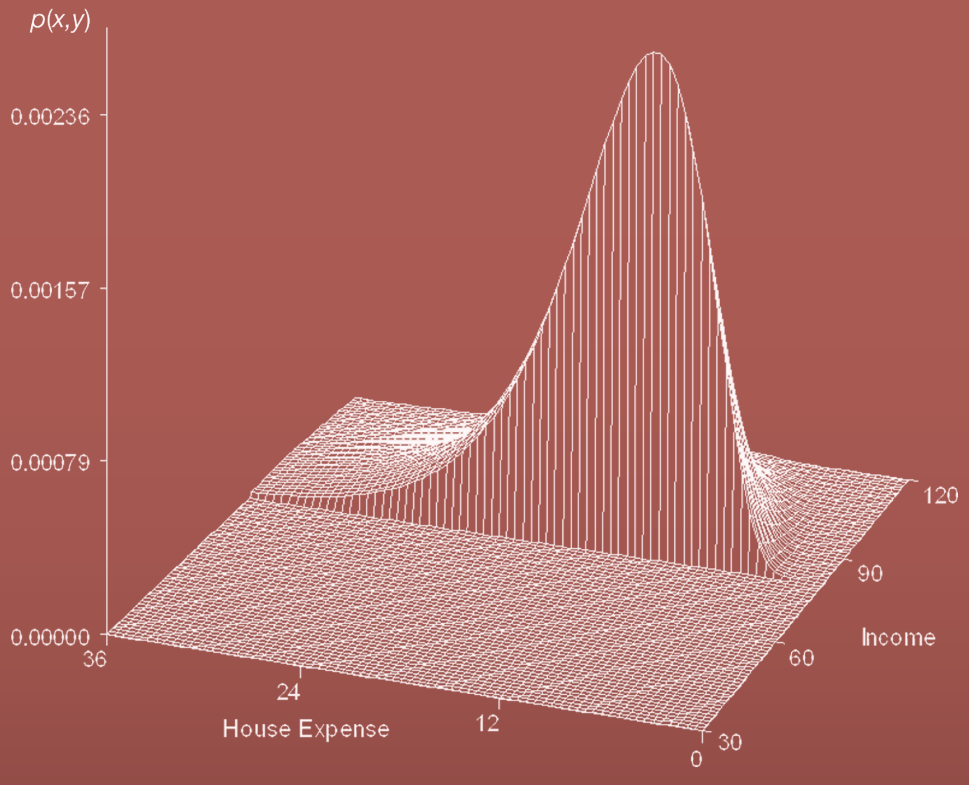
Texts in Statistical Science

Understanding Advanced Statistical Methods



Peter H. Westfall

Kevin S. S. Henning

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Understanding Advanced Statistical Methods

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Analysis of Failure and Survival Data
P. J. Smith

The Analysis of Time Series — An Introduction, Sixth Edition
C. Chatfield

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West, and J. Harrison

Applied Categorical and Count Data Analysis
W. Tang, H. He, and X.M. Tu

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Applied Statistics — Handbook of GENSTAT Analysis
E.J. Snell and H. Simpson

Applied Statistics — Principles and Examples
D.R. Cox and E.J. Snell

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Bayesian Data Analysis, Second Edition
A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians
R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Beyond ANOVA — Basics of Applied Statistics
R.G. Miller, Jr.

The BUGS Book: A Practical Introduction to Bayesian Analysis
D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis — A Bayesian Approach
J.Q. Smith

Design and Analysis of Experiments with SAS
J. Lawson

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology — Study Design and Data Analysis, Second Edition
M. Woodward

Essential Statistics, Fourth Edition
D.A.G. Rees

Exercises and Solutions in Statistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Exercises and Solutions in Biostatistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Extending the Linear Model with R — Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Generalized Additive Models: An Introduction with R
S. Wood

Generalized Linear Mixed Models: Modern Concepts, Methods and Applications
W. W. Stroup

Graphics for Statistics and Data Analysis with R
K.J. Keen

Interpreting Data — A First Course in Statistics
A.J.B. Anderson

Introduction to General and Generalized Linear Models
H. Madsen and P. Thyregod

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Introduction to Probability with R
K. Baclawski

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Introduction to Statistical Limit Theory
A.M. Polansky

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Introduction to the Theory of Statistical Inference
H. Liero and S. Zwanzig

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Linear Algebra and Matrix Analysis for Statistics
S. Banerjee and A. Roy

Logistic Regression Models
J.M. Hilbe

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures — A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics — A Practical Approach
B. Flury and H. Riedwyl

Multivariate Survival Analysis and Competing Risks
M. Crowder

Pólya Urn Models
H. Mahmoud

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Multivariate Analysis, Fifth Edition
A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research
D.G. Altman

A Primer on Linear Models
J.F. Monahan

Principles of Uncertainty
J.B. Kadane

Probability — Methods and Measurement
A. O'Hagan

Problem Solving — A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
P.S.R.S. Rao

Stationary Stochastic Processes: Theory and Applications
G. Lindgren

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control — Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory: A Concise Introduction
F. Abramovich and Y. Ritov

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants
S. Letchford

Statistics for Epidemiology
N.P. Jewell

Statistics for Technology — A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering — A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Stochastic Processes: An Introduction, Second Edition
P.W. Jones and P. Smith

Survival Analysis Using S — Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

The Theory of Linear Models
B. Jørgensen

Time Series Analysis
H. Madsen

Time Series: Modeling, Computation, and Inference
R. Prado and M. West

Understanding Advanced Statistical Methods
P.H. Westfall and K.S.S. Henning

Texts in Statistical Science

Understanding Advanced Statistical Methods

Peter H. Westfall

Information Systems and Quantitative Sciences
Texas Tech University, USA

Kevin S. S. Henning

Department of Economics and International Business
Sam Houston State University, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20130401

International Standard Book Number-13: 978-1-4665-1211-5 (eBook - PDF)

Contents

List of Examples

Preface

We wrote this book because there is a large gap between the elementary statistics course that most people take and the more advanced research methods courses taken by graduate and upper-division students so they can carry out research projects. These advanced courses include difficult topics such as regression, forecasting, structural equations, survival analysis, and categorical data, often analyzed using sophisticated likelihood-based and even Bayesian methods. However, they typically devote little time to helping students understand the fundamental assumptions and machinery behind these methods. Instead, they teach the material like witchcraft: Do this, do that, and voilà—statistics! Consequently, students learn little about what they are doing and why they are doing it. Like trained parrots, they learn how to recite statistical jargon mindlessly. The goal of this book is to make statistics less like witchcraft and to treat students as intelligent humans and not as trained parrots—thus the title, *Understanding Advanced Statistical Methods*.

This book will surprise your students. It will cause them to think differently about things, not only about math and statistics, but also about research, the scientific method, and life in general. It will teach them how to do good modeling—and hence good statistics—from a standpoint of *deep* knowledge rather than *rote* knowledge. It will also provide them with tools to think critically about the claims they see in the popular press and to design their own studies to avoid common errors.

There are plenty of formulas in this book, because to understand advanced statistical methods requires understanding probabilistic models, and probabilistic models are necessarily mathematical. But if your students ever find themselves plugging numbers into formulas mindlessly, make them stop and ask, "Why?" Getting students to ask and answer that question is the main objective of this book. Having them perform mindless calculations is a waste of your time and theirs, unless they understand the *why*. Every formula tells an interesting story, and the story explains the *why*.

Although all statistics books purport to have the goal of making statistics understandable, many try to do so by avoiding math. This book does not shy away from math; rather, it teaches the needed math and probability along with the statistics. Even if your students are math "phobes" they will learn the math and probability theory and hopefully enjoy it, or at least appreciate it.

In particular, statistics is all about unknown, algebraic quantities. What is the probability of a coin landing heads up when flipped? It is not 50%. Instead, it is an unknown algebraic quantity θ that depends on the construction of the coin and on the methods of the coin-flipper. Any book that teaches statistics while avoiding algebra is therefore a book of fiction!

This book uses calculus where needed to help readers understand continuous distributions and optimizations. Students should learn enough calculus to understand the logical arguments concerning these core concepts. But calculus is not a prerequisite. We only assume that students have a familiarity with algebra, functions and graphs, and spreadsheet software such as Microsoft Excel®. The book employs a "just-in-time" approach, introducing mathematical topics, including calculus, where needed. We present mathematical concepts in a concrete way, with the aim of showing students how even the seemingly hard math is really not so hard, as well as showing them how to use math to answer important questions about our world.

As far as probability theory goes, we employ a laser-beam focus on those aspects of probabilistic models that are most useful for statistics. Our discussion therefore focuses

more on distributions than on counting formulas or individual probability calculations. For example, we present Bayes' theorem in terms of distributions rather than using the classical two-event form presented in other sources. For another example, we do not emphasize the binomial distribution; instead, we focus on the Bernoulli distribution with independent and identically distributed observations.

This book emphasizes applications; it is not "math for math's sake." We take real data analysis very seriously. We explain the theory and logic behind real data analysis intuitively and gear our presentation toward students who have an interest in science but may have forgotten some math.

Statistics is not a collection of silly rules that students should recite like trained parrots—rules such as $p < 0.05$, $n > 30$, $\rho > 0.3$, etc. We call these *ugly rules of thumb* throughout the book to emphasize that they are mere suggestions and that there is nothing hard-and-fast about any of them. On the other hand, the logic of the mathematics underlying statistics is not ugly at all. Given the assumptions, the mathematical conclusions are 100% true. But the assumptions *themselves* are never quite true. This is the heart and soul of the subject of statistics—how to draw conclusions successfully when the premises are flawed—and this is what your students will learn from this book.

This book is not a "cookbook." Cookbooks tell you all about the *what* but nothing about the *why*. With computers, software, and the Internet readily available, it is easier than ever for students to lose track of the *why* and focus on the *what* instead. This book takes exactly the opposite approach. By enabling your students to answer the *why*, it will help them to figure out the *what* on their own—that is, they will be able to develop their own statistical recipes. This will empower your students to use advanced statistical methods with confidence.

The main challenge for your students is not to understand the math. Rather, it is to understand the statistical point of view, which we present consistently throughout this book as a mantra:

> Model Produces Data

More specifically, the *statistical model* is a *recipe* for *producing random data*. This one concept will turn your students' minds around 180°, because most think a statistical model is something *produced by data* rather than a *producer of data*. In our experience, the difficulty in understanding the statistical model as a data-generator is the single most significant barrier to students' learning of statistics. Understanding this point can be a startling epiphany, and your students might find statistics to be fun, and surprisingly easy, once they "get it." So let them have fun!

Along with the presentation of models as *producers of data*, another unique characteristic of this book is that it avoids the overused (and usually misused) "population" terminology. Instead, we define and use the "process" terminology, which is always more correct, generally more applicable, and nearly always more scientific. We discuss populations, of course, but correctly and appropriately. Our point of view is consistent with the one presented in *Statistical Science* (26(1), 1–9, 2011) by Robert E. Kass and several discussants in an article entitled "Statistical inference: The big picture."

Another unique characteristic of this book is that it teaches Bayesian methods *before* classical (frequentist) methods. This sequencing is quite natural given our emphasis on probability models: The flow from probability to likelihood to Bayes is seamless. Placing Bayesian methods before classical methods also allows for more rounded and thoughtful discussion of the convoluted frequentist-based confidence interval and hypothesis testing concepts.

This book has no particular preference for the social and economic sciences, for the biological and medical sciences, or for the physical and engineering sciences. All are useful, and the book provides examples from all these disciplines. The emphasis is on the overarching *statistical* science. When the book gives an example that does not particularly apply to you or your students' fields of study, just change the example! The concepts and methods of statistics apply universally.

The target audience for this book is mainly upper-division undergraduates and graduate students. It can also serve lower-division students to satisfy a mathematics general education requirement. A previous course in statistics is not necessary.

This book is particularly useful as a prerequisite for more advanced study of regression, experimental design, survival analysis, time series analysis, structural equations modeling, categorical data analysis, nonparametric statistics, and multivariate analysis. We introduce regression analysis (ordinary and logistic) in the book, and for this reason, we refer to the data as Y, rather than X as in many other books. We use the variable designation X as well, but mainly as a predictor variable.

The spreadsheet software Microsoft Excel is used to illustrate many of the methods in this book. It is a good idea, but not strictly necessary, to use a dedicated mathematical or statistical software package in addition to the spreadsheet software. However, we hope to convince your students that advanced statistical methods are really not that hard, since one can understand them to a great extent simply by using such commonplace software as Excel.

About Using This Book

- Always get students to ask "Why?" The point of the book is not the *what*; it is the *why*. Always question assumptions and aim to understand how the logical conclusions follow from the assumptions.
- Students should read the book with a pencil and paper nearby, as well as spreadsheet or other software, for checking calculations and satisfying themselves that things make sense.
- Definitions are important and should be memorized. Vocabulary terms are given in **boldface** in the book, and their definitions are summarized at the ends of the chapters. Strive to teach the definitions in the context of your own field of interest, or in the context of your students' fields of interest.
- Some formulas should be memorized, along with the stories they tell. Important formulas are given at the ends of the chapters.
- We often give derivations of important formulas, and we give the reasons for each step in parentheses to the right of the equations. These reasons are often simple, involving basic algebra. The reasons are more important than the formulas themselves. Learn the reasons first!
- The exercises all contain valuable lessons and are essential to understanding. Have your students do as many as possible.
- A companion website http://courses.ttu.edu/isqs5347-westfall/westfall_book.htm includes computer code, sample quizzes, exams and other pedagogical aids.

Acknowledgments

We would like to thank Josh Fredman for his excellent editing and occasional text contributions; students in Dr. Westfall's ISQS 5347 class, including Natascha Israel, Ajay Swain, Jianjun Luo, Chris Starkey, Robert Jordan, and Artem Meshcheryakov for careful reading and feedback; Drs. Jason Rinaldo and D. S. Calkins for careful reading, needling, and occasional text passages; and the production staff at Taylor & Francis Group/CRC Press, including Rachel Holt and Rob Calver, as well as Remya Divakaran of SPi for helpful direction and editing. Most graphics in the book were produced using the SGPLOT and SGPANEL procedures in SAS software.

Authors

Peter H. Westfall has a PhD in statistics and many years of teaching, research, and consulting experience in biostatistics and a variety of other disciplines. He has published over 100 papers in statistical theory and methods, won several teaching awards, and has written several books, one of which won two awards from the Society for Technical Communication. He is former editor of *The American Statistician* and is a Fellow of both the American Statistical Association and of the American Association for the Advancement of Science.

Kevin S. S. Henning has a PhD in business statistics from Texas Tech University and currently teaches business statistics and forecasting in the Department of Economics and International Business in the College of Business at Sam Houston State University.

1

Introduction: Probability, Statistics, and Science

1.1 Reality, Nature, Science, and Models

So, what is reality? Yes, this may be an odd question to start a statistics book. But reality is what science is all about: It is the study of what is real. "What is real?" is a topic that fills volumes of philosophy books, but for our purposes, and for the purposes of science in general, the question of what is real is answered by "That which is *natural* is real." Of course, that raises the question, "What is natural?"

Without delving too far into philosophy, **Nature** is all aspects of past, present, and future existence. Understanding Nature requires common observation—that is, it encompasses those things that we can agree we are observing. As used in this book, Nature includes the physical sciences (e.g., planets, galaxies, gravity), the biological sciences (e.g., DNA, medicine), and the social sciences (e.g., economics, psychology). Nature includes man-made things such as dams, as well as social constructs such as economic activity; we certainly do not limit our definition of Nature to those things that are without human intervention. In fact, most examples involving Nature in this book *do* involve human activity.

Science is the study of Nature. It involves understanding why Nature is the way that it is and using such knowledge to make predictions as to what will happen—or would have happened—under various circumstances.

Personal realities which are not commonly observed or agreed upon—for example, those of a mystical or spiritual quality—are outside the scope of science. Someone may believe that the Earth rests upon a large turtle, and while this point of view may offer comfort and meaning, it is not a common, agreed-upon observation and is therefore not a scientific proposition. The same can be said about major religions: Tenets of faith lacking agreed-upon observation cannot be subjected to measurement and testing and hence are outside the scope of science.

Statistics is the language of science. In its broadest form, statistics concerns the analysis of recorded information or data. Data are commonly observed and subject to common agreement and are therefore more likely to reflect our common reality or Nature. Data offer us a clearer picture of what Nature is and how Nature works, and statistical analyses of data allow us to reverse-engineer natural processes and thus gain scientific knowledge.

To understand Nature, you must construct a model for how Nature works. A **model** helps you to understand Nature and also allows you to make predictions about Nature. There is no right or wrong model; they are all wrong! But some are better than others. The better models are the ones you want to use, and in this book we'll help you identify them.

If you have ever played with toy trains or dolls, you are probably very familiar with the general concept of modeling. Your first toys probably only resembled their real-world counterparts in the most elementary of ways. As you grew older, however, your toys probably became more like the real thing, and hence, they became better models. For example, your first toy train might have been nothing more than a piece of wood sculpted to look like a locomotive, with no working parts, but when you got older, you may well have played with a working toy locomotive that ran on electric tracks and pulled a few miniature cars. This train was a better model because the principles behind its operation were closer to those of real trains. They were still not identical, of course. Real trains have sophisticated throttle and communications equipment and are many orders of magnitude larger than toy trains.

Trains and dolls are *physical models*. The focus of this book will be on another class of models, called *mathematical models*, which are built out of equations rather than materials. As with physical models such as the toy train, these mathematical models are not how Nature really operates, but if they are similar to Nature, they can be very informative. Thus, your model is good if it produces data resembling what Nature would produce. These models are personal: They are mental abstractions that *you* create, and that *you* use. Someone else may create and use a different model.

We will often represent models using graphs. When you see a graph, always ask yourself "What is the information that is provided in this graph?" To answer, look carefully at the axis labels and the numbers on the axes, and be sure you understand what they mean. Also, read the figure legends and the surrounding text. While a picture may be worth 1000 words, it is only worth one equation. But it is a lot more fun to look at than the equation! It is also easier to remember. When you see an equation, ask yourself, "How does that look in a graph?"

Example 1.1: A Model for Driving Time

You will drive x kilometers. How long will it take you? If you typically average 100 km/hour (or 62.1 miles/hour), then your driving time y (in hours) may be given by the model $y = x/100$; Figure 1.1 shows a graph of this equation.

Thus, if your distance is 310 km, then your driving time may be given by 3.10 hours or 3 hours and 6 minutes.

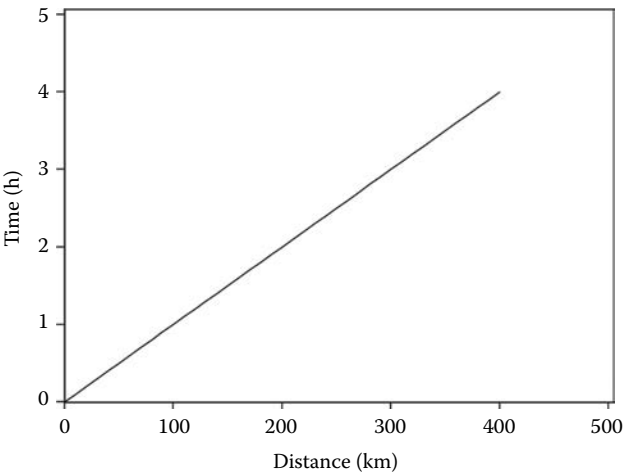


FIGURE 1.1
A model for driving time as a function of distance: $y = x/100$.

Two things you should note about the driving time model: First, a model allows you to make *predictions*, such as 3 hours and 6 minutes. Note that a prediction is not about something that happens in the future (which is called a **forecast**). Rather, a **prediction** is a more general, "what-if" statement about something that might happen in the past, present, future, or not at all. You may never in your life drive to a destination that is precisely 310 km distant, yet still the model will tell you how long it would take if you did.

Second, notice that the *model produces data*. That is, if you state that $x = 310$, then the model produces $y = 3.10$. If you state that $x = 50$, then the model produces $y = 0.50$. This will be true of all models described in this book—they all produce data. This concept, *model produces data*, may be obvious and simple for this example involving driving time, but it is perhaps the most difficult thing to understand when considering statistical models.

Of course, the model $y = x/100$ doesn't produce the data all by itself, it requires someone or something to do the calculations. It will not matter who or what produces the data; the important thing is that the model is a *recipe* that can be used to produce data. In the same way that a recipe for making a chocolate cake does not actually produce the cake, the mathematical model itself does not actually produce the data. Someone or something must carry out the instructions of the recipe to produce the actual cake; likewise, someone or something must carry out the instructions of the model to produce the actual data. But as long as the instructions are carried out correctly, the result will be the chocolate cake, no matter who or what executes the instructions. So you may say that the cake recipe produces the cake, and by the same logic, you may also say that the model produces the data.

A **statistical model** is also a *recipe* for *producing* data. Statistics students usually think, incorrectly, that the data produce the model, and this misconception is what makes statistics a "difficult" subject. The subject is much easier once you come to understand the concept *model produces data*, which throughout this book is an abbreviated phrase for the longer and less catchy phrase, "the model is a recipe for producing data." You can use data to *estimate* models, but that does not change the fact that your model comes first, before you ever see any data. Just like the model $y = x/100$, a statistical model describes how Nature works and how the data from Nature will appear. Nature is already there before you sample any data, and you want your model to mimic Nature. Thus, you will assume that your model produces your data, not the other way around.

A simple example will clarify this fundamental concept, which is absolutely essential for understanding the entire subject of statistics. If you flip a perfectly balanced coin, you think there is a 50% chance that it will land heads up. This is your model for how the data will appear. If you flip the coin 10 times and get 4 heads, would you now think that your coin's Nature has changed so that it will produce 40% heads in the future? Of course not. *Model produces data*. The data do not produce the model.

1.2 Statistical Processes: Nature, Design and Measurement, and Data

Statistical analysis requires data. You might use an experiment, or a survey, or you might query an archived database. Your method of data collection affects your interpretation of the results, but no matter which data collection process you choose, the science of studying Nature via statistics follows the process shown in Figure 1.2.

Notice that Nature produces data but only after humans tap Nature through *design and measurement*.

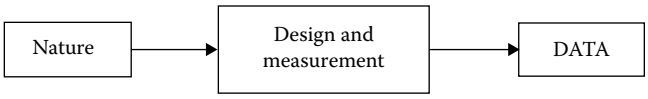


FIGURE 1.2
The statistical science paradigm.

In *confirmatory research*, design and measurement follow your question about Nature. For example, you might have the question, "Does taking vitamin C reduce the length of a cold?" To answer that question you could design a study to obtain **primary data** that specifically addresses that question. In *exploratory research*, by contrast, your question of interest comes to mind after you examine the data that were collected for some other purpose. For example, in a survey of people who had a cold recently, perhaps there was a question about daily vitamin intake. After examining that data, the question "Does taking vitamin C reduce the length of a cold?" may come into your mind. Since the survey was not intended to study the effects of vitamin C on the duration of colds, these data are called **secondary data**. Conclusions based on confirmatory research with primary data are more reliable than conclusions based on exploratory research with secondary data. On the other hand, secondary data are readily available, whereas it is time-consuming and costly to obtain primary data.

Both types of analyses—those based on primary data and those based on secondary data—are useful. Science typically progresses through an iterative sequence of exploratory and confirmatory research. For example, after you notice something interesting in your exploratory analysis of secondary data, you can design a new study to confirm or refute the interesting result.

To understand Figure 1.2, keep in mind that the arrows denote a *sequence*: Nature precedes your design and measurement, which in turn precede your DATA. The capital letters in *DATA* are deliberate, meant to indicate that your data have not yet been observed: They are potential observations at this point and are unknown or random. When we discuss data that are already observed, we will use the lowercase spelling *data*. These data are different, because they are fixed, known, and hence nonrandom.

The uppercase versus lowercase distinction (*DATA* versus *data*) will be extremely important throughout this book. Why? Consider the question "Does vitamin C reduce the length of a cold?" If you design a study to find this out, you will collect lowercase "d" data. These data will say something about the effects of vitamin C on the length of a cold *in this particular study*. However, they are not the only data you could possibly have collected, and they cannot describe with absolute perfection the nature of the effect of vitamin C on the length of cold. Your data might be anomalous or incomplete, suggesting conclusions that differ from the reality of Nature. In contrast, the as-yet unobserved DATA include all possible values. Statistical theory is all about generalizing from data (your sample) to the processes that produce the entirety of DATA that could possibly be observed. With proper statistical analyses, you are less likely to be misled by anomalous data.

In other statistics sources, *DATA* refer to a *population*, and *data* refer to a *sample from the population*. If it helps you to understand the DATA/data distinction, go ahead and think this way for now, but we suggest that you avoid the "population" terminology because it is misleading. You will learn much more about this in Chapter 7.

You will never see all the DATA; generally, it is an infinite set of possible outcomes of everything that could possibly happen. On the other hand, you do get to see your data. A main goal of statistical analysis is to use the data that you have observed to say something accurate about the potential DATA that you have not observed.

Definitions of Terms in Figure 1.2

- **Nature** is the real situation. It might refer to a phenomenon in biology, physics, or human societal interactions. It is there whether you collect data or not.
- **Design** is your plan to collect data. Broadly speaking, design involves deciding how you are going to study Nature. You could directly observe the phenomenon of interest, conduct an experiment, or analyze existing measurements contained in a database; the design refers to the methods you will use to collect your data. Think of design as something that happens before you get the actual numbers.
- **Measurement** refers to the *type* and *units* of the data that you will record and use; for example, a measurement could be height in feet, rounded to the nearest inch. The binary "yes" or "no" choices on a questionnaire is another example of a measurement. A measurement can also be a processed number such as the average of responses to questions one through five on a questionnaire, where each response is 1, 2, 3, 4, or 5.
- **DATA** are the potential data that you might observe. At this point, you should visualize a data set that will be in your computer (e.g., in a spreadsheet), but you don't know what the numbers are.

Example 1.2: The Statistical Science Paradigm for Temperature Observation

"How about the weather today?" is a common elevator topic. Suppose you designed a simple study to measure temperature. In this case:

- *Nature* refers to weather.
- *Design* refers to your plan to get the data. For example, you may plan to go outside and look at your thermometer. Or, you may plan to go online and see what a weather website tells you.
- *Measurement* refers to the type and units of the data you will actually collect. If your thermometer measures temperature in Celsius, then the measurement will be temperature in the Celsius scale. Further, if you plan to report the temperature to the nearest degree, the measurement can further be refined to be temperature in the Celsius scale rounded to the nearest integer.
- *DATA* refer to the actual number you will observe, before you have observed it. It could be any value, so you must represent it algebraically as Y (a capital letter). Once you actually observe a temperature—say, 15°C—then that's your lowercase "d" data, $y = 15$°C (note the lowercase y).

This example would be more interesting if your design were to collect data over 365 consecutive days, in which case your data set would include 365 numbers instead of just one.

Example 1.3: The Statistical Science Paradigm for Presidential Approval Polling

What do people think about the current president? In this case, the elements are as follows:

- *Nature* is public opinion.
- *Design* is your plan to collect the data. This plan should be much more elaborate than in the weather example, Example 1.2. For instance, you may hire a staff of phone interviewers, obtain a list of randomly selected telephone numbers, write a script for the interviewers to explain what they are doing to the people who answer the phone, decide how many times to follow up

TABLE 1.1

A DATA Set

ID	Response
0001	Y_1
0002	Y_2
0003	Y_3
0004	Y_4
…	…

if no one is home, decide on how many people to call, and decide on how many responses to obtain.

- *Measurement* refers to the type of data that will be collected to measure opinion about the president. If you ask the question "Do you like the president?" then the measurement is simply a yes-or-no outcome. This type of measurement is common in statistics and is sometimes called a **binary response**. Or, you might ask respondents a battery of questions about different aspects of the president's performance, on which they rate their preference according to a 1, 2, 3, 4, 5 scale. In this case, the measurement might be average preference using a collection of questionnaire items, sometimes called a **Likert scale**.
- *DATA* refer to the actual numbers that will be in your spreadsheet or other database. For example, in the simple "yes/no" measurement, the data might look like as shown in Table 1.1.

The DATA values are as-yet unknown, so you have to represent them algebraically as Y_i rather than as specific values. Once you observe the data, they become specific data values such as y_1 = "yes," y_2 = "no," y_3 = "no," y_4 = "no," and so on, assuming the measurement is a binary yes-or-no outcome.

Example 1.4: The Statistical Science Paradigm for Luxury Car Sales

How are luxury car sales these days? In an era of expensive gas prices, people tend to shy away from gas-guzzling luxury cars. If you were studying trends at a dealership, the elements might be defined as follows:

- *Nature* is car purchasing behavior.
- *Design* is your plan to collect data. You may plan to contact people in the car industry and request annual sales figures. You will need to define specifically what is meant by a luxury car first.
- *Measurement* refers to the type of data you will record. In this case, that might be annual U.S. sales (in millions of dollars) of luxury cars. Alternatively, you might decide to measure numbers of cars sold (in thousands). Or, you might decide to measure both dollar sales and car count; this would be called a **bivariate measurement**.
- *DATA* refer to the values you will collect. See Table 1.2.

Prior to observation, the DATA are random, unknown, and hence indicated algebraically as Y_i. Once you collect the data, you can replace the uppercase DATA values Y_i with the actual numbers.

TABLE 1.2

Annual Sales DATA for Luxury Cars

Year	Annual Sales
2000	Y_1
2001	Y_2
2002	Y_3
…	…

1.3 Models

A statistical model is an abstraction of Figure 1.2. It is a simplification that allows you to both explain how Nature works and make predictions about how Nature works. To explain, and make predictions, the process by which data are produced in Figure 1.2 is represented using the model shown in Figure 1.3.

The simplest case of a probability model $p(y)$ is the coin flip model: $p(\text{heads}) = 0.5$ and $p(\text{tails}) = 0.5$. As a data producer, this model produces the outcomes heads or tails randomly, just like coin flips. It can produce as many random coin flips as you want. Isn't that handy! The model $p(y)$ can be an automatic coin flipper!

Your model $p(y)$ substitutes for both Nature and design and measurement shown in Figure 1.2 and states that the mathematical function $p(y)$ produces the data, as shown in Figure 1.3. Your real DATA are produced from Nature, as tapped through your design and measurement. Your probabilistic model $p(y)$ also produces DATA; you will see examples of this repeatedly throughout the book, where we produce DATA* from models $p(y)$ using computer random number generators. When we use the computer to generate DATA, we call the resulting values DATA*, designated with an asterisk*, to distinguish them from real DATA.

Probabilistic models $p(y)$ are usually *wrong* in one way or another, as they are oversimplifications, just like a toy train is an oversimplification of the real train. But the model is *useful* if it is good, meaning that the DATA* it produces look like your real DATA. The more similar your model's DATA* are to Nature's own DATA—as tapped through your design and measurement—the better your model. By analogy, the model train is *good* if it faithfully represents the real train, but the model train is obviously *wrong* in that it is not the real train.

Just as Figure 1.2 shows how Nature's data are produced, the model shown in Figure 1.3 also *produces* data. Note that the term *model* is used in two senses here: First, Figure 1.3 itself is a model for how Nature works, and second, the function $p(y)$ is called a **probability model**. To summarize these two meanings in a single sentence, your *model* for reality is that your DATA come from a *probability model* $p(y)$. So the statement, *model produces data*, is itself a model—your model—for how your DATA will be produced.

The dual meanings of the word *model* are so important; they need a shout out.

The Dual Meanings of the Term "Model"

Your *model* for Nature is that your DATA come from a *probability model* $p(y)$.

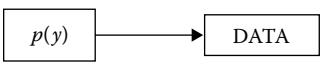


FIGURE 1.3
The model for the statistical science paradigm shown in Figure 1.2.

Such a model $p(y)$ can be used to predict and explain Nature. Again, the term *prediction* here does not necessarily refer to predicting the future, which is called *forecasting*. Rather, a prediction is a guess about unknown events in the past, present, or future or about events that may never happen at all. The best way to understand prediction is to think of what-if scenarios: "What if I invest $100,000 in this mutual fund? How much money would I have at the end of the year?" Or, "What if the state had issued 1000 more hunting licenses last year? What would the deer population be today?" These are examples of predictions, covering the past, present, future, or none of the above.

At this point, the meaning of $p(y)$ may be unclear, especially in relation to the example of driving time of Example 1.1, where the model was $f(x)$. The following sections clarify the distinctions between the deterministic model $y = f(x)$ and the probabilistic model, which is represented by the following expression:

$$Y \sim p(y)$$

The symbol ~ can be read aloud either as "produced by" or "distributed as." In a complete sentence, the mathematical shorthand $Y \sim p(y)$ states that your DATA Y are produced by a probability model having mathematical form $p(y)$. The expression $Y \sim p(y)$ is just a shorthand notation for the graphical model shown in Figure 1.3.

1.4 Deterministic Models

A **deterministic model** is a model where an outcome y is completely determined by an input x. It is a mathematical **function** $y = f(x)$ that allows you to make predictions of y based on x. Here, $f(.)$ is used rather than $p(.)$ as shown in Figure 1.3 to underscore the distinction between deterministic and probabilistic models. The driving time model of Example 1.1 is an example of a deterministic model: There, $f(x) = x/100$. This type of model is deterministic because y is completely determined if you know x: Given $x = 310$, there is one and only one possible value for y according to this model, namely, $y = f(310) = 3.10$.

In case you have forgotten, and since we use a lot of functions in this book, here is a refresher on the meaning of a mathematical *function*. A function is a mapping of values of x to values of y such that, given a particular value in a relevant range of x values, there is one and only one resulting y value. For example, $y = x^2$ maps $x = 4$ to only one value, namely, $y = 16$, and is a function of the form $y = f(x)$. But if $y^2 = x$, then $x = 4$ corresponds to two values of y, namely, $y = 2$ and $y = -2$; hence, $y^2 = x$ is not a function of the form $y = f(x)$. A deterministic model $y = f(x)$ states that for a given x, there can be one and only one possible value of y, namely, $y = f(x)$.

Example 1.5: A Deterministic Model for a Widget Manufacturer's Costs

Suppose that y represents the total yearly cost of your business, x represents how many widgets you will make in your company per year, c is your collective fixed cost, and m is your cost to produce each widget. Then a simple model that relates the number of widgets you make to your total cost y comes from the slope–intercept form of a deterministic straight-line model:

$$y = c + mx.$$

You are probably quite familiar with deterministic models like the slope–intercept equation earlier from your previous math courses. These models are often useful for describing basic relationships between quantities. However, these models have a major weakness in that they do not explicitly account for variability in Nature that we see and experience in every second of our lives. Because there is variability in the real world, deterministic models are obviously wrong. They tell you that things are perfectly predictable, with no variation. While probabilistic models are not exactly correct either, they are more realistic than deterministic models because they produce data that vary from one instance to another, just as you see in Nature. Deterministic models, on the other hand, produce data with no variability whatsoever.

As side note, if you happen to have read something about *chaos theory*, then you know that there are deterministic models that look a lot like probabilistic models. Go ahead and have a look—there are plenty of fun things to discuss about chaos theory, probability, determinism, and free will, perhaps with your colleague Hans while enjoying a pint at the local pub!

1.5 Variability

Do you eat precisely the same food every day? Shop at exactly the same stores? Arrive at work at precisely the same instant? Does Hans prefer the same brand of toothpaste as Claudia? If you saw somebody jump off a cliff, would you do it too? The answer to all these questions is, of course, a resounding "No!" And aren't we lucky! If everything were the same as everything else, imagine what a dull world this would be.

Variability is everywhere. Every time you drive 310 km, it takes a different amount of time. Every day the stock markets go up or down, different from the day before. Hans does not buy the same toothpaste as Claudia. Everybody lives to a different age. If you roll a die 10 times, you won't get the same result every time. One spoonful of vegetable soup is not identical to another spoonful. Variability is so real, you can taste it!

Deterministic models are obviously wrong because the data they produce do not exhibit variability. Every time you plug in $x = 310$ in the equation $y = x/100$, you will always get $y = 3.10$. Try it a few times: Plug $x = 310$ into the equation $y = x/100$ and calculate. Repeat, repeat, repeat. Do you ever get a different y?

You must use **probabilistic** (or **stochastic**) **models** to account for natural variability. In Example 1.1, your actual driving time Y is variable, because your average speed changes depending on variables like road conditions, city versus highway driving, your attitude about speeding, and on your need for bathroom breaks! Thus, your driving time Y is not precisely equal to $x/100$; rather, it deviates from $x/100$ by a variable amount.

Are deterministic models ever true? Perhaps in the physical and engineering sciences? Rarely, if ever! In physics, you will see deterministic models that purport to govern the physical universe, but these models have idealized assumptions that are not precisely true, leading to actual outcomes that vary from the model's predictions. For example, the models used by NASA to predict the location where a Martian rover will land will be wrong every time (although not by much), because of numerous uncontrollable factors. Further, such deterministic models often break down completely at the quantum level, where variability and randomness take over. Finally, experimental validations of physical models of the universe result in measurements that vary from experiment to experiment, again requiring statistical and probabilistic models to analyze the data.

For an engineering example, consider the maximum stress that a dam can bear. This value cannot be predicted perfectly. It depends on many unknown variables, such as the type and preparation of concrete down to the atomic level, the exact quality of the construction, and the behavior of the dam in its environment in real time. It is impossible to characterize this information so completely as to arrive at a deterministic prediction of the maximum stress level that the dam can bear at any given moment. Neither the physicist nor the engineer can tell you precisely what will happen, despite all of their wonderful deterministic mathematical models.

But, truth be told, deterministic models are often at least approximately correct in the physical and engineering sciences and give reasonably accurate predictions in spite of their failings. This happens when the variability (called *noise* in their jargon) is tiny relative to the deterministic component (called the *signal*). Thus, while deterministic models are wrong, they can still be useful.

In the social sciences, on the other hand, deterministic models are usually just plain silly. Can you precisely determine tomorrow's Dow Jones Industrial Average? No. Can you precisely determine how much money Jae Hwa will spend on a particular trip to the market? No. Can you precisely determine what answer Alejandra will enter on a survey, when asked about her coffee preference? No.

Nor are relationships deterministic in the biological and medical sciences. Can you precisely determine whether a diseased patient will survive 5 years? No. Can you precisely determine how many deer will be born in a drought season? No. Can you precisely determine whether a child born of two brown-eyed parents having recessive blue eye color genes will have blue eyes? No. On the other hand, you can predict all of these things very well by using probability models, but only in an aggregate sense—not individually.

In summary, you need probabilistic models in all areas of science. Deterministic models are obviously wrong because the data they produce lack variability, unlike the real data that you see. If the variability is tiny relative to the deterministic component, then you can still use the deterministic model; otherwise, you should use a model that includes a probabilistic component if you want realistic predictions.

Probabilistic models assign likelihoods to the outcomes of interest, rather than assigning a determination that a certain outcome will occur with 100% certainty. And while 100% certain determinations can be more comforting, likelihoods are more realistic and are quite useful for making decisions. For example, if 95% of stage II ovarian cancer patients survive 5 years when given therapy A, and if only 80% of them survive 5 years when given therapy B, then, all other things being equal, you would choose therapy A for treatment. This does not mean that, in hypothetical worlds (also called *counterfactual* worlds) where you could somehow play out your potential futures using either therapy, you would always live longer with therapy A. What it does mean is that you have a better *chance* of living 5 years with therapy A. In these counterfactual worlds, you estimate that in 95% of them you would live 5 years or more with therapy A, while in only 80% of them you would live 5 years or more using therapy B. You decide: Do you want A or B?

You might find probability models challenging because they have a strong conceptual component. Just look at the previous paragraph: While the choice of therapy A seems obvious, the rationale for preferring therapy A involves potential, counterfactual worlds and is therefore quite conceptual. Guess what: Statistics and probability require imagination! You probably didn't think you would have to use your imagination in a statistics class, did you?

Most students learn about Nature in a categorical style that emphasizes "right" and "wrong." For instance, your teachers may have said things like "*Answer A* is wrong,

Answer B is right, *Answer C* is wrong, and *Answer D* is wrong. Therefore, you should fill in the bubble for *Answer B*." Categorical thinking is so natural, fundamental, and well rehearsed for people (and probably animals) that probabilistic thinking may seem unnatural in comparison. Indeed, probabilistic investigation as a science is much more recently developed in human history. It is not well rehearsed in daily life and must be learned through a course of study such as you will find in this book. Although deterministic models can be used in an attempt to assign absolute truths, such as "If I drive 310 km at 100 km/hour, then it will take me precisely 3.10 hours to reach my destination," these kinds of determinations are in fact 100% false! You will *never* arrive *precisely* 3.10 hours later, at least when time is measured precisely, say, by using a stopwatch. Probabilistic models are much more realistic, giving you predictions such as "If I drive 310 km, I will arrive in less than 3.50 hours 90% of the time."

Before discussing probability models more formally, we must introduce the concept of a *parameter*, a concept that applies to both deterministic and probabilistic models.

1.6 Parameters

Whether deterministic or probabilistic, models have parameters that govern their performance. A **parameter** is a numerical characteristic of the data-generating process, one that is usually unknown but often can be estimated using data. For example, suppose you don't know the sales tax rate. A model for the amount you pay is as follows:

$$y = \text{Round}\{(1+\rho)x\}$$

Here, the variable x is the price of the object before the tax.

The variable ρ is the Greek lowercase letter rho, pronounced "row," and is a parameter of the model; you can estimate it using transaction data. Here and throughout the book, unknown parameters are denoted by Greek letters such as ρ, θ, μ, σ, β, λ, δ, and π.

Note that, even though ρ is unknown, this model still produces the data y. This is the ordinary situation: Models produce data, but they have unknown parameters.

Much of the statistics you will do or see involves **statistical inference**. Statistical inference is the science of using data—produced by Nature as tapped through design and measurement—together with assumptions about the data-generating process (which this book covers), to make defensible conclusions about Nature. The probability that is present in statistical models comprises an essential component of statistical inference, as it allows you to quantify the effects of chance variability on your data and thereby separate the real conclusions from those that are explainable by chance alone. We will devote considerable time in the chapters ahead to deciding whether your statistical results are explainable by chance alone.

Here, we come to the main Mantra that will be repeated throughout this book, again and again. Memorize it now!

Mantra #1:

Model produces data.
Model has unknown parameters.
Data reduce the uncertainty about the unknown parameters.

Here and throughout the book, the Greek lowercase letter theta, θ, denotes a **generic parameter**. Thus, θ could represent a tax rate, a slope, an intercept, or another quantity, depending on the specific application. In our Mantra, there may be more than one parameter (e.g., mean and standard deviation) yet we still call the parameters, collectively, "θ." In the case where θ is comprised of a list of values, it is called a parameter **vector**. It will be clear from context whether θ refers to a single parameter or a list of parameters.

While it may seem abstract to use algebraic symbols like θ to denote parameters, there really is no other way, because *model has unknown parameters*. You can reduce your uncertainty about the values of these parameters, but you cannot eliminate your uncertainty outright. Instead, you need to use probabilistic analysis: You can never make a claim like "The parameter is certainly equal to 4.5," but you will be able to state something like "The parameter is most likely between 4.3 and 4.7." *Data reduce the uncertainty about the unknown parameters*. Data do not eliminate the uncertainty about the unknown parameters.

1.7 Purely Probabilistic Statistical Models

A purely probabilistic statistical model states that a variable quantity Y is generated at random. This statement is represented mathematically as $Y \sim p(y)$, where $p(y)$ is a **probability distribution function (pdf)**, a function that assigns relative likelihoods $p(y)$ to the different observable values of the data y. The function $p(y)$ tells you what kind of numbers you will see: If $p(y_1)$ is large, you will see relatively many values of Y near y_1; if $p(y_2)$ is small, you will see relatively few values of Y near y_2. In Figure 1.4, a model for time it takes to drive 310 km, you can see an example: When the function $p(y)$ is large, for example, when $y = 3.1$, many of your driving times are near $y = 3.1$. When $p(y)$ is small, for example, when $y = 3.2$, few of your driving times are near $y = 3.2$.

NOTE: The expression $Y \sim p(y)$ is quite different from the equation $Y = p(y)$. The expression $Y \sim p(y)$ states that Y is *produced by the function* $p(y)$, while the expression $Y = p(y)$ states that

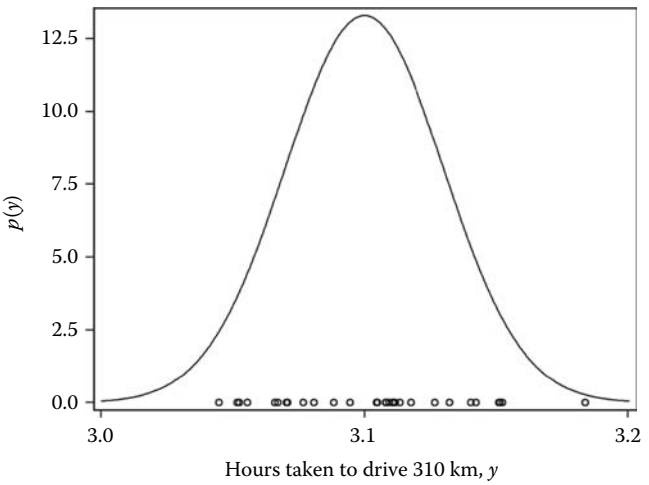


FIGURE 1.4
A model that produces data ($p(y)$, solid curve), and a sample of data produced by that model (circles).

Y is *equal to the function* $p(y)$. In Figure 1.4, the data values of Y appear on the horizontal axis, while the function $p(y)$ is the curve. In this book, you will *never* see the expression $Y = p(y)$.

The parameters of the model are never known, because *model has unknown parameters*. Figure 1.4 shows a precise curve $p(y)$ that produces Y, but in practice you will never know this precise curve. A main goal of statistical analysis is to estimate the parameters that make this curve what it is (*data reduce the uncertainty about the unknown parameters*). A statistical model, then, is a statement that your data are produced by a model with unknown parameters. In the purely probabilistic case, the definition is as follows:

Definition of a Purely Probabilistic Statistical Model

A purely probabilistic statistical model states that a variable Y is produced by a pdf having unknown parameters. In symbolic shorthand, the model is given as $Y \sim p(y|\theta)$.

Note the distinction between Y and y in the expression $Y \sim p(y|\theta)$. Capital Y refers to a single random outcome, and lower case y refers to fixed realization of Y. Earlier in the discussion, we referred to DATA in uppercase letters, and this upper case Y is equivalent to DATA, because it refers to the case where the data are not yet observed. The circles on the horizontal axis of Figure 1.4 are observed and therefore constitute lowercase "d" data.

This distinction between uppercase "D" and lowercase "d" is extremely important for your understanding of probability models and for your understanding of how to think in probabilistic terms. But what *is* probability, anyway? You often see percentages used to communicate probabilities, and this is indeed a good way to think about them. If the probability of A is 40%, then in (roughly) 40 out of 100 instances, A will occur, and in the other 60 instances, A will not occur. For example, if the probability of a die showing 1 is 1/6, or 17%, then in (roughly) 17 out of 100 rolls of the die, you will see a 1. You can also see this using the computer and a spreadsheet program: If you produce DATA* from a model where $p(1) = 1/6$, then roughly 17 out of 100 Y*s will have the value 1.

As with any model, the probability model is a *mental* conception. With the die, you *imagine* that about 17 out of 100 rolls will produce a 1, but this is only based on your mental assumption that the die is fair. What if the die is a trick die, with no 1 on it? Or what if the die is loaded so that the 1 rarely comes up? Then your mental model is wrong. A more believable mental model would be one that states that the probability of seeing a 1 is an unknown parameter θ. You can never know the precise numerical value of this parameter θ, but you can estimate it using data (*data reduce the uncertainty about the unknown parameters*).

In some cases, the 100 instances (e.g., rolls of the die) that you can use to understand probability are completely in your mind, as opposed to being real-world actions such as physically rolling the die. For example, what is the probability that the best football team in the league will beat the worst one in tomorrow's game? Here, the 100 instances would have to be repeated plays of the game under identical circumstances, much like rolls of a die. But it is impossible to play the game over and over with exactly the same people, weather, fan support, etc. Instead, you have to imagine *potential futures*: In 100 *potential future* plays of the game that you can imagine, how many times will the best team win? The number of wins in 100 potential futures depends on your personal judgment. So, what do you think? You might understand what you think a little better by putting some money on the line! If you are willing to bet 1 (dollar, euro, pound, etc.) in hopes of winning 10 (so your net earnings is $10 - 1 = 9$), then you think the probability is 10% that the underdog will win: In 10 out of 100 of your potential futures, you will net 9, for a total of 90 won, and in the remaining 90 out of 100 of your potential futures, you will lose 1,

for a total of 90 lost. Thus, over all the potential futures that you can imagine, you will come out even. This type of balancing of payouts is the way that professional oddsmakers assign probabilities.

A probability model $p(y)$ does not care whether the 100 instances correspond to physical or mental reality. It's just a model for how the future data will appear, no matter whether the futures are potential or actual. Either way, $p(y)$ will produce data for you when you use the computer—for example, you can use the computer to play out a future football game repeatedly under identical conditions, getting different outcomes from one potential future to the next.

The probability model allows you to make what-if predictions as to the value of Y, but unlike the deterministic model, it does not presume to know what the precise value of Y will be. For example, in the car driving time example, a probability model would not produce $y = 3.10$ (hours) when $x = 310$ (km); rather, it would produce random values in a neighborhood of 3.10, such as 3.09, 3.14, 3.14, 3.08, 3.19, 3.13, 3.12, …, as shown in Figure 1.4. This model is much more realistic than the deterministic model, because in repeated driving of the 310 km distance, your driving times will vary similarly.

Example 1.6: A Probability Model for Car Color Choice

Suppose you wish to predict whether the next customer will buy a red car, a gray car, or a green car. The possible values of Y are *red*, *gray*, and *green*, and the distribution $p(y)$ might have the form shown in Table 1.3.

Probability distributions are best understood using graphs. Figure 1.5 shows a **needle plot** of the distribution. A **bar chart**, possibly called a *column chart* by your spreadsheet software, is another similar, commonly used graph to depict a probability distribution.

In Figure 1.5, the vertical lines (the "needles") are in place to make the graph easier to read and are not technically part of the function $p(y)$. The pdf could have simply been depicted using only the solid dots on top of the lines, with no vertical lines.

The model of Table 1.3 does not tell you precisely what the next customer will do; the model simply says it is random: Y could be red, gray, or green. However, the model does allow you to make aggregate what-if predictions as follows: "If I sold cars to the next 100 customers, then about 35 of them would buy a red car, 40 would buy a gray car, and 25 of them would buy a green car." You should say "*about* 35" because the actual number is unknown. However, the *law of large numbers*, covered in Chapter 8, states that the sample proportion from actual data gets closer to the true probability from the model as the sample size increases.

TABLE 1.3

Probability Distribution of Color Choice

Color Choice, y	$p(y)$
Red	0.35
Gray	0.40
Green	0.25
Total	1.00

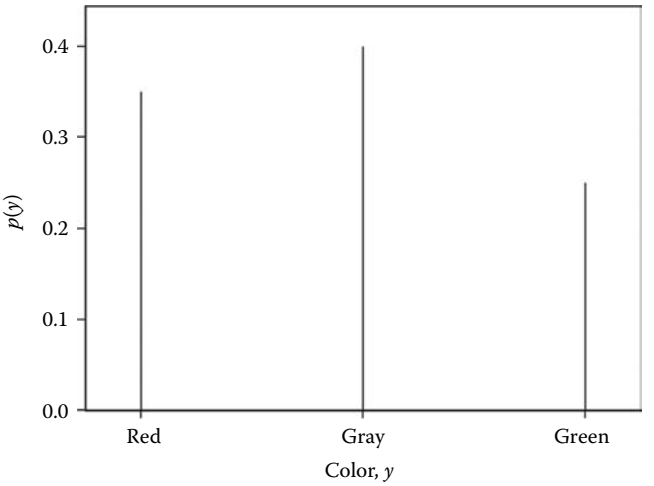


FIGURE 1.5
Graph (needle plot) of the probability distribution of *Y*.

Again, this prediction does not necessarily concern an event in the past, future, or even the present. It is simply a hypothetical, what-if statement about what would be likely to happen in a given scenario.

This model contains a very consequential assumption about reality: It assumes that only three possible choices of car color are possible. This implies that, in the universe described by this model, no customer will ever choose blue, orange, brown, white, or any other color for their car. Is this a good **assumption**? It might be, if you model the sales of cars at a dealership that sells only red, gray, and green cars. But it is a bad assumption if you model a dealership that offers more variety and a completely useless one if you model a dealership that only sells black and white cars.

This is a good time to restate the basic concept for evaluating the quality of a model. A model is good if the data it produces (recall that this is denoted as DATA* when generated by a computer) look like the data (denoted as DATA previously) produced by Nature. In this example, if you go through the sales records and notice that a brown car was sold on Thursday, then you would question the usefulness of the model, because the only DATA* you get from the model will be red, gray, and green.

The model of Table 1.3 is also bad if a sample of sales records data show drastically different percentages for the various colors, such as 10%, 10%, and 80%, rather than the 35%, 40%, and 25% anticipated by your model. This is a distinction between probability models and statistical models. Probability models assume specific values of the parameters, statistical models do not. Instead, in statistical models, the probabilities are always unknown parameters. (*Model has unknown parameters*.) This makes statistical models more believable in that the probabilities could be any numbers. If you think about it, how could you possibly know what the percentages in Table 1.3 really are? You can't. These percentages are always algebraic unknowns; we'll call them θ_1, θ_2, and θ_3 rather than 35%, 40%, and 25%; or 10%, 10%, and 80%; or anything else. It is believable that the true percentages are *some* numbers θ_1, θ_2, and θ_3, but it is not believable that the percentages are specific values like 35%, 40%, and 25%.

Thus, the requirement for a good statistical model is that the DATA* produced by the model look like the actual DATA for *some settings of the parameters*. You do not have to know what those parameter values are.

1.8 Statistical Models with Both Deterministic and Probabilistic Components

The model with both deterministic and probabilistic components is a **regression model**, which is a model for how the distributions of Y change for different X values. The regression model is represented as follows:

$$Y \sim p(y \mid x)$$

The symbol $p(y|x)$ is read aloud as "the probability distribution of Y given a particular X." The symbol | is shorthand for "given" or "given that." The model $Y \sim p(y|x)$ reads, in words, as follows:

> For a given $X = x$, Y is generated at random from a probability distribution whose mathematical form is $p(y|x)$.

While more cumbersome, the following notation is a more specific and more correct shorthand to represent the regression model:

$$Y \mid X = x \sim p(y \mid x)$$

In the example with Y = driving time and X = distance, the model states that "For a given distance $X = x$, driving time Y is generated at random from a probability distribution that depends on $X = x$, whose mathematical form is $p(y|x)$." In other words, there is a different distribution of possible driving times when X = 100 km than when X = 310 km (shown in Figure 1.4). This makes sense: While the relationship between Y and X is not deterministic, it is certainly the case that the time Y will tend to be much longer when X = 310 km than when X = 100 km; hence, the distributions of Y differ for these two cases.

In the regression case, the parameters of the model are also never known. Hence, the definition of the statistical model is as follows:

Definition of Statistical Model with Both Deterministic and Probabilistic Components

> This model states that, given the value of a variable $X = x$, a variable Y is produced by a pdf that depends on x and on unknown parameters. In symbolic shorthand, the model is given as $Y \mid X = x \sim p(y|x, \theta)$.

This model also allows you to make what-if predictions as to the value of Y. Like the deterministic model, these predictions will depend on the specific value of X. However, since it is also a probabilistic model, it does not allow you to say precisely what the value of Y will be; as shown in the previous example, probabilistic models only allow you to make what-if predictions in the aggregate.

Take the car example previously. If X = age of customer, then the distribution of color preference will depend on X. For example, when X = 20 years, your distribution might be as shown in Table 1.4 and graphed in Figure 1.6.

But when X = 60 years, your distribution might be as shown in Table 1.5 and graphed in Figure 1.7.

The model does not tell you precisely what the next customer will do, but does allow aggregate what-if predictions of the following type: "If I sold cars to the next 100

TABLE 1.4

Probability Distribution of Color Choice for 20-Year-Old Customers

y	$p(y \mid X = 20)$
Red	0.50
Gray	0.20
Green	0.30
Total	1.00

TABLE 1.5

Probability Distribution of Color Choice for 60-Year-Old Customers

y	$p(y \mid X = 60)$
Red	0.20
Gray	0.40
Green	0.40
Total	1.00

20-year-old customers, then about 50 would buy a red car, 20 would buy a gray car, and 30 would buy a green car." Similarly, you can say "If I sold cars to the next 100 60-year-old customers, then about 20 would buy a red car, 40 would buy a gray car, and 40 would buy a green car."

There are so many models to learn—probabilistic, deterministic, and the combination of the two. But really, it's easier than you might think: Just memorize the combination model of this section. The purely probabilistic model is a special case of it, one where the distribution of Y does not depend on X. And the deterministic models that the physicists and engineers use so much are also special cases, ones where the distributions have no variability.

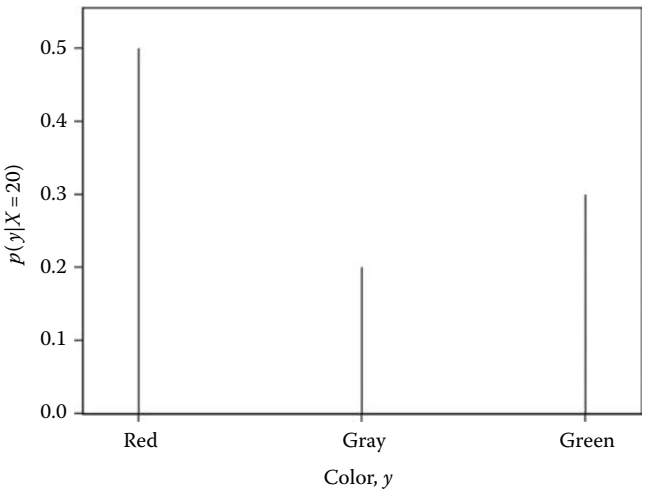


FIGURE 1.6
Graph of the probability distribution of Y when $X = 20$ years.

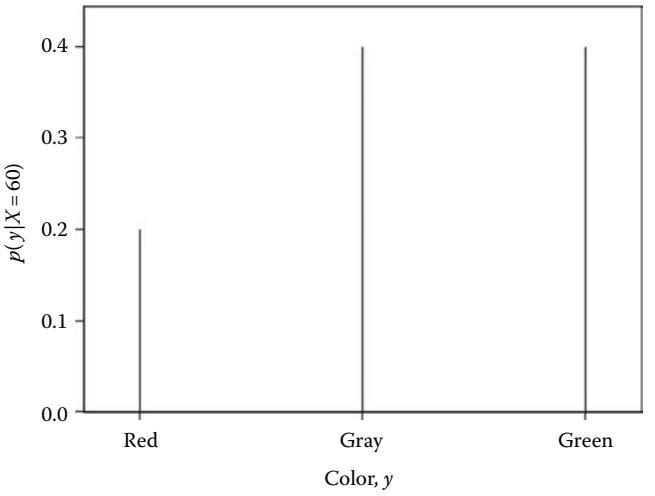


FIGURE 1.7
Graph of the probability distribution of Y when $X = 60$ years.

For example, the deterministic driving time model states that when $x = 310$, then y is equal to 3.10 with 100% probability, and there is no variability. For another deterministic example, a model that states that 100% of 60-year-olds buy gray cars also has no variability in the distribution. And these models are obviously incorrect, like all deterministic models! No matter which models you like, they are all special cases of the combination model. Thus, you can say that the model with both deterministic and probabilistic components is the *mother of all models*.

1.9 Statistical Inference

Recall the Mantra: *Model produces data. Model has unknown parameters. Data reduce the uncertainty about the unknown parameters.*

How does this work? The simple coin toss thought exercise provides the entire mental framework you need to understand even the most complex of statistical models. So if you toss a coin, it's a 50% heads and 50% tails, right? Wrong! It's close to 50–50, but due to slight imperfections in every coin—resulting in one side being just a tad heavier than the other—the probability of "heads" is not exactly 0.5000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000… with the zeroes continuing ad infinitum, but instead something slightly different, such as 0.5000000000000000000000000000000000000000000000000000000000000000000320213245342222342007 88…

For all intents and purposes, this number is so close to 0.5 that you could go ahead and assume 0.5, but while highly accurate, this assumption is not actually true.

Let's make the example more interesting. Take a pair of pliers and bend the coin. Now what is the probability of heads? You can no longer assume it's 0.5. All you can say is that it is simply π, some unknown probability. This π is an example of an unknown parameter in the Mantra *Model has unknown parameters* and is also an example of the generic θ mentioned previously. The model that produces the data is therefore as given in Table 1.6.

TABLE 1.6

Probability Distribution for a Bent Coin

Outcome, y	$p(y)$
Tails	$1-\pi$
Heads	π
Total	1.00

We apologize for the abuse of notation here, as the Greek letter π is more commonly used as the famous trigonometric constant $\pi = 3.14159\ldots$. We will use π that way later when we discuss the normal pdf—the famous bell curve graphed in Figure 1.4. Meanwhile, in this coin toss example, π is simply a number between 0 and 1, the unknown probability of getting heads when you flip the bent coin.

How can you learn about this model? The simple answer is "Collect some data!" (*Data reduce the uncertainty about the unknown parameters.*) Flip the bent coin many times, and count how many tosses turn up heads. If the proportion is 3 out of 10, or 30%, you now have a better idea about π: It is somewhere near 0.30. Your uncertainty about the unknown parameter π is reduced when you have data.

However, you are still uncertain: The true π is not 0.30; it is still the same unknown value that it was before. By analogy, if you flip a fair coin 10 times and get three heads, you shouldn't think the probability is 0.30; you should still think it is 0.50 (or darn close to 0.50 as discussed previously).

The model still produces the data. The data do not produce the model. If you think the data produce the model, then you would think, based on 10 flips and three heads, that suddenly the coin's Nature has changed so that it now will give heads in 30% of the subsequent flips. The true π is not 0.30 anymore than it is 0.50 for a fair coin; it is still the same unknown value that it was before. The data you have collected only suggest that the probability of getting heads is *near* 0.30, not that it is *equal to* 0.30.

Now, how to apply the lowly coin toss example to something that resembles typical research? Simple. Refer to Table 1.3, the example of car color choice. The statistical model looks, in reality, as shown in Table 1.7.

Here, the numbers π_1, π_2, and π_3 are the unknown model parameters, again an example of a generic parameter vector $\theta = (\pi_1, \pi_2, \pi_3)$. The model is good in that DATA* produced by the model will look like the DATA that you actually see, *for some settings of the parameter* $\theta = (\pi_1, \pi_2, \pi_3)$. You do not have to know what the parameter values are to know that the model is good.

By collecting data, you can easily reduce your uncertainty about the parameters π_1, π_2, and π_3, although you can never determine them precisely.

TABLE 1.7

Probability Distribution of Color Choice

y	$p(y)$
Red	π_1
Gray	π_2
Green	π_3
Total	1.00

Model produces data. Data do not produce the model. Instead, data reduce your uncertainty about the unknown model parameters. The reduction in uncertainty about model parameters that you achieve when you collect data is called **statistical inference**.

A note on notation: While Table 1.7 shows the model as $p(y)$, we sometimes represent the model as $p(y|\theta)$ to emphasize that the model depends on the unknown parameter(s) θ. Usually, $p(y|\theta)$ is the more correct notation. We often use the notation $p(y)$ rather than $p(y|\theta)$, just for the sake of simplicity.

1.10 Good and Bad Models

Compare Figures 1.2 and 1.3. The model of Figure 1.3 is "good" if, for some parameter settings, the DATA* produced by the model "look like" the DATA that you see in reality (Figure 1.2). But why the quotes around the words *look like*? What does that mean, specifically?

To answer, make Figure 1.3 specific to the coin toss case. Also, assume a fair coin whose probability is exactly 0.5. (This example is hypothetical, since such a coin does not exist!) A model for this process is $Y \sim p(y)$, where Y can be either *heads* or *tails* and where $p(y)$ is given as in Table 1.8.

This distribution is closely related to a special distribution called the **Bernoulli distribution**. (In Chapter 2, we cover this distribution and others in more detail.) The Bernoulli distribution produces 0s and 1s instead of heads and tails, but you can easily recode a 0 as tails and 1 as heads to arrive at the distribution in Table 1.8. You can do this in Microsoft Excel, after adding in the Data Analysis toolpack. Once you select "Random Number Generation" from the "Data Analysis" menu, the screenshot should look something like Figure 1.8.

Click OK and the result looks as shown in Figure 1.9. (Note that your numbers may differ due to randomness.)

You can recode the zeroes and ones to tails and heads as shown in Figure 1.10.

So the result is the sequence of heads, tails, heads, heads, tails, tails, heads, heads, heads, and tails. This is an example of the DATA* that can be produced by computer random number generators. However, since the data are in hand, as opposed to being in a potential future, you should call them data* instead of DATA*.

Figure 1.10 shows an example of DATA* for the coin toss case. But what do the actual DATA look like? Well, the DATA are what you would get if you actually flipped the coin 10 times. For example, a real sequence of coin flips—from actual coin tossing, not from computer generation—might be heads, heads, tails, tails, tails, tails, tails, heads, heads, and tails. This is an example of what is meant by DATA, and since these values are now in hand (as opposed to being in a potential future), you should call them data.

TABLE 1.8

Probability Distribution for a Fair Coin

y	$p(y)$
Tails	0.5
Heads	0.5
Total	1.0

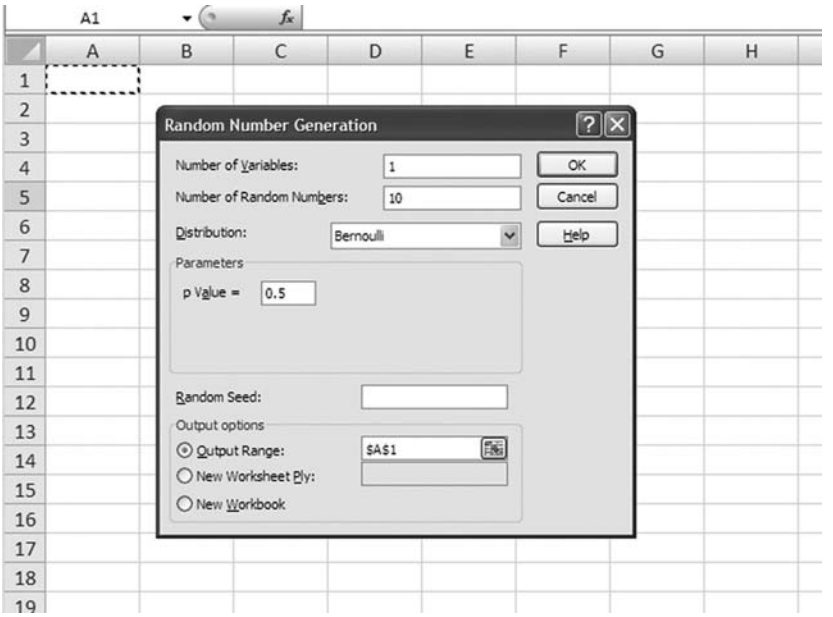


FIGURE 1.8
Generating Bernoulli random numbers using Microsoft Excel®.

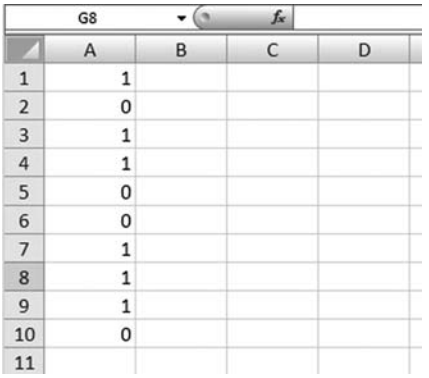


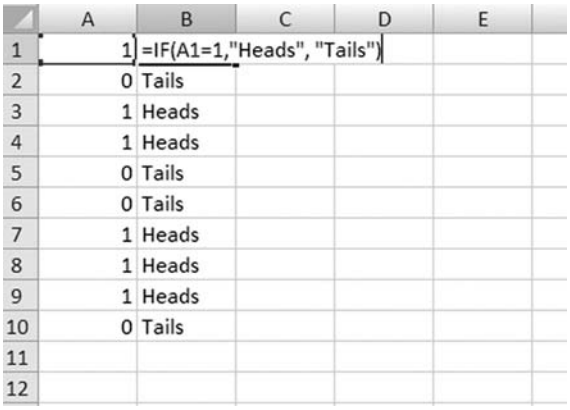
FIGURE 1.9
A sample of $n = 10$ observations produced by the Bernoulli(0.5) distribution.

The model is good if the DATA* produced by the model look like the real DATA. But if you compare the two actual sequences of computer-generated data* and the actual coin-tossed data, they won't match, flip for flip. So is the model still "good"? Yes! In fact it is an excellent model.

Definition of a Good Model

A model is **good** if:

a. For some parameter settings, the *set* of possible outcomes produced by the model well matches the set of possible outcomes produced by Nature, design, and measurement.
b. For some parameter settings, the *frequencies of occurrences* of the specific outcomes, as well as successive combinations of outcomes, well match the *frequencies of occurrences* of the specific outcomes and successive combinations of outcomes produced by Nature, design, and measurement.



	A	B	C	D	E
1	1	=IF(A1=1,"Heads", "Tails")			
2	0	Tails			
3	1	Heads			
4	1	Heads			
5	0	Tails			
6	0	Tails			
7	1	Heads			
8	1	Heads			
9	1	Heads			
10	0	Tails			
11					
12					

FIGURE 1.10
Recoding the Bernoulli data to create coin toss data.

The Bernoulli(0.5) random number generation in Excel is a model that passes on both counts: (a) the set of possible outcomes is {heads, tails}, exactly the same as that in Nature, and (b) the frequencies of occurrences are reasonably similar—both near 50% heads. Note that with more data and data*, these frequencies can be ascertained better; in the previous example, there are simply not enough data to make a firm judgment. Nevertheless, the model does a very good job of meeting our criteria (a) and (b) for a good model, and it doesn't really matter that it consistently fails to produce exactly the same sequence of heads and tails that you would get if you manually tossed the coins. It would actually be kind of creepy if that happened, right?

What does a "bad" model look like? Here are two examples of "bad" models for the coin toss process.

Bad Model #1: For toss i, where $i = 1, 2, 3, \ldots$, the outcome is heads if i is odd and tails if i is even. The sequence is thus alternately heads, tails, heads, tails, heads, and so forth. This model seems okay at first: The set of values produced is {Heads, Tails}, just like in Nature, and the frequency of heads is 0.5 as it should be. Where it fails is in the frequencies of occurrences of *successive* outcomes. The successive outcome "heads followed by heads" is impossible with this model, but very frequent in reality: In 25% of adjacent flips, both will be heads.

Bad Model #2: The Bernoulli distribution does a very good job of modeling coin flips. What about another distribution? The *normal* distribution is the most commonly assumed distribution in all of statistics. How does it work here? You can use the normal random number generator and produce some values as shown in Figure 1.11.

Figure 1.12 shows a sample from the normal distribution. Your numbers may vary due to randomness.

The numbers shown in Figure 1.12 are another example of DATA*—that is, data produced by a computer's random number generator. Is this model good? Clearly not since the set of outcomes produced consists of numbers filling a continuum between approximately −3 and +3, which do not at all match the discrete, whole integer outcomes {0, 1}.

Figure 1.4 shows another example of the normal distribution. It has a famous "bell curve" shape, producing DATA* values in the middle of the curve more often and DATA* at the extremes less often.

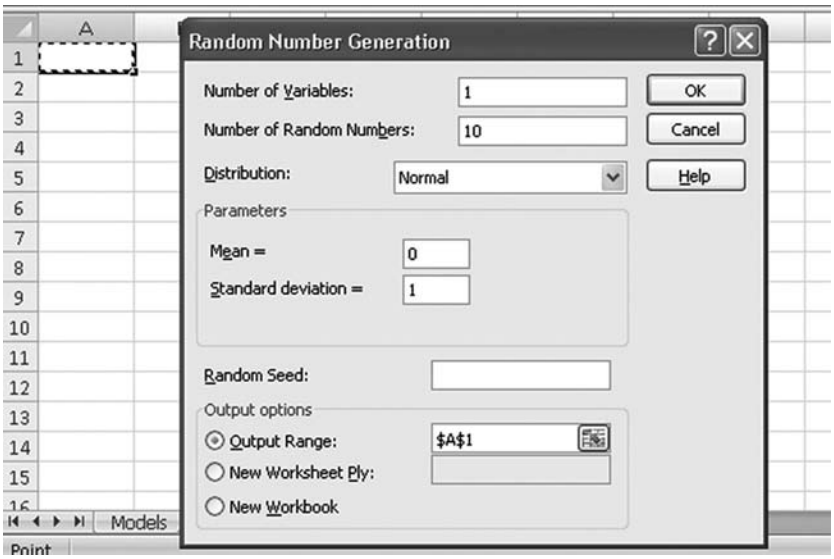


FIGURE 1.11
Generating values from a normal distribution.

A1 f_x 0.400232238462195

	A	B	C	D	E	F
1	0.400232					
2	-2.09131					
3	0.111575					
4	-0.81301					
5	0.625888					
6	-0.32393					
7	2.890592					
8	-0.20003					
9	2.56714					
10	-1.76746					
11						

FIGURE 1.12
A sample of data* produced by a normal distribution.

You do not have to know the model's parameter values to know that it is a good model. That is a relief, because *model has unknown parameters* anyway. For example, a good model for a bent coin is the Bernoulli(π) model, since the Bernoulli(π) model produces 0s and 1s that look like the bent coin results (heads = 1, tails = 0) for some values of π between 0 and 1. For example, you could specify the parameter settings π = 0.20, 0.25, and 0.30 and have the computer produce Bernoulli data for each of these settings. The resulting DATA* would look like the results of flipping a coin with a particular kind of bend. Thus, the criterion for a model being good is that, for *some parameter settings*, the DATA* produced by the model look like the DATA that are actually observed. You don't have to know the actual parameter values; that's what you use the DATA for: *data reduce the uncertainty about the unknown parameters.*

In cases where the model has both deterministic and probabilistic components, there is an additional criterion that is sometimes used: A model may be called "good" if the probabilistic component is small relative to the deterministic component. Again, imagine you have a model that is able to predict precisely what color a person would choose for their car with 100% certainty: You would say this is a good model! In Chapter 17, we define the *R-squared* statistic, which is a measure of the size of the deterministic component relative to the probabilistic component in regression models.

1.11 Uses of Probability Models

Suppose you are comfortable that a model is good. "So what?" you should ask. "What in the world am I supposed to do with this model?" The answer is simple and very important: You can make predictions! You can do this by **simulation**, which means using the computer to produce DATA* from the model.

Example 1.7: Estimating the Probability of Getting 50% Heads in 10 Flips

If you flip a coin 10 times, you should get heads 5 times, right? Wrong! To test this, you could flip coins over and over again, generating DATA, and note how often you get 5 heads out of 10 flips. If you repeated the process 1000 times—more than 1000 would be even better for greater accuracy—you should get a very good estimate. But that would be tedious! Instead, you can let the computer do the work, generating DATA* instead of DATA: Create 10 columns and 1000 rows of Bernoulli values with $p = 0.5$ to simulate 1000 instances of flipping the coin 10 times. Then count how many of the rows, out of 1000, yield exactly 5 heads. Figure 1.13 shows how to generate the data in Excel.

Figure 1.14 shows the tallying of the number of heads in Column K of the spreadsheet. The command " = COUNTIF(K:K, " = 5")/1000 counts how many of the 1000 samples have exactly 5 heads and divides that number by 1000. We got 0.228, but your number may differ slightly due to randomness. Thus, only about 22.8% of the time will you get exactly 5 heads out of 10 flips. The true probability can be calculated here to be 24.6% using the *binomial distribution*, which we will not discuss. As this example shows, simulation provides an excellent and useful approximation to the true probability.

It may be surprising that, even with a fair coin, the probability of getting exactly 50% heads is somewhat low—since the probability is only 24.6%, most of the time you will *not* get 50% heads. Challenge question: What is the probability of seeing 50% heads when there are 100 flips?

Example 1.8: Choosing an Optimal Trading Strategy

Who cares about flipping coins? Let's *earn* some coins! You have heard the phrase, "Buy low, sell high," right? Makes sense! Suppose you buy shares of stock in a company that looks promising. When do you sell? When do you buy? Suppose you debate two strategies for buying and selling this particular stock over the next 250 trading days (roughly one calendar year).

Strategy 1: Buy and hold.

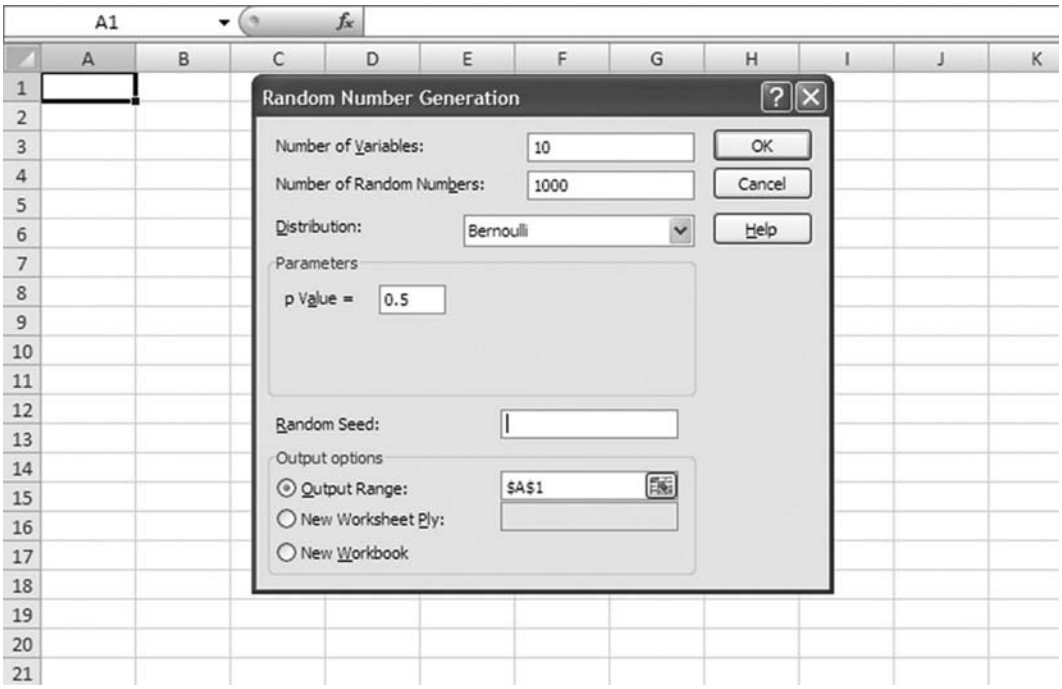


FIGURE 1.13
Generating 1000 samples of $n = 10$ Bernoulli observations per sample.

	A	B	C	D	E	F	G	H	I	J	K
1	0	1	0	1	0	1	0	1	0	1	=SUM(A1:J1)
2	1	1	0	1	1	0	1	0	0	0	5
3	0	1	1	1	0	0	1	0	1	1	6
4	1	0	1	1	0	1	1	1	1	0	7
5	0	1	0	0	0	0	1	1	0	1	4
6	0	0	1	0	1	0	1	0	1	1	5
7	1	0	0	0	1	0	0	1	1	0	4

FIGURE 1.14
Counting the number of heads per sample.

Strategy 2: Sell when there are three consecutive days where the stock price rises. Buy when there are three consecutive days where it drops. Otherwise, hold.

The rationale behind strategy 2 is your gut feeling that "what goes up must come down," and also, in the case of stock prices, "what goes down must come back up."

But is your gut feeling right? Which strategy is better? To determine the best strategy, you can create a realistic model to simulate possible future values of stock prices. You can then simulate 1000 potential futures, each containing the results from 250 consecutive trading days. Then you can try both strategies and compare your earnings after 250 trading days using each strategy. You will have 1000 values of earnings when using strategy 1 and 1000 values of earnings when using strategy 2. Then you can compare the two to find out which works better, on average, and pick the winner.

Here are the mathematical details. Let Y_t denote the price of the stock at time t. Suppose $t = 0$ is today, 6:00 p.m., so you know the price of the stock today from the financial reports. It is Y_0 and might be, for example, 23.32 dollars per share. Tomorrow, the price will be as follows:

$$Y_1 = Y_0(1 + R_1) \tag{1.1}$$

A little algebra shows that

$$R_1 = \frac{Y_1 - Y_0}{Y_0} \tag{1.2}$$

This is called the price **return** at day 1. Note that, since Y_1 is in the future and therefore unknown, R_1 is also unknown.

Extending (1.2) into the future, the model for future prices is

$$Y_t = Y_{t-1}(1 + R_t) \tag{1.3}$$

Thus, you can generate all the future price values Y_t if you only knew the future returns R_t. Of course, you don't know the future R_t values, but financial theory says that they behave remarkably like the coin tosses generated by Excel—except that, instead of being produced by a Bernoulli distribution, the return DATA look more like DATA* produced by a normal distribution, such as the one shown in Figure 1.12.

Figure 1.15 shows how to generate return DATA* and hence possible future trajectories of the stock price.

In Figure 1.15, you will notice that the normal distribution depends on two parameters, the mean and the standard deviation. These are very important statistical parameters and will be described in much greater detail later, starting with Chapter 9.

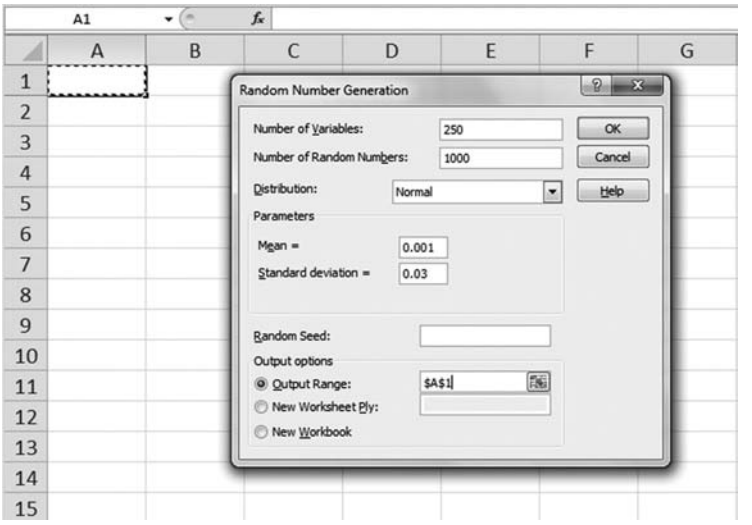


FIGURE 1.15
Generating 1000 potential future return sequences for the next 250 trading days.

For now, just suppose they are parameters that govern the particular normal distribution that you assume to produce your data: If you pick different parameter values, you get a different normal distribution. Here, θ = (mean, standard deviation) is an example of a parameter *vector*.

In Figure 1.15, the mean and standard deviation are set to 0.001 and 0.03, but these are just hypothetical values. Those parameters are never truly known, not by all the Economics Nobel Prize winners, not by billionaire financier Warren Buffett, and not even by the musician Jimmy Buffett, because *model has unknown parameters*. On the other hand, *data reduce the uncertainty about the unknown parameters*, so you can use historical data on stock returns to suggest *ranges* of plausible values of θ = (mean, standard deviation) and then perform multiple analyses for the parameters within those ranges, which is also called **sensitivity analysis**.

The simulated returns look as shown in Figure 1.16; your numbers may vary due to randomness.

Supposing today's price (at $t = 0$) is $y_0 = 23.32$, the potential prices are calculated as shown in Equation 1.3: $y_1 = 23.32(1 + r_1)$, $y_2 = y_1(1 + r_2)$, $y_3 = y_2(1 + r_3)$, etc. You can do this in another tab of the spreadsheet as shown in Figure 1.17.

At the end of the analysis, each row shown in Figure 1.17 is a potential future trajectory of the stock prices over the next 250 trading days.

A1 -0.00800696477401652

	A	B	C	D	E	F	G	H	I	J	
1	-0.00801	-0.03733	0.00833	0.03929	0.03695	0.05299	-0.06451	-0.00603	0.03385	-0.0316	-0
2	0.0152	0.02853	-0.00108	-0.00062	0.00744	-0.00956	0.02884	-0.01383	0.0713	0.02332	0
3	0.02219	-0.02165	0.01033	-0.03678	0.06951	0.02584	-0.03274	0.04577	-0.00717	-0.04164	-0
4	-0.00953	-0.01672	-0.00996	0.02639	-0.00277	0.05546	-0.01222	-0.03397	0.02017	-0.00207	
5	-0.01993	0.01759	0.02893	-0.02836	0.0408	-0.00226	0.02206	0.01128	-0.01662	-0.02543	-0
6	0.0254	0.05137	0.01859	0.01544	-0.0047	0.0024	-0.02688	-0.01987	0.03148	-0.00959	-0
7	0.03706	0.05832	0.01709	0.028	-0.02405	-0.00335	-0.01138	0.03185	-0.04959	0.0272	-0
8	-0.01813	-0.01812	0.05204	0.0054	0.04013	-0.01544	0.01625	0.02259	-0.01519	-0.01753	0
9	0.00441	0.05627	-0.0138	0.01377	-0.01423	-0.02384	0.02206	0.01765	-0.00524	0.03535	-0
10	-0.03806	0.02397	-0.025	0.05071	-0.0606	-0.02172	0.00522	0.01634	-0.00123	-0.01651	

FIGURE 1.16
Potential future return trajectories.

SUM =A2*(1+Sheet1!A1)

	A	B	C	D	E	F	G	H
1	0	1	2	3	4	5	6	
2	23.32	=A2*(1+Sheet1!A1)	22.4552	23.3375	24.1998	25.4823	23.8	
3	23.32	23.6745	24.35	24.3235	24.3086	24.4894	24.2553	24.9
4	23.32	23.8374	23.3213	23.5623	22.6956	24.2731	24.9004	24.0
5	23.32	23.0977	22.7114	22.4852	23.0785	23.0145	24.291	23.9
6	23.32	22.8551	23.2573	23.9302	23.2514	24.2001	24.1453	24.
7	23.32	23.9123	25.1407	25.6079	26.0032	25.881	25.9431	25.2
8	23.32	24.1843	25.5948	26.0323	26.7611	26.1175	26.0299	25.7
9	23.32	22.8972	22.4822	23.6522	23.7799	24.7342	24.3523	24.7
10	23.32	23.4228	24.7408	24.3994	24.7353	24.3833	23.802	24.3

FIGURE 1.17
Potential future price trajectories.

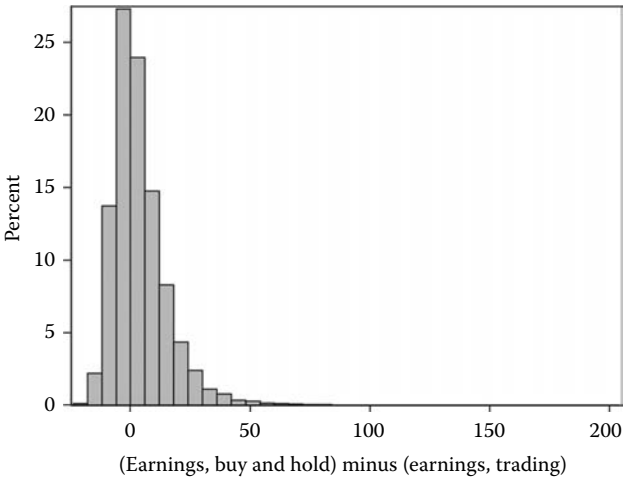


FIGURE 1.18
Histogram of earnings differences between buy and hold versus trading strategies.

With these future trajectories, you can try out each trading strategy to see which one nets you the most cash. For some potential futures, strategy 1 will work better, and for other potential futures, strategy 2 will work better. You want to pick the one that gives you more money on average, *over all potential futures*. Figure 1.18 in the following shows the distribution of the difference of your earnings, over 1000 potential futures, using the buy and hold versus trading strategy.

Figure 1.18 shows the **histogram**—an **estimate** of the probability distribution $p(y)$; see Chapter 4 for further details—of the difference between potential future earnings using strategies 1 and 2. For any particular future, the difference may be positive, meaning strategy 1 is preferred, or negative, meaning strategy 2 is preferred. Contrary to what your intuition might say, it seems that there are greater possibilities for much higher earnings with strategy 1—buying and holding—since Figure 1.18 extends much farther to the right of zero than to the left of zero. In fact, the average difference calculated from the 1000 potential futures is 3.71, meaning that you earn 3.71 more on average using strategy 1 than using strategy 2. So, on average, strategy 1 earns more. However, this does not guarantee that you will be better off using strategy 1 for the next 250 trading days. It only means that strategy 1 is better on average, over all potential futures, based on this model.

Example 1.9: Predicting a U.S. Presidential Election Based on Opinion Polls

Another example where simulation is very useful is in predicting the results of elections. In the United States, we do not directly elect our president. Rather, there is an *electoral college*, a system by which each of the 50 states and the District of Columbia contributes a certain number of votes based on its population. For example, at the time of writing of this book, California, the most populous state, contributes 55 electoral votes out of 538 total votes in the electoral college, whereas Delaware, a small state, contributes 3 votes. A presidential candidate needs to win a total of 270 electoral votes—a simple majority of 538—in order to win the election. Electoral votes are awarded based on the popular vote in each state. When a candidate wins the popular vote in a state, that state awards all of its electoral votes to that candidate. The other candidates get nothing. (Maine and

Nebraska use a different system, but this little detail has never altered the outcome of a presidential election.)

It can happen—and has happened—that a presidential candidate loses the overall popular vote but wins the election. Winning the presidency comes down to winning 270 or more electoral votes by winning the popular vote in any combination of states whose electoral votes total at least 270. This is why political analysts often talk about "battleground" states, where the vote is likely to be close and where the candidates would most benefit from spending their time and money. But how do the analysts suspect which states will have a close vote? Opinion polls!

Opinion polls provide a measure of the likelihood of a candidate winning a state. Using Bayesian calculations, as shown in Chapter 13, you can calculate the probability of winning based on the estimated proportion and the margin of error. For now, don't worry about too many details—we're just introducing the main ideas. Suppose there are two candidates and an opinion poll states that 47% of registered voters favor candidate A while 53% favor candidate B, with a 4% margin of error. In that case, assuming the opinion poll is accurate, then the probability that candidate A will win is 6.7%. In 6.7% of potential future scenarios, candidate A will win, and in the other 93.3% of the same scenarios, candidate B will win.

When you consider the combinations of all such future scenarios, with each state having different win probabilities and with different numbers of electoral votes at stake, along with the resulting potential range of future electoral college vote totals, the math seems daunting. But by using simulation, you can easily generate, say, 10,000 potential future scenarios, based on polling data for each state, and make an informed estimate as to who is likely to win. Each scenario gives a different electoral college tally. Figure 1.19 shows a histogram of 10,000 plausible values of the electoral college outcome, based on opinion polling data.

This simulation-based methodology is explained in much more detail by Christensen and Florence in their article "Predicting Presidential and Other Multistage Election Outcomes Using State-Level Pre-Election Polls," published in the journal *The American Statistician* in February, 2008.

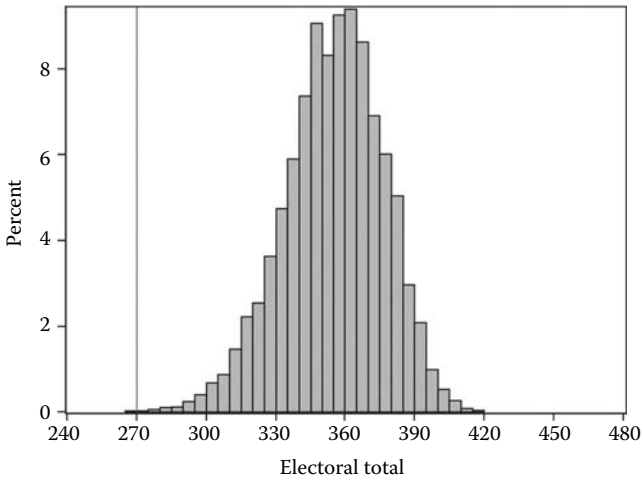


FIGURE 1.19
Histogram of the number of potential electoral votes for a hypothetical candidate, based on polling data. If the polls are accurate, the candidate is almost certain to win, since the number of votes will most likely be greater than the required 270, represented by the vertical line.

Vocabulary and Formula Summaries

Vocabulary

Nature	What is, was, will be, or might have been.
Statistics	The study of Nature using data.
Model	A mathematical representation of the outcomes of the processes of Nature, design, and measurement.
Prediction	A statement about something that might happen in Nature, be it in the past, present, future, or not at all.
Forecast	A statement about what will happen in the definite future.
Statistical model	A probabilistic recipe for how data are produced, one that depends on unknown parameters.
Primary data	Data you collected for a stated purpose.
Secondary data	Data collected for a different purpose; see *primary data*.
Design	A plan to collect data.
Measurement	The type of data to be collected.
DATA	As-yet unseen information produced from Nature, design, and measurement, also called Y.
data	The information after they are collected, also called y.
DATA*	The information to be produced by the model, also called Y^*. (See *simulation* in the following.)
data*	The information that has been produced by the model, also called y^*.
Binary response	Data that are dichotomous, such as 0 or 1 and yes or no.
Likert scale	A response scale used in surveys to indicate degrees of preference, typically comprised of items measured on a 1, 2, 3, 4, 5 scale.
Bivariate measurement	A measurement that consists of two numbers simultaneously.
Probabilistic model	The mathematical function called a pdf, typically written as $p(y)$; also a statement that DATA are produced by such a model.
Function	A mapping of values of x to values of y such that, given a particular value in a relevant range of x values, there is one and only one resulting y value.
Deterministic model	A model that always produces the same output, given the same inputs, typically written as $y = f(x)$.

Stochastic model	A *statistical model*, a *probability model*. Typically discussed in the context of time sequence data.
Parameter	A numerical characteristic of a natural process or model, usually fixed and unknown, indicated using the generic symbol θ.
Statistical inference	The method by which you learn about unknown parameters using data.
Generic parameter	Any parameter, denoted by the symbol θ.
Vector	A list of values.
Probability distribution function (pdf)	A function that assigns relative likelihoods to the different observable values of the data, typically written as $p(y)$.
Needle plot	A graph used to depict a discrete pdf. See also Chapter 2.
Bar chart	A graph used to depict a discrete pdf; see also *needle plot*.
Assumption	Something you stipulate about the model that you assume to produce your DATA.
Regression model	A model for how the distributions of Y change for different X values, written as $p(y\|x)$ or $p(y\|x, \theta)$.
Statistical inference	The reduction in uncertainty about your model parameters that you experience after you collect and analyze your data.
Bernoulli distribution	A probability distribution that produces the values 0 or 1.
Good model	A model where (a) the set of possible outcomes produced by the model well matches the set of possible outcomes produced by Nature, design, and measurement and (b) the frequencies of occurrences of the specific outcomes, and successive combinations of outcomes, well match the frequencies of occurrences of the specific outcomes produced by Nature, design, and measurement.
Bad model	One that is not good.
Simulation	Using the computer to produce DATA* from the model.
Return	The relative change from one time period to the next.
Parameter vector	A vector whose values are all parameters.
Sensitivity analysis	Multiple analyses with different plausible values of the parameters.
Estimate	A guess at the value of some entity.
Histogram	An estimate of a pdf.