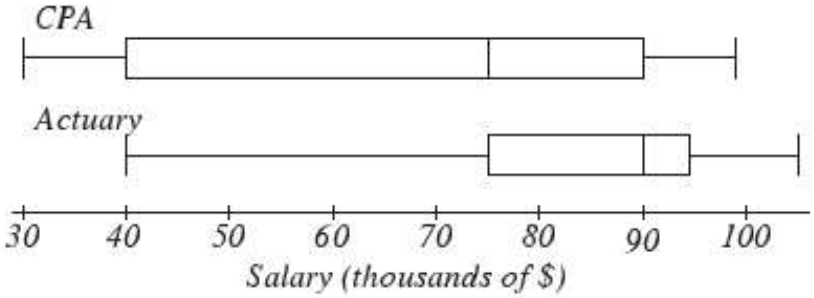


Figure 5.21[94]

11. Referring to the boxplot above, what percentage of actuaries makes more than the median salary of a CPA?

12. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected. Create comparative boxplots of the data.

    - [Researcher 1]
      3, 4, 11, 15, 16, 17, 22, 44, 37, 16, 14, 24, 25, 15, 26, 27, 33, 29, 35, 44
      13, 21, 22, 10, 12, 8, 40, 32, 26, 27, 31, 34, 29, 17, 8, 24, 18, 47, 33, 34
    - [Researcher 2]
      3, 14, 11, 5, 16, 17, 28, 41, 31, 18, 14, 14, 26, 25, 21, 22, 31, 2, 35, 44
      23, 21, 21, 16, 12, 18, 41, 22, 16, 25, 33, 34, 29, 13, 18, 24, 23, 42, 33, 29

13. Take a look at the three data sets below.

$$A = \{10, 20, 30, 40, 50\}$$
$$B = \{0, 15, 30, 45, 60\}$$
$$C = \{28, 29, 30, 31, 32\}$$

    Note that each set consists of 5 numbers, has the same mean (30), and different standard deviations. Use this as an example to make up three data sets (different from these above) with 5 numbers each that have the same mean but different

[94] Describing Data by David Lippman is licensed under CC BY-SA 3.0 US

standard deviations. Explain why the standard deviations in your sets are different from each other.

14. See the previous question. Now, make up three data sets with 5 numbers each that have the same mean but different medians.

15. See the last two questions. Now, make up three data sets with 5 numbers each that have the same median but different means.

16. You asked five people walking down Hollywood Boulevard how much money they made last year. The answers were as follows: $0, $0, $0, $0, $5 million.
    (a) What is the mean of the five data values you got?
    (b) What is the median?
    (c) What is the mode?
    (d) Is there an outlier? If so, which one(s)?
    (e) Which measure of central tendency, if any, is appropriate to describe this data set? Why?

Figure 5.22[95]

[95] Image by Diliff is licensed under CC BY-SA 3.0

5.5 Normal Distribution[xii]

You have probably seen a "**bell-shaped curve**" before. It turns out that many data sets, particularly when the sample size is sufficiently large, are distributed symmetrically, with most data values around the center (mean) and few outliers, resulting in a bell-shaped curve. This is called a "**normal distribution**."

A Normal Distribution

A normal distribution is a type of data distribution which is **symmetric**, with the mean at the center, and is characterized by a bell-shaped curve.

Is the normal distribution (bell-shaped curve) with 0 as the mean and 1 as the standard deviation?

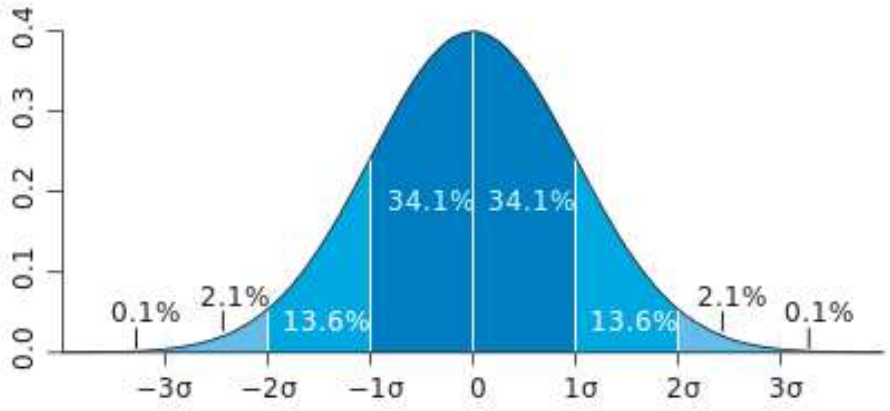


Figure 5.23[96]

Because this distribution is based on a bell-shaped function extensively studied by the German mathematician Carl F. Gauss, the normal distribution is sometimes referred to as the **Gaussian distribution** (and the associated bell-shaped curve as the **Gaussian curve**). The technical definition is much more rigorous; given the mean μ and the standard deviation $\sigma > 0$, it is the continuous random variable defined by the probability density function

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

Of course, you will *not* need to know this complicated function, but it is quite remarkable that this function, extremely important in statistics (and thus in many business and policy decisions in real life) involves irrational numbers π (which you studied in geometry) and e (which you studied in compound interest).

[96] Image by M.W. Taves is licensed under CC BY 2.5

As mentioned above, many quantitative data are **normally distributed**. For instance, the weights of all the people in the world are likely to be normally distributed, and so are their heights. Normal distributions can describe all sorts of data, including people's cholesterol level, blood pressure, duration of pregnancy, IQ (intelligence quotient) score, credit score, standardized test scores, and many more. For instance, the pregnancy (gestation) period for humans is known to have an average of 266 days (38 weeks) with a standard deviation of 16 days (approximately). The average IQ score is defined as 100 with a standard deviation of 15 (with some variations). The SAT and other standardized test scores are defined in similar ways.

But what does it mean, and how does this information help us? When data values are distributed normally, given its mean and standard deviation, we can then standardize the normal distribution and calculate various percentages related to scores/values. For example, given that the IQ scores have a mean of 100 and a standard deviation of 15, one can conclude that only about 2.3% of the population have an IQ of 130 or higher.

In fact, if you assume that data values are normally distributed, and if you know the mean and the standard deviation, there is a way to find out what percent of the data values are below a given number x (this is called the percentile for x). This is how, for instance, a pediatrician can tell you what percentile an infant's weight is at, compared to all babies of the same age.

Here is how this works. First, let us define the "**Standard Normal Distribution**."

The Standard Normal Distribution

The standard normal distribution is the normal distribution (bell-shaped curve) with 0 as the mean and 1 as the standard deviation.

It turns out that there is a simple way to "**convert**" any normal distribution to the **standard** one. It involves what is known as the "z-score."

The $z-$score measures how many standard deviations a score is away from the mean. The $z-$score of a value x in a distribution whose mean is μ and whose standard deviation σ is given by

$$z = \frac{x - \mu}{\sigma}.$$

Since σ is always positive, z will be positive when x is greater than μ and negative when x is less than μ. Also, $z = 0$ only when the value x is equal to the mean.

Example 1

On a standardized arithmetic test the mean was 65 and the standard deviation was 10. If Robert scored 81, what was his $z-$score?

[Solution]

Plugging his score of 81 as x, $\mu = 65$, and $\sigma = 10$, we get

$$z = \frac{x - \mu}{\sigma} = \frac{81 - 65}{10} = 1.6.$$

This means Robert's score was 1.6 standard deviations above the mean.

Do not let algebra or the Greek letters scare you. This is not hard! Think this way. If the average (mean) is 65 and the standard deviation is 10, then obviously the score 75 is exactly one standard deviation above the mean, so the z -score for 75 is 1.0. Similarly, the score 45 is exactly 20 below the mean, which implies that it is 2 standard deviations below the mean, i.e., $z = -2.0$. In the above example, Robert's score was 16 points above the mean, so 16/10 = 1.6 was his z -score, meaning that his score is 1.6 standard deviations above the mean.

Example 2

On a college exam, the mean was 70 and the standard deviation was 8. Helen's z-score was −1.5. What was her exam score?

[Solution]

This time, we need to solve for Helen's exam score x, given the corresponding z-score.

$$z = \frac{x - \mu}{\sigma} = \frac{x - 70}{8}.$$

Since we know that $z = -1.5$, we have the equation $z = -1.5 = \frac{x-70}{8}$. Multiplying both sides by 8, we get

$$x - 70 = -12,$$

which gives us that $x = 70 - 12 = 58$. So Helen's score was 58.

Now, if you prefer not to do all this algebra, there is another way to do this—involving more reasoning, perhaps. $z = -1.5$ means that Helen's score was 1.5 standard deviations below the mean. Well, the standard deviation was 8, so Helen's score was $1.5 \times 8 = 12$ points below the mean of 70. That is, of course, 58. Here is the key: *understand what the z-score stands for.*

Now we will see how z−scores are used to determine the **percentile** and related probability.

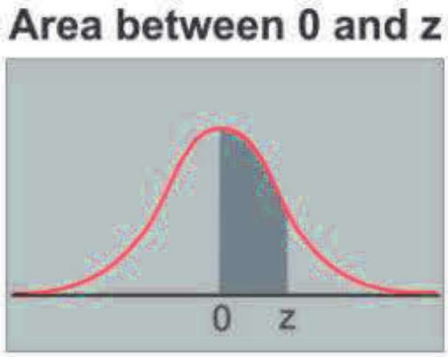


Figure 5.24[97]

z-value	Probability (area)
0.00	0.0000
0.50	0.1915
1.00	0.3413
1.50	0.4332
2.00	0.4772
2.50	0.4938
3.00	0.4987 (almost 50%)

The graph above shows the standard normal distribution (with the mean 0). Each decimal number in the **table** represents the **probability (percentage)** that a data value is between 0 and the corresponding z–score.[98] For instance, about 34% of all data values lie between $z = 0$ (mean) and $z = 1.0$. Using **symmetry**, then, we can conclude that about 68% (twice 34%) of all data values are within 1 standard deviation of the mean ($-1 < z < 1$). The standard normal distribution has been carefully analyzed using advanced mathematics, and the probability (percentage) can be calculated for any z–score. This table just shows a few of those numbers. It is known, for instance, that 99.74% of values lie within 3 standard deviations of the mean ($-3 < z < 3$), and virtually every data value is within 4 standard deviations of the mean.

Example 1

Suppose that the cholesterol levels are normally distributed with a mean of 160 (mg/dl) with a standard deviation of 40 (mg/dl).

[97] Image by CK-12 is licensed under CC BY NC 3.0.

[98] Throughout this section, we use terms like "within," "less than," "more than," and "between." A natural question may be "Do we include the endpoints?" In the theory of probability, since normal distribution is *continuous*, it is assumed that the probability that the random variable is *one specific value* (a real number) is 0. As a consequence, it makes no difference whether we say "less than" or "less than or equal to." Essentially, therefore, the endpoints do not matter that much.

(1) What is the z−score corresponding to the cholesterol level of 200 (mg/dl)? How about 260 (mg/dl)?
(2) What percent of people have a cholesterol level between 160 and 200? How about 260 or higher?

[Solution]

(1) Note that $\frac{200-160}{40} = 1$, so the z−score for 200 (mg/dl) is 1.0. Similarly, $\frac{260-160}{40} = 2.5$, so the z−score for 260 (mg/dl) is 2.5.
(2) x−values (cholesterol levels) between 160 and 200 would correspond to the z−score between 0 and 1. According to the table, this amounts to 0.3413, or about 34%. In other words, about 34% of all the people have a cholesterol level between 160 and 200. However, 260 mg/dl corresponds to $z = 2.5$. The table shows that 49.38% of the people belong to the interval between $z = 0$ and $z = 2.5$. **Adding the 50%** of those who are below average ($z < 0$), this gives us 99.38% below $z = 2.5$. Since $100 - 99.38 = 0.62\ (\%)$, only 0.62%, i.e., less than 1% of the population, has a cholesterol level 260 or higher. This person should go see the doctor immediately.

This naturally leads to **percentiles**, a great tool to know the relative standing of a data value. Percentiles basically indicate the percentage of values below a given value.

Percentile

The k^{th} percentile is the score (data value) x such that k% of all the data are below x (and k is typically the largest integer value satisfying this condition).

In other words, the 40th percentile is the data value such that 40% of all the data are below that value. If an infant is in (or at) the 80th percentile in weight, then 80% (but not 81%) of all babies weigh less than that infant. If 62.47% of scores are below x, then x is in the 62nd percentile. In the previous example, we found out that 99.38% of all people have a cholesterol level below 260 mg/dl; therefore, this person is in the 99th percentile. By the way, no x can have 100% of all data values below it (since x itself is in the data set), meaning that no score can ever be in the 100th percentile, technically speaking. (It should be noted that the definition of percentile varies slightly from book to book, so beware.)

Example 2

Suppose that IQ scores are distributed with a mean of 100 and a standard deviation of 15. What are the probabilities that one's IQ score is
(1) 100 or below?
(2) 115 or above?
(3) 130 or below?
(4) 77.5 or above?

Also, for each score above, find the percentile.

[Solution]

(1) 100 is the average, so by symmetry, exactly 50% of the population has an IQ score below 100. The score 100 (the mean) is the 50th percentile by definition and by symmetry.

(2) 115 is one standard deviation above the mean, i.e., $z = 1.0$. So, by the table, 34.13% of the population has an IQ score between 100 and 115. Since 50% is supposed to be above the average of 100 (by symmetry), this means $50 - 34.13 = 15.87\ (\%)$ has an IQ score *above* 115. This means $100\% - 15.87\% = 84.13\%$ are *below* 115, so the score 115 is in the 84th percentile.

(3) 130, which is 30 (that is 2 standard deviations) above 100, corresponds to $z = 2.0$. The table has .4772, or 47.72%, between 100 and 130. But here, we are asked for the percentage *below* 130, so we need to add the bottom half, 50%, to this number: $47.72\% + 50\% = 97.72\%$ of people are below 130. This means that the score 130 is in the 97th percentile.

(4) 77.5 corresponds to $z = -1.5$ because $z = \frac{77.5-100}{15} = \frac{-22.5}{1.5} = -1.5$. The table does not directly give probabilities for negative z-scores, but we use symmetry. Between the mean ($z = 0$) and $z = 1.5$, we have the probability 0.4337 on the table. This means that 43.37% of IQ scores are between $z = 0$ and $z = 1.5$; by symmetry, 43.37% of IQ scores are between $z = -1.5$ and $z = 0$. This time we are asked to find the percentage *above* 77.5, so we need to add the higher half (50%) to this number: $43.37\% + 50\% = 93.37\%$ of the people have an IQ score above 77.5. For the percentile, we calculate the percentage *below* this score: $100\% - 93.37\% = 6.63\%$. Hence, the score 77.5 is in the 6th percentile.

It should be noted that various details of IQ testing have changed from time to time, and one's IQ score also varies over time. Many controversies exist for IQ scores, and some even consider IQ scores unreliable. Interested readers can search for and find a lot of information on this topic.

Finally, you may be wondering, "What about those data values that are ***not*** **normally distributed**?" Indeed, many quantities are not normally distributed. However, here is a surprising fact. Remember that a population is often too large and we almost always use samples? So suppose you want to study some data, which may or may *not* be normally distributed. You pick a random sample and find its sample mean, which is a statistic. Someone else could pick another sample (of the same size) and find its sample mean. Someone else can pick a third sample, etc. You can expect that all those sample means are different since the data in each sample are different. So now what you have is a set of **sample means**, where each

sample is of the same size and taken from the same population. These sample means have their own distribution (some sample means are small; others are large). What is amazing is that, regardless of the original distribution of the data you are studying (which may *not* be normally distributed), **these sample means are always normally distributed**, provided that the sample size is large enough (like 30). This is actually a theorem, a statement proved true by a mathematically rigorous proof; it is called the "**Central Limit Theorem**," one of the most important and celebrated theorems in all applied mathematics (with an elegant proof!). It is this theorem that makes the z-score extremely important. Given a sample mean and a sample standard deviation, you can **reject** certain claims or **hypotheses** with a fixed level of confidence. Using such a method, for example, one could make a statement like, "With 98% confidence, that claim is rejected." Many decisions in society, ranging from public policy, health and welfare, education, business, and politics, are made based on these statistical analyses called **hypothesis tests**.

Exercises 5.5

1. Scores of a standardized test are normally distributed with a mean of 500 and a standard deviation of 100. Find the z-score for each test score below:
   (a) 300
   (b) 450
   (c) 600

2. Suppose that the systolic blood pressure levels of a certain group of older adults are normally distributed with a mean of 130 and a standard deviation of 23. Find the z-score for each systolic blood pressure reading below:
   (a) 130
   (b) 160
   (c) 117

3. Scores of a standardized test are normally distributed with a mean of 500 and a standard deviation of 100. Find the x-score (the test score) for each z-score below.
   (a) $z = 0$
   (b) $z = -0.7$
   (c) $z = 2.5$

4. Suppose that the systolic blood pressure levels of a certain group of older adults are normally distributed with a mean of 130 and a standard deviation of 23. Find the systolic blood pressure level for each z-score below.
   (a) $z = 0$
   (b) $z = -1$
   (c) $z = 3.1$

5. Suppose that credit scores of a certain group of home owners are normally distributed with a mean of 730 and a standard deviation of 40. Find the probability that a randomly chosen home owner in this group has a credit score
   (a) Above 730
   (b) Above 810
   (c) Below 770

6. Suppose that human pregnancies have an average duration of 270 days and standard deviation of 14 days. Find the probability that a pregnancy lasts for
   (a) Less than 270 days
   (b) More than 263 days
   (c) More than 284 days

7. Suppose that in a youth basketball league, the average points for a team was 63 per game, with standard deviation 5. Find the probability that a team ends a game with
   (a) More than 68 points
   (b) Less than 53 points
   (c) Less than 73 points

8. Suppose that the batting averages in baseball are normally distributed so that the mean is .258 and the standard deviation is .040. Find the probability that a player has a batting average
   (a) Higher than .258
   (b) Lower than .298
   (c) Higher than .318

9. In a certain country, the people's heights are normally distributed with a mean of 66 inches and a standard deviation of 3 inches. Find the probability that a person in that country is
   (a) Shorter than 60 inches
   (b) Taller than 72 inches
   (c) Shorter than 57 inches

10. Suppose that scores of a standardized test are normally distributed with a mean of 500 and a standard deviation of 100. If a student's score was 650, what percentile is the student in?

11. Suppose that human pregnancies have an average duration of 270 days and standard deviation of 14 days. If a pregnancy lasts for 263 days, what percentile is that in?

12. Suppose that in a youth basketball league, the average points for a team was 63 per game, with standard deviation 5. If a team scored 73 points in a game, what percentile is that score in?

13. Suppose that the batting averages in baseball are normally distributed so that the mean is .258 and the standard deviation is .040. If a player has an .318 career average, what percentile is he in?

14. In a certain country, the people's heights are normally distributed with a mean of 66 inches and a standard deviation of 3 inches. If a woman is 60 inches tall, what percentile is she in?

15. The normal distribution table in this book is not as thorough as those found in other books (although most scientific calculators are able to calculate probabilities found in such tables as well). If you look at more detailed tables, you will see that for the z-score 3.89, the probability is 0.4999; for the z-score 3.90 and beyond, the probability is 0.5000. Explain, in your own words, what these numbers mean.

16. Based on the previous exercise, if scores of a standardized test are normally distributed with a mean of 500 and a standard deviation of 100, what is the highest score you can possibly get? What is the lowest score possible?

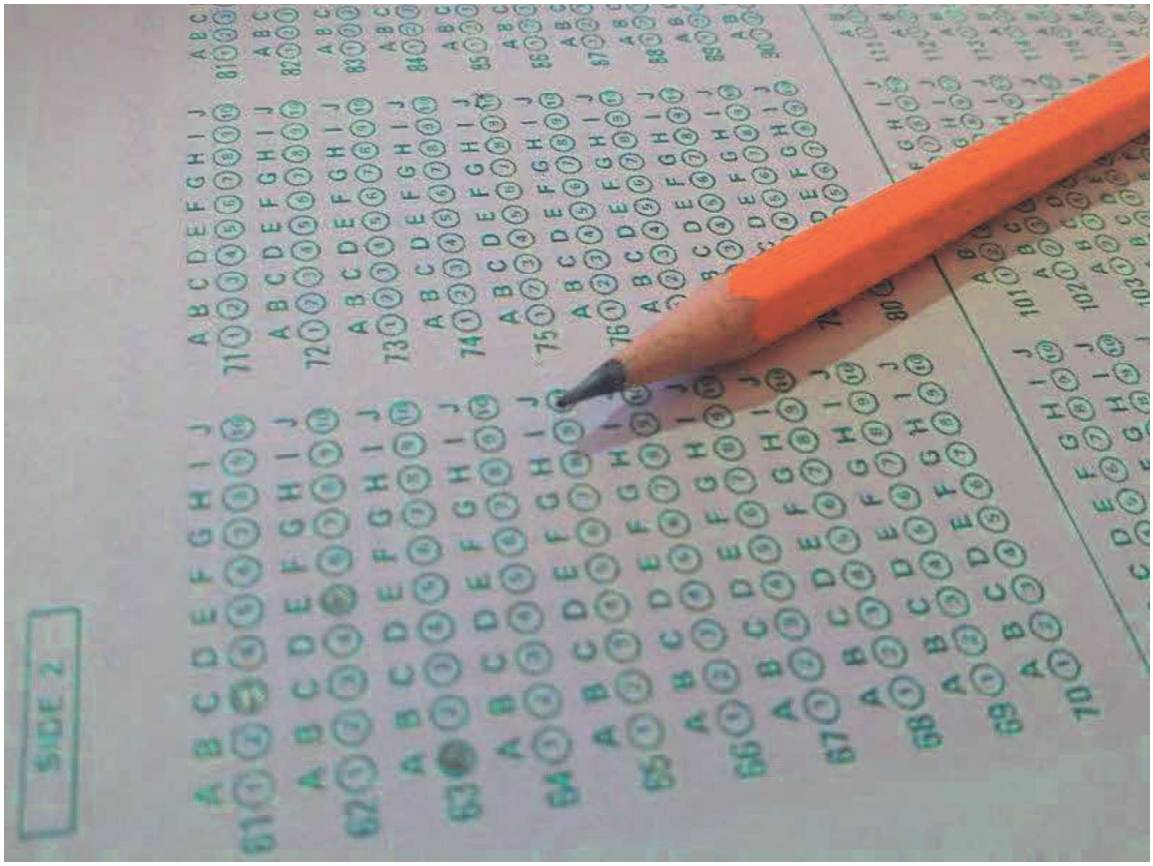


Figure 2.5[99]

[99] Image by lecroitg is in the public domain

For Further Explorations…

A common notion associated with normal distribution is what is known as the **Empirical Rule**, also known the **68-95-99.7 Rule** (this strange name will make sense as explained below). Note the parts of the z-score table below:

z-value	Probability (area)
1.00	0.3413
2.00	0.4772
3.00	0.4987 (almost 50%)

Remember that these probability values indicate the percentages of data on just one side (the right-hand side) of the mean. So if you double each of these values, we get the percentages of data within 1, 2, and 3 standard deviations of the mean *on both sides*, respectively.

$$0.3413 \times 2 = 0.6826$$
$$0.4772 \times 2 = 0.9544$$
$$0.4987 \times 2 = 0.9974$$

These numbers can be summarized as follows:

The Empirical Rule (The "68-95-99.7 Rule")

If data values are normally distributed, the following holds:

(1) About 68% of the data are within one standard deviation of the mean.
(2) About 95% of the data are within two standard deviations of the mean.
(3) About 99.7% (basically all) of the data are within three standard deviations of the mean.

The following picture illustrates this rule.

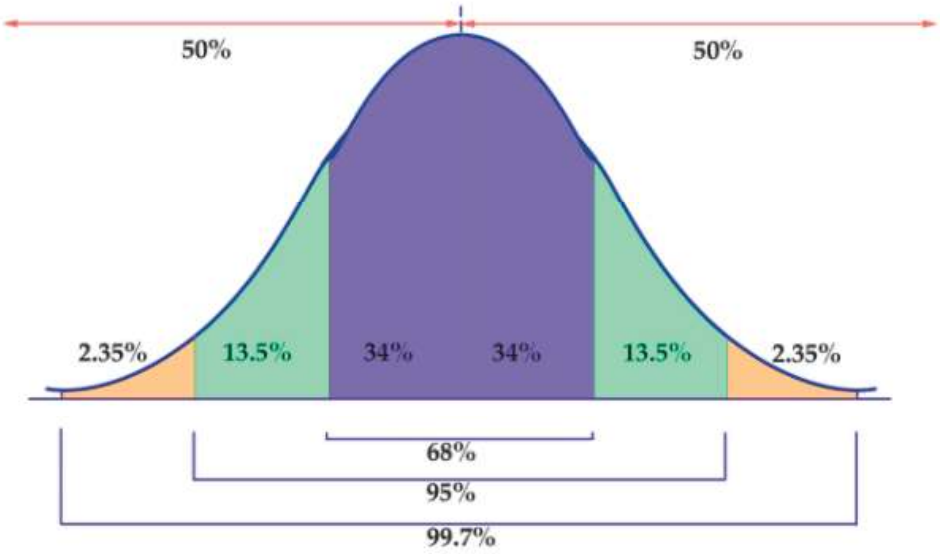


Figure 5.26[100]

Of course, we do not have to use this rule since we have a more precise table (and various tools such as calculators and software) to find these percentages as well as percentiles; however, the Empirical Rule may come in handy when you want to find rough estimates of these percentages.

[100] Image by LumenLearning is licensed under CC BY-NC 4.0

Chapter 6: Growth and Decay

"The human brain is really bad at thinking about exponential things."—Baiju Bhatt, a billionaire entrepreneur

We live in a changing world. Populations of people, animals, and other things are growing all around us. Money grows. The amount of information (data) grows. By understanding how things grow, we can better understand what to expect in the future. In Chapter 3, we already saw how compound interest is calculated. In this chapter, we focus on time-dependent change other than in finance. A special focus will be placed on exponential growth, arguably the most important type of growth in our society.

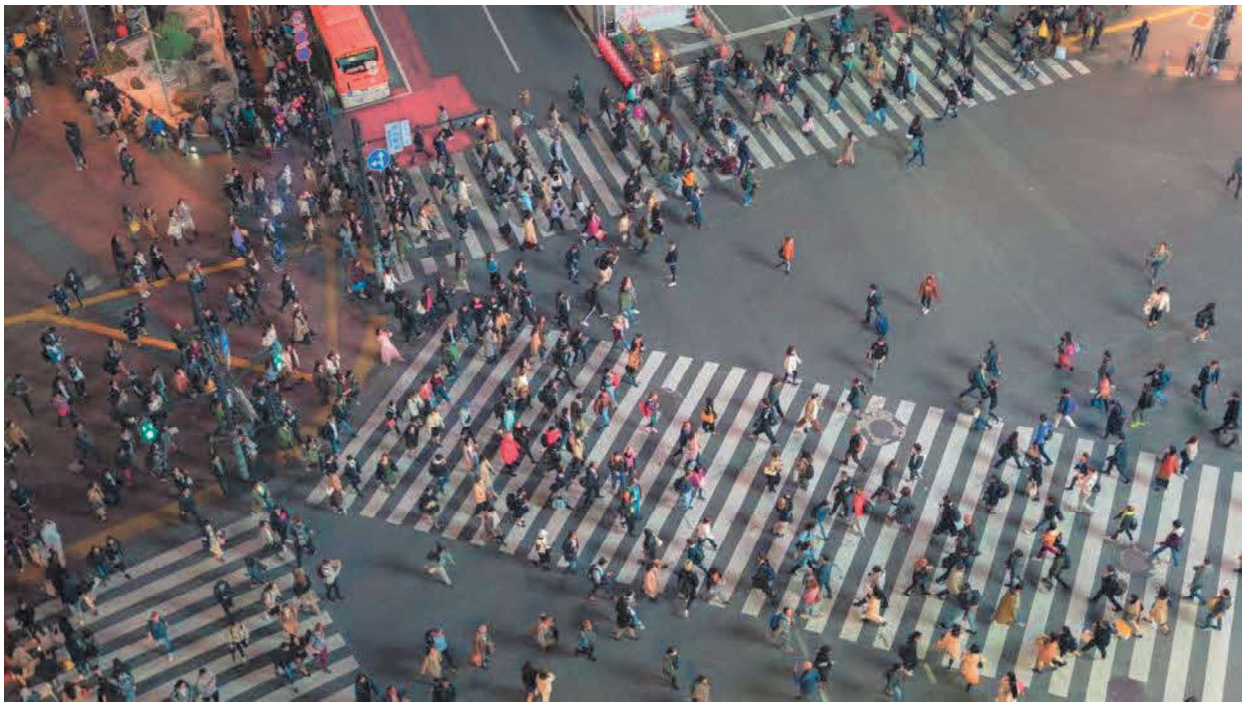


Figure 6.1: Shibuya Crossing in Tokyo, Japan[101]

[101] Photo by Grzegorz Kaliciak on Unsplash

6.1 Growth Models[xiii]

In Chapter 3 we already saw how money "grows" in two different ways—with simple interest or with compound interest. The former is growth by a fixed quantity (dollars) and is linear; the latter is growth by a fixed ratio (percentage) and is faster than linear, especially given a long period of time. These two types of growth are now generalized so that they can describe quantities outside of the financial world.

Linear Growth

Marco is a collector of antique soda bottles. His collection currently consists of 437 bottles. Every year, he budgets enough money to buy 32 new bottles. Can we determine how many bottles he will have in 5 years, and how long it will take for his collection to reach 1000 bottles?

While you could probably solve both of these questions without an equation or formal mathematics, we are going to formalize our approach to this problem to provide a means to answer more complicated questions.

Suppose that P_n represents the number, or population, of bottles Marco has after n years. So P_0 would represent the number of bottles now, P_1would represent the number of bottles after 1 year, P_2 would be the number of bottles after 2 years, and so on. Since 32 bottles are added every year, we could describe how Marco's bottle collection size is changing as follows:

$$P_0 = 437$$
$$P_n = P_{n-1} + 32$$

This is called a **recursive relationship** (the second line is called a **recursive equation**). A recursive relationship is one in which the value in a sequence is defined in terms of the previous value(s). Here, the number of bottles in year n can be found by adding 32 to the number of bottles in the previous year, P_{n-1}. Using this relationship, we could calculate:

$$P_1 = P_0 + 32 = 437 + 32 = 469$$
$$P_2 = P_1 + 32 = 469 + 32 = 501$$
$$P_3 = P_2 + 32 = 501 + 32 = 533$$
$$P_4 = P_3 + 32 = 533 + 32 = 565$$
$$P_5 = P_4 + 32 = 565 + 32 = 597$$

We have answered the question of how many bottles Marco will have in 5 years. However, solving how long it will take for his collection to reach 1000 bottles would require a lot more calculations.

While recursive relationships are excellent for describing simply and cleanly how a quantity is changing, they are not convenient for making predictions or solving problems that stretch far into the future. For that, a **closed** or **explicit form** for the relationship is preferred. An **explicit**

equation allows us to calculate P_n directly, without needing to know P_{n-1}. While you may already be able to guess the **explicit equation** for this example, let us derive it from the recursive formula. We can do so by selectively *not* simplifying as we go:

- $P_1 = 437 + 32$ $= 437 + (32 \times 1)$
- $P_2 = P_1 + 32 = 437 + 32 + 32$ $= 437 + (32 \times 2)$
- $P_3 = P_2 + 32 = 437 + 32 + 32 + 32$ $= 437 + (32 \times 3)$
- $P_4 = P_3 + 32 = 437 + 32 + 32 + 32 + 32$ $= 437 + (32 \times 4)$ …

You can probably see the pattern now, and you generalize this to

$$P_n = 437 + 32n.$$

This equation is essentially saying that we start with 437 and add 32 "n times" to calculate the number after n years. Using this equation, we can now directly calculate how many bottles he will have after 5 years:

$$P_5 = 437 + 32 \times 5 = 437 + 160 = 597.$$

This is the same answer we found earlier, of course. We can also solve for when the collection will reach 1000 bottles. We substitute in 1000 for P_n and solving for n.

$P_n = 1000 = 437 + 32n$, or more simply,

$$437 + 32n = 1000.$$

Subtracting 437 from both sides, we get

$$32n = 1000 - 437 = 563.$$

Finally, we divide both sides by 32 to get

$$n = 563 \div 32 \approx 17.59.$$

So Marco will reach 1000 bottles in 18 years (17 years is not quite enough).

In the previous example, Marco's collection grew by the **same number** of bottles every year. This constant change is the defining characteristic of **linear growth**. Plotting the values we calculated for Marco's collection, we can see the values form a straight line, the shape of linear growth.

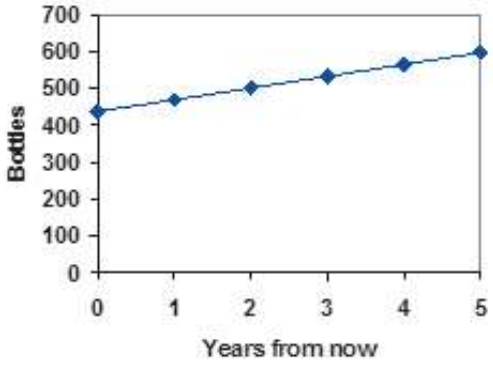


Figure 6.2[102]

This may remind you of simple interest, where the amount added remains constant every year. Simple interest is an example of linear growth as well. The constant rate of growth is called the **common difference (d)** of the linear growth.

Linear Growth

If a quantity starts at size P_0 and grows by d every time period, then the quantity after n time periods can be determined as follows:

$$P_n = P_0 + (d \times n).$$

In this equation, d represents the common difference, the amount that the population changes each time n increases by 1. The quantity P_0 is called the **initial value**.

Connection to Algebra

You may recognize the common difference, d, in our linear equation as **slope**. In fact, the entire explicit equation should look familiar – it is the same **linear equation** you learned in algebra, probably stated as $y = mx + b$. In an algebraic class, recall b was the y-intercept, or the y-value when $x = 0$. In the form of the equation here, we are using P_0 to represent that **initial value**.

In the $y = mx + b$ equation, recall that m was the slope. You might remember this as "**rise over run**" or the change in y divided by the change in x. Either way, it represents the same thing as the common difference d we are using. This is the amount the output P_n changes as the input n goes up by 1.

The equations $y = mx + b$ and $P_n = P_0 + (d \times n)$ both mean the same thing and can be used the same ways.

[102] Growth Models by David Lippman is licensed under CC BY-SA 3.0 US.

Example 1

The population of elk in a national forest was measured to be 12,000 in 2016 and was measured again to be 15,000 in 2020. If the population continues to grow linearly at this rate, what will the elk population be in 2027?

[Solution]

To begin, we need to define how we are going to measure n. Remember that P_0 is the population when $n = 0$, so we probably do not want to literally use the year 0. Since we already know the population in 2016, let us define $n = 0$ to be the year 2016. Then $P_0 = 12{,}000$.

Next we need to find d. Remember d is the growth per time period, in this case growth per year. Between the two measurements, the population grew by $15{,}000 - 12{,}000 = 3000$, but it took $2020 - 2016 = 4$ years to grow that much. To find the growth per year, we divide:

$$3000 \div 4 = 750 \text{ (elk per year).}$$

Equivalently, you can use the slope formula from algebra to determine the common difference, noting that the population is the output of the formula, and time is the input.

$$d = \text{slope} = \frac{15{,}000 - 12{,}000}{2007 - 2003} = \frac{3000}{4} = 750.$$

We can now write our equation:

$$P_n = 12{,}000 + 750n.$$

To answer the question, we need to first note that the year 2027 will be when $n = 11$ since 2027 is 11 years after 2016.

$$P_{11} = 12{,}000 + 750 \times 11 = 20{,}250.$$

This is the number of elk expected in 2027.

Example 2

Gasoline consumption in the U.S. was increasing steadily for a while. Consumption data from 1992 to 2004 is shown below. Find a model for this data and use it to predict consumption in 2023. If the trend continues, when will consumption reach 200 billion gallons?

Year	Consumption (billions of gallons)
'92	110
'93	111
'94	113
'95	116
'96	118
'97	119
'99	125
'00	126
'01	128
'02	131
'03	133
'04	136

[Solution]

Plotting this data, it appears to have an approximately linear relationship.

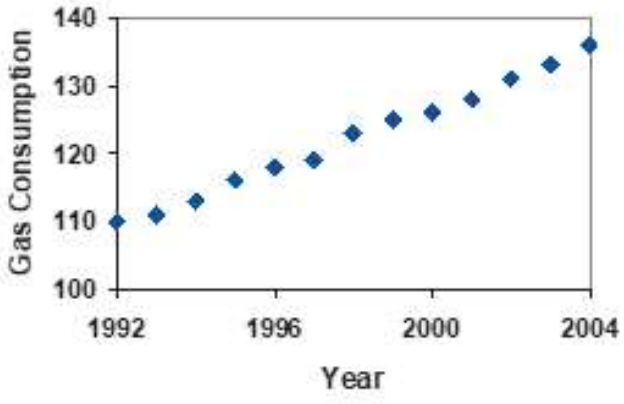


Figure 6.3[103]

[103] Growth Models by David Lippman is licensed under CC BY-SA 3.0 US.

While there are more advanced statistical techniques that can be used to find an equation to model the data, to get an idea of what is happening, we can find a linear equation by using two pieces of the data – perhaps the data from 1993 and 2003.

Letting $n = 0$ correspond with 1993, we let $P_0 = 111$ billion gallons.

To find d, we need to know how much the gas consumption increased each year, on average. From 1993 to 2003 the gas consumption (in billions of gallons) increased from 111 to 133, a total change of 133 – 111 = 22, over 10 years. This gives us an average change of 22 billion gallons / 10 years = 2.2 billion gallons per year.

Equivalently,

$$d = \text{slope} = \frac{\text{change in output}}{\text{change in input}} = \frac{133 - 111}{10 - 0} = \frac{22}{10} = 2.2 \text{ (billion gallons per year)}$$

So our formula is as follows (in billions of gallons):

$$P_n = 111 + 2.2n.$$

(This equation, and hence all subsequent calculations, would be different if other years were used to determine our model instead of 1993 and 2003.)

We can now use our model to make predictions about the future, assuming that the previous trend continues unchanged. To predict the gasoline consumption in 2023, note that $n = 30$ since $2023 - 1993 = 30$ (years). Then,

$$P_{30} = 111 + 2.2 \times 30 = 177.$$

Our model predicts that the country will consume 177 billion gallons of gasoline in 2023 if the current trend continues.

To find when the consumption will reach 200 billion gallons, we would set $P_n = 200$ and solve for n:

$P_n = 200.$	Replace P_n with our model.
$111 + 2.2n = 200.$	Subtract 111 from both sides.
$2.2n = 200 - 111 = 89.$	Divide both sides by 2.2.

$n = 89 \div 2.2 \approx 40.45$. This tells us that consumption will reach 200 billion about 41 years after 1993, which would be in the year 2034.

By the way, the actual U.S. consumption in 2018 was 143 billion while this formula would give 166 billion. Further, 2019 saw the number going down to 142.71 billion gallons. Then, "2020 happened." The consumption was way down due to the COVID-19 pandemic. For these reasons, the model used in this example is now considered obsolete.

Example 3

The cost, in dollars, of a gym membership for n months can be described by the explicit equation $P_n = 70 + 30n$. What does this equation tell us?

[Solution]

The value for P_0 in this equation is 70, so the initial cost is \$70. This tells us that there must be an initiation or start-up fee of \$70 to join the gym. The value for d in the equation is 30, so the cost increases by \$30 each month. This tells us that the membership fee for the gym is \$30 a month.

Try It Now 1

The number of companies in a certain state has been growing steadily, and the trend is fairly linear. Use the data from 1986 and 2020 to find an explicit formula for the number of companies. Then use it to predict the number in 2040.

Year	# of Companies
1986	20,610
1994	28,725
2001	43,530
2010	47,665
2020	53,555

When using mathematical models to predict future behavior, it is important to keep in mind that very few trends will continue indefinitely. In fact, most mathematical models will break down eventually. For instance, a child's height may grow linearly for a while, but certainly, we should not expect anyone to grow at the same rate year after year. As we saw with the gas consumption example, many models become obsolete after a while.

Exponential Growth

Suppose that every year, the number of fish in a lake grows by 10%. If there were 1000 fish in the lake last year, there would now be 1100 fish. Then, the following year, 10% of 1100 (that is 110) will be added, so there will be 1210 fish. Absent of any inhibiting factors, populations of people and animals tend to grow by a certain percentage of the existing population each year. Whenever a quantity grows by a **fixed percentage** rather than a **fixed amount**, the growth is **exponential**, like compound interest.

Suppose our lake began with 1000 fish and the annual rate of increase is 10%. Since we start with 1000 fish, $P_0 = 1000$. How do we calculate P_1? The new population will be the old population plus an additional 10%. Symbolically,

$$P_1 = P_0 + 0.10P_0.$$

Notice that this is simply 110% of the original quantity, so we can also write this as $P_1 = 1.10P_0$. While 10% is the **growth rate**, 1.10 is the **growth multiplier**. As in the case of compound interest, this factor serves as our **base** for exponential growth.

For our fish population, $P_1 = 1000 \times 1.10 = 1100$. More generally, since the number grows to 110% of the previous quantity each year, we can express this growth by an initial value and a **recursive equation**:

$$P_0 = 1000 \text{ and } P_n = 1.10\, P_{n-1}.$$

To get a better picture of how this **percentage-based growth** continues, we need an explicit form. This should remind you of our annual compounding studied in Chapter 3. As we did for the linear model, we will start building from the recursive equation:

$P_1 = P_0 \times 1.10 = 1000 \times 1.10$

$P_2 = P_1 \times 1.10 = 1000 \times 1.10 \times 1.10 = 1000 \times 1.10^2$

$P_3 = P_2 \times 1.10 = 1000 \times 1.10^2 \times 1.10 = 1000 \times 1.10^3$

$P_4 = P_3 \times 1.10 = 1000 \times 1.10^3 \times 1.10 = 1000 \times 1.10^4$

Observing a pattern, we can generalize and obtain the explicit form:

$$P_n = 1000 \times (1.10)^n.$$

Note that this is exactly like the compound interest formula with an APR of 10% annually.

From this, we can quickly calculate the number of fish in 10, 20, or 30 years (rounded to the nearest whole number):

$P_{10} = 1000 \times (1.10)^{10} = 2{,}594$; $P_{20} = 1000 \times (1.10)^{20} = 6{,}727$; and $P_{30} = 1000 \times (1.10)^{30} = 17{,}449$.

Graphing these points reveals a shape that is definitely **not linear**. If our fish population had been growing linearly (by 100 fish each year) the population would have only reached 4000 in 30 years compared to almost 18,000 with **exponential growth**. This is exactly identical to the difference between 10% simple interest and 10% compound interest.

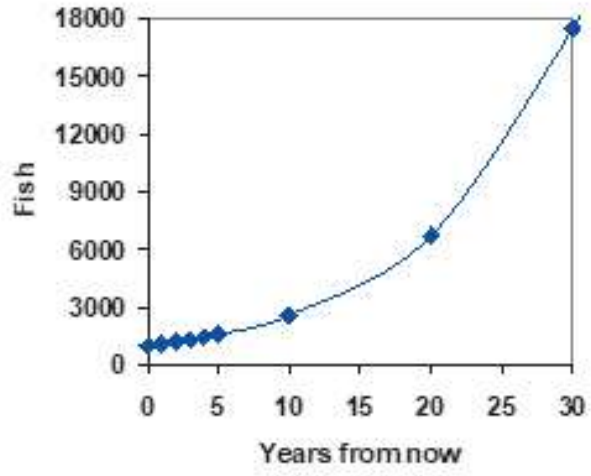


Figure 6.4[104]

In exponential growth, the population grows in proportion to the size of the population; therefore, as the population gets larger, the same percentage growth will yield a larger numeric growth.

Exponential Growth

If a quantity starts at size P_0 and grows by a fixed percentage (written as a decimal $\boldsymbol{r}$) every time period, then the quantity after n time periods is expressed as follows:

$$P_n = P_0(1 + r)^n$$

r is called the **growth rate**.

Example 4

A city with a current population of 245,000 is expected to grow by 3% every year. What will the population of this city be in 6 years?

[Solution]

We know the growth rate is 3%, giving $r = 0.03$. Using the explicit form, we get

[104] Growth Models by David Lippman is licensed under CC BY-SA 3.0 US.

$$P_6 = 245{,}000 \times (1.03)^6 \approx 245{,}000 \times 1.19405 \approx 292{,}542.$$

The model predicts that in 6 years, the city would have a population of about 292,542.

The mathematics is exactly the same as calculating the balance on a $245,000 one-time deposit invested at 3% annually for 6 years, as we saw in Section 3.1. Remember that, on your scientific calculator, your exponential key is labeled in one of the following ways.

$\boxed{\wedge}$, $\boxed{y^x}$, or $\boxed{x^y}$

Try It Now 2

India is the second most populous country in the world, with a population in 2019 of about 1.366 billion. The population is growing by about 1.34% each year. If this trend continues, what will India's population be in 2030?

Example 5

A friend is using the equation $P_n = 12{,}800\ (1.072)^n$ to predict the annual tuition at a local college. She says the formula is based on years after 2018. What does this equation tell us?

[Solution]

In the equation, $P_0 = 12{,}800$, which is the initial value of the tuition when $n = 0$. This tells us that the tuition in 2018 was $12,800. The **growth multiplier** is 1.072, so the growth rate is 0.072, or 7.2%. This tells us that the tuition is expected to grow (exponentially) by 7.2% each year.

Example 6

In 2010, the residential energy use in a certain nation was responsible for 962 million metric tons of carbon dioxide emissions. By the year 2020, that number had risen to 1.182 billion (i.e., 1,182 million) metric tons. If the emissions grow exponentially and continue at the same rate, what will the emissions grow to by 2070?

[Solution]

Let us associate $n = 0$ with 2010 as that is the year for the first piece of data given. That will make $P_0 = 962$ (in million metric tons of CO_2). In this problem, we are not given the growth rate, but instead we are given that $P_{10} = 1182$. When $n = 10$, the explicit equation gives us

$$P_{10} = P_0(1 + r)^{10} = 962 \times (1 + r)^{10} = 1182.$$

We are to find r in this equation. To that end, we first divide both sides by 962 and get

$$(1+r)^{10} = 1182 \div 962 \approx 1.22869.$$

At this point, you can use an algebraic method to find r, or you can use trial-and-error to estimate the rate. Try with $r = 0.01, 0.02$, etc. and repeat if you want more precision. For instance, $(1.02)^{10} = 1.21899\ldots$, so you may guess r is a number slightly bigger than 2%. But $(1.021)^{10} = 1.230998\ldots$, so r must be less than 2.1%, i.e., between 2.0% and 2.1%.

The algebraic method requires what is known as **radicals** or **rational exponents**; taking the 10th root of both sides (or raising both sides to the power of 0.1),

$$(1+r)^{10} = 1.22869 \text{ becomes } 1 + r = \sqrt[10]{1.22869} \approx 1.020808.$$

So $r \approx 0.0208$, or about 2.08%. This is the annual rate of increase. So the emissions are growing by about 2.08% per year. We can now predict the emissions in 2070 by finding P_{60} (since 2070 corresponds to $n = 60$).

$$P_{60} = 962 \times (1.0208)^{60} \approx 3{,}308.4.$$

Therefore, one can expect the emissions to be about 3.3084 billion metric tons in 2070.

As in Chapter 3, you must be careful to avoid rounding as much as possible in exponential calculations. In this example, if we had rounded the rate to 2.1% (rather than 2.08%), our answer would have been 3,156, which means the emissions will be 3.156 billion metric tons. The difference would be 152.4 million metric tons! Just as a very small percentage rate on a mortgage can save you tens of thousands of dollars (Section 3.4), a small percentage difference can significantly alter your answer in exponential-growth calculations.

One more note on the last example: if you decide to use the algebraic method, you can take the 10th root of an expression on your calculator by simply raising the expression to the exponent $\frac{1}{10}$, which is 0.1. Generally, if you are to find the n^{th} root, just raise the expression to the power of $\frac{1}{n}$. You can carry out this operation by entering **^ (1/n)** on your calculator.

Rational Exponents and Radicals

If $x > 0$, then its (principal) n^{th} root can be found by using a rational (fraction) exponent:

$$\boldsymbol{(x)^{\frac{1}{n}} = \sqrt[n]{x}}$$

This is the positive number whose n^{th} power is x.

Try It Now 3

The number of users on a social networking site was 45,000 in February when it officially went public, and it grew to 60,000 by October. If the site is growing exponentially and the growth continues at the same rate, how many users should they expect two years after they went public?

Example 7

Looking back at the last example, for the sake of comparison, what would the carbon emissions be in 2070 if emissions grow linearly at the same rate?

[Solution]

Again we will let $n = 0$ correspond with 2010, giving $P_0 = 962$. To find the common difference d, we could take the same approach as earlier, noting that the emissions increased by 220 million metric tons in 10 years, giving a common difference of 22 million metric tons each year. For the emissions in 2070, we can use the explicit formula:

$$P_{60} = 962 + (22 \times 60) = 2282 \text{ (million metric tons).}$$

You will notice that this number, 2.282 billion metric tons, is substantially smaller than the prediction from the exponential growth model (3.3084 billion metric tons). Plotting more values helps illustrate the differences.

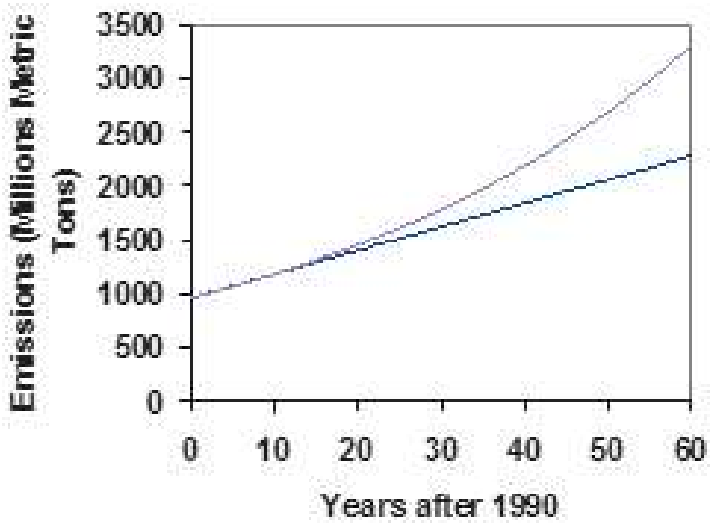


Figure 6.5[105]

[105] Growth Models by David Lippman is licensed under CC BY-SA 3.0 US.

So how do we know which growth model to use when working with data? There are two approaches which should be used together whenever possible:

1) Find more than two pieces of data. Plot the values and look for a trend. Does the data appear to be changing like a line, or do the values appear to be curving upwards?
2) Consider the factors contributing to the data. Are they things you would expect to change linearly or exponentially? For example, in nature, absent other inhibiting factors, most things tend to change **exponentially**.

There is a **mathematical reason** that many quantities in nature (including human, animal, and virus populations) change **exponentially**. Scientists have discovered that the *rate of change of a quantity is often directly proportional to its current quantity*. It is beyond the scope of this textbook to explain this statement fully (as it involves what is known as **differential equations** based on calculus), but simply stated, it says, "The bigger the quantity (e.g., population) is, the faster it grows." So as a population grows, the rate of growth also increases. This translates precisely to an exponential growth.

It should also be noted that any exponential growth (i.e., with a fixed percentage) has a fixed amount of time in which the quantity doubles. This is called the **doubling time**. For instance, if the world population is growing at 2% annually, the population doubles about every 35 years:

$$P_0(1.02)^{35} = 1.9999P_0.$$

The doubling time can be estimated by what is known as the **Rule of 70**:

Rule of 70

If a quantity is increasing exponentially with an annual growth rate of $r\%$, then the doubling time is about $\frac{70}{r}$ years. (Here, r is a percentage, not in decimal form.)

If this rule is applied to the example above, the rule gives $\frac{70}{2} = 35$ (years), the correct answer. (A more precise solution would require logarithms.)

Just think about this for a moment. If there are 8 billion people in the world in 2021, this would mean that 35 years later, in 2056, there would be 16 billion people on earth; 35 years after that, in 2091, there would be 32 billion. Another 35 years, and we would have 64 billion. Fortunately, the world population is growing at a rate closer to 1.1%, but this is something to keep in mind; **do not underestimate exponential growth**. This may affect many real-life issues as well, including climate change, sustainability, food shortage, welfare, and more.

Example 8

Here are some scenarios where the doubling time is very short. Answer each question below.

(1) The number of cells in an experiment culture doubles every hour. Suppose one cell was placed in a container at noon. How many cells will there be 15 hours later? How about 30 hours later?
(2) A **magic penny** is a fictional penny that clones itself every day. If you have one at home today, there will be two of them tomorrow, four on Day 2, eight on Day 3, etc. How many will you have on Day 15? How about on Day 30?
(3) Suppose you start your own religious group today and you are the only person. At the end of Year 1, though, you get one friend to join your group, so there are 2. At the end of Year 2, each of you will bring one more person, so there will be 4. If the number of members doubles every year, how many people will be in the group at the end of Year 15? How about Year 30? How about Year 35?

[Solution]

You probably figured out that all these are essentially the same questions. For (1), there is one cell at noon. At 1 pm (one hour later), there will be 2. At 2 pm there will be $2^2 = 4$. At 3 pm there will be $2^3 = 8$. Therefore, 15 hours later (at 3 am), there will be $2^{15} = 32{,}768$ cells, and 30 hours later there will be $2^{30} = 1{,}073{,}741{,}824$ cells. Yes, more than 1 billion. The answers are exactly the same for (2). There will be 32,768 pennies on Day 15 and over 1 billion pennies on Day 30. For (3), again, there will be 32,768 members at the end of Year 15 (not a big religious movement), but by the end of Year 30, there will be over 1 billion. By the end of Year 35, the number will be $2^{35} \approx 34.36$ billion, which is 4 times the world's population. Once again, do not underestimate exponential growth.

The above example illustrates something else—the power of mathematics. Because it is abstract, the same mathematical principles can explain a wide variety of phenomena. You have already seen that exponential growth is seen in finance, biology, and population growth. In Chapter 7, you will see an application of it in music. Remember, *the power of mathematics lies in its abstract nature.* Learn to use it. Learn to think mathematically. Every day.

Finally, while we focused on **exponential growth** in this section, where the growth multiplier is greater than 1, we can also use multipliers **less than 1** to describe quantities that decrease exponentially. This is called **exponential decay**, which describes many phenomena in nature as well. Just as exponential growth has a fixed doubling time, exponential decay has a fixed **half-life**, the amount of time it takes the quantity to become half. For quantities decaying by a rate of r, we would use the formula

$$P_n = P_0(1 - r)^n$$

For instance, if something is decreasing by 20% every year, we would use

$$P_n = P_0(1 - 0.2)^n = P_0(0.8)^n$$

In this case, since $(0.8)^{3.1} \approx 0.5007$, the half-life would be about 3.1 years. While exponential decay is not a main topic in this section, some exercises below involve this concept.

Exercises 6.1

1. Alex currently has 20 tulips in his yard. Each year he plants 5 more.
   (a) Write a recursive formula for the number of tulips Alex has in n years.
   (b) Write an explicit formula for the number of tulips Alex has in n years.
   (c) When would he have 55 tulips?

2. Pam collects CDs. Every year she buys 13 new albums to add to her collection. She currently owns 450 albums.
   (a) Write a recursive formula for the number of albums Pam has in n years.
   (b) Write an explicit formula for the number of albums Pam has in n years.
   (c) When would she have more than 550 albums?

3. A store's annual sales (in thousands of dollars) grow according to the recursive rule

$$P_0 = 40 \text{ and } P_n = P_{n-1} + 15.$$

   (a) Calculate P_1 and P_2.
   (b) Find an explicit formula for P_n.
   (c) Use your formula to predict the store's sales in 10 years.
   (d) When will the store's sales be $100,000?

4. The number of houses in a village is growing every year according to the recursive rule

$$P_0 = 200 \text{ and } P_n = P_{n-1} + 30.$$

   (a) Calculate P_1 and P_2.
   (b) Find an explicit formula for P_n.
   (c) Use your formula to predict the number of houses in 12 years.
   (d) When will there be more than 400 houses in the village?

5. A population of beetles is growing according to a linear growth model. The initial population (week 0) was 3, and the population after 8 weeks was 67.
   (a) Find an explicit formula for the beetle population in week n.
   (b) After how many weeks will the beetle population reach 200?

6. The number of streetlights in a town is growing linearly. Four months ago ($n = 0$) there were 130 lights. Now, 4 months later, there are 146 lights. If this trend continues,
   (a) Find an explicit formula for the number of lights in month n.
   (b) How many months will it take to reach 200 lights?

7. A city's population in 2020 was about 200,000 and has been growing by 9% each year.
   (a) Write a recursive formula for the population of this city.
   (b) Write an explicit formula for the population of the city.
   (c) If this trend continues, what will the population be in 2026?
   (d) When does this model predict the population to exceed 400,000?

8. The population of a state in 2017 was about 568,000 and has been growing by 1.1% each year.
   (a) Write a recursive formula for the population of this state.
   (b) Write an explicit formula for the population of the state.
   (c) If this trend continues, what will the population be in 2025?
   (d) When does this model predict the population to exceed 1 million?

9. Diseases tend to spread according to the exponential growth model. In the early days of AIDS, the annual growth rate was around 190%. In 1983, about 1700 people in the U.S. died of AIDS. If the trend had continued unchecked, how many people would have died from AIDS in 2000?
10. The population of the world in 1987 was 5 billion and the annual growth rate was estimated at 2% per year. Assuming exponential growth at that rate, find the projected world population in 2025.

11. A bacteria culture is started with 300 bacteria. After 4 hours, the population has grown to 500 bacteria. If the population grows exponentially,
    (a) Find the growth rate (per hour), either by trial-and-error or by using rational exponents, to four decimal places.
    (b) Write an explicit formula for the number of bacteria.
    (c) If this trend continues, how many bacteria will there be in 1 day (24 hours after the start)?
    (d) Find the doubling time. In other words, how long does it take for the culture to double in number? You may use trial-and-error.

12. A native wolf species has been re-introduced into a national forest. Originally 200 wolves were transplanted. After 3 years, the population had grown to 270 wolves. Assume the population grows exponentially.
    (a) Find the growth rate (per year), either by trial-and-error or by using rational exponents, to four decimal places.
    (b) Write an explicit formula for the number of wolves.

(c) If this trend continues, how many wolves will there be 10 years after the original pack was transplanted?
(d) If this trend continues, how long will it take the population to grow to 1000 wolves?

13. One hundred trout are seeded into a lake. Absent any constraints, their population grows by 70% a year. The lake can sustain a maximum of 2000 trout. Assume the exponential growth model.
    (a) Write an explicit formula for the number of trout.
    (b) Calculate the number of trout after 1 year and after 2 years.
    (c) When will the lake have the maximum number it can sustain?

14. Ten blackberry plants started growing in my yard. Absent any constraints, blackberries will spread by 200% a month. Assume the exponential growth model.
    (a) Write an explicit formula for the number of blackberry plants in my yard.
    (b) Calculate the number of plants after 1, 2, and 3 months.

15. The population of a small town can be described by the equation

$$P_n = 4000 + 70n$$

where n is the number of years after 2020. Explain in words what this equation tells us about how the population is changing.

16. The population of a small town can be described by the equation

$$P_n = 4000 \times (1.04)^n$$

where n is the number of years after 2020. Explain in words what this equation tells us about how the population is changing.

17. [Exponential Decay] A new truck costs $32,000. The car's value will depreciate over time, which means it will lose value. For tax purposes, depreciation can be calculated exponentially. If the truck is worth $20,000 after 3 years, write an explicit formula for the value of the car after n years.

18. [Exponential Decay] Inflation causes things to cost more, and for our money to buy less (hence your grandparents saying, "In my day, you could buy a cup of coffee for a nickel"). Suppose inflation decreases the value of money by 5% each year. In other words, if you have $1 this year, next year it will only buy you $0.95 worth of stuff. How much will $100 buy you in 20 years?

19. Thomas Malthus was an economist who put forth the principle that population grows based on an exponential growth model while food and resources grow based on a linear growth model. Based on this, Malthus predicted that eventually demand for food and resources would outgrow supply, with doom-and-gloom consequences. Do some research about Malthus. Using simple graphs, explain what Malthus meant.

20. Suppose you have a job where your annual income doubles every year. (This would be your dream job.) But your first year, you only make $1. Next year (in one year), you will make $2. Then $4 the following year (in two years). You cannot afford to keep your job, so you thought of quitting it, but you keep it, dreaming of a bright future. How much will you be making in 10 years? 20 years? 30 years?

21. Suppose you have a very thin paper, 0.1 mm thick, and imagine that you could fold it infinitely many times. If you fold it once, the thickness doubles. Twice, and the thickness quadruples. With each fold, the thickness doubles. Assuming that the moon is about 385,000 km away (which is 385 billion mm), how many times would you have to fold that paper so its thickness reaches the moon? Use trial-and-error.

22. Imagine a very toxic substance that is spreading rapidly over the surface of the earth, killing every living thing in its path. The area of the region where the substance is found is doubling every year. Fortunately, at the present, only 0.1% of the earth is covered by the substance (and that region is uninhabitable). Because that is such a small portion of the world, many people are not heeding the warning. How many years would it take it to spread over the whole earth so the entire planet becomes uninhabitable? How much time do humans have to save the earth?

6.2 Logarithmic Scales in Natural Sciences[xiv]

Figure 6.6 [106]

In 2010, a major earthquake struck Haiti, destroying or damaging over 250,000 homes.[107] One year later, another, stronger earthquake devastated the northeastern parts of the island of Honshu, Japan, destroying or damaging over 190,000 buildings,[108] like those shown in the image above.

How do we compare the size of earthquakes? The **magnitudes** of earthquakes are measured on a scale known as the Richter Scale. The Haitian earthquake registered a 7.0 on the Richter Scale whereas the Japanese earthquake registered a 9.0. Even though both caused substantial damage, these two numbers indicate (as we will see below) that the earthquake in Japan was 100 times as strong as the earthquake in Haiti.

The **Richter Scale** is a **base-10 logarithmic scale** (known as **common logarithms**). In other words, an earthquake of magnitude 8 is not twice as great as an earthquake of magnitude 4. As we will soon see, it turns out that a magnitude 8 quake is $10{,}000$ times as powerful as a magnitude 4 quake! In this section, we will investigate the nature of the Richter Scale and the common logarithm upon which it depends.

Converting from Logarithmic to Exponential Form

In order to analyze the magnitude of earthquakes or compare the magnitudes of two different earthquakes, we need to be able to convert between logarithmic and exponential form. For example, suppose the amount of energy released from one earthquake was 500 times as much as the amount of energy released from another. We want to calculate the difference in magnitude. The equation that represents this problem is $10^x = 500$, where x the difference in magnitudes on the Richter Scale. How would we solve for x?

---

[106] Image by Daniel Pierce is licensed by CC 2.0

[107] 2010 Haiti earthquake. (n.d.) *Wikipedia*. https://en.wikipedia.org/wiki/2010_Haiti_earthquake

[108] 2011 Tōhoku earthquake and tsunami. (n.d.) *Wikipedia*. https://en.wikipedia.org/wiki/2011_T%C5%8Dhoku_earthquake_and_tsunami#Damage_and_effects

We have not yet learned a method for solving exponential equations. None of the algebraic tools discussed so far is sufficient to solve $10^x = 500$. We know that $10^2 = 100$ and $10^3 = 1000$, so it is clear that x must be some value between 2 and 3 (since $y = 10^x$is an increasing function). We can examine the graph of $y = 10^x$ below to better estimate the solution.

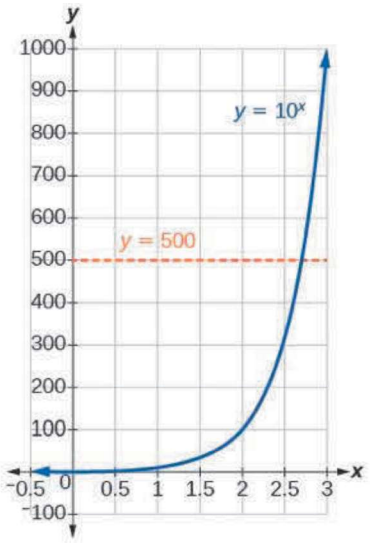


Figure 6.7[109]

When we write $10^2 = 100$ and $10^3 = 1000$, etc., we are using **exponential form** ("2" and "3" are the **exponents**). Alternatively, we can also write these as

$$\log(100) = 2 \text{ and } \log(1000) = 3.$$

This is referred to as **logarithmic form**. Note that *the value of the logarithm is the exponent*. In fact, it is the exponent to which you raise 10 in order to get the number in the parentheses.[110] For every equation in exponential form, you can write an equivalent logarithmic expression, and vice-versa. Here are a few more examples.

$\log(10) = 1$ since $10^1 = 10$.

$\log(100{,}000) = 5$ since $10^5 = 100{,}000$.

$\log(1) = 0$ since $10^0 = 1$.

[109] Logarithmic Functions by Jay Abramson et al. is licensed under CC BY 4.0.

[110] In general, for any base b ($b > 0, b \neq 1$), $log_b(x) = y$ means $b^y = x$. When a logarithm is written without an explicit base b, it is assumed to be 10, and we call it a **common logarithm**. Logarithms can be written with or without parentheses, e.g., "$\log x$" for "$\log(x)$". Thus, $\log(100) = \log 100 = 2$.

More precisely, we have the following.

Common Logarithm

$$\log(x) = y \text{ is equivalent to } 10^y = x.$$

Here, x can be any positive number. We read $\log(x)$ as "the logarithm (with base 10) of x" or "the common log of x." Note that the value y (which is the logarithm of x) is the exponent to which 10 must be raised to get x.

The logarithm *y* is the exponent to which 10 must be raised to get *x*.

Common logarithms are used to measure the Richter Scale mentioned at the beginning of the section. Scales for measuring the brightness of stars and the pH levels of acids and bases also use common logarithms. It should be noted again that a logarithmic scale does not have to use 10 as its base; however, because we use a base-10 numeration system (called the **decimal** system), it seems to make "common" sense to use base 10. This is why base-10 logarithms are referred to as **"common" logarithms**. (See "For Further Explorations…" at the end of this chapter for more details.)

If you are evaluating a logarithm for some given positive x, ask yourself, "*To what power do I raise 10 in order to get x?*" The answer to this question is the value of $\log x$.

Example 1

Evaluate $\log 1000$.

[Solution]

You can rewrite the logarithm in exponential form: $10^y = 1000$. Or, simply ask, "To what exponent must 10 be raised in order to get 1000?" We know $10^3 = 1000$, so the answer is 3. Now, be careful—the answer is *not* 10^3, which is, of course, 1000. The answer is simply 3, only that exponent part.

You may have noticed from the examples given so far that, for powers of 10 (like 10,000 or a million), the common logarithm of that number is simply the number of zeros appearing in that number. So, for 1 million, which is 1,000,000, the answer is 6. In other words, $\log 1{,}000{,}000 = 6$.

Try It Now 4

Evaluate $y =$log 1,000,000,000. (What is the common logarithm of a billion?)

In many situations, when we evaluate $\log x$, the value of x is not a "nice" power of 10, and so it will not be possible to evaluate the result mentally. In such cases, it will be necessary to use a scientific calculator instead.

Example 2

Evaluate $y = \log(321)$ to four decimal places using a calculator.

[Solution]

On a typical scientific calculator, to evaluate this, you press **[LOG]**, enter 321 (and then you may have to close the parenthesis), and press **[ENTER]**. Rounding to four decimal places, you should get $\log 321 \approx 2.5065$. This means that $10^{2.5065} \approx 321$.

Note that $10^2 = 100$ and that $10^3 = 1000$. Since 321 is between 100 and 1000, we know that log (321) must be between log (100) and log (1000). In other words, since $100 < 321 < 1000$, we know $2 < \log 321 < 3$, and it should make sense that $\log 321 \approx 2.5065$.

Try It Now 5

Evaluate $y =$log 8302 to four decimal places using a calculator.

The Richter Scale and Logarithms

Now, you might be wondering why anyone would use something as counter-intuitive as logarithms to describe earthquake strength. Here is the reason (the Richter scale has evolved over the years for better accuracy and applications; the explanations below are somewhat simplified for brevity). If the earthquake strength is to be measured, it needs to be compared to some 'standard' (or reference) energy released.[111]

Thus, the original idea was to define a very, very small energy intensity level, called μ_0 (read as "mu naught"), as the reference; this is the amount of energy that generates an amplitude of 1 micron (1/1000 of one millimeter) on a seismogram 100 km away from the epicenter. Since this is the reference with which all other earthquake energy levels are to be compared, this is defined as 1, which is 10^0 (hence, the magnitude 0 on the Richter scale). This is a motion so small that no humans can feel it. The problem is that most earthquakes have energy levels ranging from hundreds to tens of millions of times this value μ_0. However, it is not practical to use numbers like $391\mu_0$ for a small earthquake or $42{,}000{,}000\mu_0$ for a large one. (Remember, we do not easily understand very large numbers.) It is this wide range of energy levels that convinced Richter to use a logarithmic scale. This way, we do not have to use millions and billions times μ_0. The logarithm for 391 is 2.59, and the logarithm for 42,000,000 is 7.62 (approximately). So the former, smaller quake can be described as a 2.59 while the latter, larger one is a 7.62. The logarithmic scale works well when the numbers in the data range over a vast interval of numbers.

[111] The term "energy released" has sometimes been used synonymously with "intensity"; however, recently the term "earthquake intensity" is often used for the *effects* of the earthquake at various locations. In order to avoid this confusion, we will use terms like "earthquake strength" and "energy released (by an earthquake)."

The Richter Scale (Richter's Magnitude Scale)

The Richter magnitude for an earthquake is $\log\left(\frac{M}{\mu_0}\right)$, where M is the energy released by the earthquake and μ_0 is the reference energy level described above.

For example, if $M = 1000\,\mu_0$, the earthquake registers as a 3.0 since

$$\log\left(\frac{1000\,\mu_0}{\mu_0}\right) = \log\,1000 = 3.0.$$

Similarly, if a minor tremor is a 2.0 quake, it means the tremor released an energy 100 times μ_0 (since $\log\left(\frac{100\,\mu_0}{\mu_0}\right) = \log\,100 = 2.0$.

Note that $\log 100 = 2$, but $\log 300$ is not 3 times 2 (it turns out that $\log 300 \approx 2.4771$). It is important to realize that logarithmic functions are not linear, so they behave rather differently—perhaps even surprisingly.[112] What we need to learn here are a couple of basic properties of exponents. We will use simple equations to illustrate the point.

Note that $2^5 = 32$ and $2^2 = 4$. These basically say that "2 multiplied 5 times together is 32," and "2 multiplied 2 times together is 4," respectively. In other words,

$$2^5 = 2 \times 2 \times 2 \times 2 \times 2 = 32 \quad \text{and} \quad 2^2 = 2 \times 2 = 4.$$

So what is $2^5 \times 2^2$? Writing this out as we did above, we have

$$2^5 \times 2^2 = (2 \times 2 \times 2 \times 2 \times 2) \times (2 \times 2).$$

But clearly, that is 2 multiplied together 7 times. In other words,

$$2^5 \times 2^2 = 2^{5+2} = 2^7 = 128.$$

So **multiplying** two exponential expressions (with the same base) requires that you **add the two exponents**. Similarly, what is $2^5 \div 2^2$? When you write this as the fraction

$$\frac{2 \times 2 \times 2 \times 2 \times 2}{2 \times 2}$$

[112] For instance, it can be shown that $\log(xy)$ is not $\log x \cdot \log y$ but actually $\log(xy) = \log x + \log y$. (But we will not go into details concerning various properties of logarithms.)

Notice that two of the 2's cancel out, showing

$$2^5 \div 2^2 = 2^{5-2} = 2^3 = 8.$$

So **division** requires that you **subtract** an exponent from the other. These are called "properties of exponents" (with base b):

Properties of Exponents (with base b)

$$b^x \times b^y = b^{x+y}$$
$$b^x \div b^y = \frac{b^x}{b^y} = b^{x-y}$$

Comparing Logarithmic Scales

These properties of exponents, particularly the second equation involving division, help us compare two numbers expressed using logarithmic scales. To find ratios involving logarithms, you must be very careful as illustrated below.

Example 3

The 1994 Northridge Earthquake in Southern California registered a 6.7 on the Richter scale. Exactly a year later, the 1995 Kobe Earthquake in Japan registered a 6.9. Compare the amount of energy released from these two quakes.

[Solution]

We wish to know how many times stronger the Kobe quake was than the Northridge quake. To determine this, we divide the amount of energy released from the stronger quake by that of the weaker one.

$$\frac{10^{6.9}\,\mu_0}{10^{6.7}\,\mu_0} = \frac{10^{6.9}}{10^{6.7}} = 10^{6.9-6.7} = 10^{0.2} = 1.585.$$

Note that μ_0 is irrelevant as it appears on both the top and the bottom. The Kobe Earthquake released about 1.58 times the energy (58% more) released by the Northridge quake.

It is easy to see that, according to the properties of exponents, one can compare the power of two earthquakes by finding the **difference** between their Richter scale values and using that as the exponent for 10.

Example 4

The amount of energy released by a big earthquake was 500 times as much as the energy released by a smaller, 3.0 earthquake. What was the magnitude of the bigger earthquake?

[Solution]

We begin by rewriting the exponential equation $\frac{10^x}{10^{3.0}} = 500$, which is equivalent to

$$10^{x-3.0} = 500.$$

Writing this in logarithmic form, we get

$$x - 3 = \log 500.$$

Using a scientific calculator, we see that

$$\log 500 \approx 2.699,$$

so $x - 3 \approx 2.699$, and we get our answer $x \approx 5.699$.

The magnitude of the bigger earthquake was about 5.7. Note that the difference of 2 on the Richter scale always means that the energy of the bigger quake was 100 times as powerful as the smaller one, and the difference of 3 means 1000 times. Remember that every time the magnitude on the Richter scale goes up by 1 (such as from 2.0 to 3.0 or from 6.4 to 7.4), the energy released by the earthquakes becomes 10 times as much.

Try It Now 6

The amount of energy released from one earthquake was 8,500 times as great as the amount of energy released from another. The equation $10^x = 8500$ represents this situation, where x is the difference in magnitudes on the Richter Scale. To the nearest thousandth, what was the difference in magnitudes?

Exercises 6.2

1. Rewrite each equation in exponential form.
   (a) $\log 1000 = 3$
   (b) $\log 0.1 = -1$
   (c) $\log 10^{14} = 14$
   (d) $\log \frac{1}{1000} = -3$

2. Rewrite each equation in exponential form. Then, find x (use a calculator where necessary).
   (a) $\log x = 2$
   (b) $\log 10 = x$
   (c) $\log x = 2.78$
   (d) $\log 75.9 = x$

3. Describe, using your own words, what happens to the energy of an earthquake each time the magnitude on the Richter scale increases by 1.

4. As mentioned in the text, the 1994 Northridge Earthquake (in Los Angeles) registered a 6.7 on the Richter scale. In July, 2019, there was a major earthquake in Ridgecrest, California, with a magnitude of 7.1. How much bigger was the Ridgecrest quake compared to the Northridge Earthquake 25 years before? Use a calculator.

5. Here is a list of some major earthquakes in recent history.
   - The Kashmir Earthquake, Pakistan (October, 2005). Magnitude: 7.6
   - The Indian Ocean Earthquake (December, 2004). Magnitude: 9.1
   - The Haiti Earthquake (January, 2010). Magnitude: 7.0
   - The San Francisco Earthquake (April, 1906). Magnitude: 7.9
   - The South Napa Earthquake (August, 2014). Magnitude: 6.0

   (a) Out of these five, which one was the biggest? Which one was the smallest?
   (b) Compare the strength of the biggest one with the smallest.
   (c) Compare the San Francisco Earthquake with the Kashmir Earthquake. Note that the difference in the Richter magnitude is only 0.3.
   (d) Compare the Haiti Earthquake and the South Napa Earthquake. What would you say to someone who remarks, “The Haiti quake was just a little bigger than the one in Napa; the magnitude difference was only 1”?

6.3 The pH Scale[xv]

You may have heard of pH levels in chemistry or when dealing with gardening or water in a swimming pool or fish tank. pH is a measure of how **acidic** or **basic** a solution is. The pH scale is actually a logarithmic scale, similar to the Richter scale studied in the previous section. In chemistry, the level of acidity depends on the molar concentration of hydrogen ions H^+ in a given solution. This concentration is denoted by $[\mathrm{H}^+]$. The more hydrogen ions there are, the more acidic the solution is. The concentrations are extremely small fractions. Pure water, for example, has a concentration of 1/10,000,000 (1 in 10 million). The denominators range drastically, so it is convenient to use a logarithmic scale.

pH is given by the following logarithmic function of $[\mathrm{H}^+]$:

$$\mathrm{pH} = -\log[\mathrm{H}^+].$$

For example, since $\frac{1}{10{,}000{,}000} = 10^{-7}$, the pH level of water is $-\log[\mathrm{H}^+] = -\log(10^{-7}) = -(-7) = 7$.

pH is usually between 0 and 14. Knowing the dependence of pH on $[\mathrm{H}^+]$, we can summarize as follows:

pH Levels and Acidity

- f pH < 7, then the solution is **acidic** (higher $[\mathrm{H}^+]$).
- If pH = 7, then the solution is **neutral**
- If pH > 7, then the solution is **basic** (lower $[\mathrm{H}^+]$).

pH goes up as $[\mathrm{H}^+]$ goes down. This is what the negative sign does in the definition of pH. You can use pH to make a quick determination of whether a given water-based solution is acidic, basic, or neutral.

Example 1

Label each solution as acidic, basic, or neutral based only on the stated pH.

(1) milk of magnesia, pH = 10.5
(2) pure water, pH = 7
(3) wine, pH = 3.0

[Solution] (no pun intended here)

(1) With a pH greater than 7, milk of magnesia is basic. (Milk of magnesia is largely $Mg(OH)_2$. It is *not* milk. It is a laxative.)
(2) Pure water, with a pH of 7, is neutral.
(3) With a pH less than 7, wine is acidic. Acidity often means sourness.

Example 2

Identify each substance as acidic, basic, or neutral based only on the stated pH.

(1) human blood, pH = 7.4
(2) household ammonia, pH = 11.0
(3) cherries, pH = 3.6

[Solution]

Remember, pH=7 is neutral, so human blood and household ammonia are basic (ammonia is *very* basic). Cherries are acidic.

The table below gives the typical pH values of some common substances. Note that several food items are on the list, and most of them are acidic. Our stomach acid has a pH level generally between 1.5 and 3.0, so it is quite acidic, helping us digest our food.

Table 6.3: Typical pH Values of Various Substances*

Substance	pH
Stomach acid	1.7
Lemon juice	2.2
Vinegar	2.9
Soda	3.0
Wine	3.5
Coffee (black)	5.0
Milk	6.9
Pure water	7.0
Blood	7.4
Sea water	8.5
Milk of magnesia	10.5
Ammonia solution	12.5
Sodium hydroxide (NaOH)	14.0

*Actual values may vary depending on conditions.

pH is a (negative) *logarithmic* scale. A solution that has a pH of 1.0 has 10 times the $[\mathrm{H}^+]$ as a solution with a pH of 2.0, which in turn has 10 times the $[\mathrm{H}^+]$ as a solution with a pH of 3.0 and so forth.

Using the definition of pH, it is also possible to calculate $[\mathrm{H}^+]$ from pH and vice versa. The general formula for determining $[\mathrm{H}^+]$ from pH is as follows:

$[\mathrm{H}^+] = 10^{-\mathrm{pH}}$

You need to determine how to evaluate the above expression on your calculator. Because calculators vary, you should ask your instructor if you have any questions.

Example 3

What is $[\mathrm{H}^+]$ for an acid solution whose pH is 4.88?

[Solution]

We need to evaluate the expression $[\mathrm{H}^+] = 10^{-4.88}$. Depending on the calculator you use, the method for solving this problem will vary. In some cases, the "−4.88" is entered and a "10^x" key is pressed; for other calculators, the sequence of keystrokes is reversed. In any case, the correct numerical answer is as follows:

$$[\mathrm{H}^+] = 10^{-4.88} = 0.00001318 \ldots \approx 1.32 \times 10^{-5}$$

Exercises 6.3

1. The pH level of lemon juice is 2.2 while the pH level of a typical wine is 3.5. They are both acidic, but one is much more acid than the other. Which one is more acidic?

2. Compare the concentration of hydrogen ions between lemon juice with a pH of 2.2 and a typical wine with a pH of 3.5. In other words, how many times as large is the concentration of H^+ in wine as that of lemon juice? Hint: use the same method applied in the previous section to compare two earthquakes. The mathematics is the same.

3. Many orange farmers pay attention to the pH level of their oranges. If an orange produces a juice with a pH level of 4.0, how much is the concentration of hydrogen ions in that juice compared to lemon juice with a pH level of 2.2?

For Further Explorations…

1. "Common Logarithms" vs. "Natural Logarithms"

As pointed out at the end of Section 6.2, we use **common logarithms** primarily because of our familiarity with the number 10. However, from a mathematical perspective, there is another base that is much more preferred (even though it seems to make no sense at all first). In fact, you have already seen that number in Chapter 3 and in Chapter 5. Remember continuous compound interest? An indeterminant form 1^{∞}, first encountered in financial computations, led to the discovery of the number $e \approx 2.71828182845904523\ldots$ (this number is irrational, like π, so the decimal expression has no obvious patterns and does not terminate). Studies in calculus suggest that this number is extremely convenient to use as the base of a logarithm. It is quite surprising that an irrational number would be the ideal base, but it is in fact *so natural* to use this number that logarithms with base *e* are called **natural logarithms**, denoted with "***ln***" (every scientific calculator has this key). The symbol *e* is used in honor of the prolific Swiss mathematician Leonhard Euler (pronounced like "oiler"), who later in his life became totally blind (but obviously continued to see many elegant mathematical truths).

2. Logarithmic Scales for Heights and Weights?

In this chapter we have seen that (common) logarithms are used to describe earthquake size and acidity of chemical liquids. Should we use common logarithms to describe our heights and weights?

First, we need a reference number for each. The "world's smallest baby" who survived premature birth is said to be "Baby Saybie," born in December 2018 in San Diego.[113] She was reported to be about 9 inches long (22.8 cm) and weighed only about 8.6 ounces (about ½ pound, or 244 grams). If we are to use, say 5 inches as the reference height, Baby Saybie's height (in the common logarithmic scale) would be a 0.255 (since $\log\left(\frac{9}{5}\right) = \log 1.8 \approx 0.255$).

On the other hand, someone who is 6 feet tall (72 inches) would be $\log\left(\frac{72}{5}\right) = \log 14.4 \approx 1.16$. Even Lebron James, who is 6'9" (81 inches) would be described as a 1.21 on the logarithmic scale. Not very impressive or helpful, right? For weight, if we set 0.5 pound as the reference, a child who weighs 50 pounds would be described as $\log\left(\frac{50}{0.5}\right) = \log(100) = 2.0$, and someone who weighs 500 pounds would be $\log\left(\frac{500}{0.5}\right) = \log(1000) = 3.0$. Most of us would be between 2.0 and 3.0, and this is not a very practical way to compare weights. Well, you may actually like this since you will not gain much logarithmic-scale weight even when you overeat on Thanksgiving (a logarithmic scale is much more forgiving than a bathroom scale). Therefore, logarithmic scales are appropriate to use only when numbers range over a significantly large interval, such as in star brightness and sound intensity.

[113] Chiu,. A. (2019). 'She's a Miracle': Born weighing about as much as 'a large apple,' Saybie is the world's smallest surviving baby. *The Washington Post.* https://www.washingtonpost.com/nation/2019/05/30/world-smallest-surviving-baby-saybie-miracle/

3. The COVID-19 Cases in the U.S.

During the first months of 2020, the number of Americans infected with the new coronavirus (later named COVID-19) was increasing. With no specific measures or health protocols in place, this little-known virus spread as any other infections do—exponentially. If we let $t = 0$ correspond to March 1, 2021, here are the numbers of known cases in the U.S. at four dates in the spring of 2020, as compiled by the Centers for Disease Control and Prevention(CDC).[114]

Here, t is the number of days since March 1.

t	Date	Number of Cases
0	**3/1/20**	**30**
14	**3/15/20**	**3,487**
30	3/31/20	186,101
45	4/15/20	632,548

As illustrated in Section 6.1, an exponential model $P(t)$ can be uniquely determined by having two data points. Using the initial value $P(0) = 30$ and the March 15 value $P(14) = 3487$ (highlighted above), we can conclude that the cumulative number of COVID-19 cases in the U.S. can be modeled by the function

$$P(t) \approx 30\, e^{0.3396t} \approx 30\, (1.4045)^t.$$

This may not look remarkable, but this shocked the experts; the last expression suggests that the number of known cases is growing by 40% every day. It was growing at a truly alarming rate, and this was a wake-up call to the public health community and policy makers.

Why? Without any factors to slow down or inhibit the infection, the number of cases would continue to rise exponentially. If left alone, VOCID-19 would continue to spread rapidly, resulting in these numbers:

$$P(30) \approx 799{,}521.$$
$$P(45) \approx 140{,}522{,}200.$$

What do these figures mean in the real world? The first says that, absent any measures to fight the virus, there would be almost 800,000 Americans infected on March 31 ($t = 30$). The second equation shows that, by April 15 ($t = 45$), this number would be over 140 million.

[114] Centers for Disease Control. (Aug. 27, 2020). Previous U.S. Covid-19 Case Data. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/previouscases.html#anchor_1598477381

Of course, as the CDC data shows, the cumulative numbers did not go nearly as high as these calculated figures. The huge difference suggests that the drastic policies and measures implemented in mid-March are closely associated with the lower infection rate of COVID-19 (although one cannot definitively establish causation, even in this case).

By the way, get your calculator out and determine the value of $P(48)$. You would find that the answer is over 361 million. In other words, without anything to slow down the spread of this new coronavirus, *the entire country would have been infected by April 18, 2020*. Whereas the actual prediction model used by professional scientists is a simulation system far more complicated than this simple exponential one, the fact is that the virus would have continued to spread at a devastating rate. Now, *that* would have been tragic.

That is what the epidemiologists saw coming. And that is why drastic measures were taken in the middle of March, 2020.

Figure 6.8: Emergency hospital during influenza epidemic, circa 1918.[115]

[115] Image by National Museum of Health and Medicine is licensed under CC BY 2.0

Chapter 7: Mathematics and the Arts

"Film is one of the three universal languages, the other two: mathematics and music."—Frank Capra

One of the objectives of a course like Liberal Arts Mathematics is to appreciate how mathematical concepts are used in fine arts, not just in sciences, business, and statistics. After all, the term "**liberal arts**" involves all types of humanistic endeavors, including art and music, the two fields we will focus on in this chapter.

You may think of fine arts as an area associated with creativity and mathematics as an area least associated with creativity. After all, you cannot break any rules in math; if you do, you get penalized. In many of your math classes, you may have just memorized procedures, repeated them, and just moved on. "Where is creativity in math?" you may ask. Well, as you saw in problem solving, creative thinking is often a key to recognizing patterns or solving problems, where formulas and algorithms do not exist. Mathematics is an academic discipline where, over the few millennia, people exercised their creative intellect to come up with new concepts, state new conjectures, and find some elegant ways to prove them.

So there is definitely an artistic side to mathematics. But here, in this chapter, we look at some mathematical aspects of art and music. To do so, we must first go back a few hundred years (to the Renaissance) or a few thousand years (to Pythagoras).

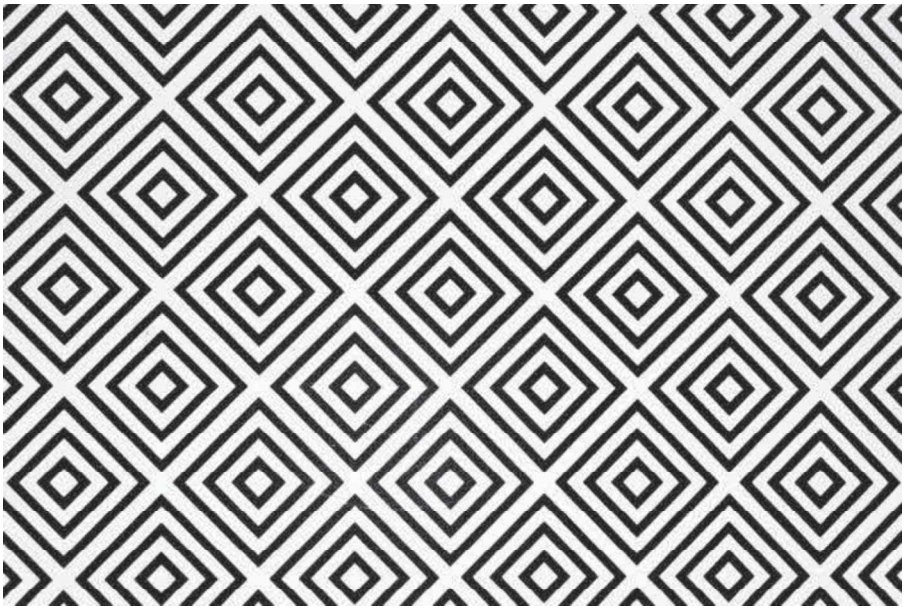


Figure 7.1: A geometric pattern[116]

[116] Photo by Akshar Dave on Unsplash

7.1 Perspective in Art[xvi]

Before drawing a 3-dimensional object on a flat surface, such as a sketch book or a canvas, one needs to learn some principles of perspective. This skill, based on **projective geometry** and developed during the Renaissance, allows the artist to draw things more realistically. All of us have seen, for instance, pictures of railroad tracks and straight highways where the two lines, though parallel in space, appear to converge to a point, referred to as the **principal vanishing point**.

Perhaps you have seen the famous art piece, *The Last Supper*, by Leonardo da Vinci (1452—1519). In that picture (Figure 1), the parallel lines on the ceiling are drawn in such a way that they, if extended, would converge in the middle of the picture, directly behind the head of Christ.

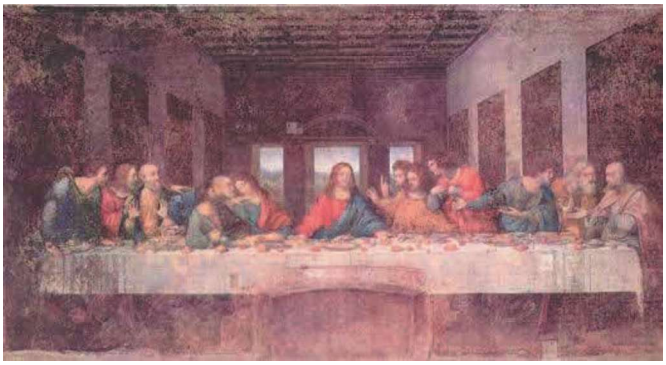


Figure 7.2[117]

During the 15th century, many more artists started to use perspective, practicing it by sketching polyhedral (like cubes, tetrahedra, and dodecahedra) and buildings, while exploring the mathematical basis and rules for such a technique. It was no surprise that da Vinci was also an accomplished mathematician, scientist, and engineer. Interested readers can look up other painters who contributed to the geometric theory of perspective, such as Filippo Brunelleschi (1377—1446), Leon Battista Alberti (1404—1472), Paolo Uccello (1397—1475), Albrecht Dürer (1471—1528), and Wenzel Jamnitzer (1508—1585). Figure 2 shows a sketch by Alberti.

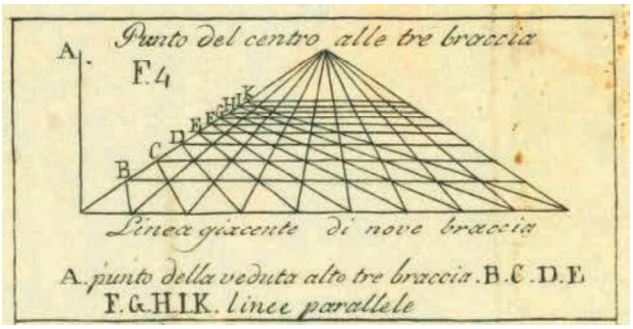


Figure 7.3[118].

[117] . The Last Supper by Leonardo da Vinci is in the public domain

[118] Leon B. Alberti's study on perspective is in the public domain.

In particular, Alberti is famous for the concept of "**visual pyramid**," which gives us the foundation for perspective, as seen in this picture:

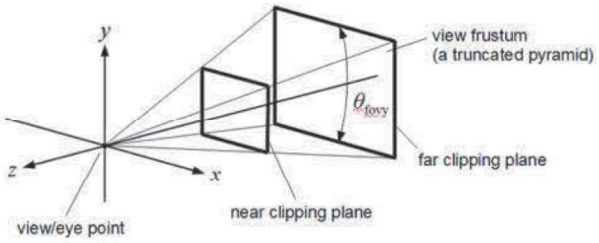


Figure 7.4[119]

Without going into all the details developed over the centuries, we now state some basic principles of perspective.

Figure 7.5[120]

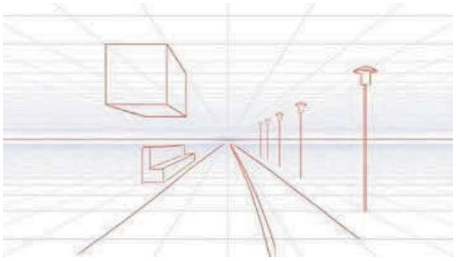


Figure 7.6[121]

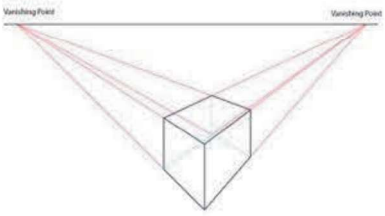


Figure 7.7:[122]

[119] Image by Martin Kraus is licensed under CC BY-SA 3.0
[120] Image by Moajjem Hossain is licensed by CC-BY-SA-4.0
[121] Image by mellowchu on Pixabay
[122] Image by Wikimedia Commons is licensed by CC-3.0

Principles of Perspective

Vanishing Points and Principles of Perspective

1. As Figures 4 and 5 show, all lines perpendicular to the canvas (the plane on which the objectives are drawn) are parallel in space and will converge at a point, called the "**principal vanishing point** (PVP)."
2. All lines parallel to the canvas, such as railroad crossties (wooden pieces below the railroad track as shown in Figure 4) and horizontal edges of the floor tiles (Figure 5), will not converge on the canvas but stay parallel on the canvas.
3. All other lines parallel in space will converge at their vanishing points, as shown in Figure 6, and all of the vanishing points are on one **horizon line** (which also contains the principal vanishing point). This line indicates the **eye level** of the viewer as well.

Knowing these three **principles of perspective** will help you draw objects in 3-dimensional space so that they look much more realistic.

Example 1

Look at the picture below; this demonstrates the principles of perspective. Identify how each of the three principles is illustrated in this figure.

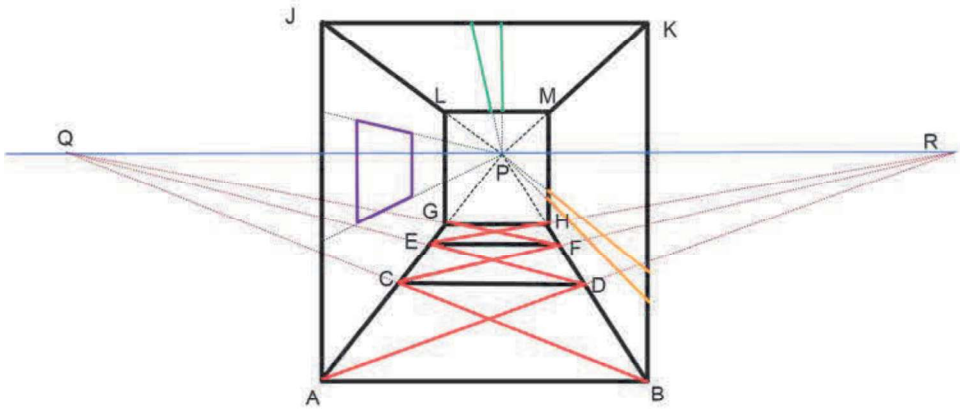


Figure 7.8: Principles of Perspective Shown

[Solution]

(1) Lines JL, KM, AG, and BH are all perpendicular to the canvas; therefore, they converge at the principal vanishing point labeled P in the picture. The same applies to the top and bottom edges of the window frame.
(2) All horizontal lines (JK, LM, GH, EF, CD, and AB) are parallel to the canvas (this page), so they remain parallel in the drawing. The same holds for all vertical lines (LG, MH, JA, KB).

(3) Floor lines AD, CF, and EH are parallel in real space, but they are not perpendicular or parallel to the canvas. Therefore, they converge at another vanishing point, R, not the principal vanishing point. The same holds for lines BC, DE, and FG, all of which converge at another vanishing point Q. Note that P, Q, and R are all on one line, the horizon line (indicating the eye level).

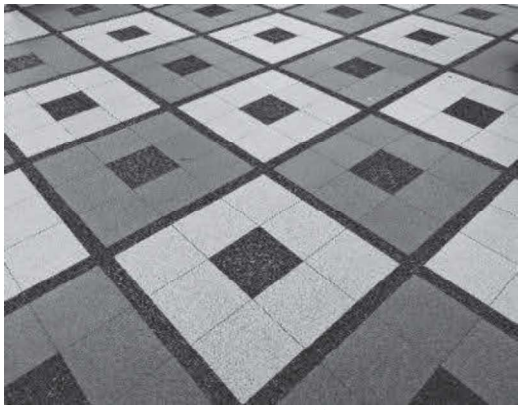


Figure 7.9[123]

Exercises 7.1

1. One side of an elliptic cylinder, a rectangular box, a triangular prism are shown below. Connect each vertex with the principal vanishing point to draw the cylinder, box, and prism using the principles of perspective.

    - (principal vanishing point)

Figure 7.10

2. Suppose you are sitting at the end of a long hallway, where the ceiling has rectangular panels, the floor has square tiles, and the two side walls have vertical lines every few feet. Draw the hallway from your perspective, with a principal vanishing point. Now, what happens if you stand up? How would your picture change? How about if you move to the left?

3. Look up some works of art by chalk artist Julian Beever on the Internet.[124] Look carefully at what he actually draws on a sidewalk and what it looks like from a

[123] Photo by note thanun on Unsplash

[124] For example, http://coolopticalillusions.com/standing-on-top-of-the-world-optical-illusion-chalk-art/.

different angle. What makes this illusion possible? Explain it using the "visual pyramid" in this section.

4. Look up Filippo Brunelleschi's famous experiment in perspective, in which he drew a famous building to study linear perspective. Do some research on this and answer each question below.
   (a) What famous building and in what city did he carry out this experiment?
   (b) What did he actually do in this experiment? How did he draw the building?
   (c) Given the time period in which this occurred, what was the significance of this experiment?

5. Look up some information on Wenzel Jamnitzer. What did he draw? What are the "five regular solids"?

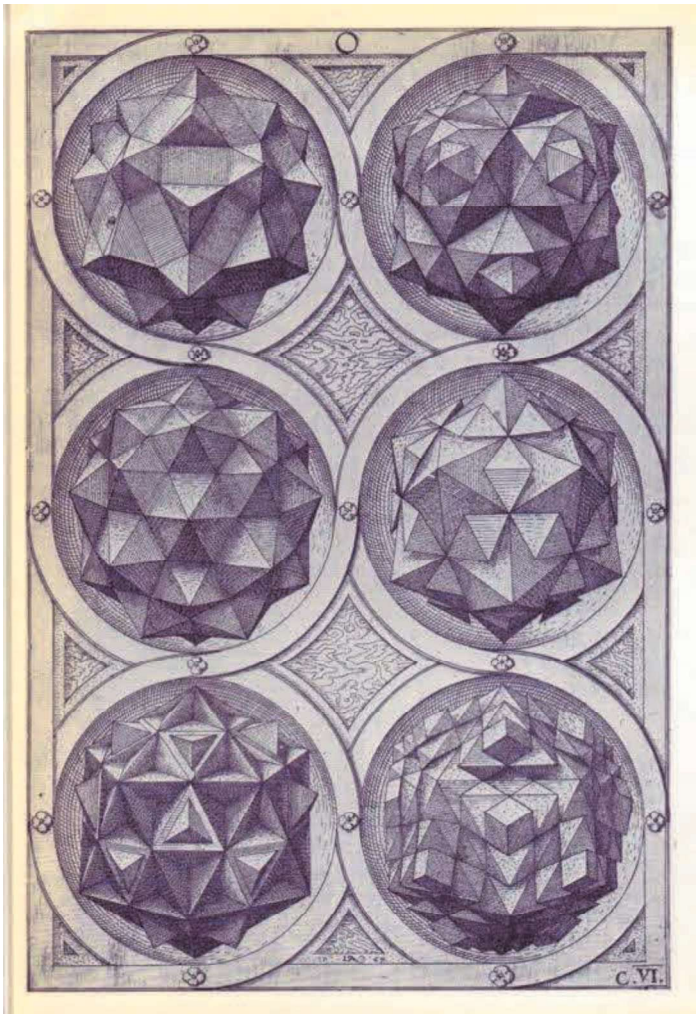


Figure 7.11[125]

[125] Image by Paul K is licensed under CC BY 2.0

7.2 The Golden Ratio and the Fibonacci Sequence[xvii]

In this section, we will discuss a very special number called the Golden Ratio. It is an irrational number, slightly bigger than 1.6, and it has (somewhat surprisingly) huge significance in the world of science, art, and music. It was also discovered that this number has an amazing connection with what is called the Fibonacci Sequence, originally studied in the context of biology centuries ago. This unexpected link among algebra, biology, and the arts suggests the mathematical unity of the world and is discussed in philosophy as well.

Golden Ratio

There is a real number widely known as the **Golden Ratio** (or the **Golden Section**), which has been studied and discussed by many art historians, architects, musicologists, and psychologists for years. What is this number, which has been a focus of attention in analyzing various creative works, and why is it so important, particularly in art? We will deal with these questions in this section. But first, let us figure out how we obtain this number and what this number is.

Mathematically speaking, here is a way to evaluate the Golden Ratio, denoted with the Greek letter ϕ (or $\boldsymbol{\varphi}$, "phi"). Suppose we have a rectangle in which the ratio of its length and its width is such that, if we add a square right next to its longer side, the larger rectangle maintains the same ratio.

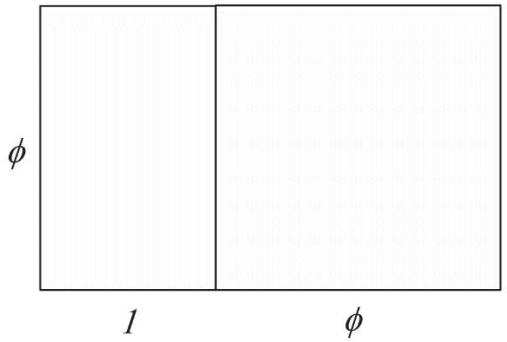


Figure 7.12

Since we are talking about a *ratio*, we can simply assume that one side (the width/base of the original rectangle) is 1 and the other side, the length, is ϕ, as shown above (start with the tall rectangle on the left). Adding a square of side ϕ right next to it creates a larger rectangle (the whole) as shown. We want this number ϕ such that the smaller rectangle and the larger rectangle have the same side ratios (i.e., similar rectangles). We can easily set up this relationship as equal proportions:

$$\frac{\phi}{1} = \frac{\phi + 1}{\phi}$$

From this, by cross-multiplying, we get

$$\phi^2 = \phi + 1$$

This is equivalent to the quadratic equation $\phi^2 - \phi - 1 = 0$, which has two solutions; however, only the positive solution is referred to as the Golden Ratio ϕ. Here is the answer:

Golden Ratio ϕ

$$\phi = \frac{1+\sqrt{5}}{2} \approx 1.6180339887 \ldots$$

$\sqrt{5}$ is the positive number which, when multiplied by itself, results in 5. Like $\sqrt{5}$, the Golden Ratio is an irrational number, so its decimal representation neither terminates nor repeats.

Golden Rectangle

So what exactly makes this number "golden"?

Rectangles with the side ratio of ϕ (like the one shown above) is called "**golden**" because, throughout history, at least since the Greek civilization, this ratio has been used quite frequently in art and architecture and is referred to as the "most elegant" ratio for a rectangle. This involves a sense of "beauty" as perceived by the human brain. Not surprisingly, therefore, psychologists have also studied this number; the **Golden Section Hypothesis** states that this number ϕ represents the rectangular ratio "most aesthetically pleasing" to the human eye.[126] The Parthenon, built around the fifth century B.C.E. in Athens, appears to have been designed using the Golden Ratio. Euclid explicitly wrote about this ratio as it appears in the regular pentagon. Further, art scholars have observed that this ratio can be seen in the works of Michelangelo, da Vinci, and other Renaissance artists. Classical music experts have also discovered this ratio used or alluded to in works of Mozart (his *Sonata No. 1 in C Major*) and Beethoven (his famous *Fifth Symphony*).

Perhaps even more surprisingly, this number, initially calculated using a rectangle and a square, can also be obtained by a very simple sequence, which was originally constructed to study populations of a certain animal. We now move to that special sequence and its incredible relationship to the Golden Ratio.

[126] See, for example, a summary article which appeared in 1980 in the *American Journal of Psychology*: https://www.jstor.org/stable/1422725?seq=1. More information can be found in many websites, e.g., http://www.yorku.ca/christo/papers/Green.golden.Perception-1995.pdf.

Fibonacci Sequence

A **sequence** can simply be thought of a list of numbers. The **Fibonacci Sequence** is a sequence which starts with the numbers 1, 1, and *each subsequent* number in the list is obtained by adding the *previous* two numbers together. Since we start with 1, 1, the next number is $1 + 1 = 2$. The next number is $1 + 2 = 3$. We now have $1, 1, 2, 3$. The next number is $2 + 3 = 5$. Then $3 + 5 = 8$, and so on. The resulting (infinite) sequence,

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots$$

is called the Fibonacci Sequence, and each of these numbers is called a Fibonacci number. Originally, Fibonacci (Leonardo of Pisa, who lived some 800 years ago) came up with this sequence to study rabbit populations! So Fibonacci would be shocked today to find out what would happen when you divide each Fibonacci number by the previous one, as seen below.

Fibonacci Number	Divided by the previous one	Ratio
1	N/A	N/A
1	1/1	1.0000
2	2/1	2.0000
3	3/2	1.5000
5	5/3	1.6666…
8	8/5	1.6000
13	13/8	1.6250
21	21/13	1.6153…
34	34/21	1.6190…
55	55/34	1.6176…
89	89/55	1.6181…
…	…	…
		1.6810…$\approx \phi$

Note that the ratio itself forms a sequence of numbers, and it appears to get closer and closer to 1.68180…, a familiar number. Here is a very surprising fact:

The Fibonacci Sequence and the Golden Ratio

The ratio of two consecutive Fibonacci numbers approaches the Golden Ratio.
More precisely, if F_n is the n^{th} Fibonacci number, then the quotient $\frac{F_{n+1}}{F_n} \to \phi$ as $n \to \infty$.

There is a beautiful way to "visualize" this fact and the table shown above. Each ratio of consecutive Fibonacci numbers can be illustrated by a rectangle in the following figure, which shows a link between the Fibonacci Sequence and the Golden Ratio; this spiral picture is known as the "**Golden spiral**." It consists of quarter circles whose radii are the Fibonacci numbers, 1, 1, 2, 3, 5, 8, 13, 21, … connected as shown below. Because the ratio of two consecutive Fibonacci numbers approach the Golden Ratio, the rectangles thus constructed will become closer to a golden rectangle.

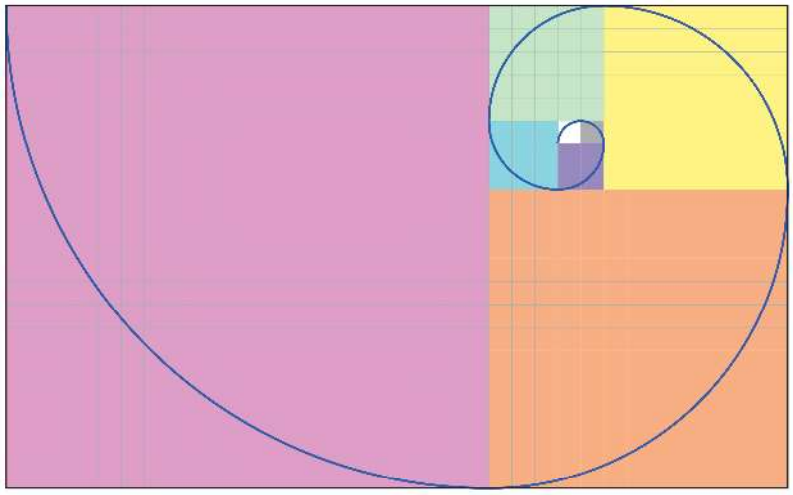


Figure 7.13

In this picture, the "Golden spiral" begins with a 1×1 white square, continues to the adjacent 1×1 gray square on the right, then to the 2×2 purple square below, the 3×3 blue square to its left, the 5×5 green square above, the 8×8 yellow square, the 13×13 orange square, and then to the 21×21 pink square. This process can be extended indefinitely to construct a golden rectangle.

This is a remarkable story. Here we have a number that is defined using a rectangle and a square, shows up in Euclid's regular pentagon, and has many interesting algebraic properties. We further find out that many artists and architects seem to be aware of this number from an aesthetic standpoint. On the hand, there is a sequence originally used to describe rabbit populations, back in Italy, during the 13$^{\text{th}}$ century. And now we discover that these two ideas, which appear to be completely unrelated, are closely associated with each other.

Well, there is more! It turns out that Fibonacci numbers show up quite often in nature. Some examples are the pattern of leaves on a stem, the parts of a pineapple, the flowering of artichoke, the uncurling of a fern and the arrangement of a pinecone.

Figure 7.14: A pinecone[127]

Math in the Real World

Ctrl click The Secret to Good Art! / The Golden Ratio / A Life Changer! to watch a short video that explores the golden ratio in nature, art, and society.

Exercises 7.2

1. Do some online research to find which artists seem to have used the Golden Ratio in their works of art. Give some examples of their art pieces.

2. Do some online research to find which composers seem to have written music that reflects the Golden Ratio. Give some specific examples of their music works.

3. Do some online research on golden rectangles. Why is this called "golden"? Where do we see golden rectangles around us?

4. The Fibonacci Sequence begins with 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, … List the next three Fibonacci numbers (the next three terms of this sequence).

5. Do some online research on the Fibonacci Sequence. Find its relationship to the biological world, such as pinecones, sunflowers, nautilus shells, and human bodies.

6. Do some online research on the history of Fibonacci and his famous sequence. Who was he? What country is he from? Why did he study these numbers in particular?

[127] Photo by Garrie van der Walt on Unsplash.

7.3 Musical Scales

”Mathematics is the music of reason.”—James J. Sylvester

The first thing musicians must do before they can play together is "tune.” For musicians in the standard Western music tradition, this means agreeing on exactly what pitch (what frequency) an "A” is, or a "B flat," and so on. Other cultures not only have different note names and different scales, they may even have different notes - different pitches - based on a different tuning system. In fact, the modern Western tuning system, which is called equal temperament, replaced (relatively recently) other tuning systems that were once popular in Europe. All tuning systems are based on the physics of sound. But they all are also affected by the history of their music traditions, as well as by the tuning peculiarities of the instruments used in those traditions.

Sound and Pitch

Many musical instruments, including pianos, guitars, and violins, make **sound** when strings are plucked, hit, or set in **vibration** in some other way. This vibration causes the air to move, creating an audible wave, which eventually reaches our ears. We call this wave “sound.” The **pitch**—how high or low the sound is—is determined by various factors such as the length of the string, its thickness, and its tension. A guitarist may play an open string, say G, and then place a finger on a fret to create sound with a higher pitch, say G#, A, or B. What is happening is that the player is shortening the vibrating part of the string to produce a higher pitch. This is based on the fundamental fact that shorter (or thinner and/or tighter) strings produce higher pitches. Similar facts are true with woodwind, percussion, and other types of instruments in general: the smaller the vibrating portion gets, the higher the sound becomes. That is why, for instance, a tuba is bigger than a French horn, and a contrabass flute is longer than a soprano flute.

We now focus our attention on string vibrations. In particular, we look at the relation between the pitch and the string length, assuming that the material, thickness, and tightness of the string remain constant. Suppose we pluck a string of a certain length and create a wave, a vibration of the air. The string will move up and down in oscillation. Each wave, whether it is a simple water wave, sound, earthquake, radio wave, microwave, or light wave, has a **frequency**, which measures how many times the oscillation occurs in a given time period, such as one second. For sound, the frequency is measured in “**cycles per second**” (abbreviated “**cps**”), sometimes called “**hertz**” (denoted “**Hz**”). Frequency is what determines the pitch of the sound generated by the vibration.

When a string is plucked and set in motion, the entire string vibrates, as shown in Figure 1. The frequency created by the whole string is referred to as the “**fundamental frequency**” of the string. It generates the lowest sound the string can possibly make. This is equivalent to playing an open string on the guitar. If the guitar is tuned in the standard way, the sixth string (the thickest one, located on the top) is the E string, meaning that its fundamental frequency

corresponds to the pitch we call an “E” (or, more specifically, “E2,” approximately 82.5 cps). This is the lowest sound a guitar can play.

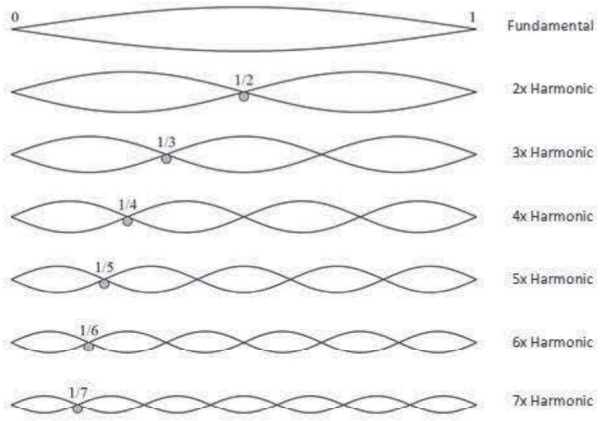


Figure 7.15[128]

Now, if we take that same string, press the midpoint of it and pluck it, only half of the string would vibrate, shortening the wave length to half and **doubling** the frequency. This, of course, results in a higher pitch. But how much higher? When the frequency is doubled, we call the interval between those two pitches an “**octave**” (meaning 8). For instance, if the open string has a frequency of 440 cps (known as A4), then the half string would have a frequency of 880 cps, and that sound/note (called A5) is an octave higher than the first. One can, of course, keep making the vibrating portion shorter and shorter to create higher and higher sounds.

Pythagoras: Numbers and Harmony

Although music has existed in different cultures for millennia, it was **Pythagoras**, an ancient Greek mathematician and philosopher, who explicitly linked sound (pitch) with frequency (number), laying a foundation of music theory based on mathematics. (It was over 2,500 years ago!) It was known to Pythagoras that *two notes (sounds with different frequencies) sound nice together (harmonious, pleasant) when the ratio of the two frequencies is a simple fraction.* For instance, two pitches that are an octave apart have the ratio of 2 to 1 as described above. 2/1 (or its reciprocal ½) is a simple fraction, and these two notes sound nice if played simultaneously. In fact, if two people sing a song one octave apart, it could sound almost like a unison. These simple fractions give rise to well-known interval names in music theory:

2 = 2.00: octave (e.g., C and the next C)

5/3 = 1.666…: major sixth (e.g., C and A)

3/2 = 1.50: perfect fifth (e.g., C and G)

[128] Image on StackExchange Physics is licensed by CC-3.0

4/3 = 1.333…: perfect fourth (e.g., C and F)

5/4 = 1.25: major third (e.g., C and E)

Pythagoras once said, "There is geometry in the humming of the strings, there is music in the spacing of the spheres."

These simple ratios became a foundation for various musical scales and musical chords. For this reason, sometimes Pythagoras is called the **father of music theory**. In fact, as a religious and philosophical leader, he taught that **ratios** create harmony which forms beauty in nature; therefore, ratios are divine. His followers, called the Pythagoreans, thus did not believe the existence of numbers that are not ratios of integers. It is then quite ironic that Pythagoras himself later proved that square root of 2 (the length of the diagonal of a unit square) is irrational, contradicting his own idea that every number is rational. (By the way, this number, $\sqrt{2}$, came about by a very theorem named after him—the Pythagorean Theorem!)

Regardless of this devastating discovery, numerical ratios played a very significant role in the composition of musical scales. A **scale** is an ordered list of musical tones that covers an octave; for instance, today the standard C-major scale consists of 8 whole tones starting with a C and ending with the next high C. If you have taken music lessons, you may have practiced these scales thousands of times.

Throughout history, there have been many musical scales in various cultures, traditions, genres, and generations. Many authors have documented a wide variety of scales and methods of tuning instruments. Interested readers can look up topics such as Pythagorean intonation, mean-tone system, and temperaments. For simplicity, we now focus on the current Western music scale structure, consisting of 12 notes and their frequencies. This scale was not so widespread (or so standard) until the days of J. S. Bach in the beginning of the 18th century. Here are some facts about today's standard musical scale:

Musical Notes, Octaves, and Frequencies

- A **scale** begins and ends with two pitches with the same name, one octave apart (meaning that the frequency of the last note is exactly twice that of the first note).
- This whole interval, an **octave**, is divided into 12 half-steps in such a way that the frequency ratio of any two consecutive notes is the same (described as an "equal distance apart"). In other words, *the frequency "grows" at a constant rate*.

Let us take a look at these requirements mathematically now. (Have no fear! You have seen this before, more than once, in fact.)

Suppose P_0 represents the frequency of the first note in a scale. Every time you move half-step to the next note, there must be a **fixed ratio increase**, i.e., growth by a fixed percentage, say, r (percentage written as a decimal). So the second note has the frequency

$$P_1 = P_0\,(1 + r).$$

Then, to go to the next note by moving half-step up, we need to multiply this new number by the same multiplier, giving us

$$P_2 = P_1\,(1 + r) = P_0\,(1 + r)(1 + r) = P_0\,(1 + r)^2.$$

Repeating, we get

$P_3 = P_0\,(1 + r)^3$, $P_4 = P_0\,(1 + r)^4$ and so on. This looks quite familiar, correct? This is what we saw in *exponential growth*. This is what we saw in compounding interest. Yes, it turns out that the mathematics that rules musical scales is the same as the mathematics that describes compound interest in finance and any exponential (fixed-percentage) growth in natural or social science! Hence, an *"octave" is the doubling time* in this exponential model.

So what is this magic number r, the common ratio that separates every half-step on the musical scale? Remember that after 12 half-steps, we have traveled an octave, doubling the original frequency. In other words, this is the number r so that the following equation is satisfied:

$$P_{12} = P_0\,(1 + r)^{12} = 2P_0.$$

If this equation were to appear in the context of finance, then this r would be the annual interest rate such that the original principal is doubled after 12 years. After dividing both sides by P_0 and taking the 12th root of both sides, we get

$$1 + r = \sqrt[12]{2} \approx 1.059463.$$

We finally conclude that the fixed ratio between any two half-steps is approximately 0.059463, which is close to **5.946%**. So, if you get this interest rate in a bank account compounding annually, you will double your balance in exactly 12 years. If you start your musical scale with a note with a certain frequency, after 12 half-steps, you will double your frequency, resulting in the note exactly one octave higher. (Think of the first two notes in the song "Somewhere Over the Rainbow.")

Below is a graph showing what happens to your bank balance if you begin with the original principal of $440 with the annual rate 5.946% for 24 years; your money doubles twice, i.e., quadruples to $1,760. This is like hitting a note two octaves higher. (You may actually be so excited about your money quadrupling that you scream out loud with a note two octaves higher.)

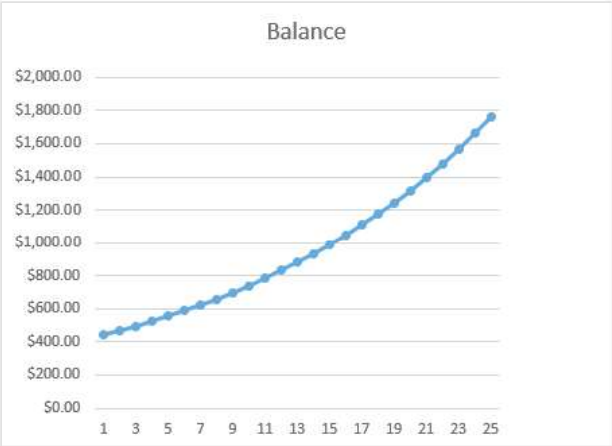


Figure 7.16

Why did I start this graph with $440? Well, in standard tuning, middle A (known as A4) is set to 440 cps. Tuning forks (old gadgets used before electronic tuners and tuning apps were invented) are designed to vibrate at this frequency. Some musicians adjust this a bit, depending on the music they play (or how they want their music to sound).

Example 1

Find the frequency of each of the 12 musical notes from A=440 cps to the next A=880 cps, spanning an octave. (This set of notes is called a chromatic scale.)

[Solution]

Increasing this frequency, 440 cps, by $r = 0.059463$ each time (for 12 times), you will get the following notes in the standard A chromatic scale. For some notes, the ratio of the frequency and 440 is also shown.

Note	Frequency	Ratio (with 440)	
A	440.00	1	
A♯	466.16		
B	493.88	1.122462	
C	523.25	1.189207	
C♯	554.37	1.259921	$\sim 5:4$
D	587.33	1.33484	$\sim 4:3$
D♯	622.25		
E	659.26	1.498307	$\sim 3:2$
F	698.46		
F♯	739.99		
G	783.99		
G♯	830.61		
A	880.00	2	$=2:1$

Figure 7.17

Take a look, for instance, at C#, with a ratio of 1.2599 compared with the frequency of A; this ratio is approximately 5:4 (that is, $5 \div 4 = 1.25$). This note in the key of A major is called "major third"; D is called "perfect fourth," and E is "perfect fifth." Because of these simple (approximate) ratios, the notes A, C#, and E sound beautiful together. The set of these three notes is called the "A major" chord. Most harmony is based on these simple ratios.

Equal Temperament: A Compromise

But each ratio is just *approximately* (and not exactly) a simple ratio, right? This is the result of **equal temperament** (devised in the 18th century); the common ratio r is the (not-so-secret) number that would allow a piece of music to sound good in every key, but the trade-off of temperament is that all ratios are now approximate, except for octaves, and thus, strictly speaking, "out of tune." However, given the difficulty in tuning big instruments like the piano, this was the preferred compromise. The alternative is to have some ratios perfect while others are way off simple ratios, resulting in some music perhaps sounding great in one key but sounding terrible if transposed to another. Equal temperament is thus widely accepted and regarded as standard now. In other words, one irrational number, $\sqrt[12]{2} \approx 1.059463$, seems to control the harmonious world of music ("one *key* number, for the 88 piano keys").

In the text of this section, sometimes a note name was followed by a number, such as in "A4." This number indicates which octave we are in; each number starts with C and goes up by 1 when the next C is reached. For instance, the C closest to the middle in the standard 88-key piano is called the "middle C" or "C4" (about 262 cps). Starting there and moving to the right, the keys are arranged as C, C#, D, D#, E, F, F#, G, G#, A, A#, and B. These twelve notes all have a "4" after that. After B4 comes C5, which is one octave higher than C4. The white key immediately to the left of C4 is called B3. The 88 piano keys range from A0 to C8 as shown below.

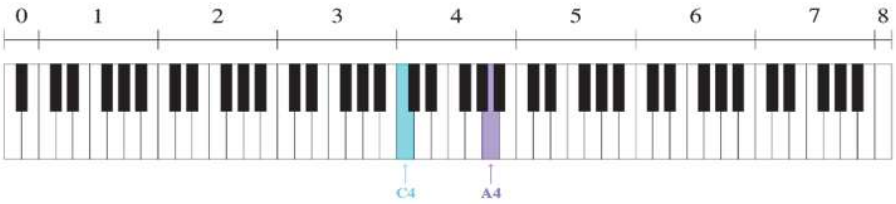


Figure 7.18

Just one more thing… it was mentioned that the **A major chord** consists of A, C#, and E. If you change that middle note and lower it by half-step, you get the notes A, C (natural), and E, which together form the **A minor chord**. Play these chords and hear the difference. The mood shifts dramatically, right? Most people can hear the difference between a song written in a major key and a minor key. In fact, it would make people laugh if you take a popular song written in a minor key and play it in the corresponding major key (or vice-versa). That huge difference, we now know, is generated by one little note dropped by one half-step, i.e., a frequency difference of about 5.946%. Such a small ratio, yet such a big difference!

Exercises 7.3

1. A guitarist tunes his first (thinnest) string at E = 330 cps (typically denoted "E4"). If his sixth (thickest) string is to be two octaves lower, what is the frequency of the sixth string?

2. We started our musical scale at A4 = 440 cps. Some people may start a scale at C5 = 520 cps. Find the frequencies of the thirteen notes which span an octave (C, C#, D, D#, E, F, F#, G, G#, A, A#, B, C), starting at C = 520 cps.

3. If A = 440 cps, the real "perfect fifth" would be E, whose frequency would be exactly 3/2 of A. What frequency is that? Compare that with the 'tempered" frequency of E.

4. If A = 440 cps, the real "perfect fourth" would be D, whose frequency would be exactly 4/3 of A. What frequency is that? Compare that with the 'tempered" frequency of D.

5. Suppose a male singer can comfortably sing from Second B (B2) to Fourth G (G4), about one and a half octave. Assuming that middle C (C4) is 262 cps (and C3 is half of it), find the frequencies that cover this range.

6. It is said that Mariah Carey can hit a note as low as F2 and as high as G7 (slightly more than five octaves).[129] Find the frequencies that cover this range. (Assume "A4" is 440 cps. A3 is half of it. A2 is half of that frequency. F2 is even lower. The sequence goes like this: F2, F#2, G2, G#2, A2, A#2, B2, C3, C#3, …)

[129] There are some video clips on YouTube in which Mariah Carey covers the range from B-flat 2 to F7, e.g., https://www.youtube.com/watch?v=CY6fGVBME8I.

7.4 Fractals[xviii]

"The beauty of mathematics only shows itself to more patient followers."—Maryam Mirzakhani[130]

Fractals are geometric objects, usually obtained through **recursion** and **self-similarity** (also referred to as "expanding symmetry" or "unfolding symmetry"). Roughly, this means that you observe the same or similar shapes and patterns as you keep zooming in on various parts of the picture. Let us begin with this concept, a basic property of fractals.

Self-Similarity

A shape is self-similar if each part has characteristics similar to the whole, showing identical or similar patterns at smaller and smaller scales.

Self-similarity can often be found in nature. In the Romanesco broccoli pictured below, if we zoom in on part of the image, the piece remaining looks similar to the whole.

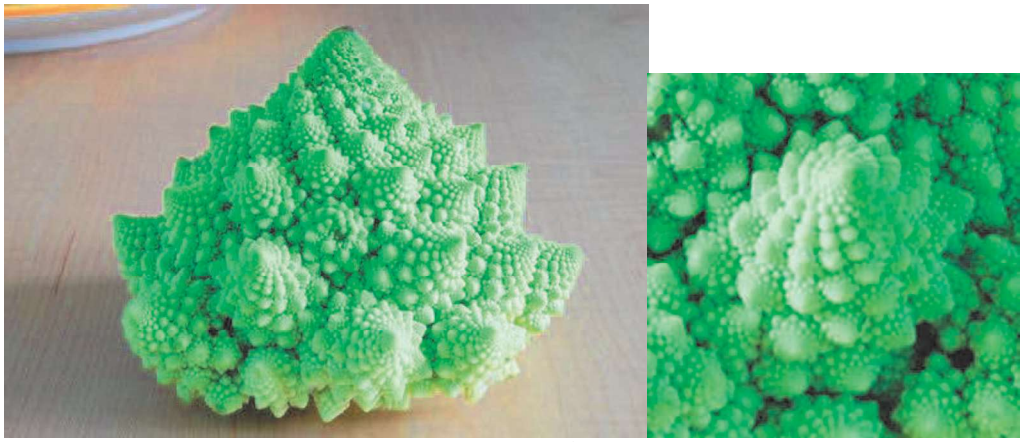


Figure 7.19[131]

[130] Maryam Mirzakhani is the first (and so far the only) woman to win the Fields Medal, often called the "Nobel Prize of mathematics." For the context of this quote and a story of her life, see https://www.theguardian.com/science/2014/aug/13/interview-maryam-mirzakhani-fields-medal-winner-mathematician. Unfortunately, she died of cancer at age 40.

[131] Image by AVM is licensed by CC BY-SA 3.0

Likewise, in the fern frond below, one piece of the frond looks similar to the whole.

Figure 7.20[132]

Similarly, if we zoom in on a typical coastline, each zoom reveals previously hidden detail; while not identical to the view from farther away, it does exhibit similar characteristics.

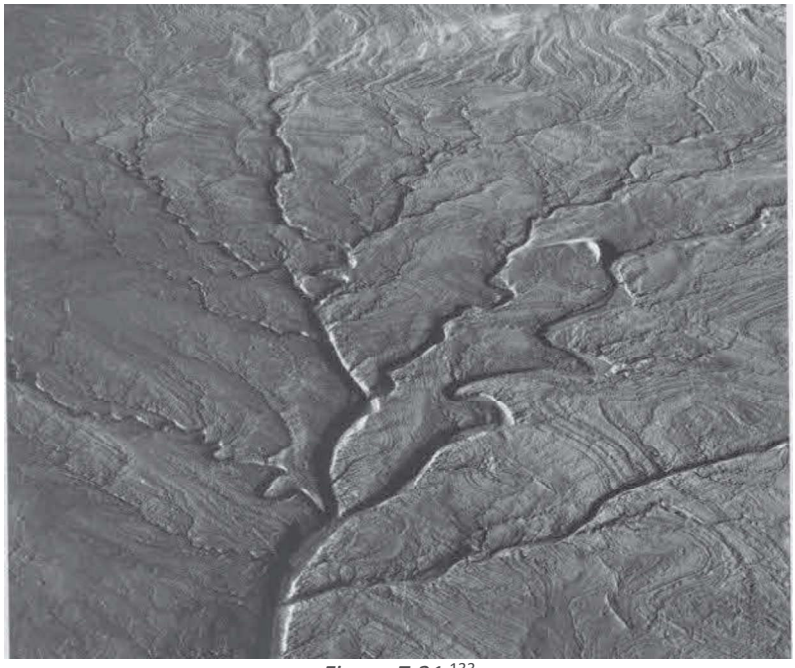


Figure 7.21 [133]

[132] Image by Carolyn Jewel is licensed by CC 2.0

[133] Image by Exploratorium Teacher Institute is licensed under CC BY-NC-SA 4.0

Iterated Fractals

This self-similar behavior can be replicated through **recursion** (or **iteration**): repeating a process over and over.

Example 1

Suppose that we start with a filled-in triangle. We connect the midpoints of each side and remove the middle triangle. Repeat this process.

[Solution]

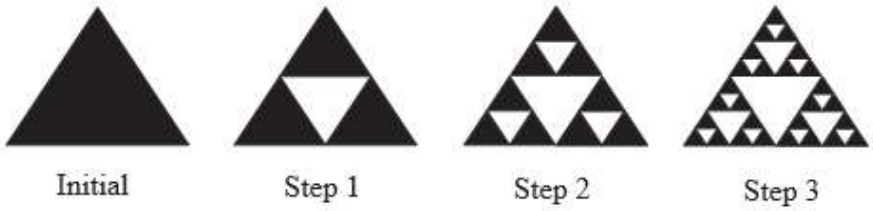


Figure 7.22[134]

If we repeat this process, the shape that emerges is called the **Sierpinski** (or the **Sierpinski triangle**). Notice that it exhibits self-similarity – any piece of the gasket will look identical to the whole. In fact, we can say that the Sierpinski gasket contains three copies of itself, each half as tall and wide as the original. Of course, each of those copies also contains three copies of itself.

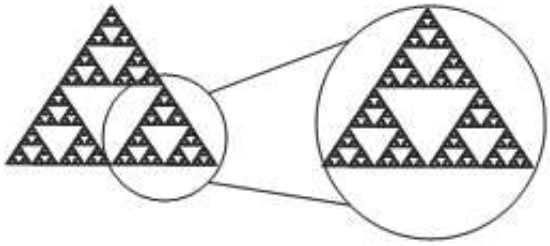


Figure 7.23[135]

Note that, if this process is carried out indefinitely (to the "end," to the true Sierpinski gasket), it does not have *any* solid black triangle left. Whatever you may think is a solid black triangle is actually one with infinitely many triangular areas missing.

Initiator and Generator

We can construct other fractals using a similar approach. To formalize this a bit, we are going to introduce the idea of **initiators** and **generators**.

[134] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[135] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

Initiators and Generators

An **initiator** is a starting shape, the stage where one begins.

A **generator** is the result of applying some action once on the initiator. This is what happens after one iteration, defining the process for the following stages.

To generate fractals from initiators and generators, we follow a simple rule:

Fractal Generation

Start with the initiator. At each step, replace every copy of the initiator with a scaled copy of the generator, rotating as necessary.

Example 2

Use the initiator and generator shown to create the iterated fractal.

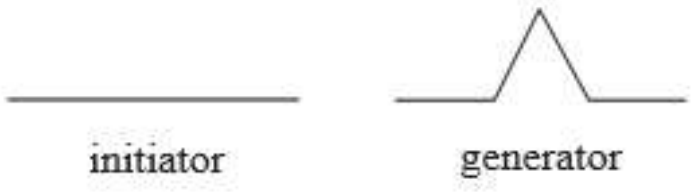


Figure 7.24[136]

[Solution]

This tells us to, at each step, replace each line segment with the spiked shape shown in the generator ("build a speed bump in the middle"). Notice that the generator itself is made up of 4 copies of the initiator (a line segment). As shown below, in step 1 (which is the generator itself), the single line segment in the initiator is replaced with the generator. For step 2, each of the four line segments of step 1 is replaced with a scaled copy of the generator:

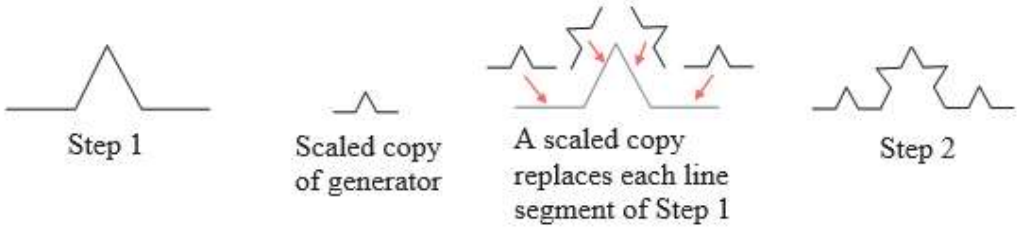


Figure 7.25[137]

[136] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[137] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

Now we have 16 short line segments. The same process is repeated to form Step 3. Again, each line segment is replaced with a scaled copy of the generator.

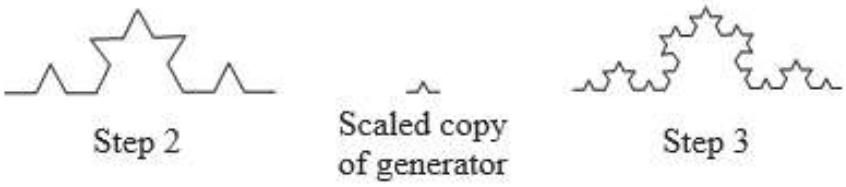


Figure 7.26[138]

Notice that since Step 0, the initiator, was only 1 line segment, Step 1 only required one copy of the generator. Since Step 1 had 4 line segments, Step 2 required 4 copies of the generator. Similarly, Step 3 required 16 copies of the generator, and Step 4 would require $16 \times 4 = 64$ copies of the generator.

Figure 7.27: Koch Curve[139]

The shape resulting from iterating this process is called the Koch curve, named for Helge von Koch, who first explored it in 1904. Here, again, the term "resulting shape" is a bit ambiguous because the process continues indefinitely (at least in theory). At each point, every segment gets shorter and shorter, so in the "end," technically there remains not a single line segment of any length on this curve.

Notice that the Sierpinski gasket can also be described using the initiator-generator approach:

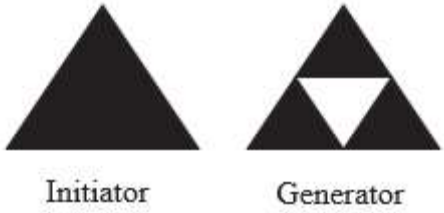


Figure 7.28[140]

[138] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[139] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[140] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

As mentioned in the previous example, sometimes the initiator is referred to as "Stage 0" and the generator as "Stage 1." In other words, Stage 1 is showing how Stage 0 gets transformed by the first iteration, and then the process is to be repeated.

Example 3

Use the initiator and generator below; however, only iterate on the "branches" (the two top splitting line segments, not the trunk). Sketch several steps of the iteration.

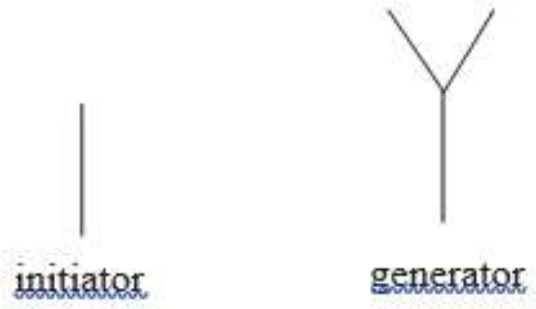


Figure 7.29[141]

[Solution]

We begin by replacing the initiator with the generator. We then replace each "branch" of Step 1 with a scaled copy of the generator to create Step 2.

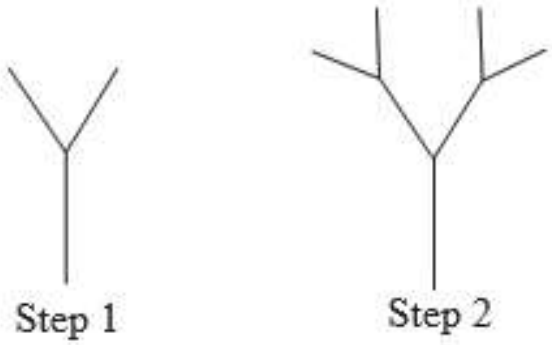


Figure 7.30[142]

We can repeat this process to create later steps. Repeating this process can create intricate tree shapes. Note that the number of branches on the end doubles each time (growing exponentially).

[141] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[142] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

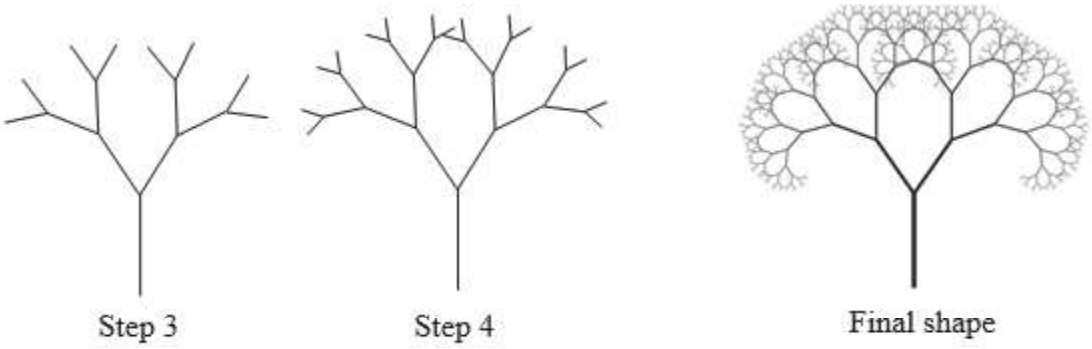


Figure 7.31[143]

Try It Now 1

Use the initiator and generator shown below to produce the next two stages. (Consider the Initiator as Step 0 and the Generator as Step 1.)

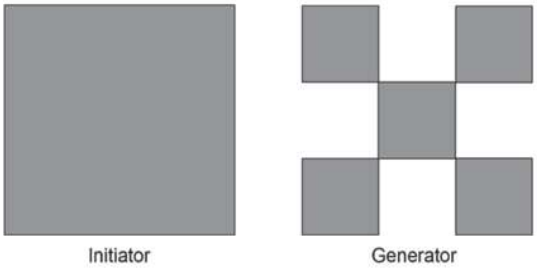


Figure 7.32[144]

When you carry out the iterative process, it may help you to verbalize (state in your own words) what is to be done. For instance, in Try It Now 1 above, you are to "divide each square into nine smaller squares and remove four of them." Some rules are more complicated, like the one below.

Try It Now 2

Use the initiator and generator shown to produce the next two stages.

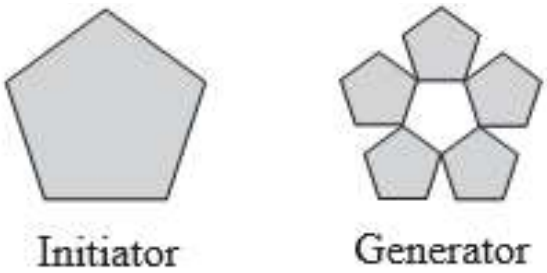


Figure 7.33[145]

[143] Image by Manuel Noah Angeja is licensed by CC 2.0
[144] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[145] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

Using iteration processes like those above can create a variety of beautiful images evocative of nature.

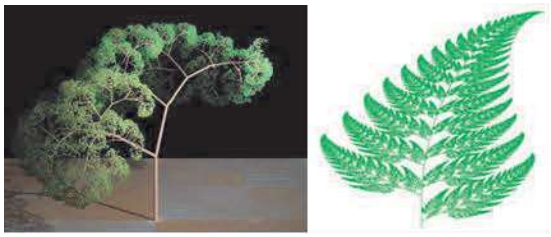


Figure 7.34[146]

The picture on the right is artificially created using an iteration process. More natural shapes can be created by adding in randomness to the steps.

Example 4

Create a variation on the Sierpinski gasket by randomly skewing the corner points each time an iteration is made.

[Solution]

Suppose we start with the triangle below. We begin, as before, by removing the middle triangle. We then add in some randomness (by changing some angles).

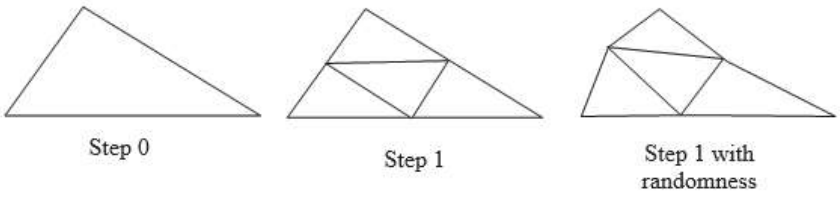


Figure 7.35[147]

We then repeat this process. Of course, there are many possible solutions to this problem since these changes are, well, random. This is just one example of what will happen. Continuing this process can create mountain-like structures.

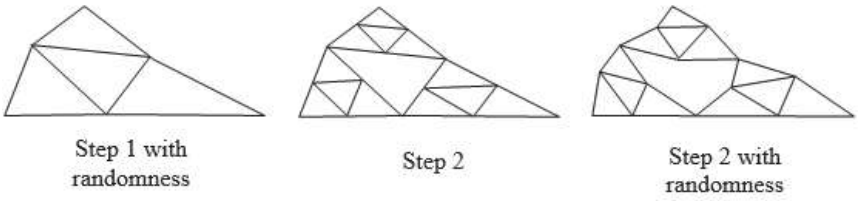


Figure 7.36[148]

[146] Image by Solkoll is in the public domain/Image by DSP-user is licensed by CC 3.0
[147] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[148] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

Fractals are also seen in works of art around the world. For instance, Hokusai, a Japanese printmaker artist (1760—1849), is famous for his great waves that are fractal-like. M.C. Escher, a Dutch artist (1898—1972), painted many art pieces containing fractal images, often reflecting advanced mathematical concepts such as tessellations, hyperbolic geometry, and transformations.

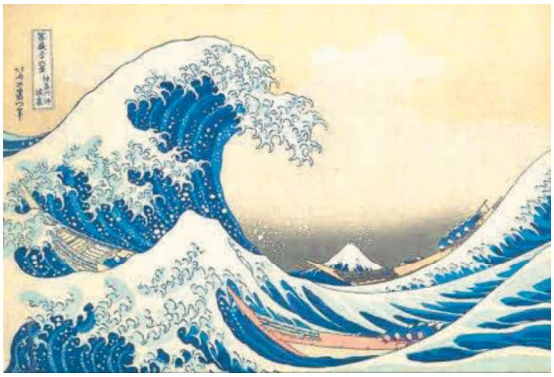


Figure 7.37[149]

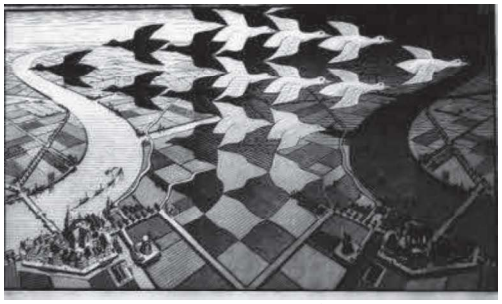


Figure 7.38[150]

More recently, equipped and empowered by high-speed computers, modern artists have been experimenting with fractals, iterations, and randomness. For instance, the landscape below was created using fractals first and then colored and textured.

Figure 7.39[151]

[149] Tsunami by Katsushika Hokusai is in the public domain.
[150] Day and Night (1938), by M.C. Escher, is licensed under CC BY 2.0.
[151] Image by The Ostrich is licensed by CC 3.0

Fractal Dimension

In addition to visual self-similarity, fractals exhibit other interesting properties. For example, notice that each step of the Sierpinski gasket iteration removes one quarter of the remaining area. If this process is continued indefinitely, we would end up essentially removing all the area, meaning we started with a 2-dimensional area and somehow end up with something less than that, but seemingly more than just a collection of 1-dimensional lines.

To explore this idea, we need to discuss dimension. Something like a line is 1-dimensional; it only has length. Any curve is 1-dimensional, even if it is located on the plane (2-dimensional) or in space (3-dimensional). Things like rectangles and parallelograms are 2-dimensional since they have length and height, describing an area. Objects like boxes and cylinders have length, width, and height, describing a volume, and they are 3-dimensional.

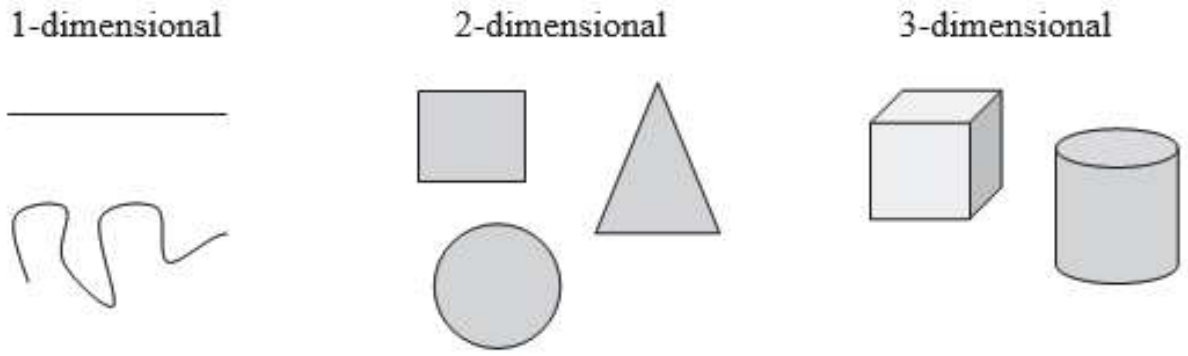


Figure 3.40[152]

Certain rules apply for scaling objects, and these rules are related to their dimension.

If you had a line with length 1 and wanted scale its length by 2, you would need two copies of the original line. If you had a line of length 1 and wanted to scale its length by 3, you would need three copies of the original.

Figure 3.41[153]

[152] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[153] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

If you had a rectangle with length 2 and height 1 and wanted to scale its length and width by 2 (doubling each dimension), you would actually need four ($2^2 = 4$) copies of the original rectangle. If you wanted to scale the length and width by 3, then you would need nine ($3^2 = 9$) copies of the original rectangle.

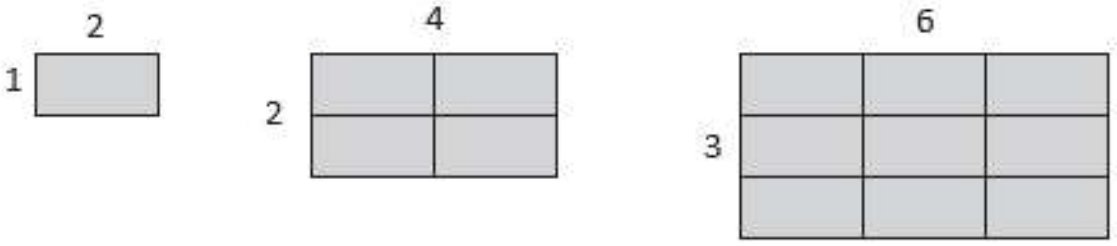


Figure 3.42[154]

If you had a cubical box with sides of length 1 and wanted to scale its length, width, and height by 2 (doubling all dimensions), you would need eight ($2^3 = 8$) copies of the original cube. If you wanted to scale the length, width, and height by 3, I would need 27 ($3^3 = 27$) copies of the original cube.

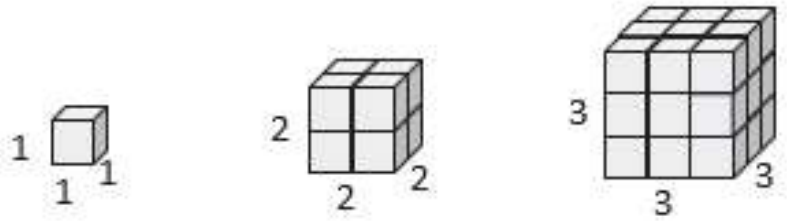


Figure 3.43[155]

We studied this back in Chapter 2. If you have a pizza, doubling its radius gives you a pizza 4 times as large. If you have a softball, increasing its radius by 50% (0.5) would make the volume more than 3 times as large as the original softball because volume $(1.5)^3 = 3.375$. Here, too, the exponent expresses the dimension.

Notice that in the 1-dimensional case, the number of copies needed is equal to the scale. In the 2-dimensional case, the number of copies needed is equal to the square of the scale. In the 3-dimensional case, the number of copies needed is equal to the cube of the scale. From these examples, we might infer a pattern. In fact, in this section, we use this pattern to define the **dimension** of an object.

[154] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[155] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

Scaling-Dimension Relation

To scale a D-dimensional shape by a scaling factor S, the number of copies C of the original shape needed is given by the following:

$$\text{Copies} = \text{Scale}^{\text{Dimension}}, \text{ or } C = S^D.$$

So the exponent required is the dimension.

Example 5

Use the scaling-dimension relation to determine the dimension of the Sierpinski gasket.

[Solution]

Suppose we define the original gasket to have side length 1. The larger gasket shown is twice as wide and twice as tall, so has been scaled by a factor of 2.

Notice that to construct the larger gasket, 3 copies of the original gasket are needed.

Figure 3.44[156]

Using the scaling-dimension relation $C = S^D$, we obtain the equation $3 = 2^D$. Since $2^1 = 2$ and $2^2 = 4$, we can immediately see that *D* is somewhere between 1 and 2. The gasket is more than a 1-dimensional shape, but we have taken away so much area that it is now less than 2-dimensional.

Solving the equation $3 = 2^D$ requires logarithms, which we do not cover in detail in this book. But it turns out that

$$D = \frac{\log 3}{\log 2} \approx 1.585.$$

Therefore, the dimension of the Sierpinski gasket is about 1.585. In fact, many fractals have irrational-number dimensions.

[156] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

Without going into the details of logarithmic operations, we can simply state the following (using the common logarithm, which we studied in Chapter 6):

Scaling-Dimension Relation to Find Dimension

To find the dimension D of a fractal, determine the scaling factor S and the number of copies C of the original shape needed, and then use the following:

$$D = \frac{\log C}{\log S}$$

$$D = \frac{\log(C)}{\log(S)}$$

Some Well-Known Fractals

We will now briefly mention another type of fractal, defined by a different type of recursion.

The first is what is known as the **Mandelbrot set**, defined using the iteration of a certain complex-valued function. The mathematical details are beyond the scope of this text, so we will not include them here. Note that this well-known set, discovered in 1980, also shows a fractal-like shape. The boundary of this shape exhibits quasi-self-similarity in that portions look very similar to the whole. In addition to coloring the Mandelbrot set itself black, it is common to color the points on the boundary of the set.

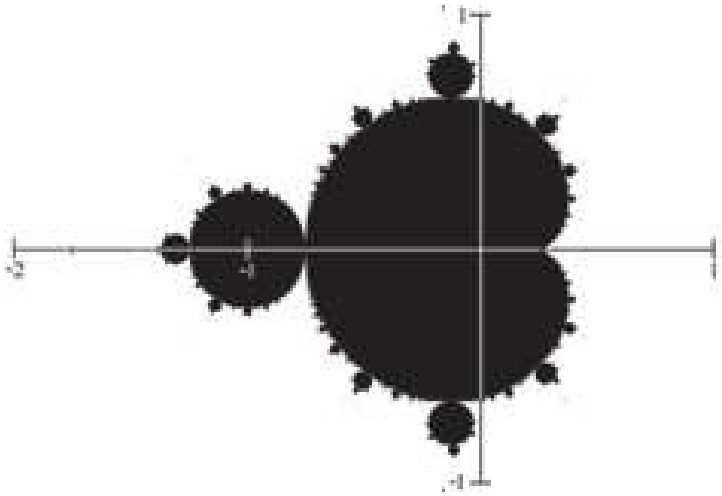


Figure 3.45[157]

[157] Image by Connelly Barnes is in the public domain

These colorings have specific mathematical meanings. Some functions were defined on the set of complex numbers, which include the imaginary unit $i = \sqrt{-1}$, and they were used to assign some colors to certain points on the plane. The results were spectacular, generating these pretty images, which astonished researchers as well as the general public all over the world.

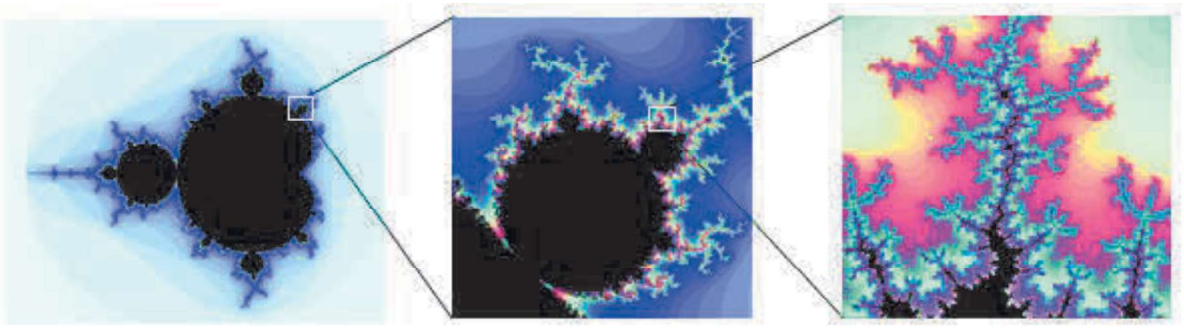


Figure 3.46 [158]

The Mandelbrot set, for having a simple definition, exhibits immense complexity. Zooming in on other portions of the set yields fascinating swirling shapes.

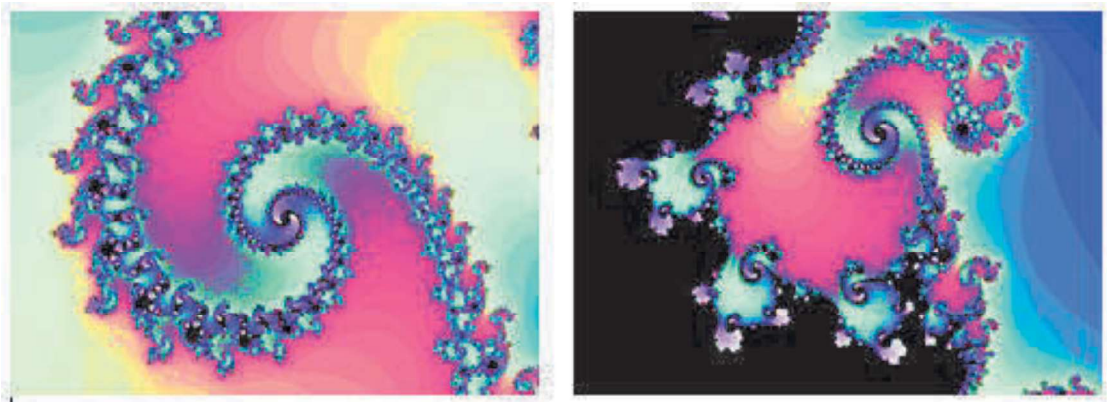


Figure 3.47[159]

The Mandelbrot Explorer site provides more details on the Mandelbrot set, including a nice visualization of Mandelbrot sequences.

[158] Series created using Scott's Mandelbrot Set Explorer, licensed by CC 3.0

[159] *Series created using Scott's Mandelbrot Set Explorer*, licensed by *CC 3.0*

Exercises 7.4

Using the initiator and generator shown for each exercise below, draw the next two stages of the iterated fractal.

1. "A little bump forming in the middle of each segment"

Figure 3.48[160]

2. "Zig-zag for each segment" (Each segment is replaced by three shorter zig-zag segments. The segment being replaced is drawn in light blue below, just as a guide.)

Figure 3.49[161]

3. "Three small branches coming from each segment" (But once three branches come out of a segment, that segment will not grow any more branches in subsequent stages.)

Figure 3.50[162]

4. "Cut out the middle third": This process generates what is known as "**Cantor's Disappearing Table**." You begin with a segment, say from 0 to 1 on the number line (Stage 0). Then, you remove the middle third, from 1/3 to 2/3 (the two end points are not removed). Now (Stage 1) you have two segments, each with length 1/3. You now remove the middle third of each of the two remaining pieces. You will then

[160] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[161] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[162] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

end up with four short segments, each with length 1/9 (Stage 2). How many segments are there, and how long are the segments in Stage n? Try to describe what the "end" result looks like if this process is continued indefinitely. The end result, the set of all remaining points, is called the **Cantor Set** (The blue dotted line shows the portion removed.)

Figure 3.51[163]

5. "Remove the middle square": This process generates what is known as the "**Sierpinski carpet**." Draw the next stage.

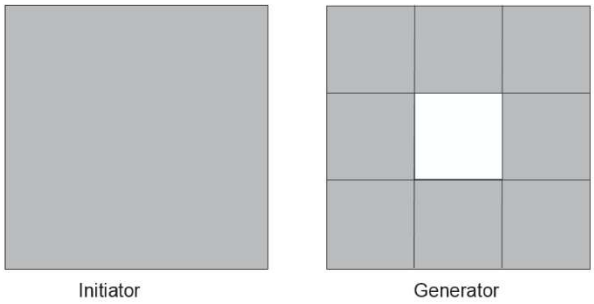


Figure 3.52[164]

6. Create your own version of Sierpinski gasket with added randomness. Draw the next stage.

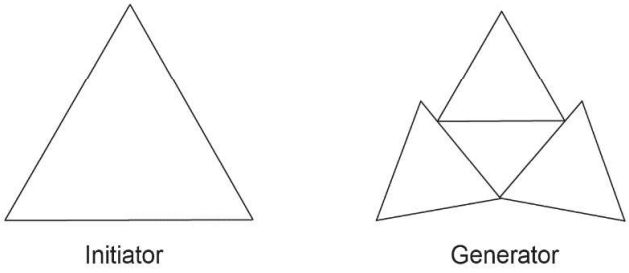


Figure 3.53[165]

7. Create a version of the branching tree fractal from Example 3 (in the text, not Exercise 3 here) with added randomness.

[163] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[164] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.
[165] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

8. Recall that the initiator and the generator of the Koch curve are as shown below. The generator has four copies of the initiator, so $C = 4$. If you compare the length of each segment, however, it has become 1/3 of the original length. In other words, to scale the original shape by scale $S = 3$, you need 4 copies of the original shape. Now, determine the fractal dimension of the Koch curve.

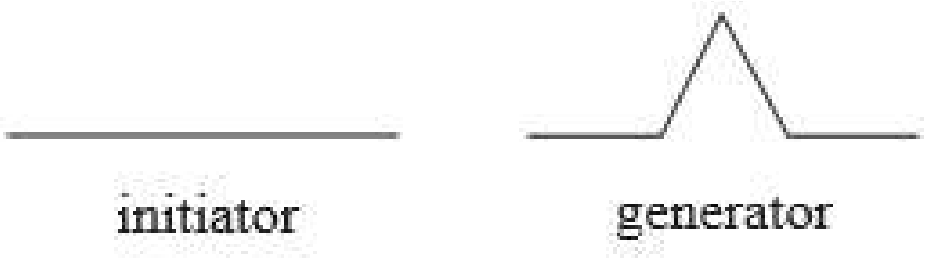


Figure 3.54[166]

9. Determine the fractal dimension of the curve generated in Exercise 1 above. (Note that here, to scale it by the factor of 3, you need 5 segments.)

10. Determine the fractal dimension of the Sierpinski carpet generated in Exercise 5. (Note that here, to scale it by the factor of 3 in each dimension, you need 8 copies of the original square.)

11. Determine the fractal dimension of Cantor's Disappearing Table (otherwise known as the "Cantor set") generated in Exercise 4.

12. Do some online research on Cantor's Disappearing Table (or the Cantor Set) generated in Exercise 4. If this process of removing the middle third of the "table" is continued indefinitely, what will be left in the end? Is there anything left? If so, how many pieces? How long is each piece left unremoved? What is the significance of this set remaining?

[166] Fractals by David Lippman & Melonie Rasmussen is licensed under CC BY-SA 3.0 US.

For Further Explorations…

In Section 7.2, we introduced the Golden Ratio ϕ, defined as the positive solution to the equation

$$\phi^2 - \phi - 1 = 0.$$

In addition to being the "most aesthetically pleasing" ratio of rectangles, this number has other interesting mathematical properties. For example, you will get the same answer when you subtract 1 from it or divide 1 by it as you see below:

$$\phi - 1 = 1.6180339887\ldots - 1 = 0.6180339887.$$

$$\frac{1}{\phi} = \frac{1}{1.6180339887\ldots} = 0.6180339887\ldots$$

In other words, the Golden Ratio minus 1 is the reciprocal (multiplicative inverse) of the Golden Ratio. This is easy to check using algebra. If you divide each term of the equation $\phi^2 - \phi - 1 = 0$ by ϕ, you get

$$\phi - 1 - \frac{1}{\phi} = 0.$$

This immediately gives the fact that $\phi - 1 = \frac{1}{\phi}$.

Also, assume that the figure below represents a golden rectangle (blue and pink together), i.e., the ratio of the whole rectangle is $\frac{a}{b} = \phi$. In the figure in Section 7.2, we *added* a square to find what ϕ is. This time, suppose we cut off a square (B) from one end of this rectangle. It turns out that the remaining rectangle (A) is also a golden rectangle as the following algebra shows:

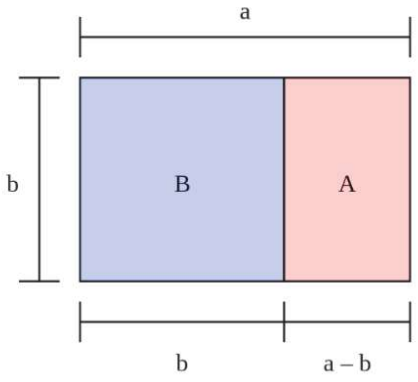


Figure 7.55 [167]

[167] Image by Peter John Acklam is licensed by CC-3.0.

Since $\phi = \frac{a}{b}$, and ϕ is a solution to the quadratic equation $\phi^2 - \phi - 1 = 0$, we have $\frac{a^2}{b^2} - \frac{a}{b} - \frac{b}{b} = 0$. Multiplying both sides by b^2, we obtain

$$a^2 - ab - b^2 = 0.$$

Then we get $a^2 - ab = b^2$, from which we also get, by factoring,

$$a(a - b) = b^2.$$

We finally divide both sides by $b\ (a - b)$ to obtain

$$\frac{a}{b} = \frac{b}{a - b}.$$

Note that both sides are equal to ϕ, proving that both A and the entire picture are both golden rectangles.

Chapter 8: Mathematics and Politics

"Give all the power to the many, they will oppress the few. Give all the power to the few, they will oppress the many."—Alexander Hamilton

One of the main goals of learning mathematics (particularly Liberal Arts Mathematics) is to become logically and quantitatively educated so that you can be well-informed, contributing members of the society. To that end, this chapter is particularly important as we will explore how mathematics is used (and misused) in politics, especially in two crucial aspects of democracy—election and representation. We will investigate some fundamental facets of politics as well as our pursuit of fairness and justice (and of happiness, of course).

We first go back in time—to the Founding Fathers of this country, and a little drama involving conflicting personalities.

Figure 8.1: Alexander Hamilton on the ten-dollar bill[168]

[168] Photo by Ryan Quintal on Unsplash

8.1 Apportionment[xix]

Say your school district has 5 schools, some with a lot more students than others. The state budget allows the district to have 18 counselors. Assuming that each counselor can only serve one school, which school gets how many? What if a healthcare provider has 4 hospitals, some serving much bigger communities than others, and the network has to allocate their 70 nurses proportionally? These are some practical applications of the apportionment problem.

Apportionment is the process of dividing up a fixed number of people or resources among groups of different sizes. While the term "apportionment" can be used for allocating resources like money or employees to various branches or offices within one organization (e.g., funds to different campuses in a college system or tellers to different branches of a bank), in **politics**, this specifically refers to allocating a limited number of representatives among voters. The best-known ways to solve this problem have their origins in the formation of the House of Representatives in the new Congress when the United States was founded. States also face this apportionment problem in defining how to draw districts for state representatives.

So, while apportionment has a wide range of applications in government and non-government sectors, we will use the language of politics and congressional representation in this section for consistency. Of course, since the mathematical principles remain the same, they can be applied in other areas.

Apportionment Rules (for Congressional Representation)

1. The number of representatives for each state must be **whole numbers** (i.e., no part-time or fractional assignments).
2. The total number of representatives is fixed (no less or no more than determined by law).
3. Each state must get at least one representative.
4. The number of representatives must be approximately proportional to the state population. (If State A is more populous than State B, then State A cannot have fewer representatives than State B.)

The current House of Representatives has 435 members (the number fixed by the Permanent Apportionment Act of 1929).

So what, exactly, is the problem? To illustrate the difficulty in apportionment, let us look at a simple example. (We will use this same example a few times.)

Example 1

Consider a tiny country with three states: Aurora (A), Bay (B), and Castle (C). Suppose that their congress/assembly is to have 41 members and wants to divide this representation based on the population of each state, shown below.

State	Population
Aurora	162,310
Bay	538,479
Castle	197,145
Total	**897,934**

Now, we are told that these 897,934 people are to be represented by 41 lawmakers. It would then make sense to find out (roughly) how many people each lawmaker should represent. So we calculate what is known as the **divisor** (or the **standard divisor**). This is the first step in all of the apportionment methods that follow.

Standard Divisor in Apportionment

If the total population P is to be represented by N representatives, then the (standard) **divisor** d is the quotient

$$d = \frac{P}{N}$$

This is the number of people represented by each representative.

For this example, we see that

$$d = \frac{897{,}934}{41} \approx 21{,}900.82927.$$

Now that we know how many people each lawmaker is to represent, we can simply divide the population of each state by the standard divisor d to determine the number of representatives for each state. That number is called the **quota** for each state.

State	Population	Quota
Aurora	162,310	7.4111
Bay	538,479	24.5872
Castle	197,145	9.0017
Total	**897,934**	

By now, you may have noticed the problem. These quotas are all decimal numbers. It is hardly surprising that these quotas are not nice, whole numbers. So how many lawmakers should each state get? If you take only the whole number part of each quota (thus **truncating** or ignoring the digits after the decimal point), Aurora gets 7, Bay gets 24, and Castle gets 9; the total would be 40, not 41. So which state gets that one last coveted seat in the congress? *That* is the problem in apportionment. Said differently, the problem is that *you cannot split people* (you cannot have 7.41 representatives, for example).

We will now go back in history to look at a few ways of solving the apportionment problem. All of them have been used at various times to apportion the U.S. Congress although the method currently in use (the Huntington-Hill method) is significantly more complicated.

Hamilton's Method

Alexander Hamilton, now made famous by the 2015 Broadway musical, proposed the method that now bears his name. His method was approved by Congress in 1791 but was vetoed by President Washington. (This, by the way, was the very first presidential veto in the United States). It was later adopted, after his death, in 1852 and was used through 1911. This may be the most intuitive method.

Hamilton's Method

1. As shown in the example above, divide each state's population by the **divisor** d to determine the **quota** for the state. These quotas are, in general, decimal numbers.
2. **Separate** the whole-number part of each quota from its decimal part. For each state, the whole-number part is called the **minimum quota** (or **lower quota**), and the decimal part is called the **remainder**.
3. Add up the minimum quotas. If the sum is equal to the total number of representatives, we are done; however, this is extremely unlikely since we have truncated (removed) all the remainders.
4. If the sum of the minimum quotas is less than the total number of representatives, assign the remaining representative(s), one each, to the states whose remainders were the largest, until the desired total is reached.

Note on rounding: Today we have technological advantages that Hamilton (and the others) could not even have imagined. Take advantage of them, and keep several decimal places.

Example 2

Consider, again, the country with three states: Aurora (A), Bay (B), and Castle (C), with 41 assembly seats. Their populations are shown below. We have already calculated the standard divisor $d \approx 21{,}900.82927$ and the quota for each state. Use Hamilton's method to determine the number of representatives to be assigned to each state.

[Solution]

Note that we already have the quota for each state, so we now add two columns: *minimum quota* and *remainder*. Simply split the quota at the decimal point.

State	Population	Quota	Min. Q	Remainder
Aurora	162,310	7.4111	7	.4111
Bay	538,479	24.5872	24	. 5872
Castle	197,145	9.0017	9	. 0017
Total	**897,934**		**40**	

Note that the sum of the minimum quotas is 40, so there is one more seat to be assigned to one of the states. According to Hamilton's method, the state with the largest remainder, Bay (with .5872), should get that extra seat, as follows:

State	Population	Quota	Min. Q	Remainder	Final
Aurora	162,310	7.4111	7	.4111	7
Bay	538,479	24.5872	24	. 5872	**25**
Castle	197,145	9.0017	9	. 0017	9
Total	**897,934**		**40**		**41**

Now, this should make sense, right? Here, Castle had a quota just a little bit above 9 while Bay had more than .5 above 24. It seems quite reasonable to award that last representative to Bay. This, to Hamilton and many other Founders, seemed quite intuitive and fair.

Example 3

Use Hamilton's method to apportion the 75 DMV employees of a state that has five DMV offices as shown.

Office	Population
Amber	49,875
Birchwood	166,158
Chester	82,888
Dogwood	626,667
Elm	126,979
Total	**1,052,567**

[Solution]

First, we find the standard divisor: $d = \frac{1{,}052{,}567}{75} \approx 14{,}034.22667$ (suggesting that there should be one DMV employee for each group of about 14,000 residents of the state). We then divide the population served by each office by this divisor to calculate the quota. Then we split that number into the minimum quota (the whole-number part) and the remainder.

Office	Population	Quota	Min. Q	Remainder
Amber	49,875	3.5538	3	.5538
Birchwood	166,158	11.8395	11	.8395
Chester	82,888	5.9061	5	.9061
Dogwood	626,667	44.6528	44	.6528
Elm	126,979	9.0478	9	.0478
Total	**1,052,567**		**72**	

We see that the minimum quotas add up to 72 while we are to apportion 75 employees. This means that three of these offices will get one extra employee each, according to their remainder. Chester (.9061), Birchwood (.8395), and Dogwood (.6528) have the three highest remainders, so each of these offices will get one more than the minimum quota. Warning: do not give *three extra* to Chester. The top three offices will get *one each*.

Office	Population	Quota	Min. Q	Remainder	Final
Amber	49,875	3.5538	3	.5538	**3**
Birchwood	166,158	11.8395	11	.8395	**12**
Chester	82,888	5.9061	5	.9061	**6**
Dogwood	626,667	44.6528	44	.6528	**45**
Elm	126,979	9.0478	9	.0478	**9**
Total	**1,052,567**		**72**		**75**

Note that even though Amber's remainder is greater than .5, it is not big enough to get an additional employee because three other offices have greater remainders.

Hamilton's method obeys something called the Quota Rule. The Quota Rule is not a law of any sort, but it is just an idea that some people, including Hamilton, think is a good one to ensure fairness.

Quota Rule

The Quota Rule says that the final number of representatives a state gets should be within one of that state's quota. Since we are dealing with whole numbers for our final answers, each state should either go up to the next whole number above its quota or down to the next whole number below its quota.

Controversy

After seeing Hamilton's method, many people find that it makes sense. It is not that difficult to use, and they wonder why anyone would want another method. The problem is that Hamilton's method is subject to several **paradoxes**. Three of them happened, on separate

occasions, when Hamilton's method was used to apportion the United States House of Representatives.

The **Alabama Paradox** is named for an incident that happened during the apportionment after the 1880 Census. (A similar incident had happened ten years earlier involving Rhode Island.) The post-1880 apportionment had been completed, using Hamilton's method applied to the new population numbers from the census. Then it was decided that because of the country's growing population (remember, this was shortly after the Civil War), the House of Representatives should be made larger. That meant that the apportionment calculations would need to be done again, still using Hamilton's method and the same 1880 Census numbers, but with more representatives nationwide. The assumption was that some states would gain another representative and others would stay with the same number they already had (since there were not enough new representatives being added to give one more to every state). The paradox is that *Alabama ended up losing a representative in the process* even though no populations were changed and the total number of representatives increased. Alabama may be a sweet home, but they were certainly not happy then!

The **New States Paradox** happened when Oklahoma became a state in 1907. Oklahoma had enough population to qualify for five representatives in Congress. But since the size of the House of Representatives overall did not increase, those five representatives would need to come from somewhere else. Five states, presumably, would lose one representative each. That happened, but another unexpected thing also happened: Maine surprisingly *gained* a representative (from New York)! Arithmetic with decimal numbers can be tricky.

The **Population Paradox** is a situation in which a state with a higher growth rate loses a seat to another state with a lower growth rate. It happened between the 1900 Census and the 1910 Census. In those ten years, Virginia's population grew at an average annual rate of 1.07% while Maine's grew at an average annual rate of 0.67%. Virginia started with more people, grew at a faster rate, grew by more people, and ended up with more people than Maine. By itself, that does not mean that Virginia should gain representatives or Maine should not (because there are lots of other states involved). But Virginia ended up *losing* a representative to Maine.[169]

Note that none of these paradoxes occurred until after the Civil War, and it is unlikely that the Founding Fathers were even aware of the possibility that these seemingly unfair incidents could take place. Remember, the first of these paradoxes was the Alabama Paradox, which occurred after the 1880 Census, almost a century after Hamilton invented the method named after him. That is partly because Hamilton's method was actually never used during his lifetime; he died in 1804, and Hamilton's method was not adopted until 1850. Why is that?

[169] As seen in this example, the fact that one state loses a seat does not mean a population *decrease*; in fact, it does not even mean that the state's population is increasing more slowly than those states which gains a seat as illustrated by Virginia and Maine in this example. Of course, slow-growing states *can* (and do) often lose a seat, such as California and New York in the 2020 census.

Well, there were a few people who did not like Hamilton all that much. Among them was a well-known and well-respected Virginian named Thomas Jefferson. An accomplished architect and a student of mathematics, Jefferson knew that Hamilton's method generally favored smaller states. So Jefferson devised another method of apportionment, one that would favor bigger states like his own, Virginia. Conveniently, even though a newly established Congress had already passed a bill approving Hamilton's method, Jefferson had a good friend, a fellow Virginian (and the first President of the United States) who could **veto** the bill. Jefferson may have said something like, "George, you need to veto this bill that favors smaller states. Let's make Virginia great again." Washington happily agreed, and in 1792 the first-ever veto in the United States prevented Hamilton's method from being used for congressional apportionment.

Here is the alternative method Thomas Jefferson came up with (and was actually used from 1792 to 1840).

Jefferson's Method

Thomas Jefferson argued that it is the ratio of people to representatives that is the critical thing and thus apportionment methods should be based on that fact. Of course, we now know the paradoxes listed above also provide mathematical reasons for not using Hamilton's method. While Jefferson's method might or might not be the best one to replace it, at least Jefferson provided an alternative.

Jefferson's method starts off in ways quite similar to Hamilton's method. We find the standard divisor d, use it to find the quotas, and split each quota into the minimum quota and the remainder. The sum of the minimum quotas is very likely to be less than the total number of representatives.

The difference is in how Jefferson resolved that problem. He argued that since we ended up with a sum that is too small, our divisor must have been too big. Instead of using the remainder, Jefferson reduced the divisor itself, as outlined below.

Jefferson's Method

1. Divide each state's population by the standard divisor d to determine the quota (a decimal number) and the **minimum quota** (a whole number) for each state as before.
2. Add up the minimum quotas. In the extremely unlikely event that the sum is equal to the total number of representatives, we are done.
3. If the sum is less than the total number, **reduce** the divisor (the new divisor is called a **modified divisor** or an **adjusted divisor**) and recalculate the minimum quota for each state. Repeat this process, if necessary, until you find a modified divisor that gives the sum of the minimum quotas equal to the total number of representatives.

In other words, Jefferson adjusted the divisor itself while Hamilton never did that but instead used the remainders to award extra seats.

A careful reader should be able to spot the problem immediately. Jefferson's method does not state exactly how to reduce that devisor. It involves trial-and-error. You keep lowering the divisor until a desired outcome is achieved. **Do not overdo it!** Reduce the divisor just enough to get the right total. Let us take a look at some examples.

Example 4

We will return to an earlier example involving three states in a country with 41 legislative seats. Apply Jefferson's method now.

[Solution]

We begin, as we did with Hamilton's method, by finding the quotas (and minimum quotas) with the original divisor, $d \approx 21{,}900.82927$. Below, all the numbers are identical as before:

State	Population	Quota	Min. Q
Aurora	162,310	7.4111	7
Bay	538,479	24.5872	24
Castle	197,145	9.0017	9
Total	**897,934**		**40**

We need 41 representatives, and this divisor (21,900.82927) gives only 40. We must now reduce the divisor until we get 41 representatives. Let us try 21,500 as a modified divisor:

State	Population	Quota	Min. Q
Aurora	162,310	7.5493	7
Bay	538,479	25.0455	**25**
Castle	197,145	9.1695	9
Total	**897,934**		**41**

This gives us the required 41 representatives, so we are done. If we had fewer than 41, we would need to reduce the divisor more. If we had more than 41, we would need to choose a divisor less than the original but greater than the second choice.

Remember, the total number (41 in this example) should *not* be changed. Notice that with the new, lower divisor, the quota for Bay (the largest state in the country) increased by *much more* than those of smaller states.

Example 5

Apply Jefferson's method for the DMV example.

[Solution]

The original divisor of $d \approx 14{,}034.22667$ gave these results:

Office	Population	Quota	Min. Q
Amber	49,875	3.5538	3
Birchwood	166,158	11.8395	11
Chester	82,888	5.9061	5
Dogwood	626,667	44.6528	44
Elm	126,979	9.0478	9
Total	**1,052,567**		**72**

We need to apportion 75 employees but have only 72 here, so we need to use a smaller divisor. Let us try 13,500:

Office	Population	Quota	Min. Q
Amber	49,875	3.6944	3
Birchwood	166,158	12.3080	12
Chester	82,888	6.1399	6
Dogwood	626,667	46.4198	46
Elm	126,979	9.4059	9
Total	**1,052,567**		**76**

Oops! We have gone too far. We need a divisor that is greater than 13,500 but less than 14,034.22667. Let us try 13,700:

Office	Population	Quota	Min. Q
Amber	49,875	3.6405	3
Birchwood	166,158	12.1283	12
Chester	82,888	6.0502	6
Dogwood	626,667	45.7421	45
Elm	126,979	9.2685	9
Total	**1,052,567**		**75**

This works.

Divisor-adjusting methods like Jefferson's are not guaranteed to follow the Quotient Rule.

Webster's Method

Daniel Webster (1782-1852), to be distinguished from Noah Webster, the lexicographer, proposed a method similar to Jefferson's in 1832. It was adopted by Congress in 1842 but was replaced by Hamilton's method in 1852. It was then adopted again in 1901. Webster **rounded (up or down)** the quotas to the nearest whole number rather than truncating (rounding **down**) the remainders. If that does not produce the desired results at the beginning, we adjust the divisor until it does. (In Jefferson's case, the first adjustment will always be to make the divisor smaller. That is not always the case with Webster's method.)

Webster's Method

1. As before, divide each state's population by the standard divisor d to determine the quota, generally a decimal number, for each state.
2. Round (**up or down**) all the quotas to the nearest whole number. Add them up.
3. If the sum is equal to the number of representatives, we are done. If the sum is less than the total number of representatives, **reduce** the divisor and recalculate the quotas (rounded to the nearest whole number). If the sum is larger than the total number of representatives, **increase** the divisor and recalculate the quotas (rounded to the nearest whole number).
4. Continue this, if necessary, until the sum is equal to the total number of representatives. The divisor we end up using is called the **modified** divisor or **adjusted** divisor.

As in Jefferson's method, this last process requires some trial-and-error and may have to be repeated. Imagine how time-consuming this was, given that there were 50 states and there were no calculators or spreadsheet programs available for use.

Example 6

Use Webster's method for the 3-state country example above, with 41 congressional seats.

[Solution]

Again, we need to apportion 41 lawmakers with an initial divisor of $d \approx 21{,}900.82927$.

State	Population	Quota	Rounded Q.
Aurora	162,310	7.4111	7 (Rounded down)
Bay	538,479	24.5872	25 (Rounded up)
Castle	197,145	9.0017	9 (Rounded down)
Total	**897,934**		**41**

This gives the required total, so we are done. No adjustment was necessary here.

Example 7

Apply Webster's method to the DMV example above, apportioning 75 employees.

[Solution]

Again, with an initial divisor of 14,034.22667:

Office	Population	Quota	Rounded. Q
Amber	49,875	3.5538	4
Birchwood	166,158	11.8395	12
Chester	82,888	5.9061	6
Dogwood	626,667	44.6528	45
Elm	126,979	9.0478	9
Total	**1,052,567**		**76**

Note that this time, all the remainders except the last one for Elm were at least 0.5, so those quotas were all rounded up to the next whole number. This is too many, so we need to *increase* the divisor. Let us try 14,100:

Office	Population	Quota	Rounded. Q
Amber	49,875	3.5372	4
Birchwood	166,158	11.7843	12
Chester	82,888	5.8786	6
Dogwood	626,667	44.**4445**	**44** (Rounded *down* now)
Elm	126,979	9.0056	9
Total	**1,052,567**		**75**

This works, so we are done.

Like Jefferson's method, Webster's method carries a bias in favor of states (or offices) with large populations, but rounding the quotas to the nearest whole number greatly reduces this bias. (Notice that the Dogwood Office, the largest, is the one that loses an employee because of the increased quota.) Also, like Jefferson's method, Webster's method does not always follow the Quota Rule, but it can be shown to follow the Quota Rule much more often than Jefferson's method does. (In fact, if Webster's method had been applied to every apportionment of Congress in all of American history, it would have followed the Quota Rule every single time.)

So far, we have studied three methods of apportionment (and there are more), and each appears to be less than perfect (some could result in paradoxes, and some violate the Quota Rule). What is a perfect then? Is there an apportionment method that guarantees fairness? In 1980, two mathematicians, Peyton Young and Mike Balinski, proved what we now call the Balinski-Young Impossibility Theorem.

Balinski-Young Impossibility Theorem

Any apportionment method which always follows the Quota Rule is subject to the possibility of paradoxes like the Alabama, New States, or Population Paradoxes. Methods that avoid those paradoxes exist, but they may violate the Quota Rule.

In other words, Balinski and Young proved what people were afraid of all along; there is no apportionment method that always guarantees fairness (where "fairness" is defined in terms of certain rules and paradoxes).

When one says, "Life is not fair!" there is some truth to that statement. Perhaps Alexander Hamilton may have seen this coming; he said, "I never expect to see a perfect work from an imperfect man."[170]

Huntington-Hill Method

In 1920, no new apportionment was done because Congress could not agree on the method to be used (a gridlocked congress is not a new thing). They appointed a committee of mathematicians to investigate, and they recommended a new method called the **Huntington-Hill Method** (named after Edward V. Huntington and Joseph A. Hill). They continued to use Webster's method in 1931, but after a second report recommending Huntington-Hill, it was adopted in 1941 (during World War II). This is the method used for apportionment in Congress ever since.[171] (This method is referred to the "**Method of Equal Proportions**" in Title 2 of the U.C. Code, Sections 2a and 2b.)

The Huntington-Hill Method is similar to Webster's method, but it attempts to minimize the percent differences of how many people are represented by each Congress member by using what is known as the **geometric mean** of two numbers.

Geometric Mean

The geometric mean of two positive numbers, x and y, is the square root of their product, $\sqrt{xy}$.

If there is a *rectangle* with length x and width y, and if you want to construct a *square* with the same area as that rectangle, how long should each side of the square be? The answer is $\sqrt{xy}$. This is why the number is called the **geometric mean**.

What the Huntington-Hill Method does is to replace the **standard rounding** (which is based on the arithmetic mean) used in Webster's method with "**rounding based on the geometric mean**." When "rounding based on the geometric mean," any number between the consecutive whole numbers n and $n+1$ is rounded (down) to n if it is less than the geometric

[170] Alexander Hamilton. *Federalist Papers No. 85.* Retrieved from https://en.wikiquote.org/wiki/The_Federalist

[171] The Huntington-Hill Method, as defined in this section, is similar to Webster's method. However, for congressional apportionment, the United States uses a slightly modified version of the Huntington-Hill Method because of the requirement that every state must have at least one representative. See the "For Further Explorations…" section at the end of this chapter fore more details.

mean of n and $n+1$ and rounded (up) to $n+1$ if it is more than the geometric mean. In other words, the basis for rounding is the geometric mean $\sqrt{n(n+1)}$.

For instance, 2.47 is between 2 and 3; it is rounded (down) to 2 if you are practicing standard rounding. The rule is that we round "down" if the digit after the decimal point is 4 or less and round "up" if it is 5 or more. However, note that the geometric mean of 2 and 3 is $\sqrt{2 \times 3} = \sqrt{6} \approx 2.44949$. Since 2.47 is greater than this number, 2.45 is rounded (up) to 3 based on the geometric mean.

The geometric mean of n and $n+1$ approaches the arithmetic mean of n and $n+1$ as n gets bigger. The difference between these two rounding methods is rather minor, but this does make a difference when it comes to apportionment.[172]

Huntington-Hill Method

The Huntington-Hill Method is almost exactly like Webster's method, except "rounding based on the geometric mean" is used instead of standard rounding to calculate the rounded quotas.

Example 8

Use the Huntington-Hill Method to apportion the 41 lawmakers in the 3-state country used in a few examples above.

[Solution]

Again, as in earlier examples, we start with an initial standard divisor of 21,900.82927. For convenience, we add a new column for the geometric mean for each minimum quota. For instance, the geometric mean of 7 and 8 is $\sqrt{7 \times 8} \approx 7.48331$.

State	Population	Quota	Min. Quota	Geom. Mean	Rounded. Q
Aurora	162,310	7.4111	7	7.48331	7
Bay	538,479	24.5872	24	24.49490	25
Castle	197,145	9.0017	9	9.48683	9
Total	**897,934**				**41**

This gives the required total, so we are done.

[172] The geometric mean of 1 and 2 is $\sqrt{2} \approx 1.414$ …, which helped Montana to gain a congressional seat (from 1 to 2) following the 2020 census. The "bar" needed for rounding *up* is lower for smaller states.

Example 9

Apply the Huntington-Hill Method in the DMV example to apportion the 75 DMV employees.

[Solution]

Again, we start with an initial standard divisor of 14,034.22667:

Office	Population	Quota	Min. Quota	Geom. Mean	Rounded Q.
Amber	49,875	3.5538	3	3.46410	4
Birchwood	166,158	11.8395	11	11.48913	12
Chester	82,888	5.9061	5	5.47723	6
Dogwood	626,667	44.6528	44	44.49719	45
Elm	126,979	9.0478	9	9.48683	9
Total	**1,052,567**				**76**

This is too many, so we need to increase the divisor. Let us try 14,100:

Office	Population	Quota	Min. Quota	Geom. Mean	Rounded Q.
Amber	49,875	3.53723	3	3.46410	4
Birchwood	166,158	11.7843	11	11.48913	12
Chester	82,888	5.8786	5	5.47723	6
Dogwood	626,667	44.4445	44	44.49719	44
Elm	126,979	9.0056	9	9.48683	9
Total	**1,052,567**				**75**

This works, so we are done.

Example 10

Consider a small country with 5 states, two of which are much larger than the others, as shown below. We need to allocate 70 representatives. Apportion them using both Webster's method and the Huntington-Hill Method.

State	Population
A	300,500
B	200,000
C	50,000
D	38,000
E	21,500
Total	**610,000**

[Solution]

The total population is 610,000. Dividing this by the 70 representatives gives the divisor: $d = 610{,}000 \div 70 \approx 8714.286$. Dividing each state's population by the divisor gives the quotas

State	Population	Quota
A	300,500	34.48361
B	200,000	22.95082
C	50,000	5.737705
D	38,000	4.360656
E	21,500	2.467213

Webster's Method

Using Webster's method, we round each quota to the nearest whole number (using standard rounding).

State	Population	Quota	Rounded Q.
A	300,500	34.48361	34
B	200,000	22.95082	23
C	50,000	5.737705	6
D	38,000	4.360656	4
E	21,500	2.467213	2

Adding these up, they only total 69 representatives, so we adjust the divisor down. Adjusting the divisor down to 8700 gives an updated allocation totaling 70 representatives:

State	Population	Quota	Rounded Q.
A	300,500	34.54023	**35**
B	200,000	22.98851	23
C	50,000	5.74126	6
D	38,000	4.367816	4
E	21,500	2.471264	2

Huntington-Hill Method

Using the Huntington-Hill Method, we round down to find the minimum quota and then calculate the geometric mean based on each minimum quota. If the quota is less than the geometric mean, we round down; if the quota is more than the geometric mean, we round up.

State	Population	Quota	Min. Quota	Geometric Mean	Rounded Q.
A	300,500	34.48361	34	34.49638	34
B	200,000	22.95082	22	22.49444	**23**
C	50,000	5.737705	5	5.477226	**6**
D	38,000	4.360656	4	4.472136	4
E	21,500	2.467213	2	2.44949	**3**

These allocations add up to 70, so we are done.

Notice that this allocation is different from what was produced by Webster's method. In this case, state E got the extra seat instead of state A.

In this last example, note that state E, with a quota of 2.467213, would have been rounded *down* to 2 under standard rounding (used in Webster's method), but this number was bigger than the geometric mean of 2.44949 (which is significantly less than 2.5), so the quota was rounded up. This is what made the difference. Hence, *the use of geometric means actually favors smaller states*, which was exactly what Huntington and Hill were trying to achieve. There are a few other apportionment methods known, but we will not mention any more in this book.

Exercises 8.1

1. A college offers tutoring in Math, English, Chemistry, and Biology. The number of students enrolled in each subject is listed below. If the college can only afford to hire 15 tutors, determine how many tutors should be assigned to each subject, using (a) Hamilton's, (b) Jefferson's, (c) Webster's, and (d) the Huntington-Hill methods.

   Math: 330 English: 265 Chemistry: 130 Biology: 70

2. Re-apportion the previous problem assuming that the college can hire 20 tutors.

3. The number of salespeople assigned to work during a shift is apportioned based on the average number of customers during that shift. Apportion 20 salespeople given the information below, using (a) Hamilton's, (b) Jefferson's, (c) Webster's, and (d) the Huntington-Hill methods.

Shift	Average Number of Customers
Morning	95
Midday	305
Afternoon	435
Evening	515

4. Re-apportion the previous problem assuming that the store has 25 salespeople.

5. Three people invest in a treasure dive, each investing the amount listed below. The dive results in 36 gold coins. Apportion those coins to the investors, using (a) Hamilton's, (b) Jefferson's, (c) Webster's, and (d) the Huntington-Hill methods.

   Alice: $7,600 Ben: $5,900 Carlos: $1,400

6. Re-apportion the previous problem assuming that 37 gold coins are recovered.

7. A small country consists of five states, whose populations are listed below. If the legislature has 119 seats, apportion the seats, using (a) Hamilton's, (b) Jefferson's, (c) Webster's, and (d) the Huntington-Hill methods.

   A: 810,000 B: 473,000 C: 292,000 D: 594,000 E: 211,000

8. A small county consists of six districts, whose populations are listed below. If the legislature has 30 seats, apportion the seats, using (a) Hamilton's, (b) Jefferson's, (c) Webster's, and (d) the Huntington-Hill methods.

A: 3,411 B: 2,421 C: 11,586 D: 4,494 E: 3,126 F: 4,962

9. A small city consists of three regions, whose populations are listed below.

A: 6,000 B: 6,000 C: 2,000

(a) If the legislature has 10 seats, use Hamilton's method to apportion the seats.
(b) If the legislature grows to 11 seats, can Hamilton's method be used to apportion the seats? Why or why not?

10. A state with five counties has 50 seats in their legislature. Using Hamilton's method, apportion the seats based on the 2000 Census, then again using the 2010 Census. Which apportionment paradox does this illustrate?

County	2000 Population	2010 Population
Jefferson	60,000	60,000
Clay	31,200	31,200
Madison	69,200	72,400
Jackson	81,600	81,600
Franklin	118,000	118,400

11. A school district has two high schools: Lowell, serving 1715 students, and Fairview, serving 7364. The district could only afford to hire 13 guidance counselors.
(a) Determine how many counselors should be assigned to each school using Hamilton's method.
(b) The following year, the district expands to include a third school, serving 2989 students. Based on the divisor from above, how many additional counselors should be hired for the new school?
(c) After hiring that many new counselors, the district recalculates the re-apportion using Hamilton's method. Determine the outcome.
(d) Does this situation illustrate any apportionment issues?

12. Adams' method is similar to Jefferson's method, but it rounds quotas *up* rather than *down*. This means we usually need a modified divisor that is smaller than the standard divisor. Re-work Exercise 7 using Adams' method.

8.2 Gerrymandering

In the current government system of the United States, there are 50 states, and each state is divided up into counties. There are 100 seats in the Senate, two from each state, and the 435 seats in the House of Representatives are apportioned according to the Huntington-Hill Method we covered in the previous section. Some states have only one representative while a large state can have over 50. So how do these large states apportion their representatives? By county so the more populous counties will have more representation than the less populous ones? That may make sense, but states are in reality divided up into **congressional districts**; the idea is to reduce the representation gap, i.e., we want to minimize, as much as possible, the gap in the number of constituents each representative is to represent. It would not be fair or ideal, for instance, if a representative in one county represents 60,000 while another representative in another county has 120,000 constituents.

The apportionment process, then, involves the **drawing of the legislative districts**, with the goal of having each district include approximately the same number of constituents. Because of this goal, a geographically small city may have several representatives while a large rural region may be represented by one legislator.

When populations change, it becomes necessary to re-draw these districts which the legislators represent. The process of re-districting is typically done by the legislature itself, so not surprisingly it is common to see a politically manipulative practice called **gerrymandering**.

Gerrymandering

Gerrymandering is an intentional act whereby legislative districts are drawn based on the political affiliation of the constituents to the advantage of those drawing the boundary, usually a political party.

Re-drawing of congressional districts is truly a major and controversial task. In April, 2021, when it was discovered that the state of California would, for the first time ever, lose a seat in Congress, a former politician was quoted saying, "It's like going into a war where you don't know where the battlefield is, who your opponent is or when the battle is going to begin."[173]

Gerrymandering is actually quite old. The term came from the fact that, in 1812, Elbridge **Gerry** the Governor of Massachusetts re-drew the congressional districts around Boston in such an unusual (creative?) way that one district looked like a **salamander**. His goal was to gain advantage for his party, the Democratic-Republicans. (He went on to become the Vice

[173] Mason, M., & Mehta, S. (2021, April). California to lose a congressional seat, according to new census data. *Los Angeles Times.* https://www.latimes.com/politics/story/2021-04-26/census-data-redistricting-delay

President under James Madison.) This led to the coining of a new word, based on "Gerry" and "salamander."[174]

Example 1

Consider three districts, simplified to the three boxes below. On the left there is a college area that typically votes Democratic. On the right is a rural area that typically votes Republican. The rest of the people are more evenly split. The middle district has been voting 50% Democratic and 50% Republican.

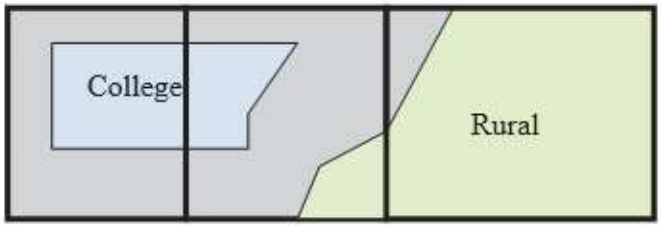


Figure 8.2[175]

As part of a re-districting, a Democratic-led committee could re-draw the boundaries so that the middle district includes less of the typically Republican voters, thereby making it more likely that their party will win in that district while increasing the Republican majority in the third district (which is already Republican).

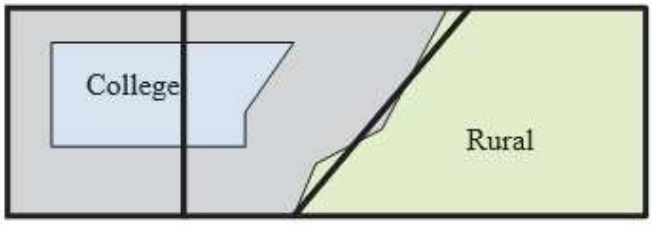


Figure 8.3[176]

■

Gerrymandering works by wasting votes. It puts more votes of winners into the district they will win anyway so the losers there could win in another district.

Suppose we have Group A and Group B both trying to win in a district. Group A has 40 votes while Group B has 50. In another district Group A knows Group B will win for sure. Group A changes the voting district so that 11 of the 50 votes are moved to that other district. Now Group A will win 40 votes to 39.

---

[174] Today, we pronounce this verb as "**gerrymander**" (with a soft "g," more like "jerrymander"), but since the name of this politician was actually pronounced with a hard "g" (as in "Gherry"), some people still maintain that our current, more common, pronunciation is incorrect.

[175] Apportionment by David Lippman is licensed under CC BY-SA 3.0 US.

[176] Apportionment by David Lippman is licensed under CC BY-SA 3.0 US.

Math in the Real World – *Using Big Data to Gerrymander*

Click How Technology and Data-Driven Gerrymandering Helps Politicians Win Elections - The Atlantic - The Atlantic to watch a short video that explains how big tech companies are making it easier to gerrymander voting districts. What does the use of data technology in this way mean for democracy?

Example 2

The map below shows the 38th congressional district in California in 2004. This district was created through a bi-partisan committee of incumbent legislators. This gerrymandering led to districts that are not competitive; the prevailing party almost always wins with a large margin. Note the very unusual shape resulting from gerrymandering. (The current district maps are quite different.)

Figure 8.4 [177]

Generally speaking, each congressional district must be geographically connected. In other words, one cannot have two separate and disconnected geographical "bubbles" and call them one district (unless they are islands or something). But this led to even more creativity and manipulation.

[177] Image from National Atlas of the United States is in the Public Domain

The map below shows the 4th congressional district in Illinois in 2004. This district was drawn to contain the two predominantly Hispanic areas of Chicago: the largely Puerto Rican area to the north and the southern Mexican area! They were only connected in this districting by a piece of the highway (part of I-294) to the west.

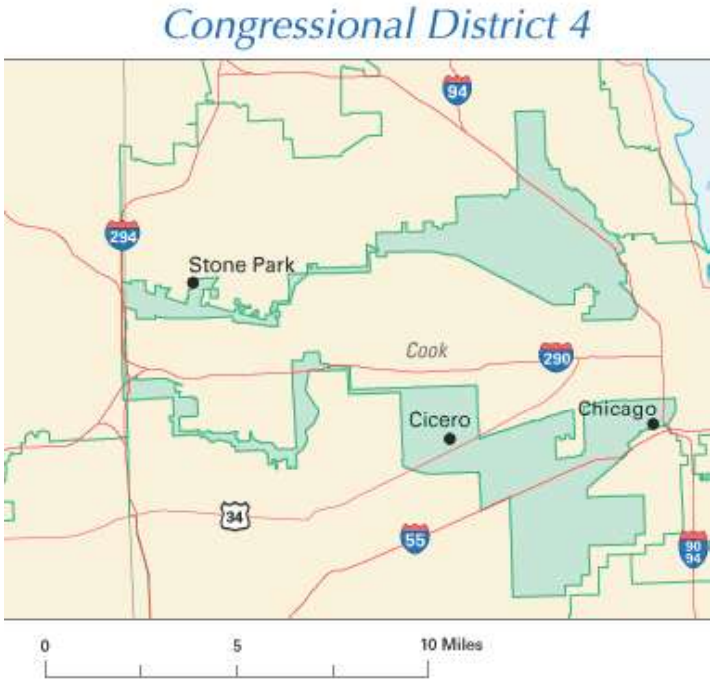


Figure 8.5 [178]

We will give one specific example with the calculations shown. Note that addition is the only operation involved in these questions, so the mathematics is really not all that difficult. One just has to pay close attention to which numbers to be added together. The challenge (if you are the one gerrymandering or trying to figure out what other gerrymandering people may do) is to come up with several plans to group (and re-group) various regions together to create legislative districts and compare the resulting numbers to see which party would win each.

Example 3

The map below shows a town with four regions that need to be grouped together to form two legislative districts. The total number of registered voters shown is 3560. To guarantee approximately equal representation, each district must have at least 1,700 voters.

[178] Image from National Atlas of the United States is in the Public Domain

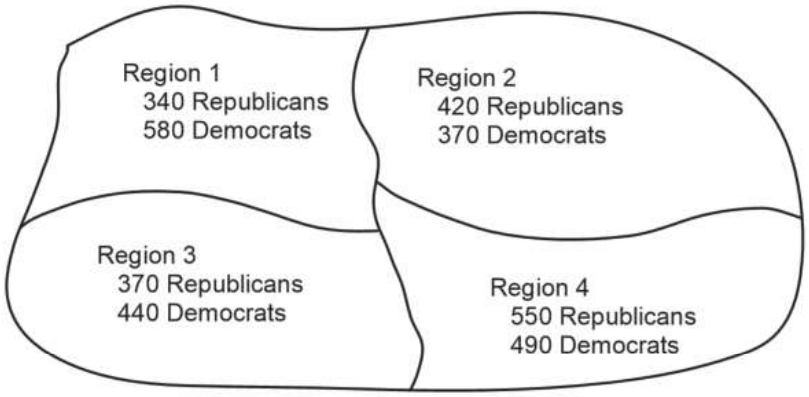


Figure 8.6

Currently, the First District consists of Regions 1 and 2 and the Second District consists of Regions 3 and 4. However, a newly elected assembly has come up with a plan to make the First District consisting of Regions 1 and 3 while the Second District consists of Regions 2 and 4. Verify that the current districts and the new plan both satisfy the minimum-number requirement and determine if the new plan would change the party representation of the town.

[Solution]

First, let us check that each district meets the minimum-number requirement (each district must have 1700 or more voters). Note that Region 1 has $340 + 580 = 920$ registered voters and Region 2 has $420 + 370 = 790$. Similarly, Region 3 has 810, and Region 4 has 1040. Therefore, currently the First District (Regions 1 and 2) has $920 + 790 = 1710$ while the Second District (Regions 3 and 4) has $810 + 1040 = 1850$. Each district has at least 1700, satisfying the requirement.

Under the proposed plan, the First District (Regions 1 and 3) would have $920 + 810 = 1730$ while the Second District (Regions 2 and 4) would have $790 + 1040 = 1830$, again both satisfying the requirement.

Now, let us see who wins each district. Here are the current numbers:

First District (Regions 1 and 2):
- $340 + 420 = 760$ Republicans
- $580 + 370 = 950$ Democrats

Second District (Regions 3 and 4):
- $370 + 550 = 920$ Republicans
- $440 + 490 = 930$ Democrats

Hence, the Democrats control both districts. Under the proposed plan,

First District (Regions 1 and 3):

- $340 + 370 = 710$ Republicans
- $580 + 440 = 1020$ Democrats

Second District (Regions 2 and 4):

- $420 + 550 = 970$ Republicans
- $370 + 490 = 860$ Democrats

So, under the proposed plan, the Republicans will gain the Second District. The new plan must have been devised by the Republicans.

This example, and gerrymandering in general, illustrates the kind of "numbers games" played in politics to gain advantage. Apportionment in Section 8.1 and voting methods in Section 8.3 also provide examples. The fact is that these numbers games play a huge role in today's political strategies. It is clear that we must all think, analyze, and vote with a mathematical mindset.

Exercises 8.2

1. A county with the following political map is to be divided up into two legislative districts. The total number of registered voters shown is 21,500. To guarantee approximately equal representation, each district must have at least 10,000 voters. Answer each of the following questions.

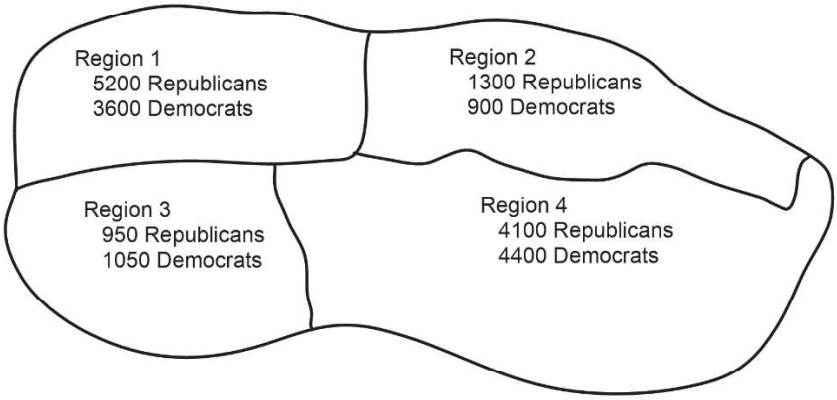


Figure 8.7

(a) Suppose the First District consists of Regions 1 and 2 and the Second District consists of Regions 3 and 4. Does each district have the minimum number of voters? Which party wins each district?

(b) Suppose the First District consists of Regions 1 and 3 and the Second District consists of Regions 2 and 4. Does each district have the minimum number of voters? Which party wins each district? Did the results change from (a) above?

2. A county with the following political map is to be divided up into two legislative districts. The total number of registered voters shown is 20,500. To guarantee approximately equal representation, each district must have at least 10,000 voters. Answer each of the following questions.

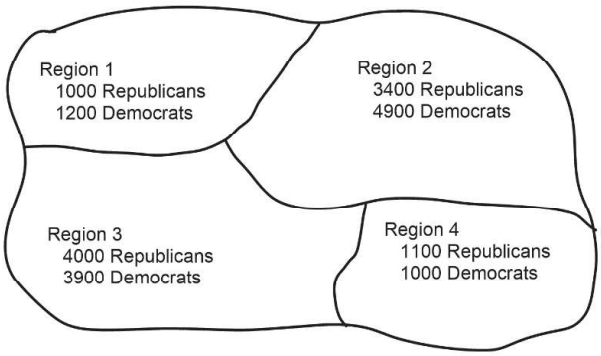


Figure 8.8

(a) Suppose the First District consists of Regions 1 and 2 and the Second District consists of Regions 3 and 4. Does each district have the minimum number of voters? Which party wins each district?
(b) Suppose the First District consists of Regions 1 and 3 and the Second District consists of Regions 2 and 4. Does each district have the minimum number of voters? Which party wins each district? Did the results change from (a) above?

3. A county with the following political map is to be divided up into two legislative districts. The total number of registered voters shown is 10,950. To guarantee approximately equal representation, each district must have at least 5,200 voters. Answer each of the following questions.

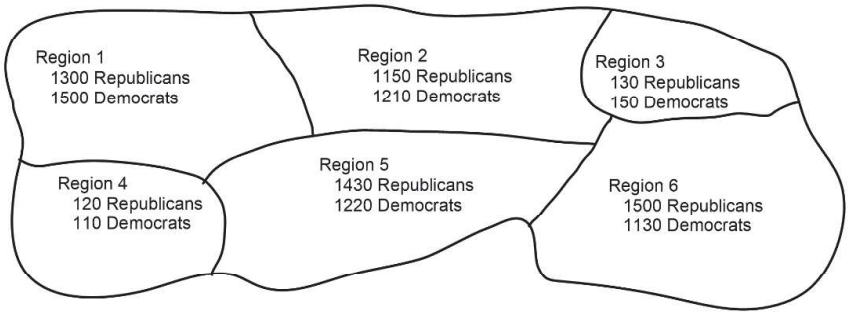


Figure 8.9

(a) Suppose the First District consists of Regions 1, 2, and 4 while the Second District consists of Regions 3, 5, and 6. Does each district have the minimum number of voters? Which party wins each district?
(b) Someone suggested we create one district consisting of only Regions 2 and 3 (which would be almost guaranteed to vote Democratic). Why is this idea not allowed?
(c) Suppose the First District consists of Regions 1, 4, and 5 while the Second District consists of Regions 2, 3, and 6. Does each district have the minimum number of voters? Which party wins each district? Did the results change from (a) above?

4. A county with the following political map is to be divided up into two legislative districts. The total number of registered voters shown is 9400. To guarantee approximately equal representation, each district must have at least 4,500 voters. Answer each of the following questions.

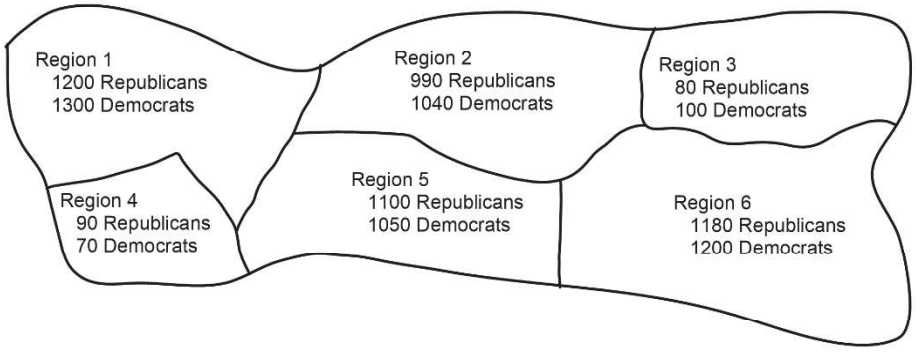


Figure 8.10

(a) Suppose the First District consists of Regions 1 and 2 while the Second District consists of Regions 3, 4, 5, and 6. Does each district have the minimum number of voters? Which party wins each district?
(b) Someone suggested we create one district consisting of only Regions 4 and 5 (which would be almost guaranteed to vote Republican). Why is this idea not allowed?
(c) Suppose the First District consists of Regions 1, 4, and 5 while the Second District consists of Regions 2, 3, and 6. Does each district have the minimum number of voters? Which party wins each district? Did the results change from (a) above?

8.3 Theory of Voting[xx]

In many decision-making situations, it is necessary to gather the group consensus. This happens when a group of friends decides which movie to watch, when a company decides which product design to manufacture, and when a democratic country elects its leaders.

While the basic idea of voting is fairly universal, the method by which those votes are used to determine a winner can vary. Among a group of friends, you may decide upon a movie by voting for all the movies you are willing to watch, with the winner being the one with the greatest approval. A company might eliminate unpopular designs first and then vote again on the remaining. A country might look for the candidate with the most (first-place) votes.

In deciding upon a winner, there is always one main goal: to reflect the preferences of the people in the fairest way possible. So which of the methods contained in this section is best from the fairness point of view? How can we guarantee fairness? What is fairness anyway?

Here is one way to define "fairness" in an election. In order to define what is fair and what is not, we must first agree that any *fair* election must pass/guarantee certain standards. These standards are called **fairness criteria** ("criteria" is the plural form of "criterion"). In other words, a **fairness criterion** is a statement that should be guaranteed in any fair election. We will introduce a few of these. This set of fairness criteria would then give us a working definition of "fairness" in the theory of voting. In other words, we decide an election method is "fair" if and only if it satisfies all of these criteria all the time.

We now look at various methods of election and ponder these questions.

Preference Schedules

We start with more information than what a traditional ballot normally provides. A traditional ballot usually asks you to pick **one** favorite candidate from a list of choices. This ballot fails to provide any information on how a voter would rank the alternatives if their first choice was unsuccessful.

Preference Ballot

A preference ballot is a ballot in which the voter ranks all the candidates in order of preference. A result can be tabulated in what is known as a **preference schedule**.

Example 1

Suppose ten members of an organization cast ballots on where they want to go for a conference: Anaheim (A), Hawaii (H), and Orlando (O). Each person orders these three destinations from 1 to 3 based on their preference. Below is the summary, shown in a tabular form. There are a total of $3! = 6$ ways to order these three destinations, but only four permutations appear here. Construct a **preference schedule** of the result.

	Bob	Ann	Marv	Alice	Eve	Omar	Lupe	Dave	Tish	Jim
1st	A	A	O	H	A	O	H	O	H	A
2nd	O	H	H	A	H	H	A	H	A	H
3rd	H	O	A	O	O	A	O	A	O	O

[Solution]

These individual ballots are typically combined into one **preference schedule**, which shows the number of voters in the top row that voted for each option (each permutation):

	1	3	3	3
1st	A	A	O	H
2nd	O	H	H	A
3rd	H	O	A	O

Notice that, by totaling the vote counts across the top of the preference schedule, we can recover the total number of votes cast: $1 + 3 + 3 + 3 = 10$ (votes).

It is important to know that a preference schedule itself does not determine the winner. Rather, the winner is chosen based on one of the many pre-determined methods, which we will now introduce.

Plurality

The voting method we are most familiar with in the United States is the **plurality method**.

Plurality Method

In this method, the candidate with the most first-preference votes is declared the winner. Ties are possible and would have to be settled through some sort of run-off vote. Second and lower preferences are not considered at all in this method.

This method is sometimes mistakenly called the majority method, or the "majority rule," but it is not necessary for a candidate to have gained a majority of votes to win. A **majority** is anything over 50% (not exactly 50%) of the votes; it is possible for a winner to have a **plurality** without having a majority.

Example 2

In our election from above, we had the preference table:

	1	**3**	**3**	**3**
1st	A	A	O	H
2nd	O	H	H	A
3rd	H	O	A	O

For the plurality method, we only care about the first-choice options. Totaling them up:

- Anaheim: $1 + 3 = 4$ votes
- Orlando: 3 first-choice votes
- Hawaii: 3 first-choice votes

Therefore, Anaheim is the winner using the plurality method.
Here, since 10 ballots were cast, 6 would make a majority (5 out of 10 is exactly 50%, and you need *more than* 50% for a majority). Notice that Anaheim won with 4 out of 10 votes, 40% of the votes, which is a plurality of the votes, but not a majority.

Try It Now 1

Three candidates are running in an election for County Executive: Gutierrez (G), McCarthy (M), and Baillieu (B). The preference schedule is shown below. Which candidate wins under the plurality method?

	44	14	70	22	80
1st	G	G	M	M	B
2nd	M	B	G	B	M
3rd	B	M	B	G	G

What Could Go Wrong with Plurality?

The election from Example 2 may seem totally clean, but there is a problem lurking that arises whenever there are three or more choices. Looking back at our preference table, how would our members vote if they only had two choices?

Anaheim vs. Orlando: 7 out of the 10 would prefer Anaheim over Orlando (look at the three columns where "A" appears above "O").

	1	*3*	3	*3*
1st	*A*	*A*	O	*H*
2nd	*O*	*H*	H	*A*
3rd	*H*	*O*	A	*O*

Anaheim vs. Hawaii: 6 out of 10 would prefer Hawaii over Anaheim (look at the two columns where "H" appears above "A").

	1	3	*3*	*3*
1st	A	A	*O*	*H*
2nd	O	H	*H*	*A*
3rd	H	O	*A*	*O*

This does not seem right, does it? Anaheim just won the election, yet 6 out of 10 voters, a majority, would have preferred Hawaii over Anaheim. That hardly seems fair. In 1785, right before the French Revolution, Marquis de Condorcet, a French philosopher, mathematician, and political scientist, wrote about how this could happen, and for him we name our first of several **fairness criteria**.

Condorcet Criterion

If there is a candidate that is preferred in **every one-on-one (pairwise) comparison**, that candidate should be the winner. We call this winner the **Condorcet winner**, or Condorcet candidate.

This naturally leads to our second method of election.

Condorcet Method

The key idea of Condorcet was one-on-one comparison, in which every pair of candidates are compared for preference in pairwise analysis. The method of election which declares the Condorcet candidate the winner of the election is (naturally) called the Condorcet method.

Condorcet Method

In the Condorcet method of voting, the Condorcet candidate (the candidate preferred over everyone else via one-on-one comparison) is declared the winner (if such a candidate exists).

The obvious problem with the Condorcet method is that sometimes there is no Condorcet candidate. For instance, if three candidates are running, it is possible for A to be preferred over B, B over C, and C over A (as we will see later).

Example 3

In the election from Example 2, is there a Condorcet winner?

[Solution]

We saw above that Anaheim was preferred over Orland, and Hawaii was preferred over Anaheim. Comparing Hawaii to Orlando, we can see 6 out of 10 would prefer Hawaii to Orlando (the two columns where “H” appears above “O”).

	1	*3*	3	*3*
1st	A	*A*	O	*H*
2nd	O	*H*	H	*A*
3rd	H	*O*	A	*O*

Since Hawaii is preferred in a one-on-one comparison over Anaheim *and* over Orlando, Hawaii is the Condorcet candidate and, therefore, the winner of this election.

Warning: we are *not simply comparing the number of first-place votes* in these one-on-one comparisons. To do one-on-one (pairwise) comparisons, you must check each column carefully. To compare candidate X with candidate Y, you must find all columns where "X" appears above "Y" and add the numbers of ballots on those columns.

Also, we have now seen that the plurality winner was Anaheim even though there *was* a Condorcet candidate, namely Hawaii. Therefore, the plurality method does not satisfy the Condorcet Criterion.

Example 4

Consider a city council election in a district that is historically 60% Democratic voters and 40% Republican voters. Even though city council is technically a nonpartisan office, people generally know the affiliations of the candidates. In this election there are three candidates: Don and Key, both Democrats, and Elle, a Republican. A preference schedule for the votes looks as follows.

	342	214	298
1st	Elle	Don	Key
2nd	Don	Key	Don
3rd	Key	Elle	Elle

We can see that a total of $342 + 214 + 298 = 854$ voters participated in this election. Who was the winner by plurality? Was there a Condorcet candidate?

[Solution]

We can clearly see that Elle, the Republican, won the election by plurality as she had 342 first-place votes. In this election, the Democratic voters split their vote over the two Democratic candidates, allowing the Republican candidate Elle to win under the plurality method with 40% of the vote.

Let us do the pairwise comparisons:

- Elle vs. Don: 342 prefer Elle; 512 prefer Don $(214 + 298)$. Don is preferred over Elle.
- Elle vs. Key: 342 prefer Elle; 512 prefer Key $(214 + 298)$. Key is preferred over Elle.
- Don vs. Key: 556 prefer Don $(342 + 214)$; 298 prefer Key. Don is preferred over Key.

So even though Don had the smallest number of first-place votes in the election, he is the Condorcet winner, being preferred in every one-to-one comparison.

Insincere Voting

Situations like the one in Example 4 above, when there are two or more candidates who share somewhat similar points of view, can lead to **insincere voting**. Insincere voting refers to a person casting a ballot counter to their actual preference for strategic purposes. In the case above, the Democratic leadership might realize that Don and Key will split the vote, so they may encourage voters to vote for Key by officially endorsing him. Then, not wanting to see their party lose the election, as happened in the scenario above, Don's supporters might insincerely vote for Key, effectively voting against Elle.

Try It Now 2

Three candidates are running in an election for County Executive: Gutierrez (G), McCarthy (M), and Baillieu (B). The preference schedule is shown below. Is there a Condorcet winner in this election?

	44	14	70	22	80
1st	G	G	M	M	B
2nd	M	B	G	B	M
3rd	B	M	B	G	G

Example 5

Consider an "election" at a summer camp for an afternoon activity. Was there a Condorcet winner?

	7	9	10
1st	Soccer	Baseball	Football
2nd	Baseball	Football	Soccer
3rd	Football	Soccer	Baseball

[Solution]

On one-on-one comparisons, soccer beats baseball 17 to 9. Baseball beats football 16 to 10. But football beats soccer 19 to 7. (This is a case where A is over B, B is over C, and C is over A, completing a "circle.") Hence, there is no Condorcet winner. If these campers had agreed on the Condorcet method of election, they would be disappointed; they would need to sit down and carefully choose another method of election. (Of course, *that* would be a great afternoon activity.)

Just as you would need some tie-breaking system in plurality (where two or more candidates have the highest number of first-place votes), you would need a way to resolve a "tie" like this (the result of Example 5) in the Condorcet method.

Instant-Runoff Voting

Instant Runoff Voting (IRV)[179] is a modification of the plurality method that attempts to address the issue of insincere voting.

Instant-Runoff Voting (IRV)

In IRV, voting is done with preference ballots, and a preference schedule is generated. The candidate with the *least* first-place votes is then **eliminated** from the election, and any votes for that candidate are re-distributed to the voters' next candidate. This continues until a candidate has a majority (over 50%).

In other words, the candidate with the fewest first-place votes are **removed** from the schedule, which then gets adjusted, and the process is repeated. Note that there is still *only one election* here even though the term "runoff" could often mean that another real election is held with some candidates removed—typically all but the top two candidates. Real runoff elections are costly and time-consuming; *instant* runoffs are based on the preference schedule

[179] IRV (instant-runoff voting) is sometimes called "RCV" (**ranked-choice voting**) although, to be precise, there are several versions of RCV.

and is much more efficient. (You may recall the not-so-instant runoff election for the two Georgia senators following the general election in November, 2020.)

This voting method is used in many political elections around the world, including San Francisco, the state of Maine, and in the election for the Australian House of Representatives. A version of IRV is used by the International Olympic Committee (to select host nations) as well as for the Academy Awards. In fact, starting 2021, New York City is using IRV (though they call it "RCV" for "**ranked-choice voting**" on the city information websites) in their primary and special elections for local offices.

Example 6

Consider the preference schedule below, in which a company's advertising team is voting on five different advertising slogans, called A, B, C, D, and E here for simplicity. Find the winner by IRV.

Initial votes

	3	4	4	6	2	1
1st	B	C	B	D	B	E
2nd	C	A	D	C	E	A
3rd	A	D	C	A	A	D
4th	D	B	A	E	C	B
5th	E	E	E	B	D	C

[Solution]

IRV often takes several steps as one candidate gets eliminated in each step. If this were a plurality election, note that B would be the winner with 9 first-choice votes, compared to 6 for D, 4 for C, 1 for E, and 0 (none) for A. Also, there are total of $3 + 4 + 4 + 6 + 2 + 1 = 20$ votes. A majority would be 11 votes. No one yet has a majority, so we proceed to elimination rounds.

Round 1: We make our first elimination. Choice A has the fewest first-place votes (none), so we remove (eliminate) that choice as shown below. This is the first modification we make to the preference schedule.

	3	4	4	6	2	1
1st	B	C	B	D	B	E
2nd	C		D	C	E	
3rd		D	C			D
4th	D	B		E	C	B
5th	E	E	E	B	D	C

We then shift everyone's choices up to fill the gaps. (In some cases, two or more columns may look identical at this point. If this happens, you should combine the columns.)

	3	4	4	6	2	1
1st	B	C	B	D	B	E
2nd	C	D	D	C	E	D
3rd	D	B	C	E	C	B
4th	E	E	E	B	D	C

Round 2: We make our second elimination. B has 9, C has 4, D has 6 first-place votes. Choice E has the fewest (with just one), so we remove E everywhere and then shift everyone's options to fill the gaps.

	3	4	4	6	2	1
1st	B	C	B	D	B	D
2nd	C	D	D	C	C	B
3rd	D	B	C	B	D	C

Notice that the first and fifth columns have the same preferences (B-C-D) now, so we condense those down to one column. Be sure to add the numbers in those columns: $3 + 2 = 5$.

	5	4	4	6	1
1st	B	C	B	D	D
2nd	C	D	D	C	B
3rd	D	B	C	B	C

Still, no one has a majority, so we eliminate again.

Round 3: We make our third elimination. Now B has 9 first-choice votes, C has 4, and D has 7. Now C has the fewest first-place votes, so C is removed, and the options are shifted up again.

	5	4	4	6	1
1st	B	D	B	D	D
2nd	D	B	D	B	B

Condensing this down we get this:

	9	11
1st	B	D
2nd	D	B

D has now gained a majority and is declared the winner under IRV.

Try It Now 3

Three candidates are running in an election for County Executive: Gutierrez (G), McCarthy (M), and Baillieu (B). The preference schedule is shown below. Who is the winner using IRV?

	44	14	70	22	80
1st	G	G	M	M	B
2nd	M	B	G	B	M
3rd	B	M	B	G	G

What Could Go Wrong with IRV?

We have already seen that the plurality method does not satisfy the Condorcet Fairness Criterion, and the Condorcet method does not always produce a winner. While plurality totally disregard voters' opinions except for their first-place votes, IRV seems to take into account all their choices along with their rankings. So is IRV a better method? Consider the following example.

Example 7

Let us return to our City Council Election. Who is the winner by IRV?

	342	214	298
1st	Elle	Don	Key
2nd	Don	Key	Don
3rd	Key	Elle	Elle

[Solution]

In this election, Don has the smallest number of first place votes, so Don is eliminated in the first round. The 214 people who voted for Don have their votes transferred to their second choice, Key. After removing Don and shifting up, we have this modified schedule:

	342	512
1st	Elle	Key
2nd	Key	Elle

So Key is the winner under the IRV method.

We can immediately notice that in this election, *IRV violates the Condorcet Criterion* since we determined earlier that Don was the Condorcet winner. On the other hand, the temptation has been removed for Don's supporters to vote for Key; they now know their vote will be transferred to Key, not simply discarded.

Example 8

Consider the voting system below. Who is the winner by IRV?

	37	22	12	29
1st	Adams	Brown	Brown	Carter
2nd	Brown	Carter	Adams	Adams
3rd	Carter	Adams	Carter	Brown

[Solution]

In this election, Adams has 37 first-place votes, Brown 34, and Carter 29. Carter is thus eliminated first, and Adams, after receiving the 29 from the last column, would be the winner with $37 + 29 = 66$ votes over Brown, who will still have 34.

Now, here is a different way to look at this example. Suppose that the results were announced, but election officials accidentally destroyed the ballots before they could be certified, and the votes had to be re-cast. Wanting to "jump on the bandwagon," 10 of the 12 voters who had originally voted in the order Brown, Adams, Carter change their vote to favor the presumed winner Adams, changing those votes to Adams, Brown, Carter. Then the schedule would look like this:

	47	22	**2**	29
1st	Adams	Brown	Brown	Carter
2nd	Brown	Carter	Adams	Adams
3rd	Carter	Adams	Carter	Brown

In this re-vote, Brown will be eliminated in the first round, having the fewest first-place votes (24). After transferring votes, we find that Carter will win this election with $22 + 29 = 51$ votes to Adams, who would get only $47 + 2 = 49$ votes!

So what happened? Before the re-casting, Carter was the first to be removed, giving Adams the victory. But when the ballots were re-cast, with *ten more people* favoring Adams instead of Brown, Adams *ended up losing the election to Carter.* In other words, the candidate (Carter) who was first to be removed ended up winning while the candidate (Adams) who should have won the re-cast election lost because *more people* preferred that candidate. Makes sense? No!

Even though the only vote changes made (by the 10 voters) *favored* Adams, the change ended up costing Adams the election. This does not seem fair. We now introduce our second fairness criterion:

Monotonicity Criterion

If voters change their votes to increase the preference for one candidate, that change should not harm that candidate's chances of winning.

This criterion is violated by this election method (IRV). It does not mean that IRV always violates the criterion; it just means that IRV has the potential to violate the criterion in certain elections and, therefore, cannot ensure that this criterion is satisfied all the time.

Borda Count

Borda Count is another voting method, named for Jean-Charles de Borda, who developed the system in 1770. It is also known as the "point system."

Borda Count

In this method, points are assigned to candidates based on their ranking; 1 point for last choice, 2 points for second-to-last choice, and so on. The point values for all ballots are totaled, and the candidate with the largest point total is the winner.

Example 9

A group of managers are getting together for a conference. The members are coming from four cities: Seattle, Tacoma, Puyallup, and Olympia. The votes for where to hold the conference were as follows. What location was the winner by Borda count?

	51	25	10	14
1st	Seattle	Tacoma	Puyallup	Olympia
2nd	Tacoma	Puyallup	Tacoma	Tacoma
3rd	Olympia	Olympia	Olympia	Puyallup
4th	Puyallup	Seattle	Seattle	Seattle

[Solution]

Note that there are $51 + 25 + 10 + 14 = 100$ people who voted here. In each of the 51 ballots ranking Seattle first, Puyallup will be given 1 point, Olympia 2 points, Tacoma 3 points, and Seattle 4 points. Multiplying the points per vote times the number of votes allows us to calculate points awarded:

	51	25	10	14
1st choice 4 points	Seattle 4·51 = 204	Tacoma 4·25 = 100	Puyallup 4·10 = 40	Olympia 4·14 = 56
2nd choice 3 points	Tacoma 3·51 = 153	Puyallup 3·25 = 75	Tacoma 3·10 = 30	Tacoma 3·14 = 42
3rd choice 2 points	Olympia 2·51 = 102	Olympia 2·25 = 50	Olympia 2·10 = 20	Puyallup 2·14 = 28
4th choice 1 point	Puyallup 1·51 = 51	Seattle 1·25 = 25	Seattle 1·10 = 10	Seattle 1·14 = 14

Adding up the points, we see the following:

- Seattle: 204 + 25 + 10 + 14 = 253 points
- Tacoma: 153 + 100 + 30 + 42 = 325 points
- Puyallup: 51 + 75 + 40 + 28 = 194 points
- Olympia: 102 + 50 + 20 + 56 = 228 points

Under the Borda Count method, Tacoma is the winner of this vote.

Try It Now 4

Three candidates are running in an election for County Executive: Gutierrez (G), McCarthy (M), and Baillieu (B). The preference schedule is shown below. Who is the winner using Borda Count?

	44	**14**	**70**	**22**	**80**
1st	**G**	**G**	**M**	**M**	**B**
2nd	**M**	**B**	**G**	**B**	**M**
3rd	**B**	**M**	**B**	**G**	**G**

What Could Go Wrong with Borda Count?

You might have already noticed one potential flaw of the Borda Count from the previous example. In that example, Seattle had a majority (51 out of 100) of first-choice votes, yet it lost the election! This seems odd and prompts our next fairness criterion:

Majority Criterion

If a candidate has a majority of first-place votes, that candidate should be the winner.

The election from Example 9 using the Borda Count violates the Majority Criterion. Notice also that this automatically means that the Condorcet Criterion will also be violated as Seattle would have been preferred by at least 51% of voters in any head-to-head comparison.

Borda count is sometimes described as a **consensus-based** voting system since it can sometimes choose a more broadly acceptable option (like Tacoma in Example 9) over the one with majority support. In the example above, Tacoma is probably the best compromise location. This is an approach different from plurality and instant-runoff voting, both of which focus on first-choice votes; Borda Count considers every voter's entire ranking to determine the outcome.

Because of this consensus behavior, Borda Count, or some variation of it, is commonly used in awarding sports awards. Variations are used to determine the Most Valuable Player in baseball, to rank teams in NCAA sports, and to award the Heisman trophy.

If you are a soccer fan, you may be familiar with the point system in the first round (Group Round) of the World Cup. Each group has four teams (countries), and every team plays every other team in the same group. A team is awarded 3 points for each win, 1 point for a tie, and 0 point for a loss. Then, the two teams with most points from each group advance to the next (elimination) round. This is a variation of Borda Count. Similar point systems are used in many tournaments, elections, and competitions.

Copeland's Method

So far none of our voting methods have satisfied the Condorcet Criterion, except, of course, the Condorcet method (but that method could fail when there is no winner). Copeland's method specifically attempts to satisfy the Condorcet Criterion by looking at pairwise (one-on-one) comparisons.

Copeland's Method

In this method, each pair of candidates is compared to determine which of the two is more preferred. The more preferred candidate is awarded 1 point. If there is a tie, each candidate is awarded ½ point. After all pairwise comparisons are made, the candidate with the most points, and hence the most pairwise wins, is declared the winner.

Variations of Copeland's method are used in many professional organizations, including election of the Board of Trustees for the Wikimedia Foundation, which runs Wikipedia.

Example 10

Consider our vacation group example from the beginning of the chapter (involving Anaheim, Orlando, and Hawaii). Determine the winner using Copeland's method.

	1	3	3	3
1st	A	A	O	H
2nd	O	H	H	A
3rd	H	O	A	O

[Solution]

We need to look at each pair of choices and see which choice would win in a one-on-one comparison. You may recall we did this earlier when determining the Condorcet winner. For example, comparing Hawaii vs. Orlando, we see that 6 voters, those shaded in the first table below, would prefer Hawaii to Orlando. Note that Hawaii does not have to be the voter's first choice – we are imagining that Anaheim was not an option. If it helps, you can imagine removing Anaheim, as in the second table below. (Hawaii wins over Orlando, 6 to 4.)

	1	*3*	3	*3*
1st	A	*A*	O	*H*
2nd	O	*H*	H	*A*
3rd	H	*O*	A	*O*

	1	3	3	3
1st			O	H
2nd	O	H	H	
3rd	H	O		O

Based on this, in the comparison of Hawaii vs. Orlando, Hawaii wins and receives 1 point.

Comparing Anaheim to Orlando, the 1 voter in the first column, 3 in the second, and 3 in the last all prefer Anaheim to Orlando; Anaheim wins over Orlando, 7 to 3. Anaheim gets 1 point.

Finally, we compare Anaheim and Hawaii. The 4 voters in the first two columns prefer Anaheim to Hawaii, and the 6 in the last two columns prefer Hawaii to Anaheim. So Hawaii wins over Anaheim, 6 to 4. Hawaii gets 1 point.

Hawaii thus ends with 2 points, Anaheim with 1, and Orlando with 0. Hawaii is therefore the winner under Copeland's method. Notice this process is consistent with our determination of a Condorcet winner.

Example 11

Consider the following preference schedule for the advertising slogan election, with five slogans labeled A through E. Who wins by Copeland's method?

	3	4	4	6	2	1
1st	B	C	B	D	B	E
2nd	C	A	D	C	E	A
3rd	A	D	C	A	A	D
4th	D	B	A	E	C	B
5th	E	E	E	B	D	C

[Solution]

With 5 candidates, there are $\binom{5}{2} = 10$ comparisons to make:

- A vs. B: 11 votes to 9 votes — A gets 1 point.
- A vs. C: 3 votes to 17 votes — C gets 1 point.
- A vs. D: 10 votes to 10 votes — A gets ½ point; D gets ½ point.
- A vs. E: 17 votes to 3 votes — A gets 1 point.
- B vs. C: 10 votes to 10 votes — B gets ½ point; C gets ½ point.
- B vs. D: 9 votes to 11 votes — D gets 1 point.
- B vs. E: 13 votes to 7 votes — B gets 1 point.
- C vs. D: 9 votes to 11 votes — D gets 1 point.
- C vs. E: 17 votes to 3 votes — C gets 1 point.
- D vs. E: 17 votes to 3 votes — D gets 1 point.

Totaling these up:

A gets 2½ points. B gets 1½ points. C gets 2½ points. D gets 3½ points. E gets 0 points. Using Copeland's method, we declare D as the winner.

Notice that in this case, D is not a Condorcet winner (since A and D tie). While Copeland's method will also select a Condorcet candidate as the winner when there is one, the method still works in cases where there is no Condorcet winner.

Try It Now 5

Three candidates are running in an election for County Executive: Gutierrez (G), McCarthy (M), and Baillieu (B). The preference schedule is shown below. Who is the winner using Copeland's method?

	44	14	70	22	80
1st	G	G	M	M	B
2nd	M	B	G	B	M
3rd	B	M	B	G	G

What Could Go Wrong with Copeland's Method?

As already noted, Copeland's method does satisfy the Condorcet Criterion. It also satisfies the Majority Criterion and the Monotonicity Criterion. So is this the perfect method? Well, in a word, no.

Example 12

A committee is trying to award a scholarship to one of four students, Anna (A), Brian (B), Carlos (C), and Dimitry (D). The votes are shown below. Find the winner by Copeland's method.

	5	5	6	4
1st	D	A	C	B
2nd	A	C	B	D
3rd	C	B	D	A
4th	B	D	A	C

[Solution]
Making the comparisons:

- A vs. B: 10 votes to 10 votes A gets ½ point; B gets ½ point.

- A vs. C: 14 votes to 6 votes: A gets 1 point.
- A vs. D: 5 votes to 15 votes: D gets 1 point.
- B vs. C: 4 votes to 16 votes: C gets 1 point.
- B vs. D: 15 votes to 5 votes: B gets 1 point.
- C vs. D: 11 votes to 9 votes: C gets 1 point.

Totaling:

A has 1 ½ points. B has 1 ½ points. C has 2 points. D has 1 point. So Carlos is awarded the scholarship.

However, suppose that the committee then discovers that Dimitry was ineligible for the scholarship. Now, Dimitry did not win, so this seems irrelevant to the outcome; however, the committee decides to re-count the vote, having removed Dimitry from consideration. This reduces the preference schedule to the one below, and now there are three comparisons to be made:

	5	5	6	4
1st	A	A	C	B
2nd	C	C	B	A
3rd	B	B	A	C

Comparisons:

- A vs. B: 10 votes to 10 votes A gets ½ point; B gets ½ point.
- A vs. C: 14 votes to 6 votes A gets 1 point.
- B vs. C: 4 votes to 16 votes C gets 1 point.

Totaling:

A has 1 ½ points. B has ½ point. C has 1 point. So suddenly Anna is the winner (just because Dimitry was removed). This leads us to another fairness criterion.

The Independence of Irrelevant Alternatives (IIA) Criterion
If a non-winning candidate (an **irrelevant alternative**) is removed from the ballot or added to the ballot, that should not change the winner of the election.

Equivalently, if choice A is preferred over choice B, introducing or removing a choice C should not cause B to be preferred over A.

In the election from Example 12, the IIA Criterion was violated.

Another disadvantage of Copeland's method is that it is easy for the election to end in a tie. For this reason, Copeland's method is usually the first part of an advanced method that is more sophisticated for breaking ties and determining the winner when there is not a Condorcet winner.

So What Is the Fair Method?

At this point, you are probably asking why we keep looking at method after method just to point out that they are not completely fair. We must be holding out on the perfect method, right?

Unfortunately, no. A mathematical economist, Kenneth Arrow, was able to prove a devastating proposition in 1949. It turns out that there is *no* voting method that will satisfy all the fairness criteria we have discussed.

Arrow's Impossibility Theorem

It is not possible for any voting method to satisfy all the fairness criteria listed.

To see a very simple example of how difficult voting can be, consider the election below:

	5	5	5
1st	A	C	B
2nd	B	A	C
3rd	C	B	A

Notice that in this election, 10 people prefer A to B, 10 people prefer B to C, and 10 people prefer C to A. In other words, no matter whom we choose as the winner, 2/3 of voters would prefer someone else! This scenario is dubbed **Condorcet's Paradox**. In this election, there is no fair resolution. Plurality would result in a 3-way tie, and so would Borda Count. IRV would fail at the first step because no one can be eliminated based on this schedule. Neither Condorcet's method nor Copeland's method would produce a winner.

It is because of this impossibility of a totally fair method that dozens of election methods and variants are all still in use. Usually the decision of which method to use is based on what seems fairest for the situation in which it is being applied.

Approval Voting

Up until now, we have been considering voting methods that require ranking of candidates on a preference ballot. There is another method of voting that can be more appropriate in some

decision-making scenarios. With approval voting, the ballot asks you to mark (check) all choices that you find **acceptable**. The results are tallied, and the option with the most approval is the winner.

Example 13

A group of friends is trying to decide on a movie to watch. Three choices are provided, and each person is asked to mark with an "X" which movies they are willing to watch. The results are as follows:

	Bob	**Ann**	**Marv**	**Alice**	**Eve**	**Omar**	**Lupe**	**Dave**	**Tish**	**Jim**
Titanic		X	X			X		X		X
Scream	X		X	X		X	X		X	
The Matrix	X	X	X	X	X		X			X

Totaling the results, we find *Titanic* received 5 approvals, *Scream* received 6, and *The Matrix* received 7. In this vote, *The Matrix* would be the winner.

Try It Now 6

Our managers deciding on a conference location from earlier decide to use approval voting. Their votes are tallied below. Find the winner using approval voting.

	30	**10**	**15**	**20**	**15**	**5**	**5**
Seattle	X	X	X			X	
Tacoma	X		X	X	X	X	
Puyallup		X		X	X	X	
Olympia			X		X		X

What Could Go Wrong with Approval Voting?

Approval voting can very easily violate the Majority Criterion.

Example 14

Consider the voting schedule:

	80	**15**	**5**
1st	A	B	C
2nd	B	C	B
3rd	C	A	A

Clearly A is the majority winner. Now suppose that this election was held using approval voting, and every voter marked approval of their top two candidates. Who would win then?

[Solution]

A would receive approval from 80 voters, B from all 100 voters (because B is ranked 1st or 2nd by everyone), and C from 20. Therefore, B would be the winner.

Some argue that approval voting tends to vote the **least disliked** candidate rather than the most liked candidate.

Additionally, approval voting is susceptible to strategic insincere voting, in which a voter does not vote their true preference in order to increase the chances of their choice winning. For example, in the movie example above, suppose Bob and Alice would much rather watch *Scream*. They remove *The Matrix* from their approval list, resulting in a different result.

	Bob	**Ann**	**Marv**	**Alice**	**Eve**	**Omar**	**Lupe**	**Dave**	**Tish**	**Jim**
Titanic		X	X			X		X		X
Scream	X		X	X		X	X		X	
The Matrix		X	X		X		X			X

Totaling the results, we find *Titanic* received 5 approvals, *Scream* received 6, and *The Matrix* 5. By voting insincerely, Bob and Alice were able to sway the result in favor of their preference.

Voting in America

In American politics, there is a lot more to selecting our representatives than simply casting and counting ballots. The process of selecting the President is even more complicated (as we have seen in recent years). Instead, let us look at the process by which state congressional representatives and local politicians get elected.

For most offices, a sequence of two public votes is held: a primary election and the general election. For non-partisan offices in which political party affiliation is not declared, the primary election is usually used to narrow the field of candidates.

Typically, the two candidates receiving the most votes in the primary will then move forward to the general election. While somewhat similar to instant-runoff voting, this is actually an example of **sequential voting** - a process in which voters cast totally new ballots after each round of eliminations. Sequential voting has become quite common in television, where it is used in reality competition shows like *American Idol.*

Congressional, county, and city representatives are partisan offices, in which candidates usually declare themselves a member of a political party. As with non-partisan offices, a primary election is usually held to narrow down the field prior to the general election. Prior to the primary election, the candidate would have met with the political party leaders and gotten their approval to run under that party's affiliation.

In some states a **closed primary** is used, in which only voters who are members of the party can vote on their candidates. In other states, an **open primary** is used, in which any voter can pick the party whose primary they want to vote in. Other states use **caucuses**, which are basically meetings of the political parties, only open to party members. Closed primaries are often disliked by independent voters, who like the flexibility to change which party they are voting in. Open primaries do have the disadvantage that they allow raiding, in which a voter will vote in their non-preferred party's primary with the intent of selecting a weaker opponent for their preferred party's candidate.

Some states currently use a different method, called a **top-two primary**, in which voters select candidates from all political parties on the primary and then top two candidates, regardless of party affiliation, move on to the general election. While this method is liked by independent voters, it gives the political parties incentive to select a top candidate internally before the primary so that two candidates will not split the party's vote.

Regardless of the primary type, the general election is the main election, open to all voters. Except in the case of the top 2 primary, the top candidate from each major political party would be included in the general election. While rules vary state-to-state, for independent or

minor-party candidates to get listed on the ballot, they typically have to gather a certain number of signatures to petition for inclusion.

Exercises 8.3

1. To decide on a new website design, the designer asks people to rank three designs that have been created (labeled A, B, and C). The individual ballots are shown below. Create a preference table.

   ABC, ABC, ACB, BAC, BCA, BCA, ACB, CAB, CAB, BCA, ACB, ABC

2. To decide on a movie to watch, a group of friends all vote for one of the choices (labeled A, B, and C). The individual ballots are shown below. Create a preference table.

   CAB, CBA, BAC, BCA, CBA, ABC, ABC, CBA, BCA, CAB, CAB, BAC

3. The planning committee for a renewable energy trade show is trying to decide what city to hold their next show in. The votes are shown below.

Number of voters	**9**	**19**	**11**	**8**
1st choice	Buffalo	Atlanta	Chicago	Buffalo
2nd choice	Atlanta	Buffalo	Buffalo	Chicago
3rd choice	Chicago	Chicago	Atlanta	Atlanta

   (a) How many people voted in this election?
   (b) How many votes are needed for a majority?
   (c) Find the winner under the plurality method.
   (d) Find the winner under Borda Count.
   (e) Find the winner under the Instant-Runoff Voting method.
   (f) Find the winner under the Condorcet method (if any).

4. A non-profit agency is electing a new chair of the board. The votes are shown below.

Number of voters	11	5	10	3
1st choice	Atkins	Cortez	Burke	Atkins
2nd choice	Cortez	Burke	Cortez	Burke
3rd choice	Burke	Atkins	Atkins	Cortez

(a) How many people voted in this election?
(b) How many votes are needed for a majority?
(c) Find the winner under the plurality method.
(d) Find the winner under Borda Count.
(e) Find the winner under the Instant-Runoff Voting method.
(f) Find the winner under the Condorcet method (if any).

5. The student government is holding elections for president. There are four candidates (labeled A, B, C, and D for convenience). The preference schedule for the election is as follows:

Number of voters	120	50	40	90	60	100
1st choice	C	B	D	A	A	D
2nd choice	D	C	A	C	D	B
3rd choice	B	A	B	B	C	A
4th choice	A	D	C	D	B	C

(a) How many people voted in this election?
(b) How many votes are needed for a majority?
(c) Find the winner under the plurality method.
(d) Find the winner under Borda Count.
(e) Find the winner under the Instant-Runoff Voting method.
(f) Find the winner under Copeland's method (if any).

6. The homeowners association is deciding a new set of neighborhood standards for yard maintenance. Four options have been proposed. The votes are as follows:

Number of voters	8	9	11	7	7
1st choice	B	A	D	A	B
2nd choice	C	D	B	B	A
3rd choice	A	C	C	D	C
4th choice	D	B	A	C	D

   (a) How many people voted in this election?
   (b) How many votes are needed for a majority?
   (c) Find the winner under the plurality method.
   (d) Find the winner under Borda Count.
   (e) Find the winner under the Instant-Runoff Voting method.
   (f) Find the winner under Copeland's method (if any).

7. Consider an election with 129 votes.
   (a) If there are 4 candidates, what is the smallest number of votes that the winner could have under the plurality method?
   (b) If there are 8 candidates, what is the smallest number of votes that the winner could have under the plurality method?

8. Consider an election with 953 votes.
   (a) If there are 7 candidates, what is the smallest number of votes that the winner could have under the plurality method?
   (b) If there are 8 candidates, what is the smallest number of votes that the winner could have under the plurality method?

9. See the preference schedule below. Find the Condorcet winner if there is one.

Number of voters	14	15	2
1st choice	A	C	B
2nd choice	B	B	C
3rd choice	C	A	A

10. See the preference schedule below. Find the Condorcet winner if there is one.

Number of voters	8	7	6
1st choice	A	C	B
2nd choice	B	B	C
3rd choice	C	A	A

11. The marketing committee at a company decides to vote on a new company logo. They decide to use approval voting. Their results are tallied below. Which logo wins under approval voting?

Number of voters	8	7	6	3
A	X	X		
B	X		X	X
C		X	X	X

12. The downtown business association is electing a new chairperson and decides to use approval voting. The tally is below. Which candidate wins under approval voting?

Number of voters	8	7	6	3	4	2	5
A	X	X			X		X
B	X		X	X			X
C		X	X	X		X	
D	X		X		X	X	

13. An election resulted in Candidate A winning, with Candidate B coming in a close second, and candidate C being a distant third. If for some reason the election had to be held again and C decided to drop out of the election, which caused B to become the winner, which fairness criterion was violated in this election?

14. An election resulted in Candidate A winning, with Candidate B coming in a close second, and candidate C being a distant third. If for some reason the election had to be held again and many people who had voted for C switched their preferences to favor A, which caused B to become the winner, which fairness criterion was violated in this election?

15. An election resulted in Candidate A winning, with Candidate B coming in a close second, and candidate C being a distant third. If in a head-to-head comparison a majority of people prefer B over A as well as B over C, which fairness criterion was violated in this election?

16. An election resulted in Candidate A winning, with Candidate B coming in a close second, and candidate C being a distant third. If B had received a majority of first-place votes, which fairness criterion was violated in this election?

17. In the election shown below under the plurality method, explain why voters in the last column might be inclined to vote insincerely. How could it affect the outcome of the election? Explain.

Number of voters	**96**	**90**	**10**
1st choice	A	B	C
2nd choice	B	A	B
3rd choice	C	C	A

18. In the election shown below under the Borda Count method, explain why voters in the last column might be inclined to vote insincerely. How could it affect the outcome of the election? Explain. (Hint: Remember, for Borda Count, you need to count.)

Number of voters	**20**	**18**
1st choice	A	B
2nd choice	B	A
3rd choice	C	C

19. The **Pareto criterion** is another fairness criterion. It states this: *If every voter prefers choice A to choice B, then B should not be the winner*. Explain why plurality, Condorcet, Instant-Runoff, Borda Count, and Copeland's method all satisfy the Pareto condition.

20. Coombs' method is a variation of IRV. In Coombs' method, the choice with the most last place votes is eliminated. Apply Coombs' method to the preference schedules from Exercises 3 and 4.

21. The plurality method is used in most U.S. elections. Some people feel that Ross Perot in 1992 and Ralph Nader in 2000 changed what the outcome of the election would have been if they had not run. Research the outcomes of these elections and explain how each candidate could have affected the outcome of the elections (for the 2000 election, you may wish to focus on the count in Florida). Describe how an alternative voting method could have avoided this issue.

22. Instant-Runoff Voting and approval voting have supporters advocating that these methods be adopted in the United States and elsewhere to decide elections. Research comparisons between the two methods describing the advantages and disadvantages of each in practice. Summarize the comparisons and form your own opinion about whether either method should be adopted.

23. In a primary system, a first vote is held with multiple candidates. In some states, each political party has its own primary. In the state of Washington, there is a top-two primary, where all candidates are on the ballot and the top two candidates advance to the general election, regardless of party. Compare and contrast the top-two primary with general election system to Instant-Runoff Voting, considering both methodological differences and practical differences like cost, campaigning, and fairness.

24. In a primary system, a first vote is held with multiple candidates. In many states, voters must declare a party to vote in the primary election, and they are able to choose only from candidates for their declared party. The top candidate from each party then advances to the general election. Compare and contrast this primary with general election system to Instant-Runoff Voting, considering both methodological differences and practical differences like cost, campaigning, and fairness.

25. Sometimes in a voting scenario it is desirable to rank the candidates, either to establish preference order between a set of choices, or because the election requires multiple winners. For example, a hiring committee may have 30 candidates apply and need to select 6 to interview, so the voting by the committee would need to produce the top 6 candidates. Describe how plurality, Condorcet, Instant-Runoff Voting, Borda Count, and Copeland's method could be extended to produce a ranked list of candidates.

For Further Explorations…

In April, 2021, the Census Bureau published the results of the U.S. Census 2020, and the data was used to re-calculate the number of congressional seats for each state. Because each state needs at least one representative, the Census Bureau does not exactly follow the procedures described in this book for the Huntington-Hill Method (modifying the divisor by trial-and-error); rather, as the bureau explains on their websites,[180] they follow a precise algorithm based on the Huntington-Hill Method using geometric means.

They start the process by assigning one seat to each state, thus filling the first fifty seats. Then, for each state (say with population P), what is called the "*priority value* v_n for obtaining its n^{th} representative" is calculated:

$$v_n = \frac{P}{\sqrt{n(n-1)}}$$

Note that this is simply the population divided by the geometric mean of n and $n-1$ (and the divisor d does not appear anywhere). The idea is that this priority value indicates how big the population must be in order to "win" the n^{th} congressional seat for that state.

Once each state is awarded one seat in Congress, the values v_2 for all fifty states are compared, and the state with the highest priority value gets its second seat. Then, the process continues, now using the new value v_3 for that one state which has already obtained two seats. So, at each step, all fifty states are being compared based on its priority value to get the next seat. This iteration continues until all 435 congressional seats are filled.

Let us use the actual data from the 2020 Census.[181] The "apportionment population" (covering only the fifty states, excluding Washington, D.C. and U.S. territories) of the United States was 331,108,434. If you divide this by 435, you see that, on the average, each congressperson should represent 761,168.8 residents. This is what we called the initial divisor d.

Here are the populations of some states based on the 2020 Census:

- California: 39,576,757
- Texas: 29,183,290
- New York: 20,215,751
- Minnesota: 5,709,752

Following the algorithm described above, the 51st seat naturally goes to the most populous state, California, with v_2 exceeds 27 million, far more than the second-place state, Texas, with

[180] United States Census Bureau. (2021). Computing Apportionment. https://www.census.gov/topics/public-sector/congressional-apportionment/about/computing.html

[181] Available at https://www.census.gov/data/tables/2020/dec/2020-apportionment-data.html.

20.6 million. For the 52nd seat, we now compare v_3 of California (16.2 million) with v_2 of the other 49 states. Texas, with 20.6 million, gets the 52nd seat. For the 53rd seat, we compare v_3 of California (16.2 million), v_3 of Texas (11.9 million), and v_2 of the other 48 states. California again has the highest priority value, winning its third seat. In other words, if Congress consists of only 53 seats, California would have 3, Texas 2, and all other states would have 1 seat each. Strange, isn't it?[182]

Continuing on, the next seven to be picked are Florida, New York, Texas, California, Pennsylvania, Illinois, and California (in that order). If, therefore, Congress had only 60 seats, California would have 5, Texas 3, Florida, Illinois, New York, and Pennsylvania with 2 each, and the other 44 states with 1 each.

We now jump to the end of this process. With all but one seat filled (434 seats taken), which state gets that coveted last seat, the 435th congressperson, based on the 2020 Census?

At this point, we have California with 52 seats, Texas with 38, New York with 26, and Minnesota with 7. Here are the priority values for the next representative for these states, respectively:

- California: $v_{53} = 753{,}877.181$
- Texas: $v_{39} = 758{,}071.464$
- New York: $v_{27} = 762{,}994.353$
- Minnesota: $v_8 = 762{,}997.705$

Do you see how close the values of New York and Minnesota are? With this slight margin, the last congressional seat goes to Minnesota, which ends up with 8.

Now, here is an amazing thing. If New York had *89 more people* counted in the census (yes, just 89), then the priority values of these two states would have been as follows for that last seat:

- New York: $v_{27} = 762{,}997.712$
- Minnesota: $v_8 = 762{,}997.705$

With an even smaller margin, New York would have gotten its 27th seat, the very last seat in Congress (and Minnesota would not have gotten its 8th seat). Imagine that! With over 20 million people counted in the state of New York, 89 more would have made a difference in apportionment.[183]

[182] This scenario was described in Episode 485 (April 29, 2021) of the podcast *Opening Arguments.*

[183] Because this number is almost negligibly small in the context of a national census, it is possible that some states may file lawsuits in pursuit of more congressional seats. Legal and political ramifications of the 2020 Census results are yet to be seen.

The fact that the algorithm presented by the Census Bureau is equivalent to using the Huntington-Hill Method is rather difficult to see, and it is well beyond the scope of this book. However, it is worth noting that these priority values are closely related to the divisor we modified in Section 8.1. In fact, as stated above, the original divisor d based on the 2020 Census data is 761,168.8. However, if we apply the Huntington-Hill Method with this divisor, we would get 437 representatives (after rounding based on the geometric means). In order to get the required 435 seats, we modify it to $d = 762{,}995$ or any number between the actual (highlighted) values of v_{27} of New York and v_8 of Minnesota shown above (without the addition of 89 to New York). You can do this on your own if you can enter formulas in Excel or some other spreadsheet program.

According to the data from the 2020 Census, three states benefited from the use of geometric means instead of arithmetic means: Minnesota, Montana, and Rhode Island. These states must be very thankful for the hard work of Edward V. Huntington and Joseph A. Hill, the two mathematicians who devised this method almost a century ago.

Figure 8.11: Edward V. Huntington[184]

[184] Image is in the public domain

Chapter 9: Selected Topics

"The essence of mathematics lies in its freedom."—Georg Cantor

Throughout this book, we have presented various practical and useful topics in which mathematics plays a crucial role in solving problems. These include logical fallacies, abuse of percentages, compound interest, chances, statistics, exponential growth, perspective in art, harmony in music, and political applications. We have also seen the inspiring side of mathematics in topics like the number e, the Golden Ratio, and fractals. Unfortunately, many of these "crowning achievements" in mathematics stay hidden from students because of the traditional emphasis on algebra- and calculus-based college mathematics curricula. In this chapter, we will present a few selected topics featuring some of the "cool stuff," the elegant "gems" that you can appreciate without taking advanced mathematics.

Figure 9.1: A colorful glass pyramid[185]

[185] Photo by Michael Dziedzic on Unsplash

9.1 Four-Color Theorem[xxi]

The **Four-Color Theorem** is about maps drawn on the plane. Imagine a map of the world or of a continent. Typically, each country or territory is colored in such a way that neighboring countries have different colors so that it is clear where the boundary of each country is. Although there are almost 200 countries in the world, obviously we do not need that many colors. In fact, it is intuitive that we need just a few colors. But how many colors do we need if we want to avoid having neighboring countries with the same color?

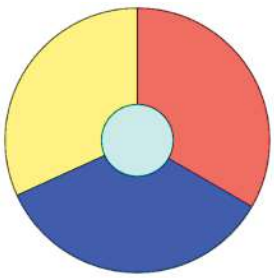
Adjacent Regions and Colors

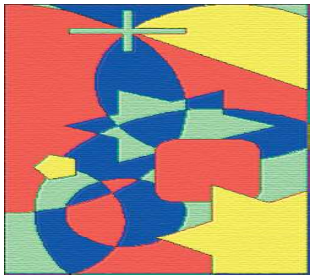
First, we define some terms so that the problem can be precisely stated. On a map drawn on the plane (called a **planar map**), we say two regions are **adjacent** if they share a common border of any length (but a single point does not count). The question posed by the Four-Color Theorem is the following:

What is the minimum number of colors needed so that adjacent regions can have different colors on any planar map?

This problem was first presented in the 1850s in England. Several mathematicians discussed it by mail, attempting to prove the conjecture that four colors would be enough for any map. But despite the effort by many intelligent people, the problem was not solved, nor were any counter-examples produced.

The answer is obviously not 1, 2, or 3 as Figure 1 shows below. Note that this simple "map" has four regions, each adjacent to all other three. In other words, at least four colors are necessary. But is there any map that requires five or more colors? The "map" in Figure 2 is extremely complicated, but still, only four colors are used. Thousands of other maps were considered, but all of them, without exception, could be colored with at most four colors. So it was conjectured that four colors would be enough for *any* map. But can we prove it? What if some clever person, maybe in the future, could come up with a very complicated map that would require 5 colors?





Figures 9.2 & 9.3[186]

[186]Three colors are not enough to color the map in figure 9.2. Figure 9.3 illustrates the potential of a four-color map. Figure 9.2 is in the public domain. Figure 9.3 is used under a CC BY-SA 3.0 license.

It took more than a hundred years, but ultimately the problem *was* solved in the United States, albeit in a somewhat controversial way. Such unique history has made this one of the most famous problems in modern mathematics. Here is the answer (as implied by the name of the theorem itself): ***four colors suffice***.

The Four-Color Theorem

For any planar map, four colors are sufficient to color all regions such that no adjacent regions have the same color.

The three-word phase "Four Colors Suffice" appeared in a postage stamp issued by the University of Illinois at Urbana-Champaign when two of their mathematicians, Kenneth Appel and Wolfgang Haken, proved this theorem in 1976. This phrase is also the title of a popular book on this subject. (In the U.K., where this theorem was first conjectured, they spell it "Four Colours Suffice.")

What is interesting is the fact that, unlike other problems in mathematics, this one is very easy to understand because we are all quite familiar with maps. In fact, it is easy for us to see that the answer is at least four (as illustrated by Figures 1 and 2). Remember, though, we are not asking how many colors are needed for a particular map; we are asking how many colors are needed for *any map with any configuration*.

So *this* was the real challenge. No one was able to come up with a map that absolutely required five colors. Sure, you *can* use five colors if you have five or more regions, but the question is about the **minimum** number of colors needed. To prove that four colors suffice, you would have to try every single map, right? Is that even possible? Maybe not. Is it possible to classify all possible maps into some finite cases?

It was not so hard to prove that five colors are enough to color any map; an elementary proof was given in the late 19th century. Proving that four colors suffice turned out to be significantly more difficult. A number of false proofs and false counterexamples have appeared since the first statement of the Four-Color Theorem at University College London in England in 1852. This story involves Augustus DeMorgan, a logician that appeared in Chapter 1 of this book.

Controversy in the Proof

This was one of the first theorems to be proved by a computer. The conclusion was established by categorizing *all possible configurations* into cases and proving each one separately. The number of cases turned out to be very large. The first proof of the Four-Color Theorem required computer analysis of 1,936 cases. This proof was controversial because most of the cases were checked by a computer program, not by traditional mathematical arguments. The shortest known proof of the Four-Color Theorem today still has over 600 cases. But what if the

program was written incorrectly? How do we know that the code and data were compiled correctly (especially back in the 1970s)? For these reasons, the mathematical community was reluctant to accept the argument as a "proof."

An easier-to-state version of the theorem uses the language of **graph theory**, a branch of finite mathematics. Each region on a map can be "shrunk" down to a point, identified as a **vertex**, and any pair of adjacent regions can be represented by two vertices connected by an **edge**. This way, every planar map can be represented by an undirected graph as shown below. The goal then is to color every vertex such that no edge has the same color on both ends. This type of abstract thinking or representation plays an important role in proofs of many mathematical theorems.

In graph-theoretic terminology, then, the Four-Color Theorem can be summarized concisely by the statement "Every planar graph is four-colorable."

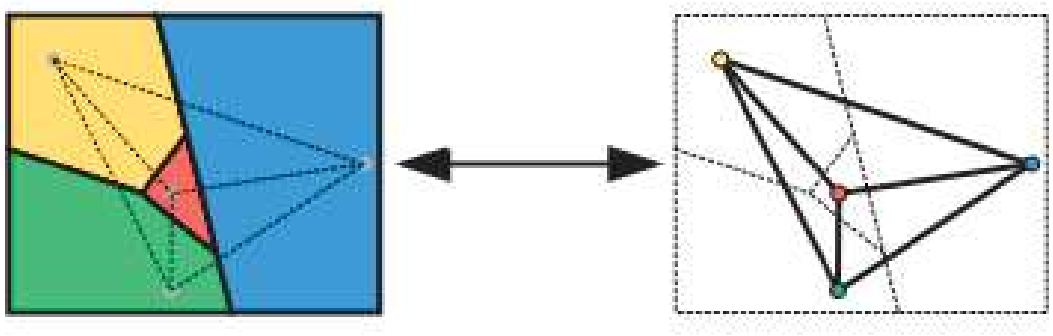


Figure 9.4: Diagram showing a map colored with four colors being transformed into a planer graph.[187].

[187] Image is public domain.

9.2 How Big Is Infinity? Or Is It "Infinities"?

The first numbers you learned as a child were probably whole numbers like 1, 2, 3, and so on. You learned basic quantitative facts like you have one nose, two ears, and ten fingers. So you learned to count before you learned how to add and subtract. You then grew up seeing a lot of objects like cars on the road, people in a concert, and stars in the sky and eventually learned, perhaps intuitively, that there is no end to this sequence of numbers {1, 2, 3, 4, 5, …}
In other words, you learned that there must be infinitely many numbers. But what exactly is **infinity**?

Whatever it is, it must be bigger than any number you can think of. And, in a sense, that is not a bad definition. You may have seen the notation "∞" as the symbol for infinity; this symbol does *not* represent a number. Let us be clear: ∞ **is not a number**. We will come back to this notation later in this section. For now, let us start with a very *natural* set of numbers.

Natural Numbers

The first set of numbers you learned, {1, 2, 3, 4, 5, …}, is called the set[188] of **natural numbers**, and the symbol $\mathbb{N}$ is used to denote this set.[189] Clearly, $\mathbb{N}$ is an **infinite set** since it contains more than any finite number of objects, more than a billion, more than ten trillion, and more than ten trillion factorial. Now, you can easily see that the sum of any two natural numbers is also a natural number. Similarly, you can always take any two natural numbers and find the product, which is also a natural number. We call this property of $\mathbb{N}$ "closure." In other words, $\mathbb{N}$ is **closed** under addition and multiplication.

But you have a problem when you try to subtract, say, 5 from 2. While both are natural numbers, the difference, 2 – 5, is not in $\mathbb{N}$. We say, therefore, that $\mathbb{N}$ is *not closed under subtraction*. For centuries, however, this did not cause any problem because this operation appeared to make no sense. After all, how can you take away five apples when you only have two? Or how can five people come out of a house which contains only two people?

Well, are there any applications for subtraction like 2 – 5? Of course there are. The temperature may go down by 5 degrees from 2 degrees (and the temperature just does not disappear). One can borrow $5 even though, for now, there may only be $2 in the account. So is there a way to make sense out of subtraction like this?

Integers

Of course, there is an easy way to fix this. Extend $\mathbb{N}$ to include numbers like $0, -1, -2,$ etc. We can then define a new set of numbers as follows:

[188] The term "set" is extremely difficult to define formally; here, we simply mean any collection of objects (like numbers). Set theory was founded by Georg Cantor, whose quote appears at the beginning of this chapter.
[189] Some mathematicians, particularly those whose areas of expertise include set theory, include 0 as a natural number and simply call the set {1, 2, 3, 4, 5, …} the "set of positive integers."

$$\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \dots\}$$

Each of these numbers is called an **integer**. Of course, every natural number is an integer. So we now have two infinite sets, $\mathbb{N}$ and $\mathbb{Z}$. Obviously, $\mathbb{Z}$ contains just about twice as many numbers as $\mathbb{N}$, so perhaps $\mathbb{Z}$ is "more infinite" than $\mathbb{N}$, right? Or are they both "just as infinite"? (This question itself sounds weird, doesn't it?

$\mathbb{Z}$ is **closed** under addition, subtraction, and multiplication. This means that the sum, difference, and product of any two integers are all integers, members of $\mathbb{Z}$. However, it is not closed under division because, for instance, 3 and 4 are both integers, but $3 \div 4$ is not.

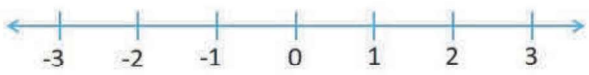
Rational Numbers

This gives rise to yet another infinite set, the set of **fractions** where both the numerator (top) and the denominator (bottom) are integers (and we assume that the denominator is not zero). This set, denoted with the symbol $\mathbb{Q}$, is called the set of **rational numbers** (the term "rational" here comes from the word "ratio"). Any integer divided by any non-zero integer is a rational number—including all of the following:

$$\frac{1}{2}, \quad \frac{-3}{5}, \quad \frac{2}{-7}, \quad \frac{-1}{-3}, \quad \frac{5}{1}, \quad \frac{-6}{2}$$

Of course, the negative sign can be put on the top, bottom, or in the front of the fraction. Two negative signs cancel each other, and some of these fractions can be reduced. Note that every integer can be written with 1 in the denominator, so *every integer is also a rational number*. In other words, $\mathbb{Q}$ is an infinite set which contains the entire set of integers (which contains the entire set of natural numbers). Thus, the size of the set $\mathbb{Q}$ is bigger than the size of $\mathbb{N}$ or $\mathbb{Z}$, right? Or are they equally infinite?

An interesting (and perhaps confusing) property of rational numbers is that these numbers appear to show up "almost everywhere." If you draw a number line (a straight line where each point represents a number), fractions appear almost everywhere while integers are "spread out" at intervals of 1.



For instance, between 0 and 1, there are no integers, but there are many fractions, i.e., rational numbers. As a matter of fact, between any two fractions, there are infinitely many fractions! Because of this property, $\mathbb{Q}$ is called a **dense set**. Thus, certainly $\mathbb{Q}$ appears to be a much, much larger set than the set of natural numbers or integers. But does this then mean

that there are *levels* of infinity? Would it make sense to say that some infinite sets are bigger than other infinite sets?

The answer is mind-boggling (just as the question may be), but we are not quite there yet.

Before we get to the answer (and more questions), we will need at least one more infinite set. It is the set of real numbers, denoted $\mathbb{R}$. Roughly speaking, a **real number** is a number represented by a point on the number line. For each point on the line, there is a real number, and vice-versa. Because each point on the line also represents the distance between that point and 0 (the origin), every real number can be thought of as the length of a line segment. Although a formal definition of a real number is quite difficult and is beyond the scope of this book, this idea of the number line gives us a good understanding of the set $\mathbb{R}$, which is obviously infinite as it contains all of $\mathbb{Q}$.

Irrational Numbers

Based on the paragraphs above, you may be under the impression that $\mathbb{R}$ is the same set as $\mathbb{Q}$, which seems to occupy virtually every point on the line already. Indeed, this was the thought of ancient Greek mathematicians before Pythagoras; they believed that every point on the line can be represented by a ratio of two integers. This idea went well with their philosophy that the world of beauty consists of ratios, and all they needed was ratios of integers, that is, $\mathbb{Q}$. Pythagoras even used ratios and rational numbers to explain musical harmony and chords and is often referred to as the father of music theory, as explained in Chapter 7. But it was also Pythagoras himself who proved that some real numbers (lengths) cannot be represented as a ratio of two integers. For example, imagine the right isosceles triangle (triangle with 45-, 45-, and 90-degree angles) where each of the two legs (the congruent sides) has length 1.

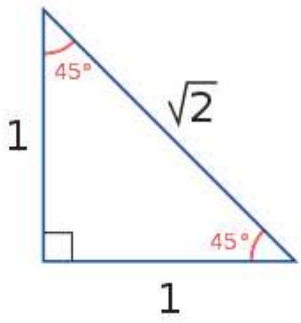


Figure 9.5

The hypotenuse, according to the theorem that bears Pythagoras' name, has length $\sqrt{2}$. Pythagoras proved that no matter how hard one tries, it is impossible to represent $\sqrt{2}$ as the ratio of two integers, thus being forced to admit that there are real numbers (lengths of line segments) that are not rational numbers. These are now called **irrational numbers**. They include not only $\sqrt{2}, \sqrt{3}, \sqrt{5}$, and many other square roots, but also numbers like π (the ratio between the circumference and the diameter of any circle) and e (the number we saw in Chapter 3).

The set of real numbers, $\mathbb{R}$, then contains both the set of rational numbers $\mathbb{Q}$ and the set of irrational numbers, both on the number line, of course. Every point on the line is either a rational number or an irrational number. Once again, it is clear that $\mathbb{R}$ contains the entire set $\mathbb{Q}$ and much more.

We now have four infinite sets, $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$, and $\mathbb{R}$, with each containing the previous set. Let us review:

- $\mathbb{N} = \{1, 2, 3, 4, \dots\}$: the set of natural numbers
- $\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \dots\}$: the set of integers
- $\mathbb{Q} = \left\{\frac{p}{q}\right\}$, where *p* and *q* are integers and *q* is not 0: the set of rational numbers
- $\mathbb{R}$, rational and irrational numbers: the set of real numbers

Georg Cantor and Set Theory

The idea of sets is so intuitive that mathematicians took it for granted for centuries. No one really paid much attention to this fundamental concept, at least formally, until Georg Cantor in the late 1800s. Cantor is one of the most interesting characters in the history of mathematics. He decided to take on the challenging problems involving infinity though many mathematicians had stayed away from infinity, calling it as "belonging to the realm of the divine." Obviously, everyone knew that there are many infinite sets, but Cantor came up with a way to compare their sizes. The size of a set S is called the **cardinality** of S. For finite sets, the cardinality is simply the number of elements in that set. For instance, the set $A = \{3, 5, 6, 7\}$ and the set $B = \{W, X, Y, Z\}$ both have four elements, so the cardinality of both of these sets is 4. For each finite natural number n, there are sets with that cardinality (a simple example is the set $\{1, 2, 3, \dots, n\}$). But how do you find the cardinality of an infinite set? If all of the infinite sets are of the same size, then there is just one infinite (or **transfinite**) cardinality, and that would be the end of the story. However, as Cantor soon discovered, this was *not* the case.
The main tool Cantor used to measure and compare the sizes of various sets is **pairing**. In mathematics, this is called a one-to-one correspondence. It is a way to pair up elements of two different sets (say A and B) in such a way that every element of A is paired up with exactly one element of B with nothing left over on each side. For example, say A is the set of five seats, labeled X1, X2, X3, X4, and X5, and B is the group of five people, Anu, Bob, Carlo, Dee, and Erika. If the five people sit in these seats, in the respective order, there is a natural pairing, or a one-to-one correspondence, in the following way:

Anu → X1
Bob → X2
Carlo → X3
Dee → X4
Erika → X5

No one is left standing, and every seat is occupied. When there is a one-to-one correspondence between two sets, you can clearly see that the two sets are of the same size (in this case, both A and B have cardinality 5). Conversely, if two sets are of the same size, then there should be a one-to-one correspondence. Hence, we have the following:

Definition

Two sets have the same size (**cardinality**) if and only if there is a one-to-one correspondence between the two sets.

With a one-to-one correspondence, if you know the size of one set, you can determine the size of the other set without actually counting the elements. For instance, Dodger Stadium has 56,000 seats. If every person is sitting in one of these seats and if each seat is occupied by exactly one person, you know that there are exactly 56,000 people in the stadium. So a one-to-one correspondence can be used to measure and compare the sizes of two sets.

This does not sound so hard, right? Well, that is true between finite sets, but things can get a little tricky when infinite sets are involved. This is where Cantor made some ground-breaking discoveries at the end of the 19th century. With this definition, he found a way to define a one-to-one correspondence between $\mathbb{N}$ and $\mathbb{Z}$ even though, at a first glance, you may think $\mathbb{N}$ is about half of $\mathbb{Z}$ (because the natural numbers make up only the positive "half" of the integers). Here is one such one-to-one correspondence:

$$1 \to 0, \qquad 2 \to 1, \qquad 3 \to -1, \qquad 4 \to 2, \qquad 5 \to -2, \qquad \text{etc.}$$

Note that the even numbers are paired up with positive numbers in $\mathbb{Z}$ while the odd numbers are paired up with negative numbers in $\mathbb{Z}$ without any duplication and with no elements of $\mathbb{Z}$ left out.

Therefore, despite our intuition, $\mathbb{N}$ and $\mathbb{Z}$ have the same cardinality, i.e., they are "**equally infinite**," or these two sets are of the same size. In fact, it is not difficult to see that all of the following proper subsets of $\mathbb{N}$ are just "as big as" $\mathbb{N}$ itself:

- $A_1 = \{2, 4, 6, 8, 10, \ldots\}$ (the set of even natural numbers)
- $A_2 = \{100, 200, 300, 400, \ldots\}$ (the set of positive multiples of 100)
- $A_3 = \{2, 4, 8, 16, 32, \ldots\}$ (the set of positive powers of 2)

This is strange because a part can be equal (in size) to the whole. In fact, A_2 contains every 100th natural number (one could say 1% of $\mathbb{N}$), but they are both equally infinite in the sense that their cardinalities are the same. This may lead one to believe that all infinite sets have the same cardinality and are thus "equally big."

But what about the dense set $\mathbb{Q}$? This set, the set of rational numbers, contains infinitely many fractions between any two real numbers (between any two distinct points on the line, there are infinitely many rational numbers). This must be much bigger than $\mathbb{N}$ or $\mathbb{Z}$, right?
Wrong!

Again, as counter-intuitive as it may be, Cantor came up with a one-to-one correspondence between $\mathbb{N}$ and $\mathbb{Q}$, thus proving that these two sets are just "as infinite" as each other, i.e., they share the same cardinality. Yes, between 0 and 2, there is only one natural number, namely 1, while there are infinitely many rational numbers. In fact, there are infinitely many rational numbers between 0.999 and 1.001. Yet, the set of natural numbers is just as big as the set of rational numbers, and this fact has been mathematically verified.

You may not even be surprised that all of these sets are of the same size because, after all, they are all infinite, correct? Well, it turns out that *not all infinities are created equal* because here is another thing Cantor was able to show: $\mathbb{R}$ is actually *strictly* bigger than $\mathbb{N}, \mathbb{Z}$, or $\mathbb{Q}$. He proved this with an original, creative method called **diagonalization**. Essentially, he began by assuming that there is a one-to-one correspondence between $\mathbb{N}$ and $\mathbb{R}$ and then showed that this assumption would lead to a contradiction (because one can always come up with a real number left over after any such pairing). Hence, we now know this truth:

Fact
The cardinality of $\mathbb{R}$ is strictly bigger than the cardinality of $\mathbb{N}, \mathbb{Z}$, and $\mathbb{Q}$. In other words, there are **different levels of infinity**.

Any cardinality that is not finite is called **transfinite**. Today, we denote the transfinite cardinality of $\mathbb{N}, \mathbb{Z}$, and $\mathbb{Q}$ as $\aleph_0$(where $\aleph$ is the first letter in the Hebrew alphabet and is pronounced '*Aleph'*). This is the first transfinite cardinality; in other words, you can say $\aleph_0$ is the "smallest infinity" though it may sound like an oxymoron (and your friends and family may wonder if you are feeling OK).

But wait. There is more.

It turns out that the cardinality of $\mathbb{R}$ can be denoted as $2^{\aleph_0}$ (the number of subsets of $\mathbb{N}$), which, as stated above, is strictly bigger than $\aleph_0$. Furthermore, it was discovered that you can always take 2 to any *aleph* number (or any cardinality) to obtain a strictly bigger cardinality. By repeating this process indefinitely, one can obtain (wait for this…) *infinite* levels of infinite numbers. Here, then, is the conclusion:

Fact
There are **infinitely many levels of infinities.**

The story does not end here, either. Cantor also went on to prove many other theorems about transfinite cardinalities. At one point, after proving another surprising result, he famously wrote to his friend Dedekind, "I see it, but I don't believe it."

Despite his outstanding contributions to mathematics, Cantor suffered from depression as his work on transfinite cardinalities was often ridiculed and criticized. Even under those circumstances, however, he continued his work on mathematics. He introduced what has become known as the "**Continuum Hypothesis**," which would (among many things) imply that there is no level of infinity between $\aleph_0$ and $2^{\aleph_0}$. This would mean that there are no sets strictly bigger than $\mathbb{N}$ but strictly smaller than $\mathbb{R}$. Is it true, or is it false? (It must be true or false, right?) Cantor actually made it his lifetime goal to prove this hypothesis and spent much of his energy in late years of his life trying to solve this problem. However, he passed away in 1918 in a mental institution without solving it. Little did he know then that a completely surprising ending would eventually put this fascinating problem to rest—but not for another 45 years. If you are interested in how this story ends (or in the life of Georg Cantor), you are encouraged to do a little research to find out more.

Oh, one more thing… earlier the symbol ∞ was mentioned. This symbol is used to represent that which is bigger than any real number (on the real line); it does not represent any number. In particular, this symbol does not express any cardinality. It simply denotes the idea of "letting a real number x get bigger and bigger without bound."

Challenging Exercises

1. Verify that each of the following numbers is rational. Then, use a calculator or a computer to express each as a decimal number. What do you notice about all of these decimal numbers?

$$\frac{1}{4} \qquad \frac{-2}{3} \qquad \frac{3}{40} \qquad \frac{2}{7} \qquad \frac{12}{4} \qquad -\frac{1}{32}$$

2. Identify the following numbers as natural numbers, integers, rational numbers, irrational numbers, or real numbers. For each number below, list all that apply.

$$0.311 \quad 262 \quad -1\frac{4}{5} \quad \frac{51}{3} \quad \frac{\pi}{2} \quad -5 \quad 0 \quad 0.11111\ldots$$

3. Try to pair up the elements of the two sets below to define a one-to-one correspondence. Describe your pairing so that each element is paired up with exactly one element and nothing is left behind.

$$A = \{1, 2, 3, 4, 5, \ldots\}$$
$$B = \{4, 8, 12, 16, 20, \ldots\}$$

4. Determine if each of the sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$, and $\mathbb{R}$ is closed under addition and multiplication.

5. It is mentioned in this section that $\sqrt{2}, \pi$, and e are not rational. What other irrational numbers are there? Try to find ways to generate more irrational numbers (we know there are infinitely many of them).

6. Following up on the previous question, how big is the set of irrational numbers? Is it as infinite as $\mathbb{N}$ or as infinite as $\mathbb{R}$? Or neither? Any guesses?

7. David Hilbert, a famous German mathematician, introduced "**Hilbert's Hotel**" in 1924. In this grand hotel, there are infinitely many rooms, numbered 1, 2, 3, … One day, every room was occupied because the hotel had infinitely many guests. At night, though, one more guest showed up. The manager, instead of turning down the guest, asked everyone to move to the next room (the guest in Room 1 moved to Room 2, the guest in Room n to Room $(n+1)$, etc.). Now that Room 1 is empty, the guest happily checked into that room.
   (a) If ten more people showed up, each demanding a separate room, is it possible to accommodate all of them? How?
   (b) If a million people showed up, how would you accommodate all of them?
   (c) If infinitely many people showed up, how would you accommodate all of them?
   (d) How big of a group of guests would this very accommodating manager have to have before turning them down? In other words, at what point would the manager say, "No, sorry, there are not enough rooms in the hotel for your group"?

8. Infinity has led to many **paradoxes**—too many to list (maybe infinitely many). Here is one curious scenario. Suppose you have infinitely many ping pong balls, each numbered with a natural number, like 1, 2, 3, … In Stage 1, Balls 1 through 100 are thrown into a bin, and Ball 1 is removed from the bin. In Stage 2, Balls 101 through 200 are thrown into the bin, and Ball 2 is removed. (If you are worried that this would take forever, imaging Stage 1 taking 1 minute, Stage 2 taking half a minute, Stage 3 taking a quarter minute, and so on, each stage taking half the time as the previous stage. Then, this "whole" process will take no more than 2 minutes.)[190]
   (a) After Stage 2, how many balls are in the bin? Which ones? How about after Stage 9?
   (b) After Stage n, how many balls are in the bin? Which ones?
   (c) At the "end" of this process, after the infinitely many stages, how many balls are in the bin? Which ones?

[190] This question also appeared as the last exercise in Section 1.5, but this is an enhanced version of it.

9.3 Seven Bridges of Königsberg[xxii]

Here is another well-known problem in the history of mathematics. The problem was naturally conceived out of an everyday activity—walking around town! Perhaps because of this fact, it is quite easy to understand. The solution is not all that difficult, either. But the historical significance of this problem is quite remarkable; it has led to the birth of new fields of mathematics, **graph theory** and **topology**.

The "Stroll" Problem

Königsberg was a city in Prussia, sort of a "college town" developed around Albertina University, a school founded in the 16th century. In fact, this city was once the capital of East Prussia. Currently located in Russia, this city has been the residence of intellectual giants such as Immanuel Kant and David Hilbert.

In the middle of this town flowed a river—the Pregel, and there were two large islands on this river, both of which were part of the city. Joining the mainland on either side of the river and those two islands there stood seven bridges as shown below.

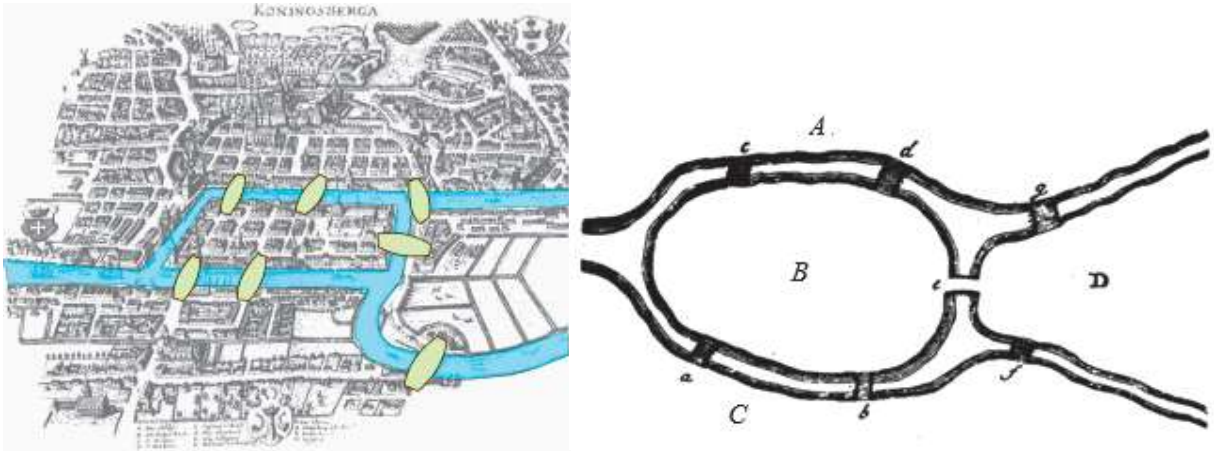


Figure 9.6[191]

It was a popular exercise among the citizens to take a pleasure stroll across the bridges. The question naturally arose: is it possible to cross **every** bridge **only once** during the course of a single walk? More to the point: could this be done such that the walkers end up back at their starting point?

Solution

Leonhard Euler, a Swiss mathematician, solved the problem in 1735 (his last name is pronounced like "Oiler"). Euler realized quickly that, although this problem first appears to involve geometry, fundamentally it is not a geometric problem. After all, we are not concerned with how big each island is, how long or wide the bridges are, or what the bridge angles are.

[191] Image is used under a CC-BY 3.0 license/Image by Leonhard Euler is in the public domain.

Instead, what was important is the relationships between the land and the bridges. So Euler described this as "geometry of position," which eventually led to a full branch of mathematics called **topology**. Over a hundred years before the Four-Color Theorem (Sec. 9.1) was initially posed, Euler figured out that one can represent this relationship using a **graph**, which consists of **vertices** (points) and **edges** (line segments) connecting those vertices. This led to the birth of **graph theory**.

For this problem at hand, each of the landmasses can be "shrunk" (contracted) to a point and thus be represented as a vertex. Each bridge connecting two landmasses can then be represented by an edge connecting those two vertices. For instance, Figure 6 shows that landmasses A and B are connected by two bridges, and so are the landmasses B and C. There is only one bridge each between B and D, between A and D, and between C and D. So the corresponding graph looks like Figure 7 below.

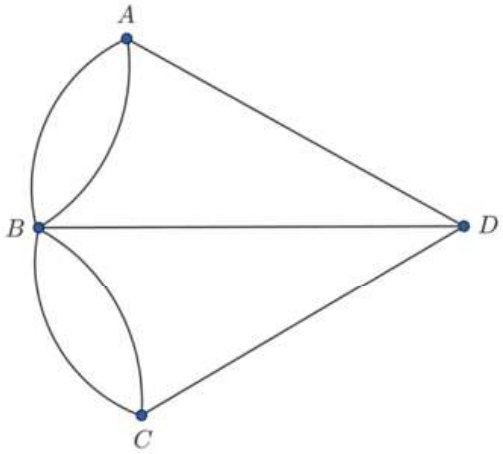


Figure 9.7 A diagram of four vertices.[192]

Notice that Figure 7 has exactly four vertices (each representing a landmass) and 7 edges (each representing a bridge). The question now evolves into these:

1. Is it possible to start at a vertex (a landmass), traverse each edge exactly once, and *return* to the *same vertex* (a vertex can be visited multiple times)?
2. If not, is it possible to start at any vertex, traverse each edge exactly once, and possibly end up at *another vertex*?

In today's language of graph theory, if the latter is possible, we call the sequence of edges traversed an **Euler path** (a path containing each edge exactly once but possibly ending at a point different from the starting point). If the former is possible, since such a path begins and ends at the same point, we call it an **Euler circuit**. These are important concepts in discrete mathematics today.

[192] Königsberg Bridges Graph. Image is used under a CC-BY 3.0 license.

Suppose you start at some vertex, say D, and end your trip at another, say A. Say from D you use the middle edge to reach B. You have to keep going, so you pick another edge from B to some other vertex. So, every time you enter B and leave B, you need two edges. When you come back to B, again you will have to leave B, so you need two more edges. In other words, B must have an even number of edges. In fact, every vertex needs an even number of edges, except at the starting point and the ending point.

In graph theory, the number of edges coming out of a vertex is called the "**degree**" (or "**valence**") of the vertex and is denoted with "*deg*." In this graph shown in Figure 7,

$$\deg(A) = 3,\ \deg(D) = 3,\ \ \deg(C) = 3, \text{and } \deg(B) = 5.$$

Remember, only the starting vertex and the ending vertex can have an odd degree. All other vertices are points to pass through, so they must have an even degree. Clearly, then, this graph in Figure 7 cannot have an Euler path (or an Euler circuit). So here was Euler's solution:

One cannot take a stroll around Königsberg crossing each bridge exactly once.

Now, in addition to crossing each bridge exactly once, if you want to return to the starting vertex, then it is easy to see that *every* vertex must have an even degree.

To summarize, here are the requirements so that an Euler path and an Euler circuit can exist for a given graph G; these are the general solutions to the two questions stated above.

Existence of Euler Circuits/Paths

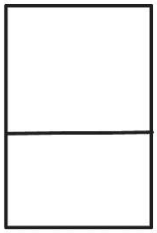
1. G has an **Euler circuit** if and only if every vertex has an even degree.

2. G has an **Euler path** if exactly two vertices have an odd degree. In such a case, an Euler path must begin and end at these two vertices.

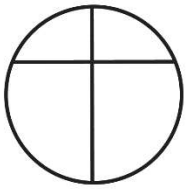
Optional Exercises:

Which of the following graphs have an Euler path? Which ones have an Euler circuit? If any exists, find one.

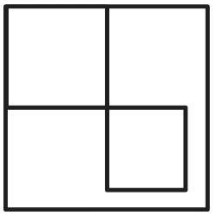
1.



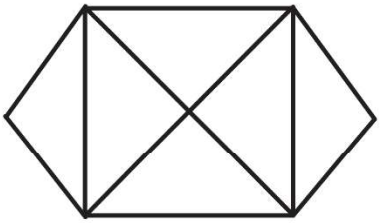
2.



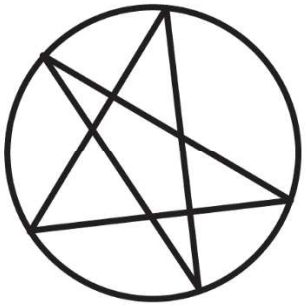
3.



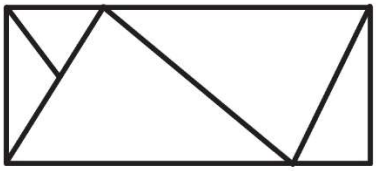
4.



5.



6.



9.4 Russell's Paradox

Russell's Paradox is a well-known logical paradox involving self-reference. It is a little tricky, so you may want to read this section *carefully* and *slowly*.

List of Lists

If you have a list of lists that do not list themselves, then that list must list itself because it does not contain itself. However, if it lists itself, it then contains itself, meaning it cannot list itself. This makes logical usages of "lists of lists" that do not contain themselves somewhat difficult.

Did you get it? Probably not. Let me explain this using a couple of examples below. Since this introduction defines Russell's Paradox with lists, let us begin by making up a few lists.

- List 1: apples, California, James, Canada, sun, basketball
- List 2: Julio, ground, beef, List 2, dachshund, washing machine
- List 3: boys, tables, anger, List 1
- List 4: earphones, correctness, hunger, winter, integers, tuna

Note that List 2 has itself ("List 2") as one of the items on the list. List 3 has "List 1" as an item, but not itself ("List 3"). Good so far?

OK, now, let us create a new list, List L, which consists of all lists that do not list themselves. List 2 lists itself as an item, so List 2 is excluded from List L. The other three are included since they do not include themselves as an item. So, we have

List L: List 1, List 3, List 4….

Here is Russell's question: is List L an item under itself? In other words, among all the items in List L (such as List 1 and List 3), do you expect List L to appear?

If List L appeared as an item under itself, then List L *would* include itself, but remember what List L was? List L was supposed to be the list of all lists that do *not* list themselves. So clearly this cannot happen. Therefore, *List L **cannot** appear as an item under itself*. Right?

Well, then suppose List L does *not* appear as an item under itself. Then, again by the very definition of List L (it is supposed to be the list of all lists that do NOT list themselves), List L *must* appear as an item under itself. Therefore, List L ***must*** appear as an item under itself.

So here is a quick summary so far: If List L appears as an item under itself, then it cannot appear as an item under itself. If, on the other hand, List L does NOT appear as an item under itself, then by definition it must appear as an item under itself.

That, my friend, is a **paradox**. This is Russell's Paradox.

The Barber's Dilemma

There is another version of this paradox which may be a bit easier to understand. This is often called "**The Barber's Dilemma**." Suppose there is a large group of men, one of whom is a barber. This particular barber was ordered to shave those men who do not shave themselves and ONLY those men who do not shave themselves. Got it? Think of dividing the group into two subgroups: Group A, the "shavers" and Group B, the "non-shavers." The barber shaves only those in Group B. So if you are one of the men (not the barber), and if you do not shave yourself (you would be in Group B), the barber will shave you. If you shave yourself (you would be in Group A), the barber will not shave you. Easy, right?

The question is this: does the barber shave himself or not? If he does shave himself, he would be in Group A, so he does not shave himself. If he does not shave himself, he would be in Group B, so he must shave himself. Yes, this guy has a serious dilemma.

You probably noticed the similarity between the two explanations above. This is a paradox generated by **self-referencing**.

By the way, you have probably seen a booklet (like a standard examination booklet) in which a page has no information on it and in the middle of the page, it says,

"*This page intentionally left blank*."

This is often done so that digital documents can be printed correctly double-sided; however, have you noticed the paradox in this? As soon as this (or any other) phrase "This page intentionally left blank" is printed there, the page is no longer left blank. It is similar to sentences like "This statement is false" or "I am lying." You can think of these as examples of Russell's Paradox as well. All of these are paradoxical because of self-reference.

There are many other forms of this paradox, and much has been written about it. See if you can come up with your own story or example that illustrates Russell's Paradox.

This paradox was originally devised in 1901 by Bertrand Russell, probably the only mathematician who won the Nobel Prize in Literature. Interested readers can find many online resources on the exciting life of Russell, a philosopher, mathematician, author, and one of the most brilliant people who ever lived.

Think About it . . . **Russell's Paradox**

Does the above explanation seem a bit confusing? Sometimes abstract concepts can take a minute to click. Watch this video Russell's Paradox - A Ripple in the Foundations of Mathematics on Russel's Paradox to further your understanding of the concept. Is this paradox helpful in understanding mathematics? Why or why not?

9.5 Non-Euclidean Geometry

In your geometry class, you probably learned that the sum of the three angles in any triangle is 180 degrees. This is a well-known theorem in geometry—more specifically, in "**plane**" or "**Euclidean**" **geometry**, which has served as the foundation for much of the learning since the time of the Greek civilization (In fact, the word "mathematics" comes from the Greek word meaning "lessons" or "things to be learned.") But it turns out that if you go away from the "plane," then plane geometry may not work. This gives rise to **non-Euclidean geometry**. It is called "non-Euclidean" because it is different from Euclidean geometry, which was developed by an ancient Greek mathematician named **Euclid** (whom we will discuss below).

An example of Non-Euclidian geometry can be seen by drawing "lines" on a sphere or other round object; straight "lines" that are perpendicular at the equator can meet at the poles. Figure 8 shows a triangular region bounded by three "lines" (the equator and two longitude lines 50 degrees apart). This "triangle" has an angle sum of $90 + 90 + 50 = 230$ degrees!

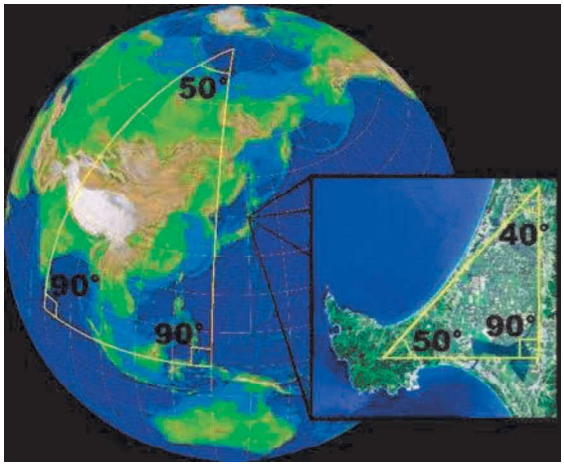


Figure 9.8 On a sphere, the sum of the angles of a triangle is not equal to 180°. The surface of a sphere is not a Euclidean plane, but locally the laws of the Euclidean geometry are good approximations. In a small triangle on the face of the earth, the sum of the angles is very nearly 180°.[193]

You may argue that these are not really *lines* and that the region is not really a *triangle* since these "lines" are actually curved in the 3-dimensional space in which the global is located. We will return to this later.

Historical Background

The birth of non-Euclidean geometry was *really* a big deal. It was truly a ground-shaking event, not only in the history of mathematics and but also in philosophy. You see, Euclid (who lived over 2300 years ago!) wrote a textbook that was so popular that practically every educated person in the world used it to study geometry for the next 2000 plus years. This subject "Euclidean geometry" (the type of geometry you studied in high school) was so popular and dominant that no one, for over two millennia, doubted its truthfulness, questioned its

[193]. Image by Lars H. Rohwedder is used under a CC BY-SA 3.0 license.

authority, or thought of coming up with an alternative. Well, almost no one. The conception and arrival of non-Euclidean geometry involved three mathematicians—one very famous and two completely unknown. The story is worthy of a movie or a play. It is that dramatic. It teaches us that:

- Common sense could be the greatest obstacle to finding truth.
- What stands the "test of time" may not be absolutely true.
- What many experts feel is offensive and repugnant may actually be true.

Here, a very abbreviated version of the story is presented. The reader is encouraged to find out more by doing a search under "non-Euclidean geometry."

Euclid, who lived around 300 B.C.E., is best known for his book ***The Elements***, a 13- volume masterpiece laying the foundations of geometry (and some number theory as well). This book may be the most widely read treatise in world history because no other books have been read longer or by more people, with the exception of the Bible. It was the standard book in geometry for over 2000 years, and there are over 1000 editions of the book in hundreds of languages.

It is too bad that the notions of copyrights and intellectual properties did not exist back then. Euclid could have been quite wealthy with all the royalties he could have earned (except he probably would not have cared—there is a story of Euclid embarrassing and humiliating one of his students who wanted to know what he would gain by learning geometry).

The Importance of *The Elements*

What made this ancient book so important? Well, for one thing, it was the first book that laid the foundation of **deductive logic**—to prove general statements (called **propositions**) by definitions, general assumptions, and already known propositions. If you remember your high school geometry, you may recall memorizing **postulates** (general assumptions) and proving theorems based on known properties and other theorems. That whole thing—which is the fundamental structure of mathematics—was first established by Euclid.

Another amazing accomplishment of Euclid was that he proved tons of propositions—465 to be exact—based on a very small number of assumptions. In fact, he started with only **five axioms**.[194] What is an axiom? It is a (self-evident) statement assumed true without proof. Just as you cannot define every word you use (because each definition uses other words, each of which also needs to be defined using other words), you cannot prove everything; some statements must be *assumed true* at the beginning. (And they had better be obvious to everyone so that no one would question them.) Here are the first four:

[194] More precisely, Euclid began with five assumptions of geometry (called "axioms" here) along with five assumptions of logic (called "common notions" by Euclid and "postulates" here); however, the words "axioms" and "postulates" can be used interchangeably or even switched in some books.

Euclid's Axioms 1 through 4

1. A line can be drawn through any two points.
2. A line can be extended indefinitely in both directions
3. A circle can be drawn with any center point and any radius.
4. All right angles are congruent (equal measure of 90 degrees).

They sound obvious, right? Join the club. Over 2200 years passed without anyone seriously challenging these.

But the last of his five axioms was much longer and more complex than the other four, and it seems that even Euclid himself hesitated using it.[195] It is his fifth axiom (thus often called "**Euclid's Fifth**," like "Beethoven's Fifth"):

Euclid's Fifth Axiom (Parallel Postulate)

Given a line L and a point P that is not on the line, there is one and only one line through P parallel to L.

(Originally this statement was more like this: "If two lines both crossing another line form two interior angles on the same side whose sum is less than two right angles (180 degrees), then the two lines, when extended indefinitely on that side, will eventually meet.")

This original statement may sound all Greek to you (may be because Euclid actually did write this in Greek); that is why we use the other, simpler statement (the first one in the box above) equivalent to the original one. But the original statement is still odd, right? This is *exactly* what caused the controversy and, eventually, an intellectual revolution.

Now, using the example at the beginning of this section (the "triangle" on the sphere), you may be able to see that this postulate is not true on the sphere. You saw the two vertical lines (both perpendicular to the equator) are parallel at the equator but end up meeting each other at the North Pole. In fact, on the sphere, there are no parallel lines. "Yeah, but these are not lines," you may say. That is understandable. You must define what a line is carefully here. But there *is* a way to do so for the sphere (though it is beyond the scope of this book); these are indeed *lines* (more precisely, "geodesics") in spherical geometry, and they indeed bound a spherical triangle. But people did not figure out for a long, long time that spherical geometry is one of the models of non-Euclidean geometry. People had that much faith in Euclid.

[195] In fact, Euclid proved the first 28 propositions without using Axiom 5. Only in proving Proposition 29 did he directly use Axiom 5. While many other proofs depend on Proposition 29 (and thus on Axiom 5), Euclid never directly invoked the fifth axiom anywhere else in his entire work. For this reason, some scholars suspect that Euclid almost foresaw other geometries; some would even go so far as to call Euclid the first "non-Euclidean" geometer.

Because this one particular axiom (the "Fifth") was so odd, many mathematicians wanted to clean up the system—more specifically, they wanted to prove that Axiom 5 follows directly from the first four axioms so they could eliminate it altogether.

But they all failed. Some of the most intelligent people who ever lived could not do it. It was not just a few folks either. Ptolemy tried it in the 2nd century, and Girolamo Saccheri in the 18th century, and many others in between. In fact, so many had attempted this that in 1763, Georg S. Klügel wrote a paper on how *not* to prove it.

Let us review the problem again. What were they trying to do? Remember that axioms (postulates) are assumptions we begin with (and accept as truths), perhaps what Thomas Jefferson (who, as a political philosopher and an architect, knew Euclid's textbook extremely well) referred to as "self-evident" truths. Everyone was fine with Axioms 1 through 4. Axiom 5, on the other hand, was quite different, lengthy, and awkward. There were over 400 statements proved by Euclid based on these five axioms. Many of them depended on Axiom 5. If someone could prove Axiom 5 from the first four axioms, then we could simply take Axiom 5 as another proposition, and all of those 400+ propositions would still be true, based on just four axioms. That would be really cool.

By the way, among the more than 400 propositions were quite obvious statements like the following:

- The sum of the three angles of any triangle equals 180 degrees.
- There are bigger and smaller triangles with the same set of three angle measures (called **similar** triangles).
- There are 4-sided polygons with four right angles (we simply call them **rectangles**).

Then Came the 19th Century

This may be hard to believe, but there were mathematicians who had dedicated their lives to solving this problem—and failed. Among them was **Wolfgang (Farkas) Bolyai**, a long-time friend and classmate of **Carl F. Gauss**, one of the most celebrated mathematicians of all times. After years of frustration, Wolfgang Bolyai wrote to his son **John (Janos) Bolyai**, telling him not to waste a lifetime as he himself had. These were the words of Wolfgang to his son in his 1820 letter:

> *I have traversed this bottomless night, which extinguished all light and joy of my life. I entreat you, leave the science of parallels alone… I saw that no man can reach the bottom of the night… I have traveled past all reefs on this infernal Dead Sea and have always come back with broken mast and torn sail… I thoughtlessly risked my life and happiness.*

Pretty depressing, right? Well, what would a boy do when his father tells him *not* to do something? Yes, that is exactly what the son Janos did. The young, unknown student Janos Bolyai worked really hard on this problem, and within a few years he wrote back to his father with these words:

> *I have discovered such wonderful things that I was amazed… Out of nothing I have created a strange new universe.*

A young crazy dreamer, correct? Maybe he was. In fact, he did *not* solve the problem. He did *not* prove Axiom 5 using the first four. In fact, he did *exactly the opposite*; he assumed that Axiom 5 was **false** and pursued what difference it would make. The difference turned out to be huge. The three "obvious" facts listed above would be all false in this "strange new universe" Janos Bolyai had created. Now, that was a bold statement. Who would question the existence of a rectangle? (Remember, though, this was all theoretical, so it would be possible to imagine such a world which is consistent with the four axioms plus the negation of the fifth.)

Recognizing his son's marvelous work (maybe as a proud and biased dad), Wolfgang Bolyai decided to publish Janos' results in 1832—but not as a book or a paper, but as an appendix to his own book he was publishing. Remember he had a famous friend? Wolfgang sent a copy of this book to Prof. Carl F. Gauss, wondering what kind of reaction Gauss would have. Perhaps Bolyai the elder was hoping that Gauss would be so impressed that the renowned professor would praise the work of his brilliant son Janos.

The response? Well, in a sense, Gauss was impressed. But this is what Gauss said about the "strange new universe" created by young Janos Bolyai: "*To praise it would amount to praising myself*."

What?? Gauss was saying that he had secretly been working on the same thing and came up with pretty much the same results as Janos Bolyai, except Gauss had stuffed all the notes in his drawer, hoping that perhaps someone would publish them after his death. He was afraid of ruining his reputation because it was *such a crazy idea*. Who would challenge the truthfulness of Axiom 5, which the entire world had accepted as absolute truth for over 2000 years? Some people claim that Gauss was especially afraid of being ridiculed by the philosophical group that had been following Immanuel Kant (1724—1804), a leading philosopher who had scoffed at anyone challenging Euclid's axioms.

So here we have a brilliant, well-known German mathematician trying to avoid criticism, his old Hungarian friend who was himself a mathematician, and that inquisitive, incredibly smart son who had developed a crazy new idea. What do you think happened to this son Janos? After receiving the reply from Gauss, this gifted young man stopped studying mathematics altogether and never published anything else, except for those few pages which have become known as the "most important appendix in the history of the world." The fact that the whole academic world pretty much ignored the appendix did not help the situation either.

Russian Connection

Meanwhile, over in Russia (yes, in a totally different country), there was another person who was independently working on the same idea, coming up with pretty much the same results. His name is **Nikolai Lobachevsky** (1792—1856). Without electronic communications or social media, this Russian mathematician had no idea what Gauss and Bolyai were up to. In 1829, two years before Janos Bolyai's appendix was published, Lobachevsky had published his work on non-Euclidean geometry in a little-known journal, written in Russian, published at Kazan University. Neither Bolyai nor Guass read Russian. Even if they had, they would not have known about this minor journal. But because of this publication, Lobachevsky became the first person to ever publish non-Euclidean geometry.

What did Lobachevsky say (in Russian) that was so controversial? Well, in a sense, he showed the world that no one would be able to prove Axiom 5 from the first four axioms because it is possible to create a new and consistent world in which Axiom 5 is false. In that *world* ("imaginary" in Lobachevsky's word), similar triangles would not exist, and neither do rectangles. In our Euclidean world, people may laugh at you for trying to construct a pentagon (a 5-sided polygon) where every angle is 90 degrees (you cannot do that). In a non-Euclidean world, people would laugh at you if you tried to describe a rectangular chocolate box (since there is no such thing as a rectangle)!

Mathematically, this meant that Euclid's Fifth Axiom was neither true nor false. Remember what an axiom is? It is a self-evident statement *assumed true* as a starting point. What Euclid showed over two thousand years ago was that if you assume all of his five axioms, then you would end up with Euclidean (plane) geometry. What Bolyai, Gauss, and Lobachevsky showed was that if you assume the first four axioms of Euclid and the *negation of the fifth axiom*, then you simply get different types of geometry, those we now call "non-Euclidean" types.

Unlike Bolyai, Lobachevsky did not have a famous friend to write to. All he had was a group of traditional, old-fashioned colleagues and scholars who thought he had gone insane. Lobachevsky was relentlessly criticized, mocked, and rejected by the academic world. His new "imaginary" geometry represented the "shamelessness of false new inventions" to them. The news eventually reached Gauss in Germany. While recognizing Lobachevsky's work as great (because it was just like his own), Gauss still did not support or endorse this new idea in fear of criticism.

Lobachevsky was fired from the university (perhaps due to his failing health, perhaps due to his crazy new idea), lost a child, became blind, and died in poverty. This obscure Russian mathematician, who had (correctly) challenged the 2000-year-old teaching and reached the surprising apex, the existence of non-Euclidean geometry, ended up leading a tragic life. Perhaps Janos' father Wolfgang was right when he wrote of the "broken mast" and "torn sail."

So What Is the Point of All This?

So why does all this matter? Is it not just a theoretical world, totally separate from the real world where we live? Well, not really. One example of non-Euclidean geometry is **spherical geometry**—the entire navigation theory is based on this. Sailors and airline pilots travel around the world along "spherical lines" (geodesics) on the earth to minimize distance. Spherical geometry is characterized by **positive curvature** while Euclidean geometry assumes **no (zero) curvature**. Another example is **hyperbolic geometry**, characterized by **negative curvature**. When Einstein was studying the structure of our universe, he needed non-Euclidean geometry. In fact, our modern astronomy would not exist without non-Euclidean geometry. Only those eccentric, free-spirited, dreaming, curious humans like young Bolyai and Lobachevsky could help pioneer our space explorations. What appears ridiculous and repugnant may not be so crazy after all.

Imagine a small village where, based on observation, everyone believes that the earth is flat. You and I could sit back and laugh at those people because, well, the earth is not a Euclidean plane; it models spherical (non-Euclidean) geometry. In addition, some research today shows that it is possible that our universe could have curvature, meaning that we may be living in a non-Euclidean universe. Imagine that! A small group of ignorant human beings who think (for 2000+ years) Euclidean geometry is the absolute and the only true geometry while living on a non-Euclidean planet, which revolves around in a non-Euclidean universe. That would be an ultimate irony.

Math in the Real World – *The 5th Axiom*

1. Watch this video The History of Non-Euclidean Geometry - Sacred Geometry - Part 1 - Extra History to get an overview of Euclidean Geometry and the 5th axiom of Euclid.
2. Watch this video The History of Non-Euclidean Geometry - A Most Terrible Possibility - Part 4 - Extra History to understand the basis of Non-Euclidean Geometry.

9.6 Mathematical Problem Solving Illustrated

We close this chapter and the book with an elegant example that illustrates various stages and strategies of mathematical problem solving. Remember, mathematical problem solving does not mean simply to look up a formula online and plug in numbers, or to find a standard procedure and carry it out.

The example below demonstrates how one is to understand and identify the problem, to consider various cases as a start, to come up with a general conjecture, to reflect on the progress and a partial solution, and to pursue the question further to see the entire picture.

You can compare this process to climbing a mountain, perhaps without any maps. You walk along a trail and may hit a dead end. You may have to turn around and try another path. You do not know the quickest way (or any way whatsoever) to get to the top. Worse yet, there may not be any way to reach the top. You are exploring various paths and strategies to move forward—sometimes you encounter something totally unexpected. You may have missed the most efficient way to arrive at the top, but you may get to see some beautiful flowers and trees precisely because you have wandered off and ended up where others have never been or imagined. You may think you have reached the apex, only to find out that there were even higher grounds you can reach. When you do get to the mountain top, you have a grand view, and you understand why your first trail led to a dead end. You may even find a quicker or safer way for your next trip. Either way, you feel fulfilled with a sense of joy and accomplishment. That is how you feel when you have solved a mathematical problem.

I invite you to join this little journey with one last problem—This is called the "Riddle of 100 Lightbulbs,"[196] which is simple enough so that all you need to understand the problem is arithmetic but sophisticated enough so that the solution is elegant and lofty, and the "view from the top" is magnificent. If you do not understand all the details, do not worry. That is perfectly OK. But be sure to give some thought first. Spend some time thinking. Allow yourself the freedom to explore. More often than not, full understanding does not come until after you have solved a problem. As you consider this problem, I trust that you will have some "aha" moments, which are appropriately called "light-bulb" moments!

Problem: The "Riddle of 100 Lightbulbs"

> Imagine 100 lightbulbs, all lined up in a large room, each with a switch labeled 1 through 100. All the lightbulbs are OFF at the beginning. Whenever a switch is "flipped," the corresponding lightbulb gets turned ON if it is OFF and gets turned OFF it is already ON. Now, Person #1 walks in and flips every switch which is a multiple of 1. That is every switch, of course. So after Person #1 leaves, every lightbulb is ON. Then, Person #2 walks in and flips every switch which is a multiple of 2. So after Person #2 leaves, every other lightbulb has been turned OFF. Then Person #3 enters and flips every switch which is a multiple of 3. This

[196] This problem appears on various websites, such as https://www.riddles.com/1002.

continues with Persons #4, #5, etc., all the way to Person #100. Of course, this last person only flips the very last switch in the room. Once Person #100 leaves, exactly which lightbulbs are ON? Can you describe them? Can you explain why?

Figure 9.9 A String of Light Bulbs.[197]

Now, before reading on, stop and think about the problem. Try different ways to see if you can find any patterns. Do whatever you can to make even small progress. Constructive struggle is absolutely essential and often leads to a breakthrough point. Even if you might think you are wasting time, you are *not*! The old saying, "No pain, no gain" accurately describes this initial stage of problem solving.

Initial Strategy: Let's Try a Few

You probably have a mental picture of a long room with 100 lightbulbs and 100 switches, along with these 100 people entering and leaving the room. As time elapses, you will see some lights turning ON while others turning OFF. Where do we start? What is a good way to *begin* thinking of this process?

Maybe we can think "step by step." After Person #2 leaves, the switches turned ON are 1, 3, 5, … Only the odd-numbered switches are ON since all even-numbered ones have just been turned OFF. But then things get a little complicated. Person #3 flips Switch 3 (turning it OFF) and flips Switch 6 (turning it ON), and so forth. So it is already difficult to explain the situation after Person #3 leaves. Perhaps this is *not* such a good idea. Let's start it over.

We know that 100 people will go through this room, so maybe we should shift our focus and consider *each lightbulb* individually. What happens to Switch 1? It gets turned ON by Person #1. But 1 is not a multiple of any bigger number. So Switch 1 stays ON the whole time! Now we know that the first lightbulb will remain ON permanently. This is good. We got at least one result.

What about Switch 2? Again, it gets turned OFF by Person #2, and no one after that person will touch this switch since 2 is not a multiple of any bigger number. So we got another result—

[197] Photo by Luis Tosta on Unsplash

Switch 2 will be OFF at the end. What about Switch 3? In a similar way, Switch 3 gets turned ON by Person #1, stays ON as Person #2 passes by, and gets turned OFF by Person #3. Then it stays OFF. So Switch 3 will be OFF at the end.

What about Switch 4? It gets turned OFF by Person #2, stays OFF as Person #3 passes by, and gets turned ON by Person #4. Then it stays ON. So Switch 4 will be ON at the end.

Do you begin to see *any* pattern?

Pattern and Generalization: Odd or Even Factors?

By this time, you are beginning to see something—a pattern, maybe something that can be generalized. This is mathematical thinking in action. Now you are thinking, not specifically about lightbulbs and people, but about factorization of whole numbers. There is *abstraction*, the essence of problem solving in mathematics.

You may want to continue on with specific switches. Switch 5 gets turned ON by Person #1 and OFF by Person #5, and that is it. The same is true with *any* number with only two factors—1 and the number itself. Wait. We have a name for these numbers, right? These are the **prime numbers**. So you have now reached a partial solution, a conjecture: "*Every prime-numbered lightbulb will be OFF at the end*." So we do not need to consider Switches 7, 11, 13, 17, 19, … They will all be OFF. Cool.

But what about all the others?

Well, Switch 6 gets turned OFF by Person #2, ON by Person #3, and OFF by Person #6. So, even though 6 is not a prime number, it will be OFF. Switch 8 gets turned OFF by Person #2, ON by Person #4, and OFF by Person #8. So Switches 5, 6, 7, and 8 will all be OFF. Now we know that even though every prime-numbered lightbulb will be OFF, its converse is not true, i.e., even if a lightbulb is OFF at the end, it may not be prime-numbered.

Switch 9 gets turned OFF by Person #3 and ON by Person #9. It then stays ON. So far, Switches 1, 4, and 9 will be ON at the end. You may or may not see another pattern at this point.

Switch 10 gets turned OFF by Person #2, ON by Person #5, and OFF by Person #10. Switch 11 is prime, so it will be OFF. Switch 12 gets turned OFF by Person #2, ON by Person #3, OFF by Person #4, ON by Person #6, and OFF by Person #12. Switch 13 is prime.

After a few more attempts, you will discover that *most* switches will be OFF at the end… a lot more than just those prime-numbered ones. Switches 14 and 15 will be OFF, too. But try Switch 16. It gets turned OFF by Person #2, ON by Person #4, OFF by Person #8, and ON by Person #16. It stays ON! So we have 1, 4, 9, and 16 as switches that will be ON at the end.

Reflection: Summarize What We Learned So Far.

Let us stop for a moment and make a list of what we know so far. This is a result of working on specific cases to discover patterns.

- Prime-numbered lightbulbs will be OFF at the end.
- Many other lightbulbs will also be OFF at the end.
- So far, those numbered 1, 4, 9, and 16 will be ON at the end.

In fact, you may have come up with this conjecture: "*Lightbulb labeled n will be ON if and only if n has an even number of factors. This includes all prime numbers*."

Great. We found a solution. This is pretty good, and an "if and only if" conclusion is the best we can get in mathematics. Here, I have used the word "conjecture," but it is really a "theorem" since a proof can easily be written: each factor of n, including 1 and itself, flips the switch, so an even number of factors means OFF while an odd number of factors means ON. We no longer need to describe ON and OFF in terms of people walking through this room; all we need is the number of factors. Simple!

At this point, it is like standing on the top of a hill, enjoying a view of the path you have taken to get there. Stop and smell some flowers. The air is fresh, too. You should feel good.

But wait. Can we go a bit farther? Is it possible?

Exploration: Go Further

My solution sounds good, but there is still something weird or interesting about those lightbulbs that will be ON at the end. So far, they are 1, 4, 9, and 16, all perfect squares. Is it possible that a switch will be ON at the end only for those numbers that are squares of a whole number?

Well, let us test this new conjecture. Maybe we should try a number with many factors. How about, say, 48? It has many factors; since $48 = 2^4 \cdot 3$, its factors are 1, 2, 3, 4, 6, 8, 12, 16, 24, and 48. It has 10 factors, so the switch will be OFF. How about 60? Since $60 = 2^2 \cdot 3 \cdot 5$, its factors are 1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, and 60. It has 20 factors, so the switch will be OFF. Neither 48 nor 60 is a perfect square, so our conjecture still holds. Let us try 64, a perfect square. Since $64 = 2^6$, its factors are 1, 2, 4, 8, 16, 32, and 64. So it has 7 factors, meaning that 64th lightbulb will be ON at the end. Our conjecture seems really good.

You probably noticed by now that our conjecture about perfect squares should be linked to the previous statement we proved; "*Lightbulb labeled n will be ON if and only if n has an even number of factors."* Could it be that n has an even number of factors only if it is a perfect square? If this is true, why?

This is like finding another path, that could possibly lead to a higher ground on this hill we are still climbing. You are not promised a great view; the trail may lead nowhere. But the trail is there. Let us see how far we can go.

Well, we have listed the factors of many numbers so far. Take a look again. In problem solving, it is a great idea to take notes; this helps you clarify your own ideas and keeps a record of what you have tried, including failed attempts.

Failures are awesome!

Failures are crucial because they could shed light on what to try next. Note-taking is so important that some mathematicians carry a journal book with them; some even sleep with a legal pad near their night lamp!

OK, so let us look at those factors again. Sometimes your previous work will jump at you, as if it had been screaming at you. You may want to start *listening* to your work or equations. Shhh. Can you *hear* your numbers trying to tell you something? The factors of 6 are 1, 2, 3, and 6. The factors of 15 are 1, 3, 5, and 15. The factors of 48 are 1, 2, 3, 4, 6, 8, 12, 16, 24, and 48. Why is there always an even number of factors? Contrast these with the factors of 16, which are 1, 2, 4, 8, and 16. Why is there an odd number of factors here?

Stare at these numbers for a few more minutes if necessary. There *is* a reason for many numbers having an even number of factors. Take 48, for example. Every time you write a factor, say 2, you notice that $48 \div 2 = 24$ is also a factor. Same for 4 and 12, 6 and 8, and so on. In other words, these factors *come in pairs!* More generally, we can state *"If p is a factor of n, then $n \div p = q$ is also a factor, so factors come in pairs."* This explains why the number of factors (for most numbers) is even! This is great, but what about those perfect squares? This should be easy now. 8 is a factor of 64, but the quotient $64 \div 8$ is 8 itself, so it does *not* come in pairs (not producing a different factor). More generally, if $n = p^2$, then the factor p gets written only once while all other factors come in pairs, resulting in an odd number of factors. You can also see that this happens only when the number is a perfect square (where the exponent is 2). Perfect cubes like 8 and 27 still have an even number of factors because they are not a square of any whole number as 25 and 64 are.

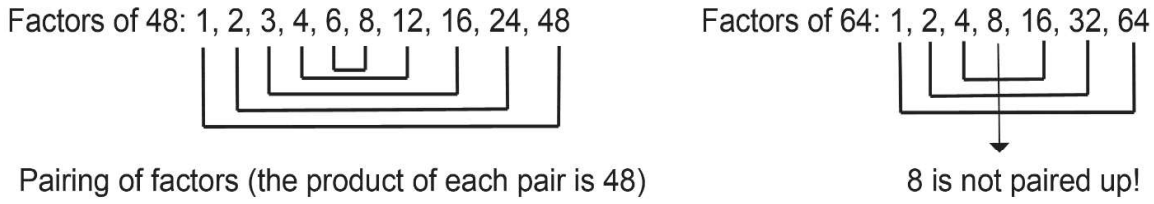


Figure 9.10

Complete Solution: A Grand View

We are now prepared to write the final answer, a complete solution to the problem: "*Lightbulb labeled n will be ON if and only if n is a complete square.*" Why? Here is the reason, which we already wrote above, but this version, a summary, is somewhat polished up so it will look pretty and elegant:

> "A lightbulb gets turned ON/OFF ("toggled") by each of its factors, so for a lightbulb to be ON at the end, there must be an odd number of factors. But for each factor of n, the quotient by that factor is also a factor, so the factors always appear in pairs, resulting in an even number of factors, except when the quotient is exactly the factor itself. This happens only when n is the square of that factor, i.e., when n is a perfect square, in which case there will be an odd number of factors, causing the lightbulb to be ON at the end."

I used the word "reason" for this explanation, but this is really what mathematicians would call a "proof." A proof is a logical argument, a series of sentences deducing a conclusion from what is given in the problem. *A proof does not have to have two columns.* While each step should be justified (as you may have learned in high school geometry), most proofs are written with sentences, and obvious justification can be omitted (such as "$A = A$ by the Reflexive Property").

I also used the word "solution" to refer to this explanation, and indeed it is. A solution to a mathematical problem may not be a number. Here, the solution is not even the list of which lightbulbs will be ON at the end. The proof itself, i.e., the series of sentences above, *is* the mathematical solution to this question. Note this solution does not even include any specific numbers, but that is completely fine.

So here it is! Finally, a complete solution, a grand view from the top of the mountain! I am using this metaphor because, hopefully, you agree that you now see things much more clearly, understanding exactly how we got to the answer. I hope you are breathing that fresh air and feel completely satisfied with the view and what you have just accomplished.

Now, take a look at the problem again. We have come a long way, right? Reading this problem now is like remembering your last vacation destination and thinking of all the fun you had there. Look at your notes. You may compare that to browsing through your photos and selfies you took on your trip. Is this problem practical? Are you ever going to see 100 lightbulbs in the room, with 100 people coming and going, flipping switches? Probably not. But that is OK, too. You see, a great mathematical problem does not have to be useful. It just has to be interesting.

Final Thought: What If…?

When a mathematical problem is solved, and after you take that deep breath of satisfaction, you may start thinking of related problems and other possibilities. "What if…?" is the type of questions I am referring to.

What if there were only 90 lightbulbs in this question, with 100 people entering and leaving as before? What if there were more than 100 people? Imagine 120 people (numbered #1 through #120) entering and leaving the room. (You can probably figure out nothing really changes in these two cases.)

What if there were only 90 people (#1 through #90) but still 100 lightbulbs? Here, the situation would be different. Switches 91, 92, …, 99 would be ON while Switch 100 would be OFF at the end. (See why?) What if there were only 10 people (#1 through #10) and 100 lightbulbs? What if there are 100 people but 150 lightbulbs? How would you change your answer? Can you come up with a more general conjecture? If you can state one and prove it, then that will be an extension, a "generalization," of your previously obtained solution.

On and on, the process could continue for a while… perhaps for a long time.

This is problem solving. It is not a boring procedure. Nor is it a process of blindly following some rules and checking the answers in the back of a textbook.

It is more like an adventure. It is a pursuit of beauty.[198]

This is mathematics.

[198] Francis Su analyzes and describes four types of beauty in mathematics in Chapter 5 "Beauty" of his book *Mathematics for Human Flourishing* (New Haven: Yale University Press, 2020).

Chapter 10: International Topics Supplement

Ideas for Projects on International or Global Mathematical Themes

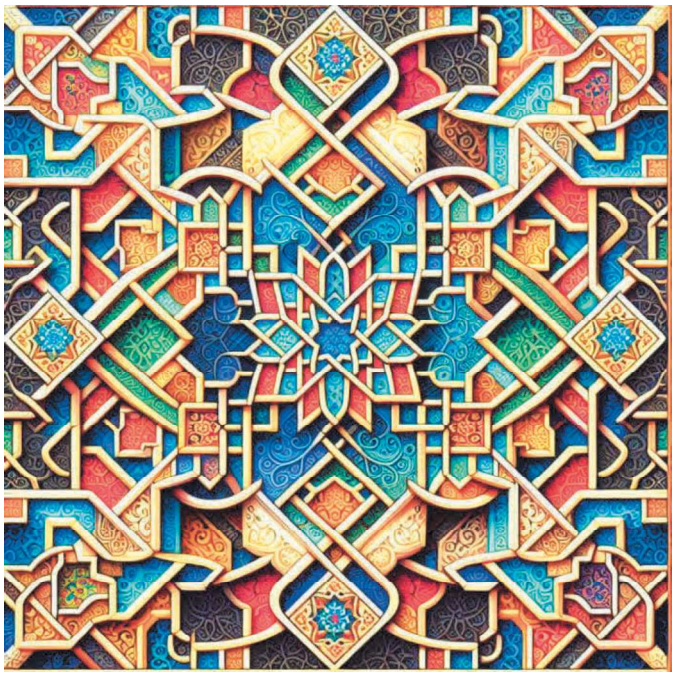


Figure 10.1 Geometric Patterns within Islamic Art.[199]

The following pages include many potential ideas and topics to pursue for projects on international or global mathematical themes. Comprehensive internationalization and global responsibility have become fundamental ideas not only on college campuses but also in our society, which is ever so complex, sophisticated, and connected. The topics are in a wide variety of categories including geometry, numbers, statistics, probability theory, exponential growth, ethnomathematics, and history of mathematics. At the end of each section is a series of questions you can explore, pursue, and find answers for (whenever possible). Some have one correct solution while many others are open questions. As you study these topics, I hope you see the global nature of mathematics and its applications as well as the history of human struggles for solving problems and finding truth.[200]

[199] Image, a geometric pattern reminiscent of Islamic art, generated by ChatGPT

[200] Special thanks go to the Education Partnership for Internationalizing Curriculum (EPIC) Community College Faculty Fellowship program, sponsored by the Global Studies Division of Stanford University, for providing support while I was one of their 2023-2024 fellows.

10.1 Project 1: The Pythagorean Theorem

A Theorem Proved Around the World

Category: Geometry, History of Mathematics

Description:

We look at the Pythagorean Theorem and some proofs of this classic theorem from among (literally) hundreds of proofs given throughout the centuries around the world.

Introduction:

The Pythagorean Theorem may just be the most well-known theorem in mathematics. Someone once met a mathematician and said, "Oh, so you are one of those $a^2 + b^2 = c^2$ kind of people, right?" Indeed, the chances are that even if you may not remember much from your high school geometry class, this may be one of the statements that sound familiar to you.

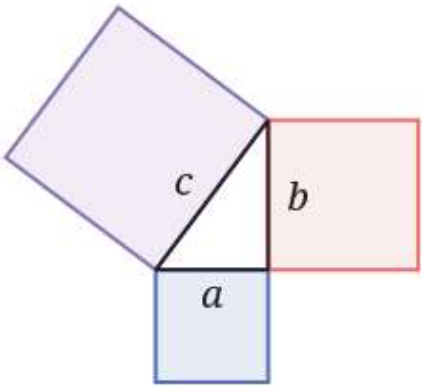


Figure 10.2 The Pythagorean Theorem[201]

The theorem is simple enough: if you have a right triangle, where c represents the length of the hypotenuse (the longest side) while a and b are the lengths of the two legs (the sides forming the right angle), then $a^2 + b^2 = c^2$. In English, the theorem states that the square of the length of the hypotenuse in a right triangle is equal to the sum of the squares of the lengths of its two legs. (In fact, it turns out that its converse is also true: if this equation is true for a given triangle, then it is the right triangle.)

1. Brief History

1.1. Origin

Pythagoras of Samos was a Greek philosopher and mathematician who lived in the 6th Century B.C.E. Yes, a long time ago. Although the famous theorem is named after Pythagoras, it is very

[201] Image by fr:Utilisateur:Steff, is licensed under CC BY-SA 3.0.

likely that the theorem had been known before him and his followers, called the "Pythagoreans," who were his students and disciples. Evidence shows that this famous equation was known in older, ancient civilizations such as India, China, Egypt, and Mesopotamia (although in somewhat different forms). The statement of the theorem also appears on a tablet from the Babylonian civilization dating somewhere in the 17th to 20th century B.C.E. It is generally believed that Pythagoras was credited for this theorem because he was the first one to give a mathematical proof of it; however, there seems to be no way to verify this claim.

1.2. Applications in History

Because $3^2 + 4^2 = 5^2$, the numbers 3, 4, and 5 (called a "Pythagorean triple") have been used to construct right angles for centuries (note that technically, this method is based on the *converse* of the Pythagorean Theorem). For instance, some writing in Egyptian hieroglyphics mentions a rope that has twelve knots equally spaced so that when the rope is tightly held in a triangular shape with sides 3, 4, and 5, the angle formed opposite the 5 knots is a right angle. Since right angles are quite common in architecture, design, and various geometric constructions, the Pythagorean Theorem is quite useful whenever one needs to find the length of one side of a right triangle, given how long the other two sides are.

In fact, Pythagoras himself must have known that the hypotenuse of the right triangle whose legs are of length 1 must be a number whose square is $1^2 + 1^2 = 2$ (today we call this number the positive square root of 2). Pythagoras' attempt to describe this number as a ratio of two integers did not go well. He (or his followers) ultimately proved that $\sqrt{2}$ is not rational, i.e., cannot be expressed as a ratio of two integers, leading them to the astonishing conclusion that not all real numbers are rational.

2. Proofs

Before we talk about proofs, we need to establish what a *theorem* is. In science and other disciplines (even in politics), we often talk about *theories*. A *theory* is different from a *theorem*. A *theory* is a system or collection of ideas in an attempt to explain something, typically in a scientific discipline; as such, a *theory* does NOT necessarily require a logical or deductive proof. Some examples are the Plate Tectonic Theory, the Big Bang Theory, and the theory of evolution by natural selection. It can even refer to an untested guess without concrete evidence, or even to an idea used to justify an action or activity, such as "conspiracy theories."

A *theorem*, on the other hand, is a precise statement, logically a proposition, which has been proved true using logical deduction. A mathematical proposition is not a theorem unless/until it is proved; without a proof, the proposition may be referred to as a *conjecture* (such as the Goldbach Conjecture and the Poincare Conjecture) or as a *hypothesis* (such as the Riemann Hypothesis and the Continuum Hypothesis). This means that once a conjecture or a hypothesis is proved, it becomes a theorem. In other words, if a statement is a theorem, then it must be true (unless the proof given turns out to be incorrect) and remains true forever. Its truthfulness will not be up for debate any longer.

The Pythagorean Theorem, as its name suggests, is a theorem, i.e., a statement that remains true forever. This means that at least one proof is known and accepted. But in fact, many, many proofs of this theorem have been given throughout history. A quick Google search will show you that there are book titles with the Pythagorean Theorem with "eight classic proofs," "24 proofs," and "4000-year history."

It should be mentioned that Elisha Scott Loomis, an Ohio-born mathematician and teacher in the early 20th century, wrote a book called *The Pythagorean Proposition,* a collection of at least 344 (367 according to another) proofs of the Pythagorean Theorem! See late Alexander Bogomolny's website "Cut The Knot," where 122 of the proofs are listed: https://www.cut-the-knot.org/pythagoras/index.shtml#Loomis.

No, we are not going to list all of them here! But let's take a look at some notable proofs by mathematicians all over the world. Truly, mathematical appeals seem universal.

2.1. Greek: Geometric Proofs by Pythagoras and Euclid

No one knows exactly how Pythagoras proved the theorem bearing his name now, but it is almost certain that he used geometric methods. Many believe that his proof involves a process of partitioning areas of the three squares formed outside the triangle, rearranging them, and putting them together for comparison.

Euclid is another ancient Greek mathematician, and he is best known for writing the 13-volume treatise *The Elements*, carefully proving more than 400 theorems (propositions). Written about 300 B.C.E., this book was used as the standard textbook in geometry for over 2000 years (talk about best sellers). While Euclid's definitions and proofs may not meet today's level of mathematical rigor, no one disputes the significance and elegance of this masterpiece. In *The Elements*, Euclid gives a geometric proof of the Pythagorean Theorem. It is assumed that his proof is based on what Pythagoreans used, i.e., dividing and rearranging various parts of the three squares to show the equality.

2.2. Islamic: Algebraic Proofs by Al-Haytham and Omar Khayyam

The global center of mathematics shifted from Athens to Rome, and then to the Middle East. As the great Library of Alexandria in Egypt lost its influence, the House of Wisdom in Baghdad gained its significance during the Islamic Golden Age, starting around the 8th century. Much of Greek literature was translated into Arabic during this period, and *The Elements* was no exception. Several Islamic mathematicians, including Al-Hajjaj ibn Yusuf and Abu'l-Hasan, translated works by Euclid. This is also where algebra (the term itself has an Arabic origin) became a major part of mathematical learning (the Greeks had been mostly focused on geometry). Thus, it is not surprising that some algebra-based proofs were given for the Pythagorean Theorem during this period. For instance, al-Nayrizi, al-Haytham, and Omar Khayyam each seem to have developed an algebraic proof of the theorem independently.

2.3. Asian: Special Case in Chou Pei Suan Ching and Others

Sometime around the 3rd-century B.C.E., a proof of the theorem for the special case (lengths 3-4-5) was illustrated in Chinese writing with the following figure.

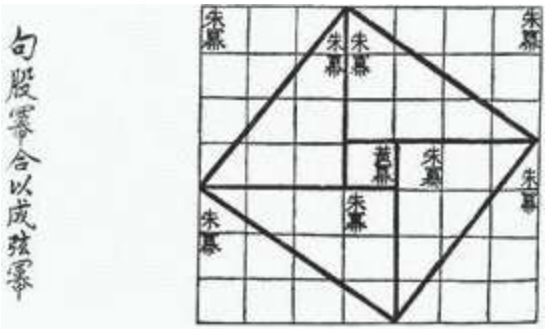


Figure 10.3 The Pythagorean Theorem illustrated in ancient China[202]

In addition, the 3rd-century mathematician Liu Hui and the 13th-century mathematician Zhu Shijie (also known as Hanquing or Songting) each provided an original geometric proof of the Pythagorean Theorem.

Other known proofs include those credited to Takakazu Seki (17th-century Japan), Jang Young Sil (15th-century Korea), Baudhayana (8th-century B.C.E. India), and Brahmagupta (7th -century India).

2.4. European and American

Last, but certainly not least, many Western mathematicians have given proofs of the Pythagorean Theorem as well. Some even use vectors and/or calculus concepts. Fibonacci (13th-century Italy) and John Wallis (a clergyman in 17th-century England) are among those who came up with their own proofs.

Furthermore, it should be noted that James A. Garfield, the 20th President of the United States, published his own proof of the Pythagorean Theorem in 1876. His proof uses a trapezoid and was published in the *New England Journal of Education*. It is said that President Garfield came up with the proof in "discussions with other members of Congress" (according to Wikipedia). Garfield and his trapezoidal proof are featured as a "mathematical treasure" in an article by the Mathematical Association of America.
https://maa.org/press/periodicals/convergence/mathematical-treasure-james-a-garfields-proof-of-the-pythagorean-theorem

Imagine, politicians discussing mathematics in Congress!

Finally, two more new proofs were announced in 2023 and widely broadcast in the media.[203] What makes these new proofs unique is that they were conceived by two U.S. high school

[202] Image is in the public domain.
[203] For instance, this story was featured in CBS's *60 Minutes* on May 5, 2024, and *Scientific American* online.

students! Calcea Johnson and Ne'Kiya Jackson, students at St. Mary's Academy in New Orleans, used trigonometry in innovative ways (in such a way as to avoid circular reasoning). This astonishing development illustrates the fact that ancient theorems are not necessarily dead but still raising interest, fascinating mathematicians, students, and many others around the globe.

Questions:

1. Do some online research on individuals from various countries who provided their proofs of the Pythagorean Theorem.
2. Which civilizations seem to have known the theorem bearing Pythagoras' name? Write a brief essay on the origin and history of the Pythagorean Theorem.
3. Study the Chinese picture above. It has 49 little squares. Explain how this picture accurately illustrates the Pythagorean Theorem for a right triangle whose leg lengths are 3 and 4.
4. Toward the end of the famous movie *The Wizard of Oz*, Scarecrow seems to quote the Pythagorean Theorem. However, if you listen carefully, what he says is not true. In fact, it is full of errors. This is a rather well-known fact, with many people pointing out the inaccuracies on the Internet. Find what Scarecrow actually said, and try to fix it so that the quote can be accurate. (Some argue that perhaps the screenwriter intentionally did this to illustrate what happens when one gets a diploma without getting a real education.)
5. Find the part of Euclid's book *The Elements* where he proves the Pythagorean Theorem. Explain the proof given by Euclid.
6. Do a little study (perhaps write a paper) on *The Elements* by Euclid. Into how many languages has it been translated? What prominent historical figures studied geometry from that book? Focus on its influence over the last 23 centuries.
7. As stated in the text, a mathematical statement becomes a *theorem* as soon as a proof is given—one proof is sufficient. Why, then, do you think there are over 300 proofs of the Pythagorean Theorem? What do you think is motivating these many people to come up with new proofs, even though they are fully aware that proofs already exist? Are they simply reinventing the wheel? Is it, as Yogi Berra said, "de ja vu all over again"?

10.2 Project 2: Musical Scales

Music Heard Around the World

Category: Mathematics and Music, Exponential Growth

Description:

While many of us are familiar and quite comfortable with traditional Western musical scales, there are many other scales used around the world. We look at these musical scales used in countries around the globe and consider their historical and cultural significance.

Introduction:

Musical scales are groups of notes (usually 12 notes per scale) that define the type/style of music being played. Each musical scale is described by listing the notes using alphabet letters. Think of a scale like an alphabet. Just as different languages use different alphabets, different types of music are based on different scales.

In traditional Western music, melody and harmony are generally grounded in a basic system of musical scales, including the two most fundamental ones—the major scale and the (natural) minor scales. However, various cultures and traditions have been using different scales that may or may not sound familiar to us, including those in our own country. What are the differences? Why do these differences exist? How do these scales influence the way music sounds in various traditions? In terms of their frequencies, measured in cps (cycles per second or Hz), what properties or characteristics do they share? Let's explore and appreciate the sound of music.

1. Scales in Traditional Western Music

1.1. Major Scale

In any given key (such as C), the major scale is a diatonic scale and has seven pitches. For example, the major scale in C consists of C, D, E, F, G, A, and B. On the piano, this particular scale (C major) has no black keys. The first of these notes in a scale, the one identifying the key, is called the tonic. A major scale is typically associated with happy, positive sound.

After the seventh note, we generally complete the scale by adding the next one, which will bring you back to the tonic (in the C major scale, add another C at the end, one octave higher than the original C). This set of 8 notes sounds "complete," and you can basically do the same starting in any of the twelve keys used in music. The interval between the first and the last of these notes (for instance, from a C to the next C, or from an A to the next A) is called an *octave*, meaning "eight." An important mathematical property is that the frequency of the sound wave associated with these notes doubles every octave. If a musical note has a frequency of 440 cps

(this is typical for the key referred to as the "middle A"), then the next A, which is one octave higher, has the frequency 880 cps, the next A the frequency 1760 cps, then 3520 cps, and so on. In other words, the frequency grows *exponentially*, doubling each octave. This is one fundamental property that governs all music.

1.2. Minor Scales

A few small changes (shall we call these "minor" changes?) made to a major scale could cause a significant difference in mood, creating a minor scale. To be exact, there are three minor scales for each key; they are referred to as natural, harmonic, and melodic minor scales. In a melodic minor scale, the notes are also different depending on whether you are going up (ascending) or down (descending). We omit the details here, but the interested reader can easily find more in any source on music theory.

Each of these three minor scales is slightly different from the other two, but one common element is that the third note (E if the scale is in the key of C) is shifted down by half step (so instead of E, the scale uses E flat), creating a rather sad, perhaps more serious, quality than the respective major scale.

It should be noted that these four scales (major, natural minor, harmonic minor, and melodic minor) are the building blocks in most Western music, including classical, popular, and jazz.

2. Musical Scales Elsewhere

Non-Western music may incorporate a wide variety of scales, many of which do not sound familiar to the Western listener. This variety reflects the rich diversity in human civilizations and traditions. Some examples of music scales used around the world are as follows.

2.1. Pentatonic Scales (East Asia)

A pentatonic scale is a musical scale that has five ("penta-") basic notes as opposed to a typical major or minor scale, which has seven (thus referred to as "heptatonic scales"). Pentatonic scales are quite common in traditional music of China, Japan, India, and other Asian countries. Pentatonic scales can be hemitonic or anhemitonic, and there are major and minor pentatonic scales. In addition, there are blues major, blues minor, suspended, and other scales within those that are described pentatonic. It is surprising to learn that, to varying degrees one can find pentatonic scales adopted in all sorts of music including bluegrass, West African, English, Tamil, Peruvian, Celtic, Nordic, Croatian, jazz, and rock.

2.2. African Scales

Various scales, such as what is known as the Nyamaropa scale, are used in African mbira music, particularly in southeastern regions of Africa. It represents the Shona culture and religion in and around Zimbabwe. Also, there is a traditional West African xylophone called the gyil, which seems to have its own scale.

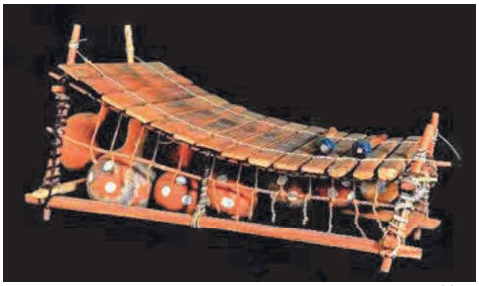


Figure 10.4 The gyil of northwestern Ghana[204]

2.3. Maqam Scales (Arabic)

The word "maqam" means "position" in Arabic, and Maqam musical scales are frequently used in Middle Eastern modal music, such as traditional Arabic, Turkish, and Persian music. They are similar to the scales used in Western music in that the scales are based on twelve-tone, equally tempered structures.

2.4. Pelog Scales (Indonesia)

Pelog scales are predominately used in Indonesian, particularly Balinese, music. While it has seven notes like in Western scales, the Pelog scale has widely different intervals and pitches. These scales are unique and exotic, perhaps characteristic of the South Pacific region.

There are many more scales known around the world, reflecting the diversity of human civilizations. These scales, as well as other elements of music, are studied widely by researchers including anthropologists, musicologists, and sociologists.

Questions:

1. Look up the interval patterns (whole and half steps between notes) found in Western major and minor scales. Describe the differences.
2. Each octave in a given scale corresponds to doubling the frequency of the sound wave. For example, if the middle A is set to 440 cps (cycles per second), the next A (an octave higher) has a frequency of 880 cps. Each half step (there are twelve of them in an octave) increases this frequency by a fixed percentage. This means the frequency grows exponentially, with a fixed ratio for each half step. Find that percentage separating each half step. [Hint] This amounts to solving the following equation:
$$(1+x)^{12} = 2$$
3. Look up one of the scales unfamiliar to you. Find the notes of that scale, play the scale, and analyze how the intervals between those notes affect the way the scale sounds. Describe how the scale sounds to you and try to explain why.
4. In traditional Western music, for any key, one can define a note called the "perfect fifth," often used in tuning instruments. For instance, in the key of C, the perfect fifth is

[204] Image by Dr clave is licensed under CC BY-SA 3.0.

G; in G, the perfect fifth is D, and so on. Find a definition of “perfect fifth” and see which non-Western scales use perfect fifths. What makes this note so special in a scale?

5. Play a simple song written in a major (Western) scale, and then try playing it using the corresponding minor scale. How would one do that? What notes are replaced with what? How does it sound? Is it possible that small changes in frequency could result in such a dramatic difference?

10.3 Project 3: Symbols and Symmetries

Symbols Seen Around the World

Category: Geometry, Mathematics and Art

Description:

We examine geometric patterns used as symbols or logos in various countries and groups worldwide, analyze their properties and symmetries, and consider their cultural significance.

Introduction:

Geometric patterns have been used as symbols, logos, and designs to represent countries, religions, dynasties, clans, families, and other historical groups around the world. Many of these patterns reflect symbolic significance based on their cultural, historical, and local background. Let's examine some examples and analyze them from geometric and cultural standpoints.

1. Examples of Symbols

1.1. Islamic Patterns

Islamic art includes a wide variety of intricate geometric patterns, many of which can be found in mosques, palaces, and other buildings, as well as in books, manuscripts, and artifacts. Complex patterns can also be used in arabesques and mosaics, which are often used to decorate walls, floors, and ceilings. Take a look at some examples below. Many mosaics utilize various types of star patterns, referred to as girih tiles.

Figure 10.5 Islamic geometry patterns[205]

[205] [Left] image by Patrickringgenberg, licensed under CC BY 3.0; [Middle] "pattern" by distrakt, licensed under CC BY-NC-SA 2.0; [Right] "Image Plate from Owen Jones' 1853 classic, 'The Grammar of Ornament'." by EricGjerde is licensed under CC BY-NC 2.0.

Many of these patterns are highly symmetric, making them quite interesting from the standpoint of transformational geometry and group theory. We will discuss the idea of symmetry later.

1.2. Celtic Knots

Knot theory is one of the newer and fascinating fields of mathematics, and it has proved (perhaps surprisingly) applicable in many scientific fields. While there are precise mathematical definitions for terms such as knots, links, and braids, for our purposes here, simply consider typical interlacing formed by tying a piece (or pieces) of string or rope.

Throughout history and around the world, there have been symbols involving knots of various kinds, but intricate knots and knotted patterns are particularly common in Celtic art. Known as "Celtic knots," these patterns are frequently used in everyday objects such as jewelry as well as in religious monuments. Some say these knots symbolize continuity, the interconnectedness of our lives, and perhaps eternity.

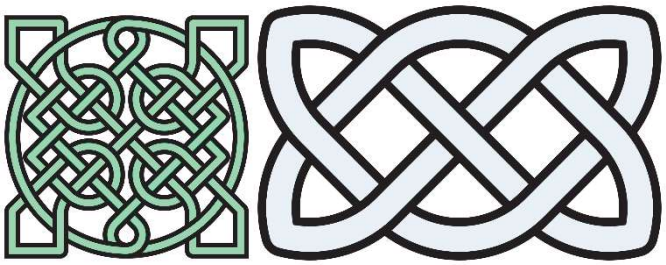


Figure 10.6 Celtic knots[206]

1.3. Native American Tribal Symbols

It is quite common throughout history for various tribes, groups, and dynasties to have their own symbols. These are seen in flags, pottery, petroglyphs, and rock carvings. All types of geometric shapes including circles, squares, triangles, and other polygons are used to represent their faith, tribal identity, and/or spiritual heritage. Native American tribes are no exception. Their visual representation involves cultural or sacred meanings in their diverse indigenous cultures throughout North America. Many interpretations may exist for these geometric patterns. Some other tribal symbols are less geometric and may include animals, bird feathers, planetary objects, and plants.

[206] Image is in the public domain. Image is in the public domain.

1.4. Mandala in Hinduism and Buddhism

A mandala, a Sanskrit-based word, is a geometric pattern or symbol representing spiritual traditions of Eastern religions such as Hinduism (also of Buddhism, Shinto, and Jainism). Polygons, perhaps symbolizing lotus flowers, are often used in these patterns.

Figure 10.7 Chenrezig sand mandala[207]

1.5. Yin and Yang

Maybe one of the most well-known symbols in the world, "Yin and Yang" (also referred to as "Yinyang"), represents a deeply philosophical concept of darkness and light (negative 陰 and positive 陽) interconnected and mutually interacting with each other. Originating in China, Yin and Yang is said to symbolize the dualistic worldview of Taoism and possibly the harmony of the universe, where the opposite forces exist in balance while connecting everything in the world. There have been many theories on this popular symbol, some even finding deep or hidden meanings.

Figure 10.8 Yin and Yang[208]

[207] Image by Colonel Warden is licensed under CC BY-SA 3.0. This is a mandala created to celebrate the Dalai Lama's visit to the U.K. in 2008.
[208] Image is in the public domain.

1.6. Spanish Mudejar Architecture

Mudejar is a style of art, decoration, and architecture seen predominately in medieval Spain with Muslim influence. The word "Mudejar" also describes the Islam followers living in Iberia in the late medieval period around the 14th and 15th centuries. While the style is based on Islamic art, many buildings built or decorated with Mudejar art may have been the work of Catholic builders in Spain.

The Mudejar style combines many features of Gothic, Renaissance, and Romanesque architecture with intricate geometric figures, which are frequently seen in Islamic buildings constructed prior to the period. Not surprisingly, girih tiles were used for decoration, and Islamic-style arabesques were incorporated.

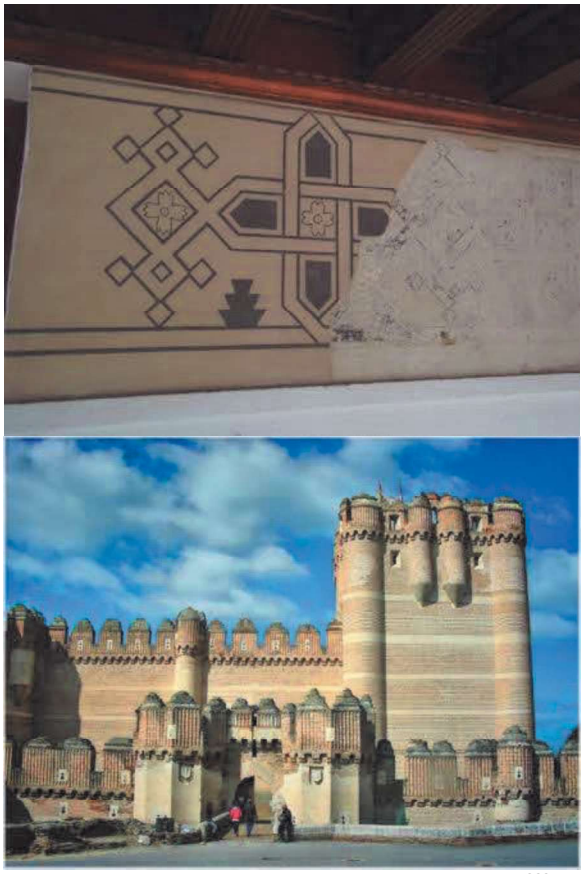


Figure 10.9 Mudejar-style architecture and design.[209]

[209] [Left] "2628-Castillo mudejar de Coca (Segovia)" by jl.cernadas is licensed under CC BY 2.0; [Right] "Palacio de Mondragon - Plaza Mondragón, Ronda - Mudejar patio - Moorish art" by ell brown is licensed under CC BY-SA 2.0.

1.7. African Adinkra Symbols

The Adinkra symbols were used in West Africa, particularly Ghana, in everyday items such as pottery, fabrics, and decorative pieces as well as in architecture. They represent ideas, concepts, specific thoughts, and beliefs. Some geometric patterns may have stood for certain philosophy or cultural wisdom. You may see some of these geometric symbols on t-shirts and other souvenirs from that region of Africa. These symbols are often based on geometric shapes like squares, circles, polygons, and combinations thereof. Sometimes many of these symbols can be placed together on one piece of fabric or cloth, such as the one shown below called the "mourning cloth." Perhaps the idea of using fabrics to reflect their way of life is similar to the tradition of American quilts, which played a major role in representing the life, culture, and history of the early American people.

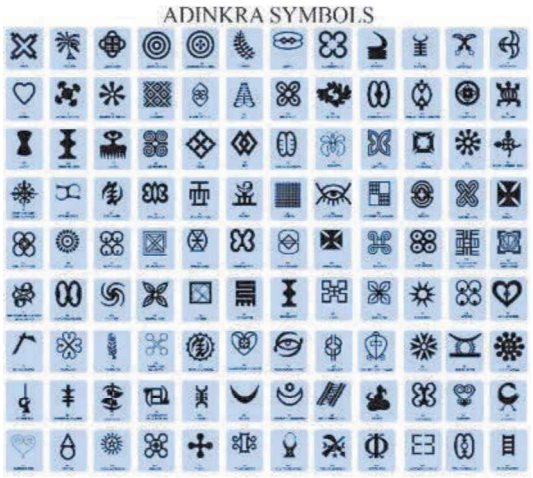


Figure 10.10 Adinkra symbols[210]

1.8. Chinese Lattice Windows

Geometric patterns are prevalent in decorating walls, floors, and doors all over the world. Chinese lattice windows provide a rich source of such geometric applications, and one can find quite striking mathematical patterns. There have been mathematical papers written about these geometric patterns seen, not only in windows, but also on silk fabrics, ceramic ware, furniture, and buildings. Chinese lattices are typically made of wood and utilize not just vertical and horizontal lines but squares, hexagons, octagons, and much more intricate shapes as well.

[210] Image by Bosomba Amosah is licensed under CC BY-SA 4.0.

Figure 10.11 Chinese lattice patterns.[211]

1.9. Aztec and Mayan Glyphs

The Aztecs, the Mayans, and other Mesoamerican peoples are known to have used many geometric symbols conveying religious and historical meanings. These symbols may have been significant in communicating seasons (various times of the year) similar to what our calendars do. In fact, it is believed that some (maybe many) of these patterns have some astronomical meaning.

Many of their symbols are highly symmetric figures based on polygons, particularly quadrilaterals (squares, rhombuses, etc.). Today, some of these patterns are used in jewelry and pottery.

It is clear from these examples that there is something universal and powerful about groups of people using geometric and mathematical patterns to symbolize and represent their identities throughout history. Humans have long tried to express abstract thoughts, concepts, and belief systems with concrete and tangible symbols, which in turn provide a rich tapestry (sometimes literally) of human history and culture.

2. Symmetries

2.1. Symmetry in Mathematics

In mathematics, an object has symmetry if there is an operation or transformation that maps the object onto itself. These transformations include reflections, rotations, and translations (in formal geometry, these are called *isometries*). Symmetry, therefore, has to do with how the pieces of the object are arranged in a proportional and balanced way.

[211] Image, showing some Chinese patterns used for window lattices and screens, is by Jukkyoutobukkyou and is under the Creative Commons CC0 1.0 Universal Public Domain Dedication.

In everyday life, when we say something is "symmetric," we normally mean that two sides (right and left, top and bottom, etc.) are identical so that the sides match up perfectly when folded along some line (referred to as the "line of symmetry"). This is a specific type of geometric symmetry called "reflective symmetry." For instance, capital letters like A, H, T, and W are symmetric with a vertical line of symmetry; capital letters like B and E have a horizontal line of symmetry.

Figure 10.12 Mercedes-Benz Logo.

The transformation can also be rotational, i.e., turning by some angle, often 180 degrees. When an object is symmetric in this way on paper, we could flip the paper upside down, and the object would still look the same. For instance, capital letters like H, I, N, O, S, and Z have this property. But rotational symmetry does not have to be by 180 degrees. A good example is the logo for Mercedes-Benz, which is symmetric with a 120-degree (or 240-degree) rotation about the center.

2.2. Symmetry Groups

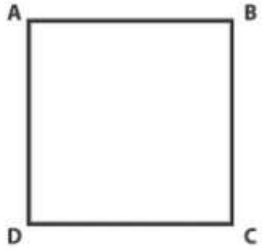


Figure 10.13 A square with four labeled corners (A, B, C, D).

Now, consider the square, and label the four vertices (corners) A, B, C, and D. How many reflections are there? What is meant by this question is this: how many lines of symmetry exist for the square so that the square would look exactly the same if the square is flipped along that line? The answer is four: two diagonal lines (AC, BD) as well as the two lines going through the midpoints of opposite sides of the square. So we can say that the square has four reflections.

Now, how can you rotate the square so that the result looks unchanged? Clearly, the square can be rotated by 90 degrees, and the square would look exactly the same. You can do it again, and again. But rotating this way four times would result in the same square, where the vertices are back at the original positions (as if you did nothing after all). This "no motion" is called the identity transformation. Here we see that doing a 90-degree rotation four times is equivalent to

the identity transformation. Similarly, flipping (reflecting) twice using the same line of symmetry is equivalent to the identity transformation.

So far, we see that the square has four reflections and four rotations. How many other symmetries (transformations) are there? It turns out that this is it. There are exactly eight symmetries for the square. With some specific examples, you can see that any reflection followed by another reflection (using a different line of symmetry) is equivalent to one of the four rotations.

This set of eight symmetries (or eight transformations, including the identity transformation) forms what mathematicians call a "group" (more specifically, a *dihedral group*). For the square, the dihedral group has eight elements. You can actually verify that there are only eight symmetries without knowing a lot of math! Vertex A can be placed at any of these four places after the symmetry (transformation), and for each position for this vertex A, you can leave the square as is or you can flip (reflect) it along the diagonal AC. Note $4 \times 2 = 8$.

If the vertices are read from the top left corner in the clockwise fashion, the eight transformations can be described as follows (the first four are rotations whereas the last four are the reflections along AC following the respective rotations):

$$ABCD, BCDA, CDAB, DABC;\ ADCB, DCBA, CBAD, BADC$$

Dihedral groups are simple and nice examples of finite groups, a specific type of algebraic structure. You may be a bit surprised to discover that these groups have practical applications in fields such as chemistry and computer security.

2.3. Symmetries in Symbols

Now, take a look again at some examples of symbols used in various cultures, religions, and countries. What types of symmetries do you see? How many? Why are so many of them symmetric? What types of symmetry groups do you see?

It seems clear that symmetry, a fundamental idea in geometry (in fact, in all of mathematics), plays a major role in expressing the beauty of shapes and patterns. Human beings have been using symmetry all around the globe throughout history. It may be a mathematical concept, but it is certainly an aesthetic factor in art, architecture, and much more.

Questions:

1. Find some more examples of geometric patterns and/or symbols used by certain organizations, religions, or peoples. What kinds of symmetries do they have?
2. Find and briefly describe, using your own words, the following symbols with religious/spiritual significance. Then, analyze them geometrically.
    a. Flower of Life
    b. Seed of Life
    c. Fruit of Life (How is this related to the Flower of Life?)
    d. Star of David (perhaps one of the most recognized symbols of a nation)
3. Look up Metatron's Cube. How many circles are there? How many triangles do you see? Do you see a hexagon (6-sided polygon)? As the name suggests, do you see a cube?
4. Find at least three interpretations of Yin and Yang. Do you think any of them is based on or inspired by mathematical ideas such as symmetries and numbers?
5. The Mercedes-Benz logo was mentioned earlier. The Mitsubishi logo is somewhat similar, and both are based on the equilateral (regular) triangle. They all share the same symmetry group. How many symmetries do you think each of these symbols has? (Find out what the name "Mitsubishi" roughly means in Japanese. Their logo should make sense.)
6. Let's generalize the previous question. It is already stated that the square has eight symmetries (four rotations and four reflections). What about the regular pentagon? The regular hexagon? How many elements are there in the symmetry group for the regular n-gon?
7. What about the symmetries for the circle? What types of symmetries are there and how many?
8. Do an internet search on other examples of symmetric symbols, possibly from your own culture/heritage. What shapes and symmetries are common in that group of symbols? Why do you think that is?
9. Look up national flags around the world. List some with symmetries. Are they rotational and/or reflective symmetries?
10. As you walk around your campus, town, neighborhood, and shops, look for various symbols and signs; consider their symmetries.

10.4 Project 4: Currencies and Denominations

Money Printed Around the World

Category: Arithmetic, Discrete Mathematics

Description:

We compare and contrast the denominations (face values) of bills (paper currencies) issued by various countries around the world. Further, we look at some exchange rates, past and present, and see how they affect international commerce and tourism.

Introduction:

Many different currencies have been used around the world, and they change from time to time. In addition, international travel and trade require conversion between two or more currencies, with ever-changing exchange rates, making the conversion rather complicated. Each currency has certain bills (paper money, more officially referred to as "banknotes") as well as coins. Some currencies have more denominations than others. Further, shapes and sizes of coins vary from currency to currency.

1. Exchange Rates and Value of a Currency

1.1. An Example: USD vs. MXN

Exchange rates fluctuate for many economic, commercial, geopolitical, social, and other reasons. In April 2020, one U.S. dollar(US$1) was roughly equivalent to 25 Mexican pesos (MX$25). Only eight months later, in December of the same year, US$1 was about MX$20. Did the value of the Mexican peso go up or down between April and December of 2020? At first, one might be tempted to think the value/power of the peso went *down* from MX$25 to MX$20, but it turns out the peso had gained power, becoming more valuable against the U.S. dollar. In other words, the U.S. dollar weakened against the peso in that eight-month period.

The following may be a simple way to think about this: In April 2020, to buy a product worth 100 pesos, one would have paid US$4. But in December, the same product would have cost US$5, meaning that more dollars were required to buy the same thing in December (compared to April). This suggests that the dollar has lost its value against the Mexican peso during the eight-month period.

Conversely, one might consider this as follows: if you were in Mexico in April 2020, and decided to send money to an American friend, say US$100, then you would have had to send MX$2500; however, eight months later, to send the same dollar amount, you only had to send just

MX$2000. Again, this suggests that the peso had become stronger while the U.S. dollar became weaker relative to the Mexican currency.

1.2. Implications

In the above scenario, we saw that the U.S. dollar became "weaker" relative to the Mexican peso from April to December 2020. Is this good news for Mexico and bad news for the U.S.? Well, that depends.

If you are an American traveling to Mexico, this decline in the value of the U.S. dollar is, indeed, disadvantageous. Imagine having to spend MX$2500 for a hotel room. As US$1 went from MX$25 to MX$20, the price of that room changed from US$100 to US$125. So everything would have cost you more, specifically 25% more, in December 2020 compared to eight months prior.

Similarly, if you are a manufacturer, farmer, or merchant in Mexico, and you have a product priced at MX$2500, your U.S. customer would have paid US$100 for it in April and US$125 in December. This may sound attractive to you at first, but there may be another producer in another country that can provide the same product at a lower price. If you have a large quantity of this product, you may lose U.S. customers. If this happens at a large national scale, the change could even impact the whole landscape of international trade.

Here is another example. Back in the fall of 1982, US$1 was hovering around 250 Japanese yen (JPY250). However, US$1 was just below JPY80 some 30 years later (in 2011) and JPY157 at the beginning of 2025. This means that a 1,000-yen bowl of ramen in Tokyo would have cost US$4 in 1982, more than $12.50 in 2011, and $6.37 in 2025. Now consider hotel rooms, guided tours, transportation costs, and miscellaneous expenses associated with traveling. You can see why exchange rates can dramatically affect tourism. On a larger scale, a 2.5-million-yen sedan in Japan would have cost US$10,000 in 1982 but more than $31,250 in 2011 and less than $16,000 in 2025. This type of volatility, not just between JPY and USD but other currencies as well, significantly affected the global auto industry, resulting in many manufacturers relocating to the U.S. (and elsewhere), merging, and/or creating international venture companies. If you add other factors like transportation, tariffs, and labor costs, the issue gets even more complicated.

2. Denominations of Paper Money

Many foreign visitors to the United States are surprised that Americans still carry and use $1 bills; many foreign countries have already stopped issuing paper money for such a small value. Below, as you look at these denominations in which various countries issue their bills (banknotes), consider what combinations of these would be required to come up with some specific amounts of money; for example, $600 (USD) would require at least six bills while €600 (euros) would require only two bills (500 + 100). Also, speculate how each country decides which denominations to issue their currencies in.

[Currency Denominations]

United States: USD	1, 2, 5, 10, 20, 50, 100($2 notes are less common.)
Most E.U. nations: EUR	5, 10, 20, 50, 100, 200, 500
Canada: CAD	5, 10, 20, 50, 100
Australia: AUD	5, 10, 20, 50, 100
United Kingdom: GBP	5, 10, 20, 50
Switzerland: CHF	10, 20, 50, 100, 200, 1000
Japan: JPY	1000, 2000, 5000, 10000 (JP2000 notes are rare.)
China: CNY	1, 5, 10, 20, 50, 100
India: INR	2, 5, 10, 20, 50, 100, 200, 500, 2000
South Korea: KRW	1000, 5000, 10000, 50000
Russia: RUB	50, 100, 200, 500, 1000, 2000

Questions:

1. Why do you think the United States still issues $1 bills? Is there a debate or movement to eliminate it? Find another currency, if any, that has a bill worth US$1 or less.
2. You may find it surprising that the biggest denomination in the U.K. is the 50-pound note. Why do you think they would not issue a 100-pound note? Do some research online.
3. Looking at this list, Japan and South Korea seem to have very large numbers as their denominations (e.g., 10,000-yen notes in Japan, 50,000-won notes in South Korea). Why is that?
4. List a few factors considered when a country decides on the denominations of their banknotes.
5. Now, consider the number of denominations issued in each currency. What is the average number of denominations? List some pros and cons of having many denominations.
6. Study the history of the euro as a currency. Before the euro, what currencies did various E.U. countries use? In particular, what did Italy use, and what was their exchange rate with other major European currencies? What are some advantages of using a shared currency? Are there any disadvantages? Summarize your observations.
7. Why do you think European countries like the U.K. and Switzerland chose not to adopt the euro as their currency? Do some online research to find some answers. What are some ramifications? Did their decision increase or decrease their consumer prices (e.g., are things more or less expensive in Switzerland and the U.K. compared to the rest of Europe)?

10.5 Project 5: The Language of Numbers

Numbers Counted Around the World

Category: Arithmetic, Number Theory, Ethnomathematics

Description:

We examine various numeration systems, numerals, and number-related words/terms used around the world and observe the diversity in counting.

Introduction:

Let's learn how to count. 1, 2, 3, … Sounds simple enough? But how do people in foreign countries count? In their language, are there some characteristics not shared in English counting? Mankind has used many different ways to count numbers, write numbers, and name numbers throughout history and around the globe. Here, we will look at three aspects of counting: numeration systems (base), numerals (written characters), and the language of numbers.

1. Numeration Systems

A numeration system is a mathematical structure used in counting and expressing numbers, primarily in written form, utilizing characters or symbols representing quantities. These symbols are called "digits." Each numeration system assumes a fixed base number (also called the "radix"), such as 10 in the decimal system, 2 in the binary system, and 16 in the hexadecimal system. Because of the way place values work in a given numeration system, the base also turns out to be the number of distinct digits required in that numeration system. For example, in the decimal system (base 10), there are ten digits (0 through 9); combining these nine digits in a systematic way, we describe each number in the decimal system—the "decimal representation" of the number. In contrast, in the binary system, only two digits (0 and 1) exist, and each number is represented with a string consisting of 0's and 1's (the "binary representation").

When we reach 9 in the decimal system, we denote the next number as "10," using two digits, 1 and 0, in this particular order. Similarly, when we write "64," we mean six 10s plus four 1s. Because any number bigger than 9 is carried over to the next position, all we need is 9 digits in the decimal system. Many numeration systems use this notion of "positional notation" (place value).

But if one uses a numeration system with a different base, then, of course, the number of digits required would be different. You may wonder, "Why would anyone do that?" But one could also ask, "Why is it that we prefer 10 so much as the base?"

1.1. Base 10

The base-10 numeration system, called the decimal system ("deci" means 'tenth'), is by far the most prominent system used throughout the world. The nine numerals we use in this system (0 through 9) are called the Hindu-Arabic numerals even though they are almost universally used today (exceptions include, ironically, some Arabic-speaking countries). Most agree that the decimal system has become prevalent because of the number of digits we have on two hands, making it easy to calculate using base 10. However, this is not the only numeration system used by mankind.

1.2. Other Civilizations in History

There is evidence that the Babylonians used a numeration system with base 60; this is similar to the way we describe time, with 1 minute equal to 60 seconds and 1 hour equal to 60 minutes. But we are not certain if they actually had 60 symbols as numerals.

The Mayans seem to have used a base-20 numeration system. Their civilization had symbols that represent 1, 20, 400 (20^2), and 8000 (20^3). We will explore this particular base later.

The Yuki people, a Native American indigenous people in Northern California, used the Yukian languages which seem to suggest that they used a base-5 numeration system. Some historical research has also shown that the ancient Sumerians, living in the Mesopotamia region about 6000 years ago, a ternary system (base 3), evident from their three numerals in cuneiform writing.

In ancient China, people used to have counting tools (like rods) made for a base-12 numeration system. This corresponds to a "dozen (12)" and a "gross (144, which is 12^2), terms we frequently use even now. Curiously enough, today we use a similar system when referring to time using A.M. and P.M.

1.3. Bases with Specialized Purposes

Theoretically, any whole number bigger than 1 can be used as the base of a numeration system. However, two bases in particular have played a very important role in the advancement of computers. Because electrical signals can be either ON (1) or OFF (0), and nothing else, it makes sense to represent data using sequences of 0's and 1's, i.e., in binary code. These two numerals used in the binary system are called "binary digits," shortened to "bits." These two symbols are the building blocks of all digital data. In other words, all data, including text, images, videos, and digital sound, is stored in binary form, and all machine instructions are fundamentally written in binary code. Naturally, then, the binary system has become an extremely important tool in the information age.

Of course, having only two numerals has its disadvantages. For example, 5 would be expressed as the 3-digit number “101” in binary (that would definitely confuse drivers in the Los Angeles area). If the answer is 42 (for some random question), the binary system would express it as “101010.” It would be nice not to have to have so many digits but still maintain some “binary” taste. This led to the application of the hexadecimal (base-16) system, in which each binary string of four symbols (0 through 15) would be represented by one single numeral. These sixteen numerals used in the hexadecimal system are 0 through 9 (as in the decimal system) as well as letters A, B, C, D, E, and F.

2. Numerals

As explained above, numerals are the symbols used to represent quantities in writing. Today, most people in the world use the Hindu-Arabic numerals. However, this has not been the case in world history. In fact, there are some notable exceptions even in the present world.

2.1. Roman Numerals

The most famous example may be the Roman numerals, where I, V, X, L, and C represent 1, 5, 10, 50, and 100, respectively. D is 500, and M is 1000. These are still used today for such purposes as page numbering in a small section of a book and listing points on outlines.

2.2. Alphabet Numerals

In many civilizations and societies, the alphabet letters were directly used as numerals. For example, in the Greek civilization (at least during a certain time period), Alpha represented 1, Beta 2, Gamma 3, and so forth while Rho expressed 100, Sigma 200, and Omega 800. (Then they ran out of alphabet letters.)

The Jewish people used their Hebrew letters as numerals. Aleph represented 1, Bet 2, Gimel 3, and so on. This was common when they referred to dates in the Hebrew calendar.

2.3. Chinese Numerals

In Chinese, the language has special symbols (characters) to represent numbers. 1, 2, and 3 are expressed with characters consisting of one, two, and three horizontal bars, respectively. Numbers larger than 3 have more complicated symbols. Just as Roman numerals, these are not necessarily digits since there is also one symbol for 10 (which looks like a cross), another for 100, another for 1000, and so on. Some symbols for large numbers include the following:

万: 10,000 (ten thousand)
億: 100,000,000 (hundred million)
兆: 1,000,000,000,000 (one trillion)

These characters are also used in nearby countries and regions, such as Japan, Korea, Taiwan, Hong Kong, and even Vietnam.

Interestingly, because the characters for one, two, three, and ten are quite simple, some official documents in countries like Japan require that these simple symbols be replaced by more complicated characters in order to avoid potential tampering and modification.

3. Language of Numbers

We often use number words in English without thinking very much, but there are a few interesting features that may not be true in other languages. Similarly, when you analyze how numbers are countered and expressed in other languages, you may notice some interesting, perhaps strange, things.

3.1. Numbers in English

One, two, three, … not very interesting, right? Well, when you get to ten, you go on… to eleven and twelve. But then, for some reason, the form changes significantly at that point. The next few numbers all end with "-teen," as in thirteen, fourteen, etc. If 16 is "sixteen" and 17 is "seventeen," why isn't 11 "oneteen" or 12 "twoteen"? It suggests that English may have been influenced by some base-12 numeration system. After all, we do have the word "dozen" to mean 12 and "gross" to mean 12^2.

3.2. Numbers in French

French, like Italian, Spanish, and Portuguese, is a Romance language, rooted in Latin. As such, French shares many features common to these other languages. Some of the properties listed below may also be true in these other Romance languages.

In French, ten is "dix" and twenty is "vingt," a single word meaning twenty, not two tens. (If you order "Venti" at your local coffee shop, you may know it has 20 ounces since "venti" in Italian means 20.) Thirty is "trente" (three tens) while forty is "quarante" (four tens), not surprisingly. Fifty and sixty are similar.

But then comes a surprise twist. Seventy is "soixante-dix," literally meaning "sixty (plus) ten." Eighty is "quatre-vingts," meaning "four twenties," and ninety is "quatre-vingt-dix," meaning (as you might have guessed it), "four twenties (plus) ten."

One hundred, in French, is quite ordinary: "cent," literally meaning "one hundred." Duh.
It is interesting to note that forty is not expressed as "two twenties" nor sixty as "three twenties." But as soon as you exceed sixty, you go to a system resembling a base-20 numeration system (but stops before you reach a hundred). Many scholars think this structure is based on some historical counting system, in which numbers from 70 to 99 have some units based on 20.

So, for example, if you want to say 87 in French, you would say, "quatre vingts sept," meaning "four twenties and seven." As strange as this may sound, this is not completely foreign to American expressions. You may recall the famous words of President Lincoln, who began his famous Gettysburg Address with these words: "Four score and seven years ago."

3.3. Numbers in Other Languages

The Danish language has an interesting numbering system as well. Fifty in Danish is "halvtreds," meaning "two and a half." This is because fifty is two and a half times twenty, or $2.5 \times 20 = 50$. Again, as in Romance languages, the number 20 seems to be a significant unit in their numbering system.

In Russian, ten is "desyat'," twenty is "dvadtsat'," and thirty is "Tridtsat'." Notice the similarity. But then, when you get to forty, the word is "sorok," completely different from the smaller multiples of ten. Perhaps in Russian, the number 20 is not as significant as in Romance languages.

These examples show that historical, linguistic, and geographical factors have influenced how people count numbers and express quantities.

Questions:

1. Review the concept of "place value" in any numeration system. For instance, in the decimal system, the number 365 means three 100s, six 10s, and five 1s added together. In other words,
$$365 = 3 \times 10^2 + 6 \times 10^1 + 5 \times 10^0$$
(note $10^0 = 1$). Now, using the same principle, verify that the binary number 101 is actually 5. Then, find out what number is written as "10111" in binary. (In mathematics, we often indicate the base using a subscript. So the question here is this: what number is 10111_2?)
2. Find the sum of 7, 5, and 4. Next, multiply 13 by 6 to find the product, showing your work on paper. Now, show both of these calculations using Roman numerals. What makes this so difficult?
3. Suppose you worked for 3:15 (3 hours 15 min.) on Monday, 4:30 on Tuesday, and 2:40 on Wednesday. How many hours did you work during these three days? Clearly, time is not expressed in base 10 (since 1 hour is 60 minutes). Is there a way to convert these durations into some decimal system and calculate your hours?
4. As seen above, the Chinese language has symbols (characters) to represent some large numbers. In English, a thousand times a thousand is a million. A million times a thousand is a billion, and a billion times a thousand is a trillion. It seems that a new word is used when the previous one is multiplied by a thousand (i.e. when three zeros are added at the end). What do you notice about the Chinese symbols? What do you think the next symbol expresses (after one trillion)? What about the one after that? Do some online research and check your hypothesis.
5. You saw the strange terms used for 70, 80, and 90 in French. Find what these numbers are called (and what each literally means) in Spanish and Italian. Compare and contrast them with the French phrases.

6. Unlike most European languages, some Middle Eastern languages go from right to left. In those languages, are numbers also read/written from right to left? If not, why not? Check, for instance, Hebrew and Arabic.
7. Look up numerals in the Hebrew language (e.g., www.hebrew4christians.com/Grammar/Unit_Eight/Letters_as_Numbers/letters_as_numbers.html). Try to figure out some patterns, keeping in mind that this language is read from the right to the left. What happens with multiples of 10? What do you see in common with numerals for 19, 29, 39, 49, etc.? Describe some patterns you observe in your own words.
8. Do some etymological research to find out why in English we say "eleven" and "twelve" instead of "—teen" like the numbers that follow.
9. In every language, you will find words that represent very large numbers, such as trillion and quintillion. What is the largest number with a word in English? What are the largest numbers expressible with a single word in other languages? Compare them. Are there good reasons why certain languages have words for much larger numbers than other languages?

10.6 Project 6: License Plates

Identification Used Around the World

Category: Probability, Counting, Discrete Mathematics

Description:

We compare various formats used in vehicle license plates in different countries and look at some restrictions that would reduce the possible number of license plates available for use in a given format.

Introduction:

Most countries around the world require automobiles and other vehicles (including motorcycles and boats) to be registered with license plates, and the format of these license plates varies significantly from country to country. Even within the United States, the formats are quite different depending on which state issues them. While it is theoretically possible to count the number of different license plates (using the Multiplication Rule), in reality some strings of letters and numbers are never used for various reasons. We examine a variety of formats used around the world and then analyze certain restrictions imposed on them to calculate how many license plates are possible in a given format.

1. Typical Formats

1.1. License Plates in the United States

In the U.S., license plate formats vary from state to state, and even within a state, the format changes from time to time, often to address the rising number of vehicles. Many states today use three alphabet letters followed by numerals. For instance, except for personalized and special license plates, the format "ABC 1234" (three letters followed by four digits) is used in multiple states including New York. This is also the current format in Texas, but for over three decades, the Lone Star state used plates consisting of six alpha-numeric characters (the format they stopped using due to the rapid population growth in Texas). The format generally used in Illinois is "AB 12345."

The standard format in California is "1ABC234" while Hawaii uses the format "ABC 123." Note that, if exceptions and restrictions are ignored, there can be up to $26^3 \times 10^3 = 17{,}576{,}000$ distinct license plates that can be made with the Hawaii format. Certainly the island state, with 1.5 million residents, does not need over 17 million different plates. However, the state of California, with a population of around 40 million, would certainly want to have more than 17 million (thus the first digit in the California license format).

1.2. License Plates in European Nations

In France and in Italy, the standard format used is "AB-123CD" (two letters, three digits, and then two more letters). In the United Kingdom, the format "AB12 CDE" is used, with the first two letters indicating the area, the two digits for "age identifier" (for the vehicle, not the driver!), and three (almost) randomly chosen letters. (Why this choice is not completely random is explained below). In Spain, the current format is "1234 BCD," starting with four digits. An interesting feature in Spain is that no vowels (except for "Y") are used in the last three letters (thus "BCD" in my notation). This is to avoid a real word, which could be awkward or offensive. (More on restrictions below.) This "only consonants" rule has been frequently used in the U.S. as well.

1.3. License Plates in Other Countries

One may wonder what a license plate may look like in countries that do not use the European alphabet. It turns out that many nations still use the standard (Hindu-Arabic) numerals along with some characters in their alphabet. For instance, in Saudi Arabia, a general license plate has three (Arabic) letters followed by four (standard) numerals. But the Arabic letters are also translated—or, more precisely, transliterated—into the corresponding Western letters while the numerals are also written in their Arabic notation. This addition is to help those visitors who may not be familiar with the Arabic letters or numerals.

There are a few problems, though. Arabic, as many Semitic languages, is read from right to left. Each Western alphabet letter is displayed right below its corresponding Arabic character, so if one reads the Western alphabet letters from left to right (as one usually does), then the order is reversed from the original Arabic. In addition, the translation is not perfectly accurate (and cannot be) because the Arabic alphabet consists of only 17 letters, not exactly matching one-to-one with Western alphabet letters.

In China, the standard license plate starts with a Chinese character (indicating the province), an alphabet letter, and five (Hindu-Arabic) digits although these five could now include non-numerals. Military and other special vehicles have their own designation and formats. Macau and Hong Kong (both special administrative regions of China), as well as Taiwan, have their own license plate format, different from those in mainland China.

The Japanese license plate format is quite different in that it has two lines. The first line starts with the municipality (prefecture, city, etc.) in Japanese, followed by a few numerals. Then, the second line begins with a small Japanese character, followed by the format "12-34" (four numerals with a hyphen in the middle). Typically, therefore, no Western alphabet letters are used on a Japanese license plate.

2. Restrictions

As mentioned above, the format determines the number of distinct license plates that can be issued. For instance, with the format "ABC123," up to $26^3 \times 10^3 = 17{,}576{,}000$ number plates (strings, i.e., sequences of alphanumeric characters) are possible, as seen earlier. Theoretically. However, not every string is desirable or appropriate, so the governing authority often has regulations to systematically exclude certain strings. What types of strings are excluded? Here are some examples.

2.1. Offensive Language

Any offensive or inappropriate language is most likely excluded from the license plates, even (especially) with personalized plates. In addition, terms that suggest obscene meanings, discriminatory connotations, or inflammatory, radical views are likely avoided. (I am not giving an explicit list of examples here, but you can probably imagine a few examples.) In fact, many agencies even prohibit certain terms that may not be offensive (or even meaningful) in the language used in that particular country but that are inappropriate in another language. Expressions promoting illegal activities may also be banned from license places.

2.2. Trademark Infringement

Registered trademarks and certain copyrighted terms (strings) are generally excluded as well for legal reasons. These may include company names, band names, song titles, and phrases used in well-known ads and products.

2.3. Confusing Characters

The alphabet letter "O" and the numeral "0" are quite similar; therefore, a jurisdiction may decide not to use one of them to avoid potential confusion. Similar things can be said about the letter "I" and the numeral "1." It is obvious that society will be well-served if all license plates are easily readable, so confusing symbols are generally excluded. For similar reasons, excessively long strings and hard-to-read fonts are typically avoided.

2.4. Other Restrictions

Some agencies may also restrict the use of excessively religious or political words and expressions. Others may reserve special characters or symbols to be used to identify some specific vehicles (such as emergency vehicles, large trucks, motorcycles, and government automobiles). They would be excluded from general use.

These principles apply to other identifier strings, such as telephone numbers and driver license numbers. For example, numbers like 000-0000 usually remain unused. Phone numbers do not begin with 911 since these three digits are reserved for emergencies.

Questions:

1. Choose a country or state not mentioned in the text here, and find their standard license plate format. Then, calculate the possible number of license plates that can be issued with that format (ignoring any restrictions).
2. In many places around the world, phone numbers (without the area code) consist of seven digits, like "555-1212." However, some have run out of numbers in this format. Do some research and find a few cities in the world where eight-digit numbers are now in use.
3. The format "ABC-1234" is quite common and is used in states like New York. How many different license plates (strings) can be issued in this format if no restrictions exist?
4. What if the five vowels (A, E, I, O, and U) are excluded from the above format? How many strings are possible then?
5. Again, using the format "ABC-1234," how many possible license plates can be issued if the five vowels are excluded AND the digits cannot be repeated (all four numerals must be different)?
6. Choose a very small country in the world. Find out how many people live there, how many automobiles are there, and what format is used for their license plates.
7. If you were in charge of writing rules on license plates, what other restrictions would you consider, if any? Is the list of strings to be excluded as described in the text above sufficient? Anything else that needs to be banned?
8. Instead of the format "ABC-1234" (without any restrictions), what would happen to the number of possible license plates if the format were "ABC-DEFG" (seven alphabet letters)? What would be some advantages and disadvantages of this format?
9. In the United States, generally a license plate is installed on the vehicle using two screws. In many other countries, at least one of those screws is specially designed, and it can be tightened only by an official government staff member. Why is this done? Is this necessary?

10.7 Project 7: Units of Measurement

Measurement Units Adopted Around the World

Category: Arithmetic, Conversions

Description:

We look at various units of measurement used in history and around the world for different quantities.

Introduction:

Measurement is all around us. It is safe to say that one of the most important roles numbers play in our everyday life is to express measured quantities—such as length, weight, volume, temperature, area, and speed, to name a few. Precise measurement is certainly important in science and commerce, but many numerical standards also ensure our safety and help us maintain quality of life.

In the United States, we mostly use what is referred to as the Imperial or U.S. Customary System (USCS), which includes units like inches, feet, gallons, pounds, and degrees Fahrenheit. If you grew up in another country, you may have felt that these units are rather weird and difficult to get familiar with. Most of the world, even the U.K. (where the "Imperial" system originated), now uses the metric system, which is said to be much simpler due to its decimal-based nature.

But throughout history and around the world, many different units have been used to measure a variety of quantities including weights of commercial goods, dimensions of buildings, and volumes of liquids. Let's examine some of those as well as a few confusing and/or interesting facts about units of measurement.

1. Units of Measurement Commonly Used Today

1.1. Metric System

The term "metrication" refers to the process of converting to and officially adopting the metric system of measurement. Virtually every country except the United States has implemented this process and thus uses the metric units today (a few other exceptions include Liberia and Myanmar).

In this system, the basic units are the meter (length), the gram (mass, weight), and the liter (volume, capacity). Each of these can take a variety of prefixes since the metric system is a decimal (base 10) system. For example, for length, one can use the kilometer (1000 m), the

centimeter (0.01 m), the millimeter (0.001 m), and so on. The introduction of nanotechnology has also ushered in smaller units such as the micrometer (one millionth of a meter), the nanometer (one billionth of a meter), and the picometer (one trillionth of a meter). Similarly, one can easily define units like the kilogram, the milligram, the kiloliter, and the milliliter. Because of this "decimal" property, many people find the metric units quite versatile and practical.

1.2. Imperial System and the USCU

The Imperial System has been used primarily in the United Kingdom and other English-speaking countries. In 1824, the British government made this system an official one for the country by enacting the British Weights and Measures Act. While there are some minor differences between this British Imperial System and the United States Customary Units (USCU), both of these systems use units such as inch, foot, yard, mile (length), ounce, pound (weight), fluid ounce, cup, quart, and gallon (volume, capacity).

Some notable differences between the U.S. system and the U.K. Imperial System include specific units of volume (e.g., gallon, fluid ounce, and bushel). But the confusion does not stop there. The U.S. ton (sometimes referred to as the "short ton") is 2,000 pounds whereas the Imperial ton (the "long ton") is 2,240 pounds. Oh, by the way, the metric ton (or "tonne") is defined to be precisely 1,000 kilograms (i.e., 1 mega gram), which is about 2,204 pounds, slightly less than the Imperial ton. Yes, when it comes to systems of measurement, there is a ton of difference.

1.3. Other Systems and Units Being Used

As local groups, tribes, and peoples all needed to communicate quantities, make comparisons, and trade goods, it is not at all surprising that various units of measurement were used throughout the world. Standardization became a necessity only when the society became more global, trade more international, and travel more popular. Every country and region, therefore, seems to have a lot of (tons of?) units that have been used throughout history.

A popular one was the Chinese system (due to the vastness of its territory, history, and population), which included units of length like Zhang, Chi, Cun, Fen, and Li; units of weight like Jin and Liang; and units of area like Mu. These have influenced units used in neighboring countries such as Japan and Korea as well.

2. Notable Units of Measurement in History

2.1. In the Ancient World

Ancient literature gives us a glimpse of many measurement units used in various civilizations and nations. One of the most well-known units may be the cubit, used to measure length. A cubit is basically the length from one's elbow to the tip of their finger, approximately 18 inches. Though it likely originated in Egypt, the cubit was a popular unit over a large region in the ancient world including Mesopotamia and Palestine. In fact, the Old Testament uses this linear unit extensively to describe items and buildings in ancient Israel.

The Greeks and the Romans, having conquered vast territories, utilized highly sophisticated measuring units. A well-known example was the stadion (plural: stadia), a unit of length. One stadium was 600 Greek feet (podes), representing around 150 to 200 meters. Herodotus, for example, describes the battle of Marathon (490 B.C.E.) using stadia to refer to distances.

Notice that, like the cubit above, many ancient units referenced parts of the human body, thus resulting in some variety of interpretations due to personal differences.

2.2. In More Recent History

Because measurement is an essential part of different aspects of human activity, many systems of measurement have been used. The league, for instance, is a unit of length used in Medieval Europe to express distances between villages. The fathom was used to describe water, but not its volume; it appears that the fathom was used strictly to represent the depth of water, about 6 feet deep. This unit was also used in the United States. When the river water marks 2 fathoms (about 12 feet) deep on a vertical scale in the water, steamboat workers would refer to the depth as "twain" (an Old English word meaning "two"), thus yelling "Mark Twain!"

3. Oddities and Facts about Units of Measurement

We already mentioned some oddities involving subtle differences between the British and the American systems. Here are some other examples.

The area is a crucial quantity in describing the size of a community. For this reason, one can make a long list of units measuring area in different countries. The Americans are quite familiar with the acre, which originally had to do with the area where a yoke of oxen can plow in a day. This is the same as the British Imperial acre but is to be distinguished from the Irish and Scottish acres, which are slightly larger. Perhaps some oxen worked harder than others.

One fascinating topic may be to do an international comparison of portion sizes of various foods and drinks sold, such as different cup sizes offered by Starbucks (or other global coffee stores) worldwide. In the U.S., Starbucks coffee typically comes in Tall, Grande, or Venti. But there are other sizes, particularly in foreign countries. See the Starbucks question below for more details.

Questions:

1. Study the history of the metric system. When was it developed? By what country or countries? When did it become globally popular? Write a brief essay answering these questions.
2. Do some research on how a meter (the principal unit of length in the metric system) was originally defined and how it is officially defined today. What is the reason for the official definition? What do you think about the way it is defined?

3. Who is Celsius? Who was Fahrenheit? Who was Kelvin? What countries did these men come from? How and why did they develop the temperature scales named after them? Also, consider some pros and cons of each scale in measuring temperature.
4. Why do you think the United States still primarily uses the USCS and not the metric system? Why do you think the British people have largely adopted the metric system?
5. Study common units of measurement used in culinary arts and cooking. See if those units have the same meaning in the U.K. and other countries. What are some differences?
6. Choose a country or region (now or in the past—you may pick your native country), and make a list of common units of measurement used there. Are there other units used historically? How were those units described before their precise definitions? Did they refer to aspects of the human body (such as the cubit) or farming (such as the acre)?
7. In the United States, typical Starbucks sizes are Tall, Grande, and Venti. How many milliliters are in these sizes? You may not be aware that Venti may be different depending on whether the drink is hot or cold. There is also a larger size than Venti. Do some research to find out what sizes are offered by Starbucks in different countries. Why do you think there are differences?

10.8 Project 8: Word Problems

Word Problems Written Around the World

Category: Geometry, History of Mathematics

Description:

We compare and contrast the types of application problems and word problems in mathematics (mostly in arithmetic) in various countries of the world and examine some unique and interesting characteristics.

Introduction:

"Word problems." These two words trigger fear and anxiety, right? You may recall those scary scenarios such as "A train leaves Los Angeles at…" or "Johnny has an uncle 3 years older than twice his age…" Some are really "application" problems while others are just dubious, outrageous, or superfluous questions. You may wonder who comes up with these crazy stories, completely artificial situations, or ridiculous riddles.

It turns out, however, that word problems are quite common in the world of arithmetic education. In fact, you would probably agree that most real-life questions are, indeed, "word problems." Therefore, it is perhaps necessary and appropriate that a mathematics curriculum include various word problems so that the student can learn how to translate those "words" into mathematical relationships, apply arithmetic or algebraic skills, and verbalize the answer using the correct language.

But what types of word problems are common in various parts of the world? Are there certain features, perhaps national characteristics? What is deemed important in one culture may or may not be relevant in another. Let us take a look at the "world of word problems" and examine similarities and differences.

1. Familiar Examples in the United States

It is not hard to list examples of word problems that are common and familiar in the United States. All one has to do is open any arithmetic or algebra book and browse through it.

1.1. Money Problems

Naturally, word problems involving money are quite common. It may be one of the most important applications of arithmetic in our lives as most of us make financial decisions every day, ranging from shopping at the store and buying automobile fuel to working to earn wages. Most word problems in the U.S. involve dollars and cents. Word problems involving money can

appear at any grade level and may require simple arithmetic operations (such as markups and discounts, percentages, profits and losses), positive and negative numbers, probability (expected value of a lottery), exponential and logarithmic equations (compound interest), geometric series (annuity and loan formulas), and even calculus (marginal costs/profits).

1.2. Sports Problems

Billions of people around the world enjoy sports, and they usually involve points, data, and statistics. It is then natural that many word problems are related to sports. In the U.S., common sports that appear in word problems include "football" (in the U.S., this means American football, not the other type more popular in the world), basketball, baseball, hockey, golf, and soccer (the other "football"). Again, these can range from simple arithmetic to probability, statistics, geometry, algebra, calculus, and even vector analysis. Many questions use measurement numbers in feet, inches, and pounds.

1.3. Transportation Problems

Another common type of word problems has to do with transportation: trains, airplanes, bicycles, and cars, for instance. It is no surprise that many American word problems involve automobiles. It is common to see two cars going in opposite or same direction. Interesting problems related to the Mean Value Theorem or the Intermediate Value Theorems can include a car traveling at varying speeds. Fuel efficiency can be used to teach ratios and proportions (as well as unit analysis); speed and distance traveled can illustrate the concept of Riemann sums and definite integrals.

1.4. Geometry Problems

There are, of course, many other types, but the last one we specifically list here has to do with geometry. Because shapes, sizes, areas, and volumes make up such an integral aspect of our lives, it makes sense to teach arithmetic using these concepts. Geometry questions often involve formulas and computations, such as perimeter, area, and volume problems. Angles and symmetries are also common topics. We also see practical questions such as using congruent or similar triangles to estimate the height of a tree or the length of a field.

2. Examples from Around the World

Now, let's compare and contrast application problems around the world and throughout history.

2.1. Application Problems in Ancient Civilizations

The Greek philosophers and mathematicians (Pythagoras, Aristotle, Archimedes, etc.) almost solely focused their attention on the pure aspects of mathematics, e.g. proofs, with very little regard for applications. The Egyptians and the Romans, on the other hand, seemed quite application-oriented, with a view to real applications such as constructing pyramids, designing bridges, and defeating enemies in battles. This is the difference between the "glory that was Greece" and the "grandeur that was Rome." In Egypt, where the Nile often flooded, it was

important to figure out where to cultivate land and do agriculture. Their necessity to "measure (meter) the earth (geo)" is the origin of our term "geometry."

It is well-known that the Babylonians used a numeration system based on the number 60, i.e. the sexagesimal system. If they had a classroom textbook, it is entirely possible that they had computation word problems in base 60. The Mayans are known to have performed sophisticated calculations involving their calendar as well as astronomy.

2.2. Application Problems Around the World

Mathematics is a universal tool with which one can measure quantities, express durations, communicate sizes, and trade goods and services. Hence, the mathematics curricula of different countries seem to reflect their own culture and tradition. Below are some examples.

2.2.1. Time/Schedule Problems

Some international questions involve time zones. Others involve summer daylight time. For instance,

- The Times Square Ball just dropped in New York to bring in a new year. What time is it in Sydney?
- You are sharing a live firework show at 9 pm on July 4th in Chicago on a social media platform with your friend living in Singapore. What time (and what day) is your friend watching it there?

In European and Asian countries, where public transportation systems have timetables (and where trains and busses generally run on time), it is common to see arithmetic questions on time and schedules. Here are some examples:

- If you take an 8-hour train at 12:10 am at Berlin Central Station to Mannheim, where you wait for an hour and take another train, taking you to Paris in 3 hours, what time do you get to Paris? (You may also want to check if the cities are in the same time zone.)
- If a 2-hour flight leaves Lisbon Portela Airport (LIS) in Portugal at 1 pm for Rome's Leonardo da Vinci-Fiumicino Airport (FCO), what time do you arrive at Rome? Remember that the two cities are in different time zones.

2.2.2. Clock Problems

It may sound similar to the "time/schedule problems," but the "clock problems" (*tokei zan*) in Japan involve angles in analogue clocks, providing unique challenges.

- How many times between noon and 6 pm do the two hands of an analog clock overlap? How many times do they make exactly 90 degrees?
- Find out exactly when (what time) between 3 and 4 pm the two hands of the clock overlap.

2.2.3. Currency Problems

Each of the following questions on bills and coins (currencies) requires the knowledge of what bills are printed and what coins are minted in the respective country or territory.

- You need to pay 85 euros at a store, and you want to use only bills (banknotes), no cards, no change. How many combinations of bills are possible for this exact amount? (Assume that, in the country where this question is asked, it is common knowledge that the values of euro notes less than 100 euros are €5, €10, €20, and €50.
- Currently, there are five types of rupee (singular "Re" and plural "Rs") coins: Re 1, Rs 2, Rs 5, Rs 10, and Rs 20. What is the least number of rupee coins required for 72 rupees? How would the answer change if there were a 50-rupee coin (Rs 50)?

2.2.4. Agriculture Problems

Everyone eats. So every country grows food. It makes perfect sense, then, that farming is a topic of many arithmetic questions in any given country. These often utilize their own units of measurement.

- Zhou is a farmer in China, with a square field 70 mu by 70 mu. If he can produce 400 kg of wheat per square mu, how much wheat can Zhou produce?
- An African farmer has a 100-hectare farm to raise two crops: maize and cassava. The yield of maize per hectare is 300 tons while that of cassava is 150 tons per hectare. If the farmer produced 25,500 tons combined, how many hectares were for maize and how many were for cassava?

In short, one can get some insight into the culture and traditions of a country by looking at the type of word problems that appear in their mathematics textbooks.

Questions:

1. The Times Square Ball just dropped in New York to bring in a new year. What time is it in Sydney? What do you need to know first to solve this problem? Do you need to consider summer daylight time? Why or why not? (Remember where Sydney is.)
2. Above, you saw this question: "You are sharing a live firework show at 9 pm on July 4th in Chicago on a social media platform with your friend living in Singapore. What time (and what day) is your friend watching it there?" Answer this question first. Then, create a couple of more questions similar to this.
3. How many times between noon and 6 pm do the two hands of an analogue clock overlap? How would you explain your thought process to someone?
4. How many times do the two hands of an analog clock make exactly 90 degrees between noon and 6 pm? This is a slightly more difficult question than the previous one. Why is that?
5. Find out exactly when (what time) between 3 and 4 pm the two hands of the clock overlap. This may seem to be a very challenging question. To do this question, you probably should first consider the angle by which each hand moves per minute. Then,

note that this question is similar to one in which one train is catching up with (and overtaking) a slower one running in the same direction.

6. Choose a country and look up in what denominations the country's bank notes (bills) are printed. Then, write a few problems similar to the currency problems listed above.
7. Pick a country or a region of the world you are interested in, or perhaps your native country. Find out what types of word problems appear in various mathematics textbooks in K-12 (or whatever the education system the country has). Compare and contrast those with the types seen in math textbooks in the United States.

10.9 Project 9: Numerology

Numerical Myths Believed Around the World

Category: Number Theory, Set Theory, Ethnomathematics

Description:

We examine what numbers are as well as the mystic or spiritual significance of various numbers as believed around the world.

Introduction:

First, a disclaimer. This is not a mathematical topic, and it does not involve serious mathematical content. In fact, the very fact that numerology does not involve much mathematics is a reflection of the truth that mathematics is really not about numbers—it is more about abstract relations, functions, patterns, and logical connections.

Of course, it is definitely possible to consider numbers as mathematical objects. For millennia, the concept of numbers was so obvious that most people, including mathematicians, did not even think about defining them. I mean, we all know what "three" means, right? Or do we?

When Georg Cantor pioneered the revolutionary field of set theory in the 1870s (yes, less than two centuries ago), the entire foundation and landscape of mathematics were transformed, and since then pretty much everything that can be defined in mathematics has been defined in terms of sets. We will not go into all the details of the foundations of mathematics, but set theory led to precise definitions of concepts such as sets of natural numbers, integers, rational and irrational numbers, real numbers, and complex numbers.

But what is a number?

In formal set theory, one begins with the most fundamental set of all—the empty set. That's right. A set with no elements. It may surprise you to learn that the number zero "0" is defined as the empty set, denoted $\emptyset$ or $\{\ \}$. The number one "1" is defined as the *successor* of 0, which is $\{\emptyset\}$ (this is different from the empty set since this set has one element, the empty set). The number two "2" is the *successor* of 1, which is $\{\emptyset, \{\emptyset\}\}$. This process can go on to define every natural number, yes, infinitely many of them. This can get quite confusing fast and is well beyond the scope of this text (even though we are just talking about numbers 1, 2, 3, …).
In other words, in the beginning was the empty set, and the entire set of natural numbers is defined using the empty set. (And this is not pointless.)

As you can see, it is already quite hard to formally define familiar numbers. But once the set of these "natural numbers" is defined, it is not hard to extend it to the set of integers and rational numbers (fractions with any integer on the top and any non-zero integer on the bottom).

However, it took mathematicians several millennia to figure out exactly how to define the set of "real numbers" (rational and irrational numbers). This is quite surprising, given that Pythagoras and his students were aware of the existence of irrational numbers (like $\sqrt{2}$) around 2500 years ago!

However, long before mathematicians attempted to give formal definitions, numbers had been used in everyday life to measure, count, compare, buy, and sell things. Also, people all over the world attached some spiritual or superstitious meanings to certain numbers for certain reasons. This is what numerology is about. But it goes beyond just abstract beliefs. As you will see below, people all over the world choose wedding dates, name their children, and make major decisions, all based on numbers.

Let's take a look at some numerology practiced around the globe.

1. Lucky & Unlucky Numbers

Thousands of couples got married all over America on Saturday, July 7, 2007. They are known as "Lucky Sevens" couples because their wedding date was 07/07/07. Indeed, The Knot, a well-known wedding website, had about 38,000 couples registered for that date. Why? Well, the number 7 is considered lucky in most Western nations. However, this is not the case everywhere. For instance, the number 8 is the lucky number in Chinese culture, perhaps because of the shape of its character (八) for that particular number.

The number 13, in contrast, is considered unlucky in the West while no such interpretation exists in other parts of the world. Instead, because the pronunciation of the Chinese character for the number 4 is quite similar to how to pronounce the word for death, the number 4 is often considered unlucky and to be avoided. Consequently, some buildings in Asia have no floor numbered 4 just as some Western buildings have no 13^{th} floor.

In some cultures, children's names are chosen based on various lucky numbers. For instance, the parents may consider the numerical values of letters in a name, the number of characters, and/or the number of strokes required to write that name. What is lucky or unlucky may even be related to the child's birthday.

Numerology.com states that the "most important and influential number in your personal Numerology is your Life Path number," which is associated with one's date of birth. Some people also use what is known as "angel numbers," which are supposed to guide them through life. They may use these numbers not just to buy their lottery tickets but also to make major decisions in life, such as whom to marry, what career to pursue, and where to live.

In Feng Shui, which originated in ancient China but has been gaining popularity internationally, some numbers are believed to bring more fortune, harmony, and peace than others. Tarot cards, believed to have originated in northern Italy around the 15th century, are also used along with numerological interpretations in divination and fortune-telling.

2. Meanings Associated with Numbers

Some people maintain that various numbers have inherent meanings or associations. Examples include the number 1 associated with independence and individuality (makes sense), 2 with balance or harmony (also makes sense), 4 with stability, 11 with spirituality, and so on. Perhaps the Hebrew alphabet provides a particularly interesting example. In the Hebrew alphabet, every letter has a numerical value. For instance, the first letter of the alphabet, א (Aleph, which is also used in mathematics to represent various transfinite numbers), has the numerical value 1. The second letter, ב, the value 2, and so on. This continues for the first ten letters. Then the eleventh letter represents 20, the twelfth 30, etc. Other letters represent larger numbers.

The book *The Bible Code* is probably the most well-known example of how these Hebrew letters and their numerical values are applied (or misapplied) to find hidden messages. Written by an American journalist Michael Drosnin and published by Simon & Schuster in 1997, this book claims to have discovered many encoded messages in the Hebrew text of the Old Testament. Drosnin based this book on a method used in the 1994 academic paper (published in *Statistical Science*) "Equidistant Letter Sequences in the Book of Genesis," authored by Eliyahu Rips, Doron Witztum, and Yoav Rosenberg.

Now, Eliyahu Rips was already a famous, award-winning mathematician known for his work in geometric group theory. However, due to the questionable methods used in the paper (obviously to prove what he and the two other authors were looking for), he faced significant criticism, including from the mathematical community. Later, a group of mathematicians including Brendan McKay (Australia) and Dror Bar-Natan (Israel) showed that the original paper by Rips, Witztum, and Rosenberg was "fatally defective," debunking the whole method by showing that similar messages could be found from any lengthy text such as *Moby Dick* or *War and Peace.*

3. Why Numbers?

A good question to ponder is this: why are numbers used by people worldwide in this manner—for divination, spiritual guidance, and fortune-telling? What is the appeal of numerology in various cultures and societies? The following reasons (and many more) have been suggested. Numbers are universal. Even though the term for the number "four" (or the numeral "4") may look different in different countries and in different historical periods, this concept of "four" represents a universal attribute, such as the answer to the question "How many legs do most dogs or cats have?" Numbers transcend national boundaries as well as linguistic barriers. They can, therefore, serve as a common tool for communication around the world. This fact may contribute to certain divination practices crossing cultural boundaries and gaining universal

appeals. Further, that which is universal, bigger than one's specific community, region, or country, can easily be considered something greater, residing in the realm of the divine, something sacred.

Also, numbers are well-structured and can be operated on arithmetically (added or multiplied, for instance). They can then create a wide variety of patterns and sequences. Certain patterns are much more unlikely to appear than others, probabilistically speaking. Many numerological methods use patterns and sequences, so numbers are easy tools that can be generated, applied, and interpreted.

In addition, numbers are (obviously) used to count things. Whether you are counting the number of days or strokes in writing one's name, it is easy to associate numbers to given circumstances.

Arithmetic operations are also objective in the sense that there is only one answer when two given numbers are added or multiplied. For example, what is the sum of your date of birth (written in a certain fixed form) and the number of letters in your full name? There is one unique answer. This may lead one to feel that the results of divination are more trustworthy since the practice involves a mathematical, bias-free calculation and the answer is not up to the practitioner.

Questions:

1. Pick a country, a culture, or a language, and find various superstitions involving numbers. Are some numbers considered more lucky/fortunate than others? Are some numbers consistently avoided? Are they based on some religious or historical reasons or how those numbers are pronounced? Try to find out why.
2. Do some research on the book *The Bible Code* (and its sequels), along with the paper by Rips, Witztum, and Rosenberg referred to above. What did the journal *Statical Science* do with this article as well as with the later paper criticizing it (written by McKay, et.al.)? Despite the inaccuracies, why did the book become so popular (even leading to a movie deal)?
3. Read about "Life Path" numbers and other numbers used commonly in numerological practices. Are there any merits to them? Why do you think so many people use these?
4. You have seen that the Hebrew alphabet letters are associated with numbers. Are there other alphabets with similar numerical associations? If so, are they used in numerological interpretations?
5. The beginning of this section describes the birth of set theory, a very foundational field of mathematics. Do some research on set theory as well as its pioneers such as Georg Cantor, Richard Dedekind, Bertrand Russell, and Bernard Bolzano.
6. What is a real number? Find some formal and informal definitions.

10.10 Project 10: Health Indices

Health Indices Checked Around the World

Category: Statistics, Health

Description:

We explore and compare various index numbers related to health and longevity as seen and used in different countries of the world.

Introduction:

Health is one of the biggest concerns worldwide, and there are many index numbers related to health, such as blood pressure, cholesterol levels, body mass index (BMI), and weight. But what is acceptable (for healthy individuals) in each of these categories may differ from country to country. Aren't we all humans with basically the same body type? Then why the difference? If your cholesterol level, for instance, is considered dangerously high in one country, should it not be considered the same way in other countries? What are the bases of such differences if any? Here are some interesting things to consider as we see international differences in the guidelines related to various health indices.

1. Blood Pressure

High blood pressure is sometimes called the "silent killer" because it can affect your heart and other organs without any explicit symptoms. It is a major risk factor for a variety of diseases such as stroke, heart failure, and kidney disorder. When blood pressure is checked, one gets two numbers: systolic pressure (the higher value, the pressure when blood is being pumped into your arteries) and diastolic pressure (between heartbeats). The systolic (top) number is mostly what healthcare professionals are concerned with since that is the value that increases as one gets older (and the arteries get more plaque). However, both numbers are relevant, and frequently they are denoted like a fraction: systolic/diastolic, e.g. "120/80" (in mm Hg, which stands for 'millimeters of mercury').

According to the American College of Cardiology and the American Heart Association, you are

- "normal" if your blood pressure is lower than 120/80 (both numbers)
- "elevated blood pressure" if the top is between 120 and 129 while the bottom is not above 80
- "Stage 1 hypertension" if the top is between 130 and 139 while the bottom is between 80 and 89
- "Stage 2 hypertension" if the top is 140 or higher while the bottom is 90 or higher.

(per 2017 Guideline for High Blood Pressure in Adults - American College of Cardiology (acc.org)).
Anything above 180/120 is considered an emergency.

Now, if you fly across the Atlantic, you will discover that the European standards are somewhat different. Their definitions are as follows:

- "Hypertension" if the numbers are 140/90 or higher (this group is further subdivided into 3 grades based on their numbers)
- "Optimal" if the numbers are less than 120/80
- "Normal" if the top is between 120 and 129 while the bottom is between 80 and 85
- "High Normal" if the top is between 130 and 139 while the bottom is between 85 and 89.

Other countries have different threshold numbers. As in the European standards, values 140/90 or greater are considered "hypertension" in most Asian countries. However, in some Asian countries, the cutoff point may be 130/80 for certain classes of patients while 150/80 for other groups. In Singapore, they distinguish one's blood pressure numbers measured at home from those measured in the office. Of course, these numbers are subject to change, based on new research and studies.

So, for example, if your blood pressure is 138/88, you will be considered "high normal" in Europe but in "Stage 1 hypertension" in the U.S. Are you in danger? Should you seek preventive measures from your doctor? Well, it looks like that decision may depend on where you live in the world.

2. Body Mass Index (BMI)

Body mass index, BMI, is an estimate of the amount of body fat one has. It is calculated based on one's height and weight. More specifically, it is your body mass (weight in kg) divided by the square of your height (in meters). Note that this formula does not directly measure your body fat and completely ignores other factors such as your muscle density and body composition. This may be one of the reasons why BMI is almost always used in conjunction with other health indicators. Nevertheless, BMI is still considered a helpful initial index to look at in terms of weight-related issues.

Quick online research may suggest that a "normal" range of BMI is between 18.5 and 24.9. If your BMI is less than 18.5, you may be considered underweight; if your BMI is above 24.9, you may be considered overweight.

For instance, suppose you are 5"10' tall and weigh 170 pounds. In metric, you are 177.8 cm tall (1.778 meters) weighing 77.11 kg. Thus, your BMI would be

$$\frac{77.11}{1.778^{\,2}} = 24.39,$$

which is within the normal range.

All right. Now, let's go overseas. Are there different standards for BMI in different countries as we saw with blood pressure earlier? It turns out that, perhaps not surprisingly, there are. Remember, BMI is used primarily to estimate the risk of weight-related health issues such as type-2 diabetes, heart disease, and hypertension. So various nations and regions, having studied correlations in their data, establish their own ranges and threshold numbers. For instance, in Asia these health risks are higher for groups of patients with lower BMI numbers than their Western counterparts. So some Asian countries have adjusted their BMI classification based on their observational studies.

As you may expect, many countries define "obesity" using BMI. But (as you may expect again) the threshold number differs from country to country. Consequently, the minimum BMI value to qualify as being "obese" depends on the country. For example, for the U.S., the number is 30 while it is 28 in China and 25 in Japan.

There are, however, worldwide definitions, the official standards established by the World Health Organization (WHO). For adults, WHO defines "overweight" as a BMI value of at least 25 and "obese" as a BMI of at least 30. The Centers for Disease Control and Prevention (CDC) in the U.S. uses the same number, 30, to define obesity, but it also subdivides it into Class 1 (30 to 35), Class 2 (35 to 40), and Class 3 (40 or higher). There are different standards for children and some other subpopulations as well.

3. Cholesterol Levels

When someone asks you what your cholesterol level is, you may want to ask for clarification since there are several different types, including total cholesterol, LDL (low-density lipoprotein), HDL (high-density lipoprotein), and triglycerides. LDL is the one often referred to as the "bad cholesterol," which can accumulate in artery walls to form plaques. Everyone seems to agree that high LDL level is considered a high-risk factor for coronary artery disease and stroke, but what level is really dangerous? At what point should you seek preventive treatment? Well, these questions depend on your gender, age, other health-related risk factors, and (of course) what country you live in.

There is no shortage of online advice. Some websites claim that (for adults) the total cholesterol level should be less than 200 (mg/dL) while other sites tell you that the total number does not matter since a high HDL ("good cholesterol") number, which is good, could raise your total cholesterol value.

Researchers and organizations around the globe have studied all of these numbers and established some standards, or cut-off values, for classifying the risk levels. Again, not surprisingly, we can see some differences, not only in these cut-off values but also in which cholesterol number to base them on. (See Cholesterol Guideline Differences in Europe and the US - American College of Cardiology (acc.org) for more.)

For example, a medical article from South Africa (South African Dyslipidaemia Guideline Consensus Statement | Klug | South African Medical Journal (samj.org.za)), instead of using total cholesterol or LDL, focuses on "LDL-C," which is obtained by an equation involving total cholesterol, HDL, and Triglycerides.

The 2018 Guideline by the American College of Cardiology (Guidelines-Made-Simple-Tool-2018-Cholesterol.pdf (acc.org)) lists a rather complicated set of answers which depend on various factors such as the person's age, health history, and whether the person is a smoker or not. For example, for healthy adults, what is considered "good" (as opposed to "borderline" or "high") is

- Less than 200 mg/dL in total cholesterol
- 60 mg/dL or higher in HDL (acceptable: 40+ for males & 50+ for females)
- Less than 100 mg/dL in LDL (70 for high-risk group)
- Less than 149 (ideal: less than 100) for Triglycerides

The European Society of Cardiology surprised many in 2019 when it set 55 mg/dL as an LDL target value for a very high-risk group of patients (European cholesterol guidelines push LDL targets below 55 mg/dL | MDedge Cardiology). An article published by American Heart Association News Ethnicity a 'risk-enhancing' factor under new cholesterol guidelines | American Heart Association states that race and ethnic background make a significant difference and recommends a "more personalized approach to risk assessment, diagnosis and treatment." According to the article, for example, South Asians in the U.S. have lower HDL numbers than White people, and "those from Bangladesh, India, Nepal, Pakistan and Sri Lanka" have higher risks for heart disease. Those of Japanese descent are "more likely to be sensitive to statin drugs used to treat cholesterol and may require lower dosing." The article also notes that since the term "Hispanic" is very broad, "even the country of origin might matter" among Hispanic patients.

As you can see, when it comes to health, data has to be very carefully examined. There is no single simple standard for any of these—whether we are talking about blood pressure, BMI, cholesterol levels, or other indices. Global perspectives are crucial and play a very significant role when we discuss health, wellness, and medical treatment.

Questions:

1. Do some online research and see what various countries consider levels of "high blood pressure," "overweight" (or "obese"), and "high cholesterol." What differences, if any, do you notice?
2. When medical organizations (such as the American Heart Association) determine how to classify patients into their risk groups like "normal" and "high risk," how do they decide the cut-off numbers? There are many papers referenced in this section, so take a look at one or two and see the process they go through to determine or update these standards. Write a short paragraph summarizing the process.

3. Measure your blood pressure. Are you in a high-risk group? What other factors do you need to consider before answering this question? Also, does your blood pressure reading change depending on the time of the day or where you measure it?
4. Based on your height and weight, calculate your BMI. What group are you classified into? Does that depend on your ethnicity?
5. Study the history of BMI. Who came up with the idea? Weight seems to be a reasonable factor to consider, but why square the height? The unit used is kg/m^2, but does that make sense? Write a brief summary of what you find concerning BMI and its applications.
6. In addition to total cholesterol (TL), HDL, LDL, and Triglycerides, what other cholesterol-based indices have been defined and used? Is there one index number that is better or more important than others? Is there one that incorporates all the information? (If you are familiar with baseball, compare these indices with statistics like BA, OBP, SLG, OPS, OPS+, and WAR. Which is the best index when evaluating a player?)

10.11 Project 11: Research on π

A Ratio Computed Around the World

Category: History of Mathematics, Computer Science, Numerical Analysis, Geometry

Description:

We examine the history of the irrational number π as well as recent and ongoing research on the number.

Introduction:

There are very few numbers as mysterious and mind-boggling as π, defined as the ratio of the circumference of a circle to its diameter. It has a long history, and even to this day there are people fascinated with it, diligently computing more and more of its decimal digits. Most people know it is a little bigger than 3; many perhaps use 3.14 as a good approximation. However, until a few centuries ago, people did not realize that it is irrational, i.e. it cannot be expressed as a ratio of two integers (thus it cannot be a terminating or repeating decimal). Let us look at its history, universal appeal, and ongoing research by those who find this number exciting.

1. History of π

1.1. Rough Estimates

The definition of π would be meaningless if different circles had different ratios of the circumference to the diameter. So someone must have discovered that the ratio is a constant, regardless of how big the circle may be. It is difficult to identify exactly who proved that this ratio is the same for every circle, but the Greek mathematician Euclid (around 300 B.C.E.) made some implicit references to the constant nature of this ratio in his classic textbook *The Elements*.

The Old Testament passage of I Kings 7:23 describes a circular "molten sea," whose diameter ("from brim to brim") was "ten cubits" and whose circumference was "thirty cubits." This, of course, implies a ratio of 3, not a very accurate estimate for π. However, it may be that they measured this basin, a circular object, using a measuring rod, a short stick. In that case, as shown below, this ratio makes perfect sense.

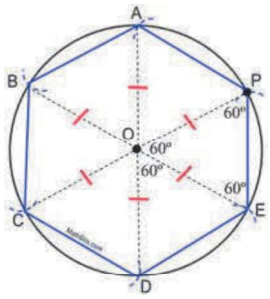


Figure 10.14 A hexagon with six labeled corners (A, B, C, D, E, P).

The point is that you take this rod whose length is OA, the radius of the circle. Say it is five units ("cubits" if you prefer). Then the diameter ("from brim to brim") would be ten units long. What then is the circumference? Well, if you also use this same rod to measure the circumference, since the rod would not bend, you would be measuring the six edges AP, PE, ED, DE, CB, and then BA. Since each of these six triangles (OAP, for instance) is an equilateral triangle, each of these edges would be exactly as long as the radius, i.e., five units. The total length of all six sides would then add up to 30 units as claimed.

If you think this is too inaccurate, how else would *you* try to figure out the answer? The exact value of that ratio was a question many ancient civilizations seemed to be interested in. Of course, they had no way of computing this ratio theoretically as we do today. All they could do was to measure the circumference and the diameter, subject to human errors. The Babylonians, around 2000 B.C.E., came up with 3.125 (25/8) as an approximate value of π. The Egyptians, according to the Rhind papyrus, may have used 256/81 (which is 3.16049…). Not bad. But not good enough.

1.2. More Precise Estimates by Archimedes

Then came a breakthrough. Of course, it came from Greece.

Archimedes (c. 287 to 212 B.C.E.), undoubtedly one of the greatest mathematicians in history, used his "method of exhaustion" to calculate the area of a circle as well as the volume and surface area of a sphere (all involving the value of π) and determined that the value must be between 223/71 (which is 3.1408…) and 22/7 (which is 3.142857…). What distinguished Archimedes from any of his predecessors is the fact that, given his method, one could theoretically find the value of π to any desired accuracy (though the calculations would get notoriously messy). In other words, he discovered an algorithm to find more and more digits of π.

Various methods were used in other civilizations, including Chinese and Arabic, to calculate the value of π more precisely (Liu Hui and Zu Chongzhi, for instance). Later, various mathematicians in India, including Aryabhata (5^{th} century), Bhaskara II (12^{th} century), and Nilakantha (15^{th} century), contributed to the study of π, and so did Islamic astronomers and scholars such as Al-Khwarizmi (9^{th} century) and Al-Kashi (14^{th} century). Yes, it was a global topic of investigation.

1.3. Computations since the Renaissance

With the Renaissance, the "re-birth" of the Greco-Roman culture, came more advancements in science and mathematics. Leonardo da Vinci may be more well-known as an artist, but he also showed his genius in science, astronomy, and mathematics including studies of π. In fact, some of his famous paintings have references to π and circles. Kepler and other astronomers also applied π in their computations of planetary motion.

Ludolph van Ceulen, a Dutch mathematician, spent many years, decades, of his life calculating π. In 1596 he published a book in which he introduced the first 20 digits of π. He later improved on his own work, finding the first 35 digits of π, which are displayed on his tombstone in Leiden, Holland. The initial tombstone was unfortunately lost; however, to recognize his achievement, it was restored in 2000 as the "Ludolph van Ceulen Memorial." If you are wondering, yes, this monument is circular.

During the last few centuries, various new ideas in calculus, geometry, real and complex analysis, algebra, and other fields of mathematics have helped us understand and appreciate the significance of π further.

2. Nature and Properties of π

A number that can be expressed as a fraction of two integers is called "rational." All rational numbers can be expressed as terminating or repeating decimal numbers, such as 0.5 (terminating) for ½ and 0.333… (repeating) for 1/3. Thus, if π is rational, then there is no need to search for more and more digits.

For centuries, many have tried to approximate the value of π with a fraction, like 7/22 or 223/71. However, despite all the effort, no one was able to precisely express π as a fraction. So is it irrational, i.e. no such fractional expressions are possible?

It turns out that π is indeed irrational. But this fact was not proved until the 1760s when Johann Heinrich Lambert gave us the first proof. Since then, different proofs have been given by several mathematicians including Charles Hermite and Mary Cartwright.

It is rather ironic that π, defined as a "ratio," is not even "rational."

OK, so π is irrational, but so are many other numbers like $\sqrt{2}$ and $\sqrt{3}$. But these radical numbers are solutions (root) to rather easy equations. For instance, $\sqrt{2}$ is a solution to the simple equation $x^2 - 2 = 0$. These numbers, i.e. solutions of a polynomial with integer coefficients, are called *algebraic* numbers. Roughly speaking, while they may not be rational, they can be obtained by *algebraic* operations like adding, dividing, and taking roots. So $\sqrt{2}$ and $\sqrt{3}$ are irrational but algebraic. Numbers that are not algebraic are called *transcendental* numbers (since they *transcendent*, or cannot be obtained by, any algebraic manipulations). It is easy to see that all rational numbers are algebraic.

But what about π? We know it is irrational already. But is it algebraic? This problem was not settled until 1882 (more than a century after π was shown to be irrational), when Ferdinand von Lindermann, a German mathematician, showed that π is indeed transcendental.

So, even among irrational numbers, π is even more "weird" or "mysterious" than radicals like $\sqrt{2}$. You probably do not know a lot of transcendental numbers, but it turns out that there are a whole lot more transcendental numbers than algebraic ones. In a sense, it is true to say that "*Most* real numbers are transcendental."

3. Modern Research on π

We now know that π is neither rational nor algebraic and that, at least theoretically, we could find any desired number of digits of π (and after van Ceulen figured out the first 35 digits by hand). What more do we need to know?

You may find it surprising, but some people are still obsessed with finding more digits of π. Whereas Archimedes (and others like van Ceulen) depended on a geometric method for finding the value of π, the invention of calculus gave rise to new ways of computing it—formulas based on infinite series.

With the arrival of computers, researchers have been carrying out calculations to an insane number of digits. While some of them are computer scientists and engineers trying to find out the processing powers of supercomputers or the efficiency of their algorithms, others are amateur enthusiasts—those who just do it as a hobby.

The website www.piday.org/million/ lists the first million digits of π. Sounds crazy? Well, not so fast.

In March 2019, Emma Haruka Iwao set a new record when she used Google Cloud to calculate a record-breaking 31.4 trillion (yes, not million or billion, but trillion) digits of π. In January 2020, Timothy Mullican of North Alabama Charitable Computing in Huntsville made news by computing more than 50 trillion digits of π. In August 2021, Swiss researchers at the University of Applied Sciences used supercomputers to find over 62.8 trillion digits. There has been so much progress (and there will be more) in computing π. On March 14 (yes, "Pi Day"), 2024, a U.S. team using Solidigm SSDs (solid-state drives) announced that they had computed 105 trillion digits of π. After 75 days of super-computing, they discovered that the 105 trillionth digit was a 6. Engineers with the same group then announced, in June 2024, that they now have over 202 trillion digits. Undoubtedly, these numbers will continue to go higher and higher.

4. Why the Quest for More?

You may be wondering why some people do what they do, such as finding trillions of digits of some irrational number. Perhaps they are irrational themselves? What do they want? Bragging rights? Maybe they want to have their names mentioned in the Guinness Book of World Records. Clearly, they have passion, curiosity, and a lot of time. But there may be more.

As recent stories show, today's efforts are mostly made by groups of computer or engineering experts. They want to optimize the efficiency of computations or algorithms in general, and finding digits of π provides some information on that. Additionally, π plays an important role in cryptology as it is used to generate encryption keys to code and decode data. So the quest for π appears to be more about computing and its applications as opposed to theoretical mathematics.

This is not to say that there is no mathematical reason for computing π. There is at least one important open question related to the digits of π. Is it a *normal* number? A number (written as an infinite sequence) is called *normal* if any sequence of digits appears with equal frequency. For example, does the digit 7 appear randomly, as much as 8 or 0? What about the sequence 123? Does that appear as often as any other three-digit sequence? At this point, π seems to pass every statistical randomness test, so it seems to be a normal number. However, statistical tests do not constitute a rigorous proof in mathematics. We may know the first 202 trillion digits of π, but we have no idea what the rest (or the next quadrillion digits) looks like. Remember, the first 202 trillion is not even 1% of its infinite sequence of numbers.

Questions:

1. Write a historical summary of the number π, including ancient civilizations, the Renaissance, and up to the modern era before the arrival of computers. Who are some of the people who made significant contributions to the understanding of π?
2. Who came up with the idea of denoting this number with the Greek alphabet letter π? Does it stand for any word or anything?
3. What is the Rhind Papyrus? What is its mathematical significance? Who discovered it and when? Where is it located now?
4. Pick a continent or a country in the world, such as Europe, China, India, or Arabic nations. See if anyone in that region contributed to the historical development of π. As seen in this section, people all over the world shared their interest in this fascinating number.
5. This section defines what it means for a number to be rational, irrational, algebraic, and transcendental. Explain why no rational number can be transcendental. List some numbers that are irrational but not transcendental. The text also states that most real numbers are transcendental (there are a lot more of those than algebraic ones). Give examples. Explain it further.
6. Do online research and find some recent news stories on the digits of π. How many digits are known now?
7. There are many books on π written for the general audience (not just for mathematicians). Find a few that look interesting and read them. Write a summary.
8. There are people, all over the world, who have memorized hundreds of digits of π. (You may be familiar with the protagonist of the story *Life of Pi*.) Find out who some of those people are and their motivations for doing so. Do *you* want to do that?

9. The last few paragraphs mention some reasons people are still looking for more digits of π. Why do you think they are so fascinated with this endeavor? Are there other reasons you can think of?

10.12 Project 12: History of Algebra

Equations Studied Around the World

Category: Algebra, History of Mathematics

Description:

We examine the history of algebra, one specific field of mathematics. As we will see, mathematicians from many parts of the world were involved in the development of algebra.

Introduction:

In the United States the term "algebra" is often used (or misused) to refer to high school courses that may consist of a wide variety of mathematical topics, not just in the proper field of algebra. The word "algebra" originally comes from Arabic (more details below), but today this word generally refers to the branch of mathematics concerned with solving equations by formal manipulations and operations on numbers and symbols. There are many subfields within algebra. For instance, in "linear algebra," matrices and vector spaces are used to solve linear systems of equations. In "abstract algebra," axiomatic sets with certain operations are defined and their various properties are studied, often without any reference to numbers. These structures include groups, rings, and fields.

1. Where It All Began

"After eating 3 apples, we only had 2 left. How many apples did we have before?" This is clearly an algebraic question, and it is equivalent to solving the equation $x - 3 = 2$. "How big is a cube-shaped box if its volume is 10 U.S. barrels?" This too is algebra and can be expressed by the equation $x^3 = 10$. The first equation can be solved easily within natural numbers (positive integers) while the second cannot be solved even if we include all rational numbers. So what types of equations can be solved? In what set of numbers, and how? How many possible answers are there? These are some of the fundamental questions in algebra that people have tried to answer, a quest spanning the entire globe and many centuries.

Ancient civilizations such as Babylon and Egypt already knew basic algebraic methods for solving simple equations, which arose from everyday life involving arithmetic and geometry. Astronomy and navigation required some basic algebraic manipulations as well. The Mesopotamian civilization (around 3000 B.C.E.), which included Sumer and Assyria, left some evidence of algebraic techniques used in applications such as trade, commerce, and taxes. (Yes, taxes existed 5000 years ago!)

In addition, many algebraic techniques were used in ancient China (10^{th} to 2^{nd} centuries B.C.E.), where surprisingly advanced algebraic concepts were written down in the book titled *The Nine Chapters on the Mathematical Art*, containing some 246 problems, many of which involve solving equations.

Perhaps it is not surprising that another region where algebra was actively studied in ancient times was India, where mathematicians introduced to the world the idea of the decimal numeration system as well as (wait for it!) the number zero. These are two of the most foundational tools in modern mathematics! The earliest Indian documents on mathematics, written in Sanskrit, date back to around the 6^{th} century B.C.E.

By the 5^{th} century B.C.E., in Greece (where else?) various mathematicians were already busy solving different types of equations, not necessarily motivated by applications but by theoretical interest and philosophical curiosity. Diophantus (for whom Diophantine equations are named) and Euclid were main contributors, but there were many more.

As the Greek Empire expanded its territory under the Macedonian king Alexander the Great, much of the scholarly work written by Greek philosophers and mathematicians ended up in a grand library named after the great emperor: the Library of Alexandria in Egypt. This was by far the largest and the most significant library in the ancient world. But as the influence of this great library declined, scholars began to flee to many other locations (many were practically kicked out during the reign of Ptolemy VIII, a son of Cleopatra). Thanks in part to this intellectual diaspora, the knowledge of algebra spread out all over the world.

2. Islamic Golden Age

Soon the center of mathematics, including algebra, shifted from Alexandria to the House of Wisdom, also referred to as the Grand Library of Baghdad, established around the 8^{th} century. Much of the scholarly work had been written in Greek, but now that the academic center has shifted to Baghdad, major academic books were translated into new languages, resulting in many well-known classic books on science available in Arabic. In addition, translation was done from other languages as well, such as Latin, Persian, Syriac (Aramaic), Sanskrit, and even Chinese. Where there are great books, more and more great scholars gather, and they write even more books, this time mainly in Arabic. This movement ushered in the Islamic Golden Age, particularly in areas such as astronomy, medicine, technology, philosophy, and mathematics. The scientific world was abuzz, not in Europe, but in the Middle East, particularly in the region we call Iraq today.

It was there at the House of Wisdom where the Persian mathematician Al-Khwarizmi, known as the "father of algebra," did a lot of research (early in the 9^{th} century). He wrote his masterpiece *The Compendious Book on Calculation by Completion and Balancing*, the title which contains the Arabic term *al-Jabr*, which ultimately gave us the English word *algebra*. Many consider this the very first book specifically written on algebra. In a somewhat ironic twist, this treatise was soon translated into Latin, received widely in Europe, and was instrumental in spreading the

knowledge of algebra all over Europe. (Talk about curriculum internationalization, 1200 years ago!)

Later, Omar Khayyam, Sharaf al-Din al-Tusi, Jamshid al-Kashi (all Persian/Iranian), and other mathematicians in the Islamic world extended the results of Al-Khwarizmi, including work on cubic polynomials (equations with x^3) and higher-degree polynomial equations—a story to be continued later. Irrational numbers involving radicals were also recognized as useful tools by scholars like Abu Kamil and al-Karaji.

The House of Wisdom played an extremely significant role in the advancement of algebra until its destruction in the mid-13th century.

3. Asian Mathematicians During the Middle Ages

During the Middle Ages (the medieval period, roughly from around 500 to 1500 C.E.), some mathematicians in the East were also quite busy working on algebra. In the 13th century, Qin Jiushao of China published his *Mathematical Treatise in Nine Sections*, where he stated the celebrated "Chinese Remainder Theorem," quite famous and relevant even today.

In India, mathematicians like Aryabhata, Brahmagupta, and Bhaskara II were taking on challenging questions on quadratic equations (polynomials with x^2) and other topics. This idea of finding general solutions to polynomial equations of various degrees would soon become one of the major endeavors of algebra. This quest happened as the central stage gradually shifted from the Middle East to the West.

And the climax—the ultimate answer—was one that nobody had anticipated.

4. Italy During the Renaissance

The world was quickly changing, with monumental shifts in the academic community as well. In Italy, the cultural, artistic, and intellectual landscapes were changing dramatically in what has become known as the Renaissance. In Germany, Johannes Gutenberg invented the movable-type printing press in 1440, making it easy for scholars to communicate. The Protestant Reformation challenged the authority of the Roman Catholic Church in the early 16th century, promoting more scientific and scholarly activities. With that, the world's center of algebraic explorations moved from the Islamic world into Europe.

Before going further, some basic algebraic terms should be explained. A polynomial equation is *linear* (or degree 1) if it is of the form $ax + b = 0$, where $a \neq 0$. This equation obviously has one (and only one) answer, namely, $x = \frac{-b}{a}$. A polynomial is called *quadratic* (or degree 2) if it is of the form

$$ax^2 + bx + c = 0, \qquad a \neq 0.$$

The solution is not immediately obvious, but there are up to two answers to this equation, given by the well-known "Quadratic Formula":

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Many American high school students learn (and dread) this equation. However, it is quite remarkable that this even exists—such a universal, direct way to solve *any* given quadratic equation. Now, that's power!

The next big challenge was, obviously, to solve *cubic* (degree 3) equations and then *quartic* (degree 4) equations. Renaissance history seems to suggest that there was some sort of contest, like today's sporting event, in which challengers were asked to find solutions to these polynomial equations. (Imagine ESPN broadcasting a polynomial competition live from Italy!) The Italian mathematician Scipione del Ferro found a partial solution to cubic questions in the early 16th century but shared it only with his student Antonio Fior. Another Italian Niccolo Tartaglia, who had discovered a more general solution to cubic polynomials, accepted a challenge by Fior and won the competition. Tartaglia kept this secret formula to himself for a while, but eventually he shared it with a man named Cardano in 1539.

Gerolamo Cardano (also known as Geronimo), a colorful character in Renaissance Italy, was also a physician, astronomer, and perhaps even an astrologer and a gambler. His book *Ars Magna* (published in 1545) introduced a systematic use of negative numbers (though not well-received at first) as well as the existence of imaginary numbers (even less well-received). But the major achievement of this book is the general solutions to cubic equations (credit given properly to Ferro and Tartaglia in the book) and to quartic equations (credit given to Cardano's student Lodovico Ferrari).

Well, Tartaglia had a major issue with this book by Cardano. According to Tartaglia, Cardano had allegedly sworn to him, with a hand on the Bible, that he would never share Tartaglia's cubic equation formula with anyone. But now, it is printed in a book for all to read. Their ugly rivalry would continue until Ferrari beat Tartaglia later in the competition.

5. Finally, to Norway

If you think the Quadratic Formula is messy and long, you should take a quick look, if you can, at the Cubic and Quartic Formulas in Cardano's book. They are extremely complicated and definitely not something you would want to memorize. But the good news is that no matter how complicated a polynomial is, you *could* solve it as long as its degree is 4 or less. This remarkable work was already accomplished by the end of the 16th century, thanks to all those mathematicians from the ancient civilizations, the Middle Ages, and the Renaissance.

The next level was obviously to find a Quintic (degree 5) Formula. Brilliant minds from all over the world were seeking one general formula that could solve any quintic polynomial. Many ambitious students tried, hoping to become the next celebrity in the history of mathematics. The Renaissance had come and gone. The Age of Reason brought more changes, like the American Independence, the French Revolution, and the Napoleonic Wars.

Then came Niels Henrik Abel, a young mathematics student from Norway. Solving quintic equations was already a 250-year-old problem when Abel was born. He finally solved the problem in 1824 (in his early 20s) but in a quite surprising way. Remember, everyone was looking for a general formula for quintic polynomials. Abel ended up proving that *there cannot be any such formula*! In other words, he solved the old problem by showing that the answer does not, and cannot, exist. Imagine how difficult it is to prove that something *cannot be done* or *cannot exist*. But he did just that—by showing that one would reach a contradiction if such a formula existed.

Yes, Abel was "able" to prove the impossibility. What a surprising end to this long and international story of algebra!

Questions:

1. Pick an ancient civilization (or a country) and do research on its contribution to the field of algebra. Focus on their effort to find solutions to equations, not just in general mathematics.
2. Find more facts (there are many interesting stories) involving Cardano, Tartaglia, Ferro, Fior, and Ferrari. You can also look up the algebra competition during the Renaissance referred to in the text.
3. This section describes Cardano as a "colorful character." Do more research on Cardano, his work, and his family. Write a 2- to 3-page paper on his life (and death).
4. Do some research and find what the Cubic and Quartic Formulas look like. Describe, in your own, words, how long or how messy they are.
5. Study the life and career of Niels Henrik Abel. He and his fellow Norwegian friends took a long trip through Europe. Where did they go? What happened to them? Was Abel planning to meet the two prominent mathematicians in Europe (Gauss and Cauchy)? Did he get to visit them? Include the sad end of Abel's life as well.
6. It appears that Abel has dealt a mortal blow to the pursuit of algebra when he proved that no general solutions exist for polynomial equations of degree 5 or greater. However, algebra has grown, even to a greater degree, since the untimely death of Abel. Look up Evariste Galois, Carl Gustav Jacobi, and William Hamilton. Find out where they are from, what they did, and how they contributed to the development of modern algebra.
7. You may have noticed that even if all the coefficients (numbers denoted with a, b, c, etc.) of a polynomial are integers, the solutions may not be integers, not even rational numbers (fractions) if the degree is higher than 1. In fact, they may not be real numbers when the degree is 2 or higher. (Consider a very simple quadratic equation, $x^2 + 1 = 0$.) Consequently, this algebraic quest has given rise to larger and larger sets of numbers.
    a. Find definitions of natural numbers, integers, rational numbers, real numbers, and complex numbers.

b. Write a brief paragraph on the Fundamental Theorem of Algebra and explain why the set of complex numbers is “complete” (no need to go further). You can also explore the life and work of the man who proved the Fundamental Theorem of Algebra.
c. Do some research on William Rowan Hamilton and his “quaternions.” Why is an unassuming, humble bridge called Broom Bridge (in Dublin, Ireland) important in this 1843 story about the birth of quaternions?
d. What in the world are “octonions”? Who invented octonions?

10.13 Project 13: Population Growth

Populations Growing Around the World

Category: Exponential Growth

Description:

We compare the population growth of various countries with exponential growth models and analyze potential reasons for the deviation.

Introduction:

Exponential growth can model global and national populations, but only to a certain degree. What can this model, based on exponential functions, do to predict the future? What other factors affect the accuracy of this model? How are populations of various countries really growing now? We consider these questions from the standpoint of exponential growth.

1. Pure Exponential Growth

As we learned in this book, exponential growth is quite different from linear growth. If a quantity $Q(t)$, a function of time *t*, grows *linearly*, the graph of this quantity is a straight line, with a fixed slope, i.e. exactly the same amount is added for each unit of time (per second, per year, etc.). On the other hand, if $Q(t)$ grows *exponentially*, the graph is not straight; rather, it grows faster and faster, bending "to the left" and "higher" as you go to the right on the graph. This is because in exponential growth, it is the rate, the percentage, which is fixed, rather than the slope. In other words, the rate of increase is proportional to the current quantity $Q(t)$, so the bigger $Q(t)$ gets, the faster it grows.

Mathematically speaking, this is precisely the assumption for exponential growth, leading to its definition. Given that the rate of increase is proportional to the current value, we get (through what is called a *differential equation*) the exponential growth formula

$$Q(t) = Q_0\, e^{rt},$$

where Q_0 is the initial quantity (when $t = 0$).

This is essentially the same as the "continuous compounding" formula introduced in Section 3.2. You can also use this equation to find the frequency of each note in a standard musical scale as seen in Section 7.3.

We now examine the global population growth along with the population growth in various countries. Are they exponential? In theory, in the absence of any external factors affecting the rate, the growth would be modeled with this equation. But the world is not an ideal place. There are myriads of social, political, and other factors; as a result, this equation does not accurately describe the population. Let's take a look at some data.

2. Global Population Growth

It is rather difficult to speculate how many people lived on the entire planet centuries ago, so we begin with the year 1600. Since then, the world population is estimated to have been increasing as follows (according to worldometers.info). Numbers since 1951 have been more precise as they are published by the Department of Economic and Social Affairs of the United Nations.

Year	World Population
1600	500,000,000
1700	610,000,000
1804	1,000,000,000
1850	1,200,000,000
1900	1,600,000,000
1927	2,000,000,000
1951	2,543,130,380
1960	3,019,233,434
1970	3,695,390,336
1980	4,444,007,706
1990	5,316,175,862
2000	6,148,898,975
2010	6,985,603,105
2020	7,975,105,156
2023	8,045,311,447

If we let the year 1600 as our initial point ($t = 0$), the graph looks like this (the last data point is for the year 2023, which is 423 years after 1600). This closely resembles an exponential graph.

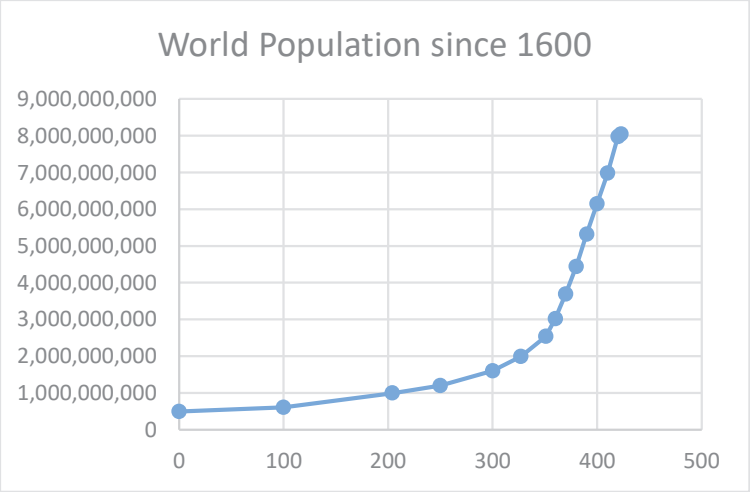


Figure 10.15 A graph of the word population since 1600).

Using the initial value along with the population number in 2000, we can calculate that, for that 400-year span between 1600 and 2000, the annual rate of increase was 0.62736%, which is far below 1%. Quite low, right? But look. Even at such a low growth rate, the planet Earth became 12 times as crowded over four centuries.

You can do this again, using data for the first two decades of the 21st century (2001—2021). Interestingly enough, the graph looks quite different for the last 20 years (here, the population, on the vertical axis, is in *billions*).

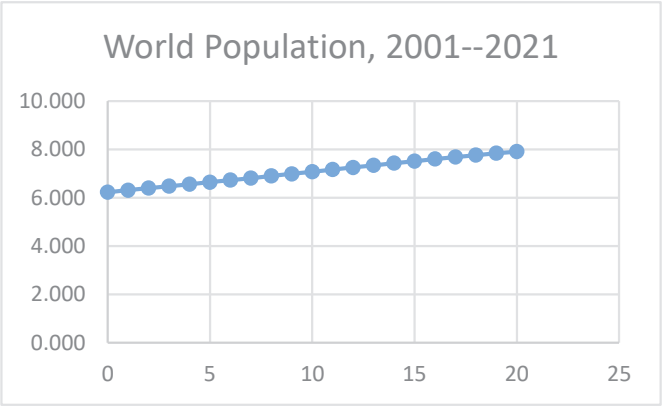


Figure 10.16 A graph of the world population between 2001 and 2021.

This graph makes it clear that the world population is no longer modeled by an exponential function, implying that there are external factors slowing down, or "inhibiting," such growth. You can go to the website World Population Clock to check the current (estimated) world population in real-time.

3. Population Growth in Some Countries

For many years, China was leading the world in population. However, with its one-child policy and other reasons, their growth rate dropped significantly while the rates in other countries stayed high. In April 2023, India surpassed China, becoming the world's most populous country.

Currently, these two countries are followed by the U.S., Indonesia, and Pakistan (note that four of the top five countries are in Asia).

While the population is growing in most countries around the world, the rate of growth is not increasing in many nations. Be sure that you make this distinction clear in your mind.

Said differently, a quantity can still increase while the rate of growth decreases. (This concept can be more precisely stated using the language of calculus.) Therefore, in any country, it is possible for the population to grow, year after year, while the growth rate drops. In fact, this is precisely what is happening in most countries these days. For instance, India has seen a steady population increase each year, but the rate of their population growth (in the last 50 years) peaked around 1963 at 2.34%; then the rate went down to 2.16% (1990), 1,84% (2000), 1.39% (2010), and 0.96% in 2020—even as they became the most populous nation in the world. Another example is Brazil, where the population has grown fairly consistently every year whereas its growth rate has been *decreasing* consistently from 3.12% (1960) to 2.38% (1980), 1.38% (2000), and 0.67% (2020).

There are exceptions, however. Yes, there are countries where the rate of population growth is still increasing. This means that not only is their population growing, but it is growing faster and faster. For instance, in Syria, with a population of about 24 million (2024), the rate of population growth was 0.1% in 2017 but grew to 2.66% in 2021 and 4.98% in 2023. Niger has also seen faster growth, with the rate (not the population) increasing from 2.69% in 1972 to 3.11% in 1980, 3.48% in 2000, and 3.8% in 2020. Their population graph, consequently, resembles an exponential curve.

In contrast, there are nations where the population is no longer increasing. In Japan, a country with more than 122 million people, the growth rate was around 1.3% in the early 1970s but kept declining ever since then, turning negative in 2010 (meaning the population began to *decrease*), then going below -0.5% in 2021. China also saw a negative growth rate in 2023. Other countries currently experiencing negative growth rates include Germany, Italy, South Korea, Portugal, Greece, Cuba, Russia, and Croatia. The country with the sharpest drop (-8.8%) in the year 2022 was, understandably, Ukraine.

4. Why Not Exponential Anymore?

As mentioned earlier, theoretically populations of any kind (not necessarily of humans) *should* grow exponentially in the absence of inhibitors. So what is causing the world population to deviate from the standard exponential model? Why is the growth rate of the world population slowing down? Why do some countries now have negative growth rates?

In addition to geopolitical factors such as conflicts and wars as seen in Ukraine, there are many social, political, medical, economic, and cultural factors. Clearly, the growth rate is the difference between the rate of increase and the rate of decrease; these are mostly determined by birth and death rates, but also immigration and other factors. In some countries, for

instance, parents are finding it difficult and expensive to raise many children and decide not to have many (or any). Politically, a nation could institute a policy limiting the number of children, such as the one-child policy of China, or conversely, promote more childbirth (for military or labor reasons). More people may die in certain countries due to medical and socio-economic reasons. Necessary resources like food, clean water, shelter (living space), clothing, and quality health care could also impact a country's demographic. Population studies, therefore, require a much more complex model than a simple exponential or linear model.

Questions:

1. Pick five or six different countries, diverse in location, population, and race, and look up their population changes since 1951 as well as the shifts in their growth rates. Summarize your data to see if you can identify any trends.
2. Estonia, Latvia, Lithuania, and Belarus are clustered together, surrounded by Russia, Ukraine, and Poland. Similarly, the countries of Slovenia, Croatia, Bosnia and Herzegovina, Serbia, Montenegro, Albania, North Macedonia, Greece, and Bulgaria are also clustered together.
    a. Every one of these countries had negative population growth between July 2023 and July 2024. What possible factors caused the decrease in these countries?
    b. Some neighboring countries saw their populations going up. What countries? Why did these countries see positive growth while their neighbors lost quite a bit?
3. In this book we saw that exponential models are used elsewhere, such as the frequencies of sound (musical notes) and compound interest. Besides human populations, what else can be described with an exponential function? Do some research on this.
4. It is mentioned that, based on data between 1600 and 2000, the annual growth rate of the world population was about 0.62736%. Assuming that the population was half a billion in 1600, this rate gives us the following formula for global population (in billions):

$$P(t) = 0.5\, e^{0.0062736t}$$

    where *t* is the number of years since 1600. Using this pure exponential model, estimate the world population in 2030, 2040, 2050, and 2100. When will the population reach 10 billion? 20 billion? Write a brief essay summarizing your findings.
5. Let's use the formula given in the previous question backward. Suppose that the growth rate has been constant since the beginning of humanity. Using this formula, find out when, theoretically, the world had only 1 million people. When were there only two people on the surface of the earth? (Is this a reasonable assumption?)

10.14 Project 14: The Golden Ratio

A Ratio Appreciated Around the World

Category: Geometry, Algebra, Sequences

Description:

We examine the Golden Ratio, ϕ (denoted with the lower-case Greek letter "phi," sometimes also appearing as φ), from many different angles, including its algebraic definition inspired by geometry, how it relates to the Fibonacci Sequence, and how it is recognized in the world of fine arts.

Introduction:

The precise definition of the Golden Ratio (or the Golden Section) is already explained in Section 7.2, where a picture of a rectangle (which is divided into a rectangle and a square) is shown. The dimensions of those two *similar* rectangles give us a proportion equation, which in turn leads to a quadratic equation whose positive solution is defined to be the Golden Ratio. We noted that it is an irrational number whose first digits are 1.618… A *golden rectangle* was then defined as any rectangle where the ratio of the adjacent sides is the Golden Ratio. Here, we look into its Greco-Roman history as well as its universal appeal as seen in various works of art in different regions of the world.

1. History of the Golden Ratio

One of the most important mathematics books ever written is *The Elements,* written by the ancient Greek mathematician Euclid around 300 B.C.E. Its importance, endurance, and relevance are already mentioned in Section 9.5, along with his famous "Fifth (Parallel) Postulate" and how it has transformed our understanding of the universe. For more details on this book and its author, read Section 9.5.

In that book, Euclid discusses a method of "dividing a line in extreme and mean ratio." It turns out the ratio he is speaking of there is the reciprocal of the Golden Ratio (i.e. 1 divided by the Golden Ratio), which is about 0.618034. So this idea indeed goes back at least 2300 years! Euclid even knew how this ratio is embedded, in multiple ways, in the pentagram (the set of all diagonals in a regular pentagon).

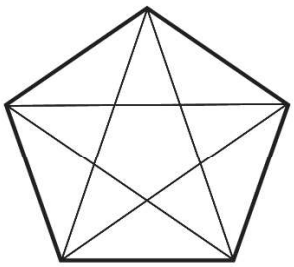


Figure 10.17 A pentagon with five corners.

But we may need to go back even further. The Greek philosopher Plato refers to this ratio as the "most binding of all mathematical relationships and the key to the physics of the cosmos"[212] in his dialogue *Timaeus*, written around 360 B.C.E. In fact, the way Plato describes this "bond" (ratio) is very similar to the way Euclid defined the Golden Ratio some 60 years later in his book. While the number ϕ has been known for centuries, it was not until the 1830s that the term "Golden Ratio" or "Golden Section" was mentioned (coined by the German mathematician Martin Ohm, a brother of the famous physicist Georg Ohm). The letter ϕ was not used until the 20th century.

As seen below, over many centuries this ratio was used, intentionally or unintentionally, in many works of art and music, perhaps hinting at its inherent quality of beauty. Its aesthetic appeal, however, was not explicitly known until the 1850s when the German psychologist Adolf Zeising suggested its intrinsic beauty, especially as it relates to the human body. Then, Gustav Fechner, another German psychologist, carried out a famous experiment in which he asked many participants to rate rectangles of various proportions. The one closest to a golden rectangle came out as the winner.

One more thing. Back in Section 7.2 we learned of a truly amazing relationship between the Golden Ratio and the Fibonacci sequence: the ratio of two consecutive Fibonacci numbers approaches ϕ. The German astronomer Johannes Kepler considers this one of the two "great treasures" of geometry (the other being the Pythagorean Theorem).
We will have more to say about Fibonacci later.

2. Examples in Nature

Along with the Fibonacci numbers, the Golden Ratio seems to pop up in many aspects of nature, sometimes surprisingly. A quick search on the Internet should bring hundreds of articles, pages, images, and videos on the topic.

Some of those resources may show you how the human body shows the Golden Ratio, something perhaps artists discovered long ago. As explained below, certainly Leonardo da Vinci's illustrations in the book *The Divine Proportion* seem to suggest his masterful knowledge of various ratios as an artist. There have been other artists, scientists, and physicians who

[212] History of the Golden Ratio - The Golden Ratio: Phi, 1.618

described ideal human proportions in terms of the Golden Ratio (for instance, the position of the navel, various face proportions, etc.).

Figure 10.18 A Natal Plum plant.[213]

Many other websites on the Golden Ratio may show spiral pictures, similar to the colorful figure that appears in Section 7.2 of this book when the Fibonacci sequence is used to construct bigger and bigger rectangles, representing the Golden Ratio. A *golden spiral* is a logarithmic spiral which grows wide by a factor of ϕ for every quarter turn. It is called "golden" precisely because the associated rectangle gets closer and closer to a golden rectangle (with a side ratio of 1: ϕ).

Golden spiral structures can be seen throughout the biological world. For instance, many seashells, pinecones, pineapples, aloe plants, sunflower seeds, chameleon tails, snail shells, and Romanesco broccoli display such spirals.

Spirals also occur as a result of physics when multiple forces are subject to one another, creating, for instance, swirling whirlpools, whirlwinds, and even tornados. Huge and dangerous storm systems like hurricanes (North Atlantic or Northeast Pacific), typhoons (northwestern Pacific), and cyclones (South Pacific and the Indian Oceans) are even bigger spiral structures. But perhaps the biggest spirals are the galaxies. So how far do we take this discussion of the Golden Ratio? Evidently, the sky is not even the limit.

3. Examples in Architecture, Art, and Music

One of the important properties of the Golden Ratio is its aesthetic appeal. Golden rectangles, whose sides represent the Golden Ratio, appear in buildings and other structures all over the world, even before Plato and Euclid gave us geometric definitions.

A word of warning first. When one says the Golden Ratio is represented or used in a particular piece of work (in art, music, or whatever), it does not mean that the precise, irrational number ϕ is intentionally used. More frequently, it means some ratio found in a work of art is fairly close to the value of ϕ, meaning it is somewhere around 1.6, or the ratio of 8:5. Therefore, the

[213] A photo taken by the author in Los Angeles, 2024.

determination of whether the Golden Ratio is used or not is often subjective and up to interpretation.

That said, perhaps the most famous example is the Parthenon in Athens, a work of classical Greek architecture. But it was already completed around 432 B.C.E., before Plato and Euclid were born, so the concept of the Golden Ratio was probably not solidified or even written down by anyone. However, this very fact—the ratio is seen prior to Plato and Euclid—may be an indication of the aesthetic quality of the number ϕ which is inherent and timeless.

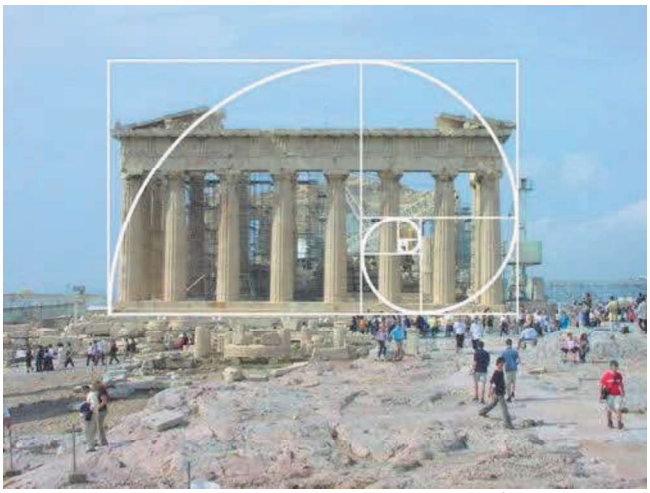


Figure 10.19 The Parthenon in Athens.[214]

Moving forward about 1900 years, a Renaissance mathematician (also a Franciscan friar) named Luca Pacioli published a book titled *De Divina Proportione* (The Divine Proportion) in 1509. This book is famous for at least two reasons. One, because it was one of the first books on the Golden Ratio. Two (perhaps more famous), because it was illustrated by a man named Leonardo da Vinci, who was Pacioli's live-in student in Milan. Da Vinci drew beautiful skeletal figures of Platonic solids, such as regular dodecahedron and icosahedron. Just as the regular pentagon does, these regular solids also include the Golden Ratio on multiple levels.

Pacioli's book contains the word "divine" since, to him, the Golden Ratio represented divine simplicity, the Trinity, and God's incomprehensibility. The book goes further into these themes.

[214] "Alain Badiou" by Peter K. Levy is marked with Public Domain Mark 1.0.

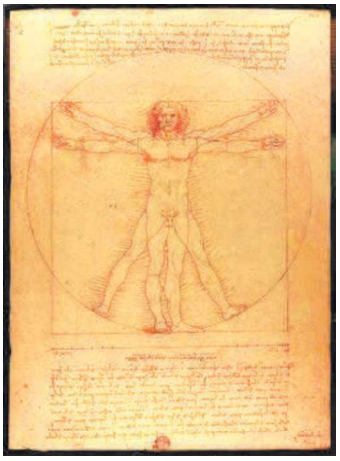


Figure 10.20 Vitruvian Man by Leonardo da Vinci.[215]

Some art historians argue that da Vinci also referenced the Golden Ratio (as it appears in the regular pentagon) in his famous work *Vitruvian Man* around 1490. Many others have superimposed golden rectangles of various sizes onto the Mona Lisa, the most famous work by da Vinci (or the most famous artwork ever, period). Similar suggestions have been made to his painting of St. Jerome and other works.

There have been many papers and articles written about the use of the Golden Ratio in ancient objects like the Great Pyramids, as well as in famous works of art, including Michelangelo's Sistine Chapel painting, *Last Supper* and *Annunciation* by da Vinci, works of Rembrandt, Petrus Christus, Salvador Dali, and even Mark Rothko.

These art pieces are not limited to Europe. The famous 1831 wood print *Under the Wave off Kanagawa* by the Japanese artist Hokusai (the same picture illustrating a fractal in Sec. 7.4) also seems to fit within golden rectangles as well as Fibonacci's spiral construction.

Figure 10.21 Under the Wave off Kanagawa by Katsushika Hokusai.[216]

[215] "'Da Vinci Vitruve' photo by Luc Viatour" by mdwombat is licensed under CC BY 2.0.

[216] Tsunami by Katsushika Hokusai is in the public domain. To see how this (and many other masterpieces) can be overlaid with the Fibonacci spiral, see, for instance, thecollector.com and the Golden Ratio Club website.

Art experts are not the only ones finding the Golden Ratio in works of fine arts. Music historians and musicologists have studied and discovered well-known music pieces that seem to have references to the Golden Ratio. For example, the numbers of measures in the two sections of the first movement of Mozart's Piano Sonata No. 1 in C (K. 279) represent the closest integer approximation (38 and 62) to the Golden Ratio. Similar analyses have been done on the first movement of Beethoven's Symphony No. 5 (the most famous one) and Debussy's introduction to *Dialogue du vent et la mer*.

It appears that the ratio first described by the ancient Greek philosophers seems to stand the test of time, used globally by artists and musicians from Italy, Holland, Spain, the U.S., Japan, Austria, Germany, France, and more.

In recent years, some product designers are applying the Golden Ratio to create popular merchandises. Chefs and culinary specialists have been using the Golden Ratio in studying meal presentation, the visual appearance of a dish, and ingredient compositions. The Golden Ratio is even used in dentistry and dental technology! Further, graphic designers, typographers, and web designers are intentionally incorporating the Golden Ratio to divide their layouts and to develop creative designs that appeal to many users. Applications of the Divine Proportion seem to be endless and quite global.

Questions:

1. See the diagram of a regular pentagon along with its five diagonals. Note each diagonal intersects two others. According to Euclid, the Golden Ratio (though he did not use this term) can be seen in this figure. Explain what parts of this picture represent the Golden Ratio.
2. Various examples of the Golden Ratio as used in art and music are mentioned here. Although not as well known, there seem to be examples from Indian and Islamic art, especially in their intricate geometric patterns, as well as their architecture. Find some examples. (You can use both the "Golden Ratio" and the "Golden Section" along with "sacred geometry" in your search.)
3. As hinted with a reference to the great pyramids, there are elements of African art and architecture that involve the Golden Ratio. Focusing on geometric patterns and African fractals, find some evidence of the Golden Ratio in African art and architecture.
4. The last paragraph in this section mentions modern areas where the Golden Ratio is applied, or at least being considered for future applications, such as product design and dentistry. Find how this is being carried out today. Further, look for some other areas or industries where this ratio is considered for practical use.
5. Find some articles or studies written on how da Vinci used the Golden Ratio (or other "divine proportions") in his *The Last Supper* as well as *Mona Lisa*. What do you think? Are these arguments valid? Use your critical thinking and analyze them.
6. As mentioned in this section, the regular pentagon (with 5 equal sides) has many instances of the Golden Ratio as first observed by Euclid over two millennia ago.

Because of this (and for other reasons) some people argue that the number “5” has some mysterious power. Whether it is true or not, you can write $\phi = \frac{1+\sqrt{5}}{2}$ using only the numeral “5” as follows:

$$\phi = .5 + .5 \times 5^{.5}.$$

Show that this indeed gives the same value as the Golden Ratio.

10.15 Project 15: Ethnomathematics

Mathematical Ideas Developed Around the World

Category: Ethnomathematics

Description:

What is ethnomathematics? What questions does it seek to answer? What are their goals and objectives?

Introduction:

The word “ethnomathematics” was introduced in the late 1970s to refer to a rather new interdisciplinary field of study closely related to mathematics education, ethnic studies, sociology, and anthropology. The general objective of ethnomathematics is to investigate the relationship between culture and mathematics; more specifically, it seeks to understand how mathematical concepts are developed, communicated, understood, and taught in various cultural groups. Researchers study archaeological evidence and historical documents as well as current practice, analyzing how culture and mathematical concepts interact with and inform each other.

One of the goals of ethnomathematics is to explore how mathematical ideas were developed in various cultures, civilizations, and countries. When one studies the history of mathematics, it is easy to see the universality of concepts such as counting and measuring since these are practical and even necessary in everyday life, no matter where and how people live in the world. However, one also notices significantly different ways various people groups have counted, measured, recorded, and communicated these quantities. This is the starting point of ethnomathematics: to recognize that there is diversity in mathematical practices among cultures. In a way, then, ethnomathematics challenges the assumption that mathematics is universally identical and culture-independent. Here are some examples of the diversity seen around the globe as we examine a few components of ethnomathematics.

1. Counting

Since the beginning of time, people have counted things. But how? Did everyone use the decimal system (base 10) we use today? Why use base 10 anyway? If humans had twelve fingers, would we be using a base 12 (duodecimal) system? This is ample evidence that the ancient Babylonians used a sexagesimal (base 60) numeration system; it was fraction-friendly since 60, unlike 10, can be divided evenly by many factors like 2, 3, 4, 5, 6, 10, 12, and more. The modern-day practice of dividing an hour into 60 minutes and a minute into 60 seconds may be based on their sexagesimal system.

Another question is the vocabulary of numbers. While there are obviously infinitely many positive integers, anthropologists have identified a tribe in the Amazon rainforest that uses only three terms: one, two, and many. Another tribal community in Namibia (southern Africa) had four terms: one, two three, and many. They had no words for bigger quantities in their languages.

As the level of human civilization evolved, more words were created, in various languages, to refer to larger and larger numbers. In English, for example, we have a new word for every three additional digits, including a billion (a one followed by nine zeros), a trillion (twelve zeros), a quadrillion (fifteen zeros), and so on. Then, skipping a whole bunch of zeros, we have the word *googol* (note the spelling, to be distinguished from the company name), representing 10 to the 100^{th} power. In Chinese (and many Asian languages), there is a new word for every four digits, such as *jing* (sixteen zeros) and *gai* (twenty zeros).

It should be pointed out that ancient numeration systems such as the Roman numerals (or those using Hebrew or Greek letters) could not easily describe very large numbers since their alphabets were all finite. It is through the Arabic numeration system, with its marvelous place values, that we can express not only the (infinitely many) integers but also fractions and (uncountably infinite) real numbers.

2. Writing, Recording, and Calculating Numbers

The Inca civilization famously used knotted strings/cords called *quipus* to express and record numbers and to carry out calculations. They did use a base-10 (decimal) system as we do today, but they expressed these numbers by a sophisticated system of strings arranged in certain configurations consisting of a main cord, pendant cords, single knots, long knots, and figure-eight knots. What an interesting way to denote quantities! This is significantly different, obviously, from Roman numerals, which had been used in Europe for centuries.

The Incas, in addition to quipus, used their own abacus called the *yupana*, a rectangular tablet used to carry out arithmetic operations.

Asian countries have long used abacuses (the exact types were slightly different in various regions) for computations. Although well-trained experts could perform highly complex computations faster on the abacus than using a calculator, the use of an abacus could be complicated. For instance, on an abacus, adding 24 to a number may involve adding 100, subtracting 80, and then add 4 by adding 5 first and then subtracting 1. These devices were used even into the late 20^{th} century, when calculators and computers were readily available. In fact, even in the 1990s, some workers were checking the accuracy of calculators or computers with their abacuses, just to make sure.

3. Measuring

Just as important as counting is in any human society, measuring has also played important roles in comparing, describing, and communicating everything from length, depth, weight, and volume to time. In an earlier section we talked about various units of measurement used around the world. Many (like feet and cubits) are based on human body parts. Others (like acres) are based on rather vague standards (e.g. the farming area a team of oxen could plow in a day). Until there were international agreements, such as the metric system established soon after the French Revolution, each nation used different units of measurement, making international commerce rather difficult.

Time is another common quantity to be measured. Because of the 24-hour rotation of the earth, the concept of a "day" is universally recognized; however, how you divide up that 24-hour period varies. The ancient Egyptians are believed to have divided the night into 12 units (like our hours) based on the astronomical patterns they observed. The Babylonians, known for their base-60 numeration system, also divided the daytime into 12 time intervals, largely influencing the later Greek and Roman civilizations. Of course, the exact duration of sunlight changes over the year, and their sundial may not have been quite precise, so the "hour" was not clearly defined as today. In other cultures, the night was evenly divided into four intervals, giving us such terms as "the first watch" (the first three hours). Different systems were used in Asian civilizations; however, somewhat surprisingly, the idea of partitioning the day (or the night) into 12 intervals seems to be a rather universal practice. The sundial, which tracks the shadow movement during the day, became a precursor to our analog clock. If the clock had been invented in the Southern Hemisphere, where the sun rises in the east, goes to the north, and sets in the west, the hands of a clock might be moving "counterclockwise" today.

4. Fractals, Patterns, and Geometry

Some cultures and peoples consider geometric patterns seen in art, craft, and architecture quite "mathematical," much more so than others. We have also seen tessellations, geometric logos, symbols, and their symmetries from mathematical standpoints. In ethnomathematics, the focus is on the relationship between culture and the geometric patterns (like fractal structures) produced by the people in that culture.

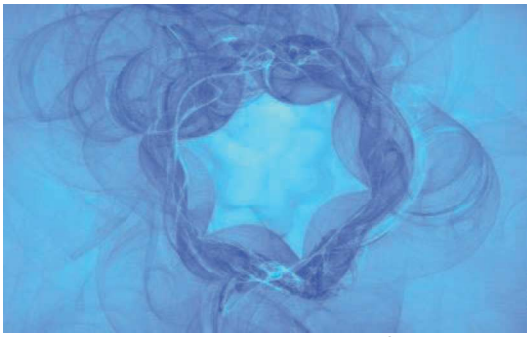


Figure 10.22 Fractal frame.[217]

[217] Image, created by Benjamint444, is used under CC BY-SA 3.0.

For instance, many fabrics discovered in Ethiopia show beautiful, intricate, and complex patterns using self-similarity seen in fractal geometry. Other works of art and sculptures in Africa also have similar recursive patterns. In addition (on a larger scale), a city built by the Kotoko people in Cameroon appears to have a fractal design, utilizing rectangular iterations. This aspect of ethnomathematics has, as you can easily imagine, involved significant work by anthropologists and archeologists.

Many tribes in Africa and other parts of the world use complicated knotting patterns in basket weaving, some of which are studied by today's mathematical experts in geometry and topology.

5. Mathematics in Education

Many people studying ethnomathematics are educators who seek to incorporate the diversity and historical context in mathematics education. Their goals may vary from person to person, however. Some are making intentional efforts to internationalize their curriculum to expand the scope of traditional mathematics, particularly in non-STEM students in institutions of higher learning. Certainly, liberal arts students can benefit from learning about diverse approaches to mathematical understanding in the world in addition to traditional mathematics consisting of quantitative reasoning and problem-solving. This approach is one of the ways whereby educators have attempted to make mathematics more equitable, accessible, and relevant to the diverse student populations, valuing their cultural backgrounds.

Another ethnomathematical approach to education may highlight different numeration systems used by various tribes and civilizations. It may identify diverse operational algorithms taught in different countries (e.g. long division may look very different in some countries) as well as what types of mathematics are taught at what levels in various countries (and why). Such an approach could be beneficial to future elementary teachers, not only to appreciate diversity but also to understand the basic notion of numbers.

In the late 1960s, Paulo Freire, a Brazilian educator, began advocating his theory of critical pedagogy, which laid the foundations for ethnomathematics. It was to be used as a tool for cultural identity and social justice, a means to solve equity gaps in education and to promote peace. This was one of the original purposes of another Brazilian educator Ubiratan D'Ambrosio, who is credited for coining the term *ethnomathematics* as part of the "People Education" movement in Latin America. The idea was to embed cultural, historical, and ethnic contexts in teaching to create more inclusive and engaging environments conducive to learning. Some have criticized ethnomathematics (perhaps justifiably so) when its topics were suggested as an alternative to (instead of in addition to) the basic algebra/analysis sequence of high school courses. These alternative courses went by names like "Rain Forest Algebra." The critics accused these courses by pointing out that algebra is nowhere to be found despite the name and traditional mathematics topics, skills, and exercises were replaced by "everything but algebra."

It should be pointed out, clearly, that while ethnomathematics involves observations about mathematical ideas in history, culture, and language, many scholars do not consider it mathematics. Parts of it belong to anthropology, sociology, education, history, linguistics, and even philosophy. It should in no way replace the fundamental concepts of proper mathematics (such as sets, functions, and equations) in formal mathematics education.

Questions:

1. The text mentions the base-60 system used by the ancient Babylonians. Find out why they chose 60 as the base for their numeration system. How did they write down their numbers, and on what? (Remember, this was around 3000 B.C.E., long before the Roman numerals were written on papyri.)
2. Find some videos on *quipus* and *yupanas* used by the Inca people. Summarize how they used different types of knots to express numbers and a *yupana* to perform basic arithmetic operations.
3. Write a short essay on the history and evolution of the metric system, including its origin (who, where, when?), how each unit was originally defined, and how they are currently defined. For instance, what was one meter supposed to define originally? What about one gram and one liter?
4. As the text suggests, it is somewhat surprising that many civilizations around the world independently decided to divide the day (daylight) and the night into 12 intervals (now called "hours"). Why do you think the number 12 was chosen, rather than 10 or 5? Were there civilizations that did not use 12 time intervals?
5. The Babylonians' sexagesimal system may have contributed to our time units (60 seconds in one minute, 60 minutes in one hour). This forces us, in some sense, to use a base-60 system (not a base-10 system) as well. Consider, and solve, the following. For each, show your work. Then, summarize some benefits and disadvantages of the sexagesimal timekeeping. Would you rather use a decimal system?
    a. How many hours did you work if you had six 40-minute shifts?
    b. A group of eight staff members need to man a front desk for 3 hours and 15 minutes. How many minutes does each member have to be at the front desk?
    c. If you are given 60% extended time in a 45-minute English exam, how long do you have to take the exam?
6. There are some examples of African fractals given in the text. Find other examples of fractal structures (characterized by recursion and self-similarity) in art, architecture, and craft in other regions and/or time.
7. Find some videos that illustrate how to use an abacus. Perform some operations like 26 + 48. Write down each step required to complete the calculation. Does it make sense?
8. What was the "People Education" movement in Latin America? How were mathematical skills and knowledge seen within the movement? What role did ethnomathematics play in that movement?

10.16 Project 16: Mathematics in Literature, Cinema, and Theater

Original Works Created Around the World

Category: Mathematics and the Arts

Description:

We examine ways in which creative authors, writers, and directors have incorporated, directly or indirectly, mathematical objects and concepts into their artistic works in areas like art, literature, movies, and theater.

Introduction:

Although authors, writers, movie directors, and playwrights are generally not big fans of mathematics, some of them incorporate mathematical ideas, objects, concepts, and principles into their works. This is by no means limited to science fiction. Sometimes these mathematical concepts play substantial and major roles, where the plot depends on them. In other situations, mathematical ideas are indirectly referenced or implied on the side. Other works may feature mathematicians and their lives. Let's take a look at some examples of literary, cinematic, and theatrical works created by writers and artists from around the world.

1. Works Featuring Mathematicians

1.1. Biographies of Mathematicians

There are a few well-known films and books featuring real mathematicians and their lives. Except for documentaries, these books and films frequently exercise "artistic license" (not every detail is factually accurate).

- *Man Who Knew Infinity* is a book first published in 1991 about the life of Srinivasa Ramanujan, an Indian mathematician without formal training or schooling. In 2015, the story was made into a British movie. Just like today's international students studying at an American college, Ramanujan left his native India to do mathematical research at Cambridge in England and faced many cultural and physical challenges. It is a fascinating story of mathematics, family, racial prejudice, cultural barriers, and personal struggles.
- *A Beautiful Mind* is an American film released in 2001, directed by Ron Howard. This too is based on a book written earlier, in 1998, by Sylvia Nasar under the same title. It is based on the life of the American mathematician John Nash, who won the Nobel Prize in Economics in 1994. His life, however, was affected by schizophrenia, resulting in hallucinations and other erratic behaviors.

- *Hidden Figures* is a SAG-award-winning drama released in 2016 about three African-American women who worked at NASA in the 1960s: Katherine Johnson, Dorothy Vaughan, and Mary Jackson. The film depicts racial and sexual discrimination as well as reconciliation at the space agency during the Cold War. The film is based on a book published in the same year by Margot Lee Shetterly; the full title of the book is *Hidden Figures: The American Dream and the Untold Story of the Black Women Who Helped Win the Space Race*.
- *Secrets of the Surface* is a short documentary about Maryam Mirzakhani, an Iranian mathematician formerly affiliated with Stanford University. In 2014 she became the first woman to win the Fields Medal, the highest and most prestigious honor in mathematics. Truly international, this movie was filmed in Iran, Canada, and the United States. In 2017, she tragically died of cancer at age 40.

1.2. Fictions with Mathematicians

There are many other works depicting real or fictional mathematicians. As fictions, these do not represent true stories.

- In the German-language novel *Measuring the World* written by Austrian writer Daniel Kehlmann in 2005, we see fictional accounts of German mathematician Carl F. Gauss, German geographer Alexander von Humboldt, and French explorer Aimé Bonpland. This international story was quite popular, selling more than 2 million copies in Europe. A movie version came out in 2012 under the original German title *Die Vermessung der Welt*.
- There is a Japanese novel *The Housekeeper and the Professor* (2019), in which a mathematician without long-term memory struggles with everyday life but shares his beloved equation $e^{i\pi} + 1 = 0$ with a housekeeper as well as her son. It won numerous awards and has been translated into English and other languages. The drama was also made into a film with the title *The Professor's Beloved Equation* (2006).
- The 2014 British film titled $x + y$ (the American title *A Brilliant Young Mind*) is a drama about a young British prodigy named Nathan with autism, based on a real person who won the silver medal at the 2006 International Mathematical Olympiad.
- *Proof* is a 2000 play by the American playwright David Auburn. In this story, which received both a Pulitzer Prize and a Tony Award, Catherine, the daughter of a mathematician with mental illness, tries to prove the authenticity of her father's proof on prime numbers.

2. Works Featuring Mathematical Objects

2.1. Möbius Strip

If you take a long strip of paper or tape and glue the two opposite ends with half a twist, you end up with a surface called the Möbius strip (or Möbius band). This is an example of what is called a non-orientable surface; if you start walking along the middle of this strip, you eventually end up on the other side, upside down, because of the twist. In other words, it has no "inside" or "outside." If you cut it along the middle line, it will not split into two pieces. Rather, the cut results in a single strip, twice as long as before. This object is named after August F. Möbius, a 19th-century German mathematician and astronomer.

The strip is connected, yet one side seamlessly leads into the "other side," making it an interesting, somewhat mysterious, link between reality and… whatever is on the "other side." Due to this intriguing property, the Möbius strip is sometimes used as a metaphor or a fictional bridge between the real and the surreal. A.J. Deutsch, an American astronomer and author, wrote a sci-fi short story in 1950 called "A Subway Named Möbius," in which a Boston-area train disappears into an unknown universe. There are a few other short stories that use the Möbius strip for a mind-twisting (pun intended) plot, including Arthur C. Clarke's "The Wall of Darkness" (1949) and George Smith's "The Mobius Trail" (1948).

The Swiss designer and artist Max Bill created a Möbius strip as a sculpture work in 1935 and titled it *Endless Ribbon*. The Dutch artist Maurits C. Escher, well-known for many eye-deceiving images and geometric pictures such as *Angels and Devils*, drew a famous work in which ants are crawling on a Möbius strip.

2.2. Two- and Three-Dimensional Spaces

One notable work of literature to be mentioned here is *Flatland: a Romance of Many Dimensions* by Edwin A. Abbott, a British teacher. A great piece of satire on the English society of the late 19th century, this novella was first published in 1884 under the pseudonym "A Square." In this story the male characters are polygons; the higher their social ranks are, the more sides they have. Women, on the other hand, are line segments. They all live in a two-dimensional world, i.e., a plane. But then a sphere, a three-dimensional character, passes through their universe, appearing as a circle and then vanishing away, causing all kinds of confusion and chaos. It has a sequel called *Sphereland* by Dionys Burger (1957) and another called *Flatterland* by Ian Stewart (2001). Both *Flatland* and *Flatterland* have been made into movies.

2.3. Other Mathematical Concepts

There are many other examples which incorporate some mathematical ideas or theorems, both in English and in other languages. Here are a few more with a brief description of each.

- *Infinite Jest* by David Foster Wallace (1996), whose reflective, non-linear narrative is based on the Sierpinski triangle, an example of fractal geometry.

- *Uncle Petros and Goldbach's Conjecture* by the Greek author Apostolos Doxiadis (1992), a story that involves a Greek boy and his uncle Petros Papachristos, who tries to prove the famous Goldbach Conjecture.
- *The Oxford Murders* (2008), a British-Spanish thriller, based on a novel (by the same title) written in 2003 by Guilermo Martinez, an Argentinian mathematician and writer. As the title suggests, it is indeed a murder mystery taking place in England, but the film includes a lot of mathematical references including Fermat's Last Theorem and Gödel's Incompleteness Theorem.
- *Arcadia*, a play by Tom Stoppard, an English playwright, premiered in 1993 in London. The story goes back between the present and the past, focusing on chaos theory, the second law of thermodynamics, and entropy (as well as love, human relationships, and history).

Questions:

1. There are many movies with mathematical contents in addition to those listed here. Find some, and describe how mathematics is used/referenced in the films.
2. The Möbius band is an example of a non-orientable surface as stated in this section. Another example is the "Klein bottle." Do some research on the Klein bottle and answer these questions:
3. How do you "construct" (theoretically) the Klein bottle?
4. What makes it so unique (weird)?
5. In what sense is it non-orientable, like the Möbius band?
6. Where did the name "Klein bottle" come from?
7. Pick one book, movie, or play mentioned in this section. Read or watch it, and write a brief paper on a mathematics topic included in that work. Does the plot depend on it? Why do you think the author/writer used that? What effect does it have on the work itself?
8. *Flatland* is a classic work, quite well-known as a satirical piece. Explain what types of characters in the story are circles, squares, triangles, and line segments. Also, explain what happens when the extraterrestrial "sphere" traverses their two-dimensional universe. You may want to draw some pictures for your explanation.
9. The "professor's beloved equation" mentioned in this section refers to the incredible equation $e^{i\pi} + 1 = 0$. What makes this equation so fascinating? What are those numbers not represented by numerals (those besides 0 and 1)? Explain what the equation means. Is there a name given to this equation?
10. Look up other artworks by M. C. Escher. (There are many images, videos, and animations based on his pictures.) Find a few you are interested in, and write a paragraph on each, commenting on the mathematical structure(s) included.

10.17 Project 17: Games and Mathematics

Games Played Around the World

Category: Game Theory, Logic, Probability

Description:

We examine different games played around the world, particularly those with significant mathematical or logical foundations. We also consider sports as games and how their rules influence the way they are played.

Introduction:

An interesting aspect of mathematics is games developed and/or played in different countries. Here, the term "game" could include recreational games, children's games, board games, and even sports. There is a well-established field of mathematics called *game theory*, but it has more to do with formal definitions and classifications of games as well as theorems, proofs, and economic applications. Rather, we focus on a more informal aspect and a looser sense of the word "game" in this section. While some of them are games of pure luck, most games involve calculated moves, probabilities, logic, and strategies that require mathematical thinking. What games are played in a country (and with what strategies) may give insight into the culture of the country. Let's take a look at some popular games played around the world.

There are various definitions of games, but most of them involve the following elements:

- Goals or objectives, which may involve points or taking certain pieces of the opponent(s)
- Rules, which restrict certain moves and establish fairness
- Turns or interaction, as players engage with each other sequentially or simultaneously

1. Popular Games and Mathematics

1.1. Chess and Similar Games

Chess has been quite popular around the world for about fifteen centuries. Having been conceived in India, where it was originally referred to as *chaturanga*, the game of chess underwent some major transformations in Persia and Europe until it was standardized in the current format and rules. But long before that, similar games had been played in China (xiangqi), Japan (shogi), Thailand (makruk), and Korea (janggi). These board games are modeled after military battles, including pieces such as kings, generals, queens, bishops, knights, rooks, and pawns, each of which has certain allowable moves.

These games have been analyzed, with enormous effort spent by many, from mathematical, computational, and logical standpoints. Computer algorithms have been applied, and artificial intelligence seems to beat some of the greatest chess champions and grandmasters of the world. Similar things can be said of other games such as *go* (Japan), *mancala* (Africa and Asia), *Otello (Reversi)*, and other strategic board games.

Many of these games are promoted as educational tools as players develop logic skills, pattern recognition, deductive reasoning, and critical/rational thinking. Indeed, many mathematical tasks such as proving theorems require comparing and analyzing various "moves" to get to a certain conclusion or argument, in ways similar to how one plays chess.

1.2. Card, Tile, and Board Games

Card games are commonly played all around the globe. One might immediately think of a standard deck of 52 playing cards consisting of four suits, with which you can play games like Poker, Bridge, Rummy, and Solitaire. However, there are many other sets of playing cards in various countries, including Ganjifa (India) and Hanafuda (Japan). Additionally, there are country-specific variations of the standard cards in Italy, Spain, Germany, and Switzerland; the number of cards in these sets varies from 32 to 48. It is clear that, whatever games you play with these cards, there is a lot of probability, combinations, permutations, and statistics involved.

Tile games from around the world include Dominoes (many Western countries) and Mahjong (China). Also, in addition to chess, the world has many popular board games including Backgammon, Catan, Dominion, Risk, and Qwirkle. All of these games require risk calculations, logical decision-making, and/or resource management. Many also have international or world championship events on a regular basis (see, for example, the World Boardgaming Championships and the World Series of Board Gaming).

1.3. Video and Computer Games

In recent years, thousands of video games have been developed and played worldwide, and many of them have been used to teach or nurture mathematical skills. For example, the very popular game *Minecraft* has an educational edition, and its "3D sandbox environment" and "block-based mechanics" can be used effectively to teach mathematical and geometric concepts such as perimeter, area, coordinates, ratios, and various geometric patterns. Studies have been conducted to analyze the educational effectiveness of *Minecraft*.

Fortnite is another extremely popular game with more than 350 million registered users worldwide (more than the entire U.S. population). Strategies that involve mathematics are abundant in this game, from estimating the trajectory of a projectile and managing resources to figuring out proper angles, distances, and other dimensions of what the player must construct. Familiarity with probability and expected values may also be helpful.

There are many other video games where mathematics is heavily immersed in the programming and it can also be used for winning strategies.

2. Popular Sports and Mathematics

Sports are a particular type of game, usually involving physical skills such as agility, coordination, athletic precision, and/or stamina. There are two essential ways in which mathematics is related to sports: (1) points (could be runs, scores, or goals in some sports) within the rules of the game, and (2) points used in competitions. Here we will focus on some sports widely played internationally, such as soccer ("football" in most countries), basketball, baseball, volleyball, and golf.

2.1. Points within the Rules

In many popular sports, the objective is to earn or score more points (or goals or runs) than the opponent. "Goals" are simple to understand in games like soccer and hockey, and so are "runs" in baseball (each time a player reaches home safely). However, in some sports the number of points awarded depends on the play. Consider the following examples:

- In basketball, a free throw gives you one point. A field goal gives you two points except for those from beyond the three-point line.
- In American football (which is not all that well-known outside the U.S.), a touchdown is worth six points; a field goal, three; a safety, two; a try (extra points), either one or two points.
- In rugby, a try earns five points; a drop goal, three; a penalty, three; a conversion, two.
- In tennis (very uniquely), points are counted 0 ("love"), 15, 30, and 40, and "game."
- In golf (using stroke play), the total number of strokes is recorded; the lower the better, which is quite different from most other sports. (There is another scoring system called "Stableford," where higher scores are better.

From time to time, sports organizations have changed the rules whereby these point systems are adjusted. Some of the changes have significantly altered the game, like the three-point rule in basketball, which was introduced in the NBA in 1979 and the NCAA in 1986.

In volleyball, the point system was changed in 1999 to adopt "rally scoring," in which a point is awarded on every rally, regardless of which team served the ball. Before that, a point was awarded only when the serving team won the rally. If the other team won the rally, the team got to serve next but was not given any points. This changed the game rather dramatically.

2.2. Points Used in Competitions

Points are also used in sports tournaments and other competitive events to determine who gets to advance in the competition. For example, in the Rugby World Cup, the teams are grouped into "pools," and only the top two teams in each pool get to advance to the next round. Within each pool, four points are given for a win and two for a draw. But a losing team can still get one point if defeated by seven or fewer points (here, the "points" refer to those

scored in one game). A winning team also gets an extra point if it scores four or more tries in the game.

A similar system is used in soccer. For years, in the first round (matches played within each group), two points were given for a win, one for a draw, and zero for a loss. But FIFA, the world's governing body of soccer, changed a new system, awarding three points for a win (other points remained the same) starting in the 1994 World Cup (held in the United States). As you can imagine, this changed the way teams played in the first round in a significant way; teams began trying to win rather than not to lose. Get the difference?

There is a lot of mathematics involved in these point systems, and one can ask many questions regarding necessary and/or sufficient conditions for a team to advance to the next round, logical rules to break ties, and potential bonus points or penalties. Some of these are in the questions below.

Questions:

1. Compare and contrast the game of chess with the Japanese game of shogi. In particular, shogi has one feature unique among other similar games. What is it, and how would it affect the mathematical/logical strategies?
2. If you draw one card randomly from a standard deck of cards, the probability that you get a spade is 1/4 while the probability that it is an Ace is 1/13. (There are many questions much more complicated.) Pick a set of playing cards from another country (those mentioned in this section), find out their compositions of numbers and suits, and find similar probabilities.
3. Pick one video or computer game played around the world. List all the mathematical or logical skills required to play that game well. Explain why.
4. One of the critical tasks of a mathematician is to come up with precise, useful, and rigorous definitions. This is often quite challenging. Try to come up with a precise definition of sport. Then, test the accuracy of your definition with activities such as walking, shooting, bowling, ice hockey, air hockey, swimming, typing, hiking, speed eating, yoga, bungee jumping, skateboarding, snorkeling, fishing, poker, and various video games (obviously some of these should satisfy your definition while others should not).
5. Choose a sport with an unconventional point/score system. Explain why such a system is used and for what purpose.
6. Select either American football or rugby, consider changing the point system, and imagine the consequences of those changes. For instance, what if a touchdown were worth only 4 points or if a safety were worth 7 points in American football?
7. The 1999 rule change in volleyball is described. Why do you think rally scoring was adopted? In what way did it change the game? Was it a good decision or not?

8. The point system was changed between the 1990 World Cup in Italy and the 1994 World Cup in the U.S. Why did FIFA decide to make the change? Did it result in intended consequences? You may want to ask soccer fans who knew the old system as well as the new one. (There may be more changes coming.)

10.18 Project 18: Gambling and Mathematics

Bets Places Around the World

Category: Probability, Statistics

Description:

We examine the history and practice of gambling around the world, including lotteries, taking into account its mathematical foundations and applications.

Introduction:

Gambling, legal or illegal, has been practiced for centuries around the world. (Wanna bet?) It exists in a wide variety of formats and types, including casino games (card games, dice games, roulettes, slot machines, etc.), sports betting (or betting on any future outcomes), horse (and other) racing, public lotteries, and even fund-raising raffles. Some would even argue that more general activities like investment and farming (even life in general) are gambling, but here we will focus our attention on more traditional money-betting activities that are commonly known and played around the world.

1. History of Gambling

1.1. Ancient Civilizations

Humans have been gambling for millennia, perhaps since the beginning of civilizations. The evidence includes stones, dice, animal bones, and other betting-related items discovered in China (3000 to 2300 B.C.E.), Egypt (2000 B.C.E.), Greece (1000 B.C.E.), and other regions. Interestingly enough, keno slips have been unearthed in China, dating around 200 B.C.E., to fund public works, possibly the Great Wall. Casting of lots is another common method used in gambling. Even the Bible mentions it as a means used for land allotment, apostolic succession, and soldiers getting Jesus' garments at the Cross. Six-sided dice were discovered in Mesopotamia as early as 3000 B.C.E. and were quite popular among Greek soldiers (even though gambling was generally illegal in ancient Greece). In addition to dice, other items of chance had been used for gambling in ancient civilizations, such as astragali (in Mesopotamia and Greece) and vibhitaka tree nuts (in India) though these items may have been possibly used for divination and superstitions.

1.2. China

Playing cards are believed to have been invented in China. These gambling cards first appeared during the 9^{th} century. Along with paper cards, tiles made of various materials were also used. Pai Gow (meaning "make nine"), consisting of 32 domino-like tiles, is a gambling game that has been played in China since at least the Song Dynasty (960 to 1279). It is believed to be the first form of dominoes in history. This game is still very popular in casinos in China and other

countries. Pai Gow has many numerical strategies based on such mathematical ideas as modular arithmetic and combinatorics. This game is featured in the 2012 action film *Premium Rush*.

1.3. Renaissance Italy

Many types of gambling, especially card games, appear to have been used as popular means to fund public works in Florence way back in the 1530s. Basset and Faro (banking games) were quite common there, and these games made their way to other major Italian city-states, including Venice. Casinos began to appear in these places as well. In fact, after failed attempts to ban many private gambling clubs, Venice opened its official gambling house "The Ridotto," possibly the first state-sanctioned casino in Europe, in 1638. Besides card games like basset, a common game at the Ridotto was Biribi ("cavagnole" in French), which was a pure game of chance using a number tableau, similar to today's roulette.

1.4. Europe and the United States

Blaise Pascal, a French mathematician and philosopher, was still a teenager when the Ridotto opened its doors in Venice. As legal gambling spread throughout Europe and beyond, by the mid-17th century, Pascal, together with his friends Pierre de Fermat and Abraham de Moivre, laid the foundation of probability theory, the mathematics behind gambling, odds, and house edge. Around the same time, roulettes gained popularity in France, Italy, and other countries. Casinos then began to appear all over Europe, and so did the social consequences of gambling. In England, where gambling had grown very popular among the aristocrats, it became a serious issue as some wealthy men would wager away money and estates. Governments tried to control, regulate, or ban gambling altogether, but the enforcement was, as one can imagine, quite difficult. (Some things never change.)

Gambling also traveled across the Atlantic. In the late 19th century, the Liberty Bell machine (a basic reel-based gambling machine) was invented by the Bavarian-born U.S. inventor Charles Fey. Soon after, slot machines were also introduced and widely distributed. Early machines were placed in such businesses as saloons and general stores. However, public concerns had been growing, with many citizens, religious leaders, law enforcement, and legislators speaking up against gambling. By 1910, virtually every form of gambling was prohibited in the United States.

2. Types of Gambling

Some gambling has outcomes that are (supposedly) completely random, without any skills or computations; examples include roulettes, lottery tickets, and dice games. Others require strategies or certain knowledge, such as poker, horse racing, and sports betting. Here are some major types of gambling forms.

2.1. Public Lotteries

A lottery is a type of gambling where players buy tickets with number(s) and prizes are given to those holding tickets with winning numbers randomly drawn. Scratch cards work in a similar way. Because of potential fraud, lotteries are banned in some countries and regions. However,

many (yes, very many) countries do allow, endorse, and even promote certain lotteries under government supervision and regulations. In fact, records show that Queen Elizabeth I chartered a lottery in 1569. Some lotteries today are specifically designed to fund public works, education, or humanitarian projects.

Most Americans are familiar with, or at least heard of, state lotteries, Powerball, and Mega Millions. Examples of lotteries in other nations include Mega-Sena (Brazil), Svenska Spel (Sweden), Takarakuji (Japan), Totoloto (Portugal), Hrvatska Lutrija (Croatia), Welfare Lottery (China), and Baba Ijebu (Nigeria).

2.2. Sports Betting

Sports provide easy and convenient opportunities for gambling. This has become a much more complex social issue in recent years; people can bet not only on who wins but also on every detail of a game or a competition, e.g. how many strikeouts the starting pitcher gets in a baseball game, if one team wins by at least three goals in a soccer match, and how many free throws a particular player will miss in a basketball game.

Various terms are used in sports begging. In-game betting is the format in which one can place wagers on a game while the game is being played. A *parlay* refers to a combination of multiple bets; to win a parlay, all of the individual bets must be won; you can think of this as the *intersection* of multiple sets, or an *and* statement in logic. A *prop* (for "proposition") bet is for a specific event or outcome instead of the final score or outcome of a game. There are, for instance, player props, team props, and game props as described in the paragraph above. These props can obviously be included in parlays. These new trends have resulted in rather unfortunate events, such as an unknown high school football player being harassed by a gambler across the country because the player just happened to intercept a pass, thereby causing the gambler to lose tons of money.

Countries around the world (and states in the U.S.) have very different policies on sports betting; some officially sanction it while others ban the practice altogether. Most countries allow sports betting with tight regulations. Some countries allow gambling only in certain sports. For instance, in Argentina, gambling is not allowed in their most popular sport, soccer, while it is permitted in horse racing. In the U.K., gambling is allowed in basically all sports, in parlors, online, and in stadiums.

2.3. Casino and Other Games Around the World

Casino games are too numerous to list, with a wide variety among establishments, regions, and countries. Here are some main categories and games.

- Gaming machines. Slot machines are quite popular universally, but there are some country-specific machines such as *pachinko* in Japan. These require neither casino workers nor opponents, providing quick and easy access to gambling. Video poker and other video games can also be included in this category.

- Table games. All card games like blackjack, baccarat, and poker are table games, and so are craps and roulettes. These require one or more players competing against the "house" (casino), not necessarily against each other.
- Random number games. These are games where the winners are chosen by random number selection. Bingo and keno are examples of this type. The fairness of such games depends on the randomness of the number generator or the equipment used to pick the winning numbers.

3. Mathematical Probabilities and Gambling

The Law of Large Numbers (also known as the "Law of Averages") is perhaps the most fundamental principle in the mathematics of gambling. It is mentioned in Sec. 4.4 and states the following:

> "When an independent trial is repeated many times, the actual probability that some event occurs approaches its theoretical (mathematical) probability."

In other words, over a certain (long enough) period of time, a particular event will occur at a frequency predicted theoretically using mathematics. For instance, if you play a game in which you win 30% of the time, you may or may not win the first game or the second. But if you play that game 100 times, you win roughly 30 of the games (but may not be exactly 30). If you flip a fair coin 1000 times, you get approximately 500 tails.

As discussed in Sec. 4.4, this is the basis for calculating the expected value, which leads to "house edge," the statistical advantage that the casino has, i.e. how much the casino expects to gain over a long period of time. This idea, along with the more general theory of probability, has been studied over the last few centuries all over the world, by leading mathematicians like Gerolamo Cardano (Italy), Pierre de Fermat, Blaise Pascal (France), Jakob Bernoulli (Switzerland), Shizuo Kakutani (Japan), Paul Erdös (Hungary), William Feller (Austria-Hungary, now Croatia), and Andrey Kolmogorov (Russia).

Questions:

1. Gambling has existed wherever and whenever there were people on the earth. Pick a country or a region of the world, and study the history of gambling there, investigating what types of items and games were used, how popular it was, and who participated.
2. There are different theories concerning the invention of the roulette wheel. Some claim that Pascal invented the roulette wheel in the 17^{th} century. Others do not agree. The modern roulette wheel did not appear until the mid-19^{th} century. Write a brief history of roulettes.
3. Today there are two distinct types of roulettes in the world: European and American. Do some research and describe the difference as well as how the difference affects probabilities and house edge.

4. Study the history and practice of state- or municipal-sponsored lotteries around the world. What types of games are available? What does the government or municipality use the profit for? What are some pros and cons of public gambling?
5. Pick one sport popular in the world. Do some research on global gambling on that sport. How extensive (how much money) is the gambling related to that sport? What countries are the major players?
6. Fill in the blank on this famous statement: “The only way to win at a casino is to _____.” Do you agree or disagree? Why? What needs to be considered besides the house edge and the Law of Averages to answer this question?
7. Give some examples of parlays involving Premier League soccer matches in the U.K. Pick one example of a multi-game parlay and another of a same-game parlay.

10.19 Project 19: Fields Medal and Its Winners

Mathematicians Recognized Around the World

Category: History of Mathematics

Description:

We examine the origin and the history of the Fields Medal as well as the winners of this prestigious prize.

Introduction:

The Fields Medal is often called the "Nobel Prize of Mathematics," referring to the fact that it is one of the highest honors a mathematician can receive (albeit with some significant differences from the Nobel Prize as described below). Who has won this award, and what did they do? What countries do they represent? Who is "Fields" anyway? Let's explore these questions from an international perspective.

1. Purpose and History of the Fields Medal

1.1. Eligibility and Conditions

The Fields Medal is said to be equivalent to the Nobel Prize as it represents one of the highest internationally recognized awards in a specific discipline. However, there are major differences. The Fields Medal is awarded every four years (possibly to multiple awardees) while the Nobel Prize is given every year. The monetary award is considerably less than that of the Nobel Prize, and it does not get nearly as much media attention.

The biggest difference may be its purpose. The Fields Medal was established in order to recognize and support young research mathematicians with major contributions; therefore, it is given only to those who are under the age of 40. In addition, once you receive a Fields Medal, you will not be eligible for another. The age limit is consistent with the wish of the Canadian mathematician John Charles Fields, for whom the medal is named. He wanted to recognize the accomplishments of young researchers and encourage further academic pursuits.

So far, the youngest recipient has been Jean-Pierre Serre of France, who won the Fields Medal in 1954 at age 27.

1.2. Origin and History

John C. Fields had begun his plans to establish the award, but he passed away in 1932 in Toronto before the first medal bearing his name could be awarded. Despite the fact that Fields specifically requested that the medal NOT be named after him, the award has become known simply as the "Fields Medal."

Supported by the funds left by Fields (an endowment from his estate), the International Mathematical Union (IMU) awarded the first Fields Medals in 1936 to Lars Ahlfors (Finland) and Jesse Douglas (the United States). Except under some extenuating circumstances, such as wars, conflicts, and pandemics, the Fields Medal is given at the International Congress of Mathematics (ICM), which is held every four years. A total of 64 people have won the Fields Medal as of 2024.

1.3. The Medal

The medal itself is made of 14-karat gold, depicting Greek mathematician and philosopher Archimedes, along with his name written in Greek. Also, the Latin inscription on the front roughly states, "Surpassing one's understanding and mastering the world." The back side states, "Mathematicians have gathered from the whole world and awarded [this] for outstanding writings" along with a background picture of an olive branch and a sphere inscribed in a cylinder (representing Archimedes' tomb).

2. Recipients of the Fields Medal

2.1. Selection

The winners of the Fields Medal are chosen by a committee within the IMU in accordance with the strict rules of the prize and its requirements. Since it is given only every four years, the committee is encouraged to pick four recipients from a variety of mathematical fields. Under this policy, the list of all Fields Medal winners represents the breadth of mathematics while, at the same time, reflecting what is "trendy" in mathematical research as seen below.

2.2. Mathematical Areas Recognized

With more than 60 individuals awarded the Fields Medal thus far, it is not surprising that a wide variety of mathematical fields (pun intended) is represented. Their areas of expertise include geometry, analysis (real and complex), abstract algebra, number theory, topology, set theory, knot theory, combinatorics, mathematical physics, chaos theory, and operator theory. Some areas combine a few, such as geometric topology, statistical physics, and algebraic geometry. Sometimes these researchers' work opens up entirely new areas of research, combining technical ideas and tools as well as pioneering new disciplines. Some areas begin as mathematicians get interested out of pure curiosity without any regard for real-world problems, only to discover that the new field has many surprising applications in a variety of disciplines. For example, number theory turns out to be fundamental in encryption and cybersecurity. Knot theory has seen many applications in scientific fields such as biology and genetic engineering.

In 1982, Edward Witten became the first physicist to receive the Fields Medal (for his remarkable work in applied mathematics).

2.3. "Silver Plaque"

The age limit caused a slight problem in 1994 with the English mathematician Andrew Wiles, who proved one of the most famous mathematics problems in history—Fermat's Last Theorem, which had not been solved for centuries. He presented his ground-breaking proof of this theorem in June 1993 (two months after his 40th birthday), in Cambridge. However, a gap (an error) was found about three months later. Wiles, working with his collaborator Richard Taylor, eventually (and profoundly) fixed the problem a year later (on September 19, 1994, to be precise). A revised proof was submitted in October 1994 and published in a prestigious mathematics journal in May 1995.

His proof solved a 358-year-old mystery in number theory!

But do you see the problem here? While a large portion of his work, significantly valuable in and of itself, was achieved before he turned 40, this major milestone was not announced until he was 40 and was not fixed until he turned 41 (and not published until he was 42). Because of his monumental work leading up to this great proof, he was considered a favorite to win the Fields Medal in 1994 (though even then he was slightly over the age limit). However, the discovery of the gap did not help. He missed the 1994 prize given in Zurich.

But because of the importance of his marvelous proof and its historical significance, the Fields Medal Committee recognized Wiles, not with a Medal, but with the first-ever "silver plaque" in 1998. Whatever Fermat had in mind when he wrote his quick comment (what is called the "most famous marginal note in history"), he could not have imagined such a dramatic end to the story of his "theorem."

2.4. Some Highlights

Naturally there are geopolitical and other global events that could cause interruptions in the award presentation. The Fields Medal is generally given at the ICM as mentioned above; the award ceremonies have been held in many countries and regions including Europe, the U.S., the U.S.S.R., Canada, Poland, Japan, China, India, Korea, and Brazil.

Just like athletes in the Olympic games, some people boycotted award ceremonies while others were not allowed to go to certain countries to receive their awards. In 1966, the German-French mathematician Alexander Grothendieck refused to attend the ICM in Moscow, protesting Soviet aggressions in Eastern Europe. During the Cold War, the government of the U.S.S.R. refused to let two of its recipients travel to receive the prize: Sergei Novikov (1970 in France) and Gregory Margulis (1978 in Finland). The 1982 ICM, scheduled to be held in Warsaw, was postponed due to the country's martial law which went into effect in December 1981.

In 2022, the global COVID-19 pandemic forced the award to be presented online in an event live-streamed from Finland. Originally this ceremony had been planned at St. Petersburg in Russia, but the plan changed after Ukraine was invaded in February 2022.

Questions:

1. Look up the list of all Fields Medal winners in history. There is one (and only one) who *declined* to receive the Fields Medal. Who is he, and why did he decline? What did he decide to do after he refused to accept this honor? (And where is he now?)
2. Write an essay on the life and work of Maryam Mirzakhani, the first female recipient of the Fields Medal (2014). Where was she from? What did she accomplish as a teenager?
3. You may find it strange that mathematics is not included as a Nobel Prize discipline, such as physics, chemistry, physiology/medicine, and literature. No one knows exactly why, but there are some speculations. Find some possible reasons why Alfred Nobel did not include mathematics. See if you can find any myth involving a Swiss mathematician named Gösta Mittag-Leffler. Is the myth supported by evidence?
4. Not to compete with the Fields Medal, there is another very prestigious honor in mathematics called the Abel Prize, named after Niels Henrik Abel of Norway. Answer the following:
    a. What did Abel do? Write a brief essay on the exciting (yet tragic) life of Niels H. Abel.
    b. What are some differences between the Fields Medal and the Abel Prize?
    c. Who has won the Abel Prize (and what countries are they from)? This list is not that long since the Abel Prize was not given until 2003.
5. We already mentioned that Maryam Mirzakhani became the first woman to receive the Fields Medal (2014). Who was the first recipient
    a. From South America (2014)?
    b. Of Indian origin (2014)?
    c. Of Korean ancestry (2022)?
    d. From Ukraine (2022)?
6. The picture depicted on the reverse side of the Fields Medal has a sphere and a circular cylinder, representing what Archimedes called his “most proud” accomplishment. What exactly did he prove by the sphere inscribed in the circular cylinder? How many years ago did he make this discovery?
7. Answer the following questions concerning Fermat’s Last Theorem (for which Andrew Wiles received special recognition as described here).
    a. What does this famous theorem say? Does the statement make sense to you? Try to paragraph it in informal language.
    b. What exactly was the “marginal note” Fermat left in a book? What significance does it have?
    c. Technically, Fermat’s Last Theorem was not a “theorem.” When did it become a theorem? Why?

10.20 Project 20: Fibonacci, Leonardo of Pisa

Thanks, Leo!

Category: History of Mathematics

Description:

We examine the life and work of Fibonacci, a pre-Renaissance Italian mathematician, also known as “Leonardo of Pisa.”

Introduction:

The name Fibonacci is primarily associated with the famous sequence that bears his name. But he did a whole lot more than just introducing that sequence, which he used to count the number of rabbits. Here, we examine the lesser-known life and work of Fibonacci, Leonardo of Pisa. While his career is buried in the shadow of the great mathematicians who came after him, during the Renaissance and beyond, the accomplishments of Fibonacci cannot be understated. His work has advanced and illuminated many fields of mathematics, and the results exemplify what can be referred to today as “internationalization” in mathematics. It is fitting, therefore, that we conclude this international/global chapter of the book with the life and work of a genuinely global mathematician.

1. Early Life of Fibonacci

It is widely believed that Fibonacci was born around 1170 to an Italian merchant father Guglielmo, who lived in the Tuscan city-state, or the Republic, of Pisa. Yes, it is the place known for its iconic “Leaning Tower.” In fact, the very foundations of the tower were laid in 1173 (evidently with a slight angle), making that tourist attraction almost exactly as old as Fibonacci himself.

Sometime after Fibonacci’s birth, Guglielmo moved to Bugia, a Mediterranean port city in Algeria (on the North African coast) as an Italian customs officer. This turned out to be a defining move for Fibonacci, for it was there that he studied science and mathematics under an Arab scholar. Remember, the world’s capital of science and mathematics shifted to the Hall of Wisdom in Baghdad during the Middle Ages, and the Islamic world dominated scholarly activities, especially in mathematics. It is therefore not surprising that Fibonacci had an Arabic instructor.

Having received a great education there, he began traveling around the Mediterranean coast as a young man, meeting merchants from a variety of countries. Where did he visit? Many of the Mediterranean regions including Egypt, Greece, Sicily, Provence, and even as far east as Syria. An avid traveler indeed.

2. His Bestseller

What makes Fibonacci famous today is the sequence that bears his name. Starting with 1, 1, each term of the sequence is the sum of the two previous numbers: {1, 1, 2, 3, 5, …}. Fibonacci, according to his own writing, was using this sequence to count the theoretical number of rabbits, but the fact is that this sequence had been known for centuries, even in the 2nd or 3rd century B.C.E. in the works of Acharya Pingala, an ancient Indian mathematician, studying (of all things) the syllable meters (patterns) in Sanskrit poetry! Fibonacci then introduced it in 1202 in his book *Liber Abaci*, perhaps one of the most consequential mathematics books in the Middle Ages. (Remember, this was before the printing press was invented, so it had to be copied by hand.)

In *Liber Abaci* ("Book of Calculation"), Fibonacci introduced something *very* significant to the Western world: the Hindu-Arabic numeral system, which had been introduced to him by his Arabic teacher back in Algeria. This system was far superior to and more practical than the Roman numerals used throughout the Roman Empire. It is like our current system with place values. In that book Fibonacci called the new system *modus Indorum* (method of the Indians). It included 0 as a place holder! Now, imagine someone asking you a simple arithmetic question, in Roman numerals, "What is $XLII - XXIX$?" Can you give the correct answer and say, "$XIII$"? Without place values, arithmetic operations would be much more complicated. This new system, very little known in the West before that time, helped the Europeans make significant progress in banking, commerce, and trade.

This book also became a great reference book as Fibonacci showed how to compute such relevant and practical things as profit margins and interest as well as convert currencies using exchange rates.

In addition, Fibonacci also discussed in *Libre Abaci* certain principles of pure mathematics, such as the existence of irrational numbers and the importance of prime numbers. Not bad for a book written before the Renaissance (whoever calls Fibonacci's period the "Dark Ages" should reconsider that term)!

3. More Books

Much of the knowledge Fibonacci mentioned in *Libre Abaci* appears to be what he had learned as a student in Algeria. But that knowledge had originated in ancient Greece and Rome (yes, Italy), then transferred to Alexandria (the city with a great ancient library) in Egypt, to the Islam world, and now it has come around in a full circle, right back to the Mediterranean coast of Italy through a son of an Italian merchant.

Incidentally, this book was widely copied and read, like today's bestseller. In fact, it got the attention of the Holy Roman Emperor Frederick II, who invited Fibonacci as a guest in his court. Evidently Frederick II enjoyed science and mathematics quite a bit (as all emperors and kings should). This led to a series of many regular conversations with the emperors' scientists and scholars.

In 1220, Fibonacci wrote another book, *Practica Geometriae*, in which he stated many theorems of geometry in the style of Euclid, along with practical problems that can be solved using geometry and trigonometry. Then, five years later, in 1225, he published another book, *Liber Quadratorum* (Book of Square Numbers), and dedicated it to Frederick II. This book is considered Fibonacci's masterpiece, a major contribution in number theory, long before Pierre Fermat was born (in 1601).

Remember, back then there was no unified country called Italy. His home, Pisa, was its own republic, a city-state. In 1240, the government of Pisa publicly recognized the accomplishments of Fibonacci.

4. His Legacy in the History of Mathematics

Many historians agree that Fibonacci, or Leonardo of Pisa, was the most significant, talented, and crucial mathematician in the Middle Ages. The name Fibonacci, however, was not used until much later; it appears that he was not referred to as Fibonacci until 1506 when he was mentioned as "Lionardo Fibonacci" in the Holy Roman Empire and again in 1838 simply as "Fibonacci" as a short form for the "son of Bonacci."

A statue of Fibonacci was erected in the 19th century, and it can be seen in the Camposanto, a cemetery on the Piazza dei Miracoli ("Square of Miracles") in Pisa, the same location as the famous Leaning Tower.

Today, there is a scientific journal, *The Fibonacci Quarterly*, dedicated to articles and studies on the Fibonacci sequence. It has been published by the Fibonacci Association since 1963. Numerous new results and proofs have appeared in academic papers submitted to this journal from all over the world.

So, you see, in a sense, mathematics has always grown, thrived, and flourished through internationalization, thanks to hard-working and diligent mathematicians like Fibonacci, who lived over eight centuries ago.

Questions:

1. Go back to Sec. 7.2 to review what the Fibonacci sequence is and how it is related to the Golden Ratio. It turns out that Fibonacci never mentioned the Golden Ratio. Who discovered the link between this sequence and this ratio? Do some online research to answer.
2. Look up and study "Fibonacci's Rabbits," the original problem posed by Fibonacci in introducing the sequence. Explain the problem and how the Fibonacci sequence provides the answer to that question.
3. Write an essay on the history of the Fibonacci sequence, taking into account some ancient evidence that the sequence was known long before Fibonacci wrote about it.

4. Explain, using examples, why the Hindu-Arabic numeration system is superior to the Roman numeral system. Be sure to mention what “place values” mean.
5. In Sec. 7.2, references are made to the Golden Spiral and how it can be drawn using the Fibonacci sequence. Explain that process using your own words.
6. People have acknowledged numerous ways in which Fibonacci numbers appear in nature. List a few examples.
7. Do some research on how music (keys, instruments, classical compositions) is related to the Fibonacci sequence, Fibonacci numbers, and/or the Golden Ratio.

Appendix A - Try It Now Answers

Chapter 1 Try It Now Answers

1. Fallacy types are not comprehensive or mutually exclusive, so these are not the only correct solutions.
    a. Circular reasoning
    b. Correlation-causation (This is not *post hoc,* or false cause because one did not precede or follow the other.)
    c. *Post hoc* (false cause)
    d. Appeal to consequence
    e. Straw man—distorted representation of an opinion easier to attack
    f. False dilemma (limited choice)—she might like neither coffee nor tea.

2. Be careful with the quantifiers.
    a. "Some Icelandic children do not learn English in school." or "There is at least one Icelandic child who do not learn English in school."
    b. "Some French children speak English." or "There is at least one French child who speaks English."
    c. "No Brazilian children speak Japanese."

3. $C \vee \sim S$

4.

A	***B***	***~A***	***~A ⋀ B***
T	T	F	F
T	F	F	F
F	T	T	T
F	F	T	F

5.

A	***B***	***~A***	***~A ⋀ B***	***~B***	***(~A ⋀ B) ⋁ ~B***
T	T	F	F	F	F
T	F	F	F	T	T
F	T	T	T	F	T
F	F	T	F	T	T

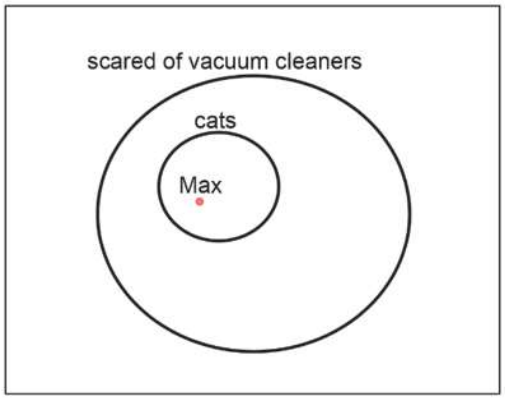
6. (b) is the contrapositive, so that is the answer. ((a) is the inverse, and (c) is the converse.)

7. (b) is the negation. Note that the negation contains neither "if" nor "then."

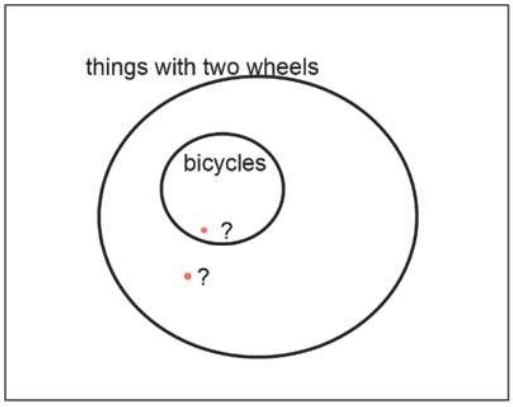
8. (a) and (b) are false; (c) and (d) are true.

9. The negation of “born as a citizen AND at least 35” is “EITHER not born as a citizen OR not at least 35.” Therefore, the conditions for disqualification is this: “Either you were nor born as a U.S. citizen or you are younger than 35 years old.”

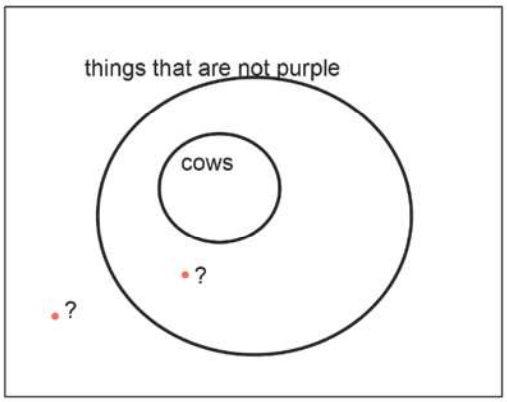
10. This argument is valid.

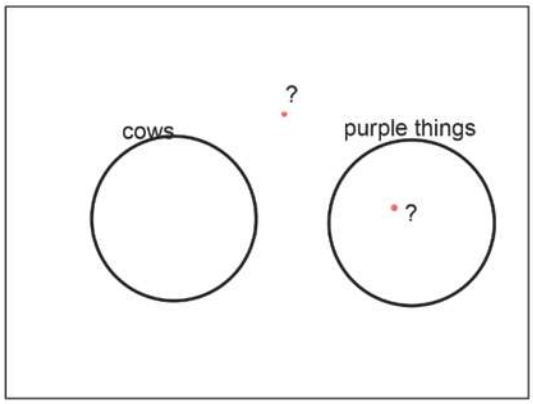


11. This argument is invalid. The fact that “this Harley-Davidson” has two wheels would put this Harley in the *outside* circle, not necessarily in the *inside* circle as shown.

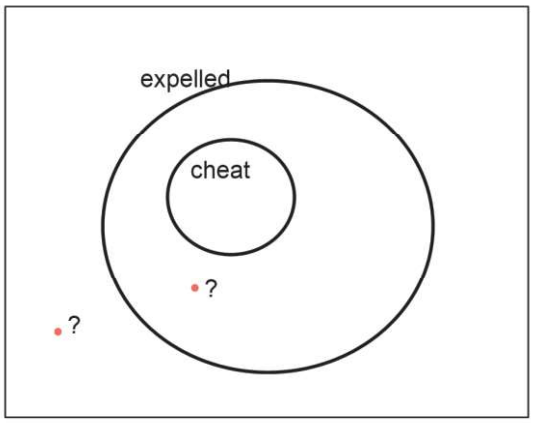


12. This argument is invalid. Because Fido is not a cow, Fido is outside of the *cow* circle, but we do not know if Fido is in or out of the purple circle (or non-purple circle). Either of these diagrams will show the invalidity.

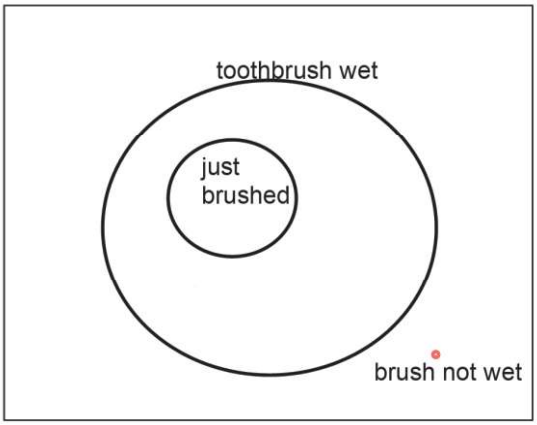




13. This argument is invalid. The fact that "you did not cheat" will put you outside the *inner* circle, but we are not sure if you are in or out of the *outer* circle.



14. This argument is valid by *modus tollens* (the Law of Contraposition). The following diagram also shows the validity. You are *outside* of the outer circle, so you cannot be in the *inner* circle.



15. Valid by the Transitive Property (hypothetical syllogism).

16. Valid by disjunctive syllogism.

17. Invalid by the Fallacy of the Converse.

18. Invalid by the Fallacy of the Inverse.

19. Invalid. If you go to the party, you will get to see your friends but will be really tired tomorrow. Now, if you do not see your friends (as stated in the antecedent of the last statement), then that would mean that you did not go to the party. However, that does not imply that you will not be tired tomorrow. You may be tired from something else. This last part is the Fallacy of the Inverse.

20. She has 180 chickens and 70 goats.

21. The 300th number is 1203. The first number is 3+4, and each time you go to the next one, 4 is added. So the 300th number will have 4 added 300 times to 3, so it is 1203.
22. Stephanos is an obnoxious 14-year-old. Multiply 436 by 3 to get 1308. Then, subtract 6 to get 1302. Finally, divide it by 93.
23. There are 9 ways as shown below.

Dimes	Nickels	Pennies
0	0	23
0	1	18
0	2	13
0	3	8
0	4	3
1	0	13
1	1	8
1	2	3
2	0	3

Chapter 2 Try It Now Answers

1. First, 30% of \$800 is \$240, so there is a \$240 discount. The sale price is \$560. Then, the sales tax is $560 \times 0.092 = 51.52$. This amount is added to 560 for the total. The answer is \$611.52.

2. [2016 to 2017] The absolute change: $20.24 - 19.57 = 0.67$ (trillion), which is \$670 billion. The relative increase is $0.67 \div 19.57 \approx 0.034236$ …, or about 3.42%.

   [2019 to 2020] The absolute change: $26.95 - 22.72 = 4.23$ (trillion), which is \$4.23 trillion. The relative increase is $4.23 \div 22.72 \approx 0.18618$ …, or about 18.6%. Therefore,

2019 to 2020 saw a much larger increase in federal debt, both in the absolute change and in the percentage change.

3.

1. [Virginia and Mississippi]
The absolute difference is $8{,}518{,}000 - 2{,}987{,}000 = 5{,}531{,}000$.
$5{,}531{,}000 \div 8{,}518{,}000 \approx 0.649$. Mississippi has 64.9% less people than Virginia.
$5{,}531{,}000 \div 2{,}987{,}000 \approx 1.852$. Virginia has 185.2% more people than Mississippi.
$2{,}987{,}000 \div 8{,}518{,}000 \approx 0.351$. The population of Mississippi is 35.1% of that of Virginia.

$8{,}518{,}000 \div 2{,}987{,}000 \approx 2.852$. The population of Virginia is 285.2% of that of Mississippi.

[Michigan and Wyoming]

The absolute difference is $9{,}996{,}000 - 578{,}000 = 9{,}418{,}000$.

$9{,}418{,}000 \div 9{,}996{,}000 \approx 0.942$ Wyoming has 94.2% less people than Michigan.
$9{,}418{,}000 \div 578{,}000 \approx 16.294$. Michigan has 1,629.4% more people than Wyoming.
$578{,}000 \div 9{,}996{,}000 \approx 0.0578$. The population of Wyoming is 5.78% of that of Michigan.

$9{,}996{,}000 \div 578{,}000 \approx 17.294$. The population of Michigan is 1,729.4% of that of Wyoming.

2. [California and Wyoming]
$578{,}000 \div 39{,}557{,}000 \approx 0.0146$. The population of Wyoming is 1.46% of that of California.

$39{,}557{,}000 \div 578{,}000 \approx 68.438$. The population of California is 6,843.8% of that of Wyoming. This means that California is about 68.4 times as populous as Wyoming.

4. First, note that 19.8 pounds is $19.8 \times 16 = 316.8$ ounces. Then, $316.8 \div 1000 = 0.3168$ (ounces per foot) is the unit weight of the wire. Now, note that 18 inches is 1.5 feet. So we have $0.3168 \times 1.5 = 0.4752$ (ounces).

5. We set up the equation

$$\frac{8 \times 8 \times 1(ft^3)}{48\ (\text{bags})} = \frac{6 \times 4 \times 1\ (ft^3)}{x\ (\text{bags})}$$

This suggests $x = 48 \times 24 \div 64 = 18$ (bags). So you will need 18 fifty-pound bags of sand. The left-hand side is saying that 1.333… ft^3 an be filled per bag. So you can also find the answer by $24 \div 1.3333 \ldots = 18$ (bags).

Chapter 3 Try It Now Answers

1. Use the annuity formula:

$$S_{10} = 150 \times \frac{\left(1 + \frac{0.03}{12}\right)^{12\times 10} - 1}{\left(\frac{0.03}{12}\right)} = 150 \times \frac{(1.0025)^{120} - 1}{0.0025}$$
$$= 150 \times 139.7414189 \ldots$$

This gives you $S_{10} = \$20{,}961.21$. Since your total payments will be $150 \times 120 = 18{,}000$, the total interest earned will be $\$20{,}961.21 - \$18{,}000 = \$2{,}961.21$.

2. Use the loan formula:

$$3{,}000 = d \times \frac{1 - \left(1 + \frac{0.16}{12}\right)^{-12\times 2}}{\left(\frac{0.16}{12}\right)} = d \times 20.423539.$$

This gives us $d = 3000 \div 20.423539 = 146.89$.

So Ciara will be paying $146.89 a month for 24 months, which may not sound so bad. However, her total payments will be $\$146.89 \times 24 = \$3525.36$. The total interest paid for the 2-year period will be $\$3525.36 - \$3000.00 = \$525.36$.

3.

a. Dylan is saving, setting aside $200 every month for 3 years, for his future car. This is an annuity problem.

$$S_3 = 200 \times \frac{\left(1 + \frac{0.03}{12}\right)^{12\times 3} - 1}{\left(\frac{0.03}{12}\right)} = 200 \times 37.62056 \ldots \approx \$7{,}524.11.$$

b. The company wants to invest money *now* (once) for the future, not a series of deposit. This is a compound interest problem with one deposit. They want $30,000 in 6 years, with the money earning 5% annually.

$$A_6 = 30{,}000 = P \times (1 + .05)^6 \approx 1.34009564\ P.$$

$$P = 30{,}000 \div 1.34009564 = 22{,}386.46.$$

So the non-profit company must deposit $22,386.46 *today* to have $30,000 in 6 years.

c. Brittany is paying $100 a month for 4 years in her loan, so this is clearly a loan problem.

$$P = 100 \times \frac{1 - \left(1 + \frac{0.02}{12}\right)^{-12\times 4}}{\left(\frac{0.02}{12}\right)} = 100 \times 46.0933 = 4609.33.$$

She can borrow $4,609.33 today.

d. Here, you are to save money every month for two years, so this is a savings annuity problem.

e.

$$S_2 = 5000 = d \times \frac{\left(1 + \frac{0.04}{12}\right)^{12\times 2} - 1}{\left(\frac{0.04}{12}\right)} = d \times 24.942888 \ldots.$$

$$d = 5000 \div 24.942888 \ldots = 200.46.$$

You would have to set aside $200.46 every month for two years to save up $5000.

Chapter 4 Try It Now Answers

1. On a digital clock, the minutes reading can be anything from :00 to :59. There are 60 possible outcomes here.
   a. The probability that it is :15 is 1/60.
   b. "Less than :15" means from :00 to :14, which consists of 15 outcomes. Thus, the probability is 15/60, or 1/4.

2. In this case, the first card is placed back, and the deck is shuffled, making the two trials independent. Therefore, the probability is

$$P(\text{two aces}) = \frac{4}{52} \times \frac{4}{52} = \frac{1}{13} \times \frac{1}{13} = \frac{1}{169}$$

3. For both pairs to be white, the first must be white, and the second one needs to be white *given* that the first was white. So the probability is

$$P(\text{1st is W}) \times P(\text{2nd is W} \mid \text{1st is W}) = \frac{6}{10} \times \frac{5}{9} = \frac{1}{3}$$

Here, this second probability is 5/9 because after the first white pair has been removed, there are 5 white pairs left out of 9.

4.

	Tested Positive (P)	Tested Negative ($\sim P$)	Total
Disease (D)	50	0	50
No Disease ($\sim D$)	398	9,552	9,950
Total	448	9,552	10,000

If we start with 10,000 people, only 50 of them (0.5%) have the disease. All 50 of them will test positive. Out of the 9,950 who do not have the disease, 4% will test positive; that is 398 people. So there will be a total of 448 who will test positive (the first column), and only 50 of them actually have the disease. So the probability that a person who tests positive has the disease is 50/448, or 11.16%.

5. By the Multiplication Rule, $8 \times 11 \times 5 = 440$ different meal options are available.

6. There are 26 letters in the alphabet.
   a. If repetition is allowed, $26^5 = 11{,}881{,}376$.
   b. If no repetition is allowed, $26 \times 25 \times 24 \times 23 \times 22 = 7{,}893{,}600$.

   ${}_{13}C_3 = \frac{13 \times 12 \times 11}{3 \times 2 \times 1} = 286$ combinations.

7. This is a tough question, but it is possible to think our way through this. First, imagine someone guessing all the questions. Each question has 5 choices, independent of one another, and there are 10 questions. Hence, there are $5^{10} = 9{,}765{,}625$ ways to answer this quiz. Now, if you are to get only Problem 1 incorrect, you have to pick one incorrect answer for that question (4 choices), followed by 9 correct answers (only 1 way to do that). Hence, there are 4 ways to answer this quiz so that you get Problem 1 incorrect but everything else correct. Note that the same argument works for missing Problem 2 only. Since you can miss exactly one of the questions, and there are 4 ways to do that for *each* question, there are 40 ways to answer this quiz so that you get exactly 9 questions right. Hence, the answer is $\frac{40}{9{,}765{,}625}$, which is 0.0004096%.

   (This question is on what is known as a **binomial probability**. As such, you can obtain the answer by calculating ${}_{10}C_1\ (0.2)^9(0.8)^1 = 0.000004096$).

8. As described in the text, there are ${}_{52}C_5 = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} = 2{,}598{,}960$ ways in total to choose 5 cards from a deck of cards. To get three out of the four aces, there are ${}_4C_3$

$=\frac{4\times3\times2}{3\times2\times1}=4$ ways, and to get two kings out of the four, there are ${}_4C_2=\frac{4\times3}{2\times1}=6$ ways. So the probability is $\frac{4\times6}{2{,}598{,}960}\approx 0.00000923$, which is about 0.000923%.

9. The probability that all ten have different birthdays is

$$\frac{365}{365}\times\frac{364}{365}\times\ldots\times\frac{356}{365}=\frac{{}_{365}P_{10}}{365^{10}}=0.883052.$$

So the probability that at least two share the same birthday is $1-0.883052\approx 0.117$, or 11.7%.

10. If you purchased all 2000 tickets, you would pay $10,000 altogether and win $4000. That is a net loss of $6,000. Divide this loss by 2000, we get a loss of $3 per ticket. This is the expected value. In terms of a weighted sum, the winner pays $5 and wins $4000, so the winner's net gain is $3995, with the probability 1/2000. So the expected value is

$$3995\times\frac{1}{2000}+(-5)\times\frac{1999}{2000}=-3.$$

11. There are altogether $6^3=216$ outcomes, but there are only $6\times5\times4=120$ ways in which the dice will all have different (distinct) values. So the probability that you lose $1 is $\frac{120}{216}=\frac{5}{9}$ (slightly more than half). The probability that you win $2 is $1-\frac{5}{9}=\frac{4}{9}$; the expected value is
12.

$$\left(-1\times\frac{5}{9}\right)+\left(2\times\frac{4}{9}\right)=-\frac{5}{9}+\frac{8}{9}=\frac{3}{9}=\frac{1}{3}$$

So every time you and your friend play this game, you can expect to win 33.3… cents.

You can also imagine playing this game 90 times. You will lose $1 approximately 50 times but win $2 approximately 40 times. Hence, $-50+80=30$ (dollars) will be your winnings. Dividing this by 90, you get 1/3.

Chapter 5 Try It Now Answers

1. Since the college would have access to data on all students (population), this is probably a parameter. However, if this value is a sample average, it would be a statistic.

2.
   a. Color is categorical (qualitative).
   b. Quantitative. Temperature is numerical, and an average can be calculated meaningfully.
   c. Quantitative. Income is numerical, and an average can be calculated meaningfully.

d. Occupation is categorical (qualitative).

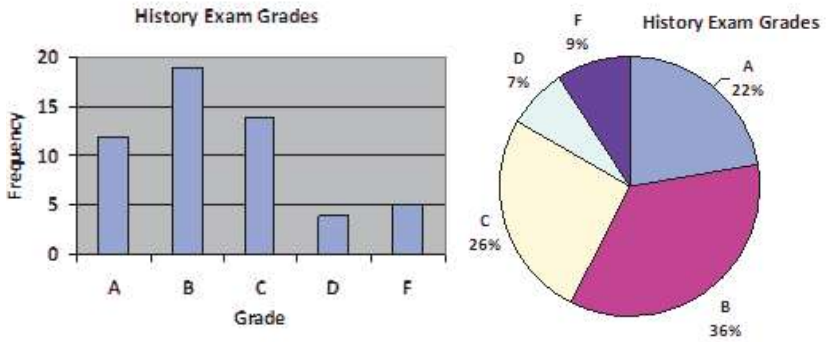
3.

a. Systematic ("every 4th person")
b. Stratified
c. Voluntary-Response
d. Random
e. Cluster

4. The answers may vary.
   a. Voluntary-response bias (some may not respond), response bias (response may not be entirely accurate since it is very personal), perceived lack of anonymity, and possibly non-response bias (many may refuse to answer this).
   b. Voluntary-response bias (Participants are self-selected.)
   c. Sampling bias (the sample may not be representative of the whole class), response (some may not be honest)
   d. Lack of anonymity (since under-age drinking is against the law), non-response (many may refuse to answer)
   e. Self-interest
   f. Loaded question

5. The difference is whether or not one group is "influenced" by the researcher (e.g. given a treatment):
   a. Observational. Measurement only.
   b. Experiment. One group is asked to do jumping jacks.
   c. Observational since they seem to be natural coffee drinkers and tea drinkers. (Later in this chapter, we will study a type of observational study in which two naturally defined groups are compared—these are called "case-control" studies. This particular example about coffee and tea drinkers is a classic case of such studies. Technically, this is observational since neither group is "influenced," but this has a hint of an experiment since two groups are compared.)

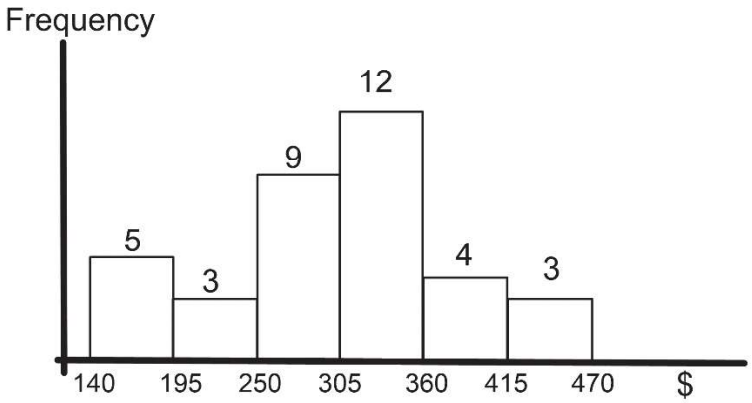
6.



7. While the pie chart seems to accurately depict the percentage of the people agreeing with each candidate, the chart is confusing since the percentages do not add up to 100%. Typically, a pie chart represents 100%.

8. First, create a table of the classes. Then, construct a histogram.

Cost Interval	Frequency
$140-194	5
$195-249	3
$250-304	9
$305-359	12
$360-414	4
$415-469	3



9. $\frac{3.29+3.59+3.79+3.75+3.99}{5} = \frac{18.41}{5} = 3.682$. The mean is $3.682.

10. Order from the smallest: $3.29, $3.59, $3.75, $3.79, $3.99. The median is $3.75.

11. Extending the table, we get the following:

Rating	Frequency	Product	Cumulative Freq.
1	4	4	4
2	8	16	12
3	7	21	19
4	3	12	22
5	1	5	23
Total	23	58	

The mean is $58/23 \approx 2.52$, the median is 2 (the 12th value), and the mode is 2 (frequency 8).

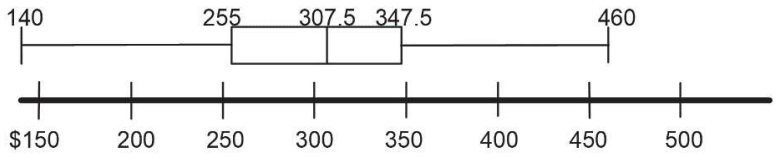
12. A table will help when calculating the standard deviation by hand (calculator). We have already calculated that the mean is \$3.682 (Try It Now 9), so the deviation is the difference between each price and this mean. Note $n = 5$ here since there are 5 prices.

Price	Deviation	Deviation^2
3.29	-0.392	0.153664
3.59	-0.092	0.008464
3.79	0.108	0.004624
3.75	0.068	0.011664
3.99	0.308	0.094864
Total		0.27328

Now, dividing 0.27328 by $n - 1 = 4$, we get the variance, 0.06832. Taking its positive square root, we see that the standard deviation is 0.261381.

13. These 36 values are already ordered. The median is the mean of the 18th and the 19th values: $\frac{305+310}{2} = 307.5$. Hence, Q2 is 307.5. There are 18 values in the lower half, so its median is the mean of the 9th and the 10th values: $\frac{250+260}{2} = 255$. This is Q1. There are also 18 values in the higher half, and its median is the mean of the 9th and the 10th values of the top half: $\frac{345+350}{2} = 347.5$. This is Q3. Of course, the minimum is 140 while the maximum is 460. Hence, the five-number summary is "140, 255, 307.5, 347.5, 460."

14. Based on Try It Now 13, we have this boxplot.



Chapter 6 Try It Now Answers

1. Since 1986 is the first year for which data is given, let 1986 be $n = 0$. Since 2020 corresponds with $n = 34$, we see that the average annual growth is $\frac{53{,}555-20{,}610}{34} \approx 968.97$. The explicit formula then is $P_n = 20{,}610 + 968.97n$. Now, since 2040 is when $n = 54$, we have $P_{54} = 20{,}610 + 968.97 \times 54 = 72{,}934.38$, approximately 72,934 companies.

2. If we let 2019 be $n = 0$, then 2030 corresponds to $n = 11$. The population will then be

$$P_{11} = 1.366 \text{ billion} \times (1.0134)^{11} \approx 1.366 \text{ billion} \times 1.15768 \approx 1.5814 \text{ billion}.$$

This may not look much bigger, but the increase in 11 years will be over 200 million people.

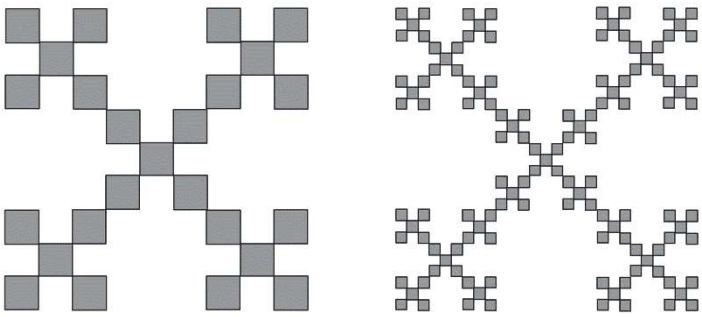
3. In 8 months, the number increased from 45,000 to 60,000. Using n as the number of months since going public, we see that the explicit formula gives us $P_8 = 45{,}000\ (1+r)^8 = 60{,}000$. Dividing both sides by 45,000, we get $(1+r)^8 = 60{,}000 \div 45{,}000 = 1.3333\ldots$ At this point, you can find the r by trial-and-error or take the 8th root of each side to obtain $1 + r \approx 1.0366$. So $r \approx 0.0366$ (meaning exponential growth by about 3.66% per month). Two years corresponds to $n = 24$, so the number of users will be $P_{24} = 45{,}000\ (1.0366)^{24} \approx 106{,}630$.

4. The answer is 9 since $10^9 = 1$ billion. So $y = 9$. (Be careful *not* to say that $y = 10^9$.) For any positive power of 10, the (common) logarithm of that number is exactly the number of 0s.

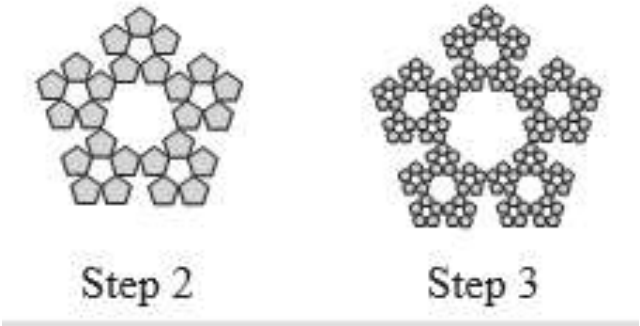
5. Enter "log 8302" on your calculator. The answer is 3.9192 (rounded to 4 decimal places). You knew it was bigger than 3 since $10^3 = 1000$ but less than 4 since $10^4 = 10{,}000$.
6. From the definition of a common logarithm, the answer is clearly $x = \log 8500 \approx 3.929$.

Chapter 7 Try It Now Answers

1.



2.



Chapter 8 Try It Now Answers

1. G has $44 + 14 = 58$ first-place votes, but M has $70 + 22 = 92$, and B has only 80. The winner is McCarthy.

2. Note here that the total number of voters is $44 + 14 + 70 + 22 + 80 = 230$.

   [G vs. M] G is preferred over M by $44 + 14 = 58$ while M is preferred over G by $70 + 22 + 80 = 172$ people. So M wins over G.

   [G vs. B] G is preferred over B by $44 + 14 + 70 = 128$. This is a majority, so G wins over B.

   [M vs. B] M is preferred over B by $44 + 70 + 22 = 136$, a majority. So M wins over B.

   On one-on-one comparisons, M beats G and B, so McCarthy is the Condorcet winner.

3. As seen in Try It Now 1, G has 58 first-place votes, M has 92, and B has 80. So G gets eliminated first. Without G, the schedule would look like this:

4.

	44	14	70	22	80
1st			M	M	B
2nd	M	B		B	M
3rd	B	M	B		

   But the columns can be combined to this:

	136	94
1st	M	G
2nd	B	B

   So G is eliminated, and McCarthy wins by IRV.

5. G: $44 \times 3 + 14 \times 3 + 70 \times 2 + 22 \times 1 + 80 \times 1 = 416$.
   M: $44 \times 2 + 14 \times 1 + 70 \times 3 + 22 \times 3 + 80 \times 2 = 538$.
   B: $44 \times 1 + 14 \times 2 + 70 \times 1 + 22 \times 2 + 80 \times 3 = 426$.

   Once again, McCarthy is the winner.

6. In Try It Now 2, we already did all the pairwise comparisons: M wins over G and B, so M gets 2 points. G wins over B, so G gets 1 point. B gets 0 point. Therefore, McCarthy is the winner by Copeland's method.

7. Seattle gets 60 approvals; Tacoma 85; Puyallup 50; and Olympia 35. Thus, Tacoma is the winner by approval voting.

Appendix B: Solutions to Exercises

Section 1.1 Answers

For this section, the answers may vary. Pick from one or more of the following: *ad hominem*, appeal to ignorance, appeal to authority, appeal to consequence, false dilemma, circular reasoning, *post hoc*, straw man, and correlation-causation.

1. False dilemma
2. *Post hoc*
3. Straw man or *ad hominem*
4. Circular reasoning
5. Straw man or false dilemma
6. Appeal to ignorance
7. Appeal to consequence
8. Correlation-causation
9. *Post hoc*
10. Appeal to authority
11. *Ad hominem*
12. Circular reasoning
13. Circular reasoning

Section 1.2 Answers

1. {5, 15, 25, 35, 45, …}
2. {13, 14, 15, 16, 17, 18}
3. {2, 4, 6, 8, 13, 15, 17, 19}
4.
   (a) Yes.
   (b) No, unless he is also a full-time college student in California.
   (c) Yes, unless she somehow earns at least $20,000 a year without working.
   (d) No.
   (e) Yes, unless she is younger than 18 years old.
   (f) Yes (because this person is at least 18 with an income less than $20,000 a year).
   (g) No, unless she is also a full-time college student elsewhere in California.
5.
   (a) Statement. True.
   (b) Not a statement because it is not a complete sentence.
   (c) Not a statement because it is not a declarative sentence (note the subjunctive mood).
   (d) Not a statement because it is not a declarative sentence (note the imperative mood).
   (e) Statement. False.
   (f) Statement. True.
   (g) Not a statement because it is not a complete sentence (there is no verb).
   (h) Statement. True. (Note there *is* a verb: "*is* equal to.")

6. The earth is not flat.
7. I did not stop smoking.
8. The sun is red.
9. Some did not fail the quiz today. (There is at least one person who did not fail the quiz today.)
10. No one in the car needs to use the restroom.
11. You must find at least one person who ate at the restaurant *and* did not get sick.
12. You must prove that no one who ate at the restaurant got sick.
13. You must find at least one baseball player who can excel at both pitching and hitting.
14. You must find at least one student who does not need an ID number before registering for classes.

Section 1.3 Answers

1.
   (a) Either Elvis is alive or Elvis gained weight.
   (b) It is not true that both Elvis is alive and Elvis has gained weight. (This is equivalent to "Elvis is not alive or Elvis did not gain weight.")
   (c) If Elvis gained weight, then Elvis is not alive.
   (d) Elvis is alive if and only if Elvis did not gain weight.
   (e) Elvis is alive and Elvis did not gain weight.
   (f) It is not true that either Elvis is alive or Elvis gained weight. (This is equivalent to "Elvis is not alive and Elvis did not gain weight.")
2.
   (a)

A	B	$\sim B$	$A \wedge \sim B$
T	T	F	F
T	F	T	T
F	T	F	F
F	F	T	F

   (b)

A	B	$\sim A$	$\sim A \vee B$	$\sim (\sim A \vee B)$
T	T	F	T	F
T	F	F	F	T
F	T	T	T	F
F	F	T	T	F

(c)

A	B	C	$A \wedge B$	$(A \wedge B) \rightarrow C$
T	T	T	T	T
T	T	F	T	F
T	F	T	F	T
T	F	F	F	T
F	T	T	F	T
F	T	F	F	T
F	F	T	F	T
F	F	F	F	T

(d)

A	B	C	$A \vee B$	$\sim C$	$(A \vee B) \rightarrow \sim C$
T	T	T	T	F	F
T	T	F	T	T	T
T	F	T	T	F	F
T	F	F	T	T	T
F	T	T	T	F	F
F	T	F	T	T	T
F	F	T	F	F	T
F	F	F	F	T	T

3.

A	B	$A \vee B$	$A \wedge B$	$\sim (A \wedge B)$	$(A \vee B) \wedge \sim (A \wedge B)$
T	T	T	T	F	F
T	F	T	F	T	T
F	T	T	F	T	T
F	F	F	F	T	F

4.

(a) Exclusive (you probably cannot have both)
(b) Inclusive (you can bring both)
(c) Exclusive
(d) Exclusive
(e) Inclusive
(f) Exclusive

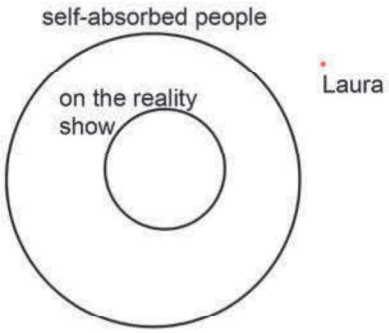
5.

(a) If you cannot watch this movie, then you are under 17.
(b) If you are at least 17, then you can watch this movie.
(c) If you can watch this movie, then you are at least 17.

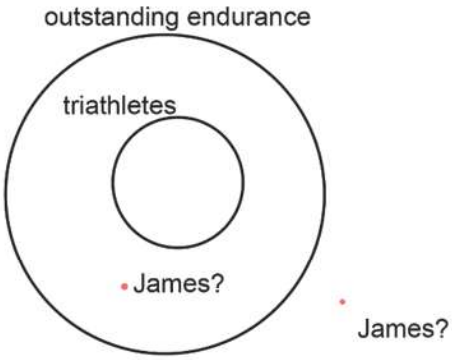
6.
   (a) If you are rich, then you have a house in Beverly Hills.
   (b) If you do not have a house in Beverly Hills, then you are not rich.
   (c) If you are not rich, then you do not have a house in Beverly Hills.
7. Note that (a) is the inverse and (c) is the converse, so they do not have to be true. (b) is the contrapositive, so this must be true.
8. Luke faces Vader, *and* Obi-Wan interferes.
9. You look both ways before crossing the street, *and* you get hit by a car.
10. You were not talking, *and* I missed the instructions.
11. You score a goal now, *and* we do not win.
12. Neither (a) nor (b) could happen according to the biconditional promise.
13. Tina does not like Sprite, *and* Tina does not like 7-Up.
14. The father of that child is not required to testify, *and* the mother of that child is not required to testify (i.e., neither parents are required to testify).
15. Either the House did not pass the bill, *or* the Senate did not pass the bill.
16. Either I do not need a car, *or* I do not need a house.

Section 1.4 Answers

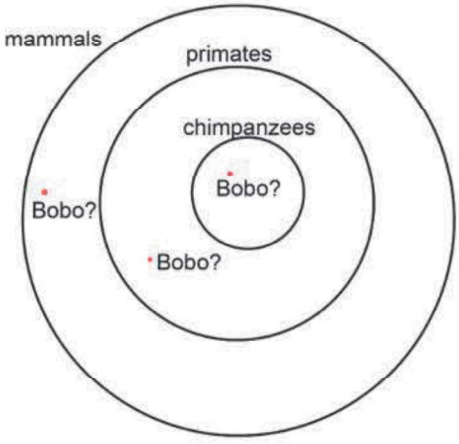
1.
   (a) Inductive
   (b) Deductive
   (c) Deductive
   (d) Inductive
   (e) Inductive
   (f) Inductive
2. Valid. It is probably not sound because there is at least one person who gets a science degree but does not get a good job.
3. Not valid (the fallacy of the converse).
4.
   (a) Valid by *modus ponens*.
   (b) Valid by *modus tollens*.
   (c) Invalid by the fallacy of the converse (affirming the consequent).
   (d) Invalid by the fallacy of the inverse (denying the antecedent).
5. Invalid. Marcie could have bought both a pencil and a pen. “Or” is interpreted as inclusive unless otherwise specified.
6. Valid by disjunctive syllogism.
7. Valid. Laura is outside of the outer circle, so she cannot be in the inner circle.



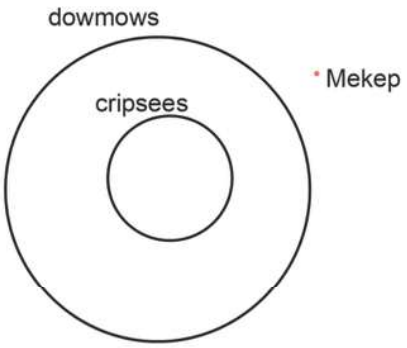
8. Invalid. LeBron James is outside the inner circle, but that does not determine if he is in or out of the outer circle.



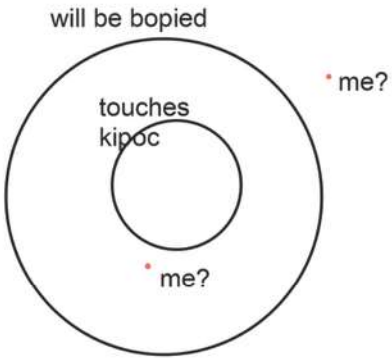
9. Invalid. There are three circles here, and we do not know which circle Bobo is in.



10. Valid by *modus tollens*. Whatever "Mekep" is, it is outside of the outer circle, so it cannot be in the inner circle of cripsees.



11. Invalid by the fallacy of the inverse. You ("I") are not inside the inner circle, but it is not certain if you are in or out of the outer circle. Note the meanings of these words are irrelevant to validity.



Section 1.5 Answers

1.
   (a) $20 \times 13 = 260$ (cents). They do not add up to \$4.10.
   (b) $(20 \times 5) + (45 \times 8) = 460$ (cents). It is more than \$4.10.
   (c) $(20 \times 10) + (45 \times 3) = 335$ (cents). It is less than \$4.10.
   (d) It must be some number between 5 and 10. Also, there must be an even number of 45-cent stamps. Try $(20 \times 7) + (45 \times 6) = 410$ (cents). It is exactly \$4.10. Therefore, there are seven 20-cent stamps and six 45-cent stamps.
   (e) Note that every time you replace a 20-cent stamp with a 45-cent stamp, the amount (value) goes up by 25 cents. From Part (a), where we got \$2.60 assuming all 13 stamps were 20-cent stamps, we see $\$4.10 - \$2.60 = \$1.50$ more is needed. Since $\$1.50 \div 0.25 = 6$, we needed to replace six 20-cent stamps with six 45-cent stamps. This leaves us with seven 20-cent stamps (and, of course, six 45-cent stamps).
   (f) You *can* get the correct answer by solving the system of equations $20x + 45y = 410$ and $x + y = 13$. However, you do not need to do that if you can solve it by guessing the answer or recognizing the pattern described under (e) above.
2. Recognize the pattern in each.
   (a) $1, -1$. The numbers are decreasing by 2.
   (b) $22222, 222222$. A new digit 2 is added each time.

(c) 48, 96. The numbers double each time, starting with 3.
(d) 21, 28. The differences between two consecutive numbers are 1, 2, 3, 4, 5, …
(e) 7, 5. Two 5s are followed by one 7, and the pattern is repeated.
(f) 720, 5040. The n^{th} number is $n!$ For instance, $24 = 4 \times 3 \times 2 \times 1$; $120 = 5 \times 4 \times 3 \times 2 \times 1$.

3. The correct answer is (c). One needs to subtract the $700 (difference between $3,200 and $2,500) from (a) to reach the correct answer.
4. Conjecture: "The sum of the first n positive odd integers is equal to n^2." A proof would require some algebra (or, more formally, *mathematical induction*). Here is an algebraic proof.
[Proof] $1 + 3 + 5 + \cdots + (2n-1) = \frac{1}{2} \times n \times \big(1 + (2n-1)\big) = \frac{1}{2} \times 2n^2 = n^2$ by the arithmetic series formula. There is also a nice way to see this conjecture as a picture of an $n \times n$ square.
5. The surgeon was the boy's mother. (Many people have the pre-conceived notion that a surgeon must be male.)
6. (a) If you happen to know the nursery rhyme, you have probably assumed "H.D." was Humpty Dumpty, an egg. However, the nursery rhyme never explicitly states that Humpty Dumpty was an egg. Here, "H.D." is never even identified with Humpty Dumpty.
(b) From this statement, you can conclude that H.D. is something that is capable of sitting on a wall, falling, and breaking. It could, for instance, be an animation character that is a plate, a fine china, or a robot.
7. Sun/Tue, Sun/Wed, Sun/Thu, Sun/Fri, Mon/Wed, Mon/Thu, Mon/Fri, Mon/Sat, Tue/Thu, Tue/Fri, Tue/Sat, Wed/Fri, Wed/Sat Thu/Sat
Altogether, there are fourteen choices.
8. The mother-in-law must be the wife's mother while a brother-in-law may be the wife's brother or a sister's husband.
9. They will start at 5 am. Remember, 6 am PDT is simply a re-naming of 5 am PST.
10. The average speed is not 50 mph. This is because you drive longer at a lower speed. The trip to your friend's house takes $300 \div 60 = 5$ hours while your return trip takes $300 \div 40 = 7.5$ hours. The average speed for the entire trip is $600 \div 12.5 = 48$ mph.
11. Basically, there are 20 black socks and 18 white socks.
(a) 22 since the first 20 may be black.
(b) 20 since the first 18 may be white.
(c) 3 because, even if the first two have different colors, the third will be either black or white.
12.
(a) The baby girl is George's granddaughter as she is his son's daughter.
(b) She is also the George's half-sister-in-law as she is his wife's half-sister.
(c) If Gloria has a brother (or Michelle had any other son), he would be the baby's half-brother. If Mike had a male child by someone else, he too would be the baby's half-brother.
(d) Gloria is the baby's half-sister since they share the same mother, Michelle. If Mike had a female child by someone else, she too would be the baby's half-sister.

13. To describe any interracial person, the fractions must add up to one (e.g., ½, ¼, and ¼), and the denominator of each must be a power of 2 (e.g., 2, 4, and 8).
    (a) Possible. One parent is Japanese; the other parent is half Russian and half Irish.
    (b) Possible. One example would be if three of the four grandparents are fully Scottish, Vietnamese, Spanish, and the last one is ½ Scottish ½ Vietnamese (or any situation where, three generations ago, there were 3 Scottish, 3 Vietnamese, and 2 Spanish ancestors).
    (c) Impossible. The fractions do not add up to 1.
    (d) Impossible. The fractions add up to 1, but you cannot have 3 or 6 as a denominator.
    (e) Impossible. The fractions do not add up to 1.
14. She is John's daughter. "My mother's child" must be John since John has no siblings. If her father is John, then she must be John's daughter.
15.
    (a) Balls numbered 1 through 200 have been thrown in while Balls 1 and 2 have been removed. There are 198 balls, numbered 3 through 200.
    (b) There are 891 balls, numbered 10 through 900 (since Balls 1 through 9 have been removed).
    (c) By the end of Stage n, $100n$ balls have been entered while n balls have been removed. The difference is the number of balls at that point, $99n$ balls.

    Theoretically, therefore, there will be more and more balls in the bin. However, every ball is eventually removed at some point. Read Section 9.2 for the rest of the story.

Section 2.1 Answers

1. 16%
2. About 7.8%
3. $1890
4. €57.50
5. $83,550
6. $25,000
7. Absolute increase: 1050. Relative increase: 32.3%.
8. Absolute decrease: 33 million. Relative decrease: 22.4%.
9. (a) $1870 or less (b) Yes, $1700 would represent a decrease of more than 15%.
10. (a) At least 104 per day. (b) Yes, 120 per day would represent an increase of more than 30%.

Section 2.2 Answers

1. $4200
2. $35
3. By 3 percentage *points*, but the rate increased by 75%.
4. By 26 percentage *points*, but the rate increased by 78.8%.
5. 3200
6. $210 million

7. Not comparable as (a) is about those without any health insurance while (b) is about employer-provided health insurance. They are certainly compatible but not equivalent.
8. Equivalent (For example, the wholesale price may be $80, and the retail price is $100.)
9. In conflict. If the number has doubled every year since 1950, it cannot have only doubled from 1950 to 1994.
10. They want the dropout rate to be 24%, so that is a 6 percentage-point drop.
11. Increased by 7.14%
12. This could mean there were 1 homicide two years ago and there are 4 this year. Or maybe there were 100 two years ago and 400 now. Sometimes the absolute numbers are more informative.
13. It is possible that it increases by 150%. For instance, from $100 to $250. It is not possible that it decreases by 150% because the price cannot be negative.
14. No. Suppose the shoes were originally $100. With a 30% discount, they were on sale for $70. The employee discount would let you buy them for $63. The total discount is $37, or 37%.
15. If your pay was $100 prior to this, the 50% pay cut would make it $50. A 60% pay raise would give you $80. You still got a 20% pay cut. The two percentages have different base numbers.
16. The absolute difference is 30,510 (students).
    (a) 64.3%
    (b) 39.1%
    (c) 60.9%
    (d) 164.3%
17. The absolute difference is 0.32 seconds.
    (a) 3.2%
    (b) 3.3%
    (c) 103.3%
    (d) 96.8%
18. You end up paying 28% of the original price. Start with $100. The 60% mark-down means the item is $40. The additional 30% off means now it is $28.
19. Same. In both cases, the net loss is 9%. Start with $100. Up 30% means $130. Then, down 30% means it is now $91. On the other hand, the second one, starting also at $100, will go down to $70 on Monday and will go up to $91 on Tuesday.
20. The season batting average may not be .280. We do not know how many at-bats the player had in each half.

Section 2.3 Answers

1. The potatoes are 40 cents a pound, or equivalently, 2.5 pounds per dollar.
2. Note 4 min. 45 seconds is “4.75 minutes (4 and ¾).” Joel ran at 315.8 meters per minute.
3. 15
4. 4
5. 96 inches
6. 6000 grams
7. $6000

8. 720 calories
9. About 55.56 meters in 2 seconds (27.78 meters per second)
10. 3.33 pounds
11. 2.5 cups of flour
12. $67.20
13. About 73.3 turbines (74 turbines)
14. 2667 dump trucks
15. Note the speed of sound is 0.208333 miles per second.
    (a) 0.8333 miles, or 4400 feet
    (b) $1100n$ feet
16. The speed of sound amounts to 1100 feet per second (as seen in Problem 15).
    (a) 275 feet (The sound of the horn has to travel to the building *and back*.)
    (b) $550n$ feet
17. We would need to know the radius of each cup and the depth of each fairly accurately. From the picture, if the skinny cup (on the right) has a radius of 1 inch, the bigger cup appears to have a radius of about 1.5 inches (50% bigger); then, the depth of the liquid in the skinny cup is about 3.6 inches while the depth of the other is about half, 1.8 inches. (All these are rough estimates.) This makes the volume of the liquid in the skinny cup 11.3 cubic inches while the volume in the other cup 12.7 cubic inches. So, it appears that the cup on the left has slightly more, but not much more.
18. The shape is similar, but the larger avocado appears about 20% taller. $(1.20)^3 = 1.728$, so the volume of the larger avocado is 72.8% bigger than the volume of the smaller one. But the price is 90.9% more. Based on the size only, it appears that the smaller avocado is the better deal. Of course, this is just an estimate, and the size of the seed is not considered (nor the freshness or the taste).

Section 2.4 Answers

1. The queen-size mattress is 64.1% bigger, so it would take about 8.2 minutes.
2. The hose speed is 10 seconds per gallon.
   (a) 1800 seconds, or 30 minutes
   (b) 424 minutes, or about 7 hours. (This circular pool has a volume of 339 cubic feet, or 2545 gallons.)
3. You need 100 cubic feet of dirt. Since 1 cubic yard is 27 cubic feet, you will need 3.7 cubic yards. You will need to order 4 cubic yards of topsoil.
4. If 5 scoops of mix is used, you will need 2.5 cups of milk. So add half a cup of milk.
5. You will need to increase everything else by 33%, which is tricky. You will need to add 0.66 egg yolks (not sure how to do that). As for white wine, instead of ¼ cup, you will need 1/3 cup, which is 16 teaspoons.
6. Half a cup is 8 tablespoons. You will need 4 lemons. That's $2.
7. 15 by 20 feet amounts to 300 sq. ft for $2,275. The 18 by 22 patio is 32% larger, so it will cost about $3003.

Section 2.5 Answers

1. 4.9 million barrels would produce 93.1 million gallons of gasoline. The average car uses 600 gallons a year, so this amounts to 155,167 cars.
2. The pool capacity is about 2880 gallons, or 46,080 U.S. cups. Each dollar pays for 4 cups, so this would cost $11,520.
3. The distance to the moon is about 0.258% of the distance to the sun. Equivalently, the sun is about 387.5 times as far as the moon is from the earth.
4. The population density of the U.S. is about 87.5 (persons per square mile) while that of India is 1094.3. India is 12.5 times as dense as the U.S.
5. GDP per capita in China is about $10,243 while that in Sweden is $52,059. The DGP per capita in Sweden is more than 5 times that of China.
6. It takes 50 billion drops of water to fill an Olympic-size pool. Up to 140 billion characters are tweeted every day, so that is 2.8 times an Olympic-size pool if each character is a drop of water. Multiplying it by 365 days, we get up to 1022 Olympic-size pools filled up every year if each character represents a drop of water.
7. If each picture is 2MB, an 8GB drive can contain 4000 pictures. Therefore, 1.5 million of those 8GB drives are necessary to store 6 billion images.
8. A millionaire in the U.S. typically means someone whose net worth is at least $1 million, so certainly it depends on the currency. If one's net worth is $1 million, the person is not quite a "millionaire" in Europe since $1 million is less than a million euros.
9. Tell her that 1.3 grams is 1300 mg, far more than 900 mg.
10. A million inches is equivalent to about 15.8 miles, so 20 miles is longer. One million inches may sound harder to walk.
11.
    (a) An upper bound is 12 million hairs.
    (b) A person with 12 million hairs would have 4000 square inches of skin, each square inch filled with 3000 hairs. Let us call him "Harry." He has dense hair everywhere on his body.
    (c) There are about 8 billion people in the world.
    (d) There are definitely people with the same number of hairs. Imagine 12 million rooms, each labeled with a number 1 to 12 million, and every person is to enter the room labeled with his/her number of hairs. 8 billion people cannot fit into 12 million rooms unless some rooms contain more (many more) than 1 person. The chances are there are many rooms with many people.

Section 3.1 Answers

1. $210
2. T-notes earn simple interest. The interest is $90 a year, so the total interest is $450.
3. You invest $9800 for an annual interest of $800, so that is about 8.16%.
4. $488.67
5. $3,869.68
6. During the 12^{th} year (more than 11 years)
7. Note 3% compounded monthly means 0.25% each month.
   (a) $3,641.51 (b) $1,641.51

8. Note 4% compounded quarterly means 1% each quarter.
   (a) $27,048.14 (b) $17,048.14
9. $3,717.14
10. $16,394.93
11. You will need a high annual rate of approximately 22.5%.
12. You will need about 13.94% compounded monthly.

Section 3.2 Answers

1. $2000 is the principal.
   (a) $101,009.90
   (b) $108,472.35
   (c) $109,196.30
2. $27,331.78
3. $3,288.93
4. $15,381.57
5. $406,569.70
6. Since $e^{0.69315} \approx 2.000$ (almost exactly 2), any amount invested doubles every year. If you start with $100, it will become $200 in one year, $400 in two, and $800 in three.

Section 3.3 Answers

1.
   (a) $116,547.38
   (b) $72,000
   (c) $44,547.38
2.
   (a) $14,486.56
   (b) $10,000
   (c) $4,486.56
3.
   (a) $58,131.59
   (b) $24,000
   (c) $43,131.59
4.
   (a) $37,871.91
   (b) $6,000
   (c) $31,871.91
5.
   (a) $630.31 (He will end up with $800,005.49.)
   (b) $226,911.60
   (c) $573,093.89
6.
   (a) $1,186.35 (You will save up $31,999.97.)
   (b) $28,472.40
   (c) $3,527.57

7. $115.68

Section 3.4 Answers

1.
   (a) $403,221.95
   (b) $612,000
   (c) $208,778.05
2.
   (a) $12,625.50
   (b) $15,000
   (c) $2,374.50
3. $542.38
4. $265.61 per month. Your total payments will be $12,749.28; your total interest is $749.28. (The amounts may be off by a few pennies because of rounding.)
5.
   (a) $180,000
   (b) $1,073.64
   (c) $1,199.10
6.
   (a) $1,618.79
   (b) Her total interest is $312,764.40.
7. His monthly payments are $431.25. He has 2 more years, i.e., 24 more payments. The sum of the present values of the next 24 payments is his current balance: $10,033.45.
8. Her monthly payments are $2,518.11. She has 15 more years. Her balance is $298,404.89. She is not even close to paying off half of the $420,000 mortgage yet since she still owes almost $300,000.

Section 3.5 Answers

1. A deficit is the difference between how much is how much is spent (expenses) and how much is collected (revenues) *during a given period* (when the expenses are bigger than the revenues). A debt is the accumulation of deficits over a number of years.
2. Federal deficit is the difference defined as follows: the amount the government collected (revenues, mostly from taxes) subtracted from the amount it spent during a given fiscal year. Federal deficit results whenever the spending exceeds the revenues.
3. (a) The difference: $780 billion (b) $21.035 trillion (c) 103.9% (d) 3.86%
4. Absolute change: 780 billion – 155 billion = $625 billion. Relative change: 403.2% increase (or “The national deficit in 2017 was 503.2% of the deficit in 1987.

Section 4.1 Answers

1. (a) $\frac{6}{13}$ (b) $\frac{2}{13}$
2. $\frac{5}{26}$

3. Assuming the sample was representative and the participants were honest, the probability is $150/335$, which is about 44.8%.
4. Assuming the sample was representative and the participants were honest, the probability is $205/405$, which is about 50.6%.
5. $3/6$, which is $1/2$, or 50%.
6. Since it is not possible, the probability is 0.
7. The probability that the student is female is $26/65$, which is $2/5$, or 40%. Since 40 people did *not* earn a C, the probability that the student did not earn a C is $40/65$, which is $8/13$, or 61.5%.
8. The probability that the person has no credit cards is $27/81$, which is $1/3$, or 33.3%. The probability that the person is male is $33/81$, which is $11/27$, or 40.7%.
9. $5/6$
10. $2/6$, which is $1/3$, or 33.3%.
11. $2/52$, which is $1/26$, or 3.85%.
12. $1/4$
13. $11/12$
14. $51/52$
15. $48/52$, which is $12/13$, or 92.3%.
16. $3/4$
17. There are 13 spades, 4 aces, and 1 both. So the total number of cards that are either spades or aces is $13 + 4 - 1 = 16$. That means $52 - 16 = 36$ cards are neither spades nor aces. Hence, the probability is $36/52$, or 69.2%.
18. The probability is $0.18 \times 0.18 = (0.18)^2 = 0.0324$, or 3.24%, assuming that the two picks are independent. (Here, technically these are dependent since the probability would change once the first person has been picked—the second person is to be chosen without replacement. However, when dealing with the population of the world, such a change is extremely small and negligible, so independence can safely be assumed.)
19. These mutually exclusive events are **never** independent because they definitely influence each other. If A happens, then B cannot happen (and vice-versa). Mathematically, observe that two events A and B are independent if and only if $P(A \text{ and } B) = P(A)P(B)$; if these are mutually exclusive, the left-hand side is 0, contradicting the conditions given in the problem.

Section 4.2 Answers

1. $\frac{17}{49} \times \frac{16}{48} \approx 11.6\%$
2. $\frac{5}{26} \times \frac{4}{25} \approx 3.08\%$
3. (a) $\frac{1}{13} \times \frac{1}{13} \approx 0.005917 \approx 0.59\%$ (b) $\frac{1}{13} \times \frac{12}{13} \approx 0.071 \approx 7.1\%$ (c) $\frac{12}{13} \times \frac{12}{13} \approx 0.852 \approx 85.2\%$ (d) $\frac{1}{4} \times \frac{1}{4} \approx 0.0625 \approx 6.25\%$ (e) $\frac{12}{13} \times \frac{3}{4} \approx 0.692 \approx 69.2\%$
4. (a) $\frac{1}{13} \times \frac{1}{6} \approx 0.0128 \approx 1.28\%$ (b) $\frac{1}{4} \times \frac{5}{6} \approx 0.2083 \approx 20.8\%$ (c) $\frac{3}{13} \times \frac{1}{2} \approx 0.115 \approx 11.5\%$ (d) $\frac{1}{2} \times \frac{1}{6} \approx 0.08333 \approx 8.33\%$
5. $\frac{4}{52} \times \frac{4}{51} \approx 0.603\%$

6. $\frac{13}{52} \times \frac{12}{51} \approx 5.88\%$
7. (a) $\frac{11}{25} \times \frac{14}{24} \approx 25.7\%$ (b) $\frac{14}{25} \times \frac{11}{24} \approx 25.7\%$ (c) $\frac{11}{25} \times \frac{10}{24} \approx 18.3\%$ (d) $\frac{14}{25} \times \frac{13}{24} \approx 30.3\%$ (e) Same as (d), 30.3%.
8. (a) $\frac{11}{25} \times \frac{14}{24} \times \frac{13}{23} \approx 14.5\%$ (b) $\frac{14}{25} \times \frac{11}{24} \times \frac{10}{23} \approx 11.2\%$ (c) $\frac{14}{25} \times \frac{13}{24} \times \frac{11}{23} \approx 14.5\%$ (d) $\frac{11}{25} \times \frac{10}{24} \times \frac{9}{23} \approx 7.17\%$ (e) $\frac{14}{25} \times \frac{13}{24} \times \frac{12}{23} \approx 15.8\%$
9. (a) $\frac{10}{65} \approx 15.4\%$ (b) $\frac{(26+22-4)}{65} \approx 67.7\%$ (c) $\frac{39}{65} \approx 60.0\%$ (d) $\frac{10}{18} \approx 55.6\%$ (e) $\frac{8}{39} \approx 20.5\%$
10. (a) $\frac{19}{81} \approx 23.5\%$ (b) $\frac{57}{81} \approx 70.4\%$ (c) $\frac{49}{81} \approx 60.5\%$ (d) $\frac{19}{39} \approx 48.7\%$ (e) $\frac{9}{33} \approx 27.3\%$
11. ½
12. 1/6
13. 3/14
14. 2/5
15. Construct a table first.

	Tested Positive (P)	Tested Negative ($\sim P$)	Total
Infected (V)	9	1	10
Not Infected ($\sim V$)	399	3591	3990
Total	408	3592	4000

(a) $P(V|P) = \frac{9}{408} \approx 2.2\%$
(b) $P(\sim V|\sim P) = 3591/3592 \approx 99.97\%$

16. Construct a table first.

	Tested Positive (P)	Tested Negative ($\sim P$)	Total
Infected (V)	95	5	100
Not Infected ($\sim V$)	9,995	189,905	199,900
Total	10,090	189,910	200,000

(a) $P(V|P) = \frac{95}{10{,}090} \approx 0.94\%$
(b) $P(\sim V|\sim P) = 189{,}905/189{,}910 \approx 99.997\%$

17. Construct a table first.

	Tested Positive (P)	Tested Negative ($\sim P$)	Total
Infected (V)	282	18	300
Not Infected ($\sim V$)	3,988	95,712	99,700
Total	4,270	95,730	100,000

(a) $P(V|P) = \frac{282}{4{,}270} \approx 6.60\%$

(b) $P(\sim V|\sim P) = 95{,}712/95{,}730 \approx 99.98\%$

18. Construct a table first.

	Tested Positive (P)	Tested Negative ($\sim P$)	Total
Infected (V)	92	8	100
Not Infected ($\sim V$)	2,997	96,903	99,900
Total	3,089	96,911	100,000

$$P(V|P) = \frac{92}{3{,}089} \approx 2.98\%$$

19. Construct a table first.

	Tested Positive (P)	Tested Negative ($\sim P$)	Total
Cancer (C)	72	8	80
No Cancer ($\sim C$)	694	9,226	9,920
Total	766	9,234	10,000

$$P(C|P) = \frac{72}{766} \approx 9.39\%$$

20. Construct a table first.

	Tested Positive (P)	Tested Negative ($\sim P$)	Total
HIV (H)	9,999	1	10,000
No HIV ($\sim H$)	9,999	99,980,001	99,990,000
Total	19,998	99,980,002	100,000,000

$$P(H|P) = \frac{9{,}999}{19{,}998} = 50.0\%$$

Here, because the percentages are so low (both for the infection rate *and* the false positive/negative rates), it is best to start with a very large number. You will get only 1 false negative out of 100 million men tested. Because the test results are so accurate, the probability that the person is infected given that he has tested positive is rather high (exactly 50%) even though the infection rate is very low.

Section 4.3 Answers

1. (a) The order is important since it is a code; repetition is allowed since you can use the same digit more than once. (b) The order is not important since it is a committee; repetition is not allowed since 4 different people must be chosen. (c) The order is important since it is a playlist; repetition is not allowed since you are to pick 4 different songs. (d) The order is important since it is a code; repetition is allowed since you can use the same digit more than once.
2. (a) The order is not important; repetition is not allowed. You are picking 3 different children in any order to go on the bus. (b) The order is important since it is for seat assignments; repetition is not allowed since each student is to sit in just one seat. (c) The order is not important; repetition is not allowed. You are picking 2 different books in any order. (d) The order is important since license plates are basically codes; repetition is allowed since you can use the same letter or digit more than once. (e) The order is important since it is a batting order; repetition is not allowed since each player must appear only once on the list. (f) The order is important since it is a phone number; repetition is allowed since you can have the same digit more than once.
3. $8 \times 3 \times 2 \times 2 = 96.$
4. $3 \times 5 \times 2 = 30.$
5. $4 \times 4 \times 4 = 64$ if repetition is allowed. $4 \times 3 \times 2 = 24$ if not.
6. $6^4 = 1296$ if repetition is allowed. $6 \times 5 \times 4 \times 3 = 360$ if not.
7. $26^3 \times 10^3 = 17{,}576{,}000.$ (more than 17 million)
8. 26^8, which is more than 208.8 billion.
9. $4! = 24$
10. $210 \times 209 \times 208 = 9{,}129{,}120$
11. $7 \times 6 \times 5 \times 4 = 840$
12. ${}_{25}P_{10} = 25 \times 24 \times \ldots \times 16$, which is over 11.86 trillion
13. ${}_{12}C_4 = 495$
14. ${}_{17}C_5 = 6188$
15. ${}_{50}C_6 = 15{,}890{,}700$
16. ${}_{10}C_3 = 120$
17. ${}_{27}C_{11} \times 16 = 208{,}606{,}320$
18. ${}_{12}C_8 \times {}_{11}C_7 = 495 \times 330 = 163{,}350$
19. (a) $15 \times 14 = 210$ (b) $15 \times 14 \div 2 = 105$ (c) $15 \times 15 = 225$ (d) If the toppings are different, by Part (b), we know there are 105 choices. But the toppings can be the same, adding 15 more options. So the answer is 120 choices.
20. ${}_{16}P_5 = 524{,}160$ permutations, and only one order is alphabetical, so the answer is 1/524,160.
21. ${}_{14}C_{12} = 91$ ways exist to choose 12 out of the 14 men. Altogether, there are ${}_{27}C_{12} = 17{,}383{,}860$ ways to choose 12 out of 27. The answer is 91 divided by this number, which is about 0.000523% (91 divided by over 17 million).
22. ${}_6C_5 = 6$ ways exist to get 5 of the 6 numbers correct, and there are 42 ways to pick one non-winning number (since 6 out of 48 are winning numbers). So there are $6 \times 42 = 252$ ways to get second prize. The total number is ${}_{48}C_6 = 12{,}271{,}512$. The

answer is 252 divided by this number, which is about 0.00205% (252 divided by over 12 million).

23. ${}_6C_4 = 15$ ways exist to get 4 of the 6 numbers correct, and there are ${}_{42}C_2 = 861$ ways to pick two non-winning numbers (since 6 out of 48 are winning numbers). So there are $15 \times 861 = 12{,}915$ ways to get second prize. The total number is ${}_{48}C_6 = 12{,}271{,}512$. The answer is 12,915 divided by this number, which is about 0.105% (12,915 divided by over 12 million).
24. ${}_{13}C_5 = 1287$ ways exist to choose 5 out of 13 hearts. The total number of possible 5-card hand is ${}_{52}C_5 = 2{,}598{,}960$ ways, so the answer is $1287 \div 2{,}598{,}960$, about 0.0495%.
25. There is only one way to get all 4 out of the 4 aces available, and there are 48 non-ace cards from which your last card can be chosen. So $48 \div 2{,}598{,}960$ is the answer, which amounts to less than 0.00185%.

Section 4.4 Answers

1. If you play this game 37 times, you can expect to get gold 3 times ($9), silver 6 times ($12), and black 28 times (-$28), the net loss of $7. The answer is $-7 \div 37 \approx -0.1892$, about -19 cents.
2. If you play this game 60 times, you can expect to get a 6 about 10 times ($30), a 5 about 10 times ($0), and a number less than 5 about 40 times (-$40), the net loss of $10. The answer is $-10 \div 60 \approx 0.167$, about -17 cents.
3. ${}_{23}C_6 = 100{,}947$ ways exist to choose 6 out of 23 numbers. If you play this lottery 100,947 times, each time picking a different combination, you can expect to win $30,000 but not before spending $100,947 to play the game many times. Your net loss is $70,947, so the expected value is $-70{,}947 \div 100{,}947 \approx 0.703$, about -70 cents.
4. ${}_{52}C_2 = 1326$ ways exist to pick 2 cards out of 52. There are ${}_4C_2 = 6$ ways to get two aces, 6 ways to get two 2's, two kings, etc. So there are $6 \times 13 = 78$ ways to get two cards of the same value. If you play this game 1326 times, you can expect to win 78 times ($390) but lose the other 1248 times ($-\$1248$). Your expected value is $\frac{-858}{1326} = -0.647$, about -65 cents.
5. If the company sells 1000 warranties, the revenues will be $48,000. Out of those, 7 of them will cost the company $350, so the cost is $2450. So the net profit is $48000 - 2450 = 45550$. Dividing this by 1000, we see that the expected value is $45.55.
6. If the company sells 10,000 policies, the revenues will be $1 million. Out of those, 13 of them will cost the company $60,000 each, so the cost is $780,000. So the net profit is $1{,}000{,}000 - 780{,}000 = 220{,}000$. Dividing this by 10,000, we see that the expected value is $22.
7. There are $3! = 6$ permutations of the names, and only 2 of them are such that no one draws their own name (BCA and CAB if their names are A, B, C). So the answer is 2/6, which is 1/3 (33.3%). With four people, the probability is 9/24, which is 3/8 (37.5%). With five people, it is 44/120 (36.66%).
8. It is better to switch. See https://en.wikipedia.org/wiki/Monty_Hall_problem. Instead of calculating probabilities, one can present a quick, simply explanation: play this game 300

times. Your guess is right 100 times and wrong 200 times. If you decide *not* to switch your choice, you will win 100 times and lose 200 times. Now, suppose your initial guess is wrong (the other 200 times). That means the real winning door is A (one of the three doors), but you picked Door B (not A). In that case, the host *must* open Door C. The host cannot open Door A or Door B. You originally picked Door B, but you can switch to Door A and win. This happens 200 times out of 300 times. So you **double your chance of winning by switching**.

9. Play this game 8 times. You can expect to win once to get \$6 but lose 7 times to pay \$14. Altogether, you will have lost \$8 in 8 games. The expected value is $-\$1$.
10. Play this game 6 times. You can expect to win once and collect \$10 but lose 5 times to pay \$10. This game is, therefore, even. The expected value is \$0.
11. Play this game 6 times. You can expect to get a 1, 2, or a 3 three times, a loss of \$6. You can expect to get a 4 or 5 two times, a loss of \$10. But you can expect to get a 6 once, a gain of \$20. So, in six games, your net win will be \$4. Your expected value is $\$4 \div 6 \approx 0.667$, or 67 cents (positive). Yes, you should continue to play this game.
12. Play this game 6 times. As in Question 11, your total net win will be \$4, but you will have spent \$6 to play this 6 times. So the net loss will be \$2. Your expected value is $-\$2 \div 6 \approx -0.333$, or -33 cents.
13. (a) 50 cents (b) You will win \$350. Since you will have spent \$700 buying all the tickets, the answer is consistent with (a).
14. Your expected loss per ticket is \$1. If you purchase all 1000 tickets, you will win \$1000 but will have spent \$2000 buying the tickets.
15. On each roll, the probability of winning is 5/6. (a) \$32 (b) \$1024 (c) On the first $n-1$ rolls, you have accumulated $\$2^{n-1}$. (d) \$32,768 (e) It is $\left(\frac{5}{6}\right)^n \times 2^n = \left(\frac{5}{3}\right)^n$ dollars. (f) It is up to you when you quit, but the expected value increases as n gets bigger. In other words, never quit. But eventually, you are almost certain to lose everything. (g) The answers do not change at all since the die rolls are independent events. The only difference is that, if the die is rolled sequentially, you will know exactly when you have lost everything you have accumulated, a devastating moment.
16. Answers may vary. Use common sense.

Section 5.1 Answers

1. (a) The population is the 106 members of the assembly. (b) 106 (c) 28 (d) The sample statistic here is 14/28, or 50%. (e) Half, or 53 members.
2. (a) The 9500 registered voters of the city (b) 9500 (c) 350 (d) $\frac{112}{350} = 0.32 = 32\%$ (e) 3040
3. Loaded question. This question is phrased in such a way as to lead to one answer.
4. Sampling bias. The sample does not represent all voters.
5. Response bias. Honest answers may be used against them. There could also be perceived lack of anonymity.
6. Voluntary-response bias. Many people may not pick up or answer many questions.
7. Loaded question
8. Sampling bias. Women are not represented in the sample.

9. Self-interest bias.
10. Quantitative
11. Categorical
12. (a) categorical (b) categorical (c) quantitative (d) quantitative
13. (a) categorical (b) quantitative (c) quantitative (d) categorical
14. Quantitative. Weight values can be averaged (and other meaningful arithmetic operations can be carried out).
15. Stratified
16. Random
17. (a) All AIDS patients (with symptoms) (b) The data will probably not be identical since different people react differently to drugs. (c) Stratified (d) Convenience. This sample may be biased because only those patients the researcher knew were included, possibly from certain geographical regions, only from certain age groups, etc.
18. (a) The double negative in the question can be (and was) confusing. When one says, "It seems impossible that the Holocaust never happened," the person is affirming the Holocaust. (b) 22% is more than "1 in 5," so their conclusion can be accurate, *provided that the participants understood and answered the question correctly*. As the new survey suggests, many people did not answer it in the way they believed. (c) The second part, "do you feel certain that it happened" makes it much clearer compared to the original question. This change would alleviate much confusion. (d) The difference in the results is due to the re-phrasing. Eliminating the double negative and separating the two positions made it clear to those taking the survey.

Section 5.2 Answers

1. Observational study
2. An experiment
3. (a) Group 1 (b) Group 2 (c) single-blind (d) No, there is no placebo used here.
4. (a) the second group that received the test medicine (b) the first group that received an inert pill (c) This appears to be a single-blind test since there is a rather objective measure of effectiveness—weight. Those staff members checking the participants' weights could know which patients belong to which group; that knowledge does not affect the assessment. Fitness and nutrition advisors do not really matter in this blind experiment. (d) placebo-controlled experiment
5. (a) This is a census if every person in the class was surveyed. (b) observational study (c) Response study, perhaps, if some students may not want to be honest for any reason. There could also be perceived lack of anonymity.
6. They can plant peas in two separate areas (with similar soil and conditions) and play those songs in one area and not the other. Placebos are not necessary or desired since plants do not have psychological effects.
7. A case-control study could be one where the researcher simply finds those children who played team sports and compare their personality with those who did not. An experiment would be difficult and not so practical; the researcher would have to randomly pick some children to play team sports and others *not* to play team sports and then follow them for years and compare their personality in the future.

8. (a) It should be an experiment since this particular plant must be given to some participants and not to others. (b) The treatment group should be given some other plant that may look like the plant being tested. (c) That similar-looking plant could be the placebo. (d) It should be a single-blind test. Subjects should not know which group they belong to. However, blood pressure is an objective measure, so there is no need for a double-blind test.
9. (a) This is a case-control study since two self-chosen groups (as opposed to randomly assigned groups) are compared. (b) A placebo would be some sort of fake social-distancing, which is not possible. (c) They cannot demand certain people *not* to practice social-distance because of ethical reasons.
10. Because the groups were not randomly chosen, there may be other factors making the difference in infection rates, e.g., the socially distancing group may be eating healthier foods, may be more cautious generally, older, exercise more, etc.
11. N/A
12. N/A

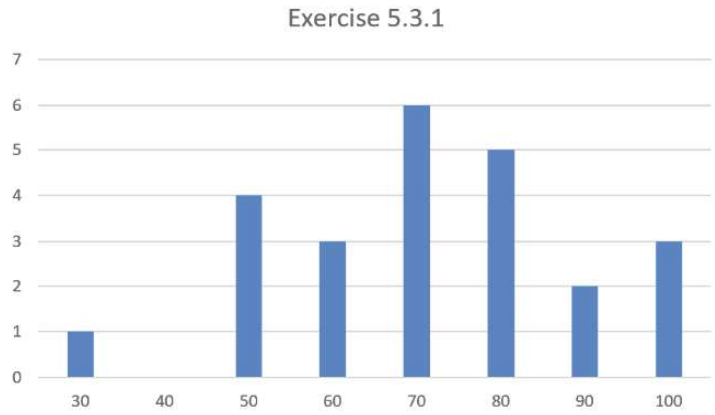
Section 5.3 Answers

1.

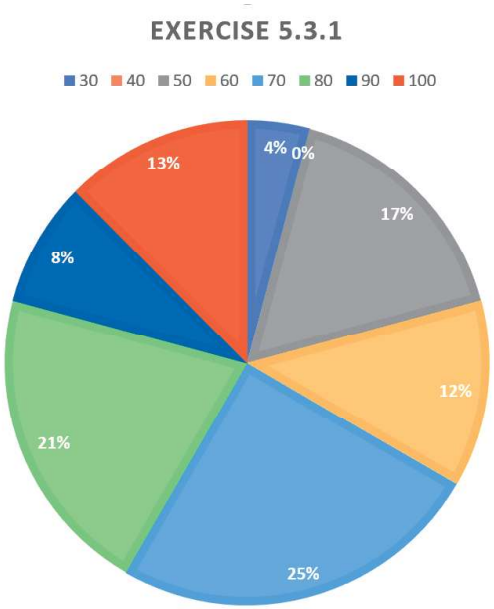
(a)

Score	Frequency
30	1
40	0
50	4
60	3
70	6
80	5
90	2
100	3

(b)



(c)

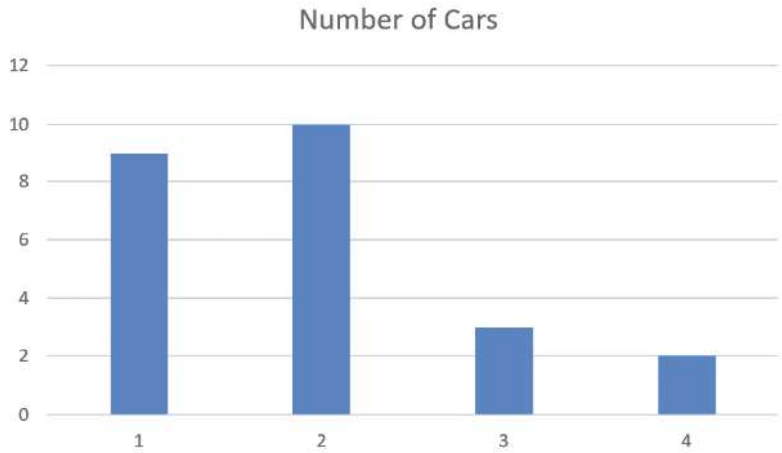


2.

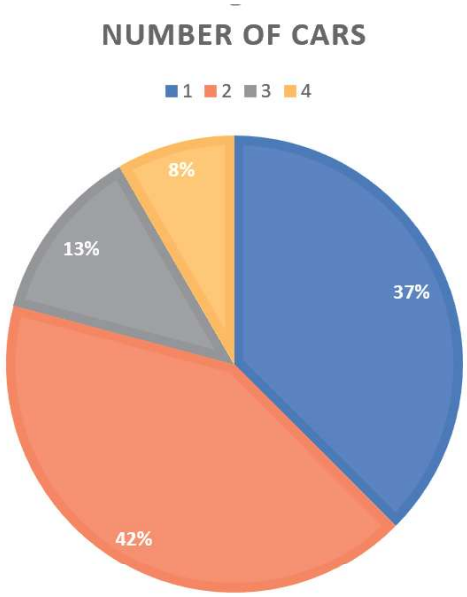
(a)

# of cars	Frequency
1	9
2	10
3	3
4	2

(b)



(c)



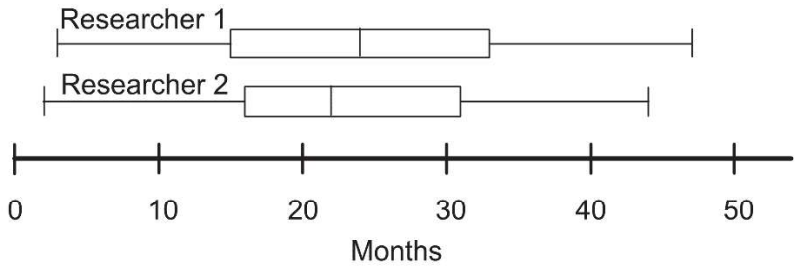
3. (a) 15 (b) 5 out of 15, so 33.3%.
4. (a) 19 (b) 4 out of 19, or about 21.1%.
5. 25% of 20 students; that is 5 students.
6. 31% of $2600 is $806.

Section 5.4 Answers

1. (a) 24.444 (b) 18 (c) 48
2. (a) 6.9 (b) 8 (c) 10 (d) 12
3. (a) $7.81 (b) $7.75 (c) $2.00
4. (a) 21.03 (b) 20.2 (c) 14.2 (d) {15.2, 18.8, 20.2, 22.1, 29.4} (e) The standard deviation is 4.34; it is calculated as follows:

Time	Deviation	Dev^2
15.2	-5.83	33.9889
18.8	-2.23	4.9729
19.7	-1.33	1.7689
20.2	-0.83	0.6889
21.8	0.77	0.5929
22.1	1.07	1.1449
29.4	8.37	70.0569
	Sum of Dev^2	113.2143
	Variance	18.86905
	Standard Dev.	4.343852

5. Here, multiply each rating by its frequency and add them up; the sum is 1032. You divide it by the total number of ratings, 360. The mean is 2.867. The median is the mean of the 180th and the 181st ratings; they are both 3, so the median is 3. The mode is 4.
6. Here, multiply each number of trips by its frequency and add up them; the sum is 127. Divide it by the total number of responses, 85. The mean is 1.494. The median is the mean of the 42nd and the 43rd ratings, both of which are 1. So the median is 1. The mode is 1. The range is 10.
7. B has the bigger mean (50). The standard deviation is the same.
8. The mean and the median are both 42. The range is 0, and the standard deviation is 0.
9. Answers vary. One example is {0, 10, 15, 17, 20, 25}.
10. The median salary for an actuary is $90K while the Q1 salary for a CPA is $40K. So Andrew makes about $50K more than Kelsey according to this data.
11. Q1 of actuary salary is about the median of CPA salary (about $75K), so about 75% of actuaries make more than the median salary of a CPA.
12. The five-number summary for each data set is as follows:
    Researcher 1: {3, 15, 24, 33, 47} Researcher 2: {2, 16, 22, 31, 44}
    The resulting boxplots are shown below:



13. Answers vary. One example of three data sets satisfying these conditions is given below. Here, the mean is 100 for each data set. The way these data values spread out in each set results in different standard deviations—largest for B and smallest for C.
$$A = \{80, 90, 100, 110, 120\}$$
$$B = \{0, 50, 100, 150, 200\}$$
$$C = \{98, 99, 100, 101, 102\}$$
14. Answers vary. One example is given below. To get the same mean, the sum of the values in each set must be equal. In this case, the mean is 30 for each set (the sum is 150 in each set).
$$A = \{10, 20, 30, 40, 50\}$$
$$B = \{0, 0, 0, 0, 150\}$$
$$C = \{10, 10, 10, 10, 110\}$$
15. Answers vary. One example is given below. To get the same median, the middle value (when ordered) must be the same. To get different means, the sum must be unequal.
$$A = \{10, 20, 30, 40, 50\}$$
$$B = \{0, 0, 30, 40, 50\}$$
$$C = \{20, 20, 30, 100, 1000\}$$

16. (a) $1 million (b) $0 (c) $0 (d) The outlier is $5 million. (e) The mean is clearly misleading. The median and the mode tell the story that most of these people had no income; however, this is still not the entire picture. No one measure of central tendency seems to do justice.

Section 5.5 Answers

1. (a) -2 (b) -0.5 (c) 1
2. (a) 0 (b) 1.304 (c) -0.565
3. (a) 500 (b) 430 (c) 750
4. (a) 130 (b) 107 (c) 201.3
5. (a) 50% (b) 2.28% (c) 84.13%
6. (a) 50% (b) 69.15% (c) 15.87%
7. (a) 15.87% (b) 2.28% (c) 97.72%
8. (a) 50% (b) 84.13% (c) 6.68%
9. (a) 2.28% (b) 2.28% (c) 0.13%
10. Based on the normal distribution table, 93.32% of the scores are below 650, so this student is in the 93rd percentile.
11. Based on the normal distribution table, 30.85% of pregnancies are shorter, so this is in the 30th percentile.
12. Based on the normal distribution table, 97.72% of games end with less than 73 points for this team, so this game is in the 97th percentile.
13. Based on the normal distribution table, 93.32% of hitters are below the .318 average, so this player is in the 93th percentile.
14. Based on the normal distribution table, 2.28% of the people in this country are shorter, so this woman is in the 2nd percentile.
15. The probability that a data value is more than 3.89 standard deviations above the mean is 0.0001, which is 0.01% (1 in 10,000); the probability that a value is more than 3.90 standard deviations above the mean is less than that, i.e., basically none.
16. The highest would be 900 while the lowest would be 100; however, in many standardized tests, the range is defined to be from 200 to 800.

Section 6.1 Answers

1. (a) $P_0 = 20$ and $P_n = P_{n-1} + 5$. (b) $P_n = 20 + 5n$ (c) In 7 years
2. (a) $P_0 = 450$ and $P_n = P_{n-1} + 13$. (b) $P_n = 450 + 13n$ (c) In 8 years
3. (a) $P_1 = 55$ and $P_2 = 70$ (both in thousand dollars). (b) $P_n = 40 + 15n$ (c) $190,000 (d) In 4 years.
4. (a) $P_1 = 230$ and $P_2 = 260$ (b) $P_n = 200 + 30n$ (c) 560 (d) In 7 years.
5. (a) The common difference is $\frac{67-3}{8} = 8$, so the formula is $P_n = 3 + 8n$. (b) In 25 weeks.
6. (a) The common difference is $\frac{146-130}{4} = 4$, so the formula is $P_n = 130 + 4n$. (b) 18 months.
7. (a) $P_0 = 200{,}000$ and $P_n = 1.09 \times P_{n-1}$. (b) $P_n = 200{,}000 \times (1.09)^n$ (c) 335,420 (d) 2029

8. (a) $P_0 = 568{,}000$ and $P_n = 1.011 \times P_{n-1}$. (b) $P_n = 568{,}000 \times (1.011)^n$ (c) Note that in 2025, $n = 8$. $P_8 = 568{,}000 \times (1.011)^8 \approx 619{,}951$ (d) In 2069, when $n = 52$. Here, it takes a while since the growth rate is very low.
9. The formula is $P_n = 1700 \times (1 + 1.90)^n$. Here, for 2000, $n = 17$, so $P_n = 1700 \times (2.90)^{17} \approx 123.37$ billion. (The world population is much smaller than this.)
10. The formula (in billions) is $P_n = 5 \times (1.02)^n$. Here, for 2025, $n = 38$, so $P_n = 5 \times (1.02)^{38} \approx 10.61$ (billion).
11. (a) $r \approx 0.1362$. (b) $P_n = 300 \times (1.1362)^n$. (c) about 6427 bacteria (d) About 5.43 hours
12. (a) $r \approx 0.1052$. (b) $P_n = 200 \times (1.1052)^n$. (c) about 544 (d) About 16.1 years
13. (a) $P_n = 100 \times (1.70)^n$ (b) $P_1 = 170,\ P_2 = 289$ (c) In Year 6, the number reaches 2000. (More precisely, when $n \approx 5.646$.
14. (a) $P_n = 10 \times (1 + 2)^n = 10 \times 3^n$ (b) $P_1 = 30,\ P_2 = 90,\ P_3 = 270$.
15. The growth is linear, starting at 4000, with 70 people added to the town population each year.
16. The growth is exponential, starting at 4000, with 4% added to the town population each year.
17. This is still exponential, so start with the formula $P_n = 32{,}000 \times (1 + r)^n$. We know $P_3 = 32{,}000 \times (1 + r)^n = 20{,}000$. This gives us $r \approx -0.145$. The fact that this is negative indicates that this is exponential decay, not growth. The value decreases by 14.5% each year. So the explicit formula is

$$P_n = 32{,}000 \times (1 - 0.145)^n = 32{,}000 \times (0.855)^n$$

18. Here, the formula is $P_n = 100 \times (0.95)^n$. Plugging in $n = 20$, we get \$35.85. In other words, \$100 will only be worth \$35.85 in 20 years. (Another way of understanding this is the following: suppose you have enough money to buy 100 cups of coffee today. That amount will buy less than 36 cups of coffee in 20 years.)
19. N/A
20. In 10 years, your income will be $2^{10} = 1024$ (dollars). In 20 years, you will be making \$1,048,576 (yes, over a million dollars a year). In 30 years, you will be making \$1,073,741,824. Yes, over one billion dollars a year.
21. Solve $0.1 \times 2^n = 385\ billion$. The answer is approximately 41.8. So if you *could* fold this paper 42 times, the thickness goes beyond the moon.
22. 0.1% now, but doubling each year means 0.2% in 1 year, 0.4% in 2 years, etc. In 9 years, the substance will cover 51.2% of the earth, and before Year 10 is done, it will fill the earth.

Section 6.2 Answers

1. (a) $10^3 = 1000$ (b) $10^{-1} = 0.1$ (c) $10^{14} = 10^{14}$ (obviously) (d) $10^{-3} = 1/1000$
2.
   (a) $10^2 = x$. Therefore, $x = 100$.
   (b) $10^x = 10$. Therefore, $x = 1$.
   (c) $10^{2.78} = x$. Therefore, $x \approx 602.56$ (by calculator).
   (d) $10^x = 75.9$. Therefore, $x \approx 1.88$ (use the "LOG" key).

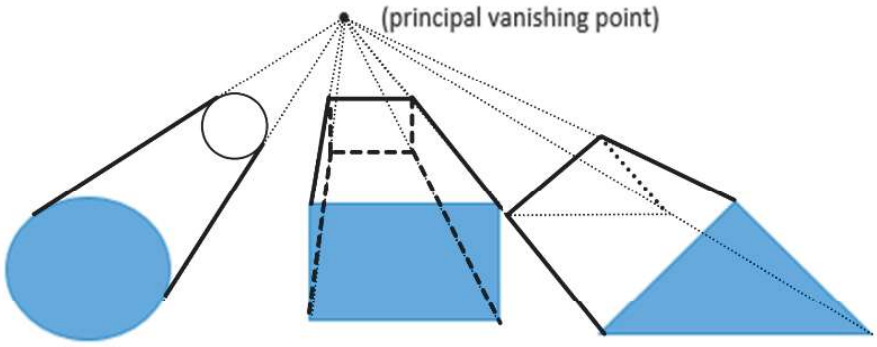
3. The energy becomes 10 times as much.
4. $10^{7.1-6.7} = 10^{0.4} \approx 2.51$. So the Ridgecrest earthquake was 2.51 times as large as the Northridge earthquake (in terms of energy released).
5.
   (a) The largest was the 9.1 (Indian Ocean); the smallest was the 6.0 (South Napa).
   (b) $10^{9.1-6.0} = 10^{3.1} \approx 1259$. The Indian Ocean quake was 1,259 times as intense as the South Napa quake.
   (c) $10^{0.3} \approx 1.995$. The San Francisco quake was almost twice as intense as the Kashmir quake.
   (d) $10^{7.0-6.0} = 10$. The Haiti quake was 10 times as intense as the South Napa quake (the energy released was 10 times as large).

Section 6.3 Answers

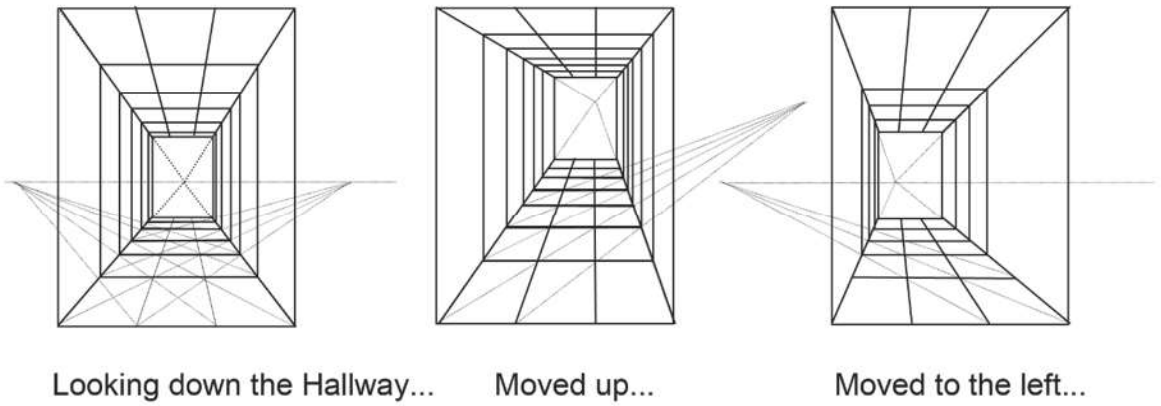
1. The lemon juice (with the lower pH level) is more acidic.
2. $10^{3.5-2.2} \approx 19.95$. One could say that, in terms of the hydrogen ion concentration, the lemon juice is almost 20 times as acidic as a typical wine.
3. $10^{4.0-2.2} \approx 63.1$. The concentration of hydrogen ions in lemon juice is about 63 times that in the orange juice. In other words, if the volumes are equal, the lemon juice contains 63 times as many H^+ ions as the juice squeezed from these oranges. (The lower the pH level is, the higher the concentration. This is due to the negative sign in the definition of pH level.)

Section 7.1 Answers

1. First, draw dotted lines from the vertices to the principal vanishing point as shown.



2. Here are the three hallway drawings:



3. N/A
4. (a) the famous baptistry building in Florence (b) He used mirrors to sketch the building and study perspective drawing. (c) This occurred at the beginning of the Renaissance, introducing the principles of perspective (3-dimensional drawing) to many Renaissance artists, primarily in Italy and the surrounding areas.
5. The five regular solids are the regular tetrahedron, cube, octahedron, dodecahedron, and icosahedron. These are shown to be the only polyhedrons with congruent sides that are regular polygons. Sometimes they are referred to as the Platonic solids.

Section 7.2 Answers

1. Check artists such as Seurat, da Vinci, Dali, Michelangelo, and others. Check the Parthenon as an architectural example.
2. Check classical examples such as Mozart's Piano Sonata No. 1 and Beethoven's famous Fifth Symphony.
3. N/A
4. 144, 233, 377, …
5. N/A
6. There are many online resources on Fibonacci and his famous sequence.

Section 7.3 Answers

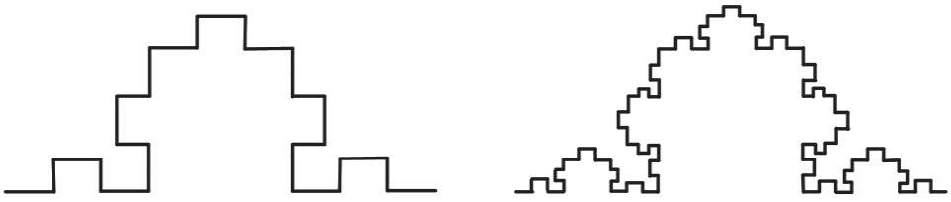
1. 82.5 cps
2. Start at 520 and use exponential growth with $r = 0.05946$:

Note	Frequency (cps)
C	520.0
C#	550.9
D	583.7
D#	618.4
E	655.2
F	694.1
F#	735.4
G	779.1
G#	825.4
A	874.5
A#	926.5
B	981.6
C	1040.0

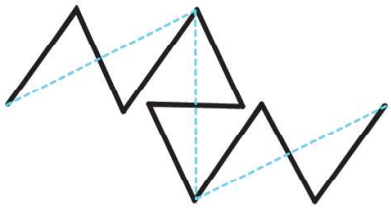
3. 660 cps. The tempered frequency of E is 659.26, so it is very close.
4. 586.7 cps. The tempered frequency of D is 587.33, so it is very close.
5. C4 is given as 262 cps, so C3 is 131 cps. B2 is 5.946% lower, so B2 is 123.6 cps. From C4, you go 7 half-steps to reach G4, so the frequency of G4 is $262 \times (1.05946)^7 \approx 392.5$ cps. So the range is from 123.6 to 392.5 cps.
6. A4 is 440, so A3 is 220, and A2 is 110. F2 is 4 half-steps lower, so $110 \times (1.05946)^{-4} \approx 87.3$. Going on to the high end, A5 is 880, A6 is 1760, and A7 is 3520. G7 is 2 half-steps lower, so $3520 \times (1.05946)^{-2} \approx 3136.0$. So Mariah Carey is said to cover any pitch from 87.3 cps to 3136.0 cps.

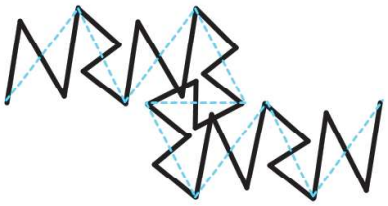
Section 7.4 Answers

1. The next two stages:

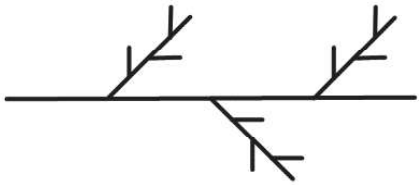


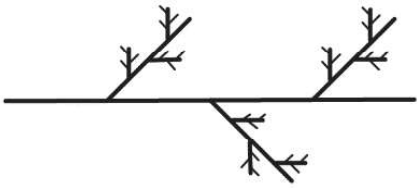
2. The next two stages: note that each segment in one stage is replaced by three zig-zag segments in the next stage. Light blue dotted lines indicate the segments being replaced in each stage.





3. The next two stages:

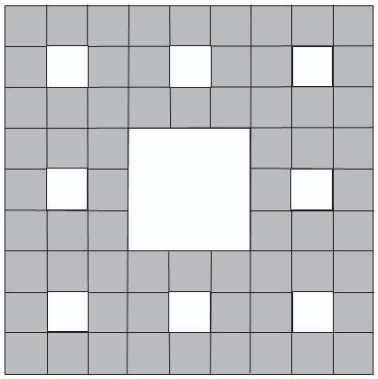




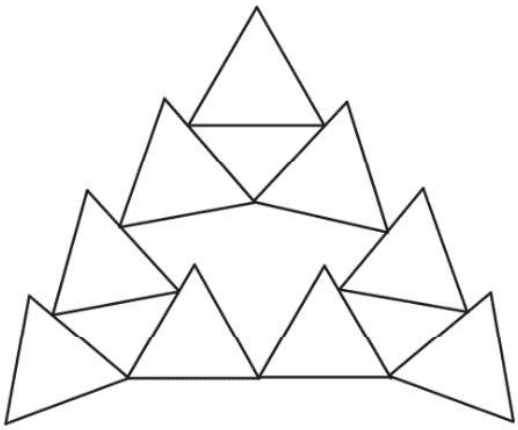
4. Stage 2 and Stage 3 look like these:

As mentioned in the problem, Stage 2 consists of 4 segments of length 1/9. Stage 3 consists of 8 segments of length 1/27. Following this pattern, we see that Stage n consists of 2^n segments of length $\left(\frac{1}{3}\right)^n$. Note that in Stage 1, we removed 1/3 of the original segment. In Stage 2, we removed 2 segments of length 1/9. In Stage 3, we removed 4 segments of length 1/27 each. In other words, in Stage n, we remove 2^{n-1} segments of length $\left(\frac{1}{3}\right)^n$. The "end" result, as you can imagine, is a bunch of points, including all the end points of segments removed. This is known as the **Cantor set** (or Cantor ternary set). You can see that there is no segment left in the end. It is also interesting that the total length removed turns out to be exactly 1 (the length of the original segment), yet there are uncountably many points left in the set.

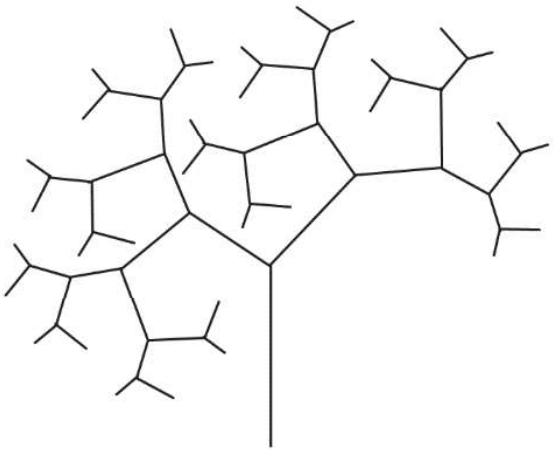
5. The next stage looks like this:



6. The next stage looks like this (the answer may vary, depending on the angle of randomness):



7. The answer may vary. Here is an example.



8. $\frac{\log 4}{\log 3} \approx 1.262$
9. $\frac{\log 5}{\log 3} \approx 1.465$
10. $\frac{\log 8}{\log 3} \approx 1.893$
11. Instead of starting with a segment of length 1, if you start with a segment of length 3, in your first stage, you remove the middle third, leaving you with two segments of length 1. Then, for each segment left, you begin the same procedure as in Exercise 4. So, to scale it by 3, you needed 2 copies. Hence, the dimension is $\frac{\log 2}{\log 3} \approx 0.631$.
12. See the answer for Exercise 4 above.

Section 8.1 Answers

1. There are 795 students in total. Since there are 15 tutors, the standard divisor is 53.

Department	# Students	Quota	Min. Q.
Math	330	6.226415	6
English	265	5	5
Chemistry	130	2.45283	2
Biology	70	1.320755	1
Total	795		14

   (a) The minimum quotas add up to 14. Chemistry has the largest remainder, so it gets the extra tutor. Math 6, English 5, Chemistry 3, and Biology 1.
   (b) $d = 47$, for instance. Math 7, English 5, Chemistry 2, and Biology 1.
   (c) With unusual rounding, the minimum quotas still add up to 14. So try $d = 51$, for instance. Math 6, English 5, Chemistry 3, and Biology 1.
   (d) Same as (c).
2. There are 795 students in total. Since there are 20 tutors, the standard divisor is 39.75.

Department	# Students	Quota	Min. Q.
Math	330	8.301887	8
English	265	6.666667	6
Chemistry	130	3.27044	3
Biology	70	1.761006	1
Total	795		18

   (a) The minimum quotas add up to 18. Biology and English have the two largest remainders, so they get the extra tutors. Math 8, English 7, Chemistry 3, and Biology 2.
   (b) $d = 36$, for instance. Math 9, English 7, Chemistry 3, and Biology 1.
   (c) With unusual rounding, the minimum quotas add up to 20. So no adjustment is necessary. Math 8, English 7, Chemistry 3, and Biology 2.
   (d) Same as (c).

3. There are 1350 customers in total for 20 salespeople, so the standard divisor is 67.5.

Shift	# Customers	Quota	Min. Q.
Morning	95	1.407407	1
Midday	305	4.518519	4
Afternoon	435	6.444444	6
Evening	515	7.62963	7
Total	1350		18

   (a) The minimum quotas add up to 18. Evening and midday have the two largest remainders, so they get the extra. Morning 1, Midday 5, Afternoon 6, and Evening 8.
   (b) $d = 62$, for instance. Morning 1, Midday 4, Afternoon 7, and Evening 8.
   (c) With unusual rounding, the minimum quotas add up to 20. So no adjustment is necessary. Morning 1, Midday 5, Afternoon 6, and Evening 8.
   (d) Same as (c).
4. There are 1350 customers in total for 25 salespeople, so the standard divisor is 54.0.

Shift	# Customers	Quota	Min. Q.
Morning	95	1.759259	1
Midday	305	5.648148	5
Afternoon	435	8.055556	8
Evening	515	9.537037	9
Total	1350		23

   (a) The minimum quotas add up to 23. Morning and midday have the two largest remainders, so they get the extra. Morning 2, Midday 6, Afternoon 8, and Evening 9.
   (b) $d = 50$, for instance. Morning 1, Midday 6, Afternoon 8, and Evening 10.
   (c) With unusual rounding, the minimum quotas add up to 26. So try $d = 55$, for instance. Morning 2, Midday 6, Afternoon 8, and Evening 9.
   (d) Same as (c).
5. The total amount invested is $14,900, and there are 36 coins. The standard divisor is 413.89.

Name	Invested	Quota	Min. Q.
Alice	7600	18.36242	18
Ben	5900	14.25503	14
Carlos	1400	3.38255	3
Total	14900		35

   (a) The minimum quotas add up to 35. Carlos has the largest remainder, so he gets the extra. Alice 18, Ben 14, and Carlos 4.
   (b) $d = 400$, for instance. Alice 19, Ben 14, and Carlos 3.

(c) With unusual rounding, the minimum quotas add up to 35. So try $d = 410$, for instance. Alice 19, Ben 14, and Carlos 3.
(d) Same as (c).

6. The total amount invested is $14,900, and there are 37 coins. The standard divisor is 402.70.

Name	Invested	Quota	Min. Q.
Alice	7600	18.87248	18
Ben	5900	14.65101	14
Carlos	1400	3.47651	3
Total	14900		35

(a) The minimum quotas add up to 35. Alice and Ben have the largest remainders, so they get the extra. Alice 19, Ben 15, and Carlos 3.
(b) $d = 390$, for instance. Alice 19, Ben 15, and Carlos 3.
(c) With unusual rounding, the minimum quotas add up to 37. So no adjustment is necessary. Alice 19, Ben 15, and Carlos 3.
(d) Here, the quota for Carlos, 3.47651, is bigger than the geometric mean of 3 and 4, which is 3.4641. Hence, Carlos' quota also gets rounded *up*, and so do the other two. The sum then would be 38, too many. So try $d = 405$, for instance. Alice 19, Ben 15, and Carlos 3.

7. There are 2,380,000 people in total for the 119 legislators, so the standard divisor is 20,000.

State	Population	Quota	Min. Q.
A	810000	40.5	40
B	473000	23.65	23
C	292000	14.6	14
D	594000	29.7	29
E	211000	10.55	10
Total	2380000		116

(a) The minimum quotas add up to 116. D, B, and C (in that order) have the largest remainders, so they get the extra. A 40, B 24, C 15, D 30, and E 10.
(b) $d = 19{,}500$, for instance. A 41, B 24, C 14, D 30, and E 10.
(c) With unusual rounding, the minimum quotas add up to 121. So try $d = 20{,}100$, for instance. A 40, B 24, C 15, D 30, and E 10.
(d) If $d = 20{,}100$, the quota for E, 10.49751, would be rounded *up* since the geometric mean of 10 and 11 is 10.48809. This change makes the sum of the quotas 120, too many. So try $d = 20{,}120$, for instance. A 40, B 24, C 15, D 30, and E 10.

8. There are 30,000 people in total for the 30 legislators, so the standard divisor is 1,000.

District	Population	Quota	Min. Q.
A	3411	3.41	3
B	2421	2.42	2
C	11586	11.59	11
D	4494	4.49	4
E	3126	3.13	3
F	4962	4.96	4
Total	30000		27

(a) The minimum quotas add up to 27. F, C, and D (in that order) have the largest remainders, so they get the extra. A 3, B 2, C 12, D 5, E 3, and F 5.
(b) $d = 895$, for instance. A 3, B 2, C 12, D 5, E 3, and F 5.
(c) With unusual rounding, the minimum quotas add up to 29. So try $d = 980$, for instance. A 3, B 2, C 12, D 5, E 3, and F 5.
(d) With the original divisor, the quota for D, 4.49, is greater than the geometric mean of 4 and 5, which is 4.4721. So D gets rounded *up* to 5. Then the quotas do add up to 30. So no adjustment is necessary. A 3, B 2, C 12, D 5, E 3, and F 5.

9. The total population is 14,000, and there are 10 seats. The standard divisor is 1400.

Region	Population	Quota	Min. Q.
A	6000	4.2857143	4
B	6000	4.2857143	4
C	2000	1.4285714	1
Total	14000		9

(a) A 4, B 4, and C 2.
(b) With 11 seats, the table looks like this. A and B are tied, but fortunately there are two extra seats available. So A gets 5 seats, B also gets 5 seats, and C gets 1.

District	Population	Quota	Min. Q.
A	6000	4.7142857	4
B	6000	4.7142857	4
C	2000	1.5714286	1
Total	14000		9

10. The total population is 360,000 according to the 2000 census, and there are 50 seats. The standard divisor is 7200.

County	Population	Quota	Min. Q.	Seats
Jefferson	60000	8.3333333	8	8
Clay	31200	4.3333333	4	4
Madison	69200	9.6111111	9	10
Jackson	81600	11.333333	11	11
Franklin	118000	16.388889	16	17
Total	360000		48	50

Madison and Franklin have the largest remainders, so they get extra seats. The right-most column shows the answer for 2000. Now, according to the 2010 census, the table is shown below. With a new population of 363,600, the divisor is now 7272.

County	Population	Quota	Min. Q.	Seats
Jefferson	60000	8.2508251	8	8
Clay	31200	4.290429	4	5
Madison	72400	9.9559956	9	10
Jackson	81600	11.221122	11	11
Franklin	118400	16.281628	16	16
Total	363600		48	50

Note that the population stayed the same except in Madison (by 3200) and Franklin (by 400). The remainders are the largest in Madison and Clay, so these counties add one seat each. The results are on the right-most column. Compared with 2010, Clay gained a seat while Franklin lost one. This is an example of the **Population Paradox**.

11. There are 9,079 students in total for 13 counselors. The divisor is 698.38.

School	# Students	Quota	Min. Q.	Counselors
Lowell	1715	2.4556669	2	2
Fairview	7364	10.544333	10	11
Total	9079		12	13

(a) The minimum quotas add up to 12. The extra counselor goes to Fairview. The right-most column shows the answer.
(b) The divisor was 698.38, meaning that there should be 1 counselor for each group of 698 students. Hence, the addition of 2,989 students would imply there should be 4 more counselors (the quotient is 4.28).
(c) Now, with 17 counselors, the divisor is 709.88. The table is as follows. Lowell picks up an extra counselor while Fairview loses one. School #3 gets 4 as it is supposed to.

School	# Students	Quota	Min. Q.	Before	After
Lowell	1715	2.4158933	2	2	3
Fairview	7364	10.37355	10	11	10
School #3	2989	4.2105568	4	N/A	4
Total	12068		16	13	17

(d) This is an example of the **New State Paradox**.

12. Using numbers from Exercise 7, we get the following. There are 2,380,000 people in total for the 119 legislators, so the standard divisor is 20,000.

State	Population	Quota	Rounded Up
A	810000	40.5	41
B	473000	23.65	24
C	292000	14.6	15
D	594000	29.7	30
E	211000	10.55	11
Total	2380000		121

So the quotas rounded up add to 121, which is 2 too many (there are 119 seats). So we would need to make an adjustment to the divisor. Try $d = 20{,}500$, for example. After rounding each quota *up* using this new divisor, we get A 40, B 24, C 15, D 29, and E 11.

Section 8.2 Answers

1. (a) Both districts have enough voters (11,000 and 10,500). The Republicans win in the First; the Democrats in the Second. (b) Both have enough voters (10,800 and 10,700); the Republicans win in both.
2. (a) Both districts have enough voters (10,500 and 10,000); the Democrats win the First, and the Republicans win the Second. (b) Both have enough voters (10,100 and 10,400); the Democrats win both.
3. (a) Both districts have enough voters (5390 and 5560); the Democrats win the First while the Republicans win the Second. (b) Only 2640 voters, not enough for the required 5200 voters. (c) Both have enough voters (5680 and 5270); the Republicans will both districts.

4. (a) Both districts have enough voters (4530 and 4870); the Democrats win the First while the Republicans win the Second. (b) Only 2310 voters, not enough for the required 4500. (c) Both have enough voters (4810 and 4590); the Democrats win both.

Section 8.3 Answers

1.

	3	3	1	3	2
1st	A	A	B	B	C
2nd	B	C	A	C	A
3rd	C	B	C	A	B

2.

	2	2	2	3	3
1st	A	B	B	C	C
2nd	B	A	C	A	B
3rd	C	C	A	B	A

3. (a) 47 (b) 24 (c) Atlanta (d) Buffalo wins with 111 pts. (e) Chicago, with only 11 first-place votes, gets eliminated first. Then Buffalo is preferred over Atlanta by 28 people, so Buffalo wins. (f) Buffalo wins by the Condorcet method.
4. (a) 29 (b) 15 (c) Atkins (d) Cortez wins with 60 pts. (e) Cortez, with only 5 first-place votes, gets eliminated first. Then Burke is preferred over Atkins by 15 people (Atkins over Burke by 14), so Burke wins. (f) Cortez wins over Burke (16 to 13), and Cortez wins over Atkins (15 to 14), so Cortez wins by the Condorcet method.
5. (a) 460 (b) 231 (c) A (d) D wins with 1240 pts. (e) B, with only 50 first-place votes, gets eliminated first. After combining the 40 and the 100 votes of "D, A, C," we eliminate D with 140 first-place votes (C has 170; A has 150). With B and D removed, A wins over C by 290 to 170. So A is the winner by IRV. (f) B over A; A over C; D over A; C over B; D over B; C over D. So A and B get 1 point each while C and D get 2 points each. So there is no winner by Copeland's method; C and D must compete in some sort of a tie-breaker.
6. (a) 47 (b) 24 (c) A wins with 16 first-place votes. (d) B wins with 128 pts. (e) C gets eliminated first with only 5 first-place votes. After combining the 8 and the 7 votes of "B, A, D," we eliminate B with 15 first-place votes (both A and D have 16). With C and B removed, A wins over D by 31 to 16. So A is the winner by IRV. (f) B over A; C over A; A over D; B over C; D over B; D over C. So A and C get 1 point each while B and D get 2 points each. So there is no winner by Copeland's method; B and D must compete in some sort of a tie-breaker.
7. (a) 33 because one candidate could get 33 votes while all others get 32 each. (b) 17 because one candidate could get 17 votes while all others get 16 each.
8. (a) 137 because one candidate could get 137 votes while all others get 136 each. (b) 120 because one candidate could get 120 while all others get 119 each.
9. B is preferred over A and C, so B is the Condorcet winner.
10. B is preferred over A and C, so B is the Condorcet winner.

11. A gets 15; B gets 17; C gets 16. B is the winner.
12. A gets 24; B gets 22; C gets 18; D gets 20. A is the winner.
13. The Independence of Irrelevant Alternatives (IIA) Criterion
14. Monotonicity Criterion
15. Condorcet Criterion
16. Majority Criterion
17. By plurality, A wins. However, those 10 voters do not like A; they would rather see B elected instead of A. Hence, to prevent A from winning, those 10 voters could choose B as their top choice because that would help B win the election.
18. A wins with 96 points (B has 94, and C has 38). However, if those 18 voters really want B to win under Borda Count, they could rank the candidates B, C, A, in that order. Then, B would win with 94 points while A would get only 78 points.
19. If every voter prefers A to B, then B cannot get *any* first-place votes. So under plurality, B cannot win. The Condorcet method clearly favors A, so B cannot win. Since B has no first-place votes, B gets eliminated first under IRV. So B cannot win. With Borda Count, A will have more points than B, so B cannot win. Note that in each column of the preference table, A is above B. So even if B is preferred to C, D, etc. by some voters, all those candidates would lose to A, so A would have more points under Copeland's method than B. So B cannot be the winner.
20. For Exercise 3, Chicago, with 28 last-place votes, gets eliminated first. Then Atlanta gets eliminated, so Buffalo wins. The process is quite similar to IRV. For Exercise 4, however, the situation is very different. Atkins gets eliminated first with 15 last-place votes. Then, Burke gets eliminated with 16 last-place votes (after Atkins has been removed). Therefore, Cortez wins. Note that Cortez was the first one to be eliminated under IRV.
21. N/A
22. N/A
23. N/A
24. N/A
25. N/A

Section 9.2 Answers

1. Each of these numbers is an integer over a non-zero integer, so by definition, they are all rational numbers. All of these are either terminating or repeating decimals.
2. All of these are real numbers. In addition, 0.311 is a rational number (311/1000). 262 is a natural number, an integer, and a rational number. $-1\frac{4}{5}$ is a rational number. $\frac{51}{3}$ is actually 17, so it is a natural number, an integer, and a rational number. $\frac{\pi}{2}$ is an irrational number. -5 is an integer and a rational number. 0 is an integer and a rational number (although some mathematicians, particularly those in set theory, also include 0 as a natural number). 0.11111… is a repeating decimal, equivalent to 1/9, so it is a rational number.
3. For each element n in A, associate the element $4n$ in B. In function notation, $f: A \to B$ by $f(n) = 4n$. This is a one-to-one correspondence.
4. Each of these sets is closed under addition and multiplication.

5. Take any rational number and add it to (or subtract it from, multiply it by, or divide it by) $\sqrt{2}, \pi$, or e to get new irrational numbers. Also, squares, cubes, and roots of many irrational numbers are also irrational.
6. It turns out that the set of irrational numbers has the same cardinality as $\mathbb{R}$, so it is strictly bigger than $\mathbb{N}$.
7. (a) Yes, each current guest moves down 10 rooms so that the first 10 rooms will be unoccupied. (b) Each current guest moves down 1 million rooms (Room 1 to Room 1,000,001, etc.) so that the first million rooms will be unoccupied. (c) Each current guest in Room n can move to Room $2n$ so that all odd-numbered rooms will be unoccupied. (d) As long as the level of infinity does not exceed the cardinality of $\mathbb{N}$, there is always a room for everyone.
8.
   (a) After Stage 2, balls numbered 1 through 200 have been thrown in while Balls 1 and 2 have been removed. There are 198 balls, numbered 3 through 200. After Stage 9, there are 891 balls, numbered 10 through 900 (since Balls 1 through 9 have been removed).
   (b) By the end of Stage n, $100n$ balls have been entered while n balls have been removed. The difference is the number of balls at that point, $99n$ balls. These are balls numbered $n+1, n+2, n+3, \ldots, 100n$.
   (c) Theoretically, therefore, there will be more and more balls in the bin. However, every ball is eventually removed at some point. So, on the one hand, you would expect to see infinitely many balls in the bin, but on the other hand, try to name a single ball that is in the bin at the "end." You cannot because every single ball has been removed (Ball n gets removed in Stage n). So the bin should be empty (while having infinitely many balls).

Section 9.3 Answers

1. There are 2 vertices of degree 3. So there are Euler paths but not Euler circuits.
2. There are 4 vertices of degree 3. So there are neither Euler paths nor Euler circuits.
3. There are 2 vertices of degree 3. So there are Euler paths but not Euler circuits.
4. There are no vertices of odd degree, so there are Euler paths and Euler circuits.
5. There are no vertices of odd degree, so there are Euler paths and Euler circuits.
6. There are 4 vertices of degree 3. So there are neither Euler paths nor Euler circuits.

References

[i] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Logic is licensed under CC BY-SA 3.0 US.

[ii] LibreTexts Mathematics for Elementary School Teachers: Problem Solving Strategies is licensed under CC BY-NC-SA 4.0.

[iii] Hartford, T. (2021). *The Data Detective: Ten Easy Rules to Make Sense of Statistics*. New York: Riverhead Books.

[iv] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Problem Solving is licensed under CC BY-SA 3.0 US.

[v] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Finance is licensed under CC BY-SA 3.0 US.

[vi] Dr. Fisher Burns' webpage on Continuous Compounding is licensed under CC BY-NC 2.5.

[vii] Lumenlearning Macroeconomics: Federal Budgets and National Debt is licensed under CC BY 4.0.

[viii] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Probability, licensed under CC BY-SA 3.0 US.

[ix] Field Epidemiology Manual: Odds, by Alain Moren, Marta Valenciano, & Arnold Bosman. Reproduction is authorized by the European Union with attribution.

[x] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Statistics

[xi] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Describing Data

[xii] CK-12: The z-Score and the Central Limit Theorem

[xiii] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Growth Models is licensed under CC BY-SA 3.0 US.

[xiv] Lumenlearning Algebra and Trigonometry: Logarithmic Functions is licensed under CC BY 4.0.

[xv] Opentextbc Introductory Chemistry, Chapter on The pH Scale by Jessie A. Key licensed under CC BY-NC-SA 4.0.

[xvi] Wikibooks' Webpage on Projective Geometry is licensed under CC BY-SA 3.0.

[xvii] Simple English Wikipedia's pages on the Golden Ratio/Fibonacci number are licensed under CC BY-SA 3.0 US.

[xviii] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Fractals, licensed under CC BY-SA 3.0 US.

[xix] Math in Society: A survey of mathematics for the liberal arts major by David Lippman, Chapter on Apportionment licensed under CC BY-SA 3.0 US.

[xx] Math in Society: A survey ,of mathematics for the liberal arts major by David Lippman, Chapter on Voting Theory, licensed under CC BY-SA 3.0 US.

[xxi] Wikimedia's Webpage on the Four Color Theorem is licensed under CC BY-SA 3.0 US.

[xxii] Wikimedia's Webpage on the Seven Bridges of Königsberg is licensed under CC BY-SA 3.0 US.