Thinking About Uncertainty: An Introduction to Probability and Statistics

David Draper

University of California, Santa Cruz

Draft 10 (June 2005)

©2005 David Draper

All rights reserved. No part of this book may be reprinted, reproduced, or utilized in any form or by any electronic, mechanical or other means, now known or hereafter invented, including photocopying and recording, or by an information storage or retrieval system, without permission in writing from the author.
## Contents

**Preface**

**Part I. INTRODUCTION**

Chapter 1. What's the Point?

1. A Brief Overview of the Whole Subject 3
2. Summary 13
3. Problems 13
4. Notes 13

**Part II. UNIVARIATE DESCRIPTIVE METHODS**

Chapter 2. Graphical Descriptive Methods

1. Variables and the Values They Take On 17
2. Raw- and Relative-Scale Histograms 20
3. Density-Scale Histograms 26
4. Stem-and-Leaf Plots 26
5. Graphical Methods for Qualitative Variables 26
6. Summary 26
7. Problems 26
8. Notes 26

Chapter 3. Numerical Descriptive Methods

1. Measures of Center 27
2. Measures of Spread 32
3. Percentiles and Boxplots 35
4. Numerical Descriptive Methods for Qualitative Variables 35
5. Summary 35
6. Problems 35
7. Notes 35

Chapter 4. The Normal Curve

1. Some History 37
2. The Form of the Normal Curve 38
3. Using the Normal Curve to Approximate Relative Frequencies 41
4. The Quality of the Normal Approximation 45
5. Summary 45
6. Problems 45
7. Notes 45

**Part III. GATHERING DATA**

Chapter 5. Experimental Design

1. Controlled Experiments 50
2. Observational Studies 58
3. Summary 58
4. Problems 58
5. Notes 58
Part V. INFERENCE AND PREDICTION

Chapter 11. Building a Statistical Model

1. Sampling Models for Parametric Inference
2. Causal Inference Models
3. The Role of Randomization, and the Nature of Randomness
4. Summary
5. Problems
6. Notes

Chapter 12. Point and Interval Estimates

Introduction to Bayesian and frequentist inference: confidence intervals in the one-sample problem for continuous outcomes -- z and t; the t distribution. Posterior distributions. Sensitivity analysis exploring the information content of the prior distribution.

Chapter 13. What to Do When the Standard Models Don’t Fit

What happens when the off-the-shelf models don’t fit. How to tell when they don’t fit: time trend plots, boxplots, normal quantile-quantile plots. What to do when they don’t fit: the data-analytic, model-expansion, and robust/nonparametric strategies.

Chapter 14. Significance Testing

Basic ideas; the z and t tests; the relationship between significance tests and confidence intervals; pitfalls of significance testing.

Chapter 15. An Introduction to Prediction

Prediction is different from inference. Conditional exchangeability judgments. Predictive uncertainty.

Part VI. INFERENTIAL AND PREDICTIVE APPLICATIONS: CONTINUOUS OUTCOMES

Chapter 16. The Two-Sample Problem

Inference and prediction in paired and unpaired comparisons.

Chapter 17. Sample Size Calculations

The estimation approach. The significance testing approach.

Part VII. INFERENTIAL AND PREDICTIVE APPLICATIONS: DISCRETE OUTCOMES

Chapter 18. Small-Sample Methods: The Binomial and Poisson Distributions

Quantitative discrete outcomes. The binomial distribution. Permutations and combinations. The Poisson distribution.

Chapter 19. Large-Sample Methods

The normal approximation to the binomial. Using the Central Limit Theorem: inference on proportions. The one- and two-sample problems.
Part VIII. QUANTIFYING RELATIONSHIPS

Chapter 20. Correlation

Scatterplots; correlation. Association and causation revisited. When is a correlation large?

Chapter 21. Simple Linear Regression

Least squares and simple linear regression, in controlled experiments and observational studies; residual analysis. The connection with the one- and two-sample problems. Prediction in regression models.

Chapter 22. Analysis of Qualitative Outcomes

Relative risk. Chi-square methods. Connections with inference on proportions and logistic regression.

Appendix A1. Areas Under the Standard Normal Curve
Preface

This is a book (SDA/1 for short) about statistics and data analysis. It is the first of several I hope to write, going from basic ideas and methods right through to rather complicated analyses. The subject offers an interesting blend of theory, applications, and philosophy, and I will try to say something about all three, but the emphasis will mostly be on what actually works in helping people solve problems in the world, why it works, and when it doesn’t work. I like an informal approach, so I have tried to write as if you and I were having a conversation about the ideas that form the core of the subject. My classroom style is based on question-and-answer, in which I lay out the problem and then ask the class questions about how to solve it, so some chapters have come out like Socratic dialogues, usually between a researcher and a devil’s advocate who enjoys poking holes in the researcher’s plans.

People familiar with other textbooks on statistics will find three points of novelty here: a strong emphasis on case studies, a lot of ink devoted to predictive methods (as opposed to inference), and an attempt to tell both the Bayesian and frequentist stories side by side. People not familiar with other statistics texts will have no idea what the last sentence means, but that will be remedied by around Chapter 14.

I am a big believer in case studies, for two reasons: most learning takes place by seeing what somebody else did in a situation similar to yours and making suitable modifications, which is exactly suited to case studies; and in case studies, the methods are seen to flow from the applied problems that prompted people to invent them in the first place, which gives people reading them a sense of the discovery process. So most chapters in this book after the introduction have the following four-part structure: I describe a real-world problem which can be solved by applying the method that is the chapter’s topic, and try to get you interested in solving this problem; we go through the reasoning that led to the method’s development by its inventor in the first place, decades or centuries ago; we use the method to solve the real-world problem at the beginning of the chapter; and then we step back and think about the method we’ve “invented” more generally: if you were to use it in other problems, under what conditions would it work well and poorly? If you thought of a method as a product you could buy in a store, like a washing machine, what would its owner’s manual say about how to get the best use from it and what to avoid to keep it from breaking?

Statistics is a branch of applied mathematics, so we will need to do some math along the way in the book, but the math won’t be as important as where it
gets us. It has been my experience that when people are given both the statistical ideas and the math underlying them, they tend to fixate too much on the math. I think this is mainly due to a key difference between statistics and mathematics: (pure) mathematics is about making assumptions that may or may not have any relevance to how the world is and seeing what you can conclude from them, whereas statistics involves quite concretely trying to figure out how the world really is. This is hard work, with a lot of judgment calls, which makes pure math seem more "objective" and cleaner-looking. Statistics uses math but involves a lot of qualitative reasoning as well. Sometimes all you can legitimately conclude with the information you have is that something is bigger than something else, without being able to say how much bigger, and reasoning of that type is a valuable skill in practical problems.

All of this is a long-winded way to say that there will be some math here, but if you're not very mathy you should still be able to get a lot out of the discussion. For example, I will do some calculus, both derivatives and integrals, and there will be a fair amount of summation notation and algebraic manipulation of symbolic sums, but if you're willing to take my word for why something is true rather than proving it along with me, you can just gloss over the details. It would be nice for you to have had a calculus class for another reason, though: getting comfortably through such a class implies a level of familiarity with the manipulation of symbols and equations that I would like to rely on. I don't insist that everybody be able to derive all the equations, but I do insist that everybody know (a) what all the symbols mean in a given equation, (b) why the quantities the symbols stand for ought to be in the equation intuitively, and (c) at least qualitatively whether those quantities are playing a reasonable role in the equation. For example, in the equation \( SE = \frac{SD}{\sqrt{n}} \), does it make sense intuitively that as \( n \) gets bigger, \( SE \) goes down, and as \( SD \) goes up, \( SE \) goes up? (Yes; see Chapter 10.)

It turns out that the only way to learn this subject is to do it, so I have provided a lot of problems. There is a real danger of thinking you know more than you do if you just read the material here—if I do my job as a writer well, you will go along pretty smoothly, nodding your head yes a lot, and then you will find when you get to the problems that that's where the real learning takes place.

I will ask you to do a fair amount of data analysis, for which a computer is indispensable. A good statistical computing package frees you from the drudgery of hand calculation that used to doom many statistics courses, and makes it easy to do something quite important in data analysis: draw a lot of pictures of the data. You should have access to a statistics program, preferably one with high-quality graphics support built in; your goal is to get good enough using it by
about the middle of the book that your choice of methods in doing data analysis is not constrained by your inexperience with your computing environment. I am not going to insist which program you should use; all of the datasets in the book are on the diskette distributed with it, so that you can read them into your favorite package and go from there. I will be illustrating the discussion with output from several packages, including S and Stata.

I owe thanks to a lot of people for helping me put this book together. It started out as a set of xeroxed lecture notes for a course called Statistics 220 at the University of Chicago in 1982, and I have been trying to finish the damn thing ever since. Dozens of doctoral students in policy analysis at The RAND Graduate School of Policy Studies have struggled through draft versions of the material here, much of it in my wretched handwriting; my thanks to them for their patience and suggestions. A number of colleagues and students read drafts of the book and made valuable comments, including [John Adams, Bob Bell, Sally Morton, Dan Relles, John Rolph,]. Heartfelt thanks to The RAND Corporation and The RAND Graduate School, in the persons of Jim Thomson and Charlie Wolf, respectively, for providing me with funding to help finish the book. Finally, thanks, using words I don’t know how to say out loud, to my wife, Andrea Steiner; without her help this would still just be inside my head.

A note to readers of this draft version of the first eight or nine chapters: from time to time you will see something in brackets that makes no sense to you. These things are little notes trying to remind me of something else I want to say when I get time to revise further; please ignore them with my apologies. Also, each chapter is only about 2/3 finished, so you’ll be reading along pretty smoothly (I hope) and suddenly the text will drop you off a cliff. Again, my apologies; it seemed better to get 2/3 of the way through nine chapters than all of the way through six. Finally, for now tables and figures in each chapter are at the end of the chapter rather than where they belong (close to where they are first mentioned in the text); I haven’t figured out how to trick the fancy word-processing program that produced this text to do better yet.
Part I. Introduction
1. What's the Point?

My goal in this chapter is to briefly mention most of the main ideas we're going to look at, so that you can decide if this book and subject are for you. To do this I have to wander around a bit; I ask you to bear with me in the journey. I will wrap it all up at the end of the chapter, in the section called Summary, and I will try to put in little signposts in bold along the way (for example, Science vs. Decision-Making) so that it doesn't get too confusing. You may find it useful as you go along in the book to rescan this chapter from time to time, to get another dose of the big picture and to keep from getting snowed under by the details.

1. A Brief Overview of the Whole Subject

This is a book about statistics and data analysis, so I guess I should start by telling you what I think the discipline of statistics is about. Statistics is the study of uncertainty—how to measure it, what to do about it. It exists as a field because of the following sad fact, which we all observe almost on a daily basis: things done over again under what you think are pretty much the same conditions don't always come out exactly the same. When I drive to work, things like the number of red lights I have to wait at, how long it takes me to get there, and so on vary from day to day, even if I use the same route and go at the same time every day. Sometimes (when I have to get something done by a certain time, for example) I wish this weren't true; I wish I understood what causes the variability in commute time, so that I could more accurately predict when I'll get to work.

Uncertainty, Variability, Causation, Prediction. I just stacked the deck so that a lot of the themes in this book were all mentioned in the last few sentences: uncertainty, variability, causation, prediction. It's a general feature of life that we would like to be able to accurately predict what will happen in the future under given sets of conditions, because that gives us a measure of control over the world around us. Sometimes we can do this quite well: if I drop a hammer off the roof of my house, it will fall to the ground every time. That's an example of a causal (or cause-and-effect, or deterministic) relationship—I cause something to happen, and the effect is always the same. These are the best kinds of relationships to know about, from a practical point of view; your prediction error is always zero. As time goes on, people add more and more of these causal relationships to our collective state of knowledge, but at any given moment there are actually damn few of them compared with the number of statistical, or stochastic, or random, relationships—relationships where a given input can lead
to one of several different outputs in a not-completely-predictable way. [some remarks on uncertainty vs. variability?]

Statisticians are the people who make a living helping other people quantify and manage unpredictability. If all relationships were causal, there wouldn’t be any statisticians. Such a world would work a lot better than the one we find ourselves in but would probably be a lot more boring. It turns out—psychologists have studied this in interesting ways [Kahneman and Tversky]—that people process statistical information a lot more poorly than they do causal information. [Interesting example?] My goal in writing this book is to help you think more sensibly than you now do about uncertainty.

Science Versus Decision-Making. Statistics turns out to be useful in two main areas of human activity: science and decision-making. The distinction I want to make is that science is knowledge for its own sake, while decision-making is putting that knowledge to work to actually do something practical. Most knowledge in science takes the form of structural information: statements about how the world works. Such statements come in two forms: facts and relationships. An example of a fact would be for a physicist to say, "I think the speed of light \( c \) is around 299,792,500 meters per second, give or take about 100 meters per second," whereas in stating a relationship this same physicist might say "It looks to me like \( E = mc^2 \) to a good approximation, where \( E \) is the total energy in a piece of matter and \( m \) is its mass."

I have made a point of expressing these statements in personalistic terms like “I think” to emphasize another theme of this book: all human activity is subjective. People try to pretend that science is totally objective, but in fact it is just as much based on opinion and belief as anything else people do. By saying that, I don’t mean to imply that science is on the same footing with reading tea-leaves or gambling at the racetrack—what we think of as “science” or “the scientific method” has been around now for about 500 years [give a bit of history], and people have worked quite hard during that time to figure out what may be validly concluded from a given body of information and what may not be. But in all that time nobody has been able to figure out a way to divorce science completely from human judgment, and nobody ever will. After all, if we look back on what the best minds of the day 300 years ago thought about almost anything you want to mention, we see their view of “the truth” as partly right and partly wrong, and we’re no smarter now than they were then—some of what we believe right now about how the world works will be understood in the 22nd century to be wrong. The trick is to develop a way to quantify strength of evidence and strength of belief, so that a given body of new evidence
can appropriately modify someone’s beliefs about the thing under investigation, increasing his or her uncertainty appropriately. Part IV of this book is about probability, the part of mathematics devoted to quantifying how much uncertainty you have given the information you possess.

By contrast with science, decision-making is about choosing between alternative futures. When you make a decision about how to act, your thought process basically goes like this: “If I act this way the future will turn out like this; but if I act this other way the future will turn out like that. I like the second future better than the first, so that’s how I’ll act.” Policy analysis, the thing people do at The RAND Corporation (where I work) and other places like it, is a special case, where the goal is to try to help the government figure out how to run the country. The other main kind of decision-making besides policy analysis is business. The distinction I’m attempting here is one of public versus private point of view: in business, the goal is often what you might call local (or private) maximization of profit, whereas in policy the goal is a more global (or public) kind of optimization in which societal needs, not always quantifiable in profits and losses, are balanced against peoples’ individual attempts to maximize profits.

A good example is the issue of the rainforests in Brazil. A person who owns land in the rainforest may want to cut the trees down to sell the timber and build something more money-making in their place. But the rainforest has special value for more than just that person: it supplies a nontrivial fraction of the world’s oxygen, and its destruction would lead to the extinction of many species of plants and animals, with unforeseeable consequences. World ecological groups say that Brazilians have a responsibility that goes beyond the profit motive, but the government of Brazil asks in reply if Brazil is to be condemned indefinitely to third-world status by not being allowed to develop its resources. The (short-term) business answer is clear (sell the lumber); the policy answer is not so clear. (Maybe the world community should offer Brazil an amount of money equal to the commercial value of its rainforests in return for the creation of a permanent world nature preserve.) I have had much more experience in policy and science than in business, and the case studies in this book will reflect that.

Science and decision-making share one key attribute and differ along several key dimensions. The knowledge of structure, which is central to science, is useful—sometimes vital—in decision-making, because if you don’t know how things are related to each other, your views on what the future might hold could well be wrong. But in science you don’t have to take any actions (except perhaps deciding what experiment to perform next), whereas in decision-making you do. We may not know all that we would like to know about how acid rain is formed,
for example (science), but we cannot indefinitely postpone action that will lead to its reduction (policy). The problems you work on in science are yours to choose, but the problems in policy are demand-driven by the crises of the nation and world.

**How People Respond to Uncertainty.** The way people respond to uncertainty in science and decision-making is also different. In science, your uncertainty will usually either be about the precise value of some physical constant, like the atomic weight of hydrogen, or about the range of conditions under which a relationship like $E = mc^2$ holds true to a given accuracy, and the only way these kinds of uncertainty affect how you behave is (as I mentioned above) to suggest to you which experiment to perform to most increase your knowledge. In decision-making, by contrast, the main role uncertainty plays is to tell you how cautious to be: Is it more sensible to be bold and put all your eggs in one basket, or to be cautious and hedge against unpredictability (to borrow a term from investing)?

An example I’ve thought about [Draper, Hodges, Learner, Morris, and Rubin] is a question from economic policy (and business): What is the right amount of money for the United States (and private banks) to lend to third-world countries like Mexico over the next 20 years, say? When you lend somebody money and you want to be pretty sure of getting it back, you need to secure the loan with some tangible assets the person borrowing the money has. In the case of Mexico, one asset that everybody always points to is the crude oil under the ground that Mexico has not yet pumped up to the surface and sold. To figure out what those oil reserves will be worth in the future, you have to predict what will happen to crude oil prices from now until the year 2010, say. A number of smart people [e.g., EMF, EIA] tried to do this back in the early 1980s, when the average world spot price for crude oil was about $[] a barrel. Almost everybody forecast steadily rising oil prices with a narrow range of uncertainty (Figure 1); for example, a popular guess in 1980 for the price of oil in 1985 was about $40, give or take only about $10. This suggested a short-term investment strategy that involved a large amount of lending to Mexico; in other words, both the U.S. government and a lot of banks felt comfortable putting all their eggs in the Mexico-lending basket. Billions of dollars were lent based on this line of reasoning, and it didn’t turn out too well: the actual 1985 oil price was about $17 (Figure 2; it was still less than $20 in 1990), and Mexico is having a hell of a hard time paying back even a fraction of the loans.

In retrospect, everybody did far too little hedging against their real uncertainty about the future price of oil. With the wisdom of hindsight, a better strategy would have been more *adaptive*; for example, they could have loaned part of the money to Mexico at the beginning and waited awhile to see if the price of oil made it
safe to loan some more later on. One theme of this book is that you’ve only done part of the job when you’ve made a single numerical guess at something like the speed of light or the price of oil five years from now; you also need to attach a good give-or-take to your guess, so that you know how much hedging to do against your uncertainty. I will try to give you some practice in decision-making under uncertainty as we go along.

Prediction. One thing you can see from how people make choices in science and decision-making is that prediction plays an important role in both. If you think you have figured out a deterministic relationship in science, you had better be able to very accurately predict what would happen in the future under conditions that you specify (“Given that the distance from the roof to the ground is 15 feet, I think it will take about 0.68 seconds for the hammer to hit the ground, give or take about 0.01 seconds”). An important hallmark of a successful scientific theory, in fact, is its ability to predict something nontrivial that we can all go out and verify. In decision-making, by contrast, good prediction isn’t just desirable, it’s crucial, because you can’t choose well among alternative possible futures without accurately saying what those futures might be like. Approaching problems from a predictive viewpoint is another of the themes in this book (Chapters [14] and [20]). I am going to try to give a number of examples of how thinking predictively helps to clarify the issues.

Data. As important as prediction is, it’s only one of the things people do when they’re reasoning statistically; there are two others: description, and inference. Before I say what these things are and how they differ from prediction, I have to say a few things about data. I want to do so in the context of a real simple and therefore slightly boring example. I get home from the grocery store with, among other things, a “one-pound” package of butter, and I wonder whether I really got one pound. Spot me a digital scale in my kitchen with a dial I can turn that controls how many digits of accuracy I get in the readout. I set the dial to the nearest ounce and weigh the package: 16 ounces (no surprise). How can I tell if the weighing process is deterministic or statistical? I can’t tell until I do it again (and again), under what I’ll try to make sure are identical conditions; if I always get the same answer, it’s deterministic, at the level of detail at which I’m currently observing it. The results of replication of this little weighing experiment as I vary the dial setting might be as in Table 1.

When the dial is set to the nearest ounce, it’s boring; I always get 16. Even at a tenth of an ounce, it always comes out 16.0. But at 0.01 ounce, or any finer gradation on the weight scale, it’s not deterministic anymore; all of a sudden it’s statistical. All measurement processes have this character: at a high enough level
of detail you always get the same thing, but when you hold the magnifying glass close enough, so to speak, the answer varies from replicate to replicate for what looks to you like no good reason.

The numbers in Table 1 are data: numerical observed values of some quantity of interest. In fact, each column in Table 1 is a little dataset that summarizes what might happen with different dial settings. Things that you might want to measure as part of a data-gathering activity are called variables, because they at least have the potential to vary from replicate to replicate of whatever you’re studying. Most data-gathering activities are predictive in spirit, and in prediction variables logically fall into two categories: outcomes, or response variables, or dependent variables, which are the things you want to predict; and predictors, or independent variables, which are the things you use to predict the outcomes. In the 0.01 column in Table 1, for example, there is only one variable—weight of the package of butter—and it’s an outcome; there are no predictors. To see what a predictor might look like in this example, imagine that on the fourth weighing, a truck went by on the street outside my kitchen window, causing the scale to jiggle and produce a reading of 16.34 ounces, but on the fifth reading things were back to normal and I got 15.98. This dataset would look like Table 2.

This is how I’m always going to present data in this book: as a table with rows and columns, in which the columns are variables, the stuff to the left of the vertical bar is outcomes, and the stuff to the right of the bar is predictors. After awhile I won’t put the vertical bar in anymore; it will be obvious from the context. In this book there will always be exactly one outcome variable, but the number of predictors can be anything from zero on up. We will start out with no predictors because it’s simpler that way; there’s enough to talk about without any predictors, in fact, that by the end of the book we will still be covering methods with only one. Multiple predictor variables will be the main subject in SDA/2. The part of statistics that deals with multiple outcomes is called multivariate methods, and that logically belongs even later.

In the example of Table 2 there’s one of each kind of variable—weight is the outcome of interest, as before, but now we have a predictor variable: whether or not a truck went by when the measurement was taken. (Does it look to you as though this variable is useful in explaining why the measurements came out the way they did?) We’ll talk more in Chapter 2 about variables and the kinds of values they can take on; you might notice for now that the outcome and predictor variables in Table 2 are fundamentally different in the scales on which they are observed (yes/no versus numbers). For people who know about matrices, Table
2 should look like a 5 by 2 matrix to you, and in fact matrices are quite helpful in theoretical work in statistics, but we’re not going to need them in this book.

The rows in Table 2’s dataset represent replications of the weighing process under what I was trying to arrange as “identical” conditions; there’s one row for each weighing, or observation. It’s a good idea to try for identical replications when you’re studying something that may turn out to be statistical rather than deterministic; otherwise, if the conditions under which you observe the process vary in ways you don’t intend, you’re shooting at a moving target. It also turns out to be a good idea to try to make the replications independent, in the sense that you get an entirely new piece of information each time you take another measurement. If for some reason the next observation depends at least partially on the value of the last observation, then every time you take two readings you’re only getting (say) 1 1/2 real pieces of information about what the package of butter actually weighs. Later on (in Parts III and IV) we’re going to call data-gathering of this kind “iid,” which stands for “independent identically distributed” observations (the word “distributed” has a technical meaning that will become clear later). Thinking about how to gather data before you actually gather it is an important step in statistics; there’s a whole subspecialty called experimental design devoted to it, which we will look at in Part II (Chapter 6) and again in Part V. [why you get a different answer each time; randomness; key modeling equation in statistics]

Sampling. The measurement setup giving rise to datasets like the one in Table 2 turns out to be only one of three ways data can arise; the other two are called the sampling, or cross-sectional, setup and the time-series, or longitudinal, setup. The idea behind the sampling setup is as follows. Pretty often you have a well-defined group of people or things you would like to know something about, and it’s hard to measure the outcome of interest on all of them. So you choose a sample, or subset, of them in some way, measure the outcome only on the sampled individuals, and try to use the sampled values to guess at what you would have gotten if you had measured everybody. Doing this is called conducting a sample survey.

The whole idea of sampling sounds a little funny when you first hear about it—if you want to know the story for everybody, don’t you have to measure everybody? Here’s an example. It’s October 1988, and I’m wondering who’s going to win the Presidential election in November. This is equivalent to wondering what percentage of the vote George Bush is going to get running against Michael Dukakis; the percentages of the two leading party candidates will add up to almost 100%. Notice that I don’t need to know the individual voting story for every single person who will cast a ballot, just the ratio \( \frac{M}{N} \), where \( N \)
is the number of people who will vote and $M$ is the number of those voters who will pull the Republican lever. It would indeed be a pain to figure this out by interviewing all [125,000,000] registered voters in America, trying to guess which $N$ of them will actually vote, and asking all those people how they feel about Bush. It would not only be a pain, it would waste a lot of money to do so: it turns out that, if you do it right, a sample of only about 1500 people can provide a guess, or estimate, for Bush's national percentage that's quite likely to be right to within about 2.5 percentage points, which is good enough unless the race is really close. How people like the Gallup polling organization do this is an interesting story, which we will take up in the chapter on sampling in Part II. The datasets the Gallup people compile look like the one in Table 3.

I'm going to use the letter $n$ to stand for the number of rows, or observations, in a dataset, and I'll use $p$ to stand for the number of columns, or variables. In Table 3, $n$, which is also called the sample size, is 1500, and $p$ is 5. The choice of the letters $n$ and $p$ is conventional, and the convention of using $n$ for sample size is particularly well assimilated: you might hear somebody in the halls at RAND telling somebody else about the study he or she is working on, for example, and the listener will ask, "What's your $n$?"

Just as with measurement data, it's a good idea to try to choose the sampled people so that you get one independent piece of information from each. Unless you're prepared to make allowances for it in the analysis of your data [define data analysis earlier?], for example, you probably don't want to talk to multiple members of a given family, because they probably share at least somewhat similar political views. The result would be a nominal sample size of 1500, say, but an effective sample size that's potentially quite a bit smaller than 1500, which intuitively would mean less information and therefore more uncertainty about Bush's national level of support. (Intuitively, if you're doing your sampling right, the bigger the sample, the more you know about the quantity of interest.) In the worst possible case, where everybody in each family has identical opinions about Bush, the effective sample size would just be the number of sampled families, since in that case it would be completely redundant to interview more than one person per family. There are times when people make this sort of "mistake" in sampling—choosing a sample of families, and interviewing more than one person per family—on purpose to good effect; it's called cluster sampling and we'll look at it in Chapter 5. The amount of redundancy within families never turns out in practice to be as big as in the worst case scenario above, so it works out that cluster sampling can actually save you money. [Too much detail here? Maybe move to Chapter 5.]
The outcome of interest in Table 3 is of course people’s stand on George Bush, and in this dataset there are four predictor variables. Before leaving this table, you might want to think about why the Gallup people would try to find out if somebody voted in the last Presidential election; what’s that got to do with how a person feels about Bush?

**Time-Series Data.** The third type of data, to go along with data from measurement processes and sample surveys, is *time-series* data. A good example is the situation above involving oil prices and loaning money to Mexico. The data you would collect in that example might look like that in Table 4.

In Table 4 there’s one row for each unit of time at which you took an observation (in this case months). That’s quite different than in the sampling setup, where there was one row for each individual you sampled. The main difference is that in the sampling setup, with the sample taken so that the individuals are independent of each other (as we agreed above was a good idea), there is no reason for the data in one row of the dataset to depend on the data in another row, whereas in the time series setup that’s almost certainly not true: the oil price in February 1981 is probably pretty closely related to the price in January 1981, and so on. Time series data are called *longitudinal*, because you’re following something along at multiple points in time; by contrast, sampling data are usually *cross-sectional*, which means that you have taken a snapshot of a lot of people at one moment in time. One tipoff as to whether a given dataset has a time series or sampling character is this: if you scrambled the order in which you wrote the data down, would you lose any information you care about? With sampling data like that in Table 3, the answer is no; the order was irrelevant. But with time series data like that in Table 4, the order of observation is key. It’s harder to analyze longitudinal data than cross-sectional data; we will mainly leave the analysis of data with time trends to a later book. Note that measurement data (like that in Table 2) looks like it has a time series character, since there definitely is an order in which the weighings occurred, but if you are successful in setting it up so that there is no reason for the value of one observation to depend on the value of another, the time order can be ignored.

**Description and Inference.** A long time ago I said that there were two other statistical things besides prediction that people do—*description* and *inference*—and then I launched into that long-winded story about the different kinds of data. [to be continued]

[description and inference: factuals and counterfactuals]

Everyday life is filled with speculations about counterfactuals: If I had taken the freeway instead of going up Wilshire Boulevard (typical of a big concern in
L.A.), would I have gotten there faster?

examples of inference: estimation, uncertainty assessment, significance testing
examples of description: numerical, graphical
associational vs. causal relationships

a few historical remarks about science (the scientific method) and statistics

4 steps in statistical work: (1) description/exploratory data analysis, summary of design of data-gathering activity (experiment, sample survey); (2) model-building based on (1); (3) inference and/or prediction, including uncertainty assessment, based on (1) and (2); (4) interpretation of findings, based on (1-3) (recommended decision, summary of current scientific knowledge)

2 useful tricks: (1) To figure out what standard methods are sensible, *visualize the data*; (2) to see how to build a model, visualize what ideal data you would have liked to have had to answer the question – the model is just a story relating the data you got to the data you wish you had (imputation of missing data) [belongs in Chapter 10?]

3 comments on model-building: (1) statistics models quite different from economics models (economics: start with theory for how data should come out based on behavioral principles (rationality, incentives, etc.), see if data fits; statistics: start with data, see what models (structural stories, theories, etc.) fit). (2) Why model-building doesn't always work: statistics ≠ {mathematics, economics} (if A then B). Easy to say "If this is the way the world is (A) (model), then the data should have come out like this (B). 2 problems: (a) even if data came out that way, that might not be the only possible story: if A1 then B, if A2 then B. (b) What if the data *didn't* quite come out that way? This kind of reasoning is good for shooting models down but not necessarily for suggesting how to change the modeling assumptions. (3) Most textbooks are real strong on how to do step (3) above (in the 4 steps) in "optimal" ways, which is boring for most people (it's specialists' work: If you make such-and-such modeling assumptions, then what's the best way to estimate so-and-so?). The real intellectual interest in the subject is in the other 3 steps, particularly the modeling step. Easy to lose sight of this sometimes in the middle of complicated calculations. [belongs in Chapter 10?]

mention all parts of the book II gathering data III probability IV inference and prediction V continuous outcomes VI discrete outcomes VII correlation and regression

flow diagram]
What is statistics good for, and not good for? Science vs. policy; description vs. inference vs. prediction; the nature of statistical models (deterministic vs. stochastic); measurement error.

2. Summary

3. Problems
### Tables for Chapter 1

Set dial to nearest

<table>
<thead>
<tr>
<th>1 ounce</th>
<th>0.1 ounce</th>
<th>0.01 ounce</th>
<th>0.001 ounce</th>
</tr>
</thead>
<tbody>
<tr>
<td>repetition under 16</td>
<td>16.0</td>
<td>16.01</td>
<td>16.013</td>
</tr>
<tr>
<td>&quot;identical&quot; conditions 16</td>
<td>16.0</td>
<td>15.96</td>
<td>15.958</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.99</td>
<td>15.987</td>
</tr>
</tbody>
</table>

At this and all higher levels it's deterministic
At this and all lower levels it's statistical

Table 1: Deterministic vs. statistical measurement processes

### Table 2: A little measurement dataset with five rows and two columns

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight at 0.01 ounce setting</td>
<td>Truck went by?</td>
</tr>
</tbody>
</table>

| 16.01 | no |
| 15.96 | no |
| 15.99 | no |
| 16.34 | yes |
| 15.98 | no |

Table 3: Hypothetical Polling Dataset
<table>
<thead>
<tr>
<th>Monthly average world spot price of crude oil</th>
<th>Year</th>
<th>Month</th>
<th>Monthly average OPEC oil production</th>
</tr>
</thead>
<tbody>
<tr>
<td>one row for each moment in time</td>
<td>1981</td>
<td>Jan</td>
<td>[]</td>
</tr>
<tr>
<td></td>
<td>1981</td>
<td>Feb</td>
<td>[]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Hypothetical oil price data
Fig. 1: Typical forecast of the world average spot price of oil (dollars per barrel) from 1980 on based on data through 1979.

Fig. 2: Actual price of oil compared with forecasts, 1980-1986 (from Symes, 1987). Dotted lines are uncertainty bands.
Part II. Descriptive Methods
2. Graphical Descriptive Methods

The point of this chapter is to derive several plots that are useful in descriptively summarizing one variable at a time. To set the stage for this, I need to say a few things about the different kinds of values variables take on and how their values help to determine the best ways to describe them.

1. Variables and the Values They Take On

We agreed in Chapter 1 that variables are things you’re interested in studying that vary from subject to subject or replication to replication of the process leading to individual data values, like how much the package of butter weighed each time I measured it. When it’s subjects that interest you rather than replications, it will often be people, but it doesn’t have to be; it could be objects like things made in a factory. Some typical variables involving people might be education level, religion, and blood pressure. Typical variables involving things made in a factory might include whether the thing is defective or not and customer satisfaction with the product (Table 1).

**Qualitative Versus Quantitative.** The main thing that’s different about the variables in Table 1 is that some of them—education level and blood pressure—have possible values with both a natural *ordering* and a built-in notion of *distance* between them. Such variables are called *quantitative* or *numerical*, because you could imagine locating the values of these variables on a number line and there would be no ambiguity about where they should go. In contrast, religion, defective-or-not, and customer satisfaction (at least as it was measured in Table 1) have no natural numerical labels for their possible values and no unique distance notion between them. Variables like that are called *qualitative* or *categorical*, because about all you can do with them is keep track of the categories into which their values fall. The simplest possible categorical variables are those that take on only two values, like defective-or-not; such variables are called *dichotomous*.

**Discrete Versus Continuous.** People have found two other distinctions among variable types useful, one for quantitative variables, the other for qualitative ones. Among quantitative variables, some of them have gaps between their possible values, of a magnitude that’s relevant to what you’re studying. An example from Table 1 is education level, at least as it is commonly measured: years of schooling completed. Only the integers from 0 to some post-doctoral value like 24 are possible; if you drew the values of this variable on the number
line, as in the left part of Figure 1, you would see big gaps. Quantitative variables like this are called discrete. Other quantitative variables like blood pressure (the right part of Figure 1) have no conceptual gaps between possible values. When you actually measure it, blood pressure will come out to the nearest millimeter of mercury or whatever, so in practice it will have small gaps between its possible values, but conceptually blood pressure could have been anywhere on the number line from [60] to [250], say (above or below those values, you’re in serious medical trouble). Variables like that, with no conceptual gaps between possible values, are called continuous.

Some variables are inherently discrete, like the number of children in your family. Many variables are conceptually continuous but are made discrete by the measuring process, like education level. Many other variables are conceptually continuous, and the measuring process technically cuts them up discretely, but the gaps created in measuring them are so small that people would treat them as continuous. Two good examples are blood pressure recorded to the nearest tenth of a millimeter or annual income to the nearest dollar. It’s a bit of a judgment call, but you’ll see as we go along that it’s not a very hard one, and in any case it turns out that many of the methods we’ll look at work equally well with all quantitative variables, whether they’re discrete or continuous.

Nominal Versus Ordinal. With qualitative variables, the other distinction I mentioned above that people make has a different character. Consider religion. Its values have no natural place on the number line—you could make up numerical labels for its values if you wanted to (1 = Protestant, 2 = Catholic, and so on), and in fact people often do make up such numerical labels when they’re doing computer work, since it’s easier to store numbers than words in a computer. But if one person chose one such set of numerical labels and somebody else chose another, nobody could say that one of these choices is better than the other; they’re both just convenient. Now contrast this with customer satisfaction. This variable may not have a unique notion of distance between possible values (I could make up the numerical labels -3, -1, 0, 1, 3 for the values from “hate it” to “love it” in Table 1, and you could insist instead on the labels -2, -1, 0, 1, 2, and neither choice would be absurd on its face), but there is a definite ordering to its possible values, and this makes it conceptually different from religion. Qualitative variables that have no inherent ordering, like religion or a person’s marital status, are called nominal, and qualitative variables like customer satisfaction that do have a natural ordering are called ordinal or ordered categorical.

One of the themes we’ll see emerge later is that the information content of a variable for answering a science or decision-making question goes up with the
number of possible values it takes on, as long as most or all of those values are actually observed in practice. This means that, when you have a choice in how to measure something, quantitative data is better than qualitative data. There is no free lunch here; it is harder work to collect data with many possible values than data with just a few. To take an example based on the defective-or-not row in Table 1, suppose that instead of just recording whether or not an item is defective you could measure its compliance with recommended performance standards (a quantitative continuous variable, with lots of possible values, say, from 0% to 100%), and defective-or-not (a qualitative dichotomous variable) just corresponds to being below or above a certain percentage on this scale, say 10%. It is obviously more informative to be told the actual percent compliance figure for a given item than just to be told whether the percent compliance is above or below 10%.

One consequence of this quantitative-better-than-qualitative fact is that when people gather ordinal data, like the customer satisfaction scale in Table 1, they often try to attach numerical scores to the categories (like -2, -1, 0, 1, 2) and then analyze the variable as if it were quantitative. You might complain that there is nothing sacred about (-2, -1, 0, 1, 2), and you would be right, but it is an empirical fact [ref. Cochran] that as long as you and I are both reasonable in how we assign the numerical scores, we will get pretty much the same answers when we each treat our ordinal variable as if it were quantitative. For this reason we will not look much at ordinal data in this book; in fact, if you look at the table of contents, you will see that we will not look nearly as much at qualitative data of any kind, nominal or ordinal, as we will at quantitative data. The two kinds of variables that get the most ink here, because they are the most frequently encountered in practice, are quantitative (continuous or discrete) variables and dichotomies.

Figure 2 is a flow chart that summarizes all these distinctions between types of variables.

What’s the Point? Now that I have made you sit through this long speech about different kinds of variables, what is the point of all of these distinctions? It turns out that the values variables take on determine to a large extent how to analyze them. Many methods you would naturally use on quantitative data would be completely silly to use on qualitative variables, and vice versa. This makes it fairly straightforward to organize methods along these lines—with this kind of variable you should do such-and-such—and I have taken advantage of this fact a bit in organizing the book.

There is a danger in relying too much on this sort of if-it’s-Tuesday-this-must-be-Paris logic, though: it may make it look as though all you have to figure
out is what types of variables you’re working with and the right methods will be automatic. Unfortunately, there’s more too it than that; otherwise somebody would have written a computer program to do everybody’s statistical analyses by now, and we could all go home. (People have tried to do this – the resulting computer programs are called expert systems [reference] – and there have been some limited successes, but there’s too much judgment involved in doing this sort of work well to turn much of it over to a machine, given our present state of knowledge.) A mathematical way to say the problem here is that identifying which are the standard methods to use based on the types of variables you have is necessary but not sufficient for good results—if you get the types of variables you’re working with wrong, you will almost certainly produce a flawed analysis, but just getting the types of variables right doesn’t guarantee sensible answers. For that you have to read the rest of the book, and practice.

2. Raw- and Relative-Frequency Histograms

Many American cities, including Chicago, hold civil-service examinations to choose people for jobs like city engineer. In 1966, 223 people showed up to take the Chicago city engineer exam, even though only 15 jobs were available. The test scores of these people, on a scale from 0 to 100, with high scores better than low ones, are given in Table 2. (These data are from a book called Social and Political Inquiry, by Bemesderfer and May [Duxbury, 1972]; I first saw them in an exercise in Chapter 3 of a wonderful introductory textbook called Statistics, by David Freedman, Robert Pisani, and Roger Purves [Norton, 1978].) Solely on the basis of the information I have given so far, the city was charged with rigging the test. Why?

Well, nobody can look at a table of 223 numbers and see much of anything. Maybe a picture would help. The point of this section (and the one after it) is to see how to make some plots called histograms, which are useful in graphically summarizing a single quantitative variable like exam score. There are three kinds of histograms—raw-frequency, relative-frequency, and density-scale—and we will look briefly at all three. Histograms turn out to be really useful; along with scatter diagrams, which will come up in Chapter 19, they are probably the single most useful graphs in introductory statistical work. Note that the test scores in Table 2 are discrete, but that doesn’t matter; histograms work both on discrete and continuous variables.

Before we go through the logic underlying the picture that will solve the Chicago mystery, it’s worth getting clear on the nature of the question we’re
trying to answer. When we figure out why the examiners were accused of rigging the test, are we doing description, inference, or prediction? [gentle readers of the rough draft: in chapter 1 material you haven’t seen, I have set the stage for the next few sentences.] It looks like a factual rather than a counterfactual to me, so it must be description. Nobody is asking us to use these data to say anything about test scores of other applicants in the past or present (inference) or in the future (prediction). This is a hallmark of histograms; they are graphical descriptive tools.

Raw-Frequency Histograms. Okay, so how do you make a plot that will reveal somebody’s guilt in this case study? The first thing to do with a big batch of numbers like the raw test scores in Table 2 is sort them from smallest to largest. All you’re giving up in doing so is the order in which they were written down, which is presumably irrelevant here. If the order potentially mattered, for instance if these were measurements of something, like the weighings of the package of butter in Chapter 1, and the data were written down in the time order in which they were taken, a histogram would not be a good idea unless you could convince yourself that the measuring process unfolded haphazardly over time. We will talk in Chapter [12] about how to make pictures called time trend plots that examine the relationship between the data values and the order in which they were gathered.

Anyway, I sort the test scores from smallest to largest and look at them (Table 3). Now I start to notice a few things: the smallest score was 26, the largest 95, and there are many duplicate values, for example four 27’s. The duplicates mean that it’s redundant to store the data as in Table 3—it’s simpler to just keep track of the different values the data take on and the frequencies with which those values occur (Table 4).

In this table I have kept track of the frequency information in both of the two ways you can think about frequencies: in raw terms—counts—and in relative terms—percentages—which you get by dividing the raw frequencies by the sample size, \( n = 223 \). Taken together, the columns called “values” and “raw frequency” in Table 4 are called a raw frequency distribution of the exam score variable, because they monitor how the data are distributed across the different values the variable takes on (how many 26’s, how many 27’s, and so on). Similarly, the “values” and “relative frequency” columns taken together form a relative frequency distribution of the exam scores. Sometimes people just abbreviate these phrases and say “distribution” for short; you hear people ask things like “How is that variable distributed?” and one way to answer them is to give them a table like Table 4.

From this table it’s an easy step to the picture I’ve been promising, as follows. Concentrate just on raw frequencies for the moment. That gives you two pieces
of information, or *dimensions*, to keep track of: the values that the variable "exam score" takes on, and the raw frequency with which it takes on those values. Two is a handy number of dimensions for plotting, since that's how many a sheet of paper has: horizontal and vertical. So let's plot values along one dimension and frequencies along the other. It's entirely conventional which you put on which axis. Most people put the values along the horizontal and the frequencies along the vertical, although it's easier when making low-resolution computer plots to do it the other way around, so you see it both ways in practice. In this book I'm always going to plot the values along the horizontal and the frequencies along the vertical, as in Figure 3. This plot is called a *raw frequency histogram* or just a *histogram* for short.

Notice that, while I could have plotted the frequencies along the vertical scale as dots or stars or something, instead I plotted them as heights of bars. A histogram is a special kind of *bar chart*, which is an example of the sort of presentation graphics (like *pie charts*) that business people and newspapers use. Many statisticians don't like most pie and bar charts because they are so easy to use inappropriately; histograms are the exception. [A bit more on what's wrong with pie charts, etc., here or later?]

To make a histogram you just have to decide how many bars you want and how wide they should be, and then you can draw it by hand or get a computer to do it for you. Most of the time you will want to have all the bars be the same width and have the left- and right-most bars just barely span the observed range of the data from smallest to largest (as in Figure 3), which if you think about it a moment means that all you really have to decide is how many bars you want. You can make the bars have variable widths if you want (we will do so in a little while), but it's often not necessary and just leads to a particular form of technical headache that we will have to solve below. Here we get the most informative plot for solving the Chicago mystery by making all of the bars one point wide on the 0-100 point exam score scale, as in Figure 3. Before reading further, you should look carefully at that figure now and try to figure out why anybody should be accused of cheating on the basis of it.

The key is in two separate things to notice: (1) there is a big gap toward the right of the plot, five points wide in fact, when the largest gap anywhere else in the picture is only one point wide; and (2) if you read the raw frequencies off the vertical scale in Figure 3 or look back at them in Table 4, it becomes clear that the number of people in the little bulge on the right of the plot, separated by the five-point gap from the bulk of the data, is 15, exactly the number of jobs to fill. One of these facts on its own might have been just barely palatable, but the two
taken together are too hard to swallow. One plausible story is that the examiners created the five-point gap by moving the test scores of the people around 85 down a bit and the scores of the people around 89 up a bit to make it impossible for people to argue about who should get the jobs (in fact the bar at 84 is unusually high in relation to the other bars around it). The picture does not have anything to say about who the 15 “successful” applicants were—friends or bribers of the examiners, for instance—but it does emit a sufficiently unpleasant odor to compel further investigation. [mention the outcome of this investigation]

**How Many Bars?** I want to emphasize that, because of the element of choice in the number of bars, histogram-making is both an art and a science; there are lots of different histograms you could make with the same variable, and some are better than others. You could make a histogram of the data in Table 2 that had one enormous bar stretching from 26 to 95, or, even worse, from 0 to 100, if you wanted; the first of these plots would be useless for solving the mystery because all of the important “local” detail (meaning in the immediate vicinity of a given value on the horizontal axis) is washed away, and the second plot would not only suffer that defect but would also mislead people into thinking the range of the observed data from smallest to largest was larger than it really was.

To best see the detail needed to solve the Chicago mystery, we went to the opposite extreme, using the largest possible number of very skinny bars. That worked in this problem but often produces bad histograms, in the following important sense: often the batch of numbers you’re making a histogram of is a sample, in the sense we discussed in Chapter 1, from some population or measuring process, and you could imagine doing the sampling again and getting somewhat different numbers (because of the statistical nature of the process you’re studying).

**Replication Principle:** Whenever you draw a conclusion from some data, ask yourself how well you think your conclusion would stand up under independent replication of the process leading to your dataset this time. There are two kinds of mistakes to avoid from this point of view: (1) failing to extract a feature of the data that is universal, in the sense that (almost) every time the data-gathering process is replicated the feature would be present in the resulting dataset (“underfitting the data”); and (2) drawing a conclusion that is too narrowly tailored to the single dataset at hand (“overfitting the data”), so that the next time you did the data-gathering the “feature” you thought you had identified would disappear. You want to try to walk the line between these two extremes.
It's like holding a magnifying glass up to the dataset as if it were an object: if you hold it too far away, the picture is blurry and you miss important features, but if you hold it too close you end up seeing things (that are the metaphorical equivalent of little bumps and scratches on the surface of the object) that would not be in exactly the same place if you were to do the data-gathering experiment over again.

Well, this Replication Principle is all rather noble-sounding, but you're probably just wondering how many histogram bars is the right number for a given variable. It should intuitively depend on the sample size, $n$, in such a way that as $n$ increases the number of bars, $k$ let's say, also increases; as you get more and more observations, you can afford more bars without overfitting the data. This topic has been the subject of some theoretical work, and one rule that has come out of this work [reference] is as follows:

$$k = \begin{cases} 
[2\sqrt{n}] & n \leq 100 \\
[10 \log_{10}(n)] & n \geq 100 
\end{cases}$$

where $[x]$ denotes the integer part of $x$.

This rule does correctly suggest increasing the number of bars as $n$ goes up, but you may well feel (as I do) that it recommends too many bars for small $n$. Another rule I've seen advocated [reference] says:

$$k = [1 + \log_2(n)].$$

The first of these rules tends to recommend too many bars with small sample sizes for my taste (20 bars with 100 observations, for example, and 8 bars with only $n = 16$), the second probably too few (only 9 bars with 300 observations). Maybe a sensible course of action is somewhere in between. A definitive story on how to draw histograms is hard to come by, but one thing is clear: it doesn't hurt, and will often help, to make several histograms with the same variable, experimenting with two or three values of $k$ until you have something that seems to be relatively faithful to the Replication Principle. A corollary of this is that you don't have to automatically accept the histogram your computer program gave you; it used some rule to figure out its default value of $k$, after all, and you may not agree with this choice.

**Relative-Frequency Histograms.** All of this has been based on the raw frequencies in Table 2; what about the relative frequencies? Intuitively, since you get from raw to relative frequencies by dividing by a constant (the sample size, $n$), one of these should give the same information about the distribution of exam
scores as the other. In fact, that’s right, and for a reason that will come up again later, so we might as well look at it for a moment now.

Dividing by a constant is a special case of linear rescaling, in which you replace some value by \( ay + b \) for some numbers \( a \) and \( b \). In the raw-to-relative frequency case, if \( y \) is the raw frequency with which some value occurs in the data, the relative frequency is \( \frac{y}{n} \), so that \( a = \frac{1}{n} \) and \( b = 0 \). Linear rescaling comes up in everyday life perhaps most prominently in the collision between the old-fashioned English measuring system we in the United States cling to (feet, pounds, degrees Fahrenheit) and the metric system the rest of the world uses (meters, kilograms, degrees Centigrade). Intuitively whether I measure the length of something in feet or meters should be irrelevant for all comparative judgments: if something is twice as long as something else when I measure it in feet, it had better be twice as long when I measure it in meters. What this means graphically on the number line (the left part of Figure 4) is that linear rescaling amounts only to a renaming of the points along the line that preserves all comparative relationships. This is in sharp contrast to nonlinear rescaling, for example working with the logarithms of a list of numbers rather than the numbers themselves (the right part of Figure 4)—20 is twice as big as 10, but \( \ln(20) = 2.996 \) is not anywhere near twice as big as \( \ln(10) = 2.303 \).

The point of this for raw-versus relative-frequency histograms is that the linear rescaling of the vertical (frequency) axis will have precisely zero effect on the shape of the histogram. To get the relative-frequency histogram of the exam scores that corresponds to Figure 3, I just relabel the vertical axis so that, for example, 1 now corresponds to \((1/223) \times 100\% = 0.45\%\), 2 corresponds to \((2/223) \times 100\% = 0.90\%\), and so on (Figure 5). The two plots are identical.

So if raw- and relative-frequency histograms are identical, why did I bring them both up? It turns out that they’re not necessarily identical when you’re comparing two or more histograms; in that case relative-frequency histograms have a distinct advantage. Imagine comparing the raw frequency histogram in Figure 3 based on 223 people to a raw-frequency histogram of the exam scores in another year when 446 people showed up for the test instead of 223 (Figure 6). Notice that the stories the two plots have to tell about how people did on the exam are identical, but if I draw them on the same horizontal and vertical scale (always a good idea when you’re comparing two or more plots), one of them will be twice as tall as the other just because it has twice as many people in it. This is a defect of raw-frequency histograms that is completely remedied by drawing the pictures instead on the relative-frequency scale: on that scale the
two plots are identical, because every value has the same relative frequency in
the two batches of test scores.

Now you have a right to ask a different question: If raw- and relative-
frequency histograms are identical except when you’re comparing two or more
batches of data with different sample sizes, in which case relative-frequency
histograms are better, why did I bring up raw-frequency histograms in the
first place? Because most statistical computing packages produce raw-frequency
histograms, even if that’s not quite what you want. The best kind of histogram is
the third kind I mentioned a long time ago, density-scale histograms, which we
haven’t even talked about yet. Almost no computer programs produce density-
scale histograms automatically, and most of them won’t even do it if you beg.
They’re the best because even relative-frequency histograms have a defect that
density-scale histograms remedy, and also because they tie in best with the theory
that underlies the inference and prediction we’re going to do later.

3. Density-Scale Histograms

   density-scale example (health care expenditures)

4. Stem-and-Leaf Plots

   [I will do boxplots in chapter 3, after medians and interquartile ranges]

5. Graphical Methods for Qualitative Variables

   [pie charts, bar charts]
   [reorganize the headings and text a bit to make the qual- vs. quant- distinction
clearer]

6. Summary

7. Problems

   empirical cumulative distribution functions?
### Tables for Chapter 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education level (years of schooling completed)</td>
<td>0, 1, 2, ..., up to 24 or so</td>
</tr>
<tr>
<td>Religion</td>
<td>Protestant, Catholic, Jewish, other</td>
</tr>
<tr>
<td>Blood pressure (systolic, say, measured in millimeters of mercury)</td>
<td>Numbers like 120.082..., 98.3826..., and so on</td>
</tr>
<tr>
<td>Defective or not</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Customer satisfaction</td>
<td>Maybe you’d use a five-point scale:</td>
</tr>
<tr>
<td></td>
<td>Hate it, don’t like it, it’s ok, like it, love it</td>
</tr>
</tbody>
</table>

**Table 1: Some variables and the values they take on**

<table>
<thead>
<tr>
<th>61</th>
<th>48</th>
<th>58</th>
<th>27</th>
<th>57</th>
<th>39</th>
<th>37</th>
<th>57</th>
<th>47</th>
<th>60</th>
<th>55</th>
<th>67</th>
<th>51</th>
<th>42</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>84</td>
<td>61</td>
<td>53</td>
<td>91</td>
<td>93</td>
<td>33</td>
<td>33</td>
<td>49</td>
<td>80</td>
<td>74</td>
<td>52</td>
<td>71</td>
<td>74</td>
<td>55</td>
<td>45</td>
</tr>
<tr>
<td>68</td>
<td>35</td>
<td>43</td>
<td>30</td>
<td>45</td>
<td>83</td>
<td>60</td>
<td>36</td>
<td>63</td>
<td>61</td>
<td>37</td>
<td>82</td>
<td>58</td>
<td>51</td>
<td>58</td>
</tr>
<tr>
<td>62</td>
<td>64</td>
<td>26</td>
<td>37</td>
<td>84</td>
<td>41</td>
<td>40</td>
<td>54</td>
<td>56</td>
<td>46</td>
<td>43</td>
<td>93</td>
<td>45</td>
<td>39</td>
<td>69</td>
</tr>
<tr>
<td>60</td>
<td>37</td>
<td>53</td>
<td>84</td>
<td>29</td>
<td>54</td>
<td>80</td>
<td>34</td>
<td>47</td>
<td>45</td>
<td>31</td>
<td>44</td>
<td>56</td>
<td>66</td>
<td>95</td>
</tr>
<tr>
<td>59</td>
<td>49</td>
<td>59</td>
<td>33</td>
<td>31</td>
<td>44</td>
<td>46</td>
<td>76</td>
<td>43</td>
<td>69</td>
<td>47</td>
<td>61</td>
<td>95</td>
<td>60</td>
<td>32</td>
</tr>
<tr>
<td>90</td>
<td>33</td>
<td>42</td>
<td>37</td>
<td>69</td>
<td>39</td>
<td>56</td>
<td>58</td>
<td>48</td>
<td>27</td>
<td>49</td>
<td>81</td>
<td>84</td>
<td>46</td>
<td>47</td>
</tr>
<tr>
<td>50</td>
<td>48</td>
<td>60</td>
<td>57</td>
<td>27</td>
<td>27</td>
<td>32</td>
<td>42</td>
<td>91</td>
<td>54</td>
<td>65</td>
<td>58</td>
<td>82</td>
<td>36</td>
<td>61</td>
</tr>
<tr>
<td>57</td>
<td>62</td>
<td>69</td>
<td>34</td>
<td>51</td>
<td>55</td>
<td>44</td>
<td>62</td>
<td>61</td>
<td>43</td>
<td>51</td>
<td>42</td>
<td>93</td>
<td>80</td>
<td>56</td>
</tr>
<tr>
<td>37</td>
<td>36</td>
<td>39</td>
<td>58</td>
<td>47</td>
<td>30</td>
<td>30</td>
<td>67</td>
<td>59</td>
<td>54</td>
<td>34</td>
<td>63</td>
<td>51</td>
<td>73</td>
<td>52</td>
</tr>
<tr>
<td>66</td>
<td>52</td>
<td>42</td>
<td>69</td>
<td>84</td>
<td>43</td>
<td>67</td>
<td>83</td>
<td>44</td>
<td>83</td>
<td>80</td>
<td>44</td>
<td>56</td>
<td>78</td>
<td>33</td>
</tr>
<tr>
<td>47</td>
<td>92</td>
<td>39</td>
<td>48</td>
<td>58</td>
<td>58</td>
<td>30</td>
<td>75</td>
<td>48</td>
<td>76</td>
<td>53</td>
<td>48</td>
<td>54</td>
<td>44</td>
<td>46</td>
</tr>
<tr>
<td>46</td>
<td>66</td>
<td>45</td>
<td>37</td>
<td>69</td>
<td>72</td>
<td>46</td>
<td>33</td>
<td>93</td>
<td>84</td>
<td>92</td>
<td>90</td>
<td>84</td>
<td>53</td>
<td>81</td>
</tr>
<tr>
<td>53</td>
<td>90</td>
<td>52</td>
<td>45</td>
<td>31</td>
<td>43</td>
<td>43</td>
<td>45</td>
<td>75</td>
<td>68</td>
<td>39</td>
<td>43</td>
<td>71</td>
<td>81</td>
<td>48</td>
</tr>
<tr>
<td>59</td>
<td>91</td>
<td>35</td>
<td>60</td>
<td>74</td>
<td>67</td>
<td>48</td>
<td>52</td>
<td>39</td>
<td>69</td>
<td>69</td>
<td>92</td>
<td>49</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2: Chicago city engineer exam data**
Table 3: Sorted Chicago city engineer exam scores

<table>
<thead>
<tr>
<th>frequency value</th>
<th>raw relative</th>
<th>frequency value</th>
<th>raw relative</th>
<th>frequency value</th>
<th>raw relative</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>1</td>
<td>48</td>
<td>8</td>
<td>68</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>1.8%</td>
<td>49</td>
<td>4</td>
<td>69</td>
<td>8</td>
</tr>
<tr>
<td>29</td>
<td>0.4%</td>
<td>50</td>
<td>2</td>
<td>71</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>1.8%</td>
<td>51</td>
<td>5</td>
<td>72</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>1.3%</td>
<td>52</td>
<td>5</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>0.9%</td>
<td>53</td>
<td>5</td>
<td>74</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>2.2%</td>
<td>54</td>
<td>5</td>
<td>75</td>
<td>2</td>
</tr>
<tr>
<td>34</td>
<td>1.3%</td>
<td>55</td>
<td>3</td>
<td>76</td>
<td>2</td>
</tr>
<tr>
<td>35</td>
<td>0.9%</td>
<td>56</td>
<td>5</td>
<td>78</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>1.3%</td>
<td>57</td>
<td>4</td>
<td>80</td>
<td>4</td>
</tr>
<tr>
<td>37</td>
<td>3.1%</td>
<td>58</td>
<td>8</td>
<td>81</td>
<td>3</td>
</tr>
<tr>
<td>39</td>
<td>3.1%</td>
<td>59</td>
<td>4</td>
<td>82</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>0.4%</td>
<td>60</td>
<td>6</td>
<td>83</td>
<td>4</td>
</tr>
<tr>
<td>41</td>
<td>0.4%</td>
<td>61</td>
<td>6</td>
<td>84</td>
<td>7</td>
</tr>
<tr>
<td>42</td>
<td>2.2%</td>
<td>62</td>
<td>3</td>
<td>90</td>
<td>3</td>
</tr>
<tr>
<td>43</td>
<td>3.6%</td>
<td>63</td>
<td>2</td>
<td>91</td>
<td>3</td>
</tr>
<tr>
<td>44</td>
<td>2.7%</td>
<td>64</td>
<td>1</td>
<td>92</td>
<td>3</td>
</tr>
<tr>
<td>45</td>
<td>3.1%</td>
<td>65</td>
<td>1</td>
<td>93</td>
<td>4</td>
</tr>
<tr>
<td>46</td>
<td>2.7%</td>
<td>66</td>
<td>3</td>
<td>95</td>
<td>2</td>
</tr>
<tr>
<td>47</td>
<td>2.7%</td>
<td>67</td>
<td>4</td>
<td>95</td>
<td>1</td>
</tr>
</tbody>
</table>

Total 223 100.0%

Table 4: Raw and relative frequency distributions of Chicago test scores
Figures for Chapter 2

Fig. 1: (left) Possible values of the variable education level, measured in years of schooling completed. (right) Possible values of the variable blood pressure, in millimeters of mercury.

Fig. 2: Flow chart categorizing variable types into quantitative versus qualitative and discrete versus continuous.

Fig. 4: Linear (left) versus nonlinear (right) scales.

20 is twice as big as 10, but \( l_u(20) \) is much smaller than
Figure 3: Raw-Frequency Histogram of Chicago Exam Scores
Plots are identical, but the histogram on the right is twice as tall because it has twice as much data.

Fig. 6: Comparison of raw-frequency histograms based on samples of size $n = 223$ and $446$. The shapes of the two
3. Numerical Descriptive Methods

The idea behind this chapter is to look at four or five numerical methods people have found useful in descriptively summarizing variables one at a time. These methods complement but do not replace the graphical methods, like histograms, we examined in Chapter 2. A histogram is a great summary of any quantitative variable that has no time ordering, but sometimes, usually on grounds of convenience or saving space, people want to boil a variable down even further, to only one or two numbers plus a few words. Some care is required if we are to do that in a way that does not involve a big loss of information.

Center, Spread, and Shape. If you wanted to look at the histogram of some variable I had and I only wanted to give you a couple of numbers summarizing it, what kinds of information would you demand so that you could at least roughly reconstruct the histogram? The left part of Figure 1 shows two histograms that are identical except for their location on the number line. So I guess one thing you’d need would be a measure of location, or center, for the data. But that’s not enough: the middle part of Figure 1 shows two histograms with the same center and basic shape, but the two distributions differ in how spread out the values are around their middle. So another thing you’d need would be a measure of spread for the data.

Even those two are not enough: the right panel in Figure 1 gives two histograms with the same center and spread but fundamentally different shape, so I guess I have to add shape to the list. That should do it: center, spread, and shape. People have invented numerical descriptive summaries for all three of these things, and numerical ways to measure the first two—center and spread—are quite helpful, but I don’t find numerical measures of shape very useful in day-to-day data analysis; I prefer to describe the shape of a distribution with words. So I propose in this Chapter to tell you a bit about the most popular measures of center and spread, and arm you with the most frequently used words to describe the shape of a variable’s histogram.

1. Two Measures of Center—the Mean and the Median—and Some Words Describing Histogram Shapes

I live in the Los Angeles area, where the air quality is unusually bad by U.S. standards. Everybody in L.A. is quite conscious of this problem, because it’s important enough to actually change your behavior: on a bad day there are times
around rush hour when it would not be much fun to run five miles, say, or take a strenuous bike ride—your lungs burn, and you feel worse after you’ve exercised than before. The level of consciousness about air pollution is high enough, in fact, that the *Los Angeles Times* publishes a detailed summary every day of the previous day’s air quality and an in-depth forecast of the current day’s likely smog levels, and people actually study this information and plan their days accordingly. Figure 2 shows the *L. A. Times* smog report for Sunday, May 6, 1990.

For each of ten communities around the L.A. basin, the *Times* prints a graph called a *time series plot* that gives hourly levels of three pollutants—ozone (O3), nitrogen dioxide (NO2), and carbon monoxide (CO)—from 7 a.m. to 6 p.m. on the previous day, and a chart forecasting the maximum level of each of these pollutants, along with another one called particulates (PM10), on the current day for each of 13 locations. They also tell us how to interpret the numbers: each variable is measured on a scale called the *Pollution Standard Index* (PSI), with breakpoints at 50, 100, 200, and 275. From 0 to 50 is considered good air quality, 51-100 is moderate, 101-200 is called unhealthful, 201-275 is “very unhealthful,” and above 275 is “hazardous.” 200 and 275 have major behavioral consequences for businesses in the area as well as people: above 200, which is called a “first-stage alert,” some businesses are required to curtail their air pollution emissions, and above 275, a “second-stage alert,” certain businesses are actually shut down to try to diminish the problem. May 5th, 1990, the day whose ozone levels are plotted in Figure 2, was not terrible by L.A. standards, but May 6th (the day to which the box marked “Today’s Forecast” in Figure 2 applied) was predicted to be rotten, with many communities likely to have ozone levels over 100 and two places forecast to have first-stage alerts.

I want to focus on ozone, which is the major offender on the L.A. scene. Table 1 gives maximum daily ozone readings in metropolitan Los Angeles on 120 typical days in 1989. A histogram of these values is given in Figure 3. [gentle readers of the draft: I made these numbers up to serve a pedagogic point. I’m going to get the real numbers later.]

As I look at data like those in Table 1, I wonder things like the following: What was the typical ozone level in 1989? Give or take about how much? What percentage of the time was a first-stage alert called? In this section, which is about measures of center for quantitative variables, we will answer the first question several ways. Section 2, on measures of spread, answers the give-or-take question, and Section 3, on *percentiles*, answers the last question and introduces a useful graph called a *boxplot*. Section 4 covers numerical summary measures for qualitative variables.
The Arithmetic Mean. How should we define what we mean by a “typical” number on a given list? Probably the most frequently used definition is the one you heard about back in grade school—add the numbers up and divide by how many there are. The result is called the average, or arithmetic mean, or just the mean for short. The word “arithmetic” is in there to distinguish it from several other kinds of means we will run into later in the book, but this kind of mean is far and away the most often used, so whenever I just say “mean” it will be a shorthand for add-‘em-up-and-divide-by-$n$. Everybody uses the same notation for the mean of a variable: take the name of the variable, say $y$, and put a bar over it to get a symbol, $\bar{y}$, that’s read “$y$-bar.”

It’s convenient here and in the rest of the book to use summation notation when talking about means and things like them. As in Chapter 1, let’s call the first observation on the list $y_1$, the second $y_2$, and so on to the last, $y_n$ (here $y_{120}$). A generic element of the list, say the $i$th one down from the top, is then $y_i$. With this notation the mean comes out

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$ 

Here the sum of the 120 ozone values works out to 10729, so the mean ozone level across this sample of 120 days was $10729/120 = 89.4$ PSI. So that’s one answer to the question about the “typical” ozone level: ozone was around 89.

The Median. Another answer that you may have heard of is also frequently used. Imagine sorting the data from smallest to largest (Table 2), and then picking the middle number—the value such that half the data is below that number and half above. That ought to be another way to define “typical,” and it is: it’s called the median of a list of numbers.

Here when I sort the ozone data in ascending order and start looking for the middle number, I notice something a little funny about the idea: with an even number of observations ($n = 120$ in this case) there isn’t really a middle number, there are two middle numbers. To see what I mean, consider the smaller lists {47, 48, 49} and {47, 48, 49, 51} for concreteness. With the first of these lists, when $n$ is odd, there’s no problem; 48, the second number in from the left or the right, is unambiguously the middle. But with the second list, where $n$ is even, there are two numbers that could both lay claim to being the “middle”: 48 is the second in from the left, and 49 is the second in from the right. What people usually do to resolve this tiny conflict when $n$ is even is to average the two middle numbers. So the median of the 4-number list is 48.5, the average of the second-
and third-biggest values, and the median of the ozone data is 79.5, the average of the 60th- and 61st-biggest observations.

Defining the median symbolically in general requires a somewhat different notation than the symbolism we used for the mean, because we need somehow to denote the sorted data values rather than the raw data \( y_i \). People use the notation \( y_{(1)} \) to stand for the first observation after sorting the data from smallest to largest, \( y_{(2)} \) the second-smallest observation, and so on up to \( y_{(n)} \), the biggest data value. The sorted data values \( y_{(1)}, \ldots, y_{(n)} \) are called the order statistics; they play a role in the section on percentiles below and in the construction of many of the graphs we will encounter as we go along.

Defining the order statistics notation is a good beginning in specifying the median, but we still need to figure out which observation or observations the median is based on as a function of the sample size \( n \). There are evidently two cases: \( n \) odd and \( n \) even. Take \( n \) odd, and think about some examples. With \( n = 3 \) observations, as in the little dataset above, we used the second-smallest value; with \( n = 5 \) you would use the third-smallest; and so on. After some experimentation I guess the pattern is using the \( \frac{n+1}{2} \)-st smallest value. How about \( n \) even? In the 4-observation dataset above we used the average of the second- and third-biggest values; with \( n = 6 \) you would use the average of the third and fourth; and so on. So I guess the pattern with even \( n \) is to use the average of the \( \frac{n}{2} \)-st and \( \frac{n}{2} + 1 \)-st smallest values. Putting this together gives the official definition of the median of a variable \( y \), which we might as well call \( \tilde{y} \):

\[
\tilde{y} = \begin{cases} 
  y_{\left(\frac{n+1}{2}\right)} & n \text{ odd} \\
  \frac{y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)}}{2} & n \text{ even} 
\end{cases}
\]

Comparing the Mean and Median. It's a little bit interesting to compare the mean and median of the ozone data: the mean was about 89.4, and the median came out noticeably smaller, 79.5. To see why, it helps to work out the graphical meaning of both of these measures of center. It's easier to do this for the median, which we set up to be the place on the number line where half the data is below it and half above. Another way to say that is the median is the place where the relative frequency of data both to the left and the right of that place is 50%. When you recall that relative frequency shows up in density-scale histograms as area under the histogram, you see that the median is the 50/50 point in the distribution—the place that divides the area under the histogram in half (Figure 4).

The mean is trickier graphically. In fact, unless you've met the idea of center of gravity in physics I have to just tell you the answer, which depends
on an interpretation of the mean that’s useful in its own right. Developing this interpretation requires a bit of algebra. Recall the definition of the mean:

\[ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \]

Subtract \( \bar{y} \) from both sides, and multiply both sides by \( n \):

\[
0 = \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right) - \bar{y}; \quad n \cdot 0 = n \left( \frac{1}{n} \sum_{i=1}^{n} y_i - \bar{y} \right) = \left( \sum_{i=1}^{n} y_i \right) - n \bar{y}.
\]

To make my point I want to get the \( \bar{y} \) inside the summation somehow, which I can do by remembering that a fancy way to get \( n \bar{y} \) is to add \( \bar{y} \) to itself \( n \) times:

\[
n \bar{y} = \sum_{i=1}^{n} y_i, \text{ so } 0 = \left( \sum_{i=1}^{n} y_i \right) - n \bar{y} = \left( \sum_{i=1}^{n} y_i \right) - \left( \sum_{i=1}^{n} \bar{y} \right) = \sum_{i=1}^{n} (y_i - \bar{y}).
\]

The quantity \( (y_i - \bar{y}) \) inside the summation is called the deviation of the \( i \)th observation from the mean. It represents how far a given data value is from the mean, with a plus or minus sign to show whether the value is above or below the mean. In words the last statement above, \( \sum_{i=1}^{n} (y_i - \bar{y}) = 0 \), says that the mean is the place where all the deviations balance to zero. The connection with physics is made by imagining a long piece of plywood in place of the number line, with a bunch of bricks of equal weight placed on top of the plywood, one at each data value. Now imagine putting the whole thing on a sawhorse and looking for the balance point. The mean is the place where this contraption would balance, because the deviations represent the amount of force in each direction exerted by the bricks (data points) and the mean is where all these forces would cancel out. This idea is harder to say than to see: Figure 5 is a histogram of another small dataset, \( \{1, 2, 9\} \), with the deviations from the mean drawn in as arrows. The histogram balances at \( 4 = (1 + 2 + 9)/3 \), which is where the negative deviations -3 and -2 balance out the positive deviation +5.

Now I can finally finish comparing the mean and median. To do so requires thinking about the different sorts of shapes histograms could take on.

**Histogram Shapes.** What would have to be true for the mean and median to coincide? Somehow the 50/50 point in area would have to be the same as the balance point. If you sketch a few histograms and try to make that happen you will convince yourself that the only way to do it is to make the histogram have what people call a center of symmetry: a place where, if you folded the histogram in half vertically, the left and right halves would be mirror images. Such histograms are called symmetric, as in Figure 6.
Both the histograms in Figure 6 are symmetric, but they have quite different shapes: the one on the left has only one tallest point, whereas the one on the right has two. The tallest place on a histogram is called the mode of the distribution. It seems from Figure 6 that we ought to have some words that distinguish the left plot in that figure from the right plot, so people call the left histogram unimodal and the right one bimodal. Bimodality, or more generally multimodality (with three or more tallest places), often means that there's another variable lurking around that could be used to divide your data into several groups, which you could then study separately. An example is the righthand plot in Figure 6, which is my freehand attempt at a histogram of height of all adults: women are on average about 5 inches shorter than men, so the histogram has two modes, one around 5 feet 4 inches and another around 5 feet 9 inches. If you can find a variable like this you will usually want to use it to make separate unimodal histograms of each group, so I won't have much to say about multimodal distributions. But there will be a lot about symmetric unimodal histograms like the left panel in Figure 6; they are the types of distributions that have received the most study, in part because there's no ambiguity in how to summarize their centers: mean = median = mode = center of symmetry.

If a histogram is not symmetric (if it's asymmetric, people sometimes say) but unimodal, there's only two possible ways it could depart from symmetry, as in Figure 7: either the right tail is longer than the left, as in the right panel, or vice versa, as in the left plot. There's a variety of names to describe asymmetry: people call the histogram on the right positively skewed (since the positive deviations from the mean tend to be bigger than the negative ones), or skewed right, or they just simply say that such histograms have a long righthand tail. I like the last of these terms because it's so descriptive, so that's what I'll mainly use. The corresponding terms for the left histogram in Figure 7 are negatively skewed, skewed left, or having a long lefthand tail. [relationship between mean, median, and mode with unimodal skewed histograms; comparative *robustness* of median]

2. Measures of Spread: The Standard Deviation

I have tried to convince you that the typical ozone level in the dataset in Table 1 is either around 80 or around 90, depending on how you define “typical.” But give or take how much? What I mean by that question, which is about the spread of the ozone variable, is this: What is the typical amount by which the numbers in Table 1 differ, or deviate, from the middle of the dataset?
Well, just as there were several reasonable ways to define the center of a list of numbers, there are several plausible ways to turn this question into a numerical measure of spread. I need to specify three things: what I mean by “typical amount,” how I want to quantify the “deviations,” and which measure of “middle” I have in mind. If I only specified two alternatives for each of these choices, that would define \(2^3 = 8\) possible spread measures right there, and in fact people have studied at least three such possibilities in some detail. Only one of them is going to play a prominent role in the rest of the book. Which one (it’s called a “standard deviation,” which sounds a little like a built-in contradiction), and why it is the “chosen” one, will come out as we go along.

**Defining the Standard Deviation.** To get things going, let’s look again at Figure 5 and its little dataset \(\{1, 2, 9\}\). The measure of center used in that figure was the mean, 4, and that’s the choice people usually make to define the “middle” of the data when measuring spread. Focus again on the deviations from the mean in that figure—\((1 - 4) = -3, (2 - 4) = -2,\) and \((9 - 4) = +5\)—and recall that these appear graphically as arrows pointing away from the mean. Your first thought in summarizing how big the deviations are might be to average them, but that doesn’t work:

\[
\frac{(-3) + (-2) + (+5)}{3} = 0.
\]

We already know why it didn’t work—we proved a little while ago that the average of the deviations is always zero. This idea evidently needs fixing; any “measure of spread” that is always zero no matter what the dataset is could use a little work. So what went wrong?

I guess what happened is that we allowed the positive and negative values to cancel each other. In fact, when measuring the typical size of the arrows the plus and minus signs are irrelevant; we should get rid of them. Two easy ways to do that come to mind: take absolute values, or square the damn things. Simplest is to use absolute values:

\[
\frac{|-3| + |2| + |5|}{3} = 3.3.
\]

This is a legitimate measure of spread: it’s always nonnegative, and is only zero when all the numbers in the dataset are the same. It’s called the *average absolute deviation*, or AAD for short (its official name is the “mean absolute deviation from the mean,” to emphasize the measure of center used in defining the deviations). The AAD had some popularity in the early 1900s; probably its biggest fan was a famous British astronomer named Sir Arthur Eddington, who used it a lot to measure variability in [1]. But nobody uses it today; it’s been completely dominated by a spread measure based on the other idea for getting rid of the plus and minus signs—squaring the deviations instead of working with their absolute values.

You might think, given that you were going to use the squared deviations,
that (by analogy with the AAD) people would just average the squares:
\[
\frac{(-3)^2 + (-2)^2 + (+5)^2}{3} = \frac{38}{3} = 12.7.
\]

And in fact this is what people do under some conditions. But the official thing to do most of the time, for reasons that we'll get into below, is to divide not by the number of observations, \(n\), but by \((n-1)\) in forming the "average" of the squared deviations. The resulting measure of spread is called the sample variance, or just the variance for short:
\[
V(\{1, 2, 9\}) = \frac{(-3)^2 + (-2)^2 + (+5)^2}{2} = 19.
\]

So the AAD came out 3.3 and the variance of the same set of numbers is 19. That doesn’t sound to me like they’re measuring even roughly the same thing. When you look back at how they’re defined you see that they’re not—the problem is that to get rid of the plus and minus signs with the variance we squared all the numbers. That means that the AAD and the variance don’t even have the same units: if you have a variable measured in dollars, the AAD would come out in dollars but the variance would be in dollars squared, whatever that would mean. The fix is to take the square root of the variance, to get the units right again. The result is called the sample standard deviation, or just the standard deviation, and is the most frequently used measure of spread today. Everybody abbreviates it SD:
\[
SD(\{1, 2, 9\}) = \sqrt{\frac{(-3)^2 + (-2)^2 + (+5)^2}{2}} = \sqrt{19} = 4.4.
\]

This is still noticeably bigger than the AAD, but that’s okay; it turns out that they’re measuring somewhat different aspects of spread, so there’s no reason they should be the same. For many datasets the AAD is only about 80% as big as the SD, and in fact that’s pretty close to how it came out this time: AAD/SD = 3.3/4.4 = 0.76.

**Computing a Standard Deviation.** Symbolically on an arbitrary variable \(\{y_1, \ldots, y_n\}\) the formula for the SD implicit in the last equation is
\[
SD(\{y_1, \ldots, y_n\}) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}.
\]

From the viewpoint of calculation it’s too bad that the SD is so popular, because it’s not much fun to calculate SDs by hand. My suggestion for learning how
to do all the graphical and numerical things in this book is to do them once by hand to see how they go and then turn the detail work over to your statistical computing program from that point on.

If you have to compute an SD by hand it's particularly unpleasant to do so with the formula above, since you have to make two passes through the data: once to get the mean, and once to compute all the deviations from the mean and square them. By expanding out the squared term inside the sum and mucking about a bit with algebra (it's tedious to check this, so of course I've assigned it as a problem at the end of the chapter), you can get an alternate formula that's easier to use by hand:

\[
SD = \sqrt{\frac{\frac{1}{n} \left( \sum_{i=1}^{n} y_i^2 \right) - \frac{1}{n} \left( \sum_{i=1}^{n} y_i \right)^2}{n - 1}}.
\]

This way you only have to make one pass through the data, keeping track of the running sum of the \( y \)'s and the running sum of their squares. Table 3 lays out the computations for the ozone data.

So one possible answer to the first two descriptive questions I asked back on page [xx] is: In 1989 ozone in L.A. was around 89 PSI (the mean), give or take about 37 PSI (the SD). But what exactly do I mean by "give or take"? It would be nice to have a graphical interpretation of the SD to go along with those for the mean and median.

A Graphical Interpretation of the SD: the Empirical Rule. [Fisher and his dispute with Eddington; the triumph of the SD for the wrong reason; why people divide by \((n-1)\) instead of by \(n\): degrees of freedom. population SD and variance instead of sample SD and variance]

3. Percentiles and Boxplots

4. Numerical Descriptive Summaries for Qualitative Variables

5. Summary
6. Problems

[work in some history: EDA vs. "classical" descriptive methods, etc.]

Measures of center (mean, median, mode), measures of spread (average absolute deviation, standard deviation, interquartile range); percentiles; five-number summaries.
Tables for Chapter 3

| 57  | 83  | 92  | 237 | 51  | 70  | 53  | 165 | 116 | 78  | 161 | 129 | 70  | 74  | 156 | 93  | 53  | 66   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 82  | 127 | 65  | 58  | 77  | 88  | 65  | 49  | 86  | 68  | 143 | 75  | 100 | 77  | 103 | 63  | 145 | 82   |
| 87  | 66  | 76  | 64  | 100 | 95  | 80  | 70  | 156 | 67  | 110 | 64  | 105 | 85  | 65  | 79  | 59  | 76   |
| 85  | 70  | 198 | 81  | 65  | 81  | 70  | 94  | 65  | 72  | 56  | 53  | 57  | 47  | 118 | 91  | 167 | 94   |
| 152 | 95  | 110 | 80  | 222 | 105 | 59  | 83  | 86  | 53  | 50  | 57  | 154 | 56  | 72  | 69  | 55  | 106  |
| 117 | 83  | 52  | 49  | 80  | 168 | 78  | 161 | 63  | 57  | 79  | 89  | 53  | 145 | 123 | 83  | 65  | 119  |
| 69  | 134 | 74  | 62  | 92  | 69  | 72  | 102 | 92  | 55  | 48  | 107 |

Table 1: Maximum daily ozone values (in PSI) in L.A. on 120 typical days in 1989

<table>
<thead>
<tr>
<th>47</th>
<th>48</th>
<th>49</th>
<th>49</th>
<th>50</th>
<th>51</th>
<th>52</th>
<th>53</th>
<th>53</th>
<th>53</th>
<th>53</th>
<th>53</th>
<th>53</th>
<th>53</th>
<th>53</th>
<th>55</th>
<th>55</th>
<th>56</th>
<th>56</th>
<th>57</th>
<th>57</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>57</td>
<td>58</td>
<td>59</td>
<td>59</td>
<td>62</td>
<td>63</td>
<td>63</td>
<td>64</td>
<td>64</td>
<td>65</td>
<td>65</td>
<td>65</td>
<td>65</td>
<td>65</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>67</td>
<td>68</td>
<td>69</td>
<td>69</td>
<td>69</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>70</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>74</td>
<td>74</td>
<td>75</td>
<td>75</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td>77</td>
<td>77</td>
<td>78</td>
<td>78</td>
<td>79</td>
<td>79</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>81</td>
<td>81</td>
<td>82</td>
<td>82</td>
<td>82</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>83</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>85</td>
<td>86</td>
<td>86</td>
<td>87</td>
<td>88</td>
<td>89</td>
<td>91</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>94</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>103</td>
<td>105</td>
<td>105</td>
<td>106</td>
<td>107</td>
<td>110</td>
<td>110</td>
<td>116</td>
<td>117</td>
<td>118</td>
<td>119</td>
<td>123</td>
<td>127</td>
<td>129</td>
<td>134</td>
<td>143</td>
<td>145</td>
<td>145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>152</td>
<td>154</td>
<td>156</td>
<td>156</td>
<td>161</td>
<td>161</td>
<td>165</td>
<td>167</td>
<td>168</td>
<td>198</td>
<td>222</td>
<td>237</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The ozone readings, sorted from smallest to largest

<table>
<thead>
<tr>
<th>observation number (i)</th>
<th>y</th>
<th>y</th>
<th>2</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i</td>
<td>i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>57</td>
<td>3249</td>
<td></td>
<td>10729</td>
</tr>
<tr>
<td>2</td>
<td>83</td>
<td>6889</td>
<td></td>
<td>1121637</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>10729/120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1121637 - 10729 /120</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SD = sqrt(---------------) = 36.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>119</td>
</tr>
<tr>
<td>120</td>
<td>107</td>
<td>11449</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Using the shortcut formula to get the SD of the ozone data.
Fig. 1: Three ways to contrast histograms: (a) identical except for center; (b) same center and shape but different spread; (c) same center and spread but different shape.

Fig. 2: The Los Angeles Times air quality report on Sunday, May 6, 1990 (reprinted with permission).
Figure 3: Histogram of the 120 L.A. Ozone Readings

Fig. 4: The median is the 50/50 point in area under the histogram.
Fig. 5: The mean is the place where the histogram balances.

Fig. 6: Two symmetric histograms. The one on the left is unimodal, the one on the right is bimodal.

Fig. 7: Two skewed histograms, one with a long lefthand tail (left), the other with a long righthand tail (right).
4. The Normal Curve

In the last two chapters I have tried to sell you on the usefulness of two types of methods—graphical and numerical—for descriptively summarizing one variable at a time. The main thread that I hope emerged for quantitative variables was starting with a histogram or stem-and-leaf plot, noting the shape of the distribution, and then choosing numerical summaries that are reasonably well matched to that shape: means and standard deviations for variables that are more or less symmetric unimodal, medians and other percentiles for variables with long tails (you may also want to report the mean and SD for skewed distributions, since people are so accustomed to center and spread being summarized in that way). I mentioned in Chapter 3 that the histogram shape that has been most studied so far is symmetric unimodal. It turns out that there is one particular form of symmetric unimodal histograms that comes up so often it merits a whole chapter of its own. It’s called the normal, or Gaussian, distribution (or sometimes just the normal curve or bell curve when people are speaking informally), and the point of this chapter is to gain some experience in working with it. The idea is that if you get this stuff under your belt now, there will be one less thing to worry about later, when it will be only one of many ingredients in the solution to more complicated problems.

1. Some History

In the mid-1800s a Belgian mathematician and astronomer named Adolphe Quetelet tried to make a catalogue of histogram shapes for lots of variables that came up in real life, like people’s heights and the lengths of their forearms. (Yes, I know you probably don’t think it’s very interesting to go around measuring people’s forearms, but Quetelet had a good reason for doing it—people were in the early stages of trying to figure out how human heredity worked, and it was useful to look at things like the relationship between the length of a child’s forearm and the lengths of his or her parents’ forearms. That doesn’t make it any more interesting to do the measuring, but it does at least make it worthwhile.) Quetelet noticed that, wherever he looked, he kept getting the same histogram shape—a smooth curve with the same functional form kept providing a quite good approximation to his histograms, when drawn on the density scale.

Quetelet didn’t have to invent this curve; it had been around since the early 1700s, when a French mathematician named Abraham de Moivre and a couple of Swiss mathematicians from the same family, Jakob and Daniel Bernoulli, had used it to prove the single most important theorem in statistics, a result called the
Central Limit Theorem (the subject of Chapter 9). Later, around 1800, a German mathematician named Carl Friedrich Gauss and a French mathematician, Pierre Simon de Laplace, working independently of each other, used the same curve as the basis of an extremely useful theory called least squares that we will look at in Part VIII. Quetelet put the curve to a different use than these other people did (we will see what the others did with it later). He did things like making a histogram of the heights of 5,738 Scottish soldiers, after someone else had painstakingly put together the dataset, and approximating the histogram with the curve. He would have called the de Moivre-Bernoulli-Gauss-Laplace curve he used for this purpose the "error law" (we will see why he used this name for it later, when we talk more about measurement error models), but we use two different names for it today: the normal, or Gaussian, distribution. Don't ask me why only Gauss gets the credit in the modern choice of names.

2. The Form of the Normal Curve

The formula for the normal curve is interesting. In the left panel of Figure 1 I have simulated Quetelet's dataset and drawn the normal curve that he might have used, to approximate the heights of his Scottish soldiers, on top of the histogram of the simulated data. I have guessed that the average height of these soldiers might have been about 5 feet 7 inches (this was only the mid-1800s, remember) with a standard deviation of about 2.5 inches.

The first thing we need to get straight about the normal curve is that, if it is to be a good approximation to a wide variety of variables, there can't just be one normal curve; there have to be a lot of them, one for each choice of where the variable is centered and how spread out it is. After all, if I want to use it, as psychologists sometimes do, to approximate the distribution of IQ scores of adult Americans, say (mean 100 IQ points, SD 20 IQ points), and Quetelet wants to use it on the heights of his soldiers (mean 67 inches, SD 2.5 inches), we're going to need normal curves with different means and SDs. One nice thing about normal curves is that's all you need; in other words, if I tell you the mean and the SD of a variable and say "the histogram follows the normal curve pretty closely," we will see below that you can reconstruct the variable quite well, in the sense that you could use a computer to simulate the variable and it would be hard to tell the simulated variable from the real one.

To motivate the form of the normal curve, I want to think about a simpler normal curve than Quetelet's, namely the one with mean 0 and SD 1, which is drawn in the right panel of Figure 1 (why I made those choices for mean and SD
will come out in a little while). If I asked you to figure out the functional form of this curve from its plot, how would you do it? (The next few paragraphs in smaller print are mathy; you can skip them if you’re not interested.)

Well, to begin with, it’s symmetric about 0, so, if I call the values along the horizontal axis \( y \) and the height of the curve \( f(y) \), I want a function \( f \) such that \( f(-y) = f(y) \) (a so-called even function). I also want it to have its highest point at 0, so its first derivative should be zero at \( y = 0 \), and if you look at its concavity/convexity (its bowl-shaped-up- and down-ness, that is) you see that its second derivative should be negative to the left and right of zero out to some points \( c \) and \(-c\), and positive further away from zero than that, so that \( c \) and \(-c\) are inflection points (from Figure 1(b) \( c \) looks to be just about exactly 1). The curve can’t ever be negative, because relative frequency is never less than zero. It may look from Figure 1(b) as though the function is zero for \( y \) bigger in absolute value than around 3.5, but it’s more convenient mathematically to just let it die down toward zero as \( y \to \pm\infty \) without ever reaching zero (in other words, it’s asymptotic to the horizontal axis: \( f(y) > 0 \) for all \( y \), but \( \lim_{y \to \pm\infty} f(y) = 0 \)). That’s about as far as you can go qualitatively in pinning down the functional form of the curve.

Interestingly, this turns out to be enough to narrow things down considerably. In fact, if you sit down with a good graphics program and try to build a function that satisfies all these constraints out of simple familiar functions (I encourage you to spend a little time doing this if you have access to such a program), you would see if you worked at it long enough that there are only two real classes of contenders: ratios of polynomials, like \( \frac{1}{(1+y^2)^n} \), and exponential functions like \( e^{-by^2} \) or \( e^{-\frac{y}{s(1+e^{-\frac{y}{s}})}} \). If you try it, don’t feel bad if it takes you quite a while to arrive at this view. One of the leading statisticians of the late 1800s and early 1900s, a British scientist named Karl Pearson (whom we will revisit several times in later chapters), tried to make a catalogue of all the useful curves that closely approximate histogram shapes actually arising in practice. It took him years to do so, and when he was done these two classes of functions were the only ones he could find that approximated symmetric unimodal distributions using everyday functions as building blocks.

If you had the patience to play around with the three functions I mentioned in the last paragraph, varying \( k, b, \) and \( s \), you would find that you can choose those constants so that it’s quite difficult to tell the three curves apart, so this is as far as I can go (for now, at least) without just telling you which curve Gauss, Laplace, and Quetelet all found so compelling (the envelope, please):
\[ f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}. \] (1)

This is the formula of the function plotted in the right panel of Figure 1—the normal distribution with mean 0 and SD 1. (The other two functions I mentioned two paragraphs ago also come up in statistics as smooth curves representing histograms. The last of the three functions in that paragraph, \( \frac{e^{-\frac{x^2}{2}}}{s(1+e^{-\frac{x}{s}})} \), is called the logistic distribution; it underlies a technique called logistic regression that I hope to cover in the sequel to this book. The other function in that paragraph, \( \frac{1}{(1+y)^r} \), is called the t distribution; it has played a big role in inference since about 1905, and we will study it in Chapter 11.) The 1/2 in the exponent in equation (1) is there so that the inflection points come out at +1 and -1, as in Figure 1(b), and the stuff in front of the exponential is to make it so that the area under the curve is 1, or 100%—in other words, so that the curve is drawn on the density scale.

We won’t use the functional form of the normal curve (equation (1) above) much in this book. The main reason has to do with the principal use to which people like Quetelet put the curve: to approximate data histograms like the heights of the Scottish soldiers (you might look back now at the left panel in Figure 1).

Mostly when you’re given a dataset like Quetelet’s and asked descriptive questions about it that go beyond simply quoting the mean (67 inches) and SD (2.5 inches), those questions take the form, “What fraction of the soldiers were over 6 feet tall?” If I gave you the raw data and asked you to answer that question, you would just count how many of the 5,738 soldiers were over 6 feet and divide by 5,738—in other words, you would work out the relative frequency with which the description “over 6 feet tall” applied to the people in the sample. If I gave you a density-scale histogram of the 5,738 heights instead of actually giving you the raw data, you would recall from Chapter 2 that relative frequency information is conveyed in such histograms by means of area under the histogram; in this case you would try to figure out the area under the histogram to the right of 6 feet. Quetelet’s idea was to use the normal curve to approximate the histogram, and to work out areas under the normal curve to answer relative frequency questions. So now I’ve finally come to why we won’t use equation (1) much in the book: getting areas under a function requires integrating the function, and it turns out that equation (1) is not integrable in closed form.

Well, that really sounds like progress: we start out wanting to answer a simple question like the fraction of people over 6 feet tall, and Prof. Quetelet converts the problem into one that has no solution. Actually, it’s not that bad—if you really want to know the area under some curve and you can’t use calculus to
work it out, you can always approximate it numerically. That’s what Quetelet did: he used numerical integration to build up a table of areas under the normal curve (in fact, somebody else had already made such a table; he just used it). Our job in the rest of this chapter is to get good at using such a table.

3. Using the Normal Curve to Approximate Relative Frequencies

Wait, there’s another problem, you say. I thought you told me that there wasn’t just one normal curve—there’s one for every possible choice of mean and SD. Wouldn’t Quetelet have to somehow generate an infinite number of tables, one for each normal curve? If so, the idea is doomed again.

Fortunately not—there’s a neat way to get by with only one table. The argument proceeds in two steps: working out the areas for one reference normal curve, called the standard normal distribution; and relating areas under any other normal curve to areas under the standard curve. I guess the first step is to choose the mean and SD that define the standard curve. Consider the mean. If this normal approximation idea is to work for all possible datasets, the range of possible values for the mean is all the way from $\infty$ to $-\infty$. Among all numbers along the entire real number line, if you had to pick just one to serve as a reference you would probably pick 0, and that’s what everybody else does, too. Now what about the SD? Again, if this approximation idea is to work with all datasets, the SD needs to range all the way from zero (a dataset with no dispersion, where all the data values are the same) to $\infty$. Among all nonnegative numbers, the one that stands out if you had to choose just one as a reference would be 1, and again that’s what everybody picks. So by general agreement the standard normal curve has mean 0 and SD 1, and somebody has worked out areas under this curve and put them in a table like Table A1 in the back of the book. You should turn to that table now and look it over.

Working with the Standard Normal Curve. When you’re making a standard normal table like Table A1, you have a choice: do you want to give people areas to the left of some place along the line, say $z$, or areas to the right of $z$, or areas in the middle between $-z$ and $z$? Fortunately, it works out that all three of these choices are equivalent—if you give me a table with areas to the left of $z$ (left tail areas, we might have said back in Chapter 2), I can work out any middle areas or right tail areas I want. This works because of two properties of the standard normal curve:
• Fact (1): Since all normal curves are symmetric about their means, the standard normal curve is symmetric about 0, so that the area to the left of \(-z\) and the area to the right of \(z\) must be the same for all \(z\) (Figure 2(a)); and

• Fact (2): Since the standard normal curve is drawn on the density scale, the total area under it is 1, or 100\% (Figure 2(b)).

In Table A1 I have chosen to give middle areas, that is, areas from \(-z\) to \(z\), for \(z > 0\). You should watch out for this when you’re reading a statistics book—some people give left tail areas, some people give the right tail, some even just give the area from 0 to \(z\). When you’re using a normal table, it’s worthwhile to pause and figure out what kind of areas it gives you before doing any calculating.

All right, let’s get some practice using the table. What’s the area from \(-1\) to \(+1\) under the standard normal curve? That’s easy: look up 1 under the “\(z\)” column and read the corresponding entry in the “area” column: 68.3\%, which for most purposes you would just report to the nearest percentage point as 68\% (Figure 3(a)). How about from \(-2\) to \(2\)? The area entry under \(z = 2\) is 95.5\% (say 95\%; Figure 3(b)). Those two numbers should sound familiar from Chapter 3: they’re the magic numbers in the empirical rule. The normal curve is evidently the reference distribution underlying the empirical rule. (There’s nothing noteworthy about the empirical rule working for the normal curve; since it was the source of the numbers 68\% and 95\%, the rule must work for the normal curve. As I said in Chapter 3, the remarkable thing about the rule is how well it works for a wide variety of histograms that look nothing at all like the normal curve.)

Here’s one that requires a bit more work: What’s the area to the left of \(-1.5\) under the standard normal curve? The table only gives middle areas—for instance, the area from \(-1.5\) to \(+1.5\) is 86.6\%—so we can’t just lazily read the number we want right from the table this time. I guess now would be a good time to try to use those two facts about the standard normal distribution I mentioned above. If I call the area I want (to the left of \(-1.5\)) \(A\), fact (1) tells me that the area to the right of \(+1.5\) must also be \(A\), and fact (2) says that the area to the left of \(-1.5\), plus the area in the middle from \(-1.5\) to \(+1.5\), plus the area to the right of \(+1.5\), must be 100\%. This yields a little graphical equation in \(A\): \(A + 86.6\% + A = 100\%\) (Figure 4), so that \(A = (100\% - 86.6\%)/2 = 6.7\%\), which you would probably report as 7\% to avoid overstating the accuracy of the calculation.

After a little practice with this, you find yourself reasoning more or less automatically as follows: “Let’s see, to find the area to the right of 2.15 (say), I look up the area in the middle (96.8\%), subtract this from 100\%, and divide by two to get 1.6\%.” The worst it ever gets is areas from \(-b\) to \(+c\), where \(b\) and \(c\) are different (for example, the area from \(-0.6\) to \(+1.7\)), which requires looking up
two middle areas in the table and mucking them about using the two normal curve facts (try it). You should practice calculations like this, for example by doing the relevant problems at the end of the chapter, to make sure you know how to get areas under the normal curve pretty smoothly; we’re going to be calculating such areas for most of the next 18 chapters.

Converting to Standard Units. So this was step (1) in the grand plan: figure out areas under the standard normal curve. All that’s left when you have a dataset with some mean different than 0 and SD different than 1 (which is virtually always) is the second step I mentioned above: establishing a correspondence between the normal curve with your mean \( \bar{y} \) and SD \( s \) and the standard normal curve. The key connection we need is based on a simple graphical device, which I will demonstrate on Quetelet’s heights of 5,738 Scottish soldiers.

If the empirical rule is to work on Quetelet’s dataset, which has a mean of 5 feet 7 inches (67 inches) and an SD of 2.5 inches, then the area one SD either way from the mean—that is, the area from 67 – 2.5 = 64.5 inches to 67 + 2.5 = 69.5 inches—should be about 68%, the same as the area from –1 to +1 under the standard normal curve. Similarly, the area two SDs either way from the mean—from 67 – 2*2.5 = 62 inches to 67 + 2*2.5 = 72 inches—should be about 95%, the same as the standard normal curve area from –2 to +2. Evidently there’s a direct connection between values in Quetelet’s dataset and values along the number line for the standard normal curve: 62 inches corresponds to –2 on the standard normal curve, 64.5 to –1, and so on up to 72 corresponding to +2. If we call the inches scale on which the data came to us the raw units scale, we can graphically convey this correspondence by drawing another line below it called the standard units scale and lining up –2 with 62, etc., as in Figure 5:

If you look at this figure for awhile and think about what’s going on, you will see that standard units just measure how far a given data value is above or below the mean, with the SD as the yardstick. A plus sign in standard units signifies that the raw value is above the mean, a minus sign that it’s below the mean. For example, 72 inches is two SDs above the mean, so it comes out +2 in standard units; 64.5 inches is one SD below the mean, so it works out to –1 in standard units; and so on. With a bit more thought you can see that the algebraic relationship between raw and standard units must be as follows:

\[
\text{standard units} = \frac{(\text{raw value}) - \text{mean}}{\text{SD}}. \tag{2}
\]

This formula captures the yardstick idea exactly: the numerator represents “how far a given data value is above or below the mean,” and the denominator
(the SD) is the yardstick. People usually use the symbol $z$ for standard units (I did so above, and in Table A1), and sometimes use the name $z$-scores instead of standard units. With $y$ as the raw value, $\bar{y}$ as the mean, and $s$ as the SD, a fancier-looking way to write formula (2) above is then

$$z = \frac{y - \bar{y}}{s}. \quad (3)$$

Notice this means that $z$-scores are pure numbers, without units: the raw values’ measurement scale, like inches for Quetelet’s heights, washes out in the conversion to standard units. In Quetelet’s case, for example, both the numerator and denominator are in inches and the units cancel. Notice also that the mean on the raw scale always comes out zero in standard units.

Applying this formula finally finishes Quetelet’s answer to the original simple question: What fraction of the soldiers were over 6 feet tall? To put all the steps together in using the normal approximation to answer such a question, first I sketch a normal curve with the same mean and SD as the data, in this case 67 inches and 2.5 inches, respectively, and then I darken in along the number line the region of interest, in this case 72 inches (6 feet) and above. Next I draw the standard units scale underneath the raw units, maybe labeling a few easy places like the mean (0 in standard units) for concreteness. Then I convert the value or values defining the left- and/or right-endpoints of the region of interest to standard units, and label them on the lower axis. Finally, I look up the relevant values in the normal table, read off the areas, and use the two facts about the standard normal curve mentioned above to get the area I want. Figure 6 summarizes these steps for Quetelet’s data.

The answer, after all these pages, comes out pretty small: only 2.5% or so of Quetelet’s soldiers were over 6 feet tall, according to the normal curve, at least. In my simulated dataset whose histogram appears in Figure 2 above, 155 of the 5,738 “simulated soldiers” were 6 feet tall or taller, and 155/5738 is about 2.7%, so the approximation was quite good in this case, but that need not be so (more on this in the next section). You may think this whole normal approximation thing was like using a sledgehammer on an ant, and if our only goal was to answer the 6-foot question in Quetelet’s dataset you would be right, but the method is much more useful than that, as I hope to show you in the chapters to come.

**Percentiles.** If we wanted to, we could run equation (2) above backwards and solve for raw units in terms of standard units:

$$\text{raw value} = (\text{standard units}) \times SD + \text{mean},$$
or, in symbols,

\[ y = z s + \bar{y} \].

This is useful for answering a kind of question that you might say is backwards from the 6-foot question above. I'll lay this out in the context of the other normal approximation example I mentioned in passing several pages ago: IQ scores of adult Americans. Given that IQ scores average 100 with an SD of 20 and follow the normal curve pretty well, what is the 90th percentile of the IQ distribution? This question is backwards from the 6-foot question, because with the IQ question I'm starting with a percent and trying to get a raw value, whereas with the 6-foot question I started with a raw value and tried to get a percent. [finish this]

4. The Quality of the Normal Approximation

["Normal" distribution named by Pearson; unfortunate name, since so many variables are non-Gaussian.

You can always do the normal approximation to any dataset; nothing guarantees that it will be any good.

Give example of skewed distribution with rotten normal approximation, e.g. something with a long right tail that starts at zero for which the normal approximation to \( P(Y < \text{small}) \) is too big because it has significant mass \(< 0\).]

5. Summary

6. Problems

Uniform and exponential distributions too? [no; put them later] References: Theodore Porter, The Rise of Statistical Thinking, 1820-1900; Stigler
Figure 1: Quetelet's normal curve, approximating his histogram of the heights of 5738 Scottish soldiers, and a simpler normal curve, with mean 0 and SD 1.

Figure 2: (a) The area under the standard normal curve to the left of \(-z\) = the area to the right of \(z\). (b) The total area under the standard normal curve is 100%.

Figure 3: The area in the middle under the standard normal curve (a) from -1 to 1 is 68%, and (b) from -2 to 2 is 95%.
Figure 4: Working out the area under the standard normal curve to the left of -1.5.

\[ 86.6\% \text{ (from Table A1)} \]

100\% - 86.6\% = 13.4\%
and area in left tail must be \[ \frac{13.4\%}{2} = 6.7\% \approx 7\% \]

Figure 5: The raw units and standard units axes with Quetelet's data.

Figure 6: Using the normal approximation to estimate the fraction of Quetelet's soldiers with heights over 6 feet.
Part III. Gathering Data
5. An Introduction to Experimental Design

This part of the book is about how to gather information to advance scientific knowledge or make a decision. Recall back in the Introduction (Chapter 1) I said that people distinguished between two ways to gather data for such purposes: experiments and sample surveys. This chapter is about how to design the former, and Chapter 6 is about how to design the latter. Part II (Chapters 2-4) was about description (factuals); this part of the book looks ahead to inference and prediction (counterfactuals).

As usual I want to bring out the ideas in the context of one or two case studies. The one I’ve chosen for the first part of this chapter concerns an important subject for both science and policy in the last two decades of the 20th century: AIDS. You probably know a fair amount about AIDS already—how could anybody living in the 1990s who reads the paper not know a lot about it? Briefly, AIDS (Acquired ImmunoDeficiency Syndrome) is a disease caused by an organism called the Human Immunodeficiency Virus (HIV) that kills people by weakening their ability to withstand opportunistic infections and by lowering their resistance to cancer. It is mainly sexually transmitted, although you can also get HIV from things like blood transfusions, and (at this writing) once a person passes from just having the virus (being HIV-positive, as people say) to manifesting symptoms of AIDS the disease is invariably fatal. In 1990 more than 5 million people worldwide were infected with HIV, and that number is expected to grow substantially over the next several decades unless a vaccine is developed. By 1989 [80,000?] people had died of AIDS in the U.S. alone.

A treatment that would keep HIV-positive people from developing AIDS was urgently needed from the moment the disease was identified in 198x. One natural idea was to find a drug that stimulated the immune system, and one such drug that had been shown in the mid-1980s to work in the laboratory (but not yet in people) was something called isoprinosine. The exact mechanism by which it caused immune system stimulation was not clear as of 1990, but it seems to work by improving the production of interleukin-1 and -2, [explain], and by enhancing the ability of immune-system components called T-cells to find and destroy invading viruses. On June 21, 1990, writing in one of the world’s premier medical journals, The New England Journal of Medicine, a group of researchers from Denmark and Sweden headed by a doctor named Pedersen announced that they had successfully used isoprinosine to delay the progression to AIDS in people infected with HIV.
We are going to think along with them as they design the experiment on which their results were based.

1. Controlled Experiments

One way to identify the ideas underlying the design of a good experiment is to look at a bunch of bad experiments and figure out what is wrong with them. So I am going to present several designs to you, each one better than the last, and ask you to close the book each time we come to the symbol *** and mentally criticize them before I do so. The narrative will take the form of a dialogue between an investigator attempting to work out a good experimental plan for the Pedersen study and a devil’s advocate who enjoys poking holes in the investigator’s ideas. From time to time I will stick in my two cents in brackets [like this]. As we go along you should try to figure out what the basic principles are—what defines good and bad in the design of an experiment. I will summarize these principles at the end of this section.

Investigator. Okay, here’s my first design.

**Design 1.** I get a bunch of people who are HIV-positive, and give isoprinosine to all of them. I wait awhile and see how many of them get AIDS.

Devil’s Advocate. You have to be more specific than that. What do you mean by “a bunch” and “awhile?” Where are you going to get your patients?

Investigator. Good point. On how many people to get, I guess it depends on how accurate I want my final answers to be—presumably if I do it right, the more people I have, the more accurately I will know the degree to which the drug prevents or delays AIDS. [This is a nontrivial topic—sample size determination—that we will take up later, in Chapter 16.] For concreteness, let’s say I did this to 400 HIV-positive people.

On where to get these 400 people, I guess that depends on how generalizable I want my findings to be. If I want my results to be relevant to all 5 million HIV-positive patients worldwide, then somehow I had better choose these 400 people so that they are “just like” the other 4,999,600 HIV-positive people I didn’t pick, in ways that are relevant to getting AIDS. That sounds hard, so for now I’m just going to get 400 patients—who agree to take part in the study—from AIDS clinics and medical centers around Scandinavia run by
investigators willing to participate. Maybe toward the end of the study I’ll try to figure out how the people I got differ from the ones I didn’t get, and if they do differ in relevant ways maybe I can adjust for the differences somehow.

On how long to wait, that depends on the disease’s incubation time. Ideally, given that I hope the drug will delay AIDS indefinitely, I should watch people for quite awhile, but if it really does work I should announce this so that HIV-positive people around the world can start taking the drug (besides, this discovery might make me famous, and I’m eager to publish). Let’s say half a year. [Pedersen and his colleagues actually used a 24-week followup period.]

So let me summarize Design 1 for you again: I get 400 HIV-positive patients who agree to participate, from medical centers in Scandinavia willing to take part in my study, I give isoprinosine to all 400 patients, and I look to see how many of them have AIDS after half a year. Let’s say only 2 of them got AIDS. The drug sounds pretty good to me. ***

*Devil’s Advocate.* Well, first of all, half a year doesn’t sound long enough. But the big problem with your design is that it’s like that old battle-of-the-sexes joke about the woman who is asked “How’s your husband?” and she replies “Compared to what?” You don’t have any point of comparison in your setup—how can we tell if an AIDS incidence of 2 people out of 400 in six months is low, when you haven’t tried to estimate what the incidence for these 400 people *would have been if they hadn’t gotten the drug*? You need what people call a control group—a bunch of HIV-positive patients who don’t get isoprinosine, to compare to your treatment group, the people who do get the drug. Maybe nobody in the control group will get AIDS, and your 2 people out of 400 won’t look so impressive.

*Investigator.* Okay, okay, you’re right. I guess I’ll need another 400 people for the control group.

*Devil’s Advocate.* You’ve just spotted yourself 800 patients. [In fact Pedersen had 831 people, 412 in treatment and 421 in control.] How come you took the 800 people and divided them 50/50 into treatment and control? What about the dose in the treatment group? What are you going to give the control people if you don’t give them isoprinosine? Shouldn’t they get some kind of pill too, so that nobody knows which patients are in which groups?
Investigator. Geez, you’re pushy. I don’t know why I picked 50/50; it just seems intuitively like the best thing to do. [Later on we will work out if this is right.] Nobody knows the best dosage of this drug—probably we should have more than one treatment group, each with a different dosage, but I don’t have enough money to do that, so I’m just going to take my best guess based on the literature: 1 gram, three times a day. As for what to give the control group, what are you worried about? Why can’t I just give them nothing?

Devil’s Advocate. Two reasons—it would be better for the patients not to know which treatment group they’re in, and it would also be better for the people evaluating who has AIDS at the end of the observation period not to know which group anybody’s in. As far as the patients are concerned, I know it sounds farfetched, and it may not be very important in your study, but sometimes people with a disease get better when all they’re given is a pill with no drug in it—what people call a placebo.

Investigator. Oh, come on.

Devil’s Advocate. No, really. The mind is sometimes capable of helping the body find healing powers it ordinarily doesn’t have access to—it’s called the placebo effect. My grandmother was a hypochondriac who could make herself physically ill in ways that mimicked whatever disease was being described to her—if the mind can make you sick, why can’t it help heal you? Anyway, I don’t care if you believe me—the phenomenon has been well documented, everybody else in science but you knows about it. I think you should give all the patients a pill that looks and tastes the same, except that the treatment people’s pills have isoprinosine in them—the official term for this is blinding the patients to their treatment status.

Investigator. All right, enough. What about the other part of your complaint—something about keeping the doctors who are going to assess who has AIDS at the end of the experiment from knowing who’s in the treatment group and who’s in control?

Devil’s Advocate. This is a more serious issue in your experiment. I’m worried that someone who consciously or unconsciously wanted the drug either to succeed or to fail could shade his or her judgments about who had AIDS, or about which group looked better in the data analysis, to make
the drug look good or bad. So I think you should blind the investigators as well as the patients to who’s in treatment and control—give them code numbers or something, with nobody knowing the code until after the data are analyzed. When both the patients and the researchers are blinded to everybody’s treatment status, it’s called running the experiment double-blind.

Investigator. Were you born this suspicious, or did somebody do something real mean to you when you were a kid?

Devil’s Advocate. No, really, double-blinding is a good idea, especially when you expect your results to be controversial. People who have a vested interest in other treatments or who are just curmudgeons like me are going to try to poke holes in your design and analysis, and you should spend some effort at design time anticipating their criticisms and doing what you can to legitimately blunt them. What use is a study with valid findings that nobody accepts?

Investigator. All right, I’ll do it, if only just to shut you up. Here’s my new design [we’ll use Pedersen’s sample sizes from now on]:

Design 2. I get 831 HIV-positive patients who agree to participate, from medical centers in Scandinavia willing to participate, I divide them roughly 50/50 into a treatment group of 412 patients who get isoprinosine and a control group of 421 people who get a placebo, I blind both the patients and the investigators to everybody’s treatment status, and I look to see how many of them have AIDS after half a year.

Let’s say only 2 of them got AIDS in the treatment group versus 17 in the control group [which is what actually happened in the Pedersen study]. Are you convinced now? ***

Devil’s Advocate. It depends on how comparable the treatment and control patients were to begin with in their likelihood of getting AIDS. You didn’t tell me how you assigned people to the experimental groups—maybe you put all the sickest people in the control group.

Investigator. There you go with that nasty mind of yours again. Of course I’m not going to stack the deck in favor of the treatment. I’ll just throw the
831 people into treatment and control any old way—isn’t that pretty much guaranteed to even things out pretty well?

_Devil’s Advocate._ You seem to agree with me that comparability of the two groups at the beginning of the experiment is a good idea. Okay, that means you have two goals—achieving this comparability, and doing so in a way that convinces other people that you succeeded. Haphazard assignment seems like it might work, and if you really do it haphazardly you might succeed in getting approximate comparability, but your critics might well not believe you, and in any case this is another good chance to protect you from yourself.

_Investigator._ So, insults too?

_Devil’s Advocate._ No, no, let me tell a story on myself that explains what I mean. When I was a freshman in college I took a psychology class in which our final project was to design and carry out an experiment. _[This really happened to me.]_ The instructor had invented a computerized logic game called “Zaps and Duds,” where you typed in a string of six X’s and O’s and the computer classified each string as a “zap” or a “dud” according to a rule it had chosen that was based on the X-O pattern in two of the six positions. Your job was to work out what rule the computer had in mind (“anything with an X in the third position and an O in the fifth position was a zap, everything else was a dud,” for example), and to do so while typing in the smallest number of strings of X’s and O’s for it to classify for you. Unknown to you, however, to minimize the role of guesswork the computer had been programmed not to have a fixed rule for what defined zaps and duds ahead of time—instead, it changed the rule as it went along. Each time you typed in a new pattern, the computer would ask itself whether it made your job in figuring out the rule harder if it replied “zap” or “dud,” and it always answered in the way that made your job harder.

The instructor had a theory that people who were told how the computer was working against them would play better, and he wanted us to test his theory in our experiment. So we made up a design that was somewhat more complicated than the designs we’ve been discussing so far, as follows. We got 30 friends in the freshmen class who knew nothing about the game and divided them “haphazardly” into two groups of 15. Everybody was taught the rules of the game, except the part about how the computer was working against them, and then they played the game ten times—five times, then a break in
the middle, then five more times. During the break the “treatment” people were told about the computer’s trickery, and the “control” people weren’t told anything. This is called a pre-post or before-after design—the advantage is that you can use each person’s performance on the first five trials as a kind of standard against which to measure any improvement in the second five trials.

Well, it worked great, or so we thought: the treatment people improved from the first five trials to the second five a lot more than the control people did. However, we also measured a number of other variables on all 30 people, and we found on closer examination of the data that, as a result of our “haphazard” assignment to treatment and control, the treatment people had much better scores on things like the SAT math test that were likely to be quite good predictors of how well people would do on the experimental task. Consciously or unconsciously (I still feel to this day that we didn’t try to do this on purpose), we had stacked the deck in favor of the treatment with our “haphazard” assignment to the experimental groups. So the whole thing was pretty much for nothing, although if we had been better statisticians we would have known that there are ways to at least approximately adjust for baseline differences like the ones we had, in an attempt at analysis time to salvage a poorly executed design. [I hope to cover such methods in SDA/2.]

Investigator. Okay, I’m convinced that haphazard assignment may not work too well, and incidentally I’m very glad to see that even a wise guy like you makes mistakes sometimes. But what should I do instead?

Devil’s Advocate. I’ll tell you the simplest thing people usually do to promote comparability of the treatment and control groups [later on in Chapter xx we’ll cover improvements to this basic idea], but it’s going to sound strange to you at first. The world is a pretty unpredictable place, would you agree with me?

Investigator. What is this, philosophy? Sure, it’s unpredictable—there’s lots of things I’d like to be able to forecast for sure, like who’s going to win the third race at Belmont tomorrow, but nobody can say for certain what’s going to happen.

Devil’s Advocate. Well, it sounds like a crazy idea, but what people usually do to try to make the treatment and control groups as similar as possible at the start of the experiment is to make the world an even more unpredictable
place than it was before the experiment was performed—to introduce some *planned* unpredictability—*randomness*—into the design.

*Investigator.* What do you mean, leave something as important as assignment to treatment and control to the toss of a coin?

*Devil's Advocate.* That's exactly what I mean, although in practice these days people don't usually use coins, they use a computer. The standard thing to do is to assign people to treatment and control at *random*—that is, in such a way that nobody can say for sure which group a given patient will fall in, and every patient has a 50/50 chance of being in each group.

*Investigator.* Of all your ideas so far, this one is the weirdest-sounding. How can it help to create comparability by tossing a coin? I have a better idea. I'm going to think up all the variables that are predictive of getting AIDS, and I'm going to personally assign each one of those 831 patients in such a way that things are exactly balanced between treatment and control on all of those variables. Then nobody can accuse me of stacking the deck.

*Devil's Advocate.* That's called *purposive assignment*, and if you had a lot of variables it would be harder than you think. There's two main knocks on your idea—first, it still lets your suspicious critics claim that you had a chance to influence the results, and second, what about any other relevant variables you forgot to include? Random assignment is actually one way to approximately produce the balance you want—since the randomization is impartial, it tends to produce comparability on all the variables you thought of, and any other variables too. I do like your idea of identifying important variables that are predictive of the outcome of interest—such variables are called *potential confounding factors*, or PCFs for short. If you wanted to, you could improve on pure randomization by taking explicit account of the PCFs in the design [*we will talk later on about how to do this*], but from the look on your face this doesn't seem like the right moment to go into that.

*Investigator.* This randomization idea is so strange that I listened pretty carefully to your little speech in favor of it, and there were a few words in it like "tends to" produce comparability and "promotes" balance that made me nervous. Are you trying to imply that random assignment doesn't always work? What are you trying to hide here?
Devil's Advocate. Well, yes, ahem, I guess this is as good a time as any to level with you. With random assignment, every time you do it you are very likely to get a different result than the last time you did it—that's the nature of the randomness. All I can guarantee about what you'll get on the one occasion you actually do randomize people to treatment and control is two things. First, on average, if you did it a lot, everything would balance out perfectly. And second, as the number of people in your experiment gets larger and larger it becomes less and less likely that you will get a "bad" assignment of patients to treatment and control, where by bad I mean that the people in the two groups differ a lot on some important variable.

Investigator. If I understand what you're saying, your first guarantee about getting things right on average isn't much consolation to me given that I'm only going to do it once. And your second guarantee doesn't sound like it would have helped me much if I only had 8 people in my experiment instead of 831.

Devil's Advocate. Now you're the one asking embarrassing questions.

Investigator. I don't hear any reassurance from you, but if this randomization idea is the best thing people have come up with so far I guess I'll join the crowd. So here's my final design.

**Design 3.** I get 831 HIV-positive patients who agree to participate, from medical centers in Scandinavia willing to participate, I divide them approximately 50/50 at random into a treatment group of 412 patients that get isoprinosine and a control group of 421 patients that get a placebo, I blind both the patients and the investigators to everybody's treatment status, and I look to see how many of them have AIDS after 24 weeks.

Now what would you say if 17 control patients got AIDS but only 2 treatment people did?

Devil's Advocate. I would say that you have conducted an example of the Cadillac of medical experiments, a randomized, double-blind, placebo-controlled clinical trial, as people say, and that your 24-week AIDS incidence in the control group was 17/421 = 4.0% versus 2/412 = 0.49% in the treatment group. This means that what people call the relative risk of getting AIDS for
control patients versus the treatment people was 4.0%/0.49% = 8.3—in other words, the control people were 8.3 times more likely to get AIDS during your 24-week followup period than the treatment people were. That sounds like a big relative risk to me, but you’ll have to do some more statistical mumbo-jumbo [we’ll talk about this later, in the chapter on “statistical significance”] to really convince the skeptics.

Some general themes on good design. Now that those two are done arguing, I want to step back and see what general principles on experimental design we can extract from their discussion. [Themes: fighting *bias*, PCFs, T = C except for the treatment, the two-numbers-on-the-forehead model, central counterfactual: the outcome the treatment people would have given you if they had been control people, and vice versa. Mention some criticisms of the Pedersen study.]

2. Observational Studies

[Another case study, maybe from health policy, e.g. the JAMA article on safety helmets and mortality in bicycle riders. Gentle readers: Can you recommend a good non-medical observational study for this section?]
6. An Introduction to Sample Surveys

[This chapter is still very rough, even more so than the first five.]

The point of this chapter is to learn a bit about how people design sample surveys. This is one of the bread-and-butter topics in applied statistics—when you ask knowledgeable people to identify the most important tools in the statistician’s bag of tricks, the phrase “you have to know how to sample” comes up over and over. The topic is closely related to experimental design—in fact, it could almost be regarded as a special case of the ideas we talked about in the last chapter, although there are enough new wrinkles that it deserves a chapter of its own—and I will try to draw out some of the connections as we go along. As a secondary goal of this chapter (which we’ll actually do first, and get to the sampling stuff later), I want to cover some key choices in the design of a big observational study, as a way to reinforce the ideas in section 2 of Chapter 5.

1. The DRG Quality of Care Study

This chapter’s case study is one that I spent four years working on at RAND, as part of a team of about 15 people: the DRG Quality of Care Study. The problem we tackled is a textbook example of health policy, the part of medicine devoted to finding cost-effective ways to deliver high-quality medical care to those who need it. Background for our study goes like this.

The policy background. In the 1930s, under the leadership of Franklin Roosevelt at the height of the Depression, the U.S. government enacted a number of social reforms, offering Americans various components of what would later come to be called the “social safety net,” including Social Security (guaranteed retirement funding) and unemployment insurance. [get historical details right] However, one key part of the social safety net was not created at that time: health insurance against crippling medical expenses, particularly toward the end of people’s lives when they were no longer working. Older Americans would have to wait more than 30 years for this basic need to be met.

In 196x, as part of Lyndon Johnson’s “Great Society” social reforms, Congress and the Executive Branch of the U.S. government set up the Medicare program to provide guaranteed health coverage for the nation’s elderly and disabled people. The Medicaid program, offering similar help to the nation’s poor people of all ages, followed shortly thereafter in 196x. One big problem with programs like this always turns out to be how to pay for them, in a way that
doesn't bankrupt the government or create weird incentives for people to behave perversely in an attempt to profit from the system. The method of funding Medicare hospitalizations they tried at the beginning—and, indeed, stuck with for almost 20 years—was called fee-for-service or retrospective payment, and here is how it worked. Let's say you're an old person and you fell down and broke your hip. Under retrospective payment you went to the hospital, the hospital fixed you and sent the government a bill, and—apart from the occasional denial of some particularly outrageous charges—the government paid the bill. This was called retrospective payment because nobody knew how much the care was going to cost until after it had been given.

It turns out that this system worked pretty well from a quality-of-care perspective, but not so well from the point of view of costs, which spiraled upward at a rate quite a bit higher than the background rate of inflation for most of the two decades of retrospective payment. It's not hard to see why, when you think about the incentives the hospitals had with this payment system, which were to overtreat the patient to make more money. After all, if you kept the patient an extra day or ran an extra test on him or her, it was a pass-along: the government would usually just pay the bill. By the early 1980s the funding problem had become acute enough that it threatened bankruptcy of the Medicare Trust Fund, the pot of money set up to pay for the system, and it became clear that a new payment mechanism was needed that would slow the upward cost spiral. The part of the government that runs Medicare, the Health Care Financing Administration (HCFA for short), was mandated in 198x to come up with a new system.

HCFA was not exactly caught unawares by this mandate, because the problem had been brewing for some time, but when the call to action came they did not have dozens of alternatives sitting on their shelves waiting for the best one to be selected and thrown into service. In fact, they only had one, and the evidence on the wisdom of turning it loose on the whole country was quite scanty: a single modest experiment run in New Jersey from 197x to 198x. Nevertheless, the government seized on this one alternative and went full steam ahead with it in fiscal 1983, with the passage of a noble-sounding piece of legislation called the Tax Equities and Fiscal Responsibility Act (TEFRA), which placed a fixed cap on how much money Medicare would pay out that year. (I say "fiscal year" because the government runs on a different calendar than the rest of us: fiscal 1983 actually began on October 1, 1982 and ran until September 30, 1983.) TEFRA was just a one-year transition to the real thing, however: the new payment method, called the Prospective Payment System (PPS), was phased in beginning in fiscal 1984.

Under the new system, the universe of all medical ailments serious enough
to put you in the hospital was divided up into roughly 475 categories called Diagnosis-Related Groups (DRGs). (Look at all the acronyms pile up—the government can’t function without acronyms.) These DRGs act like rows in a little black book with two columns, diseases and dollars, as follows. Say you’re that same old person with the broken hip as before, only now you come to the hospital under the new payment system, PPS. The hospital administrators metaphorically meet you at the door with the little black book, they look up the hip fracture DRG and look across into the payment column and they see $4012 [get the right figure], and they know that’s exactly the amount of money they’re going to get for fixing your broken hip, no more, no less. This is called prospective payment because everybody knows before the care is given how much it’s going to cost.

It’s actually a bit more complicated than I’ve just said, in three main ways. First, HCFA phased the new system in over its first four years, from 1984 to 1988, by paying a weighted average of the old dollar amount the hospital would have received under retrospective payment and the new PPS amount, with the weights gradually shifting toward 100% PPS. This was to ease the shock of introducing such a radical change in such a fundamental aspect of the Medicare system, and it was a good thing for HCFA to do—it is a fact about the world that incremental change is easier for people to cope with than drastic change. Second, most DRGs come in pairs, for less sick and more sick patients with the same ailment, and they pay more in the “more-sick” DRG in each pair (because on average the sicker you are, the more it should cost to fix you). Third, if a patient stays in the hospital a really long time, he or she becomes an “outlier,” and the hospital gets paid more than the usual flat-rate DRG payment for such patients. This is to keep hospitals from being unduly financially punished for treating especially sick patients. Leaving aside wrinkles like this, however, I hope you get the basic flavor of PPS, which is fixed payment no matter what the hospital does in treating its patients. If that sounds slightly crazy to you, you are in good company—when PPS started, a lot of people thought it was a terrible idea.

From the definition of health policy I gave awhile ago, there are evidently two things you should ask about the new system’s performance—how it’s doing on cost, and how it’s doing on quality of care. As of 1990 PPS was a modest success in containing health care costs—it had not totally halted the upward spiral, but it had noticeably slowed down the rate of increase in Medicare costs from year to year. However, on the quality of care front, let us reflect for a moment on the incentives supplied by the new payment system, which are evidently to undertreat the patient to make more money. If you keep somebody in the hospital an extra day under PPS or run an extra lab test, you (the hospital) have to eat the cost
of these things—the government isn’t going to pay you extra for them anymore. This doesn’t sound so good for quality of care—it sounds like people might try to cut corners to make a buck.

HCFA economists and others of that ilk had a different theory—they hoped instead that the new system would increase the efficiency of the health care market, as they say, by forcing hospitals to pare back their care to the barest essentials, leaving quality of care alone (or even improving it, if non-essential things done to patients worsened their care on average). The point of our study was to see who was right, the economists or the nay-sayers: Did PPS cause a decline in the average quality of care given to hospitalized Medicare patients? And what about the typical care given to important subgroups of these patients, like the very old (over 85, say), the very poor, or the very sick?

To answer these questions, we assembled a multi-disciplinary team at RAND, consisting of five physicians; two statisticians, a mathematician, and an operations research specialist; three experts in statistical computing and data base management; and two project managers who made sure the data collection went smoothly. The study we put together took about 41/2 years to finish and cost about $5.6 million of HCFA’s money, when all the design, data collection, and analysis expenses were combined. ($2.6 million each for the answers to two questions probably sounds a bit steep, but we actually answered a lot more than two questions.) I was responsible for the sampling plan and a number of the other design details, and I helped with a good bit of the analysis and the writing up of our findings.

This case study is the first large-scale example of decision-making (as contrasted with science) in the book—it turned out that we had to do some science to figure out how to measure quality of care well (as you might imagine, it’s a nebulous thing to quantify), but the real point was to help HCFA decide how, if at all, PPS should be changed if there were any quality problems. I’m almost ready to start looking at the sampling issues that are the real point of the chapter, but first (to put the sampling ideas in context) I need to go through some of the logic behind the design of our study, so bear with me for a few more paragraphs until we get to the sampling stuff.

The Design of the DRG Study. Let’s think about this PPS thing from the viewpoint of experimental design. The question is whether PPS caused a change in the quality of care, so you’ve got to compare the outcome of interest (quality of care) in two groups: people whose care was paid for under PPS (the treatment group), and people whose care was reimbursed in some other way (the control group). We saw last chapter that your best chance to figure out the effects of a given cause is typically to do a randomized controlled experiment, which makes
me want to assign people to PPS or the control group at random. But HCFA put
the whole country on PPS in 1984, and they were not about to take some of it off
PPS just to make our experimental design task easier—a controlled experiment
wasn’t feasible. So we had to settle for an observational study.

I skipped an important point in the last paragraph—what reimbursement
method should be used in the control group? Since HCFA would still be paying
for patient care under the old payment system if they hadn’t put PPS in place
in 1984, the leading candidate for the control group should be Medicare patients
whose care was paid for under the old retrospective system. This makes sense
from the perspective of the two-numbers-on-the-forehead model we talked about
last chapter: the key counterfactual for the patients treated under PPS is *what
their quality of care would have been if they had been treated under retrospective
payment*, and vice versa. To put it another way, our job in conducting the study
was to help HCFA estimate both a factual and a counterfactual: what quality of
care was like from 1984 on, under PPS (factual), and *what it would have been
like during the same period if PPS had not been implemented* (counterfactual).
The effect of PPS on quality of care is then the difference between these two
quantities. So: the treatment variable in our study was a dichotomy—PPS versus
retrospective payment—and the response or outcome variable was quality of care
(Figure 1). Note that all that placebo and double-blind stuff we talked about last
chapter is irrelevant here (why?).

So where should we get our patients for the treatment and control groups?
Simplest would be to compare quality of care for Medicare patients before 1984
(the control group, on retrospective payment) and after 1984 (the treatment group,
on PPS) —pre- and post-PPS, you might say. However, as I hope came through
clearly last chapter, any time you’re contemplating an observational study you
should be on guard against potential confounding factors, and there’s a big one
in this simple pre-post design: the passage of time. Let’s say you got all your
control patients from 1982 and all your treatment patients from 1986—then the
effects of (PPS versus retrospective payment) would be *completely confounded*
with the effects of (1982 versus 1986). What I mean by this is that any differences
you saw in quality of care might have been caused by PPS, but they could just
as easily have been caused by any other major interventions in the health care
system between 1982 and 1986, or by what people call *secular trends* in medical
care—gradual changes over time in how doctors practice medicine. We weren’t
too worried about other major interventions clouding the picture—basically there
weren’t any—but the trend problem was real.

As we saw last chapter, the best way to beat this problem of the passage of
time confounding things would have been to hold the time factor constant—to compare PPS Medicare patients after 1984 (in 1986, say) with retrospective payment Medicare patients at the same moment in time. But this brings us back again to the reason we couldn’t do a controlled experiment—since HCFA put the whole country on PPS at once, there weren’t any retrospective payment Medicare patients in 1986, so that idea didn’t work. We considered several other possible sources of control information, but none of them were very attractive, and in the end we dealt with the trend problem fairly directly: we chose a pre-post design that measured quality of care at multiple time points in both the pre-PPS and post-PPS periods, so that we could at least roughly assess any trends that were in place before PPS was implemented and adjust for them. We chose as our pre-PPS time window the entire calendar years 1981 and 1982, and as our post-PPS period the last half of 1985 and the first half of 1986.

After having made all these choices, the data-gathering problem is finally beginning to take shape: measure the quality of care received by hospitalized Medicare patients in 1981-82 and in 1985-86. But which Medicare patients? All of them? That would have cost too much—in the end, the data-collection part of our $5.6 million budget was about $2.5 million, and that only bought us data on about 17,000 of the [xx,xxx,xxx] patients treated under Medicare in 1981-82-85-86.

For one thing, we couldn’t possibly have done the study on all Medicare patients—it turns out that measuring quality of care is a very disease-specific thing. Our approach was to ask physician experts to help us identify appropriate and inappropriate medical care in considerable detail, and we estimated that it took us two months and cost about $75,000 to put together the abstraction form for each disease—the booklet that told our data-gathering people which pieces of information to copy out of each of the 17,000 medical records we examined in the study (Figure 2). In the end we were forced to limit the study to six diseases that spanned a fair amount (but by no means all) of the spectrum of care for old people: pneumonia, heart attack, stroke, hip fracture, depression, and something called congestive heart failure (a disease in which your heart-lung system gradually stops working). That narrows the data collection problem to just the people with one or more of these six diseases, although it still sounds daunting—there were still [x,xxx,xxx] Medicare patients with one of our diseases in the four years in question.

But there’s an even more basic reason why you don’t want to get data on all Medicare patients—it’s extremely wasteful. Back in Chapter 1 I mentioned that the people who run the Gallup poll can say some interesting things about the whole
country—2xx,xxx,xxx citizens, as of 1990—based only on an intelligently-chosen sample of about 1500 of those citizens. The same idea applies here, although we needed a larger sample than that to meet the accuracy goals in the DRG study: about 2,800 patients in each of the six diseases, for a total of about 17,000 patients. If you spot me the 17,000 figure for now (we’ll see where it came from in Chapter 16), we’re finally ready to begin answering the main question in this chapter: How can you possibly choose 17,000 people from a collection of [x,xxx,xxx] patients in a way that’s almost as good as having data on all [x,xxx,xxx]?

2. Exchangeability and Simple Random Sampling

First, a little terminology. Typically, in a situation where sampling makes sense, you have one or more outcomes in mind, and you’re interested in summarizing those outcomes for a (usually large) collection of people or things on the basis of data gathered on a subset of this collection. People call the overall collection of people or things of interest the population, and the subset the sample. Here the population is all [x,xxx,xxx] Medicare patients with our diseases in 1981-82-83-86, and the sample is the 17,000 people we got data on. Back in Chapter 1 we called the “people or things” you’re examining subjects, but samplers often refer to them as units, and (although the term is rather impersonal) I will occasionally do so too.

I like to visualize the population as a big dataset, with one row for each subject or unit and one or more columns for the outcome(s) of interest and any predictors of the outcome(s) you happen to have collected. The sample is then a smaller dataset with the same columns as the population but with only the rows that were chosen in the sampling process. Figure 3 illustrates this for an extremely simple version of the DRG data in which all we focus on is the outcome, quality of care (on the 0-100 scale I used in Figure 1), and a single predictor: whether the patient was treated in the pre-PPS or post-PPS period.

Now as I said back in Chapter 1, when most people are first presented with the idea of trying to get a sample of the population you’re interested in to stand in for the whole population, the idea strikes them as somewhat strange: how can this possibly work? But if you suspend disbelief for a moment and begin thinking about how you might actually try to do it, it’s clear that there are only two things to specify: how many units should you get in the sample, and how should you choose them? We have agreed to postpone the first question to the chapter on sample size determination, so I will move on briskly to the second question. The phrase “stand-in” provides a clue—in the DRG study, somehow we want to pick
the 17,000 patients so that they stand in for, or represent, all [x,xxx,xxx] patients as well as any collection of 17,000 patients can do so. This in fact is the key goal in the simplest forms of sampling—people say that you want the sample to be representative of the population. But what exactly does this mean?

Well, consider the outcome you’re interested in, in this case quality of care. Certainly you’d want the sampled subjects to be similar to the whole population with respect to this variable—if not, you would draw misleading conclusions about the population from the sample. By “similar” I mean things like the outcome column in Figure 3 in the sample dataset being descriptively similar to the outcome column in the population dataset—things like the mean, the SD, and the histogram of the two columns should be similar. Actually, you’d want more than this—you’d want the sample and population to be similar not only on the outcome of interest but also on all important predictors of that outcome, because if you got the predictors wrong in the sample you’d expect the outcome to come out wrong too. Two examples of predictors of quality of care we found in the DRG study are the age of the patient and the urbanicity of the hospital in which the patient was treated—on average, the older you are the worse quality of care you get in American hospitals, even accounting for how sick you are (a kind of age discrimination), and quality of care is also noticeably better in urban hospitals in the U.S. than in rural facilities (probably due to the slow dissemination, from research centers and so-called teaching hospitals where doctors are trained, into outlying areas of information about the best ways to treat patients). So to achieve representativeness of our sample in the DRG study we would want to make sure that the age and urban/rural distributions of our sampled patients and hospitals matched the nation well.

The moral of the last few paragraphs is that in any given problem you can make the term “representative” come to life by specifying the outcome you’re interested in and the important predictors of that outcome, and trying to choose the sample so that the sampled subjects are “just like” the whole population with respect to those variables.

There is actually a little cleaner way to say this that reveals a key relationship between sampling and experimental design. When you say that you want the sample to be “just like” the population in a certain way, all you really mean is that you want the sampled and unsampled subjects to be similar to each other in that way, because the sample is a subset of the population and is similar to itself (in fact, identical to itself) by definition. (This is worth thinking about for a minute. What I’m saying, for example, is that if you’re trying to judge if the two sets of objects $S_1 = \{A_1, A_2\}$ and $S_2 = \{A_1, A_2, B_1, B_2, B_3\}$ are similar, you
really ought to just concentrate on $S_1 = \{A1, A2\}$ and $S_2^* = \{B1, B2, B3\}$—if $S_1$ and $S_2^*$ are similar then so are $S_1$ and $S_2$, and vice versa.) Another way to put it is that representativeness is equivalent to the sampled and unsampled units being *interchangeable* with each other as far as the outcome and its predictors are concerned. As we saw back in Chapter 5, the technical term used to describe this sort of interchangeability is *exchangeability*: the sampled and unsampled units are exchangeable (for sampling purposes) if they are “just like” each other with respect to the outcome and its main predictors. To summarize all of this:

When you say that, in your judgment, a sample is representative, what you mean is that, in your judgment, the sampled and unsampled units are exchangeable with respect to the outcome variable of interest and all variables that are strongly predictive of that outcome.

I have used phrases like “in your judgment” here to emphasize that *things like exchangeability and representativeness are not properties of the world, they are judgments about the world*. There’s a big element of subjectivity in defining “similar to” and in specifying “all variables that are strongly predictive of the outcome of interest”—every time you change your mind about these things, you are entitled to change your mind about the representativeness of somebody’s sample, and two reasonable people can disagree about how to specify such things. Saying this doesn’t make the results of sample surveys any less valid, it just underscores the role of judgment in designing and critiquing them.

So, okay, the goal is exchangeability of the sampled and unsampled units. How do you accomplish this? The key is to notice that this appeal to exchangeability makes sample surveys a lot like controlled experiments. Remember in the last chapter when we had a bunch of people (we could have called them *experimental units* back then if we had wanted to use an impersonal term for them), and we wanted to assign them to treatment and control in such a way that they were “just like” each other except for the treatment? This is equivalent to saying that we want to assign people to treatment and control so that the two groups are exchangeable with respect to the outcome of interest (getting AIDS, back in Chapter 5) and all good predictors of that outcome. This sounds so similar to what we’re trying to do with sampling that I can feel an analogy coming on—if we could only set up a correspondence between sample surveys and controlled experiments, we could use the solution we already worked out last chapter. If you enjoy figuring things like this out, you should put the book down now and think about it, because I’m going to tell you the right answer in the next paragraph.
[To be continued: Answer—what if the “treatment” in sample surveys is getting sampled and the control group is the units that aren’t sampled? Sources of bias. Work in cross-sectional versus longitudinal?]

3. Cluster Sampling and Stratification

4. Summary

5. Problems
Figures for Chapter 6

Treatment: PPS (treatment group) vs. retrospective payment (control group)

Response: quality of care (measured, let's say (for the moment) on a scale from 0 to 100, with high values meaning good quality of care)

<table>
<thead>
<tr>
<th>Patient Number</th>
<th>Payment System</th>
<th>Quality of Care if Treated Under PPS</th>
<th>Quality of Care if Treated Under Retrospective Payment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PPS</td>
<td>60</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>PPS</td>
<td>35</td>
<td>?</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>8500 (say)</td>
<td>PPS</td>
<td>82</td>
<td>?</td>
</tr>
<tr>
<td>8501</td>
<td>Retrospective</td>
<td>?</td>
<td>74</td>
</tr>
<tr>
<td>8502</td>
<td>Retrospective</td>
<td>?</td>
<td>29</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>17000 (say)</td>
<td>Retrospective</td>
<td>?</td>
<td>51</td>
</tr>
</tbody>
</table>

Figure 1: Treatment and response variables for the DRG study, and the two-numbers-on-the-forehead (causal inference) model in this case.

<table>
<thead>
<tr>
<th>population</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of Care</td>
<td>Quality of Care</td>
</tr>
<tr>
<td>Pre- or Post-PPS</td>
<td>Pre- or Post-PPS</td>
</tr>
<tr>
<td>94 Pre</td>
<td>32 Pre</td>
</tr>
<tr>
<td>68 Pre</td>
<td>55 Pre</td>
</tr>
<tr>
<td>32 Pre</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>55 Pre</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>71 Pre</td>
</tr>
<tr>
<td>71 Pre</td>
<td>.</td>
</tr>
<tr>
<td>80 Pre</td>
<td>40 Post</td>
</tr>
<tr>
<td>26 Post</td>
<td>.</td>
</tr>
<tr>
<td>59 Post</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>88 Post</td>
</tr>
<tr>
<td>40 Post</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>19 Post</td>
</tr>
<tr>
<td>88 Post</td>
<td>.</td>
</tr>
</tbody>
</table>

Figure 3: A simplified version of the population and sample in the DRG study.
30. Other Medical Conditions

Prior to the admission did the patient ever have:

Yes  No/ND

i) Chronic obstructive pulmonary disease (COPD), emphysema, chronic bronchitis or asthma .......... 1  9
ii) Current cigarette smoker .......................... 1  9
iii) Cirrhosis, portal hypertension, or ascites ...... 1  9
iv) Current alcoholism ................................. 1  9
v) Diabetes ............................................. 1  9
vi) Chronic renal failure ............................... 1  9
vii) Systemic lupus erythematosus (SLE) .......... 1  9
viii) Temporal arteritis ................................. 1  9
ix) Deep vein thrombosis (DVT) ...................... 1  9
x) Carotid endarterectomy ............................ 1  9
xi) Bypass surgery of leg arteries (femoral, iliac, popliteal) or amputation of leg, foot, or toes ............ 1  9
xii) Prior hospitalization within 6 months .......... 1  9
xiii) Episodes of psychosis or depression .......... 1  9

Check here if none of the above apply or no data ............ □

30a) If any questionable responses to Q30, copy exact remarks.

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

Figure 2: An excerpt from the abstraction form for congestive heart failure, one of the six diseases chosen in the DRG study.
Part IV. Probability
7. Basic Probability, from the Frequentist Point of View

This section of the book is about probability, the part of mathematics devoted to quantifying uncertainty. It's a topic that everybody knows at least a little bit about: for example, there's the person on your local television news program who is always saying things like, "The chance of rain tomorrow is 40%." Not too surprisingly, I guess, there's a fair amount more to the subject than the use we make of it in ordinary conversation, including a few surprises and some genuinely hard philosophical questions. It turns out that there is a controversy at the heart of statistics that comes down to philosophical issues in probability—for instance, what exactly does the weather person mean by that 40% figure above? The two sides in this controversy are referred to as frequentists, a name derived from the word "frequency," and Bayesians, named for a man called Thomas Bayes who [proved an important theorem] about 250 years ago. These two names correspond to the two main ways to define precisely what the weather person means: the frequency and Bayesian theories of probability. In this chapter I will lay out some basic groundwork in probability from the frequentist point of view; we will look at the Bayesian story in Chapter 8.

Before starting in on the frequentist approach, I want to make a distinction between probability and statistics that helps to explain the order of topics in the book. The distinction I have in mind is an old one—it relates to the difference between inductive and deductive reasoning, a concept that goes back at least to the ancient Greeks. If you look up the words "induction" and "deduction" in the dictionary, it will tell you that deduction is the process of reasoning from the whole to the part, induction the opposite process, from the part to the whole. (I will say what "whole" and "part" mean in this context in a minute.) It turns out that most of the uses of probability we will make in this book are examples of deductive reasoning, whereas the inferential and predictive applications we will study later in the book are all inductive.

This will all become clear (I hope) if you think back to the main problem we looked at in Chapter 6—sample surveys. In that setup (1) you have some population of interest, say the people who will vote in the next presidential election; (2) it's not practical to ask all of them who they'll vote for, so you take a sample, using randomization (chance) in some way to promote representativeness of the sample for the population; and (3) you use the sample to estimate, say, the fraction of people in the population favoring the Democratic candidate. In the
terminology of Chapter 1 this is a problem of inference or prediction rather than description—we want to infer from the sample to the population, or equivalently (as we saw in Chapter 6) if you want to you can think of this as trying to predict how the unsampled people would have responded if you had interviewed them in addition to the sampled people. Figure 1 makes the key distinction between probability and statistics in this setup.

In this case the “whole” the Greeks had in mind is the population, and the “part” is the sample. As we will see in Chapters [xx] and [xx], if you know the population fully and you choose a sampling method like simple random sampling that incorporates chance in a controlled way, you can use probability ideas to accurately predict what kind of sample you will get, and how closely it will approximate the population—that’s deduction. If all you have is a sample, and you are trying to infer backwards from it to the entire population, that’s induction, and it’s statistics, not probability. It’s easier to go from the whole to the part than the other way around, so we will spend a few chapters on probability before going on to the heart of the statistical work in the book.

1. Probability as Relative Frequency

So what exactly does the weather person mean by saying that “The probability of rain is 40%”? There are many possible silly answers to this question—for example, out of every 10 people walking down the street, 4 will get wet and the other 6 will stay dry. If I gave you access to lots of weather information over the past five years, say, in your town, and asked you to do the best job you could of telling me whether I should bring my umbrella tomorrow, eventually you would probably be drawn to a strategy something like the following, which is a lot like what the National Weather Service actually does:

1. You would figure out which variables were relevant to the prediction of rain,
2. You would find all the days in the past on which you had data that were “just like” what you thought tomorrow would be like with respect to these variables, and
3. You would see on what fraction of such days it rained in the past.

If there were 261 such days, say, and it rained on 104 of them, you would announce the chance of rain as 104/261 = 40%. (To put it more technically, in the language of Chapters 5 and 6 “just like” is shorthand for “exchangeable with.”) Note that there is an underlying assumption of what might be called persistence of structure in this strategy—an implicit belief that the future will be enough like the past, at least as far as the weather is concerned, that good guidance to
future rainfall behavior may be obtained by examining past data.) Since the way your chance, or probability, of rain was arrived at was by working out a relative frequency (104 out of 261), this is an example of the frequency or frequentist approach to defining probability.

Let me give another example, somewhat different but simpler. I pull a coin out of my pocket and flip it, letting it fall to the floor. What is the chance that it will come up heads? To use the same approach to defining this probability as in the weather example, we would have to observe a series of such coin flips, preferably under "identical" conditions (or as close to identical as we could make them, with respect to anything that might determine which side landed up), and we would just see what fraction of the time heads actually appeared. But maybe there's a shortcut, using common sense. If the coin is what people call fair—that is, if it's not two-headed, and if I have no particular skill in making it come up heads—it makes sense to notice that there are two possible outcomes of the experiment of tossing the coin once (heads and tails) and that there is no particular reason for one of them to be likelier than the other, so the chance of heads should be 1 out of 2 or 50%. Evidently this is also a kind of relative frequency approach to coming up with a probability: you figure out how many ways the experiment you are thinking of performing could come out (here two: heads and tails), you count how many of them are favorable to the event of interest to you (in this case one: heads), and—as long as common sense suggests that all the possibilities should be equally likely—you define the probability of the event of interest as the ratio or relative frequency

\[
\frac{\text{number of ways favorable to the event}}{\text{total number of ways the experiment could come out}}.
\]

This is called the equally-likely model for defining probabilities, and we will be making a lot of use of it.

Mathematically, the framework I've just sketched goes something like this. You have an experiment you're thinking of performing, like tossing a coin or waiting until tomorrow to see if it will rain. The collection of all possible ways the experiment could come out is a set, called the sample space and usually denoted by the Greek letter \(\Omega\). (In the coin-tossing example \(\Omega = \{\text{heads, tails}\};\) in the weather example \(\Omega = \{\text{rain tomorrow, no rain tomorrow}\}.\)) There's some particular event—that is, some particular way the experiment could turn out—whose probability interests you. Events are usually denoted by capital letters, often chosen to remind you of the outcome they stand for; mathematically events are subsets of the sample space \(\Omega\). (In the coin-tossing example the event of
interest was \( H = \{ \text{heads} \} \); in the weather example it was \( R = \{ \text{rain tomorrow} \} \). 

The **probability** of an event is a number, a relative frequency, arrived at in one of two ways:

1. If you have a lot of relevant data, the ratio 
   \[
   \frac{\text{number of times event occurred in your data}}{\text{number of repetitions of the experiment}}, \text{ or}
   \]

2. If you don’t have any data, the number of times it *ought* to occur if you were to repeat the experiment a lot, based on common sense and equally-likely reasoning like that in the coin-tossing example above.

Usually the probability of an event \( A \) is denoted \( P(A) \), so that symbolically a statement like “The chance of rain is 40%” comes out \( P(R) = 40\% \).

(Technical notes: Mathematically a probability is a function \( P(.) \) that when given an event \( A \) produces a number \( P(A) \). The first definition above, in which a frequentist probability is based on data, can’t be quite right technically, since (for instance) in the weather example each time you get another day’s worth of data the probability of rain changes a bit. To come up with a rigorous definition along these lines, people focus on the *limit* of the ratio mentioned in (1) above as the number of repetitions goes to infinity, but for our purposes it is enough just to think about what you would get if you had a lot of data.)

The point of all this is that to assign a probability to an event using the frequentist approach, you either have to have data on how many times the event occurred in many past repetitions of the process you’re studying—together with the conviction that the past and future are sufficiently similar that past experience is a good guide to future outcomes—or you have to be able to imagine repeating the process a lot and use common sense to say what you *ought to* get for the relative frequency of interest.

---

Repetition under “identical” conditions, either real or imagined, is at the heart of the frequentist approach to defining probability.

---

Okay, enough philosophy for a while. How about some history? I bring this up now (a) because it’s interesting and (b) to explain a pattern you will notice in the examples and problems in this chapter: a lot of them are about gambling. There’s a reason for this that goes all the way back to the 1600s: the ideas behind frequentist probability were developed to settle gambling disputes.
The noblemen in the court of King [so-and-so] in France in the middle of the 17th century didn't have much to do that interested them more than gambling, so they gambled a lot, and (since they were by and large pretty rich) the stakes were rather high. They invented all manner of elaborate ways to wager on the fall of cards, dice, and coins, and a lot of interest focused on how much money should change hands on a given outcome so that the gamble was fair—that is, so that on average, if the game were repeated a lot, neither player would come out ahead. For instance, if you and I agree to gamble on a single toss of a fair coin, and I say that I will give you $10 if the coin comes up tails, then it's pretty obvious that you should give me $10 if it comes up heads to make the gamble fair, because (as we have agreed above) heads and tails are 50/50 outcomes. But what if I want you to give me $10 if I get a six, say, on a single roll of a fair six-sided die—how much should I have to give you if I don't get a six to make things fair? We will work out details like this in Chapter [xx]; for now, you can probably see intuitively that everything hinges on how likely the two outcomes {six} and {not six} are, and that therefore the French noblemen routinely needed a way to figure out probabilities of various gambling-related outcomes, some of them rather complicated.

Even the King got interested in this problem, and eventually commissioned a well-known French mathematician of the day, a man named Blaise Pascal, to work on it. Pascal laid out the ideas behind frequentist probability, worked out all the stuff we will spend the rest of the chapter developing (and a lot of other stuff, too), and settled all the noblemen's disputes. One of the arguments was over whether or not two dice-rolling outcomes had the same probability—the noblemen suspected they did not. Pascal was able to show that one of the outcomes had a 50% chance and the other one about a 51% chance—it was remarkable that the noblemen had been able to discern any difference between them.

Anyway, to get back to the story here, I have chosen (along with just about every other writer of an introductory statistics book) to use a fair number of gambling examples in this chapter, not because they come up so often in real life but mainly for the same reason that led such examples to be the genesis of frequentist probability: the essence of the gambling setup is repetitions of the same thing again and again under close-to-identical conditions, so gambling problems are tailor-made for the frequentist story. Later in the book we will see something that people began to notice in the 1800s when they tried to apply Pascal's ideas to scientific and social problems: the frequentist setup does not always work quite as splendidly in real-world arenas, other than gambling, in which people encounter uncertainty and wonder how to behave sensibly in the face of it.
I want to spend most of the rest of the chapter telling you about three rules that are a consequence of the frequentist definition just given, because they’re so useful in actually computing the chances of various events of the types that come up in science and decision-making. In practice it is not always necessary to directly observe the relative frequency of a given event to figure out its frequentist probability; often, particularly if the event is complicated, you can get frequency data on simpler events and use the rules in the rest of the chapter to work out the probability you want from the probabilities of the simpler events. The first rule below is easy and comes in two parts; the second two rules are a bit more complicated.

First of all, if you repeat something \( n \) times, the smallest number of times the event of interest could occur is 0 and the largest is \( n \), so I guess the smallest and largest possibilities for a frequentist probability are \( \frac{0}{n} = 0\% \) or 0 and \( \frac{n}{n} = 100\% \) or 1, respectively. (I will use percentages and decimals interchangeably for probabilities.) In other words,

\[
\text{For any event } A, \ 0 \leq P(A) \leq 1. \tag{1}
\]

In set terminology, the only event with probability 0 is the null set, and the only event with probability 1 is the whole sample space \( \Omega \). It makes sense that \( P(\Omega) = 100\% \), since \( \Omega \) is supposed to be an exhaustive list of all possible outcomes, and something has to happen when the experiment you’re interested in is performed.

The second fact that’s pretty easy to establish concerns the relation between the probability of something happening and the probability of it not happening. If the event you’re interested in happened \( k \) times out of \( n \) repetitions, say, then it must have failed to happen the other \((n - k)\) times, so for any such event \( A \)

\[
P(A) + P(\text{not } A) = \frac{k}{n} + \frac{n-k}{n} = 1. \tag{1.1}
\]

Combining equations (1) and (1.1) gives

| The Easy Rule: For any event \( A \), |
| \( 0 \leq P(A) \leq 1 \), and \( P(\text{not } A) = 1 - P(A) \). \tag{2} |

This makes pretty good sense: anything you’re interested in has to either happen or not happen, so if something has a 70\% chance of happening, say, and
there's 100% worth of probability to go around it must have a 30% chance of not happening. It may not seem to you that this observation advances the ball much, so to speak, but in fact it's a quite useful trick to remember when trying to compute a probability: if you're having a hard time figuring out the chance of some event, try to work out the chance of it not happening, and subtract from 100%. I will give an example of this in section 5 below.

Before going on to the other two rules I mentioned, I want to finish off this section with two pictures related to the Easy Rule. The first picture (Figure 2) helps to explain why I said at the beginning of the chapter that probability is the part of math devoted to quantifying uncertainty.

It is a strong statement to say that the probability of something is 0, or that it's 1—these two extremes on the probability scale represent complete certainty about the occurrence or non-occurrence of the event A in question, whereas the probability values in between these extremes represent varying degrees of uncertainty about whether or not the event A will happen. How much uncertainty? We will talk more about this later on, in Chapter [xx], when we have sufficient machinery to answer the question well, but two things can be seen now just based on what we already know:

1. Since you can't have less than 0 uncertainty, and uncertainty about the occurrence of the event A in Figure 2 is zero at the edges, uncertainty must increase as we move in from the edges of the probability scale toward the middle.

2. Since knowing whether the event (not A) occurs is equivalent to knowing whether A occurs, the amount of uncertainty must be symmetric about the middle of the probability scale, 1/2—that is, if something has a 20% chance of happening, I have the same amount of uncertainty about it as if its chance of happening were 80% (think about it).

It's natural to conjecture, putting these two facts together, that your uncertainty about an event is largest when its probability is 1/2. We will see later on, in Chapter [xx], that this is true.

The second picture I have in mind (a) illustrates the Easy Rule and (b) has the added virtue of finally getting some use out of something you were probably taught in grade school that must have seemed completely useless at the time. Do you remember Venn diagrams? (If you ever had what educators for a while called the "new math" inflicted on you, the answer should be yes. If not, you're lucky, and I will try to make this exposure to one of its key subjects as painless as possible.) The idea is to draw a rectangle representing all the different ways
your experiment could come out, and one or more blobs inside the rectangle representing the various events you're interested in (Figure 3). In these pictures, probability is represented by area, with the total area of the rectangle set to 1 since it represents the whole sample space. Equation (2) above is now pretty obvious from Figure 3: the area of the blob representing the event A and the area of everything outside the blob have to add up to the total area of the rectangle.

2. The Addition Rule for Working with Or

If Figure 3 were the best use people could make of Venn diagrams in probability, it would not be worth mentioning them here, but in fact we can use them to pretty easily demonstrate another basic probability fact that's very useful in computing complicated probabilities: a result called the Addition Rule. To set the stage for this rule, let me give an example of a slightly complicated probability of the type we will study in more detail later, in Chapter xx [binomial]. In baseball, a batter who gets a hit 3 times out of 10 is considered quite a good hitter (if you do this you are said to be "battling .300"). In a game in which he or she comes to bat three times, what is the chance of a .300 hitter getting at least 2 hits?

The way you figure out probabilities like this one is to break down the event of interest into a bunch of smaller events that are equivalent to it and whose probabilities you can figure out directly. Since (as I said above) mathematically events are sets, the way people accomplish this is to re-express the complicated set of interest in terms of simpler sets, by means of what are referred to as the elementary operations of symbolic logic: and, or, and not. We have already met not in equation (2) above; the addition rule involves or and and. (In set theory, and, or, and not correspond respectively to the set operations of intersection, union, and complementation.) In the baseball example, for instance, getting at least 2 hits in 3 tries means getting either 2 or 3 hits, so

\[ P(\text{at least 2 hits in 3 tries}) = P(\text{2 hits in 3 tries or 3 hits in 3 tries}). \quad (3) \]

Similarly, getting 3 hits in 3 tries, say, requires getting a hit on the first try, and a hit on the second try, and a hit on the third try, so

\[ P(\text{3 hits in 3 tries}) = P(\text{hit on 1st try and ... and hit on 3rd try}). \quad (4) \]

If we knew how to compute probabilities like \(P(A \text{ or } B)\) in terms of the simpler component probabilities \(P(A)\) and \(P(B)\), equation (3) would be progress, and that's where the addition rule comes in. Similarly, equation (4) suggests that
it would be nice to be able to express probabilities like \( P(A \text{ and } B) \) in terms of \( P(A) \) and \( P(B) \). That’s the point of the Product Rule in section 4.

But I’m getting ahead of myself; let’s look first at how or works. Figure 4 gives a Venn diagram with two events, A and B, in it. If you imagine shooting at random at the Venn diagram (by the end of this part of the book you may indeed want to shoot at the damn things) and recording the probability of an event as the fraction of time your shot lands inside the blob that stands for the event, it is evident that the fraction of time you’d hit either A or B in Figure 4 must just be the sum of the two quantities (fraction of time you’d hit A) and (fraction of time you’d hit B). In other words,

\[
P(A \text{ or } B) = P(A) + P(B).
\]

This is the simplest form of the Addition Rule. But does this rule always work? If you think a minute, you can see that it can’t always be right just to add the separate chances of A and B—what is to prevent the sum from going over 1, which we know from the Easy Rule is impossible? Evidently there must be something special about how the events A and B in Figure 3 relate to each other; in other words, there must be some other way for a Venn diagram involving two events to look, besides Figure 3. After you play around with it for a minute (I encourage you to do so) you will probably be led to something like Figure 5.

The difference between Figures 4 and 5 is that in Figure 4 the events A and B didn’t overlap, whereas in Figure 5 they do. To see what the overlap in Figure 5 represents, imagine shooting at the diagram, hitting the overlap, and describing to someone where your shot landed. You’d say that it was both inside A and inside B; that is, the overlap must be the event \( (A \text{ and } B) \).

Some events overlap and some do not; in other words, for some events it is possible for both to occur, whereas for other events the occurrence of one prevents the occurrence of the other. To take a simple example, imagine picking a single card out of a standard deck, and consider the following three events describing the card you get: \( A = \{\text{ace}\} \), \( Q = \{\text{queen}\} \), \( S = \{\text{spade}\} \). The events \( A \) and \( Q \) cannot simultaneously occur—it is not possible for a single card to be both an ace and a queen. But the events \( A \) and \( S \) do overlap—the overlapping event is drawing the ace of spades—and so do the events \( Q \) and \( S \). This distinction between events that overlap and events that don’t is important enough to have a name: events that have no overlap are called mutually exclusive. The events in Figure 4 were mutually exclusive, whereas those in Figure 5 were not.
Definition: Two events A and B are mutually exclusive if it is not possible for them to occur simultaneously—that is, if the occurrence of A prevents the occurrence of B, and vice versa.

It is now clear why the addition rule that worked in Figure 4 doesn’t work in Figure 5: by adding \( P(A) \) and \( P(B) \) you double-count the overlap (Figure 6). Evidently the way to fix things is to subtract out the probability \( P(A \text{ and } B) \) of landing in the overlap; this will count the event \( (A \text{ and } B) \) only once, as it should be. So the right answer, in Figure 5, anyway, is to take the formula that worked in Figure 4, \( P(A \text{ or } B) = P(A) + P(B) \), and subtract \( P(A \text{ and } B) \) from the right-hand side:

In Figure 5, at least, \( P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \).

So we seem to have two addition rules floating around: one for situations like Figure 4, another for those like Figure 5. If the world were a nice place we wouldn’t have to carry two addition rules around in our heads, and if you think about it a minute you will see that things are in fact nicer than they seem: if the events A and B are mutually exclusive, the event \( (A \text{ and } B) \) is impossible, so in that case \( P(A \text{ and } B) = 0 \) and the rule for Figure 5 coincides with that for Figure 4. In other words, the rule for Figure 4 is a special case of the one for Figure 5.

As long as there aren’t any other ways for a Venn diagram with two events in it to turn out, we have figured out how or’s work in probability. After a little reflection you can convince yourself that Figures 4 and 5 do exhaust all the possibilities, so I guess we’re ready to write down a general rule for dealing with or:

**Addition Rule for Working with Or.** For any events A and B,

\[
P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).\]

As a special case of this, if A and B are mutually exclusive, \( P(A \text{ and } B) = 0 \) and you get a simpler rule:

For any mutually exclusive events A and B,

\[
P(A \text{ or } B) = P(A) + P(B).\]
There are generalizations of this rule to three or more events, but as formulas
they’re quite ugly (there’s too many overlaps to keep track of) and we won’t
need them in the rest of the book. Problem 7.x gives the formula for three events
and asks you to explain why it’s right, if you’re interested.

To wrap up this section, let’s use the addition rule to finish the aces and
queens example above. To do so we have to specify the chance mechanism
governing how the card is picked. Suppose I shuffle the deck thoroughly (if
you use the standard method, which gamblers call a riffle shuffle (dividing the
deck in half and interleaving the cards from the two halves tightly),
shuffling “thoroughly” turns out, on the basis of some theoretical and empirical work [ref],
to mean at least seven times), and draw a card out from a haphazardly chosen
spot somewhere in the middle. This should pretty well approximate the idea we
talked about in Chapter 6 of drawing a unit from a population at random—in this
case, that is, in such a way that all cards have an equal chance of being chosen.
Then the equally-likely model I mentioned in section 1 would apply, and it’s
straightforward to figure out probabilities like P(ace) and P(spade). For instance,
there are 4 aces out of the 52 cards, so P(ace) = 4/52 = 1/13, which is about 8%
(the same argument applies to P(queen)), and there are 13 spades in the 52-card
deck so P(spade) = 13/52 = 1/4 = 25%.

Now what’s the chance of getting an ace or a queen? Well, according to the
addition rule, P(ace or queen) = P(ace) + P(queen) - P(ace and queen), but we
have already noted that the events {ace} and {queen} are mutually exclusive, so
the simpler form of the addition rule applies: P(ace or queen) = P(ace) + P(queen)
= 1/13 + 1/13 = 2/13, which is about 15%. This is not very likely; 2 out of 13 is
the same as 1 out of 6.5, so the event {ace or queen} would only happen about
once every 6 or 7 times you picked a single card at random from a full deck.

How about P(ace or spade)? The addition rule says P(ace or spade) = P(ace)
+ P(spade) - P(ace and spade), and this time we can’t ignore the overlap since it
is possible for the chosen card to be both an ace and a spade. In fact, we agreed
above that the overlap consists of exactly one card in the deck, the ace of spades,
and by the equally likely model the chance of drawing any single card of the 52
in the deck must be 1/52, so P(ace or spade) = 1/13 + 1/4 - 1/52 = 16/52, about
31%. In this case the probability of the overlap is close to negligible — 1/52 is
only about 2%—but there are lots of examples in which you ignore the overlap
to your peril (I’ll give one in section 5 below). Note that we were lucky in this
case in being able to figure out P(ace and spade) directly, since we haven’t yet
talked about how to work with and’s. That’s the topic of section 4.
3. Conditional Probability

In order to figure out how to work with and in computing probabilities, it turns out that first we need to invent a new kind of probability—a way of expressing how uncertain events are related to each other. The reason is that and is about linkages between things: to know how likely it is for both of two events A and B to occur, intuitively it ought to be necessary to know how the two events depend on each other, and the notion of probability we've looked at so far is inadequate to this task. The material in this section, which may at first look like a detour, is in fact some of the most important stuff in the chapter.

The best way to express dependence in probability terms turns out to be as follows. Suppose A and B are events that will unfold in a particular temporal order—first A, then B, say. I ask you to tell me the probability of B occurring, P(B), and then I ask you to tell me the probability of B occurring given that A has occurred, which is called the conditional probability of B given A—symbolically, P(B given A) (sometimes instead people write P(B|A) – the vertical bar | is read “given”). If B depends on A, the knowledge that A has occurred should change the chance that B will occur. So this suggests a way of quantifying how two uncertain events depend on each other: first define precisely what is meant by P(B given A), and then go around computing P(B) and P(B given A) for various events that interest you—if P(B given A) is different from P(B), then it makes sense to consider the events A and B dependent.

For example, if I draw two cards at random without replacement from a standard deck, the event B = {ace of spades on the second draw} should certainly depend on the event A = {ace of spades on the first draw}, because if you got the spade ace on the first draw and you don’t put the first card back in you can’t get it on the second draw. The value of P(B) = P(ace of spades on the second draw) may not be intuitively obvious (except to say that whatever this probability is, it’s positive, because there is some chance of getting the spade ace on the second draw), but it’s pretty clear that P(B given A), the conditional probability of getting the ace of spades on the second draw, given that you got the ace of spades on the first draw, should be zero. So P(B) > 0 but P(B given A) = 0; in other words, the information that A occurred changes the chance that B will occur, and the events A and B are dependent, which is as it should be.

I have described all of this with events unfolding in a particular temporal order in mind, but the idea is more general than that—it applies equally well, for instance, to examining the relation between two different attributes or descriptions of one single occurrence. Here’s an example.
As a teaching assistant in graduate school at Berkeley in 1979, I taught an introductory statistics course to a class of 106 students one quarter. As part of that course, to generate a dataset that we could analyze, I gave the students a questionnaire in which I asked things like “Do you think marijuana should be legalized?” You would expect a person’s attitude toward this question to be related to, or depend on, some of his or her attributes—like his/her political affiliation and religion—and not to depend on other attributes, like eye color. How do you think a person’s attitude toward legalizing marijuana relates to his or her gender?

Suppose I imagine choosing a student at random and defining the events $B = \{ \text{the chosen person favors legalization} \}$ and $A = \{ \text{the chosen person is female} \}$, say. If you buy the line of reasoning about dependence above, a good way to explore the relationship between these two attributes is to compare $P(B) = P(\text{favors legalization})$ with $P(B \text{ given } A) = P(\text{favors legalization given female}).$ Just as in the example of drawing two cards above, if the information that the person is female sharply changes the chance that he/she supports legalization of marijuana, you would say that marijuana legalization preference depends on gender, or equivalently that a person’s gender is predictive of his or her attitudes toward marijuana, at least among these 106 students.

Okay, enough preliminaries. I guess we need to get down to defining precisely what a person might mean by the conditional probability of something, given something else. This turns out to be our third and last opportunity to make good use of Venn diagrams. Figure 7 shows two such diagrams, one with a single event $B$ in it (Figure 7a), the other with overlapping events $A$ and $B$ (Figure 7b). Using the same motivation as in section 2, you can think of shooting at random at the Venn diagram in Figure 7a repeatedly, and recording $P(B)$ as the fraction of time your shot lands in the blob standing for $B$, so that $P(B) = \text{(area of blob representing } B)$. But this time, I would like to express this idea a bit differently, as follows. You will see where I’m going with this in a paragraph or two, I hope, so please bear with me for a minute.

Since $P(\Omega)$, the probability of the whole rectangle or sample space, is 1, and since dividing a number by 1 leaves the number alone, you could equivalently think of $P(B)$ as the ratio of the area of $B$ to the whole area of the rectangle—$P(B) = P(B)/P(\Omega)$. This way of looking at things makes figuring out a probability a three-part computation:

1. Identify the correct denominator—the relevant universe or sample space, as it were. For $P(B)$ it’s the whole rectangle $\Omega$;

2. Identify the correct numerator—the portion of the relevant universe within which the event of interest is true. For $P(B)$ it’s the blob representing $B$; and
(3) Compute the probability of interest as the ratio of the two areas identified in steps (1) and (2). For \( P(B) \) this is \((\text{area of blob B}) / (\text{area of } \Omega) = P(B)/P(\Omega) = P(B)/1 = P(B) \).

This three-part computation may not look like a very useful trick—for ordinary probabilities like \( P(B) \) it’s certainly more bother than it’s worth. But now consider by analogy what the chance of \( B \) given \( A \) should be, in Figure 7b. We have agreed that when I don’t know anything about whether \( A \) has occurred, the relevant universe, or denominator, for \( P(B) \) is the whole sample space \( \Omega \)—the whole rectangle. But suppose I tell you that \( A \) has occurred. Then none of the rectangle outside \( A \) is relevant any more—it’s as if the universe or sample space has shrunk just to the blob representing \( A \). In other words, the relevant denominator for \( P(B \text{ given } A) \) is the blob representing \( A \). What about the relevant numerator? According to step (2) above, it’s supposed to be “the portion of the relevant universe within which the event of interest is true,” so that must mean the part of the blob representing \( A \) in which \( B \) also occurs—in other words, the relevant numerator for \( P(B \text{ given } A) \) is the blob representing the overlap (\( B \) and \( A \)). From this step (3) is easy: \( P(B \text{ given } A) = (\text{area of } (B \text{ and } A))/(\text{area of } A) = P(B \text{ and } A)/P(A) \). We have derived the standard definition for conditional probability:

\[
P(B \text{ given } A) = P(B|A) = \frac{P(B \text{ and } A)}{P(A)}.
\]

This definition makes pretty good intuitive sense, when you think about it a bit: the probability of \( B \) given \( A \) is the fraction of time that both \( B \) and \( A \) happen, among the times that \( A \) happens. Until you have used the definition enough that it’s second nature to you, maybe you can remember something like “the chance of something given something else is the chance of them both over the chance of the one you’re given,” or you could try to remember Figure 7b and work the definition out from the picture. Note that the definition doesn’t work if \( P(A) = 0 \), because division by 0 is undefined, but that’s okay since \( P(A) = 0 \) means that \( A \) is impossible and it doesn’t make sense for the event you’re “given” in \( P(B \text{ given } A) \) to be impossible.

By now we are in dire need of something to bring all this abstraction into focus, so let’s finish the marijuana example I started at the beginning of the section. As usual, it helps to visualize the data. Recall that we are wondering how
a person’s attitude toward the legalization of marijuana depends on the person’s gender in a group of 106 students, so it sounds to me like there should be 106 rows in the dataset—one for each student—and 2 columns—one for each variable, with attitude toward legalization as the outcome variable on the left, and gender as the predictor variable on the right. But what values do these variables take on? Gender is straightforward enough—male and female—but you can measure people’s attitudes about things like marijuana legalization in a variety of ways, from a simple yes/no all the way up to (let’s say) a nine-point ordered categorical scale (1 = definitely opposed, ... 3 = moderately opposed, ... 5 = neutral, ... 7 = moderately in favor, ..., 9 = definitely in favor). Here I’m going to stick with a simple yes/no answer to the question “Do you favor legalization?” Table 1 gives the data, coded in this way.

Well, that’s about as boring a table as I’ve seen in awhile. The problem with it is that it is highly redundant—in fact, there’s only four possible data values in it, repeated over and over: (no, female), (yes, female), (no, male), (yes, male). If instead we were to keep track of how many of the 106 students fell into each of these four categories, the only information lost in reducing Table 1 to this form would be the order in which the students’ data values were written down, which is presumably irrelevant. So this reduction of the data sounds like a good idea, and the only question is how best to display the counts summarizing how the students distributed themselves across the four categories.

The standard way is by means of a table with rows standing for the values of one of the variables we’re interested in and columns standing for the values of the other variable. It is arbitrary which variable goes where; I’m going to put the predictor variable, gender, in the rows and the outcome variable, legalization preference, in the columns (Table 2).

This is called a contingency table, in this case a 2 by 2 table to be explicit about how many rows and columns it has. People also call tables like this cross-tabs—short for cross-tabulations—and often go on to say which variables are being cross-tabbed, as in the statement “Table 2 is a cross-tab of gender against legalization preference.” Notice that both variables in Table 2 are categorical—contingency tables are a good way to display the relationship between such variables. Table 2’s variables are both dichotomies, but there is nothing to prevent us from using this idea when the categorical variables of interest take on more than 2 possible values—we will look at how to analyze the general $r$ by $c$ contingency table (where $r$ is the number of rows and $c$ is the number of columns) in Chapter [xx].
The numbers around the edges of Table 2 are called the margin totals, and summarize things like how many males there were (57) out of the 106 students and how many total students opposed legalization (25). The numbers in the body of the table are the four counts we were trying to display—for instance, 29 of the 106 students were both female and said yes to the legalization question. If I pick a student at random, the equally-likely model will apply and it’s easy to work out various probabilities directly from the numbers in the table.

For example, what’s the probability that the chosen student is female? The equally-likely model says we should count the number of outcomes favorable to the event of interest (49), count the total number of ways the experiment could have come out (106), and form the ratio of these two counts: \( P(\text{female}) = 49/106 = 46\% \) (in other words, the class was not far from 50/50 in gender). How about \( P(\text{yes}) \)? By the same reasoning it’s \( 81/106 = 76\% \) (these were undergraduate college students in 1978. Undergrads have gotten more conservative since then—the fraction of students favoring liberal marijuana laws in the early 1990s would be lower, I think). Other probabilities describing the *joint distribution* (no pun intended) of the two variables, as people say, are equally easy to compute: \( P(\text{male and no}) = 5/106, \text{ about 5}\% \), for instance.

If we want, we can divide all the counts in Table 2 by the total number of subjects (106) and get what you might call a probability table (Table 3), which expresses pretty much the same information as the table based on counts but in a form that’s handier for many purposes. Notice that probabilities like \( P(\text{yes}) = 76\% \) that involve only one of the two characteristics cross-tabbed here appear in the margins of the probability table, whereas the body of the table contains probabilities like \( P(\text{male and no}) = 5\% \) that specify how the two variables relate to each other.

So what exactly do Tables 2 and 3 imply about the relationship between gender and legalization preference? At the beginning of this section I tried to sell you on the idea that a good way to answer this question is to compare the probability of answering yes to the legalization question with the conditional probability of answering yes given that you’re female, say—if information about a person’s gender changes the chance of answering yes, then the two variables are dependent. We already worked out that \( P(\text{yes}) = 81/106 = 76\% \). Let’s compute \( P(\text{yes given female}) \) two ways—from the definition of conditional probability we worked out awhile ago, and from first principles.

The definition says \( P(\text{yes given female}) = P(\text{yes and female})/P(\text{female}) \). We already computed \( P(\text{female}) = 49/106 = 46\% \), and from Table 2 or Table 3 \( P(\text{yes and female}) = 29/106 = 27\% \), so I guess \( P(\text{yes given female}) = (29/106)/(49/106) \)
= 29/49 = 59%—in other words, 59% of the women favored legalization. The first-principles way of computing this chance is to recall that the idea behind conditioning on gender, as people say—in other words, being given the information that the chosen person is female, in this case—is to redefine the relevant universe and (in this case) ignore all the male students. Conditional on having chosen a woman, it’s as if the universe has shrunk to the 49 female students in the first row of Table 2, and of those 49 students 29 favored legalization, so P(yes given female) = 29/49—the same answer as with the definition, but arrived at a little more directly. Evidently conditioning on something in a contingency table amounts to focusing in only on the relevant row or column specified by what you’re conditioning on, so by the same reasoning to get P(yes given male) you look only in the “male” row in Table 2 and get 52/57 = 91%. Another way to say this that more closely conforms to regular conversation is that among the male students, 91% favored legalization, so “among” must be a code-word for “given” in the language of conditional probability.

So what have we learned? Without knowing the person’s gender, the chance that a randomly chosen student favors the legalization of marijuana, in this group of 106 students, at least, is 76%, but this figure drops to 59% among the women and rises to 91% among the men. That sure sounds like gender and legalization preference are dependent—strongly so, in fact: being told a person’s gender sharply changes the likelihood that he or she favors legalization—but all we have shown is that these two attributes were dependent in this group of 106 students. Generalizing outward from this collection of people to some larger population is statistics, not probability, and we will get to that later, in Chapter [xx].

By the way, in this demonstration that the two attributes were dependent, was there anything special about comparing P(yes) to P(yes given female) and P(yes given male)? Would we have arrived at the same conclusion if we had conditioned in the other order, as people say—in other words, if we had compared P(female) to P(female given yes) and P(female given no)? Intuitively, it shouldn’t depend on which variable is taken as the outcome and which as the predictor—if you can use gender to predict legalization preference, you ought to be able to use legalization preference to predict gender. Well, let’s see: from Table 2, P(female) = 49/106 = 46%, P(female given yes) = 29/81 = 35%, and P(female given no) = 20/25 = 80%—in other words, just as was true when we looked at the two variables in the other order, gender and legalization preference are strongly dependent. This phenomenon is quite general—to check dependence between two attributes or outcomes, you only have to condition in one direction (Problem 7.x).

One more by-the-way: Remember the Easy Rule? We agreed back in section
1 that, for any event B, say, \(0 \leq P(A) \leq 1\), and \(P(B) + P(\text{not } B) = 1\). There ought to be a version of this rule for conditional probability as well, and there is: among all the different outcomes in which some event A happens (in other words, conditioning on the event A), any other event B must either happen or not happen, and the fraction of time both B and A happen cannot drop below zero or exceed the fraction of time just A happens (why?), so [remotive this?]

<table>
<thead>
<tr>
<th>The Easy Rule for Conditional Probability. For any events A and B,</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0 \leq P(B \text{ given } A) \leq 1, \quad \text{and} \quad P(B \text{ given } A) + P(\text{not } B \text{ given } A) = 1.]</td>
</tr>
</tbody>
</table>

4. The Product Rule for Working with And

This is maybe a good moment to remind us of the big picture in this chapter, which is as follows. To compute complicated frequentist probabilities, people re-express the event of interest in terms of simpler events with and's, or's, and not's, and then use various rules for working with these three logical operators to express the probability of interest in terms of the probabilities of the simpler events. We have covered not and or, but we haven't polished off and yet. On that score, I have some good news: Now that we have gone to all this trouble to define and get familiar with the idea of conditional probability, and is easy. In fact, all we have to do is to take the definition of conditional probability,

\[P(B \text{ given } A) = \frac{P(B \text{ and } A)}{P(A)},\]

multiply both sides by \(P(A)\),

\[P(A) \cdot P(B \text{ given } A) = P(B \text{ and } A),\]

and note that \(P(B \text{ and } A) = P(A \text{ and } B)\), since the events (B and A) and (A and B) are just two ways of expressing the same overlap set. The result is

\[P(A \text{ and } B) = P(A) \cdot P(B \text{ given } A).\]  \hspace{1cm} (5)

This rule is gratifyingly simple and involves multiplication, which matches up nicely with the word “product” in the title of this section, so I guess we're on
the right track. In words, equation (5) says: To compute the chance that both of
two events A and B will happen, work out the chance of A, and the chance of
B given A, and multiply. If you think about it a bit, this makes eminently good
sense: the fraction of time both A and B happen is the fraction of time A happens
times the fraction of time B happens among those occasions on which A happens.
(Can you give an example of this from everyday life?) Expressing things this
way makes me think that there's nothing special about starting with A and then
going on to (B given A)—I ought to be able to do it in the other order and get
the same thing. And indeed you can check that interchanging the role of A and
B in the argument above gives

$$P(A \text{ and } B) = P(B) \cdot P(A \text{ given } B) ,$$  \hspace{1cm} (6)

so it doesn't matter which event you start with. In practice people choose between
equations (5) and (6) based on which probabilities are already known or the easiest
to compute using common sense.

We're almost ready to state the product rule, but (as was true for the addition
rule) there's an important special case, in which things are a little simpler, that's
worth laying out explicitly. I have repeatedly emphasized the idea that two events
are dependent if the knowledge of how one of them turned out changes the chances
governing how the other one will turn out. I guess to be consistent I should say
that two events are independent if knowing that one of them happens doesn't
affect the chance of the other one happening. In symbols what this means is that

Two events A and B are independent if (and only if)

$$P(B \text{ given } A) = P(B) \text{ and } P(A \text{ given } B) = P(A) .$$  \hspace{1cm} (7)

To get a simpler-looking expression that gives a different insight into indepen-
dence, we can substitute the left-hand equation on line (7) into equation (5),
or equivalently the right-hand equation on line (7) into equation (6). Either way
what you will get is that

Two events A and B are independent if (and only if)

$$P(A \text{ and } B) = P(A) \cdot P(B) .$$

What all of this means is two things:

1. We have two different ways to define the independence of some events A and
   B, and now is as good a time as any to state the definition explicitly:
Definition. Two events A and B are independent if and only if

\[ P(B \text{ given } A) = P(B) \text{ and } P(A \text{ given } B) = P(A), \]

or equivalently if and only if

\[ P(A \text{ and } B) = P(A) \cdot P(B). \]

2. The product rule for working with and is simpler for independent events: evidently if A and B are independent, to get the chance of them both happening you can just work out their separate chances and multiply. This again makes good sense: if A happens 50% of the time, for instance, and B—unrelated to A—happens 10% of the time, then B will happen along with A on 10% of the 50% of the occasions A happens, and 10% of 50% is 0.1 \times 0.5 = 0.05 = 5%.

So now I'm finally ready to summarize all this by stating the Product Rule.

**Product Rule for Working with And.** For any events A and B,

\[ P(A \text{ and } B) = P(A) \cdot P(B \text{ given } A) = P(B) \cdot P(A \text{ given } B). \]

As a special case of this, if the events A and B are independent,

\[ P(B \text{ given } A) = P(B) \text{ and } P(A \text{ given } B) = P(A), \]

and you get a simpler rule: If A and B are independent,

\[ P(A \text{ and } B) = P(A) \cdot P(B). \]

The way the product rule generalizes to more than two events is interesting. I won't prove it (you can try your hand at the proof, if you want; it's not hard), or spend much time motivating it, but the general product rule for three or more events comes out looking like a kind of chain rule, a little like the one you probably learned in calculus for differentiating functions like \( f(g(x)) \). For example, with three events the product rule comes out

\[ P(A \text{ and } B \text{ and } C) = P(A) \cdot P(B \text{ given } A) \cdot P(C \text{ given } (A \text{ and } B)). \]
I bet you can see how this would generalize to an arbitrary number of events: the idea is to chain along, conditioning on everything that came before—A, then B given A, then C given both A and B, etc. What I mean by $P(C \text{ given } (A \text{ and } B))$ in this formula, by analogy with the definition of conditional probability for two events, is

$$P(C \text{ given } (A \text{ and } B)) = \frac{P(C \text{ and } A \text{ and } B)}{P(A \text{ and } B)}.$$ 

[work in independence as early as possible]

5. Some Examples

6. Summary

7. Problems
### Tables for Chapter 7

<table>
<thead>
<tr>
<th>Favor Legalization?</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>female</td>
</tr>
<tr>
<td>yes</td>
<td>female</td>
</tr>
<tr>
<td>no</td>
<td>male</td>
</tr>
<tr>
<td>yes</td>
<td>male</td>
</tr>
</tbody>
</table>

Table 1: The marijuana legalization preference raw data.

<table>
<thead>
<tr>
<th>Favor Legalization?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>29</td>
<td>20</td>
<td>49</td>
</tr>
<tr>
<td>Male</td>
<td>52</td>
<td>5</td>
<td>57</td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>25</td>
<td>106</td>
</tr>
</tbody>
</table>

Table 2: 2 by 2 contingency table summarizing the raw marijuana legalization preference data.

<table>
<thead>
<tr>
<th>Favor Legalization?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>.27</td>
<td>.19</td>
<td>.46</td>
</tr>
<tr>
<td>Male</td>
<td>.49</td>
<td>.05</td>
<td>.54</td>
</tr>
<tr>
<td>Total</td>
<td>.76</td>
<td>.24</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 3: The probability table for the marijuana data, obtained from Table 2 by dividing all the counts by the total number of subjects (106).
Figures for Chapter 7

deduction (probability)

<table>
<thead>
<tr>
<th>population</th>
</tr>
</thead>
</table>

\[ \rightarrow \]

induction (statistics)

Figure 1: Probability (deduction) versus statistics (induction).

| no uncertainty \(<--\) uncertainty --> \(\rightarrow\) uncertainty |
|---------------------|---------------------|
| \(0\)               | \(0.5\)              |
| \(^\wedge\)          | \(^\wedge\)          |

the event is certain to happen

maximum uncertainty

Figure 2: The 0-1 probability scale as a quantitative expression of uncertainty about the occurrence or non-occurrence of an event \(A\).

\[ P(A) = \text{area of } A \quad \text{and} \quad P(\text{not } A) = \text{area of } (\text{not } A) \]

Figure 3: Venn diagram illustrating the fact that \(P(\text{not } A) = 1 - P(A)\).

\[ P(A) + P(\text{not } A) = 1 \]

\[ \text{area of } A + \text{area of } \text{not } A = 100\% \]

so \(P(\text{not } A) = 1 - P(A)\).
Figure 4: A Venn diagram with two events A and B.

Figure 5: A Venn diagram with two events A and B that overlap.

Figure 6: When the events A and B overlap, adding $P(A)$ and $P(B)$ double-counts the overlap.

Figure 7: (a) $P(B) = \frac{\text{(area of B)}}{\text{(area of Omega)}} = \frac{P(B)}{P(\text{Omega})}$.
(b) $P(B \text{ given } A) = \frac{\text{(area of (B and A))}}{\text{(area of A)}} = \frac{P(B \text{ and } A)}{P(A)}$. 

91.3
8. Bayes’ Theorem, and How Bayesians Think About Probability

This chapter is about the second of the two ways people have so far invented to quantify uncertainty through probability: the Bayesian, or subjective, approach. The story with this way of defining probability is quite different than the frequentist method we looked at in Chapter 7, and the differences have been the subject of considerable dispute. With plenty of life-and-death matters to occupy the world, you might not think that anything as trivial-sounding, in the grand scheme of things, as a disagreement over how to define what people mean by statements like “The chance of rain tomorrow is 40%” would be a fighting matter, but if you thought that, you’d be wrong.

In fact, from time to time over the last 250 years or so, the dispute between the two camps has taken on an undertone of almost religious fervor, which has led to the closest thing to name-calling and rowdiness you are likely to see among statisticians (who are by and large a pretty sober lot). The argument has been simmering since about 1750, when a mathematics paper by an English [faith] minister and philosopher named Thomas Bayes was published posthumously—the event that got the Bayesian ball rolling. In this paper Bayes proved the theorem that bears his name—we will look at his result in section 4 below—and applied it to [...] its first controversial application, and by no means its last]. I think it’s important for anybody who wants to know how to think clearly about uncertainty to hear both sides of the controversy, but to lay out the issues clearly the first three sections of this chapter are pretty philosophical—we don’t get to actually computing Bayesian probabilities until section 4. It may be tempting to skip the philosophical stuff to get to Bayes’ Theorem faster, but I don’t recommend this—it is at least as important to know what the output of the Theorem means as it is to know how to use it. The chapter is long, and bears several readings, but if you can absorb it you’ll have a valuable new tool to use in your work.

I tried to tell the frequentist story in a pretty even-handed way in the last chapter, and I will try to do so with the Bayesian story here, so that you can choose for yourself what blend of the two approaches is best in your own work. Personally I am some kind of agnostic—it’s pretty clear to me that in any intellectual dispute in which both sides are still standing after 250 years of slugging it out, there must be elements of merit in both positions—but (so that you know where I’m coming from) I’m basically a Bayesian who is sympathetic to the good features of the frequentist story, rather than the other way around. I will try to convince you
that the Bayesian approach is more general and flexible than the frequentist story, but I will also try to be frank about the potential difficulties of implementing the Bayesian way of thinking in real-world applications. Fortunately, as we will see, there’s a fair amount of overlap in the actions suggested by the two approaches when you’re trying to quantify and manage uncertainty in practice, although there may still be considerable disagreement over the meaning of the results produced by the two schools of thought. That may sound like a funny state of affairs—how could people who disagree over the meaning of something fundamental in a long-running dispute consistently agree on the right course of action?—but there it is.

Well, what does it mean to be a Bayesian, anyway? The Bayesian way of defining probability is similar to the frequentist story in that its underpinnings are in gambling ideas, but the way the gambling motif is used to motivate the assignment of probability is quite different. Recall that in the frequentist world we have to be able to imagine repeating the event of interest, and observing the relative frequency with which it occurs, to define its probability, something that (as we saw in Chapter 7) is particularly easy to do with gambling: spinning the roulette wheel again and again, for example, is at the heart of the game. In the Bayesian world, no such repetitions, real or hypothetical, are needed. Instead, to assign a Bayesian probability to something, I imagine gambling with someone about whether the thing in question is true and I ask myself what odds I would need to give to, or receive from, this hypothetical gambler in order that the gamble is fair.

1. Odds and Probabilities

To make this idea clear, before going on I had better wander off on a little tangent explaining exactly what people mean by “odds.” This is also a good moment to motivate and state the relationships between odds and probabilities, because we will be using odds a fair amount from now on—Reverend Bayes’ Theorem turns out to be easier to use when expressed in terms of odds than in terms of probabilities.

If you live near a racetrack and you look in the sports section of your local paper, you will see “handicapping tables,” as they’re called, which give the odds of various horses winning in each race. These tables are filled with terms like “even money,” “2 to 1 favorite,” and “5 to 2 against” that require some decoding. For now, I’m just going to state how odds work, and hope that the reasons for the rules below will become clear a bit later. Also for now, I will use frequentist language
in presenting and discussing these rules, even though with a little rewording they are equally valid for Bayesians.

**How odds work**

1. The odds in favor of an event can either be expressed with two numbers separated by the word “to,” as in “3 to 1,” which I will write 3:1, or by a single number that is the ratio of these two numbers. If the odds $o$ of an event using the : notation are $s : t$, then $o = \frac{s}{t}$.

2. "Even money" is just a shorthand for 1:1 odds, or $o = 1$.

3. Odds of, for instance, 2 to 1 against an event are the same as 1:2 in favor of it, so there is no need to carry the words "in favor of" and "against" around, if you just adopt (as I will) the convention that whenever the directionality of an odds statement is not mentioned, it will be understood that the odds are being stated in favor of the event.

For rule (1) above to work, all pairs of numbers that have the same ratio must just be different ways to express the same odds. The gamblers say something is a "5 to 2 favorite," but they could just as well say "2.5 to 1" or "odds = 2.5," except that they don't like to work with fractions.

Now, how do odds and probabilities relate? You have probably seen that when people say something is "even money," they mean that it has a 50/50, or 50%, chance of happening. But when they say the odds in favor of something are 3 to 1, what they mean (appealing to your frequentist intuition) is that, if the process in question were repeated, for every three times the event of interest happens there will be one time it doesn't happen—in other words, the event has probability $3/(3 + 1) = 3/4$. I am looking for a general rule relating odds and probabilities, and this 3-to-1 example makes me think that if the odds $o$ are $s : t$, then the probability $p$ of the event is $\frac{s}{s + t}$. Let's see, this rule checks for even-money events—the odds are 1:1, and the probability is $1/(1 + 1) = 50%$—and it even works for events that are underdogs: if something is 2 to 1 against, then the odds in favor of it are 1 to 2 and the chance of it happening is $1/(1 + 2) = 1/3$, which makes sense. Note also that since $o = \frac{s}{t}$, another expression for the probability in terms of the odds is

$$p = \frac{s}{s + t} = \frac{\left(\frac{s}{t}\right)}{1 + \frac{s}{t}} = \frac{o}{1 + o}.$$

How about in the other direction—if something has probability $p$, what are the odds $o$ in favor of it? By solving the relation $p = \frac{o}{1 + o}$ backwards for $o$, you
can check that the inverse rule is \[ o = \frac{p}{1 - p} . \] In other words, the odds in favor of some event are given by the ratio
\[
\frac{\text{chance of the event happening}}{\text{chance of it not happening}}.
\]
For example, if something has probability 0.95 then the odds in favor of it must be \( 0.95/(1 - 0.95) = 19 \text{ to } 1 \). To summarize this little digression,

<table>
<thead>
<tr>
<th>If the odds in favor of an event are ( s : t ), where ( o = \frac{s}{t} ), then the event has probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ p = \frac{s}{s + t} = \frac{o}{1 + o} . ]</td>
</tr>
<tr>
<td>If an event has probability ( p ) then the odds in favor of it are</td>
</tr>
<tr>
<td>[ o = \frac{p}{1 - p} . ]</td>
</tr>
</tbody>
</table>

2. Bayesian Probabilities as Expressions of Judgment

Okay, enough about odds. What’s the point of all this for defining probabilities with the Bayesian approach? I said at the end of section 1 that the way a Bayesian computes the probability that something is true is to imagine betting with someone about its truth or falsehood, and asking him/herself what odds he/she would need to give to make the gamble fair. Here’s an example of what I mean, drawn from European politics in the post-Cold-War era.

In the fall of 1990, all the countries of Europe were proceeding full steam ahead toward monetary union—the surrendering of their individual currencies (like the deutschmark and the franc), in favor of a common money called the European Currency Unit. All except the British, that is: under the leadership of Prime Minister Margaret Thatcher, the government of Great Britain was reluctant to give up the sovereignty of its currency, the pound sterling. Thatcher’s foot-dragging on this and other aspects of Britain’s economic union with the European Community led people within her own party, the Conservatives, to question her leadership. What started out as back-room grumbling eventually escalated into open rebellion: on November 14, 1990, a man named Michael Heseltine, the former defense secretary of Britain, declared his candidacy to head the party,
a challenge that, if successful, would lead to Thatcher losing her job as Prime Minister.

Thatcher vowed to fight for her job, creating the prospect of a spirited contest. The British love a good fight, and love even more the opportunity to bet on such a fight. Betting of many kinds is quite legal in Britain, so much so that London bookmakers routinely quote odds on all sorts of propositions. In this spirit, as soon as Thatcher announced that she would contest Heseltine, a prominent bookmaker named William Hill made her the 2-to-1 favorite to win the leadership election. This translates directly into a Bayesian probability: as far as Mr. Hill is concerned, \( P(\text{Thatcher will win}) = 2/3 \). [footnote: it was actually more complicated than this because there was another guy—Douglas Hurd—in the race; pari-mutuel, etc. but this gives the right flavor.]

Now, coming to the last few sentences with the frequentist perspective of Chapter 7, there ought to be at least three things that sound strange:

1. All of a sudden I’m talking about propositions rather than events as the things to which we assign probabilities.
2. There doesn’t seem to be anything preventing two people with different beliefs from having different probabilities for the same thing, so that there’s no such thing as “the” probability of something at all. In the Thatcher example, for instance, it sounds as if two bookmakers would be free to quote completely different odds, leading to potentially quite different Bayesian probabilities that Thatcher would win.
3. The process of arriving at a probability seems awfully informal—you just ask yourself how you feel about something, in the sense of gambling about it, and out comes the probability.

What exactly is going on here?

Well, as I said at the beginning of the chapter, the frequentist and Bayesian stories are quite different. I’ll try to address the first two of the points above in the rest of this section; the third requires a longer discussion and will be the topic of the next section.

First of all, if the only requirement in assigning probabilities to things is to imagine gambling about them, we are freed from having to imbed the propositions we’re wondering about in hypothetical or actual sequences of repetitions, and we don’t need the machinery of sample spaces and events. This means that statements that—as we will see in a minute—are very hard to fit into the relative frequency framework, like (Francis Bacon actually wrote all the stuff we attribute to Shakespeare), or (Margaret Thatcher will win the leadership fight), are fair game.
to a Bayesian, in addition to things to which we can readily assign probabilities as frequentists, like (it will rain tomorrow). In this key sense the Bayesian story is more general than the frequentist approach.

Why exactly is it hard to approach things like \( P(\text{Thatcher will win}) \) from the frequentist point of view? Recall that there are two ways to develop a sensible frequentist probability for something—you can either appeal to a history of relevant repetitions in the past, or you can use common sense and equally-likely-type considerations to say what the relative frequency ought to be if you were to repeat the event of interest in the future. Unfortunately neither of these approaches works with something like \( P(\text{Thatcher will win}) \): there have never been any Conservative Party leadership fights exactly like the one we’re wondering about—Thatcher versus Heseltine, on the topic of Britain’s economic role in a united Europe—on which to base a frequency history, and it doesn’t help to think about what ought to happen if the leadership fight were to be repeated, since such repetitions would be so hypothetical. Evidently

- Things to which people would like to assign probabilities (I’ve been calling them propositions, and that’s as good a name as any for them) are of at least two types—*inherently repeatable* occurrences (what as frequentists we would call events), and *unique* or *one-time* occurrences; and
- What we are seeing is that it is hard to attach sensible frequentist probabilities to one-time occurrences.

Actually things are even worse than this for the frequentist approach, as far as general applicability is concerned: there’s another whole class of propositions we haven’t discussed yet, and the frequentist story is no more helpful in this new case than it was with one-time events. So far all the propositions we’ve looked at have been *occurrences* that had a definite location in time, which could either have been in the past—\( P(\text{Bacon wrote “Shakespeare”’s plays and sonnets}) \), for example—or in the future—\( P(\text{it will rain tomorrow}) \), say. But what about propositions like (There is life on other planets) or (The atomic weight of nitrogen is somewhere between [xx] and [xx])?

These are *statements of (possible) fact* without any particular appeal to time built into them, and if we are uncertain about them (I certainly am about the first one) we ought to be able to use probability to express this uncertainty. However, if you think about it, it is again hard to come up with sensible frequentist probabilities for statements of possible fact, because (for instance) the atomic weight of nitrogen is not like a coin that takes on a different possible value (heads or tails) each time I flip it: it’s just some number I’m not certain about that does not change with
repetition. [footnote: I guess if you wanted to you could imagine something like repeating the whole process of creating the universe over and over again and ask how often an element like nitrogen would appear with atomic weight between [xx] and [xx], but that's even more hypothetical than the repetitions in the Thatcher example.]

To summarize the last couple of paragraphs,

A proposition is something about which you’re uncertain and to which you would like to attach a probability as a way of quantifying that uncertainty. There are two kinds of propositions: occurrences (things that may or may not have happened in the past, or that may or may not happen in the future), and statements of possible fact without any particular reference to time. Occurrences can either be inherently repeatable or unique. The frequentist story is good for capturing uncertainty about what would happen if you did something that is inherently repeatable over and over again in the future. Other kinds of uncertainty—about unique, non-repeatable occurrences or statements of possible fact—are much harder to fit into the frequentist framework.

By contrast, no appeal to real or hypothetical repetition of an experiment with the proposition of interest as a possible outcome is needed to arrive at a Bayesian probability for a proposition. For this reason the Bayesian story is useful, at least in principle, for quantifying all kinds of uncertainty, not just uncertainty about repeatable phenomena.

Can Two People Have Different Probabilities for the Same Thing? Let’s turn now to the second thing about Mr. Hill’s Bayesian P(Thatcher will win) of 2/3 that seem funny to a frequentist eye: the possibility that different people could come up with different probabilities for the same proposition. That may sound odd at first, but once you think about it, there’s nothing wrong with it—in fact, it’s a virtue, not a defect. After all, if the two people we’re considering start with different sets of evidence on which to base their probability assessment, or if they have different judgments about what the available evidence implies, they ought to have different probabilities for it, because these are just ways of saying that those people have different levels of uncertainty about the proposition in question. A Bayesian probability is a subjective expression of belief based on (i) whatever evidence the person doing the probability assessment can bring to bear and (ii) how that evidence is relevant, in the judgment of the person doing the assessing. Since people may differ in what they know about a proposition and how they
believe their knowledge bears on it, two different people’s Bayesian probabilities for something could easily be different.

Moreover, although I didn’t mention it in Chapter 7 (in order not to muddy the waters too much early on), this same thing—the possibility of more than one probability for the same proposition—comes up when we try to reason as frequentists as well. For example, recall that at the beginning of Chapter 7 I talked about the chance it will rain tomorrow, and I said that a good way to compute this as a frequentist would be

a. to figure out what you thought tomorrow was going to be like with respect to things that are predictive of rain,
b. to find as many days in the past as you can that were just like your best guess for tomorrow in step (a), and
c. to see how often it rained on the days in step (b).

This is a frequentist probability, and it looks pretty unassailable because it’s based on data that everybody would agree is relevant. But what if you and I disagree about what we think tomorrow will be like, or about which days in the past were “just like” our best guess for tomorrow? Then we will get different frequentist probabilities of rain. Bayesians who had similar disagreements about how the data informs the chance of rain would also arrive at different probabilities.

This is not an isolated instance—it’s just one of our first examples of an important general fact (the coin-tossing example in Chapter 7 was the first). Later on in the book, beginning in Chapter xx, I am going to talk more formally about the process of building a statistical model, as we will call it then. It turns out that building such a model involves telling a story that relates the data you have to the data you wish you had. For instance, in the rain example, “the data you have” is the past information about when it did and didn’t rain as a function of things that help to predict rainfall, and “the data you wish you had” is the knowledge, available to us the day after tomorrow, of whether or not it actually did rain. Saying clearly

- what you think tomorrow will be like, as far as rainfall predictors are concerned, and
- which days in the past are just like your best guess for tomorrow

are two examples of the sorts of choices you have to make when you build a statistical model and use it to make an inference or prediction. People call choices like this your modeling assumptions. Typically, you can use the data to help make these choices in a plausible way, but the data values—and the information about how they were gathered—rarely identify a unique set of modeling assumptions.
as the only plausible ones. Most of the time you will have to bring something more than data to the party, so to speak: your judgment about which modeling assumptions seem most reasonable.

If all you want to do is describe when it did and didn’t rain in the past, your job is easier, and involves less judgment—sometimes the data are so clear-cut that everybody would agree on a good description, and individual judgment plays a minor role. But whenever you want to infer outward from what you know to what you wish you knew—whenever you want to make an inference or prediction, in other words—personal judgment in both the choice of modeling assumptions and the body of relevant information becomes much more important. Since reasonable people can disagree about things that involve the exercise of judgment, we can easily end up with two or more probabilities for the same proposition, whether we are reasoning in a frequentist or Bayesian manner.

Well, that’s a fine kettle of fish—things were looking pretty clear-cut back in Chapter 7, and now all of a sudden it’s not black and white, it’s all shades of grey. What’s worse, I’ve been telling you all along that probability was supposed to be a good way to express your uncertainty about something, and now I’m telling you that it’s perfectly possible for me to get .28 for some probability based on one reasonable set of assumptions and for you to get .32 based on a different reasonable set. In other words, because of uncertainty about the modeling assumptions, we seem to be uncertain about how much uncertainty we have, which sounds hard to handle in the probability framework. How do people deal with this new type of uncertainty?

It’s too early to give a complete answer to this question, but I can say roughly what is done in practice: you write a little report in which you say “If I assume so-and-so I get .28, but if I assume such-and-such instead I get .32. Both of these sets of assumptions seem pretty reasonable to me (here’s why: ...), so I think the probability is around .3.” People call this way of exposing the relationship between assumptions and results sensitivity analysis, because what you’re doing is varying your assumptions across a reasonable range and seeing how sensitive the answer is to what you assume. If all roads lead to Rome, so to speak—that is, if you get pretty much the same thing no matter which set of reasonable modeling assumptions you pick—then you’re lucky, and you announce the (more or less unique) answer. If not, you say what you get under each reasonable set of assumptions, and you recommend what kind(s) of additional data to gather to clear up the uncertainty about which model is right. I will give some examples of this process in section 4 below and in the later parts of the book on inference and prediction.
I bring all of this up not to make you uneasy but to try to say again clearly something I started to say all the way back in Chapter 1: everything people do, including assessing uncertainty and trying to make choices in the face of it—probability and statistics, in other words—involves judgment and is therefore inherently subjective. Numbers may look pretty objective, but they are always based on assumptions. The best we can do in both science and decision-making is to state our assumptions clearly to each other and to be prepared for the possibility that the right course of action, based on the data so far, may be to gather additional data. This is par for the course in science—indeed, it’s just one way of saying how science works—but it’s not always a popular message in decision-making, where people tend to feel better doing something that seems more active than just getting more data.

To summarize the last nine or ten paragraphs,

Bayesian probabilities are personal expressions of belief about the plausibility of the propositions of interest. Two people with different knowledge bases or different judgments about how that knowledge informs the proposition in question will have different assessments of the proposition’s plausibility, and thus will have different Bayesian probabilities of it being true.

The same thing happens when we reason as frequentists, because data are never enough to arrive at an inference or prediction—you always have to supplement the data with modeling assumptions that express your judgment about how the data are relevant to the inference or prediction at hand. Reasonable people may differ in the exercise of this judgment and arrive at different frequentist probabilities for the same proposition.

The point is that in practice, judgment plays a key role in implementing both the frequentist and Bayesian stories. It often makes people uncomfortable to acknowledge how subjective both science and decision-making really are, but rather than trying to avoid this inescapable fact, it seems best to (a) state all your assumptions clearly, and (b) use sensitivity analysis to explore the relationship between assumptions and results.
3. Coherence and Calibration

As reasonable as the Bayesian approach has sounded so far, the something-for-nothing bell should still have gone off in your head by now—after you think about it a bit, the freedom the Bayesian story implies starts to sound too good to be true. If a Bayesian probability is just an expression of somebody's opinion, what's to prevent that opinion from being flat-out wrong? For example, if I happen to believe so strongly that Shakespeare did not write the plays and sonnets we usually attribute to him that my personal $P$(Shakespeare was the author) is 0.01, but over time stronger and stronger historical and literary evidence accumulates that Shakespeare did indeed write these works, what's to prevent my stubbornly clinging to my 0.01 probability in defiance of the data? Evidently there must be some requirements imposed on how Bayesian probabilities are manipulated, and updated when new evidence comes in, to insure that the Bayesian story produces sensible results. People have identified two kinds of requirements along these lines—coherence and calibration. Discussing what these two terms mean is the goal of this section. I should say that the intent here is to give you some things to think about as you read the rest of the book rather than to demonstrate how to actually compute Bayesian probabilities; that's the topic of the next section.

Coherence. The idea behind coherence is that if you ask a Bayesian his/her opinion, in probability terms, about two or more propositions, you would like the answers you get to be internally consistent—there isn't much point listening to the opinions of people who contradict themselves. The way this concept is expressed in the theory that underlies the Bayesian story is to require a Bayesian to give his/her odds for the two or more propositions in such a way that no one betting against him/her is guaranteed to make money. A set of Bayesian probabilities that is consistent in this way is called coherent. [give a simple example]

How is coherence achieved? It sounds hard—in fact, harder and harder, the more probabilities you try to assess simultaneously. It turns out to be easier to say when someone is behaving incoherently than coherently. In this regard, remember how in Chapter 7 we found it so useful to work out complicated probabilities by expressing them in terms of simpler ones and using the rules in that chapter to manipulate the simpler probabilities (pp. xx-xx)? This technique turns out to be just as useful for Bayesians as frequentists. The reason is that people have connected the frequentist and Bayesian stories by proving the following remarkable, and reassuring, result: [footnote: Mathematicians would say that the
If, when reasoning in a Bayesian manner, you ever manipulate probabilities in a way that violates any of the rules governing frequentist probability we developed in Chapter 7, like the Addition and Product Rules, or any other rules that follow logically from them that we haven’t looked at yet—the most important of which is Bayes’ Theorem, the subject of section 4 below—the probabilities you would produce in violation of these rules would be incoherent.

rules in Chapter 7, Bayes’ Theorem, and all other rules like them that follow logically from the frequentist definition of probability are a necessary condition for coherence.]

This means that the Easy Rule, the Addition Rule, the definition of conditional probability, the Product Rule, and so on are just as important to you with your Bayesian hat on as they are when you’re reasoning as a frequentist: if, as a Bayesian, you manipulate probabilities in violation of any of the ground rules defining the frequentist game, the results are guaranteed to be stupid. This is a great relief—otherwise, Bayesian would have a whole set of probability rules distinct from the frequentist ones, and this chapter would be a lot longer than it already is. There’s evidently no need to say things like “the rules in Chapter 7” any more; we may as well just refer to them as the rules of probability.

The fact that everybody agrees on the rules of probability leads to a simple behavioral rule for Bayesians: To figure out your personal Bayesian probability of some proposition being true, use all the rules you can think of to break things down into simpler probabilities (as we did in Chapter 7), and try to identify the smallest set of simpler probabilities necessary to compute the probability of interest (in practice this set is usually quite small). Then you can use judgment and data to elicit these simpler probabilities, as people say—that is, to ask yourself what you think reasonable values for the simpler probabilities might be. This process of elicitation may at first sound difficult, and vague to boot, but it’s actually not so bad—people have done a fair amount of work on how to do it in a coherent way—and in any case when you do sensitivity analysis on the simpler probabilities, varying them across reasonable ranges, you often find that the probability of interest doesn’t depend on them too much. I will give a number of examples of this in section 4 and later in the book.

Calibration. The second of the two types of requirements people try to impose on Bayesian probabilities to insure that they are sensible is an example of how frequentist ideas can be helpful to a Bayesian. Suppose you are trying to forecast the price of oil at a number of points in the future, say 10 (recall that we
talked about this problem a bit in Chapter 1). To be useful, each of these forecasts should have something like the following form: "I am pretty sure that the price of oil in 1995 will be between $20 and $50 a barrel." People call such a statement a predictive interval, because you are specifying a range, or interval, of values within which you think the price of oil will be found at some specified moment in the future. For statements like this to be as useful as possible, the person making the statement should try to quantify the phrase "pretty sure," for instance by saying that in his or her judgment \( P(20 \leq \text{price of oil in 1995} \leq 50) = 0.9 \), say. Later on when we look at predictive intervals from the frequentist point of view, we will call the number 0.9 or 90% the confidence level of the prediction.

Suppose you make each of your forecasts at the 10 different time points with 90% confidence, and we wait and see how accurate your guesses were. In order that your predictive intervals are neither unrealistically narrow nor needlessly wide, ideally we would like around 90% of them—around 9 out of 10, in other words—to be correct, in the sense that the actual price of oil at the indicated moment in time fell somewhere inside your predictive interval for that time point. People call this kind of criterion a calibration requirement, and if we had looked at a lot of your forecasts and seen that you were right about as often as you had advertised through your confidence levels, we would say that you were calibrated.

This idea is essentially frequentist—we are just computing a relative frequency based on how often your predictive intervals include the truth—and it is a goal to which frequentists aspire, but it seems reasonable to ask it of Bayesians as well. It might seem that this goal would automatically be met by insisting that you compute your Bayesian probabilities coherently, but this turns out not to be so—the two requirements, coherence and calibration, are distinct. The reason is as follows. I said above that in practice people compute Bayesian probabilities in two steps:

1. They use the rules of probability to express the complicated probability of interest as a function of a small set of simpler probabilities, and
2. They then use judgment and data to elicit their personal values for these simpler probabilities.

Coherence covers the first of these two steps, making sure that you do the required manipulations sensibly, but it says nothing about whether the judgmental values you come up with for the simpler probabilities are sensible. Calibration is a way of keeping you honest, so to speak, in the specification of these judgmental values.
Just as coherence sounded difficult to achieve, it may sound hard to check whether somebody is calibrated in his or her probability assessments. You can see from the definition of the term that the key is building up a frequency history of relevant predictions or other statements whose truth can be checked. The idea is that if you have been right in the past about as frequently as you ought to have been, given the confidence level of your predictions and other inferences, then we have some justification for the belief that you are worth listening to in your judgments about the future. I will give an example of how to check whether somebody’s predictions are calibrated in the case study on prediction in Chapter xx.

I emphasized “relevant” a minute ago because there are some propositions about which we may be pretty interested for which it’s quite hard to produce relevant calibration information. For instance, what is your personal P(life on other planets)? A frequentist would probably say the question has no good answer—there’s no hypothetical or real sequence of repetitions of anything on which to base a decent frequentist probability. But Bayesians, by thinking about gambling with somebody over whether or not the statement “There is life on other planets” is true, would have no problem coming up with a probability—in fact, I’m sure I could find you ten people with personal probabilities of life on other planets ranging from near 0 to near 1. Unfortunately, if you wanted to bet the farm based on one of these people’s opinions, it’s hard to know whose advice to take—the gamble on which their Bayesian probabilities are based is hypothetical, since we have no way at present to find out whether or not there really is life on other planets, and if you asked each of these people to produce evidence that they had usually been right in the past when they made similar judgments, they would be pretty hard pressed to do so for the same reason.

This is not to say that calibration is an unattainable goal; there are lots of examples of people using Bayesian probabilities in settings where uncalibrated “experts” fall by the wayside. The London bookmakers I talked about in section 2, who quoted odds on Margaret Thatcher’s chances of retaining her job as Prime Minister, are a good instance of this—these people are kept honest by the forces of the marketplace. What I mean by this is that anybody who expresses judgments about an uncertain future that are retrospectively seen to be consistently poor will not last long when there’s serious money on the line.

The British insurance firm Lloyd’s of London is another example that’s actually closely related to the Thatcher case study (insurance, after all, is nothing more than a form of legalized gambling between you and the insurance company). Lloyd’s will cheerfully quote you a price for an insurance policy on the sort
of one-time rare occurrence that would make a frequentist’s hair curl. [give example, e.g. \( P(\text{movie star so-and-so will successfully complete the filming of such-and-such film}) \)] We can work backwards from the insurance policy’s payout and premium—how much you have to pay for the policy—to an implied, and thoroughly Bayesian, probability that the dire event being insured against will occur. [footnote on how to do this] Lloyd’s has been in business since 1xxx—it is pretty safe to conclude that they are paying attention to how often they get it right and adjusting their judgments accordingly.

How is it that Lloyd’s can achieve the goal of calibration when we had such a hard time doing so two paragraphs ago with \( P(\text{life on other planets}) \)? The key difference is that Lloyd’s is in the business (in effect) of quoting odds on future occurrences, for which an automatic feedback loop is present, so to speak—we just have to wait awhile to see if they were right—whereas no such feedback is possible (yet!) with the life-on-other-planets assertion. It is harder to produce demonstrably calibrated Bayesian probabilities for statements of possible fact like (Bacon wrote “Shakespeare”’s plays and sonnets) and (There is life on other planets) than for assertions about possible future occurrences.

The message I want to leave you with is that calibration of probabilities—Bayesian or frequentist—is a goal that is not always attained but that is always worth trying to attain. I say “Bayesian or frequentist” because the issue also comes up for frequentists: when you notice you’re out of calibration, the problem is typically that some judgment you’ve made somewhere along the way in producing a probability statement is (retrospectively) wrong, and (as we saw earlier in this section) judgment plays just as much of a role in the frequentist story as it does in the Bayesian world. If I make incorrect modeling assumptions in producing any probability, frequentist or Bayesian, it could easily be a poor summary of how much uncertainty I really have. As I mentioned back in section 2, our first example of this came up back in Chapter 7—if I assume that the coin you’re tossing is fair when in fact it’s two-headed, then I will make gambling decisions based on the belief that \( P(10 \text{ heads in a row}) \) is almost zero when in fact it’s one, and you will laugh all the way to the bank.

To summarize this section,
Any probability assessment, whether frequentist or Bayesian, may either be good or bad, where “good” means “an accurate reflection of how much uncertainty you have about the proposition in question.” In particular, there is nothing built into the Bayesian machinery guaranteeing that probabilities produced using it won’t be stupid (incoherent) or out of calibration with the world. When you find yourself in the role of consumer of somebody else’s probability assessment (let’s call this person the producer of the assessment), it is good form both to check the logic used to arrive at the answer and to ask for evidence that relevant probability assessments made by the producer in the past have been right about as often as might be expected, given the confidence levels with which they were announced.

By the way, did William Hill, the bookmaker whose $P(\text{Thatcher will win})$ was $2/3$, get it right? Well, almost, in a funny sense: Thatcher failed to win a first-ballot victory by only 4 votes out of XXX, and then suddenly resigned and threw her support to one of her darkhorse followers, John Major, in order to defeat her antagonist Michael Heseltine. Of course, you’ve probably heard that old saying about near misses only counting in horseshoes, which—unfortunately for Mr. Hill and the thousands of punters who followed his advice—is quite true for gambling. (This does not prove that Mr. Hill is uncalibrated, though—it just counts as one piece of data in his personal loss column. We would need a fair amount more data before we would have ample reason to seek expert opinion elsewhere.) For my part, I’m just glad nobody made me say what my personal probability of Thatcher winning was and then back it up with money—it sounded like a hard race to call.

4. Bayes’ Theorem and the Updating of Evidence

Okay, (more than) enough metaphysics; how do you compute these damn Bayesian probabilities? To answer this question well, I need one last bit of preparation: a few words about the different kinds of evidence people bring to bear when solving a problem in science or decision-making. As usual, I have an example; this one is from sports, specifically college basketball.

The Two Kinds of Evidence in Science and Decision-Making. If you follow college sports at all, you have probably noticed the multitude of polls published in the newspaper each week ranking the top teams in the country. In basketball the two top polls are conducted by the major wire services, AP and UPI; the former is a poll of leading sportswriters, the latter of top college basketball coaches. These
polls are conducted each year from late November, just before the regular season begins, until late March, just before the post-season tournament that determines the national championship team.

Let's concentrate on the AP poll. If you are a sportswriter invited to participate in it, you write down a list each week giving your opinion of the top 20 teams in the country, ranked from 1 down to 20, and mail it in to the AP. You can think of your job as predictive in nature: the AP wants to know which teams you think, at this moment in the season, will be the top 20 teams at the end of the season. The people running the poll collect the responses from you and about xx other sportswriters, award 20 points for each vote for number 1, 19 for each number 2 vote, and so on, and add everything up. The team with the highest total is number 1 that week, the one with the second-highest total is number 2, and so on. Table 1 gives the result of this process on December xx, 19xx, a point in the 19xx-xx season when each team had only played a few games. What I want to focus on here is this: if you were asked to vote in such a poll every week from preseason all the way through to the tournament, what kinds of evidence would you use in arriving at your weekly vote, and how would you use this evidence?

If you thought about it awhile (you might want to do so before reading on), I bet you'd come up with something like the following scheme. At the beginning, for the preseason vote, there's no information yet—no data—on how well each team has done this year, so your vote would be a pure expression of your judgment about who will come out on top at the end of the season. Then, each week thereafter, data would arrive, in the form of results of all the games in each team's schedule: a win by a low-ranked team against a higher-ranked opponent, for example, would be evidence that the former's ranking should be improved and the latter's downgraded. In the first few weeks of the season, your vote would still be strongly influenced by your initial preseason judgment—after all, there's not much data yet. But after awhile, enough data would have accumulated that your initial judgment would have little or no effect on your vote.

There is evidence in Table 1 that something like this plan was in use by the sportswriters who voted in the December xx, 19xx poll: look, for example, at the ranking of Louisville, a team with a record of 1-2 (one win and two losses) at the time the poll came out. Readers of this poll have every right to wonder what the hell a team with such a lousy record was doing so highly ranked—everybody else in the top 20 at that time had records like 4-0 or 5-1. The explanation is that Louisville was very highly regarded in the preseason by almost all the experts, for three reasons: the strength of their team in 19x[x-1], and the number of 19x[x-1] players who would still be playing in 19xx; the excellence of the new
crop of freshmen and transfer students they recruited for 19xx; and the ability of their coach, who had already won two national titles at Louisville in the previous xx years. This preseason judgment turned out to be sound—Louisville finished among the top 8 teams in the country in the tournament.

Now it’s pretty clear that the sportswriters were using a combination of judgment and data in arriving at their rankings in Table 1, not just data; in fact, their judgmental evidence outweighed the data evidence, or Louisville’s 2 losses would have already dropped them right out of the top 20. It’s also clear in retrospect, from the final results of the 19xx-xx season, that it would have been a mistake for the AP voters to just rely on the data about Louisville available on December xx—only 3 games, and flukey ones at that, evidently. On this occasion, at least, data plus judgment was better than data alone in predicting the season’s final outcome.

This example illustrates two general points. First, there are two kinds of worthwhile evidence when problem-solving in science and decision-making: data and judgment. It is not easy to quantify in a given problem how much of the information you have is based on judgment and how much on data, but it doesn’t do too much harm to conceive of a continuum from 100% judgment to “100% data,” like the one in Figure 1. I put “100% data” in quotes because I have already argued back in section x that such a situation is impossible—you can’t avoid expressing your judgment while problem-solving.

The other extreme—all judgment, no data—comes up quite often, whenever you’re thinking about something nobody’s collected relevant data on, either because it’s not possible to collect such data given current knowledge (P(life on other planets), for instance) or because nobody has gotten around to it yet. It’s a good description of your situation after you’ve planned how to collect some data to solve your problem but before you’ve actually collected it—that’s a good moment to weigh the available pre-data evidence and summarize it judgmentally, like the sportswriters did just before the season began. Bayesians call this your prior information, from the Latin a priori, for “before”. Often this information will be based in part on data that other people have collected that isn’t quite relevant and that you can’t quite fit into your framework—maybe these other people did an experiment similar to yours, say, but they didn’t do it exactly right (in your opinion), or they did it on people who differ somewhat from your subjects in some relevant way. Later in the chapter I will give two examples of prior information and where it comes from in practice.

The second general point illustrated by the basketball poll is that the acquisition of knowledge is typically cumulative, so that learning—attempting to decrease
your uncertainty about something unknown to you—is a process of updating your views about the unknown in the light of new information. For the sportswriters in the poll, the “unknown” was a comprehensive summary of the strongest teams at the end of the season, and the process of updating their views about this unknown took place once a week, as follows. Each weekend, new data arrived, in the form of the record (wins and losses) for each team against its competition that week (together with other relevant information like how strong the opposition was, whether the games were played on the team’s home court or other courts, and so on), and the writers modified their predictions of who the strongest teams would be at the end of the season accordingly. The process was sequential, with each week’s output (poll) serving as input to the next round of updating when another week’s data came in. Most years, the data provided as the season unfolds serves to decrease uncertainty about who the best team is, but not always—unforeseen developments late in the season, like injuries to key players, can cause this uncertainty to go back up again.

Bayesians like to acknowledge the cumulative, sequential nature of learning by breaking the process down into three steps:

- First, you summarize your prior information about the unknown thing of interest;
- Then, you gather data values relevant to the unknown, and summarize them;
- Finally, you combine the data and the prior information into what people call your posterior information about the unknown—a summary of what you know now that the new data are in (it’s called the “posterior” information because it’s what you know after the data arrive, and the Latin for “after” is a posteriori).

Schematically,

\[
(\text{posterior information}) = (\text{prior information}) + (\text{data information}).
\]

If you need to do this three-step process more than once, your starting point for the second iteration should be your ending point from the first iteration, as it was for the sportswriters—in other words, your posterior information this time becomes your prior information next time.

Well, this all sounds fine, I hope, but it doesn’t advance the ball much to make up all these Latin names for things and write the process down in Three Steps unless you can tell people how to actually do it. The $64 question seems to be what is meant by the verb “combine” in Step 3 above: how exactly should you go from prior to posterior information? That, finally, is where the Rev. Bayes
comes in: his Theorem says how to do this. As we will see, it actually says more than this, but the updating of evidence is its principal application today.

The rest of this section, and the next section as well, are about the content and use of Bayes’ Theorem. I have two examples to keep things from getting too abstract. The first one, the case study we’ll concentrate on for the rest of this section, is from business, specifically banking, and is about decision-making rather than science. It’s an example in which frequentists and Bayesians would agree completely about what evidence is available and how to use it, which demonstrates that Bayes’ Theorem isn’t just for Bayesians. The second case study, which we will take up in Section 5, is a reprise of the coin-tossing example of Chapter 7, and is an instance in which Bayesians and frequentists—strict ones, at least—would disagree completely about the main probability of interest and whether it’s even possible to attach a value to it.

A Straightforward Use of Bayes’ Theorem: Credit-Verification. A big problem facing banks and merchants alike is determining potential customers’ credit-worthiness. You have probably had the experience of trying to use a credit card to pay for something—nowadays store owners have a little electronic machine they pass your card through, and information stored in the magnetic strip on the back (mainly your credit card account number) is transmitted over a telephone line to a central credit-verification system somewhere. This system uses some sort of computer program to decide if your purchase should be approved, based on factors like how often it thinks your card has been used lately and how recently it thinks you have paid your bill. If you pass this invisible screening, a little green light goes on back in the store, and you walk out with your purchase.

Of course, any computer-based system of this kind makes mistakes sometime, because of faulty information or bad programming: sometimes credit cards that are good are declared bad, sometimes vice versa. Events of this type look random to the people at the bank trying to figure out why they happen—at least until the causes of the mistakes are determined—so it makes sense to talk about the probability that a bad credit card is declared good, and the probability that a good card is judged bad. Standard terminology is to call the kind of mistake in which you declare a good credit card bad a false positive, and the other kind of mistake—in which you label a bad credit card good—a false negative. Evidently the people who made up this terminology were thinking of “positive” in this context as equivalent to “calling a credit card bad” and “negative” as amounting to “calling the card good,” which is a little perverse, but there it is. [footnote on medical screening]
People evaluate the quality of credit-screening systems of this type by running tests in which (a) they attempt a number of fake "purchases" with some credit cards that are known to be good and some others that are known to be bad, and (b) they look to see how often the system gets it right. Suppose that in one test of this type on the system we’re going to look at, 97% of the test "purchases" with cards known to be good were labeled good by the system, and 98% of the "purchases" in which the test cards were bad were declared bad by the system. Suppose further that the system is to be used in a market in which about 1% of all attempted purchases are with bad credit cards. Somebody now walks into a store using this credit verification system and tries to make a purchase by credit card. In thinking about how to weigh the evidence supplied by the system about the validity of this card, let’s focus on two questions:

- Since so few people attempt to buy things with bad cards, and since we don’t know anything special about this particular person, I guess you would say before the system renders its verdict that the card is probably good, but how strong is the evidence for this view?
- Now suppose the system comes back with a negative opinion about this person’s credit-worthiness. If the system is any good, this should be strong evidence the other way—that the card is in fact bad, that is—but quantitatively how strong is this new evidence? In other words, how much should our opinion about the card change in light of the new information provided by the system?

Asking the questions in this way makes it sound suspiciously like I’m going to try to fit this setup into the 3-part Bayesian learning story I laid out above, and in fact I am, so I guess the next questions are: what’s the unknown, what’s the prior information, and what’s the data? Since this is your first chance to put your Bayesian hat on, you might want to close the book and answer those questions for yourself before I do so in the next paragraph.

Well, think about the temporal order of the events here: the customer walks in, we wonder about his or her credit-worthiness, then the system weighs in with its opinion, and we have the opportunity to modify our beliefs in light of what the system says. So I guess

- The unknown is this person’s actual credit-worthiness;
- We would regard what we knew about this unknown before the system expressed its opinion as the prior information;
- We would consider the system’s opinion the data; and
• We would think of what we knew about the unknown after the system announced its verdict as the posterior information.

Okay, to make progress in quantifying all of this, let's try to set things up in the language of probability. There seem to be a variety of propositions floating around here, but if you write them down and sort them out they are of only two types: statements about the true status of the card (the unknown)—good (G) or bad (B)—and statements about the system's opinion of the card (the data)—also good or bad, which we might as well denote as negative (-) and positive (+), respectively, in keeping with the terminology above. If we set up two dichotomous variables, one representing truth, the other what the system says, we can arrange the various possibilities in a 2 by 2 probability table like the one in the marijuana legalization example of Chapter 7 (Table 2).

In the marijuana example we had two 2 by 2 tables, one based on counts, the other based on relative frequencies or (frequentist) probabilities, and we saw in that chapter that you can move back and forth freely between them without loss of information, as long as you know the sample size. Here we ultimately want a probability table like Table 2 (except with numbers filled in instead of the symbols, of course), but I find it easiest to motivate what will follow by thinking about a hypothetical batch of credit-card transactions and summarizing them with a table of counts, so we'll look at both kinds of tables in the next few pages. Before going further, recall from Chapter 7 that probabilities like P(G) in the margin of Table 2 summarize one variable at a time, and probabilities like P(G and -) in the body of the table summarize the joint behavior of the two variables.

As I just said, our goal here is to replace all the symbols in Table 2 with numbers, so let's try to translate the values like 97% and 1% I mentioned in the statement of the problem—and the words that went with them—into probabilities. For example, what does the 1% figure mean in probability terms? Well, to put on your frequentist hat from Chapter 7 first (for familiarity's sake), frequentists would view the potential purchase in this example as like a random draw from a large collection of potential purchases in the market served by this credit system, 1% of which are known in the past to have been bad. Not having any reason to treat the past and future as anything other than similar (exchangeable) in this case, and not having any further information about this person or the store at which the purchase is being attempted, frequentists would take all of this to mean that P(the card really is bad) = P(B) = 0.01, and therefore that P(good) = P(G) = 1 - P(bad) = 0.99.

Bayesians would use probability to express their judgment that the credit card is good or bad given the available evidence, without having to imagine
drawing the person in question at random from some pool of purchasers. But
since the only information around (the 1% figure) is of a frequency nature, and
since Bayesians would share the italicized judgments in the last paragraph with
frequentists, everybody would arrive at the same conclusions in this case—P(G)
= 0.99 and P(B) = 0.01. For exactly the same reasons it’s going to turn out that
Bayesians and frequentists would agree on all the rest of the probabilities in this
problem, too, so I will continue to motivate them from the frequentist point of
view (because it’s easier) and we can just remember that Bayesians would get
the same thing. The only differences that will come up are in terminology: for
example, we agreed above that with your Bayesian hat on you would think of P(G)
as the prior probability that the card is good. Frequentists don’t use the terms
“prior” and “posterior”; in fact, mentioning prior information to strict frequentists
is like waving a red flag at them, for reasons I will try to explain in Section 5.

So I guess we’re ready to answer the first question I posed above, about the
strength of initial evidence that the card is good: both camps would say that the
odds in favor of it being good before the system rendered its verdict were P(it’s
good)/P(it’s bad) = P(G)/P(B) = .99/.01 = 99 to 1. This is strong prior evidence
that the card is good.

So much for the 1% figure. Now what about that statement that 97% of the
test “purchases” with cards known to be good were labeled good by the system?
Another way to say this is that among all test “purchases” that were known to be
good, the system said that 97% were good, which makes me think of conditional
probability (remember that “among” is like “given”): P(system says good given
really is good) = P(- given G) = 0.97. Similarly, the statement that 98% of the
“purchases” in which the test cards were bad were declared bad by the system
is another way to say that P(system says bad given really is bad) = P(+ given
B) = 0.98.

We can go a bit further with these two probabilities. If you know that “among
all test ‘purchases’ that were known to be good, the system said that 97% were
good,” it must follow that among those “purchases,” the system said the other
3% were bad—in other words, P(system says bad given really is good) = P(+
given G) = 1 - P( - given G) = 1 - 0.97 = 0.03 (this is just the result we called
the Easy Rule for Conditional Probability back in Chapter 7). By the same reasoning,
P(system says good given really is bad) = P( - given B) = 1 - P( + given B) = 1
- 0.98 = 0.02. In other words, continuing to use the “false positive” and “false
negative” terminology above, the false positive rate of this system is 3% and its
false negative rate is 2%, both of which sound pretty good.

So here is what we know:
<table>
<thead>
<tr>
<th>P(G) = 0.99</th>
<th>P(- given G) = 0.97</th>
<th>P(+ given B) = 0.98</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(B) = 0.01</td>
<td>P(+ given G) = 0.03</td>
<td>P(- given B) = 0.02</td>
</tr>
</tbody>
</table>

What do we want to know? In the second question above I wondered how much our opinion of the card should change if the system announced that it thought the card was bad. In probability terms, once the system renders its verdict we should take what it says as a given—in other words, we should condition on the system’s opinion and compute P(the card is good given that the system says it’s bad) = P(G given +). Since we agreed above that the system’s verdict plays the role of the data in the three-part Bayesian learning story here, Bayesians would call this probability, P(G given +), the posterior probability that the card is good given the data. Evidently the goal here is to work out this posterior probability and compare it to what we knew about the card before we heard the system’s verdict, as summarized by the prior probability P(G) (Figure 2).

One other thing is also evident now:

In probability terms, updating in light of new information just corresponds to conditioning on that information. When data arrives we move from P(unknown) to P(unknown given data).

Okay, it looks like it all comes down to P(G given +), which doesn’t seem to be on the list of things we know yet. In fact, it’s a bit irritating—I know P( + given G), but not P(G given +). Another way to put this is that

I know P(data given unknown) but not P(unknown given data).

This turns out to be typical—as we will see beginning in Chapter 9, it’s usually a lot easier to get your hands on the former than the latter. If I knew how to compute conditional probabilities in the other order—in other words, if I knew how to reverse the order of conditioning—I’d be in business. This is what Reverend Bayes showed the world how to do back in 17xx—if you have two propositions A and B you’re interested in, Bayes’ Theorem tells you how P(B given A) and P(A given B) relate. In this case what we want is to go from P(data given unknown) to P(unknown given data), and it is no accident that we find ourselves searching for a way to do this—it will become clear from Chapter 10 on that this is the central problem in inference and prediction.

Deriving Bayes’ Theorem. The simplest form of Bayes’ Theorem is pretty easy to figure out, as follows. If you want to know how P(B given A) and P(A
given B) relate, why not try writing out the definition for both of these conditional probabilities, and seeing if something good happens?

\[ P(B \text{ given } A) = \frac{P(B \text{ and } A)}{P(A)} \quad \text{and} \quad P(A \text{ given } B) = \frac{P(A \text{ and } B)}{P(B)} . \]

Multiply both sides of the left-hand equation by P(A) and both sides of the right-hand equation by P(B), and switch left- and right-hand sides in both equations:

\[ P(B \text{ and } A) = P(A) \cdot P(B \text{ given } A) \quad \text{and} \]
\[ P(A \text{ and } B) = P(B) \cdot P(A \text{ given } B) . \]

Now recall that P(B and A) = P(A and B), so

\[ P(A) \cdot P(B \text{ given } A) = P(B) \cdot P(A \text{ given } B) . \]

Divide by P(A) to get

\[
\text{The Easy Form of Bayes' Theorem.}
\]

\[ P(B \text{ given } A) = \frac{P(B) \cdot P(A \text{ given } B)}{P(A)} . \]

Well, that was rather easy. (Pretty good deal for Rev. Bayes to figure out something that simple and have his name go down in history—nice work if you can get it.) Of course, there must be a catch—nothing worthwhile is ever this easy. And there is: let’s try to apply this to the credit verification example.

\[ P(G \text{ given } +) = \frac{P(G) \cdot P(+ \text{ given } G)}{P(+)} . \]

The catch is that, while we know the two probabilities in the numerator, we don’t know the one in the denominator. It turns out that there are a number of forms of Bayes’ Theorem, dealing with situations a lot more complicated than this one, and computing the denominator is a constant source of annoyance in using the Theorem in almost all of its versions. The general fix for this problem involves another theorem called the Law of Total Probability that we’ll look at in Section 6. Right now I’m going to go through two special-case fixes that are adequate for 2 by 2 situations like the credit card example and that give a good flavor of how things go more generally. The first one is quite direct and involves completing the
probability table I introduced a few pages ago; the second one uses a neat trick based on expressing Bayes’ Theorem in terms of odds rather than probabilities.

Before I forget, though, to get a little more of the big picture let’s write down the Easy Form of Bayes’ Theorem in its main application in statistics, in which A stands for the data and B for the unknown:

\[
P(\text{unknown given data}) = \frac{P(\text{unknown}) \cdot P(\text{data given unknown})}{P(\text{data})}.
\]

Two of the quantities in this equation are already familiar—\(P(\text{unknown})\) is the prior probability and \(P(\text{unknown given data})\) the posterior probability of interest. The second term in the numerator, \(P(\text{data given unknown})\), is called the likelihood, because it measures how likely the observed data are given the unknown; if that sounds a little mysterious, that’s okay—we’ll look at this term in the equation a lot more later. It turns out that the denominator can be viewed as present in the equation only in order that everything adds up to 1, like probabilities are supposed to. Math people refer to quantities of this type as “normalizing constants,” and I have already mentioned that the denominator is an annoyance when trying to apply the Theorem, so putting this all together you get

\[
\text{posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{annoying normalizing constant}}. \quad (1)
\]

This is probably the key equation in the whole Bayesian story—it says, at a glance, how to update your information about an unknown when new data come along. We will be applying it a good deal more in the parts of the book on inference and prediction.

[small-print] To finish off this little big-picture interlude (you can skip the small-print if you’re not mathy), I said toward the end of the basketball poll example that schematically

\[
(\text{posterior information}) = (\text{prior information}) + (\text{data information}) \quad (2)
\]

If we identify the likelihood as the place where the information contained in the data resides (which we will see later on makes sense), equations (1) and (2) look
similar except that one of them involves multiplication and the other addition. But I can fix this by recalling that there's a function—logarithms—that converts multiplication into addition. Taking the logarithm of both sides of equation (1) and ignoring the annoying normalizing constant (trust me) shows that equation (2) was more than schematic:

\[(\log \text{posterior}) = (\log \text{prior}) + (\log \text{likelihood}).\]

Evidently it's on the logarithmic scale that information accumulates additively. We will take this point up again much later. [end small-print]

**Computing the Denominator in Bayes' Theorem Directly.** Okay, let's look again at the probability table for the credit card example (Table 2). I'll fill in the margins we know, P(G) and P(B), the prior probabilities, to get Table 3. Recall I'm trying to work out P(G given +), which by Bayes' Theorem is given by

\[
P(G \text{ given } +) = \frac{P(G) \cdot P(+ \text{ given } G)}{P(+)} = \frac{0.99 \cdot 0.03}{P(+)}.
\]

I can't fill in any of the other values in the table directly, because the only other probabilities I know are conditional ones like P(+ given G). But if I could only work out P(G and +) and P(B and +) and add them, I'd have that stupid denominator P(+) and I'd be done.

Fortunately, it's actually not too bad to work out P(G and +) from what we know, as follows. Pretend that the credit verification system is applied to a large number of transactions, say 10000, of which 99%, or 9900, are good and the other 1%, or 100, are bad. The information that P(+ given G) = 0.03 means that 3% of the 9900 good transactions will be (falsely) declared bad, and 3% of 9900 is 297. So I guess 297 of the 9900 good transactions would fall in the + row of the table, and the remaining 9900 - 297 = 9603 would land in the - row (Table 4).

By the same reasoning, the information that P(+ given B) = 0.98 means that 98% of the 100 bad transactions will be correctly labeled bad by the system, and 98% of 100 is 98. That leaves 2 of the 100 bad cards to be (falsely) labeled good, and I can fill in the rest of the table, including (finally) the row margins needed to finish applying Bayes' Theorem (Table 5). The Table shows that P(+) = 395/10000 = 0.0395, so at last

\[
P(G \text{ given } +) = \frac{0.99 \cdot 0.03}{0.0395} = \frac{297}{395} = 75\% \text{ (!).}
\]

Now that the table is all filled in, there is an even more straightforward way to get what we want—just read P(G given +) directly out of the second row: among
the 395 transactions the system labeled bad, 297 were in fact good, so \( P(G \text{ given } +) = 297/395 \).

This is an amazing result—it says that 3/4 of the transactions the system claims are bad are in fact good! Nobody would guess that a system with such good false negative and false positive rates would have such bad performance, but there it is. In problem 8.x I encourage you to explore this phenomenon a bit; it turns out that the main trouble here is that bad credit cards are so rare (only 1% of the transactions are bad) that in order to catch almost all of the bad cards, the system has to label a lot of good cards bad. It is interesting to think about this setup from the points of view of three different actors in the drama: the customer, the store, and the bank. What are the costs and benefits to each actor of declaring a lot of good credit cards bad in order to correctly spot most of the bad ones? [finish this]

**A Shortcut: Bayes’ Theorem in Terms of Odds.** Here’s a shortcut that gets around a lot of the hassle we just went through. The problem all along has been that damn denominator in the Easy Form of Bayes’ Theorem:

\[
P(B \text{ given } A) = \frac{P(B) \cdot P(A \text{ given } B)}{P(A)}.
\]  

(3)

It would sure be nice to come up with a trick to get around this problem, based on the lazy-person’s idea that anything too hard to do should be avoided. The secret turns out to be to apply the Easy Form of the Theorem to not \( B \) as well as to \( B \):

\[
P(\text{not } B \text{ given } A) = \frac{P(\text{not } B) \cdot P(A \text{ given } B)}{P(A)}.
\]  

(4)

If you now divide equation (3) by equation (4), the source of our irritation, \( P(A) \), cancels, and we never have to compute it directly:

\[
\frac{P(B \text{ given } A)}{P(\text{not } B \text{ given } A)} = \frac{P(B)}{P(\text{not } B)} \cdot \frac{P(A \text{ given } B)}{P(A \text{ given } B)}.
\]  

(5)

Moreover, not only does this dodge some hard work, it also has a terrific interpretation when we try to apply it to an important special case of the usual Bayesian updating story involving the unknown \( B \) and the data \( A \), as follows. Let’s say the unknown is in the form of a proposition, or *hypothesis*—\( H \), say—that’s either true or false, as in the credit card example: the card really is either good or bad. You might say in this case that the unknown is *dichotomous*. Then equation (5) (with the symbol \( l \) in place of the word “given”) becomes

\[
\frac{P(H \text{ true}|\text{data})}{P(H \text{ false}|\text{data})} = \frac{P(H \text{ true})}{P(H \text{ false})} \cdot \frac{P(\text{data}|H \text{ true})}{P(\text{data}|H \text{ false})}.
\]
The first ratio on the right, \( \frac{P(H \text{ true})}{P(H \text{ false})} \), is recognizable as the prior evidence that \( H \) is true, expressed in terms of odds, which we might call the prior odds for \( H \), and by the same token the ratio on the left side of the equals sign is the posterior odds for \( H \) given the data. That leaves the second ratio on the right, \( \frac{P(\text{data given } H \text{ true})}{P(\text{data given } H \text{ false})} \). The numerator of this ratio is what a few pages ago I called a likelihood—in this case, it measures how likely it was to get the data you got if \( H \) were true—and the denominator is also a likelihood, in this case measuring how likely the data were if \( H \) had been false. For this reason some people call the second ratio on the right the likelihood ratio, but I'm going to use the other name people give to it: the Bayes factor in favor of \( H \) given the data. By whatever name, it represents the relative likelihood of getting the data you got under each of the two stories—\( H \) true, \( H \) false—whose plausibilities you're trying to compare.

Putting all these names together gives

\[
\frac{P(H \text{ true} | \text{data})}{P(H \text{ false} | \text{data})} = \frac{P(H \text{ true})}{P(H \text{ false})} \cdot \frac{P(\text{data} | H \text{ true})}{P(\text{data} | H \text{ false})}.
\]

Or, in words,

\[
\left( \frac{\text{posterior odds}}{\text{in favor of } H, \text{ given the data}} \right) = \left( \frac{\text{prior odds}}{\text{in favor of } H} \right) \cdot \left( \frac{\text{Bayes factor}}{\text{in favor of } H, \text{ given the data}} \right).
\]

This is just about my favorite equation in the whole book. In one simple stroke it tells you how to revise your evidence about a dichotomous unknown when new information arrives: assess the prior information in odds form, compute the Bayes factor by weighing the relative likelihood of the data under each of the two stories you're comparing, and multiply to get the posterior odds. You don't come across something that's simultaneously as simple and as powerful as this too often.

Once you've used this equation to produce the posterior odds, say \( o = \frac{s}{t} \), you can convert them as usual into the posterior probability \( p \) that \( H \) is true given the data with the relations \( p = \frac{o}{1+o} = \frac{s}{s+t} \); we worked out back in Section 1.

Turning this equation loose on the credit card example is a breeze. We worked out early on that the prior odds in favor of the card being good were 99 to 1.
The Bayes factor in favor of this hypothesis, given the data that the system says the card is bad (+), is

\[
\text{Bayes factor} = \frac{P(+) \text{ given } G}{P(+) \text{ given } B} = \frac{0.03}{0.98} = 3 \text{ to } 98.
\]

In other words, the data evidence is 98 to 3, or about 33 to 1, against the card being good. So the posterior odds are

\[
\left( \frac{\text{posterior odds that the credit card is good given that the system says it's bad}}{1} \right) = \left( \frac{99}{1} \right) \cdot \left( \frac{3}{98} \right) = \frac{297}{98} = \text{roughly } 3 \text{ to } 1,
\]

and the posterior probability that the card is good given the data is \( \frac{297}{(297 + 98)} = 297/395 \), or about 75%. As advertised, we computed this using only the information available in the original problem statement—that stupid denominator \( P(+) \) that gave us so much trouble earlier never came up.

5. A Bayesian Example

As this chapter's last example of how Bayesians think about probability, I want to finish that little food-for-thought teaser on coin-tossing I left you with toward the end of Chapter 7. The setup was that some guy offers to gamble with you over the tossing of a coin—his coin, with him tossing it, and he won't show it to you. You're suspicious (if you're not, I have a bridge I'd like to sell you), so you ask him to toss it a few times, say 5. He does so, and gets 5 heads in a row. "Wow, what a coincidence," he says. "Looks like tails is long overdue. But I tell you what, I'm a sporting guy, I'll even give you the advantage—how about $10 on the next toss, tails you win, heads I win?"

Well, that makes it a little confusing. Let's see if we can use Bayes' Theorem to sort it out. Since we have only seen one example of the Theorem in action, it would probably be good to review the logic we went through in the credit verification example.

By now the Theorem's ingredients should have started to come into focus—to begin with, I need to identify the unknown and the data, and specify my prior knowledge about the unknown. What I know is that 5 heads came up in a row when the guy tossed the coin, so that must be the data, and what I don't know is the truth about this damn coin and how he tosses it.

At this point, something comes up that didn't arise in the credit verification example. In that example, once we identified the unknown as the truth about the
credit card wielded by the person who came into the store, it was clear that there were only two possible values for this unknown—the card was either good or bad. But in this case I can think of at least three possible explanations for the data:

1. Maybe the coin is fair, and 5 heads in a row just happened to turn up; or
2. Maybe it’s two-headed; or
3. Maybe it’s a regular coin but the guy is a magician who can make it land however he wants (there are people who can do this).

[finish this]

6. The Law of Total Probability

7. Summary

- To a frequentist, a probability is a relatively objective expression of how often repeatable things happen, and it is not possible to attach probabilities to things that are not repeatable.
- To a Bayesian, a probability is a subjective expression of judgment or belief about the strength of evidence for or against a given proposition, and everything is fair game.
- A good Bayesian will try to use the standard rules of probability to make his or her judgments, expressed in probability terms, reasonably coherent, and will try to keep track of how often he or she was right and modify his/her judgments accordingly in order to remain reasonably well calibrated, but these goals are not necessarily attained by all Bayesians on all occasions.
- If the only evidence available is of a relative frequency nature, then frequentists and Bayesians will agree, either exactly or pretty closely, on the probability of interest.
- In other situations, when your question is “What is the probability of so-and-so?” and “so-and-so” is not a repeatable event, a frequentist is liable to say that the question has no good answer; a Bayesian will always be able to give you an answer, at least in principle, but (in view of coherence and calibration) there is no guarantee that the answer is a good one. What I mean by “in principle” is that Bayesian probabilities can be quite hard to compute.
- Calibration requires feedback—you need to say what you think will happen in the future on a number of occasions and see how often you were right—but no feedback is possible on some propositions, like P(life on other planets), so the goal of calibration is not always attainable. [finish this; frequentists
find Bayes' Theorem just as valid as Bayesians do; they just disagree with Bayesians on how often it can be applied to good effect (where does the prior come from, etc.?)

8. Problems

1. In the little artificial population [1, 2, 7.5] we looked at in Chapter 7, when sampling 2 members of the population without replacement, use conditional probability to show that \( P(1 \text{ on the second draw}) = P(2 \text{ on the second draw}) = P(7.5 \text{ on the second draw}) = 1/3 \). (Hint: \( P(1 \text{ on the second draw}) = P(1 \text{ on the second draw} | 1 \text{ on the first draw}) \times P(1 \text{ on the first draw}) + ... \); finish this.) There is nothing special about this choice of population and sample size; it is a general fact that although the draws when sampling without replacement are not independent, they are identically distributed.

2. Food for thought: Does it make sense to say that you think the probability of something is “about 0.6,” say? If so: I thought probability was supposed to be measuring your uncertainty about the proposition in question—does it make sense to be uncertain about how much uncertainty you have? [infinite regress, etc.]

3. In an article about the preferential treatment Harvard University accords the children of its alumni in its admissions process, two sociologists writing on the editorial page in the New York Times on December 8, 1990 reported that 280 of 1602 Harvard freshmen had parents who attended Harvard and that 40% of alumni children are admitted each year as against 14% of non-alumni children. With \( S = \{ \text{successfully admitted} \} \) and \( A = \{ \text{alumni parents} \} \), work out the following probabilities: \( P(A), P(S), P(A \text{ given } S), P(A \text{ given not } S), P(S \text{ given } A), P(S \text{ given not } A) \). How much does the information that a given child’s mother or father went to Harvard change the odds that he or she will be admitted? etc. [spiff this up; easiest to use Bayes’ Theorem in odds form then work out the 2 by 2 table etc. Harder: if all we knew was “280 of 1602 freshmen had fathers who had attended Harvard” and preference is given to children of either Harvard fathers or Harvard mothers or both, how far can you get? (children applying in 1988 were born in 1970; their parents were born in, say, 1945 or earlier and went to Harvard no later than 1963, say: female admission rate then?)]

4. The law of restricted choice in bridge.

5. What assumption(s) are required to legitimately convert the observation that e.g. 30% of the time in the past so-and-so has happened to \( P(\text{so-and-so will happen in the future}) = 0.3 \)? [rewrite?; belongs in chapter 7?]
Tables for Chapter 8

Table 1: The results of the AP college basketball poll of December xx, 19xx (unavailable; sorry).

<table>
<thead>
<tr>
<th>The Truth (unknown)</th>
<th>Card is</th>
<th>Card is</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good (G)</td>
<td>Bad (B)</td>
<td></td>
</tr>
<tr>
<td>+--------------------</td>
<td>----------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>Card is Good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is Good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The System's Opinion Card is Good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is Bad</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P(G)</td>
<td>P(B)</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 2: 2 by 2 probability table relating the true status of the credit card to the verification system's opinion of the card.

<table>
<thead>
<tr>
<th>The Truth (unknown)</th>
<th>Card is</th>
<th>Card is</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good (G)</td>
<td>Bad (B)</td>
<td></td>
</tr>
<tr>
<td>+--------------------</td>
<td>----------</td>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>Card is Good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is Good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(-)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The System's Opinion Card is Good</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is Bad</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(+)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.99</td>
<td>0.01</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 3: 2 by 2 probability table relating the true status of the credit card to the verification system's opinion of the card, with the column margins filled in.
The Truth

<table>
<thead>
<tr>
<th>Card is Good (G)</th>
<th>Card is Bad (B)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card is Good</td>
<td>9603</td>
<td>?</td>
</tr>
<tr>
<td>Card is Bad (+)</td>
<td>297</td>
<td>?</td>
</tr>
<tr>
<td>Total</td>
<td>9900</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4: Applying the credit verification system to a mythical batch of 10,000 transactions, to work out $P(\text{+})$ directly. So far the only information used is $P(G) = 0.99$ and $P(\text{+ given } G) = 0.03$.

The Truth

<table>
<thead>
<tr>
<th>Card is Good (G)</th>
<th>Card is Bad (B)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card is Good</td>
<td>9603</td>
<td>2</td>
</tr>
<tr>
<td>Card is Bad (+)</td>
<td>297</td>
<td>98</td>
</tr>
<tr>
<td>Total</td>
<td>9900</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5: Applying the credit verification system to a mythical batch of 10,000 transactions, to work out $P(\text{+})$ directly. Now the information that $P(\text{+ given } B) = 0.98$ is used to fill in the second column, and this completes the table.
Figures for Chapter 8

<p>| all judgment,  |</p>
<table>
<thead>
<tr>
<th>no data</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;all data,</td>
</tr>
<tr>
<td>no judgment&quot;</td>
</tr>
</tbody>
</table>

Figure 1: A continuum of types of evidence, from 100% judgment on the left to "100% data" on the right.

<table>
<thead>
<tr>
<th>What we thought before the system rendered its verdict (before the data came in, that is)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
</tr>
<tr>
<td>P(unknown)</td>
</tr>
<tr>
<td>P(good) = P(G) = 0.99 -- in other words, 99 to 1 odds in favor of the card being good</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What we think after the system renders its verdict (after the data arrives)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posterior</td>
</tr>
<tr>
<td>P(unknown given data)</td>
</tr>
<tr>
<td>P(good given system says bad) = P(G given +) = ?</td>
</tr>
</tbody>
</table>

Figure 2: Updating the evidence about the card in light of the system's verdict.
Chapter 9

Expected Value, Standard Error, and the Central Limit Theorem

This chapter is about how to build probability models that help answer questions about random events, in science and daily life, that are hard to answer with the methods of Chapters 7 and 8. As the case study that will help indicate how to do this, I want to analyze a gambling game called Keno that people play in places like Las Vegas and Atlantic City. As we saw back in Chapter 7, gambling is a fitting choice historically for learning about probability, and if you have ever thought about trying your hand at the casinos with the goal of winning money I hope this chapter will dissuade you—it will become clear that Keno, like virtually all casino gambling games, is a losing proposition.

Case Study 9: Keno. In the casino game of Keno there is a big cylinder with balls numbered from 1 to 80 in it, and every now and then throughout the day somebody shakes up the balls in the cylinder and draws out 20 of them at random without replacement. For awhile before this drawing takes place, other people have been going around the casino offering the gamblers a chance to bet on which balls will be chosen.
Two available bets in Keno are a *single number* and a *double number*. With the single number bet you pick a number that you think is going to be among those chosen, circle it on one of the *Keno cards* that the casino management has thoughtfully placed near you, and give the Keno person your wager. If you bet $1 on a single number and it turns out to be one of the 20 chosen, the casino gives you your dollar back plus $2 in winnings—in casino language, the payoff is 2 to 1. If your number is not chosen, they keep your dollar and encourage you to try again: "Maybe your luck will be better next time." A double number is just like a single number except that you circle two numbers on your Keno card, and if they are both among the 20 numbers chosen the casino pays you 11 to 1.

Is Keno a good game to play? Well, it depends on what you mean by "good"—for example, if it gives you pleasure to play it, that ought to count for something (this is an example of a decision-making concept called *utility*, which we will talk about in Chapter xx). Right now I want to focus on a different issue: do you think you will win if you play? Notice that if you were to play Keno several times, we are uncertain how each play will come out—sometimes you will win, sometimes you will lose—but notice also that the "at random without replacement" method of picking the winning numbers means that we ought to be able to use probability to quantify our uncertainty about the outcomes. So this makes me interested in the answers to questions like the following:

- What is your chance of winning on any one play?
- If you bet $1 each time, how much money do you expect to win or lose on any single play, on average? Give or take how much?
- If you had the patience to play a large number of times, say 100, betting $1 each time, how much money would you expect to have won or lost after the 100 bets? Give
or take how much? What is your chance of coming out ahead after 100 plays?

The answers to these questions may be different for the single- and double-number betting strategies. I’ll lay out the story for single-number gambling first, and then use the machinery to address the same questions if you were to bet on double numbers instead. The argument is rather lengthy, and mathier than most of the rest of the book, but hanging on all the way through this chapter has a big payoff—the model presented here is the cornerstone for all the rest of the work that follows.

9.1 Building a Probability Model

Figure 1 is where we’re headed: it’s a probability model for single-number $1 betting at Keno. Probability models turn out to have three main ingredients, corresponding to the left, middle, and right parts of the figure: the population, the sample, and the imaginary dataset. Each of these elements of the model can be thought of as a dataset, with (as usual) rows for replicates and columns for variables. In this part of the course we will only be looking at one variable at a time, so for now each of these datasets will have only one column, but toward the end of the book we will look at situations with two or more variables. The main setup I’m going to describe here is frequentist in character, but we will see later that it’s useful in Bayesian calculations too.

Diagrams like Figure 1 looks complicated, and when you first start to do them they take awhile to finish, but I recommend writing them out in full whenever you have the patience to do so. They enforce a useful discipline in reminding us of all of the judgment calls we need to make when trying to model the real world in probability terms, and—if you do enough problems like the ones in this chapter—eventually a problem will come along in which failure to attend to all the ingredients in the diagram will cause your model to be unrealistic in some important way. The terminology imaginary dataset used here is nonstandard; the official description of it would involve words like “simulating the distribution of a random variable,” but I like “imaginary dataset” better. I have put some more information about random variables in the math interludes below, if you get interested.
9.1.1 The Population

Basic to the construction of these models is the idea that you are studying something that could come out differently each time you do it, and you are imagining doing it over and over. The population dataset records how things might come out, as measured by some variable that’s particularly interesting to you, each time you repeat the process you’re imagining repeating. In single-number wagering at Keno, for instance, we are imagining betting $1 on your favorite number over and over, and the variable we would like to keep track of is how much money changes hands on each play—what we might call “your net gain” on any single play, if we considered it from your point of view rather than from the point of view of the casino. Sometimes it’s convenient to have a symbol for the number of rows in the population dataset—let’s call it $N$.

To finish the population dataset I guess I should put some data in it. There are two things to pay attention to: the possible values the variable of interest could take on, and the probabilities of each of those possible values actually occurring. Here I said I was interested in how much money might change hands on any single play, from your point of view. With the single number strategy if you bet $1 on your favorite number and it loses, you lose your dollar—in other words, your net gain is $-1$. If your number is chosen you win $2$—in other words, your net gain is $+2$.

So far so good. Now what are the chances associated with each of these outcomes? If there are 80 numbers and the Keno person only picks 20 of them, the chance of your favorite number being in there somewhere doesn’t sound very high. It’s actually not too hard to work out this probability, so let’s take a minute and do so.

Drawing 20 numbers at random without replacement from the list 1, 2, \ldots, 80 is like writing all the numbers down on cards, one number per card, shuffling the deck of 80 cards thoroughly, so that all possible rearrangements are equally likely, and then dealing out the top 20 cards. Consider the deck of cards just before you deal out the top 20, from the point of view of your lucky number—the single number you want to bet on in Keno. There are 80 different positions this number could occupy in the deck—for example, on top, second from the top, third from the top, and so on down to the bottom—and because you
shuffled thoroughly they’re all equally likely. Twenty of them are favorable to you winning the $1 bet—any of the first 20 positions will get your number chosen—so by the equally-likely model the chance of you winning must be $\frac{20}{50} = \frac{1}{3} = 25\%$, and your chance of losing must be $\frac{3}{4} = 75\%$. We have answered the first question in the case study: your chance of winning on any one play is 25%. Single-number betting in Keno is not a very good game when viewed in this way—you probably would have wanted the odds a little closer to even money than that.

We’re not quite done with this part of the model: we still have to set up the population dataset to reflect the fact that you have 1 chance in 4 of winning $2 and 3 chances in 4 of losing $1. The simplest way to do this is to put four numbers in the population: three $-1$’s and one $+2$. (It would be equally good for what we’re going to do to put six $-1$’s and two $+2$’s in there, or any other choice that gives three times as many $-1$’s as $+2$’s, but three $-1$’s and one $+2$ is easiest.) Notice in the left-hand part of Figure 1 that it is helpful to other people trying to understand your model to put a box over the population dataset and fill it with words describing the population. In this case the population is hypothetical, or conceptual—it represents what might happen each time you make one $\$1 bet on a single number. In the statistical models we will build in Chapter xx, which are closely related to the probability models of this chapter, the population dataset will often be less hypothetical than it is here.

The last thing to specify about the population dataset is the same kind of summaries for it that we have been using to summarize a column of numbers ever since Chapter 3—center, spread, and histogram shape. In probability models like those we will look at, people always use the mean and SD to keep track of center and spread (we will see why a little later), and the best way to convey the histogram shape is to actually sketch it. So you finish off the population part of the probability model by writing the mean and SD of the numbers in the population right below the dataset itself (as in Figure 1), and below that you draw the histogram. In this case the histogram is discrete, with 3/4 of its area in a bar at $-1$ and the other 1/4 in a bar at $+2$.

It will turn out below that other parts of the model diagram will also have means and SDs, so we need some special notation for the population mean and SD to keep from getting confused. By convention
everybody gives Greek letters to population summaries, usually chosen to remind you of the first letter of the thing they summarize, so the population mean is usually called \( \mu \) ("mu," pronounced "mew") and the population SD is typically called \( \sigma \) ("sigma").

**Working Out the Population Mean and SD.** How do you calculate \( \mu \) and \( \sigma \)? It's pretty easy—you just think of the numbers in the population as data, and figure out their mean and SD, much as we did in Chapter 3. The mean is particularly easy:

\[
\mu = \frac{3(-\$1) + 1(+$2)}{4} = \frac{-3 + 2}{4} = -\$0.25. \tag{9.1}
\]

This answers part of the second question in the case study: if you played $1 single-number Keno a lot, in the long run 75% of the time you'd lose a buck and 25% of the time you'd win $2, which averages out to losing a quarter each time you play. People summarize this by saying that the **expected value** of your net gain on any given $1 play at single-number Keno is \( \mu = -\$0.25 \). You could think of this as the entertainment value of $1 wagered in this way—if you don’t get 25¢ worth of fun out of watching yourself lose $1 most of the time, maybe you shouldn’t play this game. We shouldn’t be surprised that this expected value came out negative—after all, the casinos are in business to make a profit, not to let you make one at their expense. If they had been sporting about it they could have set the payoff so that the expected value was $0—people refer to games with $0 expected values as **fair**—but that would have left them with a 50/50 chance of losing money on every gambler who walked in the door, which would be a pretty uncertain business proposition (see Problem 1).

**Math Interlude.** Before we go on it's useful to rewrite the calculation for the mean in a different, more symbolic way. Let's let \( a_1, a_2, \ldots, a_k \) stand for the possible values for the outcome of any single $1 play and \( p_1, p_2, \ldots, p_k \) stand for the probabilities of those values occurring, where \( k \) is how many such values there are. Here \( k = 2 \) and the two possibilities are \( a_1 = -\$1 \) and \( a_2 = +\$2 \), with probabilities \( p_1 = \frac{3}{4} \) and \( p_2 = \frac{1}{4} \). Then we could rearrange equation (9.1) above to read \( \mu = \frac{3(-\$1) + 1(+$2)}{4} = \)
\[ \frac{3}{4}(-81) + \frac{1}{4}(+32) = p_1a_1 + p_2a_2 \] and write the population mean \( \mu \) symbolically as

\[ \mu = p_1a_1 + p_2a_2 + \ldots + p_k a_k = \sum_{i=1}^{k} p_i a_i. \]  

(9.2)

You can see from this that \( \mu \) is a kind of weighted average of the possible values for each \$1 gamble, with the weights given by the probabilities of those values occurring.

What about the SD \( \sigma \)? The story is similar to calculating an SD with data you are thinking of as a sample, as we did in Chapter 3:

\[ \sigma = \sqrt{\frac{3[-81 - (-80.25)]^2 + 1[+32 - (-80.25)]^2}{4}} = \sqrt{\frac{6.75}{4}} = \$1.30. \]  

(9.3)

This should seem familiar, except that with samples of data in Chapter 3 we used to divide not by (the number of values) but (the number of values \(- 1\)). That's the only difference between sample and population SDs—when it's the population SD people are after, they divide by \( N \), not \( (N - 1) \). One informal way to understand why is to recall the idea of degrees of freedom from Chapter 3: when you're working with a sample of data and you don't know the population mean, you need to use the data to guess at it in calculating the SD, which "burns up one degree of freedom," so that instead of dividing by the number \( n \) of data values in calculating the sample SD you divide by \( n - 1 \); but in figuring out the population SD the population mean is known, and there is no need to lose a degree of freedom in guessing at it.

Math Interlude. A symbolic expression for \( \sigma \) is also useful, and is easy to get by rewriting equation (9.3) above as

\[ \sigma = \sqrt{\frac{3}{4}[-81 - (-80.25)]^2 + \frac{1}{4}[+32 - (-80.25)]^2}, \]  

(9.4)

which in terms of the \( a \)'s and \( p \)'s and \( \mu \) is just

\[ \sigma = \sqrt{p_1(a_1 - \mu)^2 + \ldots + p_k(a_k - \mu)^2} = \sqrt{\sum_{i=1}^{k} p_i(a_i - \mu)^2}. \]  

(9.5)
So $\sigma^2$ is also a kind of weighted average, in this case of the squared deviations of the possible values from the mean $\mu$. In parallel with the terminology of Chapter 3 people call $\sigma^2$ the population variance since it's the square of the SD.

Equation (9.3) answers the rest of the second question in the case study: Given that we have already worked out that $\mu = -$0.25, the interpretation of $\sigma = 1.30$ is that each time you wager $1 on a single number at Keno you expect to lose about a quarter, give or take about $1.30. This only roughly describes what will actually happen on any given play in this case, since the only possible values for your net gain are $-1$ and $+2$, but even with only two possible values $\mu$ and $\sigma$ give you some idea of what to expect on each play, and we will see in Chapter 10 that give-or-take statements like this become more and more accurate as the number of possible values in the population grows.

9.1.2 The Sample

Okay, so much for the population part of the probability model, which helps you visualize what might happen if you made a single $1 gamble at Keno. The next part of the model is the sample, which helps simulate a bunch of such $1 plays at a time. The idea is that the sample represents one possible realization of the overall random process you're studying. Here, for instance, I want to simulate making 100 $1 plays on a single number, so I'll let the first number in the sample stand for the outcome of the first $1 play, the second number for the outcome of the second play, and so on up to 100 plays. This means that the sample can be thought of as a dataset with one column/variable (representing your net gain on each play) and 100 rows/replications, so that the sample size is $n = 100$. The middle part of Figure 1 shows the sample part of the model diagram—I have made up some hypothetical outcomes of each $1 bet, about $3/4 of which should be $1 and $1/4 of which should come out $2.

The next thing to specify about the sample dataset is its relationship to the population—in other words, to be specific about the sampling mechanism. In Chapter 6 we talked about two kinds of samples—probability samples, and samples of convenience. The random way
in which the balls are drawn at Keno makes the sample in Figure 1 a probability sample, which (you will remember from Chapter 6) is fortunate for us because it's much harder to draw valid conclusions from samples of convenience than it is from probability samples. But what kind of probability sample corresponds to repeated $1 betting at Keno?

In Chapter 7 we looked at the two simplest probability sampling mechanisms, in which the draws from the population are either made

- At random with replacement (IID sampling, we called it, which stood for independent identically distributed draws), or

- At random without replacement (which we called simple random sampling, or SRS).

Does either of those mechanisms describe repeated betting at Keno? Well, it can't be SRS, because if you made the draws without replacement and the population only had 4 elements in it (as in Figure 1) you'd run out a long time before you got to 100 draws, so maybe it's IID. Let's see: are the individual $1 plays at Keno independent of each other? Sure—each time they draw the 20 balls, they put them back in and mix them up thoroughly before the next game, so that what happens in one Keno game can't affect other games. And are the amounts you win or lose on the individual $1 plays identically distributed, by which I mean that they have the same possible values and probabilities of taking on those values? Again, yes—each Keno game is played out, in the sense of probability, in exactly the same way. So I guess the sampling mechanism in repeated betting at Keno is IID. It's useful to add this information to the model diagram by drawing an arrow from the population to the sample, with the notation "IID" above the arrow to explicitly show anybody looking at the diagram how the sample relates to the population. We will see in later chapters that important ingredients in the process of computing things like the probability of coming out ahead in Keno depend on the sampling mechanism.

Math Interlude. Sometimes it's useful to have symbols that stand for the individual draws from the population and the values they take on. The standard symbol for outcome variables
is \( Y \), and that's as good as any other choice here, but to keep from getting confused we'd better add a subscript for which draw we're thinking of, so let's let \( Y_1 \) stand for the first draw, \( Y_2 \) for the second, and so on down to \( Y_{100} = Y_n \) for the last draw, with \( Y_i \) standing for the generic \( i \)th draw. When I say that \( Y_2 \) stands for the second draw I mean that quite literally: \( Y_2 \) represents the process of making the second draw at random with replacement from the population in Figure 1. Formally \( Y_1, \ldots, Y_n \) are called random variables, which is a pretty good name for them: if you were to repeat each draw over and over you would notice that the outcomes vary randomly.

When people are being especially careful they draw a notational distinction between random variables and the values they take on, by letting \( Y_i \) stand for the process of drawing for the \( i \)th time from the population and \( y_i \) for the outcome you actually get when you make that draw. For example, with the simulated data I presented in the middle part of Figure 1 the second Keno game resulted in the loss of a dollar, so on that occasion \( y_2 \) came out \(-\$1\). With this notation you can talk about the probability that \( Y_2 \) is \(-\$1\)—symbolically, \( P(Y_2 = -\$1) \)—but it doesn't make sense to talk about the probability that \( y_2 \) is \(-\$1\): it either is or it isn't.

If you go on to take another probability or statistics course, the book you use might well summarize a setup like the population-and-sampling story for 100 single-number \$1 Keno bets with notation something like this:

\[
Y_i \IID, \ i = 1, \ldots, n = 100, \quad P(Y_i = y_i) = \begin{cases} 
\frac{3}{4} & \text{if } y_i = -\$1, \\
\frac{1}{4} & \text{if } y_i = +\$2, \\
0 & \text{otherwise.}
\end{cases}
\]

In this book we will use model diagrams like Figure 1 instead of this notation, but the diagram and this equation stand for the same set of assumptions about how Keno works.

The idea of random variables is slightly strange at first; don't be dismayed if it takes you awhile to get the hang of them. As I said above, they stand for the process of doing something "at random," like playing Keno or some other gambling game, and—since we're not actually playing Keno, we're just imagining what
might happen if you were to play—they’re hypothetical, with no objective reality. In other words, random variables are a mathematical fiction, but they turn out to be a useful fiction. Data—what you get in the sample part of Figure 1 when you actually play Keno, or at least simulate doing so—is the only part of the story told by the figure that has any objective reality. You will never see a random variable, but—if you work in a quantitative field—you will routinely see data, and you will often find it useful to think of your data as like the outcome of a random process. To express this sort of thing you hear people speaking technically make statements like “I am thinking of my data as realizations of IID random variables,” which would probably have sounded incomprehensible before you read this, but all that people mean by such a statement is that they have a diagram like Figure 1 in mind.

Next in specifying the sample part of the diagram is how to summarize the sample dataset. Here I’m not so much interested in the 100 actual outcomes as I am in what they imply about how much money I’ve won or lost overall, so I’d like to summarize the 100 data values with a number that keeps track of my net gain or loss at the end of the 100 gambles. If you think about it a minute, you will see that the sum of the 100 draws is just what we’re looking for. (For instance, if there were only two $1 plays and you lost both times, the sample values would both be −$1 and their sum would be −$2, which is just another way of saying correctly that you would be behind $2.) We’re not sure how the sum will come out—each time you had the patience to play 100 times the sum would very likely be different—so let’s just give it a symbol, $S$, which it’s useful to write at the bottom of the sample dataset as in Figure 1. Symbolically, in terms of $S$, answering the last question in the case study about your chance of coming out ahead amounts to computing $P(S > 0)$.

The last ingredient in the sample part of the probability model is a histogram of the 100 sampled values, which can be sketched below the sample as in the middle part of Figure 1. Since we’re uncertain about how each $1 gamble will come out, the histogram is uncertain, too, but it helps to think about how it would probably look. Well, each draw has a 75% chance to be a −$1 and a 25% chance to be a
+2, so in the 100 draws there should be around three −$1's for every +$2—in other words, the sample histogram should look a lot like the population histogram. That's another way to say that IID sampling should produce a pretty representative sample, which we have already seen back in Chapter 6 makes good intuitive sense.

9.1.3 The Imaginary Dataset

Okay, we've got the population and the sample, we've figured out what it is about the sample that's of principal interest—the sum $S$ of the 100 sampled values—and we're trying to compute the probability that $S$ would come out greater than $0$ if you were to make 100 IID draws tomorrow. From Chapters 7 and 8 we have two main approaches to quantifying probabilities: frequentist and Bayesian. In problems like this the Bayesian story doesn't add anything to the frequentist approach: there doesn't seem to be any prior information lying around in this situation, and if we imagined betting with someone about whether $S$ will be positive or not, it's not at all clear what odds we should give or take to make the bet fair (except to say that $P(S > 0)$ should be less than 50% or else the casino is crazy). So let's try the frequentist approach, which involves repeating over and over the process that leads to a value of $S$—that is, repeatedly making 100 $1 single-number Keno bets—and computing the percentage of time that your net gain $S$ comes out positive.

Recall from Chapter 7 that there are two ways to carry out this relative-frequency plan in practice: we can either actually do the repetitions (or something equivalent to them), which is called the simulation approach, or we can imagine doing them and then use math to work out the relative frequency we want. The math approach is preferable, when feasible, mainly because using math to figure out how to compute probabilities in one kind of problem often indicates how to compute them in other problems as well (whereas simulation tends to be less generalizable). The simulation approach has a long and honorable tradition—from the early days in which people were interested in probability (the mid 1600s) to the present, people have used simulation to approximate probabilities that are difficult to compute any other way—and the only difference between the simulation and math
approaches is whether you actually do the repetitions or just think about doing them, so it is useful to look first at the simulation story.

The Simulation Approach. Before the advent of computers in the 1950s, people used one version or another of what might be called physical or empirical simulation. With this approach in the case of Keno, for instance, you would

- Get one of those cylinders with the 80 balls in it that I mentioned at the beginning of the chapter, or create something similar to it (like a hat with slips of paper numbered from 1 to 80 in it),

- Draw 20 numbers at random without replacement and see if your favorite number was among the 20 chosen values, recording −$1 or +$2 accordingly,

- Do the previous step a total of 100 times and add the −$1’s and +$2’s to get a value of $S$, and

- Repeat the last two steps $M$ times for some (preferably large) value of $M$, and work out the percentage of repetitions in which $S$ came out positive.

This would take a lot of time and effort, and at the end of it all you would have would be an estimate of the probability we want, because in practice—out of boredom or fatigue—I bet you would settle for a relatively small value of $M$, and (as I mentioned in Chapter 7) the official frequentist value of $P(S > 80)$ is only obtained by letting $M$ go to infinity.

These days to carry out the simulation approach people would write a computer program instead of drawing slips of paper out of a hat, which would improve on the physical approach in two ways: it would be faster, and you could get an estimate of arbitrary accuracy just by running your computer long enough to make $M$ really big. However, before we settle for simulation let’s see what’s involved in the math approach—as I said earlier, the probability model we’re building turns out to generalize pretty smoothly to almost all the rest of the problems we’re going to encounter in the rest of the book, so the effort to finish building it seems worthwhile.
The Math Approach. The math story goes just like the simulation approach I itemized a minute ago, except that you just imagine doing each step instead of actually doing it. The fundamental operation here is making 100 $1 single-number Keno bets and recording your net gain at the end of the 100 gambles, which we modeled in Figure 1 by (a) making 100 IID draws from the population to create the sample and (b) computing the sum $S$ of the 100 draws. To finish calculating $P(S > 0)$ we evidently need to imagine taking another IID sample of size $n = 100$ and computing the sum of these 100 draws, taking another IID sample and computing the sum of those draws, and so on, over and over. It is convenient to collect the values of $S$ you would get by doing this into a third dataset that might be called the imaginary dataset, as in the right-hand part of Figure 1. The first entry in the imaginary dataset is the sum $S$ based on the first sample, which I have said in the middle of Figure 1 might come out $-31$; the second entry is the value of $S$ based on the second sample of 100 IID draws, which might come out $-10$ (say); and so on. I have labeled this column of numbers “Possible Sums” in the figure to remind us of what the imaginary dataset is.

The number of rows in the imaginary dataset is just the number of times we imagine getting another value of $S$, which in the simulation approach I called $M$. I said above that to get the official value of $P(S > 0)$ you have to let $M$ go to infinity, so apparently the imaginary dataset should have an infinite number of rows. You can see that building these probability models would have been second-nature to [the Red Queen] in Alice in Wonderland, who ["always imagined two or three impossible things before breakfast."] I guess while we’re imagining the rest of this, it doesn’t hurt much more to imagine a dataset with an infinite number of values in it.

9.2 The Expected Value and Standard Error of a Sum

Okay, so we’ve got this imaginary dataset with a lot of values of $S$ in it—how do you figure out what percentage of these values are greater than $80$? We have seen that it usually seems useful when somebody
gives you a dataset to work out three things to summarize it: the mean, the SD, and the histogram. Maybe that will help here, too.

"Wait," you will be thinking by now, if this whole thing is starting to seem farfetched to you (as well it may), "How can you figure out the mean and SD of an infinite number of values of a variable you've only imagined, much less the histogram?" It may sound impossible, but in the end it just turned out to be really hard: It took people about 150 years, from roughly 1650 to 1800, to answer these questions in full generality. The main actors in the drama were [xxx, ..., and xxx (for more of the history see Stigler, 19xx)]. I'll take the three items on the list—mean, SD, and histogram—in that order.

The Expected Value of a Sum. Let's talk about the mean of the imaginary dataset first, which I'm going to call the long-run average or expected value of the sum $S$—symbolically, $E(S)$, or $E_{IID}(S)$ if we're being notationally careful to remind ourselves of the sampling mechanism. To begin with, what ingredients should go into calculating this quantity? Well, remember what $S$ is supposed to stand for—it's your net gain after 100 gambles. The long-run average or expected value of $S$ should certainly depend on how much money you expect to win or lose on each gamble, which we called $\mu$ awhile ago and which in this case came out $-\$0.25$. It should also depend on how many gambles go into the sum, which we have called $n$—after all, if you expect to lose money each time, the more times you gamble the more you should expect to lose. So I guess the expected value of $S$ should involve $n$ and $\mu$, but how? Here's a way to put it that makes the answer clear:

Q: If I expect to lose a quarter each time I do something and I do it 100 times, how much money do I expect to lose overall?

A: A hundred quarters.

In other words,

The expected value of the sum of some IID draws from a population is the number of draws times the population mean. Symbolically,

$$E(S) = E_{IID}(S) = n\mu.$$  \hspace{1cm} (9.6)
Here with \( n = 100 \) and \( \mu = -0.25 \) the expected value works out to \(-$25\), which answers part of the third question in the case study: After 100 $1 plays you expect to be behind by about $25. This looks like a good game for the casino.

**Math Interlude.** If you get interested in the random-variable end of this business you will want to know more about how to work algebraically with expected values. Here are two basic rules for doing so, together with a few words meant to make them plausible. In what follows the symbol \( c \) stands for a constant (something that does not vary randomly from one repetition to the next of whatever it is you’re studying) and upper-case symbols like \( Y \) are random variables (things that do vary randomly from one repetition to the next).

- \( E(cY) = cE(Y) \) (if you multiply something that is varying randomly around 5 (say) by 2 (say), the product should vary randomly around 10). In words, **multiplying a random quantity by a constant multiplies its expected value by the same constant**.

- \( E(Y_1 + Y_2) = E(Y_1) + E(Y_2) \) (if one thing is varying randomly around 3 (say) and another around 4 (say), the sum should vary randomly around 7). This rule applies to adding more than just two random quantities together; in words you could say that **the expected value of the sum is the sum of the expected values**, whenever the number of terms going into the sum is finite.

These rules exactly parallel those back in Chapter 3 for working with means of data values. It’s easy to use the second rule to prove the expected value formula (9.6) above: with \( S = Y_1 + \ldots + Y_n \), in which each draw has mean \( \mu \),

\[
E(S) = E(Y_1 + \ldots + Y_n) = E(Y_1) + \ldots + E(Y_n) = \mu + \ldots + \mu = n\mu.
\]

(9.7)

**The Standard Error of a Sum.** Well, after 100 plays you expect to be behind by $25, but give or take how much? Maybe the give-or-take is big enough that you still have a pretty good chance of coming out
ahead. This is where the SD of the imaginary dataset comes in. It has
a special name—it’s called the standard error of the sum $S$, written
$SE(S)$ or $SE_{IID}(S)$ if we’re being explicit about how the draws were
chosen. The SE is harder to pin down than the expected value—we
can use intuition to figure out what ingredients should go into it, and
in roughly what way, but the precise formula requires some math to
derive. Here we’ll just go through the intuition and I’ll say what the
answer is, and you can read through the math interlude below for more
information about where the formula comes from.

The issue is how much the sums in the imaginary dataset should
vary around their long-run mean $-25$. In figuring this out it helps to
remember that each value of $S$ is the sum of 100 IID draws from the
population, all of which are fluctuating around the population mean
$\mu = -25$ by an amount given by the population SD $\sigma$, which here
came out $1.30$. It makes pretty good intuitive sense that the variabil-
ity of the sum should depend on the variability of the draws going into
the sum, so I guess $SE(S)$ should involve $\sigma$. Moreover the standard
error should involve the population SD in such a way that as $\sigma$ goes
up, so does the SE, because as the individual draws become more vari-
able, their sum should, too. We will see as we go along that standard
error formulas often take the form of fractions, with a numerator and
a denominator, and that the ingredients in the formulas tend to ap-
pear either in the numerator or the denominator but not in both. I
have just argued intuitively that $\sigma$ should appear in the numerator of
$SE(S)$. This also makes sense by thinking about the units in which
$\sigma$ and the SE are measured: in Keno $\sigma$ is in dollars, the SE (which is
after all just the SD of the sums in the imaginary dataset) should also
be in dollars, and it would be a mistake to put $\sigma$ in the denominator
because then the units would come out wrong.

It turns out that there is just one other ingredient in the formula
for $SE(S)$: the number $n$ of draws going into each sum. You can see
that $n$ is relevant to the SE by thinking over the following three facts:

- Noticing that the sums in the imaginary dataset exhibit variabil-
ity around their mean is equivalent to saying that we are uncertain
about the precise value that any one of the sums will take on,

- This uncertainty arises because we are in turn uncertain about
the exact value each of the draws going into a given sum will take, and

- It stands to reason that our uncertainty about a sum should depend in part on how many uncertain things are being added together to yield the sum.

Moreover, this same argument shows that \( n \) should appear in the numerator of the standard error formula along with \( \sigma \), because the more things you add together, each of which is uncertain, the more uncertainty you should have about the sum.

That’s about as far as intuition can take us; I now have to just say the answer, which comes out like this:

\[
SE(S) = SE_{IID}(S) = \sigma \sqrt{n}.
\]  \hspace{1cm} (9.8)

Here with \( \sigma = \$1.30 \) and \( n = 100 \) the SE works out to \( \$1.30 \cdot \sqrt{100} = \$13 \). Putting this together with the expected value answers another part of the third question in the case study: the interpretation is that after 100 plays you expect to be behind by about \( \$25 \), give or take about \( \$13 \). In terms of the sums in the imaginary dataset this means that you would not be surprised to see a sum like \( -\$31 \) or \( -\$10 \) but you would be pretty surprised to see a value like \( +\$20 \), because \( +\$20 \) is \( \frac{+\$20 - (-\$25)}{\$13} = 3.5 \) SDs above average and that sort of thing doesn’t happen very often. It’s beginning to sound like our chance of coming out ahead with single-number betting in Keno isn’t too good.

\textit{Math Interlude.} You will also want to know a few rules for algebraically manipulating standard errors if you get into the random variables story more deeply. It turns out that it is easier to express what’s going on not in terms of the SE of a random quantity but in terms of the \textit{square} of the SE, which—since a standard error is just a kind of standard deviation—people call the \textit{variance} of the random quantity, in parallel with the
terminology in Chapter 3 for measures of spread for data values. Let's use the notation \( V(Y) \) to stand for the variance of a random variable \( Y \), so that symbolically the definition in the last sentence becomes \( V(Y) = [SE(Y)]^2 \) and \( SE(Y) = \sqrt{V(Y)} \). As with expected values there are two basic rules to learn about how variances work, and each rule can then be translated back into a statement about standard errors, by taking the square root of both sides of the variance formula.

- \( V(cY) = c^2V(Y) \), so that \( SE(cY) = |c|SE(Y) \) (if you multiply something that is varying randomly around its mean with a give-or-take of 3 (say) by 2 (say), the result should vary around its mean with a give-or-take of 6). In words, multiplying a random quantity by a constant multiplies its variability by the absolute value of the constant.

- If \( Y_1 \) and \( Y_2 \) are independent, \( V(Y_1 + Y_2) = V(Y_1) + V(Y_2) \), so that in this case

\[
SE(Y_1 + Y_2) = \sqrt{V(Y_1 + Y_2)} = \sqrt{V(Y_1) + V(Y_2)} = \sqrt{[SE(Y_1)]^2 + [SE(Y_2)]^2} .
\]

(9.9)

In words, when independent random quantities are added together, the variance of the sum is the sum of the variances, and the standard error of the sum follows a Pythagorean law: the SE is like the hypotenuse of a right triangle whose sides are the SE's of the terms going into the sum. We will return to this fact later in Chapter xx, where it is crucial to analyzing data gathered with two independent samples from different populations. Notice that it is on the variance scale that variability is additive, not the standard error scale. This is the only reason the term "variance" was invented in the first place—after all, as a measure of variability it has the wrong units (if \( Y \) is in dollars then \( V(Y) \) is in dollars\(^2\), which are pretty hard to spend). In practice people do math about the uncertainty expressed by random variables on the variance scale and then interpret the results by taking the square root, to get
things back onto the standard error scale where they belong.

If you spot me without proof the formula for the variance of a sum, we can now see immediately where the formula \( SE(S) = \sigma \sqrt{n} \) comes from: since \( S = Y_1 + \ldots + Y_n \) and the individual draws \( Y_i \) are IID, meaning that the \( Y_i \) are independent and all have the same standard error (SD) \( \sigma \) and hence variance \( \sigma^2 \),

\[
V(S) = V(Y_1 + \ldots + Y_n) = V(Y_1) + \ldots + V(Y_n)
\]
\[
= \sigma^2 + \ldots + \sigma^2 = n\sigma^2,
\]  
(9.10)

and you get the result we want, \( SE(S) = \sigma \sqrt{n} \), by taking the square root of both sides of this equation.

9.3 The Central Limit Theorem

This is a long calculation, and we’re almost to the end of it. We’re still wondering what \( P(S > 80) \) is, which is equivalent to wondering what percentage of the sums in the imaginary dataset are positive. If we knew what the histogram of the sums in that dataset looked like (drawn on the density scale for convenience), we could answer this question by working out the area under the histogram to the right of 80, in much the same way that Quetelet answered questions about the heights of his Scottish soldiers back in Chapter 4. So, what does the histogram of the sum of a bunch of IID draws from a population look like?

Of the three summaries of the imaginary dataset we have considered (mean, SD, and histogram), this was the hardest for people to figure out. It turns out that the answer depends on two things—the sample size \( n \) and the shape of the population histogram—which at first glance makes this whole approach sound impractical again, since there’s an infinite variety of possible shapes the population histogram could take on and an infinity of possible values for \( n \). But xxx in xxx, and people building on his work, were able to prove the following remarkable result, which has come to be called the Central Limit Theorem, or CLT for short:

**Central Limit Theorem:** Pretty much no matter what the population histogram looks like, as long as the number \( n \)
of IID draws going into a sum is big enough, the histogram of the sum $S$ will follow the normal curve pretty well.

This is one of the two most important theorems in the book, along with Bayes' Theorem—we will see that it makes feasible many probability calculations that would otherwise have been quite hard to carry out. Let's use the CLT to finally compute the chance of coming out ahead with 100 single-number $\$1$ gambles at Keno, and then I'll have more to say about when the theorem applies in practice and when it doesn't.

The picture in the lower right corner of Figure 1 finishes the calculation: according to the CLT, the histogram of the sum follows the normal curve with a mean (expected value) of $-\$25$ and an SD (standard error) of $\$13$ pretty well, so $0$ in standard units is $\frac{\$0 - (-\$25)}{\$13} = 1.9$ and the area to the right of 1.9 under the standard normal curve from Table A.1 is about 2.7%, or about 1 chance in 37. If you play single-number Keno once for $\$1$, your chance of coming out ahead is only 1 in 4; if you persevere and make 100 such plays, your chance of coming away a winner drops to only 1 in 37. Keno is a terrible game for the gamblers and a great game for the casino.

Math Interlude. The way I stated the Central Limit Theorem a minute ago makes it sound pretty vague, but it has a precise statement as well:

If $Y_1, \ldots, Y_n$ are IID draws from a population with mean $\mu$ and finite SD $\sigma$, $S = Y_1 + \ldots + Y_n$, and $\Phi(z)$ is the function that keeps track of the area under the standard normal curve to the left of $z$, then for all $z$

$$
\lim_{n \to \infty} P\left( \frac{S - n\mu}{\sigma\sqrt{n}} \leq z \right) = \Phi(z).
$$

This theorem has hundreds of variations—as you read this page, I'm sure that someone somewhere is proving another form of it—and (unlike Bayes' Theorem) the plain-vanilla version I just stated requires some fairly high-powered math to demonstrate (most people taking statistics courses don't see a complete proof of it until they are graduate students in the subject). It's a pleasant fact that we don't need to know how to prove it to make good use of it in the rest of the book.
Figure 2a shows the actual histogram of the sum of 100 IID draws from the population in Figure 1, with the normal curve specified by the Central Limit Theorem superimposed on it. The real histogram has gaps in it, because you can't get all possible numbers from \(-\infty\) to \(+\infty\) by adding together 100 \(-1\)'s and \(+2\)'s; in fact the smallest possible value for your net gain is \(-100\) (if you lost every time), the second-smallest possible value is \(-97\) (if you won only once), and so on up to the biggest possible value of \(+200\) (if you were so lucky as to win every time), with the possible net gains occurring every \$3 along the way from \(-100\) to \(+200\) (\$3 comes from the fact that \(-1\) and \(+2\) are \$3 apart). The CLT does not appear to provide a very good normal approximation here, but this is visually misleading since the normal curve is continuous and the real histogram is discrete. When I correct for this in Figure 2b, by replotting the real histogram with bars that are \$3 wide and adjusting the normal curve accordingly, you can see that in fact the normal approximation is quite good—for example, the exact value of \(P(S > 0)\) in this case is 2.75\%, which is close to the approximate value we got above from the CLT (2.71\%).

Evidently for this population histogram and this value of \(n\) the CLT has worked pretty well. But the theorem (especially as I stated it above) is unclear on an important point—how big does \(n\) have to be to get a good normal approximation for the histogram of a sum?

When Does the Central Limit Theorem Apply? This question has no easy general answer. There's a hint from the word "limit" in the name of the theorem—mathematically the CLT guarantees a perfect normal curve only in the limit as the number \(n\) of draws going into each sum \(S\) becomes infinitely large. Fortunately perfection is not needed, and—as the single-number Keno example above in Figure 2 demonstrates—values of \(n\) considerably smaller than \(+\infty\) can yield really good normal approximations. One thing should be clear, though, from the idea of taking a limit as \(n\) goes to infinity: The normal approximation for the histogram of the sum improves as \(n\) grows. If you already have a good normal approximation with \(n = 100\), as in Figure 2, it would be even better for \(n = 200\).
Note that what counts in getting a good normal approximation for the histogram of the sum is not the population size $N$, or the number $M$ of hypothetical sums in the imaginary dataset; it’s the number $n$ of draws going into each sum.

The best way to gain additional insight into when the CLT applies is by looking at a few more examples. Here are two.

- **What if the population histogram is normal to begin with?** Suppose the population has many values in it, and a histogram of those values follows the normal curve pretty well, as in Figure 3. Then the following math fact comes in handy:

  **Math Fact:** For any $n$, the histogram of the sum of $n$ IID draws from a normal population is itself normal.

This means that when the population histogram follows the normal curve pretty well to begin with, the histogram of the sum will already follow the normal curve well even with only $n = 1$! This makes sense if you think about it: With only one value $Y_1$ going into the “sum,” the sum just is that value, so the histogram of the sum is just the histogram you would get by repeatedly taking one draw at random from the population and then replacing it—but that’s just the sample histogram, which we already know is supposed to look like the population histogram, which in this case is normal. Evidently

  The closer the population histogram is to normality to begin with, the smaller $n$ needs to be to get a good normal approximation for the histogram of the sum.

- **What if the population histogram is quite skewed?** Figure 4a gives the histogram of hospital expenditures for Medicare patients in 19xx with a heart attack. You can see it has a quite long right-hand tail—the typical Medicare patient with a heart attack ran up a hospital bill of about $\$xxxx$, but some people had much higher expenses than that. An average hospital will have about 50 such patients per year. If we think of Figure 4a as a population histogram and imagine taking an IID sample of $n = 50$ of
these patients, the sum of those 50 draws would keep track for us of the total amount of money paid out by the government per year to a typical hospital to treat heart attacks in elderly people, which would be a dollar figure of some policy interest. Of course, this total would vary from hospital to hospital according to some histogram. What would this histogram look like?

Figure 4b shows the exact histogram, together with the normal approximation offered by the CLT. The normal curve doesn't fit too well in this case—with only 50 draws going into each sum, the histogram of the sum is still pretty skewed. Figure 4c shows the corresponding histograms for \( n = 200 \), and you can see that the normal approximation is much better. The moral is that

\[
\text{If the population histogram is sharply skewed, } n \text{ needs to be quite a bit bigger to get a good normal approximation for the histogram of the sum.}
\]

To summarize all of this,

\[
n = 50 \text{ or } 100 \text{ should be enough to give a workable normal approximation for the histogram of the sum of a bunch of IID draws, unless the population histogram is strongly skewed to begin with, in which case several hundred draws should be enough. Some populations are close enough to normality to begin with that } n = 5 \text{ or } 10 \text{ draws are already enough.}
\]

All general rules of this type can be violated—in other words, you can invent populations for which this rule makes you think you have a better normal approximation than you actually do—but we will see as we go along in the book that building probability models to learn about real-world phenomena is a process that involves several approximations to reality, and when the modeling process fails it rarely does so because the CLT has let you down. I will return to this point in later chapters.

Comparing Single-Number and Double-Number Gambling at Keno. Let's wrap up the chapter by working out the chance of coming out ahead for double-number gambling at Keno—this will review
all the steps in building a probability model, and there is an interesting punchline about how to decide which gamble is better. Figure 5 summarizes the results of the modeling.

First the population. Recall ages ago at the beginning of the chapter I said that the payoff for double-number wagering at Keno was 11 to 1. If we’re going to imagine making $1 bets, as we did with the single-number gambling strategy, this means that the population should have -$1’s and +$11’s in it, but how many of each? You can verify using logic similar to that in Section 1 that, just as the chance of winning any single number bet was \( \frac{20}{80} = \frac{380}{6320} \), or almost exactly 6% (see Problem xx). If we’re being careful in building the population dataset we should put 380 +$11’s and \((6320 - 380) = 5940\) -$1’s in it, which is what I did in Figure 5, but you could get a good approximation to this by using 6 +$11’s and 94 -$1’s if that seems simpler to you.

Next comes the population mean \( \mu \) and SD \( \sigma \):

\[
\mu = \frac{380(+$11) + 5940(-$1)}{6320} = \frac{-$1760}{6320} = -$0.28,
\]

\[
\sigma = \sqrt{\frac{380[+11 - (-$0.28)]^2 + 5940[-1 - (-$0.28)]^2}{6320}} = $2.85.
\]

Double-number gambling is a bit worse than single-number wagering in how much money you expect to lose each time (28¢ versus 25¢), but it’s quite a bit more variable in its outcomes (the SD is more than twice as big). What do you think these two facts imply about whether your chance of coming out ahead is bigger with the double numbers? ♠️♠️

The sample part of the model diagram is identical to that with single-number gambling—\( n = 100 \) IID draws, with attention focusing on the sum—except that the sample histogram should now resemble the population histogram on the left side of Figure 5, which is considerably more skewed than its counterpart back in Figure 1. The imaginary dataset stands for the same thing as it did earlier, with each entry representing a possible sum of 100 draws, but its long-run mean (the expected value of the sum) and SD (the standard error) are different:

\[
\text{expected value} = E(S) = n\mu = 100(-$0.28) = -$28,
\]

\[
\text{standard error} = SE(S) = \sigma\sqrt{n} = 10 \cdot $2.85 = $28.50. \quad (9.12)
\]
I expect to lose about $28 with double-number gambling, which is not much different than with single-number betting ($25), but the give or take around the expected value is quite a bit bigger ($28.50 versus $13). This means, interestingly, that my probability of coming out ahead, $P(S > 0)$, is quite a bit bigger than it was with the single numbers, as follows: (1) the Central Limit Theorem should again give us a pretty good normal approximation to the histogram of the sum; (2) "coming out ahead" with double-number gambling works out to $\frac{\$0 - (-\$28)}{\$28.50} = 0.98$ in standard units; and (3) $P(S > 0)$ therefore comes out to about 16%. (The CLT doesn't give quite as good an approximation here—the exact answer is about 14.8%—but it's good enough for our purposes.) A coming-out-ahead chance of 15% or so is more than five times larger than the corresponding value for single-number wagering (2.7%), so I guess double-number betting at Keno is better. Or is it?

Figure 6 shows the normal approximations to the two histograms for your likely net gain after 100 plays with the two strategies, plotted on the same dollar scale. The double-number approach does indeed put more area under the curve to the right of $0$, because double-number wagering is more variable (more uncertain) than single-number betting, but the something-for-nothing bell should be going off in your head if I tried to tell you that this benefit came without any cost. The figure shows what the cost is: the additional variability created by double-number betting means that your chance of losing a lot of money is also a lot bigger than it was with the single numbers. For instance, with single-number betting your chance of losing $50 or more is the same as your chance of coming out ahead—2.7%—because $-50$ is just as far to the left of the expected value ($-25$) as $0$ is to the right of it, but with double-number wagering this chance is well over 20%.

So which gamble is better? There is no single "correct" answer for everybody—it depends on how conservative you are about taking risks. People studying decision theory would call you risk-averse or risk-seeking according to your preference in situations like this (see the discussion in Chapter xx about utility functions). Which gamble do you like better? Or is the punchline in this case that you should keep your money in your pocket?
9.4 Chapter Summary

1. Probability models are useful for studying random processes like the amount of money you’re likely to win or lose at games of chance. Such models have three parts: the population, the sample, and the imaginary dataset. Each can be thought of as a dataset, with mean, SD, and histogram. The population represents what you might get each time you repeat the process you’re studying; the sample helps to simulate a number of such repetitions, and to focus attention on a summary of the repetitions like their sum; and the imaginary dataset helps to visualize what you would get if you took repeated samples and wrote down the summary value each time. See Figures 1 and 5 for examples of probability modeling diagrams that bring all these ingredients together.

2. To specify the population, you identify the possible values the random process you’re studying could produce on any give repetition, together with the probabilities associated with each of those values. In gambling models, for instance, the possible values are the dollar amounts that could change hands each time the gamble is repeated, and you can often use the rules in Chapter 7 to figure out the probabilities of each of these outcomes occurring. The population mean $\mu$ and SD $\sigma$ are computed in the same way we did back in Chapter 3 with any other lists of numbers, except that when working out the population SD you divide by the number $N$ of values in the population rather than $(N - 1)$.

3. The sample represents one possible outcome of a bunch of repetitions of the random process you’re studying. In the Keno example of Section 1, for instance, we were wondering what would happen if you played Keno 100 times, so the population modeled the individual plays and the sample collected $n = 100$ such gambles together. It’s important when describing the sample to be specific about the sampling mechanism—the relationship between the sample and the population, of which IID (drawing the sample at random with replacement, as in gambling) and SRS (at random without replacement) are the two simplest cases. The other important thing to specify about the sample is what summary of it is of particular interest in the problem at hand. In the Keno models of this chapter, for instance, the individual draws represented the amounts of money we won or lost on each individual
Keno game, and the sum of the draws in the sample stood for our net gain after all 100 gambles.

4. The imaginary dataset helps you visualize what would happen if you took repeated samples and calculated the relevant sample summary each time. With Keno, for example, you imagine making 100 IID draws and calculating their sum \( S \), making 100 more IID draws and working out their sum, and so on, writing down the sums you would get in a dataset. Calculating the chance of coming out ahead after 100 gambles—symbolically, \( P(S > 80) \)—then amounts to working out the percentage of the sums in the imaginary dataset that are positive, or equivalently computing the area under the histogram of the imaginary dataset to the right of 80.

5. This area is hard to work out exactly, but—when the summary of interest is the sum of some IID draws from the population, as it was with Keno—the probability we want may often be approximated by computing the long-run mean and SD of the imaginary dataset and working out the relevant area under the normal curve with that mean and SD.

6. The long-run mean of the sums in the imaginary dataset is called the expected value of the sum \( S \), written \( E(S) \). You can calculate it with the following rule:

   The expected value of the sum of some IID draws from a population is the number of draws times the population mean. Symbolically, \( E(S) = n\mu \).

7. The long-run SD of the sums in the imaginary dataset is called the standard error of the sum \( S \), written \( SE(S) \). The rule for calculating it is as follows:

   The standard error of the sum of some IID draws from a population is the population SD times the square root of the number of draws. Symbolically, \( SE(S) = \sigma \sqrt{n} \).

8. As long as the number \( n \) of IID draws going into a sum is big enough, the histogram of the sum \( S \) will follow the normal curve pretty well. This result is called the Central Limit Theorem, or CLT for short. The closeness of the normal curve to the actual histogram of the sum
depends on two things: the sample size $n$ and the population histogram. The closer the population histogram is to normality to begin with, the smaller $n$ needs to be to get a good normal approximation. In general, $n = 50$ or $100$ should be enough unless the population histogram is quite skewed, in which case several hundred draws going into each sum may be needed to get a good normal curve. Some populations look normal enough to begin with that already with only $n = 5$ or $10$ you have a nice normal approximation.

9. (Math) The mathematical objects that stand for things like the process of drawing at random from a population are called random variables. In the Keno example, for instance, $Y_2$ stood for the process of making the second draw from the population in Figure 1, and $S$ stood for the process of making all $100$ draws and computing their sum. $Y_2$ and $S$ are both random variables, because each time you make the second draw or add together all $100$ draws you will get a different result at random. Random variables have means and SDs just like datasets, but the names are different: the mean of a random variable is called its expected value, and the SD is called its standard error. These terms coincide with the idea, expressed more informally in this chapter, of the long-run mean and SD of the imaginary dataset. Expected values and standard errors of random variables obey the following simple rules, in which $c$ stands for a constant and $V(Y)$ is the square of the standard error of $Y$ (the variance of $Y$):

- $E(cY) = cE(Y)$;
- $E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$;
- $V(cY) = c^2V(Y)$, so that $SE(cY) = |c|SE(Y)$; and
- If $Y_1$ and $Y_2$ are independent, $V(Y_1 + Y_2) = V(Y_1) + V(Y_2)$, so that $SE(Y_1 + Y_2) = \sqrt{[SE(Y_1)]^2 + [SE(Y_2)]^2}$.

9.5 Problems

1. (gambling) What should the payoff have been in single-number wagering at Keno so that the gamble is fair? What about with double-number betting? If both gambles were fair, show that neither would have an advantage over the other with respect to the probability of coming out ahead. What other criteria would you use, and which of
the two fair gambles seems better to you when evaluated with your criteria? Explain briefly.

2. (engineering) Some traffic engineers working for the city of Los Angeles are trying to synchronize the lights on a long stretch of Santa Monica Boulevard so that someone driving along at a constant speed like 30 miles an hour would hit them all green (you can tell this problem is hypothetical—they would never do anything that sensible), so they set up a traffic survey in which they record the times at which cars pass various points along the street at various times of the day and night on weekdays and weekends. There are two main ways to study data of this type—you can count the number of cars going by in fixed-length intervals of time, or you can keep track of the times between cars. The point of this problem is that even if you choose just one of these methods you can answer questions phrased in terms of the other, so in that sense the two approaches are equivalent.

These engineers decide to focus on the times between cars, and examination of the data for one spot along the street in one direction shows that during one stable time period, on weekday afternoons from 1 to 3:30 pm, there is no systematic trend or pattern to the interarrival times and they average 5.5 seconds with an SD of 4 seconds.

One of the standard volume-of-traffic criteria used in studies like this involves the number of cars passing by in a 10-minute period: if at least 100 cars often go by in that amount of time, the spot is characterized as “busy” and treated differently in the subsequent analysis. Suppose that a car passes by this spot one day at some moment in this weekday afternoon period, and one of the engineers starts her stopwatch at that moment and begins counting cars. How likely is it that at least 100 cars go by in the next 10 minutes? Would you classify this as a busy spot? Explain briefly. (Hint: This problem is not straightforward. Try building a probability model based on the times between cars, and relate the event that at least 100 cars go by in 10 minutes to this model.)

3. (gambling) Show, using reasoning similar to that in Section 1, that the chance of winning any given double-number bet in Keno is \( \frac{20 \times 20}{80 \times 75} = \frac{380}{3325} \), or just about exactly 6%. How does this generalize to three numbers, or more? Explain briefly.
4. (probability modeling) From what you have seen in this chapter, why do people concentrate on the mean and SD as summaries of the population, sample, and imaginary datasets, rather than other measures of center and spread from Chapter 3 like the median and interquartile range? Explain briefly. (Hint: The Medicare example in Figure 4a is helpful (what is the quantity of greatest real-world interest in that example?), and so is the CLT.)

5. (gambler's ruin)

6. (lottery)

7. (roulette)

8. (the elevator problem?; more)
Figure 3: Where the population histogram is normal to begin with, the histogram of the sum will be normal too, even with only $n=1$.

(Figure 4a and 4b to be supplied later)
Figure 6. Histograms of your net gain after 100 $1 single-number Keno bets (solid curve) and 100 $1 double-number bet (dotted curve).
Chapter 10

Working With Averages

Chapter 9 was about how to build probability models for sums, and how to use them to calculate things like your chance of coming out ahead if you are so foolish as to gamble in a casino. This chapter is about how to build probability models for averages, or means. Sums and averages are closely related, so a lot of what we developed in Chapter 9 carries over here, but several important qualitative conclusions are different, in an interesting way that has major implications for the inferential work we will start next chapter.

The other main topic in this chapter is the basic measurement error model in probability and statistics, because the idea of combining a bunch of measurements by taking their mean is one of the most important uses of probability models for averages. This chapter’s main case study is an example taken from medical decision-making—it’s typical of a simple kind of measurement error problem that comes up frequently in medicine. Before we get to it I want to talk a bit about the idea of measurement errors, and how people build models for them. The material in this chapter is a bit more abstract than last chapter’s ideas—we will see that in an important way the basic measurement error model is even more hypothetical than the models we built for Keno in Chapter 9—but I will try to appeal to your common sense often enough to keep us from wandering too far away from reality.
10.1 The Basic Measurement Error Model

Measurement error models are an attempt to "explain" the frustrating fact that

| If you measure something more than once under what you think are identical conditions, you may well get a different answer each time. |

The problem arises frequently in science in fields like chemistry and physics; in its simplest form in these fields it comes up when people try to determine what are called the fundamental constants, for example the speed of light or the mass of the electron. I put the word "explain" in quotes above because it will become clear that the measurement error model doesn't really explain why you get a different value each time you measure something, it just describes the fact that you do.

Cavendish and the Density of the Earth. Here is an early example. In 1798 an English physicist named Henry Cavendish tried to experimentally estimate the (mean) density of the earth relative to that of water, using a piece of equipment called a torsion balance. He repeated his experiment 29 times, and his results are given in columns 1, 2, 4, and 5 of Table 10.1 (ignore columns 3 and 6 for now). You can see that he got a different estimate each time. Why?

Well, maybe what he was measuring changed each time he tried to measure it. That idea sounds stupid in this case, and in this case it is stupid (nobody would imagine that the density of the earth fluctuates from moment to moment), but there are other settings in which it is far from stupid (the main example I have heard of is situations in subatomic physics that are governed by what is called the Heisenberg uncertainty principle, which says that, without meaning to, you can actually change something like the speed of an electron just by trying to measure it). Putting this possibility aside, it makes sense to identify a value that we might call the true density of the earth, and to note that Cavendish was trying to estimate this true, unchanging value by gathering the data in Table 10.1.

The Basic Measurement Error Equation. This still leaves unexplained the fact that he didn't get the true value each time he repeated
Table 10.1: Cavendish's measurements of the (mean) density of the earth, in grams per cubic centimeter (g/cm³)

<table>
<thead>
<tr>
<th>Measurement Number</th>
<th>Value</th>
<th>Random Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.50</td>
<td>+0.05</td>
</tr>
<tr>
<td>2</td>
<td>5.61</td>
<td>+0.16</td>
</tr>
<tr>
<td>3</td>
<td>4.88</td>
<td>-0.57</td>
</tr>
<tr>
<td>4</td>
<td>5.07</td>
<td>-0.38</td>
</tr>
<tr>
<td>5</td>
<td>5.26</td>
<td>-0.19</td>
</tr>
<tr>
<td>6</td>
<td>5.55</td>
<td>+0.10</td>
</tr>
<tr>
<td>7</td>
<td>5.36</td>
<td>-0.09</td>
</tr>
<tr>
<td>8</td>
<td>5.29</td>
<td>-0.16</td>
</tr>
<tr>
<td>9</td>
<td>5.58</td>
<td>+0.13</td>
</tr>
<tr>
<td>10</td>
<td>5.65</td>
<td>+0.20</td>
</tr>
<tr>
<td>11</td>
<td>5.57</td>
<td>+0.12</td>
</tr>
<tr>
<td>12</td>
<td>5.53</td>
<td>+0.08</td>
</tr>
<tr>
<td>13</td>
<td>5.62</td>
<td>+0.17</td>
</tr>
<tr>
<td>14</td>
<td>5.29</td>
<td>-0.16</td>
</tr>
<tr>
<td>15</td>
<td>5.44</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Number</th>
<th>Value</th>
<th>Random Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>5.34</td>
<td>-0.11</td>
</tr>
<tr>
<td>17</td>
<td>5.79</td>
<td>+0.34</td>
</tr>
<tr>
<td>18</td>
<td>5.10</td>
<td>-0.35</td>
</tr>
<tr>
<td>19</td>
<td>5.27</td>
<td>-0.18</td>
</tr>
<tr>
<td>20</td>
<td>5.39</td>
<td>-0.06</td>
</tr>
<tr>
<td>21</td>
<td>5.42</td>
<td>-0.03</td>
</tr>
<tr>
<td>22</td>
<td>5.47</td>
<td>+0.02</td>
</tr>
<tr>
<td>23</td>
<td>5.63</td>
<td>+0.18</td>
</tr>
<tr>
<td>24</td>
<td>5.34</td>
<td>-0.11</td>
</tr>
<tr>
<td>25</td>
<td>5.46</td>
<td>+0.01</td>
</tr>
<tr>
<td>26</td>
<td>5.30</td>
<td>-0.15</td>
</tr>
<tr>
<td>27</td>
<td>5.75</td>
<td>+0.30</td>
</tr>
<tr>
<td>28</td>
<td>5.68</td>
<td>+0.23</td>
</tr>
<tr>
<td>29</td>
<td>5.85</td>
<td>+0.40</td>
</tr>
</tbody>
</table>

Mean 5.45 0.00
SD 0.22 0.22

his experiment. Statisticians find it useful to write down a simple equation that draws attention to the discrepancies between the observed values and the truth (without necessarily explaining them), as follows.

\[
\text{observed value} = \text{true value} + (\text{observed value} - \text{true value}). \quad (10.1)
\]

Well, that doesn’t seem to have advanced the ball much—all we’ve done so far is point out, to anybody who didn’t already know, that \(a = b + (a - b)\). However, now we can begin to focus on the discrepancies (observed value – true value) on the right-hand side of equation (10.1), which people call measurement errors. Why do such discrepancies arise?

About the best people have been able to do in answering this question is to identify two “sources” of discrepancy, called systematic and
random errors, as in the equation

\[
(\text{observed value} - \text{true value}) = \text{systematic error} + \text{random error}.
\]

(10.2)

We have encountered systematic error earlier, as far back as Chapters 5 and 6, where we talked about it under its other name: bias. If you ask the person who sells fish at your favorite market to weigh a piece of fish several times for you, and he or she always puts his/her thumb on the scale in a way that makes the fish seem 2 ounces heavier than it is, that’s bias—a systematic tendency to over- or underestimate the true value. Notice that systematic error (if any), like the true value, doesn’t move around from observation to observation—that’s what makes it systematic.

In the standard measurement error model, by definition what’s left after any systematic error in the measuring process has been accounted for is lumped together under the name random error in equation (10.2):

\[
\text{observed value} - \text{true value} - \text{bias} = \text{random error}.
\]

Moving the true value and bias back over to the right-hand side gives the basic measurement error equation:

\[
\text{observed value} = (\text{true value} + \text{bias}) + \text{random error}.
\]

(10.3)

Actually, when you think about applying this to Cavendish’s data, since he had 29 observed values I guess this isn’t just one equation, in disguise it’s 29 equations, one for each of his measurements.

\textit{Math Interlude.} It will help later to have some notation for the quantities in these equations, so let’s use the following:

- \( n \) = how many measurements Cavendish made (here 29);
- \( Y_1 \) = his first observed value, \( Y_2 \) = his second measurement, and so on up to \( Y_{29} = Y_n \) = his last observed value;
- \( \theta \) = the true value;
- \( b \) = the systematic error; and
- \( e_1 \) = random error \#1, \( e_2 \) = random error \#2, and so on up to \( e_n \) = random error \#n.
Putting all this notation together in equation (10.3) gives a symbolic form of the basic measurement error equation: with $i$ standing for the index of a typical observation $Y_i$,

$$Y_i = \theta + b + e_i, \quad i = 1, \ldots, n.$$  \hfill (10.4)

**Random Measurement Errors.** The idea of random errors is that people have noticed in practice, in situations where the true value and bias were eventually determined, that what's left over after (true value + bias) is subtracted from all the observed values is—often, but not always—a bunch of numbers with two properties:

- They're haphazard-looking, with no pattern or trend to them, and
- Sometimes the observed value is on the high side of (true value + bias) and sometimes it's on the low side, meaning that some of the random errors are positive and some are negative—usually in such a way that they balance at or near zero, in the sense that their average is close to zero.

Columns 3 and 6 in Table 10.1 illustrate the idea of random error with Cavendish's data. These days we think the density of the earth is about 5.52 g/cm$^3$, and we currently believe that Cavendish's measurement process was biased on the low side by about 0.07 g/cm$^3$, so I have subtracted $5.52 - 0.07 = 5.45$ from all the measurements in the middle column to get what you might call Cavendish's "estimated random errors" (I will explain the use of the word "estimated" later). Notice that, pretty much as advertised, the estimated random errors do look fairly haphazard—there isn't anything regular you can discern about them, such as a positive one always being followed by a negative one—and they do seem to balance out at zero (for instance, 15 of them are positive and 14 are negative, and their mean is 0.0).

If we wanted to we could learn more about the contribution of random error to a typical measurement by working out the SD of the estimated random errors and drawing their histogram or stem-and-leaf plot. Since I got the estimated random errors by subtracting a constant (5.45) from the measurements, and adding or subtracting a constant from a column of numbers leaves the SD and distributional shape of
the original numbers unchanged, I could just as well focus on the SD and stem-and-leaf of the original measurements themselves, as in Figure 10.1. Either way the SD comes out 0.22 g/cm³. The interpretation of the SD is that each time Cavendish took a measurement, the typical amount by which it missed the truth was about 0.22.¹ Note that the stem-and-leaf roughly follows the normal curve, although with only 29 measurements rough judgments about distributional shape are about all we can make.

```
SD
0.22

0
6 4 2 3 2
7 4 4 5 3
9 6 6 7 5 5
8 7 0 9 7 8 8 9
4 4 5 5 5 5 5 5 5 5 5 5
8 9 0 1 2 3 4 5 6 7 8
```

Figure 10.1: Stem-and-leaf plot of Cavendish's density measurements.

After all these paragraphs I still haven't said why random errors crop up in measuring things. Sometimes it's just human error—somebody writes down the wrong number, for instance by transposing two digits. Sometimes what's going on is that, despite your best efforts to make the measurements under identical conditions, there are small variations in things that help to determine the outcome of the measuring process. Sometimes you don't read your measuring instruments exactly right. And so on—if you've ever tried to measure something carefully you can probably think of two or three more "explanations." In retrospect if we try hard enough we can often identify the cause of a discrepancy between the observed and true values, but prospectively any such differences (apart from bias) may well look "random" to us. I will have more to say in Chapter 11 about the nature of randomness; this is enough for now to get us into the case study.

¹This is not quite true, since Cavendish's measuring process appears (as noted above) to have been slightly biased on the low side; the right answer is more like 0.23 in this case. We will talk in Chapter xx about how to fold bias into the assessment of the accuracy of an estimate.
Case Study 10: Hypoglycemia

From symptoms you have been exhibiting (dizziness and light-headedness a few hours after meals), your doctor thinks you may have hypoglycemia, which is a fancy way to say low blood sugar. The normal blood sugar level in humans is around 90 milligrams per deciliter (mg/dl for short), and people whose level falls below 60 on this scale are considered hypoglycemic if they have the kind of symptoms you have shown.

Your doctor is thinking of ordering some laboratory tests to compare your level with these values. She knows that the usual lab test does not always give the correct level, “because of” measurement error. The lab she works with is known from past experience to give unbiased blood sugar readings that vary around the truth with an SD of about 10 mg/dl and a histogram that follows the normal curve pretty well. Common sense says that it would be a better idea to get several readings of your blood sugar level and average them rather than just settling for one value, but how many readings should she take?

The problem is complicated by the fact that if she gets blood samples from you on several different days, your actual blood sugar level may vary a bit. To keep things simple, let’s suppose that your blood sugar level doesn’t vary much from day to day, holding pretty constant at 75, so that you don’t really have hypoglycemia. If your doctor orders one lab test, what is the chance that the test will incorrectly claim that you are hypoglycemic? Suppose instead that she has you come in for blood samples on three different days, and bases her diagnosis on the average of these three blood sugar readings. What is the chance now that you will be incorrectly diagnosed? Which is better, one reading or three? How much better?

This case study is pretty realistic except for one thing: usually when your doctor takes measurements on you the true value of whatever it is she’s trying to measure will not be known. However, as we will see next chapter, trying to figure out what the true value is from measurements estimating it is an example of statistics, not probability. Before we tackle the harder task of trying to infer the truth from data relevant to such an inference, it’s useful to make some calculations in

\(^2\)Well, two things: in practice the SD of the blood sugar readings would be smaller. I have chosen this SD to emphasize the value of replicating measurements and averaging them.
the simpler setting in which the truth is known.

The questions in this case study are about the chances of various things happening, so I guess we need a probability model to answer them. It would be nice if we could use the same model we went to all that trouble to develop in Chapter 9. Recall that last chapter's models had three parts, all of which can be thought of as datasets: the population, the sample, and the imaginary dataset. How can we fit measurements like Cavendish's observations, or the blood sugar readings your doctor is about to take on you, into this framework?

The Population in Measurement Error Models. Well (you might want to review Section 9.4 before launching into this), the population is supposed to represent what you might get each time you repeat the process of interest. In this case evidently \{the process of interest\} = \{measuring some unchanging quantity over and over under “identical” conditions\}, so I guess we should take as our population the conceptual set of all possible measurements you might get any time you make a single observation. In the case study we are told some things about how such observations have gone in the past: they have fluctuated around the truth—which for the sake of illustration here is known to be 75—with an SD of 10, no bias, and a histogram that follows the normal curve pretty well. We have no reason to treat the past and the future as anything other than similar (as far as measuring blood sugar is concerned, at least), so it makes sense to take as the population here a dataset with mean $\mu = 75$, SD $\sigma = 10$, and approximately normal histogram (see the left-hand part of Figure 10.2).

Notice that there is a difference between this population dataset and the ones in Chapter 9: the population histograms last chapter were discrete (see Figures 9.1 and 9.5) and the population datasets had a finite number of values in them (for example, the one for single-number Keno betting only had $N = 4$ numbers in it). By contrast, the normal curve we're assuming for the histogram of the blood sugar population dataset is continuous—if you look at a range of potential blood sugar measurements from 65 to 85, say, in theory (with more and more precision in the measuring process) all of the values in that range are possible, so with this population $N = \infty$. People call populations like the one in Figure 10.2 conceptually infinite to contrast them with finite populations like those in Figures 9.1 and 9.5.
In the interests of realism I should say before going on that the normal curve the case study mentions for the population histogram would in practice be a kind of best-case in measurement situations like the one we’re thinking about. Often people notice that the great majority of their data—90 to 99%, say—follows a basically normal histogram shape, but the rest of the data is made up of one or more outliers that are just too far out in the tails (on either side of the average) to be consistent with the normal curve (recall we met the concept of outliers back in Chapter 3). This is entirely in keeping with the idea that occasionally an error of some kind crops up in the measuring process. We will talk more about how to deal with outliers in measurement error situations in Chapter xx; as simple as this case study is, it’s complicated enough without them for now.

The Sample in Measurement Error Models. Okay, so much for the population; what about the sample? Recall that the sample is supposed to represent one possible outcome of a bunch of repetitions of the random process you’re studying. With $n$ standing for the number of blood sugar readings your doctor wants to take on you (either 1 or 3 in the case study), I guess the sample should just represent those $n$ readings.

Well, that was easy. The next question we tackled about the sample last chapter—how does it relate to the population?—is harder in this case. What sampling mechanism is at the heart of repeated measurements of an unknown true value?

To begin with, if you had your way you would certainly want your measurements to be independent of each other, because if not you’re not getting a full piece of information—a full degree of freedom, as people say—each time you take a reading. After all, if the measurements depended on each other in some way, today’s reading would be some combination of new information and the old information contained in yesterday’s reading, and that would not be as informative as getting a whole new piece of information each time. Is it reasonable to assume that blood sugar readings based on blood samples taken from you on different days are independent? Sure—for this particular measurement process, there’s no reason that yesterday’s reading should affect today’s. It does sometimes happen that a lack of independence creeps
into people's attempts to measure something, so it's worth noting that independence is not guaranteed in this model, it's assumed, but in this case it's an assumption that seems justified. In Chapter xx I will talk a bit about how to check this assumption.

The word "independence" in the context of a discussion of sampling mechanisms makes me think of IID, since it's a sampling method we know something about, so it's natural to wonder if the ID (identically distributed) part of IID is as reasonable to assume in measurement error models as the I (independence) part was. The phrase "identically distributed" is meant in this context to suggest two things: that

- The measurements are taken under essentially identical conditions, and that

- The quantity being measured doesn't change over time while you're trying to measure it.

Both of these things are sensible goals in measuring something—to avoid shooting at a moving target, so to speak—and, like independence, they're both assumptions rather than guarantees; you need to think about their reasonableness in each case. Here we can only hope that the lab to which your doctor sends the blood samples does their blood sugar assay under more or less identical conditions each time, and—while it is easy to imagine that your blood sugar does indeed vary from day to day, violating the second of the assumptions above—we are encouraged to assume in the case study that variations of this type are small enough to be ignorable in this case. The "ID" part of IID seems shakier in this case study than the "I" part, but I'm going to go ahead and assume IID, which is the standard assumption in measurement error models. Figure 10.2 summarizes the model so far for $n = 1$.

The last thing to specify in the sample part of the model is which summary of the $n$ data values in the sample you're interested in. In the models of Chapter 9 we focused on the sum of the IID observations sampled from the population, because the sum kept track automatically for us of the main quantity of interest—your net gain after $n$ gambles at Keno. Here we need to think about how you would summarize $n$ blood sugar readings to come up with a good guess for the true blood sugar from them. There seem to be two questions:

\[3\text{[Anecdote to be supplied later]}\]
• Each reading is itself a guess for the true value—what’s to guarantee that some summary of the \( n \) readings is better than any one of them?

• Given that several readings are better than one, what is the best way to summarize them?

In the case study I said without justification that you might take the average of the readings, but why is this a good idea? The basic measurement error equation (10.3) helps to explain why several readings are better than one, why the average is a good summary, and when averaging might not help much after all.

**Why Taking Several Measurements and Averaging Them Is a Good Idea.** I said awhile ago that equation (10.3) isn’t actually one equation, it’s really \( n \) equations, one for each observation. When you write down all these equations in words you get

\[
\begin{align*}
\text{observation 1} & = \text{true value} + \text{bias} + \text{random error 1}, \\
\text{observation 2} & = \text{true value} + \text{bias} + \text{random error 2}, \\
& \quad \vdots \\
\text{observation } n & = \text{true value} + \text{bias} + \text{random error } n.
\end{align*}
\]

Imagine averaging all these equations together. The left-hand side is easy—it would just be the average of the \( n \) observations. How about the right-hand side? I guess it would have three parts added together, one for the true value, the bias, and the random errors. The first two of these terms are similar: For example, what do you get when you average the true value with itself over and over again \( n \) times? Adding a constant to itself \( n \) times and dividing by \( n \) just gives you back the constant (check this if it’s not familiar to you), so evidently the answer is just the true value itself. The same thing happens with the bias—averaging it with itself gives it back again. The final term must just be the average of the \( n \) random errors. Putting this all together gives

\[
\left( \frac{\text{average of the}}{\text{n observations}} \right) = \text{true value} + \text{bias} + \left( \frac{\text{average of the}}{\text{n random errors}} \right).
\]

\[10.5\]
Okay, so what? Well, although it might not be obvious, something good just happened to the random errors. Since they fluctuate around 0, averaging them involves adding a bunch of numbers, some of which are positive and some negative, and what happens when you do that is that the positive and negative numbers tend to cancel each other, leaving the average closer to 0 than the numbers going into the average. The punchline is that

Averaging tends to damp down the influence of random errors on the measuring process, so that the average of a bunch of measurements tends to be closer to the truth than any one reading.

There is another way to say this using language that might be familiar to you if you ever went shopping for a good stereo receiver or FM tuner. An electrical engineer might look at the basic measurement equation with no bias,

\[
\text{observed value} = \text{true value} + \text{random error},
\]

and say, “Oh, you mean

\[
\text{data} = \text{signal} + \text{noise}.
\]  \hspace{1cm} (10.6)

The job of the FM tuner in your stereo receiver is to take in radio waves that are a mixture of \textit{signal} (what the FM station you’re trying to tune in actually broadcast) and \textit{noise} (static and other stuff that gets in the way of clear reception of the signal), and separate the signal from the noise. The analogy with unbiased measurement processes is pretty direct—the signal is the true value, and the noise is the random errors. Taking a bunch of measurements and averaging them tends to filter out the noise, allowing us to concentrate on the true value.

This is an idea with frequent implications for daily life. To take an example close to your own interests: if you are a student trying both to learn and to establish a good written record of your work (I bet that’s true), and if I’m teaching a class you’re taking and (as usual) I’m asked to grade you, I can get a much better idea of your “true” achievement in the class by measuring it with a variety of tests, homework assignments, and quizzes than by basing the evaluation all on one big final exam or paper. It may involve more work to sign up for classes with multiple tests or papers, but you will probably get a better chance in them to
accurately demonstrate your achievement than you will in the classes that rise or fall on the final exam alone.

As a second example, say I've moved into a new neighborhood and I'm trying to decide which of two stores is cheaper to do my grocery shopping at. I hit on the idea of making a list of typical things I buy all the time, purchasing everything on the list at each store within a fairly short period of time (like a week), and comparing. But, given that prices on common items at grocery stores fluctuate, to some extent haphazardly, as a result of the timing of discounts and promotions, which is better—to do the comparison once and be done with it, or to do it several times (a few months apart, say) and average the results? You tell me.

Table 10.2: A small simulation illustrating how to damp down random errors by averaging them

<table>
<thead>
<tr>
<th>Random Error Number</th>
<th>Simulation 1</th>
<th>Simulation 2</th>
<th>Simulation 3</th>
<th>Simulation 4</th>
<th>Simulation 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+6.5</td>
<td>-0.8</td>
<td>-13.7</td>
<td>+12.5</td>
<td>+18.1</td>
</tr>
<tr>
<td>2</td>
<td>-4.2</td>
<td>+2.3</td>
<td>+11.6</td>
<td>-6.2</td>
<td>-4.1</td>
</tr>
<tr>
<td>3</td>
<td>+2.1</td>
<td>+8.3</td>
<td>+2.3</td>
<td>+4.2</td>
<td>-1.4</td>
</tr>
<tr>
<td>4</td>
<td>-5.2</td>
<td>-11.5</td>
<td>-18.5</td>
<td>-1.3</td>
<td>+22.2</td>
</tr>
<tr>
<td>5</td>
<td>+3.4</td>
<td>-4.2</td>
<td>-3.7</td>
<td>-11.0</td>
<td>+20.2</td>
</tr>
<tr>
<td>6</td>
<td>+15.1</td>
<td>-1.5</td>
<td>-13.0</td>
<td>+9.0</td>
<td>+7.6</td>
</tr>
<tr>
<td>7</td>
<td>-16.2</td>
<td>-15.3</td>
<td>+3.2</td>
<td>-9.7</td>
<td>+2.4</td>
</tr>
<tr>
<td>8</td>
<td>+17.6</td>
<td>+3.2</td>
<td>+2.7</td>
<td>+3.6</td>
<td>+1.0</td>
</tr>
<tr>
<td>9</td>
<td>-3.1</td>
<td>+6.2</td>
<td>-13.3</td>
<td>+3.0</td>
<td>-5.3</td>
</tr>
<tr>
<td>10</td>
<td>+5.9</td>
<td>+1.6</td>
<td>+10.2</td>
<td>-17.3</td>
<td>+3.1</td>
</tr>
<tr>
<td>11</td>
<td>+9.0</td>
<td>-2.9</td>
<td>-10.2</td>
<td>+0.9</td>
<td>-0.4</td>
</tr>
<tr>
<td>12</td>
<td>-11.7</td>
<td>-12.6</td>
<td>+6.8</td>
<td>-3.1</td>
<td>+9.1</td>
</tr>
<tr>
<td>13</td>
<td>+8.0</td>
<td>+7.3</td>
<td>-2.9</td>
<td>-3.7</td>
<td>-16.2</td>
</tr>
<tr>
<td>14</td>
<td>-3.3</td>
<td>-1.1</td>
<td>+16.4</td>
<td>+14.9</td>
<td>+0.4</td>
</tr>
<tr>
<td>15</td>
<td>+3.6</td>
<td>-11.4</td>
<td>-10.8</td>
<td>+8.7</td>
<td>+4.2</td>
</tr>
<tr>
<td>Mean</td>
<td>+1.8</td>
<td>-2.2</td>
<td>-2.2</td>
<td>+0.3</td>
<td>+4.1</td>
</tr>
</tbody>
</table>

Table 10.2 illustrates the idea of damping down random errors by averaging them. In each of the five columns in the table I used a
computer to conduct a little simulation, by drawing 15 hypothetical blood sugar errors at random with replacement from a population with mean 0 and SD 10. You can see that the individual simulated errors do indeed fluctuate around a value near 0 (the mean of all 75 draws is 0.4), with a give-or-take of about 10 (the SD of the 75 draws is 9.5). However, notice that the column averages, each based on 15 errors, also fluctuate around a value near 0 (their mean is also 0.4), but the fluctuations are much smaller—the SD of the five column averages is only 2.7. Because of cancellation of positive and negative values, the contribution of random error to the average of 15 measurements is likely to be a lot smaller than the typical size of the random errors hidden inside the individual readings.

Does taking repeated observations and averaging them always produce a better answer? Well, yes, but sometimes only up to a point. Look back at equation (10.5), whose right-hand side had two components in addition to the true value: the average of the \( n \) random errors and the bias. I have argued (qualitatively, so far) that the average of \( n \) random errors is likely to be closer to 0 than any single random error, which means that we can make the random component on the right-hand side of equation (10.5) arbitrarily small just by getting more and more data, but in so doing we will be closing in on the truth only if there is no bias in the measuring process. This reinforces an idea I mentioned back in Chapter 6: with a biased method of gathering information you cannot make the bias go away just by getting a lot of data—the bias will shine right through.

A Few More Words About Bias. Now is as good a time as any to talk a bit more about bias, which is the most unpleasant part of the measurement error story. The least attractive thing about bias is contained in the answer to the following question: Is there any way to look just at the measurements themselves and tell if they have bias hidden inside them? (Think about the person weighing fish for you at the market—if you didn’t see him or her put a thumb on the scale, could you tell that the alleged weight was too high just by looking at the measurement itself?) Equation (10.3) says no: the true value and the bias are inextricably bound together; or (to use a term from Chapter 5) confounded, in the sum (true value + bias). There’s no way to separate the terms in this sum by looking only at the data.
The only way to detect and adjust for bias is to appeal to an external standard—that is, to bring in an independent assessment of the truth.

In the fish example, you might be able to tell that the weighing was biased on the high side if, from past experience buying fish at other markets, you knew what a pound of fish was supposed to look like, but it would be hard visually to know how big a bias adjustment to make. In science people often calibrate new measuring instruments by repeatedly measuring something whose true value is already “known” and comparing the average of these readings with the reference value (to see why I put “known” in quotes, consider Problem 10.3). Without an effort along these lines, your measuring instrument could cheerfully be giving you the wrong answer and you wouldn’t know it. Fortunately we are told in the case study that the lab has been shown in the past to produce more or less unbiased blood sugar readings, so we don’t need to do anything special about bias here.

10.2 The Expected Value, Standard Error, and Histogram of an Average

Okay, we’ve got the population and the sample, and we’ve hit on the average of the sampled measurements as the summary of interest. The case study asked us to work out the chance of an incorrect diagnosis of hypoglycemia based first on a single blood sugar reading, which is like a sample of size $n = 1$, and then to compute the same chance based on the average of $n = 3$ readings. Notice that the average of a single number is itself, so that we can speak in either case of basing the diagnosis on the average of $n$ readings (with $n$ either 1 or 3). Figure 10.2 lays out the model for $n = 1$; Figure 10.3 does the same for $n = 3$.

Now under what conditions would the blood sugar measurement process described in the case study incorrectly label you as hypoglycemic? I guess whenever by chance the average of the $n$ readings (let’s as usual call it $\bar{Y}$) happens to fall below 60 mg/dl. Symbolically we’re wondering about $P(\bar{Y} < 60)$, which is a lot like wondering (as we did last chapter) about $P(S > 0)$, the chance that the sum of the gambles in Keno would come out positive. Recall that in Chapter 9 we answered
such questions by constructing the third part of the model diagram—the imaginary dataset. It looks like we need to do that again here.

The Imaginary Dataset in Measurement Error Models. The story is a lot like it was in Chapter 9, except that we're concentrating on the sample average rather than the sum. As the middle and right-hand side of Figure 10.2 show, we need to repeatedly imagine drawing an IID sample of size \( n \) from the population, taking its average \( \bar{Y} \), and writing the \( Y \)'s down in the imaginary dataset. The probability we want, \( P(\bar{Y} < 60) \), is then just the relative frequency with which the \( Y \)'s in this dataset fall below 60. As in Chapter 9, it will be useful in approximating this relative frequency to work out the mean, SD, and histogram of the \( Y \)'s in the imaginary dataset. Recall that we had special names for the mean and SD in this part of the model—the mean was called the expected value and the SD was called the standard error of whatever it is we're keeping track of in the imaginary dataset (in this case the average \( \bar{Y} \)). Evidently to finish the case study we need to work out the expected value and standard error of an average of a bunch of IID draws, and then we need to think about what the histogram of such an average might look like.

The Expected Value of an Average Under IID Sampling. This shouldn't be too hard to figure out, since we already know the answer for sums, and averages are so closely related to sums. As in Chapter 9 we could let \( S \) stand for the sum \( Y_1 + \ldots + Y_n \) of the \( n \) IID draws in the sample and recall that the expected value of \( S \) is the number of draws times the population mean:

\[
E(S) = E_{\text{IID}}(S) = n\mu.
\]

Now if I told you the sum of the draws, to get their average you'd just divide by how many draws there were: \( \bar{Y} = \frac{S}{n} \). And if the world were a nice place, doing something simple like dividing a random quantity by a constant should divide its expected value by the constant (think about it: if \( X \) varies randomly around 10, then \( \frac{X}{2} \) should certainly vary randomly around 5). As I mentioned in one of the math interludes in Chapter 9, this simple and plausible rule is in fact correct: if \( c \) is a constant then \( E(cX) = cE(X) \). So I guess the expected value of the
average comes out like this:

\[ E(\bar{Y}) = E_{IID}(\bar{Y}) = E\left(\frac{S}{n}\right) = \frac{E(S)}{n} = \frac{n\mu}{n} = \mu. \]

To summarize,

\[ E(\bar{Y}) = E_{IID}(\bar{Y}) = \mu. \quad (10.7) \]

In words, the expected value of the average of a bunch of IID draws from a population is just the population mean \( \mu \).

In the case study \( \mu \), your true blood sugar level, is 75, so what we have learned is that with either \( n = 1 \) or \( n = 3 \) the mean of the imaginary dataset is 75. This is good news—you would like the \( \bar{Y} \)'s in the imaginary dataset to fluctuate around the true value, and this says that they do. What makes it work in this case is two things: the blood sugar measurements were taken IID, which is a good way to gather data, and the measuring process was unbiased, so that the value the \( \bar{Y} \)'s fluctuate around is not (true value + bias) (see the discussion on bias above) but the truth itself. Another way to put all of this is that

\begin{center}
On average, across independent repetitions of the experiment of gathering \( n \) observations and taking their mean \( \bar{Y} \), with IID sampling and no bias you get the right answer.
\end{center}

This doesn't guarantee that every sample will give you the true value exactly—far from it—and it's not clear how much comfort we should take from this fact given that in practice we are typically only going to take one sample, but you could think of it as a sort of reality check: if it had not been true, after assuming that the measuring process was both IID and unbiased, there would have been something wrong.

Notice that there is nothing in equation (10.7) about \( n \)—evidently with IID sampling and no bias, one reading is as good as 10 or 100 as far as getting the right answer on average is concerned. How close you typically get to the right answer each time you take one or 10 or 100 measurements is another story—as we have already noted qualitatively (and as we will soon see quantitatively, as well), the average of 10 readings will usually get you closer to the truth than any single reading, and 100 readings would be even better. To see how much better we
need to think about the standard error of $\bar{Y}$—the SD of the $\bar{Y}$’s in the imaginary dataset.

The Standard Error of an Average Under IID Sampling. This should also be easy to calculate, again because we already know the standard error of the sum of a bunch of IID draws and the average is so simply related to the sum. Recall from Chapter 9 that the standard error of the sum is just the square root of the number of draws times the population SD:

$$SE(S) = SE_{\text{IID}}(S) = \sigma \sqrt{n}.$$ 

Once again, the average $\bar{Y}$ of $n$ draws is just the sum $S$ divided by $n$: $\bar{Y} = \frac{S}{n}$. What should happen to the standard error of something random if you divide it by a constant ($c$, say)? Since the standard error is a kind of SD, we could equivalently wonder what happens to the SD of something when you divide it by $c$.

- **Q:** If $X$ is varying randomly around 10 with an SD (a give-or-take) of 6, then $\frac{X}{2}$ should vary randomly around 5, give or take what?

- **A:** I guess if its give or take is 6, when you divide it by 2 that should divide the give or take by 2.

This is another one of those if-the-world-were-a-nice-place arguments that turns out to be right. As I mentioned in a math interlude last chapter, $SE(cX) = |c|SE(X)$—in other words, multiplying or dividing a random thing by a constant multiplies or divides its SD (or SE) by the (absolute value of the) constant. Applying this to the standard error of $\bar{Y}$ gives the following equation:

$$SE(\bar{Y}) = SE_{\text{IID}}(\bar{Y}) = SE\left(\frac{S}{n}\right) = \frac{SE(S)}{n} = \frac{\sigma \sqrt{n}}{n} = \frac{\sigma}{\sqrt{n}}.$$

To summarize,

$$SE(\bar{Y}) = SE_{\text{IID}}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}. \tag{10.8}$$

In words, the standard error of the average of a bunch of IID draws from a population is the population SD divided by the square root of the number of draws.
This may well be the most important equation in the whole book. It has a name—people call it the square root law for averages. Let’s see what it implies in the case study, and then I want to take a few pages to explore its implications more generally.

In the blood sugar example the population SD $\sigma$ is 10 mg/dl. If your doctor takes a single reading on you, we have already seen that it will be some number around 75 (the expected value), but give or take how much? Equation (10.8) says how much: $\frac{\sigma}{\sqrt{n}} = \frac{10}{1} = 10$. This makes sense: the case study said that individual blood sugar readings fluctuate around the true value with an SD of 10, and with $n = 1$ the imaginary dataset just consists of a bunch of hypothetical individual readings. Of course, this still doesn’t say what the chance is of any one of those readings falling below 60, but we’re close to answering that question—to conclude our retracing of the steps in Chapter 9, all we need to know is what shape the histogram of the $\bar{Y}$'s in the imaginary dataset follows, because the chance that $\bar{Y}$ will fall below 60 mg/dl is then just the area under the histogram to the left of 60. Last chapter we saw that the histogram of the sums in the imaginary dataset often follows the normal curve, by the Central Limit Theorem. Will that be true here for the average as well?

The Histogram of an Average Under IID Sampling. This is the final example of how results for the average parallel those for the sum, and it comes down to a question similar to those we answered about the expected value and the standard error: If the histogram of a sum $S$ of $n$ IID draws from a population follows the normal curve pretty well, and the average $\bar{Y}$ of the draws is (as usual) just the sum divided by $n$, will the histogram of the average also follow the normal curve? An equivalent way to ask the question is to wonder what dividing a random quantity by a constant does to the shape of its histogram. By sketching a few examples you can convince yourself of the plausibility of a simple rule, which (like the ones above for expected values and standard errors) turns out to be true in general:

If $X$ is a random quantity and $c$ is a (nonzero) constant, the random quantity $cX$ has the same basic histogram shape as $X$. In particular, if $X$ follows the normal curve then so does a constant times $X$. 

19
This means that the Central Limit Theorem applies just as well to the average of a bunch of IID draws from a population as it does to the sum. In other words,

**Central Limit Theorem for Averages:** As long as the number of draws going into the average is big enough, the histogram of the average of a bunch of IID draws from a population will follow the normal curve pretty well.

This is an extremely useful fact, and we will be making liberal use of it in the rest of the book, but we would be stretching its applicability to the limit if we tried to use it in the case study, where the sample sizes of interest are only \( n = 1 \) and \( n = 3 \). If the population histogram doesn’t look much like the normal curve to begin with, neither will the histogram of the average when the sample size is that small. Fortunately the case study again comes to the rescue: people have kept track of the histogram of blood sugar readings in the past at the lab we’re interested in, and we are told that this population histogram follows the normal curve pretty well. So (as we noted in Chapter 9) the histograms of the average of \( n = 1 \) and \( n = 3 \) draws from this population should also follow the normal curve, and we can finally compute the chance of misdiagnosis.

As Figure 10.2 shows, if your doctor only orders one blood test the histogram of the imaginary dataset will be roughly normal with mean (expected value) 75 and SD (standard error) 10, and the chance of any single reading falling below 60 is just the area to the left of \((60 - 75)/10 = -1.5\) under the standard normal curve, which is a little less than \(7\%\) (about 1 chance in 15). Many people would consider this an unacceptably high risk of misdiagnosis. How much better is it to base things on the average of three readings?

Well, by the same reasoning the histogram of the imaginary dataset with \( n = 3 \) will also be roughly normal with mean 75, but the square root law says that the SD (standard error) of this normal curve will be smaller than it was with only one draw: with \( n = 3 \), \( SE(Y) = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{3}} \approx 5.8 \). The interpretation of this number is that if you were to take 3 IID unbiased blood sugar measurements and average them, the likely amount by which the result would differ from the true blood sugar value is about 5.8 mg/dl. The punchline (as Figure 10.3 shows) is that the
chance of misdiagnosis with the average of 3 measurements is much smaller than it was with one reading—it’s the area to the left of \((60 - 75)/5.8 = -2.6\) under the standard normal curve, which is only about \(0.5\%\) (about 1 chance in 210). By tripling the number of blood sugar measurements, we have cut the chance of incorrectly diagnosing you as hypoglycemic by a factor of about 14 (from roughly 7% to 0.5%). I’d call that a success story for the idea of taking repeated measurements.

To get full value from the idea of replication when measuring something, there are three measurement goals you have to try to achieve: ideally the measurements should be

- Independent,
- Unbiased, and
- Taken under essentially identical conditions.

In practice this is hard work, but the increase in accuracy promised by the square root law (for averages of unbiased IID measurements) makes the effort worthwhile.

Before concluding the chapter with a general discussion of the implications of the square root law, I have a decision-making question for you. This entire case study has focused so far on the benefits of replication—what about its costs?

**Q:** If the lab tests cost $25 each and you had to pay for them out of your own pocket, would the gain in diagnostic accuracy by using three tests instead of one be worth it to you?

This is a hard question to answer—it depends on your valuation, in dollar terms, of the costs of misdiagnosis, which is not an easy thing to do—and (like the question I asked at the end of Chapter 9 about which Keno gamble is best) there is no single “right” answer for everybody. However, difficulties of this type do not make the question any less worth asking. If you do not try to answer questions like this explicitly each time you visit the doctor, he or she will answer them for you implicitly without even asking you. The point is that learning how to think about uncertainty using probability and statistics can help make
you a better consumer, which is something that each of us is called on to do every day. I will return to this example in the chapter on decision-making.

10.3 The Consequences of the Square Root Law

Figure 10.4 is another look at the histograms of the averages in the imaginary datasets in the blood sugar case study with \( n = 1 \) and \( n = 3 \). I have put them on the same horizontal scale to show the square root law in action. Since the measurement process is both IID and unbiased, both histograms are centered at the true blood sugar value (75 mg/dl), but notice that the histogram is more concentrated near the truth when \( n \) is 3 than when it is 1. The square root law for averages says how much more concentrated: when the sample size is three times larger, the amount by which any given \( \bar{Y} \) can be expected to deviate from the truth is \( \sqrt{3} \) times smaller. As we will see in the next few chapters, this implies—in situations where you don’t know the true value—that our uncertainty about the truth after we gather data relevant to estimating it is also \( \sqrt{3} \) times smaller when the sample size goes up by a factor of 3.

Like most things in life, this observation contains both good and bad news: the good news is that getting more data makes your uncertainty about an unknown thing go down, and the bad news is how slowly it goes down. The square root law says that

Under the best of conditions—that is, when the data-gathering process is IID and unbiased—to double the accuracy with which you measure something you have to quadruple the number of measurements.

There are fancier ways to gather data than IID, for which uncertainty may decrease more quickly (stratified sampling, which we talked about briefly back in Chapter 6, is one example), but all of them require more prior information than we made use of in this chapter. Basically uncertainty goes down like the square root of \( n \), although we might wish for it to diminish at a faster rate, and there’s not a damn thing anybody can do about it. This fact is the basis of judgments
about how much data to gather when you’re planning an experiment or a sample survey—we will talk more about the topic of sample size determination in Chapter xx below. The typical result of sample size calculations is that everybody is disappointed at how much money it will cost to decrease their uncertainty to a manageable level, but that should come as no surprise—when was the last time something cost less than you thought it would?

The other main consequence of the square root law to note for now is evident when we write the standard error formulas for sums and averages of IID draws side by side:

\[ SE_{IID}(S) = \sigma \sqrt{n} = \frac{\sigma \sqrt{n}}{1} \quad \text{and} \quad SE_{IID}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}. \]

Both SE formulas involve the population SD \( \sigma \), and in the same way (in the numerator), and the sample size \( n \) enters into both through \( \sqrt{n} \), but the similarity ends there, because the \( \sqrt{n} \) appears in the numerator of the SE formula for sums and in the denominator of the SE formula for averages. This difference in the role played by the \( \sqrt{n} \) makes for qualitatively opposite behavior in the SE’s: as \( n \) goes up, \( SE(\bar{Y}) \) goes down, whereas \( SE(S) \) goes up. This makes sense intuitively—the SE says how variable the sum or average of a bunch of draws would be if you were to observe it over and over, which is another way to say that the SE measures how uncertain we are about the value the sum or average will have on any given occasion, and what the standard error formulas are trying to say is that

| With IID sampling, uncertainty about the average of a bunch of draws from a population goes down with the number \( n \) of values going into the average, whereas uncertainty about the sum of the draws goes up with \( n \). |

If you ever get confused about which standard error formula to use, this rule can help sort it out—ask yourself if it makes more sense for your uncertainty about the thing you’re interested in to go up or down as more numbers go into calculating it. It’s a big mistake to use the SE formula for sums when you’re working with averages, or vice versa.
10.4 Chapter Summary

1. If you measure something more than once under what you think are identical conditions, you may well get a different answer each time. Sometimes that's because what you're measuring changes from day to day or moment to moment, but sometimes it's more reasonable to think of the thing being measured as unchanging—to identify a true value for it—and to appeal to the idea of measurement error to "explain" why you didn't get the same value each time. "Explain" is in quotes because the basic measurement error model described in this chapter doesn't really explain why the measurements differ, it just describes the fact that they do.

2. The basic measurement error equation says that when you measure something unchanging more than once, each observed value can usefully be thought of as a sum of three components:

\[ \text{observed value} = \text{true value} + \text{bias} + \text{random error} . \]

(Symbolically, \( Y_i = \theta + b + e_i, i = 1, \ldots, n. \)) The bias term represents a systematic kind of error in your measuring process—a consistent tendency to over- or underestimate the truth. By definition in this equation, if you were to figure out the true value and the bias and subtract them from your observed values, what's left would be called random errors.

3. The idea behind random errors is that they're haphazard, with no pattern or trend to them, and that some of them are positive and some negative, in such a way that they balance out to zero on average. They arise for a variety of reasons, including human error (for example, transposing digits when writing numbers down), small variations in the measuring process, and misreading your measuring instruments. Notice, however, that nothing logically forces the quantities (observed value - true value - bias) to be "random"—their "randomness" is an assumption on our part, not a law of nature. We will talk in Chapter xx about how to check this assumption.

4. The appearance of the random errors in the basic measurement equation permits us to draw conclusions about future measurements and about the true value by building probability models like those in Chapter 9. As usual these models have three parts, all of which can
be thought of as datasets: the population, the sample, and the imaginary dataset. All three of these components differ somewhat from their counterparts last chapter.

5. The population in measurement error models is just the conceptual set of all possible measurements you might get any time you make a single observation. Usually this population is conceptually infinite, because—at least in theory, with better and better measuring equipment—all values continuously along the number line, in the relevant range for the outcome of interest, are possible. You can get some idea of what is reasonable to assume about the shape of the population histogram by looking at measurements taken in the past with the same measuring process. People sometimes, but not always, find that such a histogram follows the normal curve pretty well; when it doesn’t it’s usually because of a small number of outliers. We will look at how to deal with outliers in measurement situations in Chapter xx.

6. The sample in measurement error models just represents the raw readings you have taken, or are about to take, on the unknown true value. Three principles govern how to take the readings—that is, how to draw the sample—for greatest accuracy: ideally you would like the measurements to be

- Independent of each other, so that you get a new piece of information about the true value each time;
- Taken under essentially identical conditions, so that nothing confounds your assessment of the truth each time; and
- Unbiased, so that you really are measuring what you think you are.

The first two of these goals together amount to saying that the sampling should be IID from the population of potential measurements; the third goal makes the population mean and the true value coincide [say why; math interlude on \( \mu = \theta + b \). Also, comment on how weird the model is: breaking down something observed into the sum of three unobserved things, etc.].

7. Bias in a measuring process is a slippery thing to quantify, because there is no way to tell just by looking at your measurements if they have bias hidden inside them. The only way to detect and adjust for bias is to appeal to an external standard—to bring in an independent assessment of the truth.

25
8. A good way to summarize an IID unbiased sample of \( n \) measurements is by taking their average \( \bar{Y} \). Averaging tends to damp down the influence of random errors on the measurement process—to filter out the noise, so to speak—so that the average of a bunch of measurements tends to be closer to the truth than any single reading.

9. Imagine taking an IID sample of size \( n \) from the measurement population and computing the sample mean \( \bar{Y} \). Now imagine doing this over and over a whole lot of times. The imaginary dataset in measurement error models collects together the \( \bar{Y} \)'s you might get if you did this. You can answer questions about the chance of any given \( \bar{Y} \) coming out in some specified way—for instance, the chance of misdiagnosis in the hypoglycemia case study, \( P(\bar{Y} < 60 \text{ mg/dl}) \)—by working out the relative frequency with which the \( \bar{Y} \)'s in the imaginary dataset come out that way.

10. As in Chapter 9, such relative frequencies may not be all that easy to compute exactly, but you can often approximate them with the same four-step program we worked out last chapter: figure out the longrun mean (expected value), SD (standard error), and histogram of the imaginary dataset, and approximate the relative frequency you’re interested in by the appropriate area under the histogram.

11. The expected value of the average of \( n \) IID draws from a population with mean \( \mu \) equals \( \mu \) (symbolically, \( E(\bar{Y}) = E_{IID}(\bar{Y}) = \mu \)). This is true for any population, not just measurement populations. In general in measurement error models, the population mean \( \mu \) is the sum of the true value and the bias in the measuring process; in the special case in which the process is unbiased, \( \mu \) equals the true value. What this means is that on average across repetitions of the experiment of gathering \( n \) observations and taking their mean \( \bar{Y} \), with IID sampling and no bias you get the right answer. This does not mean you will get the true value exactly every time—far from it, in fact.

12. The standard error of the average of \( n \) IID draws from a population with SD \( \sigma \) equals the population SD divided by the square root of the number of draws (symbolically, \( SE(\bar{Y}) = SE_{IID}(\bar{Y}) = \frac{\sigma}{\sqrt{n}} \)). This is called the square root law for averages. Its interpretation is that with IID sampling and no bias, the SE of the average, \( \frac{\sigma}{\sqrt{n}} \), is the likely amount by which any given \( \bar{Y} \) will differ from the truth.
13. As long as the number \( n \) of draws going into the average is big enough, the histogram of the average of \( n \) IID draws will follow the normal curve pretty well. In other words, the Central Limit Theorem applies just as well to averages as it does to sums.

14. In words the square root law for averages says that uncertainty about an average goes down at a \( \sqrt{n} \) rate when the measuring process is IID and unbiased. This fact has two main consequences:

   - Under the best of conditions—that is, when the data-gathering process is IID and unbiased—to double the accuracy with which you measure something you have to quadruple the number of measurements.

   - The standard error formula for sums is \( SE_{IID}(S) = \sigma \sqrt{n} \), which has the same ingredients as the SE formula for averages except that the \( \sqrt{n} \) is in the numerator rather than the denominator. So while uncertainty about the average of a bunch of IID draws goes down with \( n \), uncertainty about the sum goes up with \( n \). You can use this fact to keep from getting the two standard error formulas mixed up—ask yourself whether uncertainty about the thing you’re interested in should intuitively go up or down as more data values go into calculating it.

10.5 Problems

1. Show speed of light data; some physicists theorized on the basis of this data that the speed of light was changing over time; show that a simpler explanation is bias in the measuring process.

2. (hard) The basic measurement error model assumes that the observations you make of the unchanging quantity of interest are independent. One way to think about whether two random quantities are independent is to ask if knowing the value of one of them helps you to predict the other—if so, they’re dependent, if not, independent. (Formally, \( X \) and \( Y \) are independent if and only if \( P(Y|X) = P(Y) \).) Think about trying to predict the first measurement \( Y_1 \) your doctor will make of your blood sugar level, and then trying to predict the second reading \( Y_2 \) given that the first one came out, say, 70. Before the first reading you may not have known much about your blood sugar level. Wouldn’t you say that the knowledge that the first reading was 70 would help
you to guess at the second reading? Why is it ok, then, for people to assume independence? Explain briefly.

3. If the only way people know that a measuring process is unbiased is to compare it to a reference measuring process that’s known already to be unbiased, how did they find out that the reference process was unbiased to begin with? [fix this]

4. Returning to Cavendish's measurements of the density of the earth in Table 10.1, there was a complication I didn’t tell you about. After his sixth observation, Cavendish replaced a suspension wire in his measuring equipment by one that was stiffer, and hence (perhaps) likelier to give more accurate readings. Work out the mean of the first 6 measurements and the mean of the last 23, and compare them to the overall mean of all 29 readings and to the currently accepted true value, 5.517g/cm^3. With the wisdom of hindsight, does it look to you as though Cavendish was right to replace the wire? What do you think he should have done with his first six readings—keep them (that is, base his estimate on all 29 data values), discard them, or use them in some other way (how?)? Explain briefly. [also, what about outliers?]

5. (more to come)
Figure 10.2: The probability model for the hyperglycemia case study with $n=1$
Figure 10.4: Histograms of $\bar{y}$ in the hypoglycemia case study with $n = 1$ and $n = 3$. More data makes the histogram concentrate more tightly around the true value.
Chapter 11

Building a Statistical Model

So far in this book, after the introduction, we have talked about three things:

- **Data-gathering**: how to collect information to *reduce your uncertainty* about something of interest—for instance, last chapter's case study on blood sugar, in which we were uncertain about whether you should be treated for hypoglycemia;

- **Descriptive methods**, which are about how to summarize *factuals*—the data you *have*—for instance, "We took three blood sugar readings on you, and they averaged 78 with an SD of 10"; and

- **Probability**, the mathematical language people use to *quantify* uncertainty—for instance, "If your real blood sugar is 70 and we try to diagnose you on the basis of a single reading, there is about a 7% chance of misdiagnosing you as hypoglycemic."

The rest of the book is about *counterfactuals*—things like "What would have happened to that patient over there if she had received the drug instead of the placebo?" or "What would we have gotten if we had taken a complete census of this population instead of a sample from it?" Recall the distinction I made back in Chapter 1: *factuals* are about *what is or was*; *counterfactuals* are about *what might have been or might yet be*. Factuals are about the *data you have*; counterfactuals are about the *data you wish you had*.

In general, in any situation in which you are uncertain about something, there is a body of information that—if you had it—would remove your uncertainty about the unknown thing. Often this body of information can be thought of as a dataset of some kind, which we may as well call the *no-uncertainty dataset*. In assessing your uncertainty about the unknown thing it is usually a good first step to try to *visualize the no-uncertainty dataset*, and to compare it with the dataset you have, which last paragraph I called the *factual data*. 
If these two datasets coincide, you’re in luck—you have no uncertainty about the unknown thing. If they don’t, any data in the no-uncertainty dataset that’s missing from the factual dataset is data you wish you had, because if you had it—together with the factual data—your uncertainty would vanish. Awhile ago I called the data you wish you had the counterfactual dataset. Evidently with this way of looking at things

\[ \text{factual dataset } + \text{ counterfactual dataset } = \text{ no-uncertainty dataset .} \quad (11.1) \]

What we will be doing for the rest of the book is trying to use the information in the factual dataset to guess at the no-uncertainty dataset. To do this you have to be able to say how the data you have—the factual—relates to the data you wish you had—the counterfactual. That’s where statistical models come in.

A statistical model is a mathematical story that relates the data you have to the data you wish you had.

This chapter is about how to build such models. In day-to-day applied work they come in many forms; we are going to talk here about the two most basic statistical models—sampling models and causal inference models. As a last bit of stage-setting for this discussion, I need to say some more about the three types of counterfactuals that come up most frequently in science and decision-making, and then we can look at the models.

Inference and Prediction: The Three Types of Counterfactuals.
Consider these three situations:

- **Situation 1.** I am interested in the quality of hospital care received by the entire population of Medicare patients admitted to U.S. hospitals in 1992 with a diagnosis of heart attack. There were about 300,000 of these patients. I can’t afford to assess the quality of care they all received, so I have taken a sample of (say) 4,000 of them and measured the quality of care given to the sampled patients. The *factual* here—the data I have—is the dataset of 4,000 quality of care numbers for the sampled patients, or maybe just a summary of these 4,000 numbers like their mean. The *counterfactual* here—the data I wish I had, to go along with what I actually do have—is the dataset of the 296,000 quality of care numbers for the unsampled patients, or (again) a summary of these 296,000 numbers like their mean. As in equation (11.1) above, if you merged these two datasets—factual and counterfactual—together, you would have the whole population—the no-uncertainty dataset—and you could answer any population-based questions that might occur to you with certainty.
• Situation 2. I have a drug that I hope will be effective in lowering the blood pressure of hypertensive patients. To measure its effect I find (say) 120 hypertensive people who agree to take part in a study, and randomize them, 60 to a treatment group (who get the new drug) and 60 to a control group (who get a placebo). I measure their blood pressures before and after (say) 12 weeks on the drug or placebo, and compute the change (after − before) for each person.

Consider somebody in the treatment group. As we talked about back in Chapter 5, the factual for this person—the data I have on him or her—is how much his/her blood pressure changed in 12 weeks on the drug. The counterfactual for this person—the data I wish I had on him or her—is how much his/her blood pressure would have changed in 12 weeks if I had given him/her the placebo instead of the drug, and everything else about the person that is relevant to blood pressure had stayed the same. You could write out a similar statement for somebody in the control group: the factual for such a person is his/her change in blood pressure having taken the placebo, and the counterfactual of interest is how much that same person would have changed if instead he or she had taken the drug.

Applying equation (11.1) to this setup, if you were to merge the factual and counterfactual datasets here, you would have the no-uncertainty dataset, which would include for each person both his or her blood pressure outcome having taken the drug and the corresponding outcome under the placebo. With this dataset you could determine with certainty for each person what effect the drug caused on blood pressure—for instance, if the two numbers were equal for somebody, the drug had no effect for that person. After looking at the no-uncertainty dataset here, we might still be uncertain about the effect of the drug on people other than the 120 men and women who took part in the experiment, but we wouldn’t have any uncertainty about the drug’s effect on those 120 people.

• Situation 3. I am trying to make a business decision that depends on the wholesale price of gasoline in Los Angeles next year. I have weekly data on gasoline prices in the past over a 30-year period (say), and evidently my task in making the decision sensibly is to use these past data—and any other information I have, judgmental or otherwise—to forecast next year’s weekly gasoline prices, or a summary of them like their average. The factual here—the data I have—is the gasoline prices in the past, and any other relevant historical information. The counterfactual here—the data I wish I had—is next year’s weekly prices. In this case, if I merged the factual and counterfactual datasets the
resulting no-uncertainty dataset would be the entire 31-year gasoline price time series, and if I could see this larger dataset I'd be able both to characterize the good and bad decisions in the past and to identify the right thing to do in the present.

These three situations are all examples of factuals and counterfactuals, but they are different enough to merit different names. In motivating these names I want to return to the discussion I started back in Chapter 7 on deduction (reasoning from the whole to the part) and induction (reasoning from the part to the whole). If we identify the "part" in this case as the factual dataset, the data we have, and the "whole" as the no-uncertainty dataset, the data we have + the data we wish we had, all three situations above are instances of reasoning from the part to the whole—in other words, they're all instances of induction.

So far, no differences. However, the third example involves the element of time—trying to guess into the future—whereas the other two do not. For this reason many people make a distinction between the first two situations on the one hand—they're referred to as examples of inference—and the third situation on the other, which is referred to as an example of prediction. The distinction is not hard and fast—you could argue, for instance, that all three of them basically involve trying to predict something, and this is in fact a very useful way to look at things, but for now I will stick with the traditional labels. In Chapter xx I will have more to say about the value of reinterpreting inferential situations in predictive terms.

I still need to say what separates the first two situations, both of which involve inference. Well, in the first situation, if I gave you a chance to observe all 300,000 quality of care numbers in the population and you said that all you were really interested in to begin with was a summary of those 300,000 numbers like their mean, you would be admitting that there was a single number of particular interest to you that you'd be willing to settle for, and you'd be in good company—people often just focus on the mean of a large population of numbers. Unlikely numerical population summaries like the mean are called parameters, and your attempt to guess at the population mean on the basis of a sample is an example of parametric inference. The second situation is quite different—we are trying to figure out what effect on blood pressure the drug causes, so people call this an example of causal inference.

So, okay, parametric inference, causal inference, and prediction. In this chapter I will talk about how to build statistical models to attempt the first two of these three forms of induction—sampling models for parametric inference, and causal inference models for figuring out the effects caused by an intervention you're interested in. Later on, in Chapter xx, we will take up models for prediction. In the next few sections I will try to make
the differences between the three kinds of counterfactuals a bit plainer, and there are some problems at the end of the chapter you can work through to make sure the distinctions are clear.

11.1 Sampling Models For Parametric Inference

The setup with sampling models is like Situation 1 above: there's a population of interest to you—the no-uncertainty dataset—and you either can't afford to take a complete census of it or you've figured out that it would be wasteful to do so, because information from a well-chosen sample will provide an accurate enough answer for your purposes. So you take a sample—which creates the factual dataset—with the hope that the sampled units will allow you to guess at the unsampled units—the counterfactual dataset. Let's look at a more in-depth example than Situation 1.

<table>
<thead>
<tr>
<th>Case Study 11 (business): The C&amp;O Freight Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>How does stuff produced in one place in the country get to consumers somewhere else? One of the main ways is by rail. In the railroad business, most things need to be shipped such a long distance that a given shipment will travel over several different railroads to get there. In cases like this, the total freight charge paid by the shipper is divided among all participating railroads. It is expensive and time-consuming to work out each railroad's share of the revenue when the freight charge is split in this way.</td>
</tr>
<tr>
<td>Back in the 1950s an Eastern railroad called the Chesapeake and Ohio (C&amp;O) was looking around for methods that would cut down on these auditing expenses. They decided to perform some experiments to see if sampling would save them money and still produce accurate results. In one of these experiments, they studied the division of revenue between the C&amp;O and another railroad, for all freight shipments of a particular kind traveling over a part of their rail lines called the Pere Marquette district during a six-month period. Information needed to compute the amounts of money the other railroad owed the C&amp;O was written on documents called waybills, one for each shipment. There were 22,984 waybills in all.</td>
</tr>
<tr>
<td>The C&amp;O did two things in this experiment: they took a simple random sample of 5,334 waybills (about a 23% sample), and used the amounts of money on the sampled waybills to estimate the average amount of money owed them on all 22,984 waybills; and (independently, as a kind of validity check) they did a complete audit of all 22,984 waybills so that they would know how close the sample estimate came to the truth. The average amount of money on the sampled waybills was $74.54 (I have re-expressed the amounts in 1992 dollars; the actual amounts in 1952 dollars sound laughably small), with an SD of $137.22. Set up a statistical model for this situation that displays the factual and counterfactual datasets and how they are related.</td>
</tr>
</tbody>
</table>

We will finish this case study next chapter, by talking about how to
infer backwards from the sample to the population; for now let’s just set up the model. The samples-and-populations story here should remind you of the probability models we constructed in Chapters 9 and 10 for Keno and hypoglycemia. There’s a good reason for that—the setup is almost identical, with one main difference: in probability models you know the population and you’re trying to figure out what the sample will look like when you take it, whereas in statistical models you know the sample and you’re trying to figure out what the population looks like.

Statistical sampling models are a lot like the probability models of Chapters 9 and 10, except that we are trying to draw conclusions in the opposite order. Probability is *deductive*—reasoning from the population to the sample; statistics is *inductive*—reasoning from the sample to the population.

Figure 11.1 gives a statistical sampling model for the setup in the C&O case study. The population of interest (the no-uncertainty dataset) is all 22,984 waybills, so the population size \( N \) is 22,984. The variable of interest, measured on each waybill, is the amount of money owed to the C&O railroad as a result of the shipping transaction documented by the waybill. The population summary of interest is the mean \( \mu \) of all these 22,984 dollar amounts, because if we knew \( \mu \) we could multiply it by 22,984 to get the bottom-line business number the C&O really wants to know, the total amount of money owed to them. The sampling was done by SRS, with a sample size of \( n = 5,334 \). The sample mean \( \bar{Y} \) came out \$74.54.

As in all sampling models, the *factual* dataset is the sample; I have emphasized this by filling in (hypothetical) known data values in the sample, consistent with the summary information that the sample mean was \$74.54 and the sample SD came out \( s = \$137.22 \). The *counterfactual* dataset is the dollar amounts on the 22,984 – 5,334 = 17,650 unsampled waybills in the population. I have noted this by putting question marks where an unsampled waybill appears, and by also putting a question mark next to \( \mu \), the unknown *parameter* of interest. We have no uncertainty about the factual dataset, but we are uncertain about the values in the counterfactual dataset, and the question marks symbolize this uncertainty. The idea will be to use the data in the sample to reduce our uncertainty about the unsampled population units and about the overall mean \( \mu \). Notice that with this way of looking at the situation the problem could be described as one of missing data: the dollar amounts on the sampled waybills are observed, but the dollar amounts on the unsampled waybills are missing.

That’s about all there is to sampling models based on SRS. IID models are identical except that you write “IID” on the arrow from the population to the sample. The model in Figure 11.1 has a finite value of \( N \), the popu-
lation size—recall in Chapter 9 I called such setups finite population models. There’s no real difference in conceptually infinite models when \( N = \infty \), as (for instance) in the blood sugar case study in Chapter 10. Figure 11.2 shows a sampling model for this setup in the more realistic situation in which (a) we don’t know your actual hypoglycemia status, (b) your doctor has taken three blood sugar readings on you to find out, and (c) we are trying to infer from those readings whether you do in fact have hypoglycemia. You might compare this picture with Figure 10.3—in Figure 11.2 the sample is known and we are trying to reason backward to the population, whereas in Figure 10.3 we pretended the population was known in order to reason forward from it to the probabilities of various things happening in the sample.

In Figure 11.2 the factual dataset is the three blood sugar values your doctor measured on you, which average \( \bar{Y} = 78 \text{ mg/ml} \) with an SD of \( s = 10 \text{ mg/ml} \), and the counterfactual dataset is all the other blood sugar readings she could have gotten if she had measured you over and over again a lot more than three times. With this way of setting things up the no-uncertainty dataset you would get by merging the factual and counterfactual data would be all possible blood sugar readings on you at the time the three observed readings were taken. Because we are assuming (as we did in Chapter 10) that the measuring process is IID and unbiased, the mean \( \mu \) of the no-uncertainty dataset would be your true blood sugar value by definition, so again this is an inference problem—to use the mean \( \bar{Y} \) in the data to guess at the truth \( \mu \) in the conceptual population of all possible blood sugar readings.

Notice that imagining this conceptual population is more hypothetical than imagining what the C&O would get if they audited all 22,984 waybills—the former can never be observed, whereas all that keeps us from observing the latter is the cost of making a complete audit and the boredom that would set in while doing so. As I mentioned in Chapter 10, conceptually infinite models—like the ones in Figures 11.2 and 10.3—rest on shakier ground than finite models, like Figure 11.1 or the one for Keno back in Chapter 9. I will have more to say about this in the section below on the nature of randomness, and we will talk a bit in Chapter xx about how to use the observed data values to question the plausibility of conceptually infinite sampling models like the one at the heart of the hypoglycemia case study.

11.2 Causal Inference Models

Okay, so much for sampling models, of the kind we might use in Situation 1 above. The story is quite different for setups like Situation 2, in which people would think of using a causal inference model. As a detailed example of how to build such models, I’d like to revisit Case Study 4 back in Chapter 5—the AIDS experiment (you might want to read it again now). To review the
particulars, some Danish scientists in the late 1980s had a theory that a
drug called isoprinosine (iso for short) might be successful in slowing the
progression to AIDS of people who were HIV-positive. To check this theory
they rounded up 833 HIV-positive people, from various treatment centers in
Scandinavia, who agreed to take part in their study and randomized them,
412 to a treatment group who took iso for 24 weeks, the other 421 to a control
group who got a placebo for the same length of time. They then recorded
for each person whether or not he or she developed AIDS at any time in the
24-week period.

Figure 11.3 presents a causal inference model for this setup. For the
412 treatment people, the factual is whether or not they got AIDS during
the observation period while taking iso, and the counterfactual of interest is
whether or not they would have gotten AIDS during that period if instead
they had taken the placebo, and all other relevant things in their lives had
been held constant. The story is reversed for the 421 control people—for
them the factual is what happened to them while taking the placebo, and
the counterfactual is what would have happened to them if instead they had
taken iso. As in Figures 11.1 and 11.2 I have emphasized what we know
and what we’re uncertain about by putting in hypothetical numbers for the
factual data and question marks for the counterfactual data, coding 1 = {got
AIDS} and 0 = {didn’t get AIDS}.

If we could see both numbers for any one person in Figure 11.3, we
would know the effect caused by taking iso for that person. For example, for
people who either got AIDS under both treatment conditions or under neither
condition, iso would have had no effect (epidemiologists call the (1,1) people
“doomed” and the (0,0) people “immune”). For everybody else iso would
have caused a different AIDS outcome: for the (0,1) people iso would have
inhibited AIDS, and for the (1,0) people iso would have actually encouraged
the disease (we would hope that there wouldn’t be many such people). The
trouble is that we can’t see both numbers for anybody—there isn’t any way to
simultaneously give people iso for 24 weeks and also give them the placebo
during the same period (see problem 11.x for an attempt at designing an
experiment to come as close to this ideal as possible). However, we ought
to be able to use the observed 1’s and 0’s in Figure 11.3 to guess at what
the other number would have been in each pair, which is another example
of inference from observed to unobserved data. In Chapter xx we will talk
about how to quantify the uncertainty documented by the question marks
in Figure 11.3. Notice that, as with the C&O case study, you could think of
the setup in Figure 11.3 as a missing data problem—the question marks
are like missing values that we wish we had observed.

It’s worth taking a moment to compare the causal inference model in
Figure 11.3 with the sampling setup in Figures 11.1 and 11.2. The main dif-
ference is that in Figure 11.3 we are not attempting to infer outward from the 833 patients in the Danish study to, say, the population of all HIV-positive patients. Although we might want very much to generalize the results of the study to everybody with HIV, the Danish scientists did not try to take a representative sample from {everybody with HIV}—they worked with the 833 Scandinavian patients who were available for study and who agreed to take part in the experiment. It would require an assumption, not yet supported by any data, to conclude that {the Scandinavian patients in the study} and {all other HIV-positive patients around the world} were sufficiently similar in relevant ways that the Danish findings provide an unbiased guess for the effects of iso on {everybody with HIV}.

This is not a devastating criticism of the Danish study, it’s just a comment on the scope of its findings. The Danish scientists may well have shown that iso had a definite effect on these 833 patients (recall that the AIDS rate for the iso patients was 0.5%, versus 4.0% for the control group), and this encourages us to think that iso might have a beneficial effect on other groups of HIV-positive patients, but we don’t yet have a good estimate of its effect on the whole population of interest. Biostatisticians refer to experiments like the one in Case Study 4, in which the investigators try to document what’s going on causally for the patients in the study without trying to generalize outward, as Stage I clinical trials [is this terminology right?], and studies in which an attempt is made to draw conclusions about the entire population of patients with a given disease are instead called Stage II trials. We will look at this distinction again in Chapter xx.

11.3 The Role of Randomization, and the Nature of Randomness

I said at the beginning of the chapter that the point of a statistical model was to relate the data you have to the data you wish you had. So how exactly are the factual and counterfactual datasets related in the examples we’ve looked at so far?

Let’s do the C&O freight study first—the question there becomes how the sampled and unsampled waybills relate to each other. To answer this question we have to think about the sampling mechanism. To infer from the sampled to the unsampled waybills we would like the unsampled ones to be similar in all relevant ways to the sampled ones, which here means similar on the outcome of interest—dollars owed to the C&O railroad—and anything that would be strongly predictive of that outcome—in other words, similar on any PCFs as well. Taking a simple random sample of 5,334 of the 22,984 waybills, as they did in Case Study 11, should arrange this automatically,
although as usual there are no guarantees—we will see next chapter that even with such a big sample it is possible for the sample estimate of the total amount of money owed to the C&O across all 22,984 waybills to be off by thousands of dollars. The point is, though, that the randomization that chose which waybills were sampled acts to make the sampled and unsampled waybills similar, and this makes the inference we’re interested in possible.

What if the sample had not been drawn with SRS—if, say, they had just taken the top 5,334 waybills on the pile? In that case we would be left wondering if the sampled and unsampled waybills really were similar in all relevant ways, which would make it hard to validly infer from the sample to the population of all 22,984 waybills. That’s why using a representative sampling method based on randomization, like SRS, is so useful—it gives people confidence in the inferential results.

Now how about the AIDS case study—what role did randomization play there? Recall from Figure 11.3 that the factual data for a typical treatment person in the experiment was his or her AIDS outcome having taken iso, and the counterfactual data for that same person was what his/her outcome would have been if he/she had instead taken placebo. The situation was reversed for the control people: (factual = outcome under placebo), (counterfactual = outcome under iso). Consider somebody in the treatment group—person A, say—and imagine trying to use the observed data to guess at his/her outcome under placebo. If we could find somebody in the control group—person B, say—who was just like this treatment person with respect to factors strongly predictive of progression from HIV-positive status to AIDS, we could use person B’s control value as a guess for what person A’s control value would have been. Even better, given what we know about measurement error, would be to find a bunch of people in the control group who are just like person A, and use the average of their control values as a guess for what person A’s control value would have been. We could try the reverse of this sort of guesswork as well—using person A’s treatment value (or the average of the treatment values of a lot of people like person A) to guess at the outcome person B would have had under iso.

For this to work well, as we noticed back in Chapter 5, we would evidently like the treatment and control people to be as similar as possible on any PCFs. This is why investigators try to use randomized controlled experiments whenever possible—the hope is that the random assignment of people to treatment and control will make the two groups comparable, so that we can use the observed data as good guesses for the unobserved values in Figure 11.3. I mentioned back in Chapter 5 that randomization is quite likely to succeed in achieving this goal, as long as the number of PCFs is not too big and the number of people being randomized into treatment and control is large. With 833 patients in the AIDS study we would expect the random-
ization to do a nice job of balancing the two groups, giving us confidence that any outcome differences we see were indeed caused by the drug.

Notice, however, that the randomized assignment to treatment and control in the AIDS study did not permit the researchers to generalize outward from their 833 HIV-positive patients, which makes randomization’s role in the AIDS study different from its role in sampling models like the one in the C&O study. [discuss internal versus external validity?] To arrive at valid causal conclusions about the average effect of iso on the entire population of HIV-positive patients worldwide, the researchers would first have had to choose their subjects in a way that was equivalent to SRS from that population, and then they would have needed to insure comparability of the treatment and control groups, for instance by performing a second randomization to allocate the chosen people to iso or placebo.

Causal Inference Models for Observational Data. The AIDS case study is an example of a causal inference model for data gathered by running a controlled experiment. You can build causal inference models for observational data, too, although from what we saw back in Chapter 5 it probably won’t surprise you to hear that it’s harder to draw valid cause-and-effect conclusions from such data.

As an example, American sociologists, political scientists, and others have argued for some time now about the effect, if any, on {the probability that someone will choose to murder somebody else} caused by {the presence or absence of the death penalty for murder in the state in which the murderer lives}. Some people claim that the death penalty has a “chilling effect” on potential murderers; others think it has no effect at all. How would you gather data to reduce your uncertainty about this question?

Well, the ideal no-uncertainty dataset for settling this debate would have one row for each of a bunch of potential murderers, maybe arranged by the state in which they live, and two columns: whether they would commit murder if there were no death penalty in their state, and whether they would do so if instead their state had the death penalty. Nobody has any data like this. A less satisfying but more available dataset that bears on the question is given in Table 11.1 [to be supplied later], which has one row for each state in the U.S. The first column in the table is the state name; the second column is its death penalty status in 19xx; and the third and fourth columns are its murder rate with and without the death penalty, with question marks for the counterfactuals.

These data are definitely observational—nobody has ever run a controlled experiment on the relationship between capital punishment and murder. This fact does not make us any less interested in knowing if capital punishment has a deterrent effect, but look at how hard it is to arrive at valid causal
conclusions from the data in the table. Yes, it's true that you can find some capital punishment states with unusually low murder rates, and some non-capital-punishment states with quite high murder rates—summarize the rates for the two types of states—but is it fair to use the murder rates in the capital-punishment states as guesses for what the murder rates would have been in the states without capital punishment, if instead they had had the death penalty? Of course not—the two groups of states differ in a number of ways that are relevant to their murder rates besides their capital punishment status, for instance their gun control laws and their proportions of people living in urban rather than rural areas. In other words, it's not hard to think up PCFs, and we don't have any randomization to bail us out by balancing the confounds in the two groups. It is possible to build more complicated models than the ones we are going to talk about in the chapters that follow—models that attempt to adjust for the PCFs and permit tentative causal conclusions to be drawn from observational data—but such methods are beyond the scope of this book.⁸

To summarize all of this stuff about the role of randomization in sampling and causal inference models,⁹

In sampling models for data gathered with SRS (or IID) sampling—or some other, more complicated method involving randomized choice of the sampled units, like stratified sampling—the purpose of the randomization is to make the sampled and unsampled units as similar as possible on the outcome of interest and all factors strongly predictive of it, so that you can infer outward from the sample to the whole population. In causal inference models for data gathered with controlled experiments, the purpose of the randomization is similar—to make the treatment and control subjects as similar as possible on all PCFs, so that you can use the observed treatment values as good guesses for what outcomes the control subjects would have had if instead they had received the treatment, and vice versa.

The Nature of Randomness

I'd like to finish the chapter with a few words on what people mean when they say something is random, because the concept is a bit slippery, and there's a distinction that's worth keeping in mind throughout the rest of the book. People tend to use the word "random" in two distinctly different ways in statistical work, as follows.

- In sampling models like the one in Figure 11.1 for the C&O study, we can talk confidently about the data being chosen randomly, because we imposed the randomness on the world ourselves—somebody picked the
sampled waybills "at random" without replacement, by which we mean that they were chosen {haphazardly, but so that they all had roughly an equal chance of being picked}. The same thing applies to the model for Keno back in Chapter 9—each time the game is played, somebody picks the Keno balls "at random"—and to the causal inference model for AIDS in Case Study 4, where the investigators randomized the assignment of patients to treatment and control.

- But in the basic measurement error setup that underlies the model in Figure 11.2 for the hypoglycemia data,

\[
\text{observed value} = \text{true value} + \text{random error}
\]  

(recall that we assumed no bias in this model), all we have done is to declare the discrepancies (observed value – true value) to be "random," by which we mean that (a) we don’t exactly know why the measurements came out different each time, and (b) the observations themselves, and therefore the discrepancies (observed value – true value), look haphazard to us.

You might use the term active randomness to describe situations in which people choose a sample using SRS or IID, or assign subjects to treatment and control in an experiment at random, because in cases like that somebody has actually gone to the trouble of introducing randomness into the world, for instance in the Keno study by using a kind of physical randomization to draw the Keno balls. By contrast the kind of "randomness" in measurement error models might be called judgmental randomness, because all that’s happened is that somebody has looked at the measurements and declared them "random." The difference is between {gathering your data by SRS or IID} on the one hand (active randomness), and {claiming that the data are like an SRS or IID sample} on the other hand (judgmental randomness).

Researchers in science and decision-making use the idea of active randomness to model data gathered with controlled experiments, or surveys based on random sampling mechanisms like SRS, but that doesn’t cover all phenomena involving uncertainty—to which people would like to apply probability and statistics—by a long shot. The great majority of data-gathering people do in their lives is observational in character, and that’s where judgmental randomness comes in.
In everyday life people talk about randomness all the time, in a loose sort of way, to mean anything that looks haphazard. In statistical work it helps to be more precise, by distinguishing between active randomness, which comes up in situations (like randomized controlled experiments) where randomness was actually used to gather the experimental or sampling data, and judgmental randomness, which arises in situations (like those in which people use measurement error models) where the data-gathering was observational and the process under study just looks random.

I said awhile ago that the concept of randomness is slippery. But active randomness is pretty straightforward—you’ve either done some sort of randomization or you haven’t—so I guess it would be more precise to say that judgmental randomness is the elusive idea. In this book we will be working more with data from randomized experiments and surveys than with observational data, because it makes sense to look at the more straightforward case—randomized data—first, and you can only cover so much ground in 600 pages. But we will run into measurement error models—and some other situations involving observational data—from time to time in what follows, so it would be good to get in a couple more points now about judgmental randomness before we go on to the next chapter. The nature of this kind of randomness comes into focus most clearly when you’re trying to predict observational data, as in the following two examples.

• Example 1. Consider this string of numbers:

79821480865132823066470938446095505822317.

Suppose somebody asks you to predict the next one. I don’t know about you, but—at first glance, anyway—these digits look “random” to me: there doesn’t seem to be any pattern to them, and each of the integers from 0 to 9 seems to come up about as often as the others. Based on this assessment, the best you could do in predicting the next one would be to say something like “I think it will be 0 with probability around $\frac{1}{10}$, 1 with probability about $\frac{1}{10}$, and so on up to 9.” But now suppose somebody points out (correctly) that these numbers are just the 100th through 140th digits in the decimal expression for $\pi$, and you can write down a relatively simple formula that generates them exactly. Notice that an interesting thing just happened: Before you knew the formula, the numbers seemed random, but after you knew the formula they weren’t random to you any more—you could predict them exactly. The only thing that changed was your state of knowledge. This means that
One way to think about judgmental randomness is that it's not a property of the world, it's an expression of your imperfect ability to predict something of interest to you. As your knowledge about whatever process you're studying increases, the goal is to reach a point where you no longer need to invoke judgmental randomness—that is, to reach 100% predictive accuracy. At that point you don't need to appeal to randomness anymore.

- **Example 2.** Let's say I'm watching you shoot free throws in a basketball game, and I want to predict whether you'll hit a particular free throw late in the game. I watch you for awhile and notice that (a) you seem to make about 7 out of every 10 shots you take and (b) your pattern of hits and misses looks haphazard. For the sake of prediction I may decide to behave as if the process by which you hit or miss your free throws is like making IID draws from a population with 7 hits and 3 misses in it, but let's not fool ourselves here: as with the digits of \( \pi \) above, this modeling story, which involves judgmental randomness, is not a causal explanation of why you hit or miss any particular free throw. It's just an expression of my incomplete understanding—about what is really causing you to hit or miss each time—that I may find useful in trying to predict what will happen.

Prospectively, before you attempt a given free throw, I may express my uncertainty about the outcome by saying that it will occur "at random," with \( P(\text{hit}) = 0.7 \). Okay; that may well be about the best I can do to predict what will happen. If you miss I might then be tempted to say that your miss was "due to chance"; you hear people saying things like that all the time. However, retrospectively if I look at the videotape I can see that chance didn't cause you to miss—what caused you to miss was that you extended your shooting hand a little too far and the ball bounced off the back rim.

With situations involving judgmental randomness, nothing is ever "due to chance." Judgmental randomness is not a causal explanation; it's a placeholder for incomplete causal understanding.

This applies to the measurement error model, too: If I'm trying to measure something exactly, like your blood sugar level right this minute, and I take several readings and notice they're not the same, I may choose to "explain" this by invoking the concept of "random" measurement error, but—as we noticed back in Chapter 10—nothing has really been explained by this appeal to judgmental randomness; it's
just an expression of my incomplete understanding of what is really
causing the measurements to come out different each time. If later on
by studying things more carefully I can identify the cause, I won't need
to appeal to randomness any more.

In this sense statistical models of observational data are like
waystations on the road to full causal understanding. They
can be useful for making predictions based on our current state of
knowledge, but we shouldn't let the act of "throwing what we can't
explain into the random error term" in equations like (11.2) make us
complacent—the goal is to reduce the errors to 0 by figuring out the
full causal story. We'll revisit this idea in Chapter xx when we talk
about regression models.

The final point about judgmental randomness for now is one that we
already noticed in Chapter 10:

| It does not follow, just because you say something looks random to
| you, that the particular story involving judgmental randomness you
| have in mind fits the facts. You have to (a) make a good case for
| your judgment substantively and (b) check to see that the data are
| consistent with your story. |

In measurement error models, for example, (a) amounts to verifying that the
observations were collected under (essentially) identical conditions and in a
way that makes them (at least close to) independent, and (b) requires some
methods that we will get to in Chapter xx.

11.4 Chapter Summary

1. In any situation involving uncertainty, there is a body of information
that—if you had it—would remove your uncertainty about the un-
known thing. Often this body of information can be thought of as a
dataset, called the no-uncertainty dataset. In assessing your un-
certainty about the unknown thing it is usually a good first step to
visualize the no-uncertainty dataset and compare it with the data
you have, the factual dataset. Any data in the no-uncertainty dataset
that's missing from the factual dataset is data you wish you had: the
counterfactual dataset.

2. A major goal in most problems is to use the factual data to guess at the
no-uncertainty data. To do this you have to say how the factual and
counterfactual datasets are connected. This is done with a statistical
model, which is a mathematical story that relates the data you have to the data you wish you had.

3. Three types of counterfactuals come up a lot in statistical work: parametric inference, causal inference, and prediction.

- In **parametric inference** there's a population you'd like to summarize, for instance by measuring a variable like the education level of a bunch of people and taking the mean of these ed level values. You can't afford to do a complete census of the whole population, or you decide that such a census would lead to more accuracy than you need for your purposes, so you take a sample from the population, with the hope that the sample summary will be a good guess for the population summary.

- In **causal inference** you'd like to measure the effect caused by some treatment of interest to you, so you get some people and assign some of them to a treatment group and some to a control group. Each person can be thought of as potentially offering two numbers: the value of the outcome he or she would have had if he/she had been in the treatment group, and the outcome the same person would have had if instead he/she had been in the control group. Comparing these two numbers for each person gives the effect on that person caused by the treatment. The hope is that the control values will provide good guesses for what the treatment people's outcomes would have been if they had been in the control group, and vice versa.

- In **prediction** you have some data on how some process of interest to you has come out in the past, and you'd like to forecast how it will come out in the future; or you'd like to guess at the unobserved value of some variable for an individual, on the basis of that person's observed value on another variable. We will talk about how to do this in Chapters xx and xx.

4. Parametric inference is usually based on **sampling models**. In a sampling model, the factual dataset is the sample, the counterfactual dataset is the unsampled individuals, and the no-uncertainty dataset is the whole population. You can build such models both in **finite population** situations like the C&O case study in this chapter and in **conceptually infinite population** setups like the measurement error model in Chapter 10.

5. By contrast, in **causal inference models** the factual dataset is the treatment and control outcome numbers you have observed, and the
counterfactual dataset is what you would have observed instead if the
treatment people had been in the control group and vice versa.

6. For accurate extrapolation from the sample to the population in sam-
ping models, the sampled and unsampled individuals have to be similar
on the outcome and all factors strongly predictive of it. An excellent
way to achieve this when the population is large is by randomizing the
choice of which individuals to sample, as in SRS and IID. Similarly, for
accurate extrapolation from the factual to the counterfactual datasets
in causal inference models, the individuals in the treatment and con-
trol groups should be similar in all relevant ways, and a good way to
promote this similarity when the total number of people in the exper-
iment is large is to randomize the allocation of subjects to treatment
and control. If this is not done—that is, if the data are observational in
character—you can still build causal inference models, but it’s harder
to arrive at valid causal conclusions because it’s not clear how the peo-
ple in the treatment and control groups differ beyond their treatment
status.

7. It’s useful to distinguish between two kinds of randomness: active
randomness describes situations—like randomized controlled exper-
iments and surveys based on SRS—in which somebody went to the
trouble of actually introducing some randomness into the world as part
of their data-gathering, whereas judgmental randomness describes
situations—like observational studies and measurement error setups—
where all that’s happened is that somebody has looked at the process
under study and declared it “random.” Judgmental randomness is not
an absolute property of the world; instead it’s an expression of your
imperfect ability to predict something of interest to you, and of your
incomplete understanding of what is causing the outcomes you see. As
your causal knowledge about the process of interest increases, you will
need less and less to appeal to randomness as part of your description
of the process, until you reach a point where your predictions are 100%
accurate and the process doesn’t look random to you any more.

8. It does not follow, just because you say something looks random to
you, that the particular story involving judgmental randomness you
have in mind fits the facts. You have to (a) make a good case for
your judgment substantively and (b) check to see that the data are
consistent with your story.
11.5 Problems

1. Problems distinguishing between parametric inference, causal inference, and prediction.

2. "Well," you might say, "Why not give everybody iso for xx weeks, and then give them the placebo for another xx weeks?" You could even randomize the order in which you did this, with half of the people getting the placebo first and the other half getting iso first. The trouble with this plan ...

3. In each of the following cases, build a sampling model based on SRS—including identifying the population of interest and explaining why the data either are or are like an SRS from that population—or explain briefly why no such model can be built.

4. I live in a place called Topanga and I work at UCLA. To get from Topanga to UCLA there are two main routes: up Sunset Boulevard, or on the freeway. I am wondering which way is faster, both on average overall and under particular conditions (at 9am versus 11am, say, or when it's sunny versus when it's raining). Design a data-gathering experiment for me that would reduce my uncertainty about the best route, and build a statistical model that would help me analyze the data. Be specific about the no-uncertainty, factual, and counterfactual datasets, and explain how under your design the factual data would provide good guesses for the counterfactual data. What kind of randomness are you invoking (if any) in your model? Explain briefly.

5. (more to come)

11.6 Notes

1. Geisser, etc.

2. Whether you should be willing to settle for a single number like the mean is another question (see the discussion on this point back in Chapter xx).

3. Case Study 11 is based on an article by John Neter called "How Accountants Save Money by Sampling" in an excellent book called Statistics: A Guide to the Unknown (Judith Tanur, Editor; Second Edition, 1978; San Francisco: Holden-Day) that contains essays on the use of statistics in many different application areas. In fact the designers of the sampling experiment in Case Study 11 took a stratified random sample of 2,072 waybills (recall that this method came up back in Chapter 6), but we're not ready yet to talk about how to analyze data from such samples. Their method was equivalent to an SRS of about 5,300 waybills with a mean of about $75 and an SD of about $137.
4. Rubin, Little, etc.

5. One logical difference between viewing the models in the C&O and AIDS case studies as missing data problems is that it is feasible to observe all the missing values in the C&O study (just by taking a complete census rather than a sample), whereas you could never exactly observe both the factual and counterfactual outcomes for people in the AIDS study. Some people (ref) think this is an important distinction; I'm not so sure.

6. Given that all the values in Figure 11.3 are 1's and 0's, it may seem a little strange to use the average value across a lot of people as our best guess of what somebody's control value would have been, since the average would almost certainly not be 0 or 1. In Chapter xx I will try to say why this is a good idea.

7. References.

8. Multiple regression, etc.

9. Some people go beyond what I have described here and quite literally base their inferences on the randomization (permutation tests, etc.). I won't talk about such methods in this book.
Figure 11.1: A sampling model for the C20 case study
Figure 11.2: A conceptually infinite sampling model for the hypoglycemia measurement error example.

The three actual readings:

\[ \begin{align*}
\text{blood sugar} & \rightarrow \begin{bmatrix} 74 \\ 92 \\ 69 \end{bmatrix} \\
\text{mean } \bar{y} & = 78 \text{ mg/dL} \\
\text{sd } s & = 10 \text{ mg/dL}
\end{align*} \]

The three measurements:

all the other possible readings your doctor could have gotten

Counterfactual dataset:

hypotheical collection of all possible measurements
<table>
<thead>
<tr>
<th>person</th>
<th>treatment status</th>
<th>AIDS outcome (T)</th>
<th>placebo (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>412</td>
<td>T</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>413</td>
<td>C</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>414</td>
<td>C</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>833</td>
<td>C</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

(1 = got AIDS, 0 = didn't get AIDS)

**Factual dataset:**

The observed outcomes

**Counterfactual dataset:**

What outcome the treatment people would have had if they had been control people, and vice versa.

**Figure 11.3:** A causal inference model for the AIDS case study.
### Table A1

<table>
<thead>
<tr>
<th>$z$</th>
<th>Area</th>
<th>$z$</th>
<th>Area</th>
<th>$z$</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0</td>
<td>1.50</td>
<td>86.64</td>
<td>3.00</td>
<td>99.730</td>
</tr>
<tr>
<td>0.05</td>
<td>3.99</td>
<td>1.55</td>
<td>87.89</td>
<td>3.05</td>
<td>99.771</td>
</tr>
<tr>
<td>0.10</td>
<td>7.97</td>
<td>1.60</td>
<td>89.04</td>
<td>3.10</td>
<td>99.806</td>
</tr>
<tr>
<td>0.15</td>
<td>11.92</td>
<td>1.65</td>
<td>90.11</td>
<td>3.15</td>
<td>99.837</td>
</tr>
<tr>
<td>0.20</td>
<td>15.85</td>
<td>1.70</td>
<td>91.09</td>
<td>3.20</td>
<td>99.863</td>
</tr>
<tr>
<td>0.25</td>
<td>19.74</td>
<td>1.75</td>
<td>91.99</td>
<td>3.25</td>
<td>99.885</td>
</tr>
<tr>
<td>0.30</td>
<td>23.58</td>
<td>1.80</td>
<td>92.81</td>
<td>3.30</td>
<td>99.903</td>
</tr>
<tr>
<td>0.35</td>
<td>27.37</td>
<td>1.85</td>
<td>93.57</td>
<td>3.35</td>
<td>99.919</td>
</tr>
<tr>
<td>0.40</td>
<td>31.08</td>
<td>1.90</td>
<td>94.26</td>
<td>3.40</td>
<td>99.933</td>
</tr>
<tr>
<td>0.45</td>
<td>34.73</td>
<td>1.95</td>
<td>94.88</td>
<td>3.45</td>
<td>99.944</td>
</tr>
<tr>
<td>0.50</td>
<td>38.29</td>
<td>2.00</td>
<td>95.45</td>
<td>3.50</td>
<td>99.953</td>
</tr>
<tr>
<td>0.55</td>
<td>41.77</td>
<td>2.05</td>
<td>95.96</td>
<td>3.55</td>
<td>99.961</td>
</tr>
<tr>
<td>0.60</td>
<td>45.15</td>
<td>2.10</td>
<td>96.43</td>
<td>3.60</td>
<td>99.968</td>
</tr>
<tr>
<td>0.65</td>
<td>48.43</td>
<td>2.15</td>
<td>96.84</td>
<td>3.65</td>
<td>99.974</td>
</tr>
<tr>
<td>0.70</td>
<td>51.61</td>
<td>2.20</td>
<td>97.22</td>
<td>3.70</td>
<td>99.978</td>
</tr>
<tr>
<td>0.75</td>
<td>54.67</td>
<td>2.25</td>
<td>97.56</td>
<td>3.75</td>
<td>99.982</td>
</tr>
<tr>
<td>0.80</td>
<td>57.63</td>
<td>2.30</td>
<td>97.86</td>
<td>3.80</td>
<td>99.986</td>
</tr>
<tr>
<td>0.85</td>
<td>60.47</td>
<td>2.35</td>
<td>98.12</td>
<td>3.85</td>
<td>99.988</td>
</tr>
<tr>
<td>0.90</td>
<td>63.19</td>
<td>2.40</td>
<td>98.36</td>
<td>3.90</td>
<td>99.990</td>
</tr>
<tr>
<td>0.95</td>
<td>65.79</td>
<td>2.45</td>
<td>98.57</td>
<td>3.95</td>
<td>99.992</td>
</tr>
<tr>
<td>1.00</td>
<td>68.27</td>
<td>2.50</td>
<td>98.76</td>
<td>4.00</td>
<td>99.993</td>
</tr>
<tr>
<td>1.05</td>
<td>70.63</td>
<td>2.55</td>
<td>98.92</td>
<td>4.05</td>
<td>99.994</td>
</tr>
<tr>
<td>1.10</td>
<td>72.87</td>
<td>2.60</td>
<td>99.07</td>
<td>4.10</td>
<td>99.995</td>
</tr>
<tr>
<td>1.15</td>
<td>74.99</td>
<td>2.65</td>
<td>99.20</td>
<td>4.15</td>
<td>99.996</td>
</tr>
<tr>
<td>1.20</td>
<td>76.99</td>
<td>2.70</td>
<td>99.31</td>
<td>4.20</td>
<td>99.997</td>
</tr>
<tr>
<td>1.25</td>
<td>78.87</td>
<td>2.75</td>
<td>99.40</td>
<td>4.25</td>
<td>99.998</td>
</tr>
<tr>
<td>1.30</td>
<td>80.64</td>
<td>2.80</td>
<td>99.49</td>
<td>4.30</td>
<td>99.983</td>
</tr>
<tr>
<td>1.35</td>
<td>82.30</td>
<td>2.85</td>
<td>99.56</td>
<td>4.35</td>
<td>99.986</td>
</tr>
<tr>
<td>1.40</td>
<td>83.85</td>
<td>2.90</td>
<td>99.63</td>
<td>4.40</td>
<td>99.989</td>
</tr>
<tr>
<td>1.45</td>
<td>85.29</td>
<td>2.95</td>
<td>99.68</td>
<td>4.45</td>
<td>99.991</td>
</tr>
</tbody>
</table>