CONTENTS

1

Life in a Random Universe

> I don't believe we know anything with complete certainty, but everything probably and to different degrees of probability.
> —CHRISTIAAN HUYGENS

June 8, 2004, was the first transit of Venus visible on Earth since October 6, 1882. Not wanting to miss the big event, I consulted a chart by Richard Proctor giving the precise minute of onset—the time when the small dot of Venus would begin its trip across the face of the sun. The chart was exactly right. Mr. Proctor drew it in 1874.[1] †

Impressive! If this kind of steel-trap precision is your idea of "science," then you would be mostly right in 1874, but considerably less so in the twenty-first century. What has become clear in the past hundred years is that steel-trap science is a luxury enterprise, a high-end shop inside a much bigger scientific emporium. A less precise but more flexible brand of science now shares the stage, helping us make sense of the world we live in, a random universe where truth most often lies concealed behind a foggy curtain of noise.[2]

1. The † mark in the margin indicates topics that are discussed further, along with references, located in the chapter's appendix.

2. "Noise" is shorthand for the kind of random disturbances that afflict observational data. Look ahead to figure 1.4 for an example.

Statistics was the first information science, preceding its upstart nephew, computer science, by two centuries. Astronomers have stars, geologists rocks, chemists elements, and statisticians have information. There is a big difference: stars, rocks, and elements are part of the natural world, but information is man-made, which means that statisticians don't have nature to fall back on as a check on their theories.

Statistical inference, what I'll be talking about in this book, is a collection of powerful ideas designed to see the truth behind the foggy curtain, and no more than the truth. The book's title, *To Think Like a Statistician*, presupposes that this is a good thing. It is! If sciences are judged by how often they appear in public media, then probability and statistics are superstars. My hope is to give you, the reader, a pleasant familiarity with the tricks of the trade, without equations, theorems, or long-winded tables.

"Tricks" isn't quite the right word. There's some underlying philosophy involved, but mainly I want to talk about a collection of intriguing topics, partly familiar in the news but still mysterious, and the methods statisticians have devised to understand them:

- Correlation and causation (Do tall children cause tall parents?)
- Bayes' Rule and its contentious history
- Reading the medical news (Is anything good for you? Bad for you?)
- Prediction (Why will your championship team probably do worse next year?)
- Accuracy and the bootstrap (What does "plus or minus" or "margin of error" really mean?)
- Crucial comparisons (How do we know the new drug is better than the old?)
- Survival and the perverse law of waiting
- Learning from the experience of others (I didn't have any accidents, so why did my car insurance premium go up?)

I have in mind a grand tour of the statistical point of view, hitting the high points while avoiding textbook prose. This chapter offers a preview of the tour, touching on several examples of statistical thinking in action in science, sociology, medicine, art, and sports.

These are powerful ideas, of 250 years' gestation. They enable scientists, and everyone else, to navigate more confidently through a random universe. The current era of Big Data (sometimes *really big*, as in the example of the next section), is testing traditional statistical methods to the limit, and adding some new ones that we'll visit later in the book. All the ideas, new and old, reflect the statistician's approach to making sense of our random universe.

1.1 Twenty-First-Century Science: An Example

Watching the construction of a new data science building across from my office impressed on me the nonchalance of girder walkers and how unsuited I'd be for that line of work. Statisticians aren't in danger of falling, but some nonchalance is called for when faced with a data set like the genomic study described next. It's so darned big! A trillion numbers, enough to fill a million thick volumes if ever printed out. The technology that produced twenty-first-century computers has also given us the twenty-first-century experimental equipment necessary for genomic research.

Figure 1.1 is what is called a *Manhattan plot* because of its supposed † skyline appearance. It illustrates the results of a study concerning the genetic basis of high LDL levels (the "bad cholesterol" that is thought to cause heart attacks). There are 2.4 million dots on the diagram, not all of which are visible, corresponding to 2.4 million positions on the human genome. That sounds like a lot, but it's a tiny fraction of the more than three billion "base pairs" on the total genome, each denoted by one of the four letters of the genetic code A-T-C-G.

But these are a special 2.4 million positions. Called *single nucleotide polymorphisms* or SNPs (pronounced "snips"), they are positions where there is substantial variation in which letter occurs: maybe G 90% of the time but C the other 10%. Neucleotide variation can cause dramatic differences in disease rates—perhaps the C people are more prone to an illness of interest—and is just the kind of thing the researchers were hoping to find. *Genome-wide association study* (GWAS) is the popular name for this kind of wide-ranging investigation ("fishing expedition" being a less kind term).

The study included 190,000 subjects, each of whom had all of their SNPs determined and their LDL measured. The basic idea was simple enough: look for SNPs where an unusual pair of letters (one from each parent), say, Cs where there are usually Gs, was associated with high LDL. A Nobel Prize kind of result would have uncovered one or two killer SNPs that perfectly predicted high LDL, but that's not at all

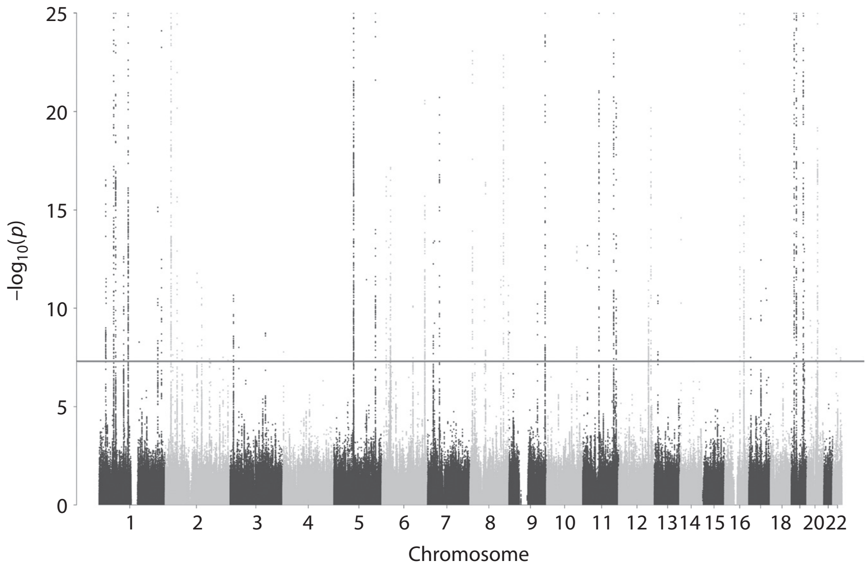

FIGURE 1.1. Manhattan plot of a study concerning the genetic basis of high LDL levels. Heights of the points measure effects on high LDL of an individual SNP. Of the 2.4 million SNPs measured, 3078 were significantly associated with high LDL, seen rising above the light horizontal line, and further concentrated in 58 *loci* to form the observed "skyscrapers." For a more in-depth discussion of figure 1.1 and the Manhattan plot, see appendix A.1.

how things turned out, as you can see from all those skyscrapers in the Manhattan plot.

This is where statistical inference comes in. The height of each dot is a measure of correlation between the letter of that SNP (A, T, C, or G) and the subject's LDL levels. High points (those above the horizontal line) indicate important SNPs, ones that increase LDL. A great deal of caution is needed here. We are searching through needles-in-haystacks land: with 2.4 million SNPs to consider, the overwhelming majority of which are not predictive, random variation alone could produce a torrent of "false positives."

The height of the horizontal line is set high enough to avoid most of those (the discussion in appendix A.1 gives some idea of how this is done); 3078 SNPs cleared the bar. These were concentrated at 58 positions, or *loci*, along the genome, shown as skyscrapers in the Manhattan plot.

There is no question that the 3078 SNPs are related to high LDL, but as with many GWAS investigations, the results were predictively underwhelming: combined with the 100 or so previously identified loci, only 12% of the observed variability of LDL levels could be explained. (The universe seems particularly noisy at the genomic level.) The GWAS can be thought of as an opening salvo in science's attack on the LDL question, helping map out the landscape for future investigation.

A GWAS is obviously "scientific," but it is not Mr. Proctor's kind of science. Transits of Venus are predicted from Newton's laws of motion, beautiful in their simplicity and precision. Medical science has made, and will continue to make, progress in understanding disease, though without any prospect of a precise Newtonian formulation. The human body is a miraculous assemblage, but compared with celestial mechanics it is a soft machine that follows its rules more statistically than punctiliously, and requires statistical analysis (that is, an analysis allowing for variability and contrary events) for its interpretation.[3]

1.2 A Small-Data Example

With 190,000 subjects and 2.4 million SNPs per subject, the LDL study makes it into anyone's definition of Big Data. Not only has data gotten big in the twenty-first century, but the technology that produces all those numbers has mushroomed, too.

Small data hasn't gone away, though. It's been a substantial part of my consulting work in the Stanford School of Medicine and is featured in many, maybe most, of the examples I use to illustrate statistical thinking in action.

Table 1.1 concerns a small-data inference problem. A European car insurance company has tabulated the number of claims per customer made in 1950: 7840 customers made zero claims each (ideal from the company's point of view), 1317 made one claim each, and so on, up to one customer, not the most careful driver, who made seven claims. Now we only have eight numbers to look at, instead of a trillion, but they still manage to conceal a surprise.

3. Celestial mechanics has its limits too. Mr. Proctor could not have accurately predicted the next 1000 transits. Henri Poincaré in the early 1900s showed that tiny disturbances would eventually propagate out of control (the "butterfly effect"). Perhaps I should say that the randomness component is much smaller for orbits than for disease occurrence, but smaller isn't zero.

TABLE 1.1. Number of claims per customer made in 1950 for a European car insurance company; 7840 customers made no claims in 1950, 1317 customers made 1 claim each, 239 made 2 each, etc.

Claims	0	1	2	3	4	5	6	7	Total
Customers	7840	1317	239	42	14	4	4	1	9461

The company wants to set premiums for 1951. It would surely help to know how many claims each customer *will* make in the coming year. One guess is that customer x will make as many claims again in 1951 as in 1950, which would suggest charging the 7840 zero-claimers next to nothing. That can't be right. . . . Or, the company could assume that everyone would make 0.214 claims in 1951, the 1950 average, but that would treat the zero-claimers no better than the seven-claimer.

Amazingly, a much better answer was discovered in 1956. *Robbins' Rule* says that the expected number of claims in 1951 for a 1950 zero-claimer is the number of 1950 one-claimers divided by the number of zero-claimers, or

$$1317/7840 = 0.168.$$

That's only about one-sixth of a claim per person, not much but definitely more than zero. In the same vein, the expected number of 1951 claims for a 1950 one-claimer is

$$2 \times 239/1317 = 0.363,$$

or a little more than twice the 0.168 of the zero-claimers, with similar formulas for 1950 two-claimers, three-claimers, etc., as explained in appendix A.1.

Table 1.2 shows Robbins' Rule for 0, 1, 2, 3, and 4 claims in 1950. (The number of 1950 claims greater than four in table 1.1 gets too small for the Rule to be accurate.) Following Robbins' Rule lets the company set their 1951 rates according to the number of expected claims—less for 1950 zero-claimers than others, but not nothing.

At this point, the zero-claimers might launch a class action lawsuit on the grounds that they are being forced to pay for the carelessness of the one-claimers. This gets us to a central precept of statistical thinking: that we can learn from the experience of others. Doing so requires us to subsume our individual experience ("I had a perfect driving record last year!") in the pooled information from a relevant group. How to

TABLE 1.2. Robbins' Rule for the expected number of 1951 claims given the customer's number of 1950 claims.

1950 claims	0	1	2	3	4
Robbins' Rule	0.168	0.363	0.527	1.33	1.43

do so in a rational and effective way is what it means to think like a statistician.

The LDL GWAS involved many billion times more data than the insurance example, but its analysis also depended on learning from the experience of others: the other 2.4 million SNPs. To be deemed "predictive," a SNP needed to clear the horizontal bar in figure 1.1, and the height of the bar was determined by the results from *all* the SNPs, which is where the learning came in. Both the LDL and insurance examples depend on Bayesian thinking (chapter 3), one of the linchpins of statistical inference. The 18 baseball players seen in section 1.6 will make the same point again, in more dramatic fashion.

1.3 Newton's Perfect World, and Ours

Isaac Newton (1642–1727) was a two-fold genius: as what we would now call a scientist (but then a "natural philosopher") and as a mathematician of the first rank. In combination, these talents produced a view of the world that has dominated science ever since. Newton believed that

1. There are inflexible laws that underlie the workings of nature.
2. These laws can be expressed in mathematical terms.

What I called "steel-trap science" is in fact Newtonian science, a symbiosis of mathematical logic and careful observation.

The success of Newton's laws had a bracing effect on humanity's intellectual confidence, as if a strong set of spectacles had snapped the world into sharp focus. Consider his Second Law of Motion:

$$\text{acceleration} = \text{force}/\text{mass}.$$

Our day-to-day world is filled with first things pushing on second things, and if you push harder they usually move faster. The second

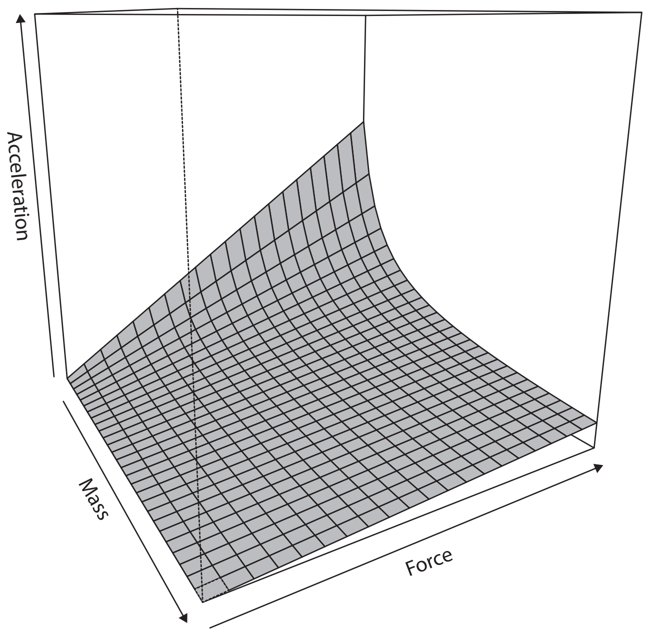

FIGURE 1.2. In Newton's perfect world, acceleration (plotted vertically) exactly equals force divided by mass.

law says precisely how this happens, depending on the force of the first thing and the mass of the second, and not on their colors or temperatures or heights above sea level or anything else. Its more usual statement, "$F = MA$" (force equals mass times acceleration), was the world's most famous equation before Einstein's "$E = mc^2$."

Figure 1.2 shows Newton's Second Law in graphical form. This is a 3D graph: you're meant to imagine the two "predictor variables" of force and mass, located on the surface of your desk, with the "response variable" of acceleration plotted vertically above them. Low mass and high force, at the far corner of the desk, produces the greatest acceleration. In Newton's hands, and that of successors like Laplace, the laws of motion enabled humans to land softly on the moon, without missing it or being smashed to bits.

How did Newton discover his laws? First of all, he was a prodigious observer (apples falling from trees, etc.). Second, he had a formidable imagination and could theorize general truths that had eluded millennia of earlier savants. Of course, he might also have done

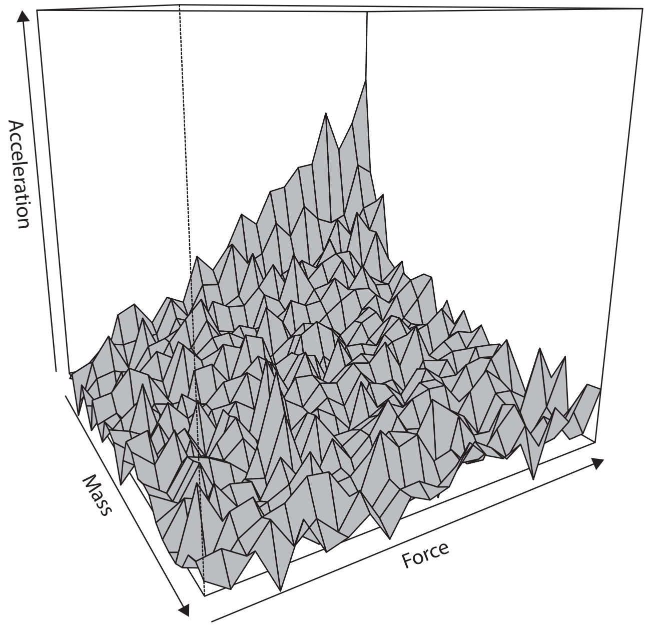

FIGURE 1.3. What Newton might have seen in the real world.

experiments, trying different forces on different masses and measuring the resulting accelerations. Perhaps not though: acceleration requires careful measurement of time, just what *wasn't* available in the mid-1600s.[4]

Figure 1.2 portrays a perfect world, one of absolute precision. This was the world Richard Proctor was working in, and much of contemporary science still thrives there. It isn't, however, the world of our book. Figure 1.3 pictures, rather fancifully, what Newton might have seen if he'd actually tried to learn the Second Law by experiment. Now we are in an imperfect world. Noise has corrupted a smooth underlying reality into what looks like a dangerous version of Disney's Space Mountain.

Newton was able to look at figure 1.3 and visualize figure 1.2. Physicists are still trying to uncover a perfect underlying reality through

4. Galileo, a half century earlier, did attempt such measurements, using marbles rolling down inclined planes, with his own pulse and a water clock as timing devices. Evidently, the effort was unfruitful.

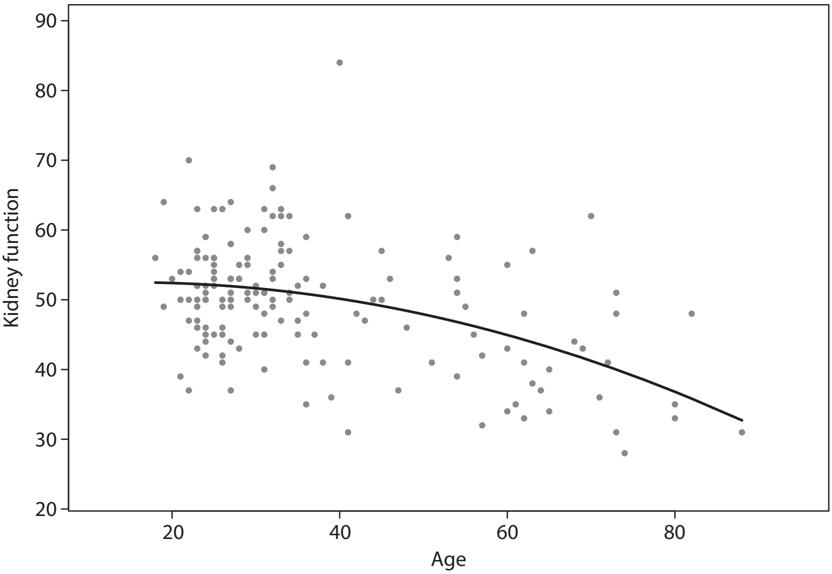

FIGURE 1.4. Kidney function of 159 healthy volunteers plotted against their age; a smooth curve has been fitted to the points. From the laboratory of Dr. Brian Myers.

keen observation and mathematical logic, string theory being a particularly heroic example, but there aren't that many Newtons to go around and reality itself may not be all that perfect in a GWAS kind of environment.

Most often, statisticians are called on to help scientists during the imperfect-world portion of a study. Figure 1.4 reports on a study of 159 healthy volunteers enrolled in a nephrology (kidney) investigation. There are 159 points, each showing a volunteer's kidney function (plotted vertically) and age (plotted horizontally). Learning how kidney function decreases with age was the goal of the study, but the term † *scatterplot* seems all too appropriate here. In the absence of Newton, I've fit a smooth curve to the points. Does this capture the hidden reality of kidney function versus age? Certainly not exactly, but the curve is evocative of a slow early decline getting steeper past age 50. Younger donors are preferred for kidney transplants, as the figure suggests, but they can be in short supply. Chapter 7 says a little more about the kidney study.

The traditional statistical worldview is typically one of a smooth and orderly underlying reality hidden from sight by corrupting layers of random noise. The job of the statistician is to strip away the noise from the picture. How this is done, and to what degree it's possible, is what I try to describe in this book.

1.4 Statistical Inference

Inference is one of those miraculous human abilities that we use, and have always used, frequently and automatically:

> "There's something rustling in the grass: maybe it's a saber-toothed tiger!"

or, more recently,

> "That car is a block away: I can cross the street safely."

Statistical inference is what statisticians do for a living. A powerful way of thinking about evidence, it depends upon hard-won theories of learning from experience, especially experience that arrives a little bit at a time in small, noisy, and sometimes contradictory pieces (as in the kidney study). We've already made a few statistical inferences, in the LDL GWAS, the insurance company's prediction problem, and the kidney study of figure 1.4.

Inferential theories—there are two main ones we'll be meeting, Bayesian and frequentist—are amalgams of mathematics, philosophy, and applied science. Here we avoid most of the math, relying on examples, charts, graphs, and figures to get across the big ideas, along with some aspirational hand-waving. The theories have been, by necessity, adapted to the computational limits of the day. In our day, computation has sped up by factors of millions. Statisticians now stride in seven-league boots, using massive numerical resources to execute statistical inference on Brobdingnagian scales (think GWAS or "Deep Learning"), but the basic inferential ideas have mostly stayed the same.

Deduction is the process of drawing specific conclusions from given axioms. *Induction* is the opposite process: arguing from observed cases back to an underlying mechanism. For instance, one inference from the 159 cases in figure 1.4 was quicker decline of kidney function after age 50. Philosophers have disagreed on whether induction is actually possible—

> "That car looks blue but I can only see one side; maybe the other side is red?"

— but life would be hopelessly muddled without it.

Statistical inference is the weaponized version of induction, putting it in action in the cause of science, medicine, society, and politics. There are obvious pitfalls lurking here. Induction can be a dangerous game compared to slow-but-safe deduction. A large part of statistical theory is defensive, aimed at avoiding inferential errors, as I'll discuss later.

1.5 Are Those Genuine Basquiats?

Not all inference problems are numerical in nature. Here is the beginning of an article from the *New York Times* of May 30, 2022:

> **FBI Investigating Basquiat Painting Show at Orlando Museum of Art**—The ongoing fascination with the life and work of Jean-Michel Basquiat shows little sign of dimming whether in the form of brisk sales for $29.99 Basquiat-theme t-shirts at the Gap, large crowds for Basquiat's latest art exhibitions, or an actual canvas by the painter auctioned last week for $85 million.

Jean-Michel Basquiat was an African American painter, a friend and protégé of Andy Warhol, who died at age 28 in 1988 during the height of international praise and fame, which has recently intensified. The same person who paid that $85 million, Yusaku Maezawa, then bought one of Basquiat's striking "Black Skull" paintings for $110 million, a record for work by an American artist, which is to say that there is big money involved here. The article continues:

> The FBI's Art Crime Team is investigating the authenticity of 25 paintings that the Orlando Museum of Art says were created by Basquiat and are on exhibit there. . . . The paintings in the "Heroes and Monsters: Jean-Michel Basquiat" exhibition were said to have been recovered from a Los Angeles storage unit in 2012. . . . The paintings' owners and the museum's chief executive Aaron De Groft say the paintings are genuine Basquiats, citing statements from art world experts commissioned by the owners. . . . The paintings are set to leave the museum on June 30 for public exhibitions in Italy.

The FBI's interest reflects allegations of tax evasion and perhaps money laundering in the high-end art world and, in this case, the tangled provenance of the 25 paintings. Worked on cardboard, they were discovered in 1982 and said to have been sold by the artist for $5000. The trove was

> . . . bought for about $15,000 by William Force, an art and antiques dealer, and Lee Mangin, a retired salesman. A third owner is the Los Angeles trial lawyer Pierce O'Donnell, who in 2016 represented Amber Heard in her divorce from Johnny Depp and Angelina Jolie in her divorce from Brad Pitt.

This is where the *Times* story gets more pointed:

> Much of the backstory establishing the paintings' origins rests largely on the word of Mangin and Force, who have both served time in prison for felony drug trafficking under different names. . . . In 1996 the Securities and Exchange Commission arrested Mangin for securities fraud, alleging that Mangin was part of a criminal ring that forged documents. . . . O'Donnell also has a criminal record. . . . Richard LiPuma, a lawyer for Mangin, said that the provenance of the paintings was "airtight," and the fact that the owners were once in trouble with the law was irrelevant to the question of whether the works are genuine. . . .

Not so fast, LiPuma. "Innocent until proven guilty" doesn't apply to the world of inference. A thoughtful oil sheikh, in considering whether to lay out $100 million for a Basquiat, would probably prefer non-criminal records for the sellers. On the other hand, the sheikh might give more credence to the owners' art experts, whom the rather one-sided article ignores: they and the museum must have had some reason for putting their reputations on the line in an obviously contentious situation.

Other evidence, more con than pro, is reported in this article, such as from a handwriting expert (Basquiat wrote a few words on his paintings) who may or may not have verified authenticity, and an "independent brand expert consulted by the *Times*," who said that the Federal Express imprint on the cardboard backs wasn't used by FedEx until 1994, six years after the artist's death.

At this point, the sheikh might wish to hire an "independent Bayesian expert." Bayes' Rule, historically the first word in statistical inference, provides a rational way to update one's opinions as new evidence arrives. "New evidence" often takes the form of numerical data, but it can also be vague and qualitative, as with the maybe-Basquiat paintings.

Chapter 3 applies the Rule to an important inferential question that arose in 2020: Was the drug hydroxychloroquine effective against COVID-19? The venue there was the interface between politics and medicine rather than art and money, but it was even more contentious. After the hydroxychloroquine discussion, I'll encourage you (with some hints) to try a Bayesian analysis of the Basquiat authenticity question.

Gathering information from different sources is a trademark of statistical thinking, Bayes' Rule being just one of the methods we'll discuss. The problem of the 18 baseball players in the next section shows how information can accrue in surprising ways.[5]

1.6 The 18 Baseball Players: Learning from the Experience of Others

Asked to free-associate on the word *statistician* (and told to avoid *nerd* and its cognates), one is likely to get "data" and "numbers" first out of the box. The Basquiat story shows that statistical evidence doesn't have to be numerical, but most often it is. Numbers are to statisticians as words are to poets: the raw material from which gold can be mined. What follows is an example of the mining process applied to the world of sports. The setting is baseball, but as will become clear, the reader who knows nothing about baseball will have an advantage. Prediction, not baseball, is the real subject here.

Early in the 1970 baseball season, Carl Morris[6] was looking through the sports pages of the *Los Angeles Times*. It listed the batting averages, the proportion of successful attempts to hit the ball, for all Major League players, several hundred of them. Wanting to deal with a smaller data set, Morris focused his attention on those players with

5. Here's a more personal inference story: I was walking downtown for lunch one day when I was approached by a young policeman who said, "Are you from the Veterans' Home?" "No," I said, "I teach statistics. Why do you ask?" "Someone has wandered away." He was gone before I realized the need for a snappy comeback.

6. Then at the RAND Corporation, later as Professor of Statistics at Harvard.

TABLE 1.3. Early batting averages for 18 Major League baseball players after 45 at-bats in the 1970 season. Two predictions of their final averages for the season are shown: the early average and the "James–Stein" prediction.

	Hits	Early average	James–Stein prediction	Final average
Clemente	18	.400	.294	?
F. Robinson	17	.378	.289	?
F. Howard	16	.356	.284	?
Johnstone	15	.333	.279	?
Berry	14	.311	.275	?
Spencer	14	.311	.275	?
Kessinger	13	.289	.270	?
L. Alvarado	12	.267	.265	?
Santo	11	.244	.261	?
Swoboda	11	.244	.261	?
Unser	10	.222	.256	?
Williams	10	.222	.256	?
Scott	10	.222	.256	?
Petrocelli	10	.222	.256	?
E. Rodriguez	10	.222	.256	?
Campaneris	9	.200	.251	?
Munson	8	.178	.246	?
Alvis	7	.156	.242	?

exactly 45 at-bats (that is, 45 attempts to hit the ball, regardless of success). There were 18 such players, whom Morris listed in decreasing order of their successful hits, as shown in table 1.3. At the top of the list was Roberto Clemente with 18 hits out of his 45 at-bats, for a batting average of

$$18/45 = 0.400.$$

(Morris chose 45 in order to include Clemente, a favorite of his.)

Morris then asked a classic inferential question: How well could he predict each player's final batting average for the remainder of the season (about 400 more at-bats)? The players bat independently of each other, so Munson's miserable performance could have no effect up or down on Clemente, and likewise for any of them. With this in mind, the best predictor for future performance seems certain to be past performance: 0.400 for Clemente's season, 0.178 for Munson's, etc.

Among other things, Morris was a leading sports statistician, but it wasn't just baseball that led him to table 1.3. The recently invented *James–Stein Rule* offered a different if controversial prediction formula that he wanted to test out. The third column of table 1.3 gives the James–Stein predictions. These are based on the same early averages in the second column but with a twist: for the James–Stein predictions, each player's early average is shrunk a certain proportion of the way to the average of the early averages (equaling 0.265 in this case). The clever algorithm that says what "certain" means is discussed in appendix A.1, but its exact expression isn't needed to see its effect here.

Time went by. In October, after the end of the baseball season, Morris was able to fill in the question marks in the last column of table 1.3. Column 3 of table 1.4 shows the season's final averages for the 18 players. We have two possible predictors for the final average: column 1 shows the early averages while column 2 shows the James–Stein predictions. Which did the best job of predicting the final averages?

Table 1.5 makes the comparison. Column 1 lists the prediction errors made by the early averages. For example, Clemente's early average was .400 while his final average was .346, for a prediction error of .054 (i.e., 0.400 – 0.346). His James–Stein prediction of .294 underestimated his final average of .346, for a prediction error of −0.052 (i.e., 0.294 – 0.346). For both methods, some of the prediction errors are positive, some are negative, but you can see that the James–Stein errors are usually smaller in size.

At the bottom of table 1.5, I've added up the squares of the prediction errors for each method (squares in order to count positive and negative errors equally). The sum of squared errors provides an overall assessment of prediction inaccuracy. It is more than three times worse,

$$0.075/0.021 = 3.6,$$

for the early averages than for the James–Stein predictors. This shows a stunning difference in predictive performance! Nobody cares much about 50-year-old baseball statistics, but if we were predicting cancer cure rates at 18 hospitals, a factor of three-times-less-error might be life-saving. This wasn't a fluke. James and Stein used inferential theory to prove that their method will always be expected to give better results.

TABLE 1.4. Now showing the players' final batting averages for the 1970 season.

	Early average	James–Stein prediction	Final average
Clemente	.400	.294	.346
F. Robinson	.378	.289	.298
F. Howard	.356	.284	.276
Johnstone	.333	.279	.222
Berry	.311	.275	.273
Spencer	.311	.275	.270
Kessinger	.289	.270	.263
L. Alvarado	.267	.265	.210
Santo	.244	.261	.269
Swoboda	.244	.261	.230
Unser	.222	.256	.264
Williams	.222	.256	.256
Scott	.222	.256	.303
Petrocelli	.222	.256	.264
E. Rodriguez	.222	.256	.226
Campaneris	.200	.251	.285
Munson	.178	.246	.316
Alvis	.156	.242	.200

What is going on here? We seem to have a paradox: the 18 players bat independently of each other and classical statistical theory shows that, individually, each player's batting average is the best predictor of his future performance. How can the James–Stein method do better?

The answer is simple but surprising: there is extra information available from the full set of 18 results. In this case, it is information about baseball. You don't have to be a baseball fan to learn from the data in table 1.3. The 18 batting averages fall between 0.156 and 0.400; there aren't any 0.000s or 0.999s, and in fact most of the batting averages are not far from the "average" average (that is, the average of the 18 averages) of 0.265. As explained in appendix A.1 (in particular figure A.2), the James–Stein method takes advantage of these facts to pull each estimate closer to the middle value of 0.265. But for our discussion here, the details aren't as important as the bigger picture of statistical

TABLE 1.5. Comparison of the prediction errors of the two methods.

	Early avg minus final avg	JS prediction minus final avg
Clemente	.054	−.052
F. Robinson	.080	−.009
F. Howard	.080	.008
Johnstone	.111	.057
Berry	.038	.002
Spencer	.041	.005
Kessinger	.026	.007
L. Alvarado	.057	.055
Santo	−.025	−.008
Swoboda	.014	.031
Unser	−.042	−.008
Williams	−.034	.000
Scott	−.081	−.047
Petrocelli	−.042	−.008
E. Rodriguez	−.004	.030
Campaneris	−.085	−.034
Munson	−.138	−.070
Alvis	−.044	.042
Sum of squared errors	.075	.021

methods scooping up information from both likely and unexpected sources.

My definition of statistics as the science of learning from experience applies doubly to the 18 baseball players. At a first level, each player's early average was a statistical summary of his first 45 at-bats. Clemente, for example, might have had this early season history:

1 1 0 0 1 1 1 0 0 0 0 0 0 0 0
1 0 1 0 0 1 0 0 0 1 0 1 1 0 1
0 0 0 1 0 0 1 0 1 0 1 1 0 1 0

with 1 indicating a hit and 0 an out. His early-season average of .400 was the average, or mean, of the 1s and 0s.[7]

Clemente's early season .400 average was based entirely on his own experience, and likewise for each of the other 17 players. The James–Stein predictions were based on a second level of statistical learning where the data was the 18 early season averages. In other words, each player was learning from the experience of others.

In my experience helping medical school researchers analyze their data, there is often a wide choice of possible "others", such as previous studies, related diseases, different age groups, etc. Statistical gems like the James–Stein estimator help point the way for the data analyst, but there is still a lot of art mixed in with the science.

If the James–Stein phenomenon seems bizarre to you, you're in good company. The appearance of their jointly authored paper in 1961 flummoxed much of the statistics community. In the James–Stein world, Munson's bad performance *does* pull down Clemente's prediction (by lowering the average of averages)! How can this be if they are performing independently? The answer has much to do with *Bayesian inference*, which we'll begin investigating in chapter 3, "Reverend Bayes and the Physicist's Twins."

Bayesian methods bring past knowledge to bear on current problems. A baseball fan would naturally use Bayesian thinking in answering Morris's prediction question. They would know that Clemente's early 0.400 was almost certainly too high for a whole season—there hasn't been a season-long .400 batter in 75 years—and that Munson's 0.178 was too low (he was one of the best hitters of his time). Sports statistics are a friendly venue for Bayesian inference, making sports fans into natural Bayesians.

Not all statistical problems have useful Bayesian knowledge at hand. My hypothetical "cancer cure rates at 18 hospitals" example very well might not, and even if it did, that knowledge might be controversial or undependable. Frequentist inference, the kind we've been tacitly following here, aims to exclude prior knowledge from its methods. We'll

7. It can be shown that the mean is the *best* single-number summary of the full history, but in fact scientific acceptance of the mean is comparatively recent. Given a set of varying observations, eighteenth-century astronomers argued over using the mean versus using only the one observation considered most reliable. Stigler gives an insightful discussion of the debate in the first chapter of *The Seven Pillars of Statistical Wisdom*.

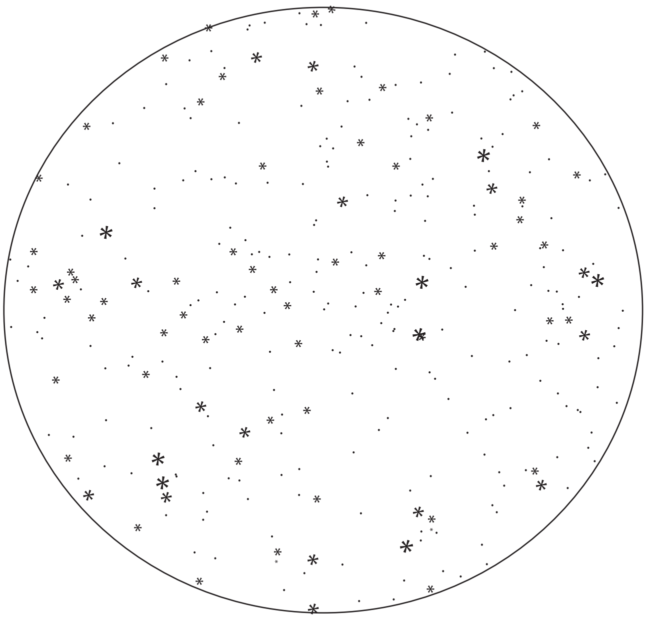

FIGURE 1.5. The northern view in a world where stellar locations and brightnesses have been selected at random. Each star has equal probability of being at any point in the circle.

be going back and forth between frequentist and Bayesian thinking in the chapters ahead.

One definition of statistics is *the science of learning from experience*, especially experience that arrives in small, sometimes contradictory pieces. In this sense, inferential theory concerns the efficient extraction of information. There were literally billions of pieces of data that went into the GWAS investigation of figure 1.1, a library's worth of hefty volumes, all of which were boiled down to the discovery of a few new groups of significant SNPs.

1.7 The Statistical Skeptic

My "peering through the fog" metaphor for statistical inference has two faces: seeing what's hidden beneath the fog and *not* seeing what's not there. We're happy if the saber-toothed tiger in the grass turns out to be a false alarm, but it's a different matter when a new cancer drug is mistakenly claimed to be effective. Some of the key ideas we'll be talking

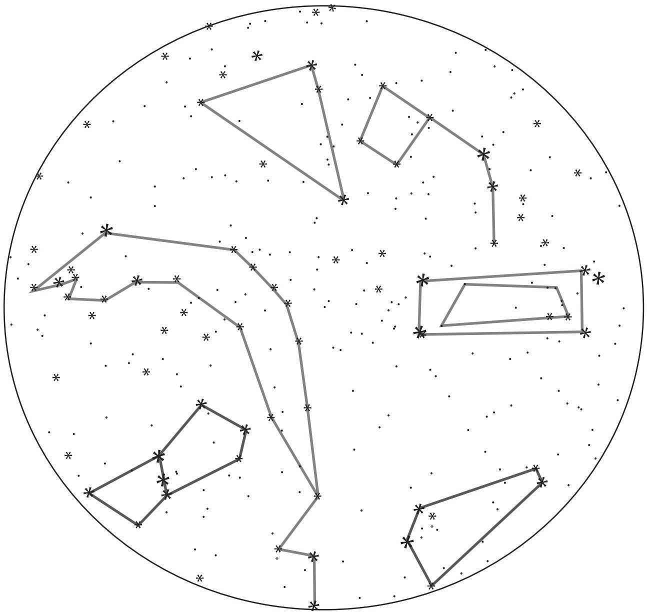

FIGURE 1.6. My constellations in the sky of figure 1.5. Clockwise from top: the Right Triangle, the Dipper, the iPhone, the Coffin, the Serpent, O'Brian and O'Brian's Belt.

about are defensive in nature, designed to protect overenthusiastic scientists from seeing only what they want to see.

Here's an example of seeing what one wants to see. Figure 1.5 pictures the northern sky in a world where the stars have been placed completely at random. I spent half an hour drawing constellations in that sky with the somewhat childish results shown in figure 1.6. Someone with a keener artistic eye could do better, but that's not the point. Our actual constellations tell thrilling stories, and people love stories, but stories get less desirable when we're trying to really understand an intricate system. A newspaper headline like "President Blamed for Inflation" means someone has done some constellation-drawing concerning national politics.

Chapter 4, "The Gold Standard of Comparisons," talks about the statistical guardrails that keep medical science on track: randomization, blinding, hypothesis testing, the famous 0.05 rejection level, statistical significance. . . . When the Hippocratic oath is quoted in the news, it's usually for its directive "do no harm." That's excellent defensive advice even 2500 years later. Human health and sickness are complicated topics, maybe the *most* complicated that science confronts.

That doesn't stop, and in fact seems to encourage, a high volume of health advice in the news: for instance, from the *New York Times* science section:

"An Hour of Running Lengthens Life Expectancy by 7 Hours"

(seeming to offer the prospect of eternal life, but the authors put in a time limit). Chapter 5 presents a statistician's guide to reading the medical news in terms of two competing viewpoints on personal well-being: the Health Mountain and the Health Plateau. It concludes with a health news scorecard you can use to rate the plausibility of stories like the one under the headline quoted above.

(continued...)

INDEX