

Introduction

What's new in the fourth edition?

For the fourth edition, new expository material was added at crucial places. For instance, the first section of chapter 26 was completely rewritten. Some other famous difficulties have been handled a little better too, and there are new problems on topics of current interest. But the principal change is to the data. Statistics, like people, show wear and tear from aging. Fortunately or unfortunately, data are easier to rejuvenate.

Why did we write this book?

The world is full of elementary statistics books. Why did we write another one? The answer is that we came to want a book which would explain the basic ideas in the subject to an intelligent but nonmathematical reader, and make the ideas vivid through real examples. These objectives seem innocent enough; achieving them turned out to be much harder than we had expected. We proceeded largely by trial and error, going through many cycles of classroom testing and revision before first publication in 1978. Each successive draft was used for a year with many hundreds of Berkeley undergraduates, in courses at different levels of difficulty (with or without a calculus prerequisite), the class sizes ranging from 30 to 300.¹

Each year, we watched the students working on the materials, listened carefully as the friendlier ones told us what was wrong with the exposition, and scribbled frantically away at the next year's draft.

Along the way, we were forced to notice some unpleasant facts. The first shock was discovering how much trouble the students had with arithmetic. In self-defense, we started giving pre-tests. By now, such tests have been given to several thousand

¹ The book was mainly developed in the Statistics 2 course at Berkeley. This course, which is divided into two or three large lecture sections, enrolls about 500–1000 students each semester, drawn mainly from the social sciences and the less-quantitative natural sciences. Still, about 40% of these students have taken a calculus course, and 20% of them have completed two or more additional college-level mathematics courses. The book is also used in Statistics 20, 21, and 131. Statistics 20 has class sizes of 30–150; virtually all students have had calculus, and about half are in quantitative fields like mathematics, statistics, computer science, and engineering. Statistics 21 is a large lecture course for business students, with about 300 students. Statistics 131 is an upper-division course for students in the social and life sciences, with class sizes in the range 30–60.

students. Here are four questions from the pre-test.¹

1. 300 is what percent of 2,000?
2. $\sqrt{100,000}$ is about:
(i) 30 (ii) 300 (iii) 1,000 (iv) 3,000 (v) can't tell
3. In the United States, 1 person out of every 500 is in the navy and one-sixth of naval personnel are officers. What fraction of the United States population consists of naval officers? Or can this be determined from the information given?
4. A quart of vodka is 40% alcohol. Write a formula for the percentage of alcohol in a mixture of V quarts of vodka and J quarts of orange juice.

Only three students in four can do the percentage in question 1, and only two in three can handle the square root in question 2. Question 3 tests whether they know when to multiply fractions; only one student in four gets it right. Many elementary statistics texts claim their sole prerequisite to be “high school algebra.” Question 4 is a very gentle probe into what the students remember from high school algebra: one student in six can write down the formula.

The pre-test even seems to understate the problem. One issue it misses is reliability. A student may be quite good at doing one-line arithmetic problems, like

$$\sqrt{2500} = \underline{\hspace{2cm}}$$

But an exercise that requires doing half a dozen steps of similar difficulty is rather a different project. Another issue is context. When students have trouble deciding which arithmetic operations to perform in response to word problems, many stop being able to do arithmetic at all. It is as if they get exhausted during the analysis phase.

Now when we started writing, we tried to teach the conventional notation,

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

and all the rest. But it soon became clear that the algebra was getting in the way. For students with limited technical ability, mastering the notation demands so much effort that nothing is left over for the ideas. To make the point by analogy, it is as if most the undergraduates on the campus were required to take a course in Chinese history—and the history department insisted on teaching in Chinese.

So we decided to try writing in ordinary English. For three probabilists, this presented some unexpected difficulties. And it led to a surprise in the classroom: the students wanted the equations, even though they found the symbolism baffling. Perhaps we shouldn't have been surprised. Nonmathematical students seem to flounder in numbers courses. They survive only by ruthless pragmatism. Their objective is to pass the final. Usually, the final is a series of word problems, and the course is seen

¹ Pretests from 1977, 1988, and 1995 are reproduced at the end of the manual. Over the period, there have been many changes in admissions standards, but the pre-test results have stayed about the same.

as a series of equations. The instructor may think that the equations express some general truths, but this tends to be lost on the students. For them, the main issue is learning how to associate the equations in the course with the word problems on the final, and recognizing which numbers in the word problem are to be substituted for which variables in the equation. There is a Berkeley student word for this syndrome: *pluginski*.¹

By the time students get to a statistics course, *pluginski* is so ingrained that anything like an equation tends to shortcut thought: students just grab the equation and run. Without equations, students really have to work at understanding the concepts in order to solve the problems. This is exactly what we want to achieve, even though the students find it irritating. As a result, whenever possible, we banish equations.² However, in many cases we do have a substitute: short summary sentences for the major points. There is a definite advantage to this approach: it is hard to memorize an English sentence without paying some attention to what the words mean.

By now, we had been through several drafts, and thought the worst was over. It wasn't. By our lights, we had succeeded in translating quite a lot of statistics into acceptable English. But, as we discovered, the students were still having a hard time with our materials. We got discouraged enough to start grumbling to colleagues in other departments, showing them the "easy" passages the students couldn't read. The colleagues couldn't read them either. Where we saw simplicity, they saw a maze of complexity.

This was a low point, but things improved from there. We realized that the problem wasn't "dumb students"; it was more a case of nonstatisticians seeing the world very differently from statisticians, needing different kinds of explanations, and wanting to learn different kinds of skills. Very few members of our audience are actually going to derive formulas, or carry out large scale data analysis. Many, however, are going to have to deal with statistical findings, because nowadays it is hard to read research journals—or even newspapers—without coming across statistical arguments.

We began to rethink our strategy. We had been making a tacit assumption, that the exposition should start from the points which were clear and obvious—"elementary"—to us, building up to more complicated and interesting ideas. However, elementary mathematical points are often rather hard, even when expressed in English. Insisting on these points just confuses things and distracts attention from the main issues. Also, we were still focusing on the procedures, leaving it to the students to infer the purposes of the activities—the scientific questions being answered. This is fine for people who find technique easy, and therefore have time to think about what they're doing. For our readers, students and nonstatistical colleagues alike, this was a failure.

We decided to start at the other end. What are the main ideas that our field has to offer the intelligent outsider? Everything else, no matter what its technical interest, had to be set aside. Then, the reader has to be persuaded that each idea is

¹ The phrase of the new millenium is *plug'n chug*.

² Some instructors who have used the book do the equations in lecture, and tell us the students accept this as complementing the text.

worth knowing. To do that, we had to make explicit the question behind the statistical procedure. Often, we were able to find some vivid example embodying the question. Similarly, many statistical concepts formalize some understanding about the world; and in many cases we were able to find the right example to crystallize this insight. Once motivated, the ideas had to be presented in reasonably smooth language, free of annoying technicalities. And it all had to be fitted together into a coherent narrative, so that at each stage the reader would know enough to appreciate the next question.

Carrying out this program turned out to be a real adventure, because it forced us to reconsider the basics of the field from a different perspective. In the end, we think we brought it off. The book covers a good set of topics for a first course, arranged in logical order, and properly illustrated by examples. It works quite well for us, and for many friends who have tried it elsewhere.¹ Sample tests, with pass rates, are reproduced below.² As far as we can see, the book is intelligible to its intended audience: nonstatisticians who want to learn some statistics in order to go about their affairs. This includes students in college classrooms—as well as professionals in other fields.

To some statisticians, the book looks like an easy read—too easy to use as a college text. This criticism is off the mark. The material is not easy. We know from our courses that students, even those with good mathematical preparation, have to work quite hard to read the book and solve the exercises. In part, this is because there are many pedagogical difficulties we just could not overcome. Then too, statistics does involve some deep ideas. Instructors who use the book will have to help their students master those ideas.

Scheduling

At Berkeley in the 1970s, Statistics 2 was taught in ten-week quarters, with three hours of lectures a week, and three hours of laboratory. In the 1980s, the university went back to fifteen-week semesters. The book can be used successfully with both calendars. It is written so that most chapters take about an hour of lecture time. However, this is a fairly quick pace. To maintain it, the more difficult sections in some chapters have to be skipped, or carried over to a second lecture.

There are 29 chapters to the book, so something has to go to fit it into a quarter. In a semester, the whole book can be covered; there may even be some time to spare

¹ Indiana, Minnesota, Sonoma State College, Stanford, UC Los Angeles, UC Santa Barbara, Utah State, Winnipeg, Wisconsin, Yale.

² In typical Statistics 2 finals, the class averages were around 60 out of 100, with an SD of 20. Students with calculus averaged around 65, each additional college mathematics course contributing around 2 points to the average. On similar tests, Statistics 20 students, who know calculus and are majoring in quantitative fields, averaged about 70. Most of the test questions were taken from exercises in the book, so the students had seen them before. An interesting sidelight: about 70% of the Statistics 2 students take the course to fulfill a requirement; the others take it voluntarily. Those taking it as a requirement only averaged about 55; the volunteers averaged over 65.

at the end, to do some of the mathematical formalism. Dependencies among the various parts of the book have been minimized in the writing, leaving instructors fairly free to pick and choose. As far as we are concerned, the logical core of the book consists of—

Chapters 1–2	Design of experiments
Chapters 3–4–5	Descriptive statistics
Chapter 13	What are the chances?
Chapters 16–17–18	Chance variability
Chapters 19–20–21, 23	Sampling

We see chapters 1, 2, and 19 as the most important. The big point is that the design of a study determines its reliability, and likewise for samples.

Sometimes when we teach the course, we cover parts I–VII, but omit part VIII on testing. At other times, we have covered everything except part III (correlation and regression). A third strategy, which we can recommend, is to cover the whole book, omitting—

Chapter 12	The regression line
Chapter 15	The binomial coefficients
Chapter 25	Chance models in genetics
Section 26.6	The t -test
Sections 27.3–4	The z -test for experiments
Chapter 28	The chi-square test

Exercises

We discovered early on that unless we could write an exercise to test a point, students were not likely to learn it. So we worked quite hard to create a variety of good exercises. Most sections close with an exercise set, the answers being at the back of the book. All chapters but 1 and 7 include a set of “review exercises.” In many chapters, the review exercises cover previous material too. This prevents the material from disappearing, and makes the students learn to judge when the different procedures apply. Answers to the review exercises do not appear in the book, but are in this manual, below. We usually make out homework assignments from the review exercises, and put some of them on the tests as well. Generally, we assign about half the review exercises in the book as homework.

Most exercise sets include a few problems which can be solved by a straightforward application of the procedures just covered in the book. However, there usually are harder problems too. Some exercises, for instance, ask the students to choose among competing procedures, or decide whether a proposed procedure is sensible. Other exercises ask the students to make rough guesses as to the magnitudes of certain quantities, still others call for qualitative judgments. Such exercises cannot be solved by mechanical application of formulas: they require understanding. In the student vernacular, these are “concept questions.”

Many exercise sets can be used as diagnostic aids, to pinpoint the difficulties students are having with the concepts. We often get the students to do the exercises in laboratory periods, working together in small groups. We go around from group

to group, talking to them about what they are doing. The exercises provide a good framework within which to discuss the ideas we want to get across.

Supervision

One key to teaching a large lecture course is supervision of the teaching assistants who handle section meetings (or “labs,” in the Berkeley vernacular). Our experience is that TAs want to lecture. Not unnaturally, they want to teach mathematics, and are a little impatient with our nonmathematical approach. On the other hand, we think we’ve already given the lectures, and just want the TAs to help the students work problems.

To make this stick, we drop in on the labs from time to time, and observe the TAs at work, or talk to the students ourselves. More formally, we meet the TAs once a week, and review with them the problems to work in lab. This means going over the statistical content of the problems, and the pedagogical issues: what does this problem illustrate? where is it discussed in the text? what will students find hard? how can you break the problem down into smaller pieces? These sessions and the lab visits were eye-openers—for us and the teaching assistants alike.

Grading

At Berkeley, the students turn in homework; this is graded by “readers,” often undergraduate majors.¹ The readers work on a very tight time-table, and come out of a tradition where word problems have numerical answers which are right or wrong. However, we want solutions to be written out in reasonable style, with the logic explained.

Some of the answers at the back of the book are quite complete, and could serve as models for students handing in assignments or tackling exam questions. Others are sketchy. Generally, we provide complete answers for some of the questions in each section, particularly those covering new material. Similar comments apply to answers in this manual.

The focus is on the concepts. When grading, we do not penalize students for minor numerical errors. For many of our students, interpolating in a table is a lot of trouble. So we tolerate rather crude rounding. This attitude may have affected some of our numerical solutions.

We tend to write out complete solutions for assigned homework, or delegate this task to the TAs. These solutions—not the Instructor’s Manual—go to the readers. Solutions are returned with the graded homework. The readers are then better able to judge what we want from the students, and there is less chance of solution files appearing in the student community. For similar reasons, we ask you not to circulate material from this manual.

¹ When budgets have to be cut, university management tends to view readers in lower-division courses as a luxury; managers think differently from the rest of us.

How to use the book

This section of the manual has detailed comments on the different chapters in the book, outlining the contents, pointing to nonstandard language and pedagogical difficulties.

Part I. Design of Experiments

The material in part I of the book is interesting and not very technical, so we find it a natural introduction to the subject. The material looks easy, and students may get the wrong impression. Some instructors may wish to start right in with descriptive statistics (part II), and talk about design issues as they come up. The book is organized with that possibility in mind.

Chapter 1. Controlled Experiments

This chapter explains the key elements in a randomized controlled double-blind design, and why each is necessary. The context is the Salk vaccine field trial. Other examples are presented to reinforce the ideas. Conventional wisdom dictates that the investigator should control the key variables and randomize the rest. The text focuses on the randomization, which is the hard idea. Some instructors will want to pay more attention to the possibility of controlling variables by stratifying subjects before randomization.

Chapter 2. Observational Studies

In this chapter, observational studies are distinguished from controlled experiments. With an observational study, it is harder to draw conclusions about cause-and-effect relationships. The “cause” and “effect” may both be the result of some hidden third factor—a confounder. We return to confounders in chapter 9. Students often seem to interpret a “confounder” as any alternative explanation for an effect. Of course, the idea is more subtle: in order for X to confound the association between Y and Z , X has to be associated both with Y and with Z : that is the point of section 5. See exercise 8 on p. 22 or exercise 10 on pp. 26–27. (Note 9 to the chapter has more discussion.)

Notes on review exercises. Exercises 1 and 2 may seem unnecessary, but many students do not realize that you take percentages to adjust for differences in group sizes. Some such students think that with a bigger denominator, the percentage will be bigger; others, perhaps more sophisticated, think the reverse. The usual recipe for computing percent— $\frac{a}{b} \times 100\%$ —obscures the idea that a percent is a rate: 10 percent means 10 per 100. (Exercises 14–15 on p. 24 teach this idea.) We regret to say, however, that the idea will probably get overwhelmed by the repetition of $\frac{a}{b} \times 100\%$.

Many of the review exercises are tough, because students don’t see any alternative explanation to the causal one; or, they find a “confounder” that is not associated with the putative cause. We keep editing the problems to make our points more sharply. Even so, many students will have trouble; they are not used to reading at all carefully. In grading, we aren’t sympathetic to rote repetition of slogans—even

ones we believe, like “association isn’t the same as causation.” Exercise 12 covers Simpson’s paradox (section 2.4).

Notes on lecturing. Instructors have asked us how we handle this part of the book in lecture, and we have done it several ways. One is to give a straightforward presentation of the material: each chapter can be covered in one lecture, omitting a section or two if time runs out. There are enough ideas here for the students to benefit from lectures as well as reading. Another approach is to bring out the main ideas in discussion. Take the Salk vaccine field trial, for example. We present the background to the trial, as outlined in the text. Then we say:

Suppose they gave the vaccine to everybody, and the incidence of polio went down. Would that show the vaccine was effective?

The class usually figures out why not. Then we present the design which puts the consent group in treatment and the no-consent group in control and ask about that. The first objection is almost always that the treatment and control groups are different sizes. After dealing with that, the class will figure out that the two groups will differ in some more important way, although they may not be able to say exactly how; we explain that polio is a disease of hygiene (p.4 of the text). Then we present the NFIP design (grade 2 in treatment, grades 1 and 3 in control), and ask for comments on that. We talk about running a proper controlled experiment, and ask the class whether the assignment should be done by the toss of a coin, or by expert judgment. Then we go on to talk about placebos and double-blinding; these ideas are hard to elicit.

If the class is too small or too large, discussion can collapse; however, we have had good discussions with classes ranging from 20 to 200 students. The length of the discussions has never been a problem: if time runs out we just drop some sections in the chapter, assigning them for reading. On the other hand, an instructor who wants to present additional material on design will find many examples in the exercises; others are cited in the footnotes.

Part II. Descriptive Statistics

For students, descriptive statistics is much easier to understand than probability or inference, and it may be a more important topic. This part of the book is about descriptive statistics for one variable—the histogram, average, standard deviation—and their relation to the normal curve.

A first pass is made at the topic of measurement error, in chapter 6. This may seem out of place in an introductory course, but it embodies one of the great lessons of statistics: every empirical number is subject to error, whether it is generated in a physics lab, a market survey, or a census. If the number is determined again, it comes out a bit different. In fact, the variability in repeated measurements is a basis for judging the likely size of the error.

Chapter 3. The Histogram

The main object of this chapter is teaching students how to read a histogram, but we found this hard to do without also teaching them how to draw one. Drawing

a histogram—or any graph—is hard work. Students will need pencil, graph paper, and eraser (or a computer with a graphics package and an eraser tool). At first, students will have to be helped with the rudimentary mechanics, like laying out axes. We talk about “class intervals”; some instructors find “bins” and “bin widths” less intimidating.

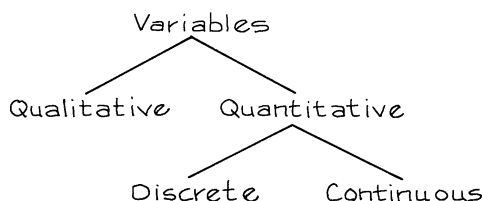
We originally tried to fudge the definition of a histogram, but kept getting caught in contradictions. Eventually, we were forced to follow the definition quite strictly, which is perhaps unusual in an elementary text. For us, percentages are represented by areas. In this setup, the height of a histogram shows crowding or *density*. The word “density” has a technical sound, and is downplayed for that reason. Moreover, the units—for instance, % per \$1000—are complicated; we couldn’t get around that. Our experience is that “% per \$1000” goes down better than “%/\$1000.”

There are two advantages to the area approach:

- There is only one kind of histogram to deal with (other books move from “frequency” to “relative frequency” to “density”).
- The histogram can be matched up against the normal curve so that area under the curve becomes intelligible.

Histograms will be used a lot in this book, so it is important to get the students used to looking at them.

Chapter 3 also introduces the idea of a *variable*, with the following classification:



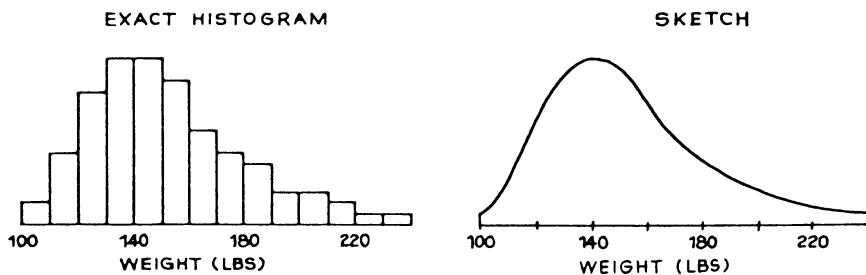
Our students didn’t seem to like this much, but then they didn’t seem to like any distinctions. (Perhaps they lack the experience needed to appreciate the usefulness of the distinctions, and don’t want to be examined on things they don’t quite grasp.)

Notes on review exercises. Exercises 1 and 4 teach the interpretation of histograms. Exercises 2 and 3 are for practice in drawing the graphs. Exercise 5 is about the density scale, and 6 shows how the histogram groups the data—and blurs distinctions within groups. Exercises 7–12 are hard. Exercise 7 makes them look at the tails of a histogram. Number 8 is about the difference between histograms and bar graphs. With number 9, students explained the spike at 2 by the fact that lots of respondents gave 2 as the GPA. This is now the answer to parts (a–b). Part (c) therefore has to have another answer. With number 10, students will prefer explanations in terms of any real factors—epidemics, immigration, whatever—to the statistical explanation in terms of digit preference. Likewise, the statistical explanation for number 12 (few very hot days) will not be obvious.

Notes on data. Some instructors use data sets of their own to illustrate the statistical techniques discussed in the book; this works out well. Some do stem-and-leaf plots on small data sets before presenting histograms, and report good results

from this approach.

Notes on graphics. Liberal use is made of smooth curves to indicate the shapes of histograms (as on p. 34), and some students will need reassurance about this. The point of sketching the histogram is usually to show some qualitative feature, such as the weight in the tails. For this, a smooth curve is just as good as the histogram, and is easier on the eye (sketch below). In general, the art work has been kept fairly informal, in the hope that working diagrams will not look too forbidding.



Chapter 4. The Average and the Standard Deviation

The chapter focuses on interpreting these two statistics. “Standard deviation” is abbreviated to “SD,” read “ess dee.” *Variance* is not introduced, for two reasons:

- Students get confused between Var and SD—“Is $SD = \sqrt{\text{Var}}$ or $\text{Var} = \sqrt{SD}$?”
- Var comes out in the wrong units, and the wrong order of magnitude.

For instance, American men average 190 pounds in weight, with an SD of 40 pounds. So the variance of weight is—1600 square pounds. To a mathematician, taking the square root is an easy fix. However, we think it is quite hard to visualize the impact of a square root (or even a linear transformation), without actually doing the arithmetic. For instance, is 17 degrees Celsius warm or cold? In a Fahrenheit world, you might reach for a calculator before answering.

So we decided to focus on the SD, deferring the concept of variance to later courses. And even before presenting the calculation of the SD, the book explains the interpretation: the SD measures how far away, on the whole, the numbers are from their average. This interpretation can be fleshed out in the usual way:

- For many lists of numbers, about 68% of the entries are within one SD of average, and 95% are within two SDs.

The book points out that this rule isn’t exact or universal. We hope it won’t be misconstrued as slavish devotion to the normal curve. In fact, it works surprisingly well for many data sets that don’t follow the normal curve at all (footnote 10 to the chapter). We often talk about the SD as the “typical” departure from average, and hope instructors will not mind the potential confusion with “probable error”—a concept not used in the book.

The root-mean-square operation is presented in section 4, as a mathematical preliminary to computing the SD. In fact, taking the r.m.s. is a basic operation in

statistics. For instance, it comes up again for the regression line (chapter 11). We used to introduce it there, but found that the students had a terrible time distinguishing between the r.m.s. error of the regression line and the SD of y . Moving the r.m.s. forward helped solve that problem—but caused a new one: some students now confuse the r.m.s. and the SD. This is easier to sort out (exercises 9 and 10 on p. 73) but the instructor should be prepared to help.

Students may ask, “Instead of doing the r.m.s., why not just drop the signs and average?” We do not have such a good answer, except to say that the r.m.s. fits in better with the theory; orthogonality is discussed in note 8 to the chapter, but that is a tough sell. Later in the course, instructors can explain that with large samples, it is the SD of the population which determines the asymptotic distribution of the sample average around the population average. Competing measures of spread, like the average absolute deviation from average, just won’t do the job (footnote 9 to chapter 18).

The technical definition of the SD, as the r.m.s. deviation from average, is presented on pp. 71–72. This reinforces the interpretation of the SD as a measure of the overall size of the deviations from average. Test results indicate that virtually all the students learn to calculate the SD correctly. But if not made to practice, they forget the algorithm within a few weeks. The book only teaches the “r.m.s. deviation” procedure for computing the SD. Another one,

$$\sqrt{x^2 - \bar{x}^2},$$

is mentioned on p. 74. We used to explain this, as well as procedures for grouped data, but only managed to confuse the students and make them learn less rather than more. They never seemed to believe that the two formulas would give the same answer, so they worried about which one to use, or combined them in unfortunate ways:

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \bar{x}^2}.$$

For us, alternative formulas represent a diversion from the main objective: teaching the students how to use the SD. After all, computers make it less important for people to learn efficient algorithms—you just have to enter the data and push a button. The trick is interpreting the output.

Notes on review exercises. Many of the exercises focus on the qualitative ideas. Exercise 3, for instance, requires students to make a rough guess as to the answer: this forces them to think, instead of rushing to the formula and plugging in. Exercise 12 is hard, because students won’t fit it into the cross-sectional vs. longitudinal framework.

Notation. When working at the blackboard, we write “ave” and “SD.” We no longer use \bar{x} , s , μ or σ in the beginning courses—too exhausting for the audience.

Which SD? The text defines the SD with n (the number of entries in the list) in the denominator, rather than $n - 1$. The $n - 1$ is introduced much later (section 26.6) as one of the modifications needed to handle small samples. We felt that in the main line of exposition, there should be only one formula for the SD. To see why we went for n , consider the average of m draws made at random with replacement

from the box $-1, +1$. When m is reasonably large, this average will be in the range $-1/\sqrt{m}$ to $+1/\sqrt{m}$ with probability about 68%. We want this interval to be of the form $\pm \sigma/\sqrt{m}$, where σ is the SD of $\{-1, +1\}$. So, the SD of $\{-1, +1\}$ has to be computed with 2 in the denominator, not $2 - 1$. In other words, when calculating the SD of a population in order to determine the asymptotic behavior of the sample average, the right denominator is n .

The conventional argument for $n - 1$ is that $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is unbiased. So it is, unless a regression is involved, in which case $n - p$ is needed. And the minute someone takes square roots to get the SD, bias comes back. We know it looks old-fashioned, but n is the right denominator for present purposes.

Chapter 5. The Normal Approximation for Data

This key chapter ties together histograms, the average, the SD, and the normal curve. The passage on pp. 80–81, which justifies the 68%–95% rule, is difficult to teach. For instance, take figure 2. The shaded area under the histogram between 60.5 inches and 66.5 inches represents the percentage of women with heights in that range, which is the interval within 1 SD of the average. By inspection, the shaded area is about equal to the area under the normal curve between -1 and 1 . This last area is 68%, justifying the rule. However, when asked, “What does the area under the histogram between 60.5 inches and 66.5 inches represent?”, many students will respond “68%.” Their anxiety to get to the numerical answer shortcuts the logic. Review exercise 1 of chapter 3 is designed to prevent this; also see review exercise 5 in the present chapter. (Review exercise 1 in chapter 4 is designed to help with the language: “the percentage of entries within 1 SD of average” isn’t exactly student English.)

Our method for teaching the normal approximation is graphical. On the blackboard, we draw diagrams just like the ones in examples 8–9 on pp. 85–87. Unless pushed, students seem to resist drawing these diagrams (or any others). Then later on in the course, with more complicated problems, they lose track of which areas they want. The diagrams help.

Section 4 takes up percentiles. It also shows that many histograms are far from the normal curve, a point which comes up again in section 6.3. The point is important, because some students take the word “normal” very literally indeed (p. 89 of the book). For this reason, we try to avoid phrases like “normal histograms,” saying instead “histograms which follow the normal curve.”

Section 5, on finding percentiles for the normal curve, will be tough going for some students. This material is used again—glancingly—in part III. However, exercises on percentiles are interspersed with later material.

Note on terminology. In this book, a histogram “follows” the normal curve if it is close to the curve.

Notes on review exercises. Exercises 10–11 are hard. To help students work exercise 11, we ask them to mark (by eye) the average on the histogram, as well as the region within one SD of the average. Then we get them to work out $\text{ave} \pm \text{SD}$, using the values given in the problem.

Chapter 6. Measurement Error

Students may confuse chance error and bias. They may also need help in seeing that the SD of a series of repeated measurements gives the likely size of the chance error in each one (pp. 100–101 of the book, and chapter 24).¹

The text has the equation

$$\text{individual measurement} = \text{exact value} + \text{bias} + \text{chance error}.$$

Some tact is needed when presenting this, because many students want to solve for the unknowns on the right, and feel cheated when they discover this to be impossible. The equation is a useful conceptual tool. Even though the unknowns cannot be precisely determined, they can often be estimated quite well.

Outliers are discussed in section 3, emphasizing the point that many histograms just do not follow the normal curve.

Notes on review exercises. The special review exercises cover most of the ideas in parts I and II. Exercise 3 may make standard units more vivid; #4 is not easy, due to the interplay between numbers and percents. Exercises 6–7 prove difficult for students who want to operate formally with the SD, instead of seeing it concretely as a measure of spread. Such students think the SD should stay the same. To help, we tell them to think about having all the men and women in a classroom, then sending the women out; what does this do to the spread in heights? Exercise 9 is a warm-up for #11 on p. 138.

With exercise 10, the HANES data are cross-sectional, so the older people in the study were born earlier, when there was more social pressure to be right-handed. This exercise, like many others in the book, may provoke students who want a self-contained mathematics course, free of background facts. Exercise 12 illustrates digit preference; also see exercise 10 in chapter 3. Exercise 13 was edited, to make it easier and to bring out the points more sharply: the elegant fact about the uniforms is now part of the exercise. Exercise 14 raises some interesting issues about the design of clinical trials. This is a hard one: students often say that the bias favors screening, “because there will be more cancers to detect and more lives saved.” In exercise 15, the tables are quoted from the source, and the numbers really do not add up.

Chapter 7. Plotting Points and Lines

The presence of this chapter may be a bit of a shock. However, many of the elementary statistics students at Berkeley have trouble with graphs. Some teachers may want to spend an hour on this chapter. Our approach is to review points and lines as we cover part III; students who need extra help can read chapter 7 by themselves.

¹ “Likely size” is just meant to convey similarity in magnitude: chance errors similar in size to the SD are common; chance errors several times larger than the SD are quite rare. The same point comes up again on pp. 16 and 20 below.

Part III. Correlation and Regression

This part of the book is about bivariate data—scatter diagrams, the correlation coefficient, and the regression line. The treatment is purely descriptive. Many teachers may wish to postpone or even skip part III. It is possible to move directly from part II to chapter 13 (probability), and then to part V (chance variability). It is also possible to do just chapters 8 and 10 from this part of the book. However, part III does follow naturally from part II, and it is easier for the students than parts IV–VIII.

In chapter 8, the correlation coefficient is presented as a key descriptive statistic needed to summarize the relationship between two variables. Then r is used to get the regression line going in chapter 10, and to determine the spread around the line in chapter 11. We used to do the regression line before introducing the correlation coefficient, but this proved too mathematical for the audience, and we had a hard time explaining r after presenting the line. For instructors who think of regression equations as invariant across data sets, with the SD of the residuals—and hence r —as situation-specific, our order of topics may seem a bit artificial (note 9 to chapter 12). If so, please bear with us.

Chapter 8. Correlation

The main job is to teach students how to read (and draw) scatter diagrams. Then, *association* is discussed carefully. Students who work exercise set A on pp. 122–24 will get comfortable with these ideas. Next, the correlation coefficient is interpreted graphically, as measuring clustering around a line. It is clearer to say that r measures clustering—rather than spread—because as r goes up to 1, clustering increases while spread decreases. Scatter diagrams are summarized by the five statistics on p. 126; the warning about outliers or nonlinear association is deferred to section 9.3. Section 8.4 gives an algorithm for computing the correlation coefficient. We do not discuss r^2 : the fraction of variance explained by a regression is at bottom a rather mysterious statistic. See D. A. Freedman, *Statistical Models* (Cambridge, 2005, §4.3).

Note on terminology. We found it helpful to introduce two nonstandard terms:

- The *point of averages* (ave. of x , ave. of y) picks out the center of the scatter diagram (p. 125).
- The *SD line* indicates the drift of the scatter diagram (section 8.3). This line goes through the point of averages, and its slope is (SD of y)/(SD of x); the sign is the same as that of r . (If r is 0, either sign can be used.)

Many nonstatisticians (and some statisticians) who fit a line to a scatter diagram by eye will approximate the SD line rather than the regression line. The contrast between the two is the regression effect (section 10.4). For us, the main point of the SD line is to help in defining the regression effect.

Notes on review exercises. The graphical interpretation of r is covered by exercises 1, 7 and 8; the computation of r , by exercise 9—although part (c) can be done qualitatively. Exercises 2 and 5 are about association. Exercises 3–4 and 11 try to get at the connection between r and linearity. Exercise 11 is not easy; to help students work it, we ask them to plot some data points. Exercise 9 on p. 106 was preparatory; so were exercises 7–8 on p. 130, but these turn out to present interesting

difficulties of their own. For instance, with #8, students want r to measure the impact of hypothetical changes in incomes, rather than the association between incomes in a fixed data set.

Chapter 9. More about Correlation

Section 1 explains that r is a pure number, invariant under scaling, symmetric in x and y . (The last point has some force, because students will interpret r as a measure of causation.) Since r is invariant under change of scale, “clustering” must be interpreted relative to the SDs. This is somewhat delicate, as indicated by figure 3 on p. 145. Section 3 explains that r may not be useful if there is a strong nonlinear association, or outliers.

Section 4 discusses the ecological fallacy—the idea that individual behavior can be inferred from group behavior. (The term “ecological” is mysterious, and is downplayed in the text.) This may be a controversial section, because many investigators in the social sciences use ecological correlations without batting an eye: see notes 3 and 4 to the chapter for some cites.

For many students, a real intellectual effort is needed to compute r . They conclude that it must be a very powerful tool. It is. But there are limits, and section 5 points some of them out.

A subliminal theme in this chapter is *attenuation*, the reduction of r due to restriction of range or measurement error. See exercise 9 on p. 144, exercises 1–2 on pp. 145–46. Ecological correlations generally exceed individual-level correlations: this is attenuation in reverse—at least, if the individual-level data are obtained from group averages by adding noise.

Notes on review exercises. Exercise 4 is on attenuation-in-reverse. Exercise 9 discusses the relationship between student evaluation of TAs and student gains in learning; the correlation is negative. Exercise 10 is a little trick. Students tend to “explain” the negative correlation between SAT scores and percentage of students taking the test by saying, “students did worse in the states where more of them took the test.” That is the answer to the first question in part (a), so a different response is needed for the rest. Exercise 11 is about ecological vs. individual correlations; exercise 12 helps students to interpret different regions in a scatter diagram.

Chapter 10. Regression

Section 1 presents a verbal equivalent of the regression equation for estimating the average of y from x . If x goes up by one SD, on the average, y does not go up by a whole SD, but only by part of an SD, namely, $r \times \text{SD of } y$. Section 2 develops a more intuitive feeling for the regression method, using the graph which displays the average of y against x . This is called the *graph of averages*. Exercise 1 on p. 163 shows the graph for incomes of husbands and wives. Section 3 takes up regression estimates for individuals, along with percentiles (which are a bit difficult). The material on percentiles can be skipped, although some of the later review exercises cannot then be assigned. Exercise 4 on p. 168 paves the way for exercise 7 on p. 567, and demonstrates that there can be some art to examining scatter diagrams.

The regression fallacy is discussed in section 4. This is the most interesting—and difficult—idea in parts II and III. When x goes up by one SD, most people want y to go up by a full SD too. The fact that it doesn't is the regression effect. The text explains that the regression effect is due to the spread of the scatter diagram around the SD line: see figure 5 on p.171 and figure 6 on p.172. People resist this statistical explanation, and want some real cause for the regression effect: that is the regression fallacy. The regression effect is implicit in section 1, but there it is kept in a very low key; we wanted the students to learn the mechanics before confronting the mystery.

Section 5 explains that there are two regression lines, one for y on x , another for x on y . There is ample room for confusion here. For example, in figure 8, the regression line of height on weight is steeper than the SD line; how come? (Answer: weight is plotted on the vertical axis.)

Notes on review exercises. Exercise 1 helps students interpret regions in the scatter diagram; also see review exercise 12 in chapter 9. Exercise 2 tries to connect regression estimates for groups and for individuals. Many students will do the same arithmetic twice—and feel puzzled; we want them to make the connection (section 10.3). Exercises 4 and 7–8 demand a real understanding of the regression effect, and are difficult. Exercises 9 and 10 are on percentiles, the latter putting another spin on the regression effect.

Note on the regression equation. The equation behind the prose treatment is

$$\frac{y - \bar{y}}{\text{SD } y} = r \frac{x - \bar{x}}{\text{SD } x}.$$

We used to teach the equation. Students would ask what r meant, as well as SD x and SD y , to say nothing of \bar{x} and \bar{y} . This was fair enough. Then they would ask what x was, at which point we got a bit discouraged. Finally, they would ask what y was. We gave the equation up as a bad job.

Note on terminology. The “graph of averages” is not a standard term, but we found it useful in discussing the regression line. In principle, this graph depends on how finely you subdivide the x 's.

Chapter 11. The R.M.S. Error for Regression

This chapter introduces residuals, as well as the formula for the r.m.s. error of the regression line: the r.m.s. error is interpreted as the amount by which a “typical” point deviates, up or down, from the regression line. (Compare pp.10–11 above, on the SD.) Students may want to know why they need both r and the r.m.s. error: one answer is that r is in relative terms—relative to the SDs—while the r.m.s. error is in the same units as y . Residual plots are taken up in section 3, although their power only becomes apparent with multiple regression.

The definition of “homoscedastic” on p.190 is a problem for some students. As far as they can see, the scatter diagram in figure 8 (p.191) shows more spread in a strip over 68 inches than in the strips over 64 or 72 inches. They are using range to measure spread. The range is bigger in the middle of the diagram, because there are

more people there. This is taken up in the text when homoscedasticity is defined; also see exercise 8 on p. 71.

For “football-shaped scatter diagrams” (bivariate normal distributions) section 5 shows how to calculate the distribution of y when x is confined to a narrow strip: of course, that is the conditional distribution of y given x . The calculation is a bit intricate. Students will have a hard time connecting the r.m.s. error and the “new SD:” the first is global, describing the whole diagram; the second is local, describing one strip. Exercises 1–3 on p. 193 are designed to make the connection. Many students will ignore the heteroscedasticity in exercise 3, and just do the arithmetic. The lesson continues with exercises 4–6 (p. 194). Exercise 4 requires the students to interpret the strip in the diagram. Exercise 5 requires estimation of averages, SDs, and r by eye. Exercise 6—which is the punchline in this series—makes you look at the local SD; part (b) is intended to ward off the obvious misinterpretation—that the SD of *any* subgroup is smaller than the SD of the whole. Exercise 7 on p. 195 is a real puzzler—the regression effect in acute form.

The focus of chapter 11 is descriptive, not inferential. The r.m.s. error measures the spread of the points around the regression line. The chapter does not consider uncertainty in the position of the regression line, which increases with distance from the point of averages; see note 5 to the chapter. Despite the relatively narrow focus, chapter 11 will take some time to teach.

Notes on review exercises. Exercise 8 is about measurement error; a common student response to (a) is “to see the regression effect.” Charitably interpreted, this isn’t so bad; the point is that the two measurements are likely to differ. Exercise 10 requires students to see that regression estimates fall on a line. Exercise 11 requires the students to look at a scatter diagram—and use what they know about U.S. schools. (Compare figure 5 on p. 39.)

Chapter 12. The Regression Line

The regression equation is presented in section 1, as an aid to computing: the exercises were set up with this in mind. The slope and intercept of the regression line are interpreted as descriptive statistics, with a warning about confounding. Section 2 discusses fitting a straight line to data in order to estimate the slope and intercept of an ideal linear relationship, and makes the point that the regression line minimizes the r.m.s. error. This material will not be easy. Section 3 restates the difficulties in drawing causal inferences from slopes. Exercise set B tests the understanding of the material in section 2; also see review exercise 8.

Review exercises 9 and 10 are hard, because students do not recognize the regression line from its description. We encourage them to sketch a scatter diagram for the income-IQ data, find the point of averages, draw the line defined by the exercise, and mark the strip corresponding to children with the given IQ. Then we ask the students to find the center of that strip. Exercise 11 makes the point that the regression line goes through the point of averages. Special review exercises 1–17 at the end of chapter 15 cover the material in chapters 1–12, and will be of interest to instructors who give midterms covering the first 12 chapters.

Part IV. Probability

As probabilists, we like the subject a lot; but students find it confusing. And whatever the advocates of the new math used to say, sets and functions make things worse for beginners. We also found that very little probability is needed to handle the statistics presented later in the book. So we went back to a more primitive approach. Chapter 13 handles the basics—independence being the most important idea—and sometimes we skip the rest of part IV. Section 14.1 on counting and chapter 15 on the binomial distribution help just a little, when setting up probability histograms in chapter 18. Students will realize that there is some depth to the material, when they hit part IV. Manifestations of “test anxiety” are to be expected.

Chapter 13. What Are the Chances?

Section 1 explains the frequency interpretation of chance. We could only afford one interpretation, and this seemed to be the smoothest. We hope that colleagues who belong to other schools of thought will not be offended. Section 2 presents conditional probabilities. Example 2(a) responds to students who have trouble thinking about the chance that the second card dealt from a deck will be the queen of hearts: “What’s the first card?” So we try to explain what an unconditional probability is, which takes a bit of work.

Section 3 does the multiplication rule. Independence—the key idea—comes in section 4. *Collins* is discussed in section 5, showing that the assumption of independence matters. This opens one of the major themes of the book. When does the theory of chance apply? What happens if the theory is used in a situation where it does not apply? The application to DNA testing is mentioned on p.234; the chapter notes give citations to the literature.

Notes on review exercises. These exercises are simple and qualitative, in order to encourage thinking about the issues. (Displays of professional cleverness are especially disastrous when teaching probability; the students just wonder how they’ll ever manage.) Exercise 2 is puzzling to some readers: we explain that it is harder to jump two hurdles than one.

Exercise 7, from Kahneman and Tversky, points to a common misconception. Exercise 9 asks for the chance of not getting 10 sixes on 10 rolls of a die. Many students will answer this by calculating the chance of getting 10 non-sixes, $(5/6)^{10}$. (From their perspective, the opposite of 10 sixes seems to be—no sixes.) To help such students sort things out, we ask them if the dice can land so as to get some sixes, but not 10 of them. We try to elicit concrete answers, e.g., 3 sixes followed by 7 aces. Part (c) is a further effort to sort out the confusion. Exercise 11 prepares for expected values and box models.

Notation. On the blackboard, we write fragments like “chance of heads” or “ch. of ace on 1st roll and ace on 2nd roll.” We try to avoid “ $P(A)$,” “ $P(\text{heads})$,” “ $A \cap B$,” “ $\text{red} \cup \text{black}$.”

Chapter 14. More about Chance

Thinking about the set of all possible ways that a chance experiment can turn out is a very useful technique, and section 1 presents it. Section 2 has the addition

rule. Example 5 is non-trivial, because many students want the chance of getting at least one ace in two rolls of a die to be $1/6 + 1/6$. The double-counting argument is a bit abstract; at this point, the sample-space representation of chances would be quite powerful, and figure 1 is a reasonable substitute.

Section 4, on the paradox of the Chevalier de Méré, is an example of how to compute probabilities using the method of complements. Students find this a bit too clever: instead of being impressed that the problem can be done at all, they are annoyed at not having a simpler way to do it.

The focus in chapters 13 and 14 is qualitative, getting across the new concepts of “independent” and “mutually exclusive” events, and trying to separate them. Students have a hard time with these two ideas. After all, both seem to express ideas of unrelatedness; there is a natural temptation to merge any two new ideas: and another temptation to think that if one doesn’t apply, the other must. Exercises are designed to ward these temptations off, with partial success; and see the “FAQs” in section 3. Moreover, basic probability really does involve fractions, and this may demoralize some students.¹ The rest of the book features decimals, which are easier.

Notes on review exercises. Exercise 3 may seem like over-kill; trust us, many students still don’t get it. Exercise 4 teaches that two chances are better than one; after all, that is why students like midterms. Exercises 5 and 6 help distinguish between “independent” and “mutually exclusive” events. The language—“all,” “not all,” “none”—is still foreign to the students; that will be the key difficulty in exercise 11. (Exercises 1–2 on p.250 may help.) Exercise 12 will get to them, because they have trouble separating conditional and unconditional probabilities. Exercises 13 and 14 are quite subtle.

Chapter 15. The Binomial Formula

This chapter explains how to calculate binomial probabilities. We skip the derivation of the coefficients; some instructors may wish to do this in class.

Notes on review exercises. Students want to scan the problem, grab the numbers, and run to a formula. You can’t do probability that way, or statistics either. Many of our problems (like number 9) are set up to defeat the student strategy. Exercise 11 brings in the sign test. The context is twin studies on the health effects of smoking. (Also see exercise 6 on pp.258–59.) The sign test is an attractive introduction to significance-testing, but there is a hitch. Students want to get the P -value by computing the probability of the observed outcome. They do not like tail probabilities, and who can blame them? We prefer to deal with this issue in a setting where the chance of any particular outcome is too small to be interesting (section 26.1).

The special review exercises cover all of parts I–IV. Exercise 3—from the *Bouman* case—is a variation on Simpson’s paradox. Exercise 6 reinforces the distinction

¹ According to the NAEP, only 68% of the seventeen-year-olds in school in the United States can add $1/2$ and $1/3$. Berkeley students can add fractions; even for them, however,

$$1/2 \text{ of } 1/3 = 1/2 \times 1/3 = 1/6$$

is rote learning—“of means times.” For proof, see the pre-test results.

between cross-sectional and longitudinal studies. (The Current Population Survey, of course, is cross-sectional.) Older people were born earlier, got less education, and their skills may have become obsolescent. Students want to “explain” the results in terms of employment status: older people work less. So, in this edition, we consider only people working full time. Exercises 7 and 9 involve percentiles; one of the difficulties in number 7 is that students will not be quite sure about the difference between “percent” and “percentile.” In exercise 10, the data were extracted from the CPS file, with no hitches. The idea is to prevent the students from saying, automatically, that every diagram in an exercise is wrong. In exercise 13, many students will want r for diagram (i) to be nearly 1, because there is a strong—nonlinear—association. (Exercise 3 on p. 148 gave some warning.) Exercise 16 (by Amos Tversky) is a cunning example of the regression fallacy. Exercises 15 and 17 cover material in chapter 11; these are hard.

Part V. Chance Variability

One famous difficulty in teaching elementary statistics is getting across the idea that the sample average is a random variable. Randomness, after all, is quite a complicated idea. It is easily overwhelmed, either by the definiteness of the data, or by the arithmetic needed to calculate the average.

In our experience, the most intelligible short explanation goes something like this:

You took a sample and computed the average. That is a number. But it could have come out a bit differently. In fact, if you did the whole thing all over again, it would come out differently.

This variability is the key point to get across, and it tends to be obscured by the technical sound of the phrase “random variable.” As a result, we have given that phrase up—and many other hallmarks of civilization too. For the phrase, at least, there is a good substitute: drawing at random from a box of tickets, where each ticket has a number written on it. This may seem crude, but conveys a clear image.

To bring variability into sharper focus, we use the idea of *chance error*. For instance, when we talk about the sample average (chapter 23 in part VI), we tell the students:

Draw some tickets at random from a box, and take the average of the numbers you get. This will be close to the average of all the numbers in the box, but it will be a little bit off. This amount off is *chance error*:

average of draws = average of box + chance error.

How big is the chance error likely to be? This question is answered by a number we call the *standard error* (abbreviated to SE, read “ess eee”). The upshot is that the average of the draws will be around the average of the box, give or take an SE or so. Technically, a “chance error” is the difference between a random variable X and its expected value $E(X)$. The “standard error” of X is $\sqrt{E\{[X - E(X)]^2\}}$. (At the risk of the obvious, the formula disappeared from the text at a very early stage, followed soon after by the random variables themselves.)

“Standard error,” of course, is not the usual term; most authors use “standard deviation” both for data and for random variables. In our experience, however, students have a lot of trouble separating the standard error for the sample average from the SD of the sample. Calling the two by the same name makes it hopeless. So in this book we are quite rigid:

- The SD is for data.
- The SE is for random variables.

Some instructors prefer the more conventional terminology; we ask their indulgence in this matter among many others.

Drawing tickets from a box, chance variability, expected values, standard errors, the normal approximation. . . . That is a lot of ideas. It takes times to get them across, and it is very hard to deal with them adequately in the middle of a complicated discussion on sampling. So we develop these ideas first, in part V, focusing on the sum of draws made at random with replacement from a box.¹ We start with the sum because chance variability is easier to recognize for sums than averages.

We handle chance variability with more care than is common in elementary books. Our pedagogical motives should be clear by now: the ideas are hard, and need time to sink in. But we also have to admit an ideological motive. We think that statistical inferences should be based on explicit chance models, for reasons given in the text; sections 21.4–5, 22.5, 23.4, 24.4, and 29.4–5.

Now students are busy people, slightly cynical, with a definite short-term goal: passing the final. Their previous mathematical education stresses arithmetic procedure, not logical deduction. It is useless to tell them, “Statistical inferences should be based on chance models.” This is empty rhetoric, with a lot of fancy words: no sensible exam question can be based on that kind of statement. We want students to take chance models seriously, so we spend course time on the topic. We also have exercises where getting the model wrong leads to the wrong answer—and losing points.

A final remark. Part V is independent of part IV. Instructors who want to spend the minimum amount of time on “pure probability” should, in our opinion, skip part IV but do part V. Part V only takes three or four hours of class time, and it is a very good investment.

Chapter 16. The Law of Averages

Students often think that with a good sample, the sample percentage will equal the population percentage. This makes it difficult for them to appreciate the standard error calculations in part VI. Part of the trouble is that they don’t understand chance variability. Section 1 of chapter 16 takes this up. We have a coin. On each toss, it is as likely to land heads as tails. Now we toss it 10,000 times. Are we likely to get exactly 5,000 heads? Surely not. As the number of tosses goes up, the difference between the number of heads and the expected number tends to get larger and larger

¹ Technically, this is our substitute for a sum of independent, identically distributed random variables. We are sacrificing some generality: our random variables only take finitely many values, with rational probabilities. That is quite enough.

in absolute terms, that is, as a number. However, the difference tends to get smaller and smaller in percentage terms, relative to the number of tosses. For many students, this distinction is new and difficult. It is central to the careful discussion of the law of averages in section 1. This section also discusses the concept of chance error, with the equation

$$\text{number of heads} = \text{half the number of tosses} + \text{chance error.}$$

The *likely size* of the chance error is used informally in the text. (The technical equivalent is the standard error.)

The balance of the chapter is spent setting up box models and introducing the sum of the draws from the box. A box model consists of draws made at random from a box of tickets; each ticket in the box shows a number. The chance variability in coins, dice, roulette wheels (and later, sampling processes) is related to the chance variability in draws from a box. Eventually, this produces real economy of thought: there is a general theory, instead of a lot of special cases. At first, students find this approach rather strange, but they quickly get used to it.

Many examples in this chapter are based on gambling at roulette: the sum of the draws from the box corresponds to the net gain. For instance, take example 1 on p.283. The net gain in 100 plays at roulette, staking \$1 on a single number at each play, is like the sum of 100 draws from the box:

1 ticket	\$35	37 tickets	−\$1
----------	------	------------	------

The phrase “is like” has a precise technical meaning: the net gain and the sum have the same probability distribution. Of course, we do not insist on this in the text, but make the point through problems like exercise 6 on p.281 or exercise 2 on pp.284–85.

Students find the gambling interesting, although a bit technical. (One touchy point is adding up negative numbers.) It is a digression from the mainline statistical issues. However, setting up a proper model for a mainline statistics problem is hard. Setting up a model for roulette is much easier, and it’s good practice. As we tell the students, the first step is to write the box down. (Of course, you can quickly generate a lot of free-floating boxes; nobody said this was an easy subject to teach.)

Notes on review exercises. Exercise 1 tests the distinction between absolute and relative errors, and will be easier for the students when translated into a problem about coin-tossing. Exercises 4, 6, and 9 are variations on the law of large numbers; the last may have some technical interest (note 6 to the chapter). Exercises 7 and 8 are about box models, and #10 foreshadows chapter 23.

Chapter 17. The Expected Value and Standard Error

This chapter presents the formulas for the expected value and standard error for the sum of draws made at random with replacement from a box. The first idea is that the sum of the draws from a box will be around its expected value, but will be off by a chance error:

$$\text{sum} = \text{expected value} + \text{chance error.}$$

The likely size of the chance error is given by the SE for the sum. As we write over and over again on the blackboard,

The sum of the draws will be around _____ give or take _____ or so.

There is a downside: some students will later view expected values as random variables, computed up to some margin of error. Among other things, after the sample has been drawn, students will want the expected value for the *parameter* to equal *the estimate*. Given our (slavish?) devotion to the frequency theory, we developed many ward-off exercises. See, for instance, exercise number 6 on p.294, number 8 on p.328, number 1 on p.366. . . .

We tell the students that chance errors of an SE or so in size are fairly common, but chance errors bigger than several SEs in size are very unusual. The SE for the sum of draws made at random with replacement from a box is computed by the square root law (p.291) as

$$\sqrt{\text{number of draws}} \times \text{SD of box}.$$

Students need help seeing what the square root means: when the number of draws goes up by a factor of 100, say, the SE for the sum of the draws only goes up by the factor $\sqrt{100} = 10$. In particular, as the number of draws goes up, the SE for the sum goes up in absolute terms, but goes down relative to the number of draws. When the number of draws is large, the normal approximation can be used (section 3), although a full discussion is postponed to chapter 18. Exercise 8 (p.297) reinforces the law of averages, and may have some appeal on its own: see note 6 to the chapter.

As a matter of style, it is wise (though cumbersome) to write “SE for sum,” not just “SE.” (We try to make the students do this, although we often sin by omission.) Later on, we will have both the SE for sums and the SE for averages. Students will want to merge those two entities. Insisting on full names helps prevent this.

Many boxes in gambling problems (roulette, for instance) have only two kinds of tickets, and there is a short cut formula for the SD of the box. More technically, if $P\{X = a\} = p$ and $P\{X = b\} = 1 - p$, the SE is

$$|a - b|\sqrt{p(1 - p)}.$$

This formula appears (in words) on p.298.

We attempt to treat standard errors in a unified way, tracing everything back to sums. In section 5, a coin lands heads with probability p and is tossed n times: what is the standard error for the number of heads? This problem fits into the general framework of sums by the 0–1 coding trick, counting heads as 1 and tails as 0. The number of heads is like the sum of n draws made at random with replacement from a box where the fraction of tickets marked 1 is p , and the fraction marked 0 is $1 - p$. The SE for the sum is, of course, $\sqrt{n} \times$ the SD of the box: now use the short cut.

Unfortunately, the 0–1 coding isn’t so simple, in part because adding up 0’s and 1’s only seems sensible to mathematicians. So the section goes through the coding in some detail. The students have trouble remembering to put 0’s and 1’s on the tickets. This isn’t so bad with coin-tossing: some numbers are needed, 0 and 1 seem

reasonable. It is harder when rolling die and counting the number of 6's, still harder when taking a sample and counting the number of high-income people. In such examples, the students may already be thinking about some quantitative variable: 0's and 1's pale by comparison. The "classifying and counting" slogan should help, and so does the cartoon on p.301.

There are two other downsides to the 0–1 coding:

- (i) When computing the SD of a 0–1 box, students insist on the factor " $1 - 0$ " in the formula $(1 - 0)\sqrt{p(1 - p)}$. They love substitution; it's what they've been trained to do in math courses.
- (ii) Students may automatically change to 0's and 1's, even for quantitative data. (The crunch comes in part VIII.) To help students use 0–1 boxes only when needed, we try to mix up the exercises a little. For instance, review exercise 9 in chapter 21 is on quantitative variables, even though chapter 21 is about qualitative variables. Conversely, review exercise 5 in chapter 23 involves qualitative data.

Section 5 closes by relating the law of averages to the square root law. It is the square root which makes the SE for the number of heads go up in absolute terms, but down in relative terms. Chapter 17 has a lot of material, and it may spill over into a second lecture. (On the other hand, chapters 16 and 18 go fairly quickly.) We put some emphasis on the idea of "observed values," introduced on p.292. Also see, for instance, exercise 4 on p.293, or 4 and 7 on p.303–4. We think this will help when it comes to statistical inference in parts VI–VIII.

Notes on review exercises. Some students have trouble getting started on exercise 4: the connection between percentages and probabilities may be problematic. Exercise 9 will be difficult for students who think that two games with the same expected value must offer the same chance of winning. This exercise should demonstrate why the SE is needed. (For a preview, see exercise 4 on p.299.) The contrast between expected and observed values is drawn in exercises 6 and 12. Number 10 focuses on the $\sqrt{\quad}$ in the square root law. Number 11 is a hard exercise on setting up box models, as is #14. Number 13 is an interesting variation on the law of large numbers (perhaps too interesting).

Chapter 18. The Normal Approximation for Probability Histograms

We introduce probability calculations for sums through the normal curve. When the number of draws is large, there is about a 68% chance for the sum to be within one SE of the expected value, and so on. This topic is broached in chapter 17 and discussed in chapter 18. The key idea is the "probability histogram"—a graph which represents chance by area. These histograms are drawn *deus ex machina*, by the computer. However, we find graphs easier to use in the classroom than hypothetical lists of all possible samples. (The sample space representation appears as a technical note on p.414; some instructors prefer this approach: our advice would be to do it in the context of quantitative data.) Probability histograms are introduced in figures 1 and 2, as the limit of empirical histograms from simulations. The reason for thinking about products (figure 2) is to see that not everything is normally distributed. The

normal curve is tied to sums. Students should work exercise set A to pin down the interpretation of probability histograms.

Sections 3–4–5 present a “local” version of the central limit theorem: the probability histogram for the sum of a large number of draws from a box will follow the normal curve very closely. However, as the chapter points out, if the distribution of tickets in the box is highly skewed, then many draws may be needed before the approximation takes hold. (This will cause some test anxiety—how can they tell when it is safe to use the curve?) The continuity correction is introduced in section 4, to estimate the chance that the sum will take a given value. The official name itself, “the continuity correction,” appears in the text. The phrase is a bit intimidating, but we wanted students to be able to look it up in case of need, and we wanted them to have some way of packaging the idea. Similarly, we have—with a little publication anxiety—the phrase “central limit theorem.”

Some instructors are troubled by the approach in chapter 18, because they want the “global” central limit theorem: a sum will be in an interval with probability close to the corresponding area under the normal curve. In our experience, students see that if the probability histogram for the sum is close to the curve, areas under the histogram—probabilities—must be close to areas under the curve. The local theorem does imply the global one, both intuitively and formally.

With our approach, probability histograms have to be put into standard units before matching them to the normal curve: that is because we only have one normal curve—with mean 0 and SD 1. The scaling is done in figure 3 on p.315, which is like figure 2 in chapter 5. The elided difficulty is non-trivial: Is the density of $a + bX$ equal to $\frac{1}{b}f\left(\frac{x-a}{b}\right)$? or is it $bf(a + bx)$? Scaling is our substitute for the equation

$$P\{S_n < x\} = P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} < \frac{x - n\mu}{\sigma\sqrt{n}}\right\}.$$

(See note 8 to the chapter.) In our experience, the equation is a loser; scaling works.

Note on the SD. The normal approximation shows why the SD is so useful. The shape of the probability histogram for the sum of a large number of draws from a box depends only on the average and SD of the numbers in the box. Other measures of spread, like average absolute deviation from average, have very little to do with it. (See note 9 to the chapter.)

Notes on review exercises. Exercise 3 tries to reinforce the idea that the histogram gives the exact answer, and the normal curve is just an approximation. Since the probability histogram is a difficult idea, students will confuse it with the histogram for the data—the draws from the box. Exercise 4 is on the continuity correction. Many students will be confused by the “and”; others will want to use the box $\begin{bmatrix} 13 & 0's & 12 & 1's \end{bmatrix}$, a confusion that may resurface in parts VI–VIII. Distinguishing between the data and the model is not so easy; exercise 5 may help. Exercise 6 previews hypothesis testing. Exercises 9 and 10 are about the number of draws needed for the central limit theorem to take over. (Also see exercises 5 and 6 on p.324; the best thing for the students is to look at some pictures.) Exercise 11 reviews observed values, and #15 previews significance testing.

Part VI. Sampling

Chapter 19. Sample Surveys

There are a lot of ideas about sampling which are obvious to statisticians but not to others, and are well worth teaching in an elementary course. For example:

- The method used to draw the sample matters.
- Some methods are terrible.
- Handpicking the sample to get a representative cross-section tends not to work very well.
- Haphazard selection may be even worse.
- The best methods for drawing a sample involve the planned introduction of chance.
- If the non-response rate is high, the results may not be trustworthy.

Jumping straight into the calculations prevents the students from coming to grips with the basic ideas. That is why chapter 19 opens with a qualitative discussion, pinned to historical examples like the *Literary Digest* poll's choice of Landon (section 2), and the Gallup poll's "election" of Dewey (section 3). Probability methods are discussed in section 4, and their success is documented in section 5.

Elementary books (ours is no exception) concentrate on simple random sampling. Of course, the technical meaning of "random" is quite a bit more specialized than the usual meaning:

"Without definite aim, direction, rule, or method."

—Webster's

An effort is required to make students appreciate the technical meaning of "random." We take our best shot in sections 19.4 and 20.1; also see the discussion of "convenience samples" in section 23.4. Review exercise 6 on p.352 may reinforce the point.

Once they know what the terms mean, students think that with a simple random sample, the sample percentage is very likely to equal the population percentage. (They are capable of thinking so, yet going on to compute 95% confidence intervals in response to word problems.) Chapter 16 was designed to prevent this confusion, and section 19.8 continues the work. Again, the chance-error language creates the image of the sample percentage coming close to the population percentage, but missing by a little:

sample percentage = population percentage + chance error.

(There is no bias term with simple random sampling.)

Real sample surveys, of course, use methods much more complicated than simple random sampling. Our book faces up to this issue. *Multistage cluster sampling* is introduced in section 4; it will be discussed again in chapter 22. Section 6 points to some of the difficulties faced by the Gallup poll, and section 7 discusses telephone surveys. Some of our teaching assistants confuse quota sampling with stratified sampling, and then wonder why we are attacking stratification. We aren't. The two methods are very different, although they start out the same way. The crucial difference is that with quota sampling, the interviewer is free to choose respondents

to make up the quota. For a stratified sample, the choice of sampling units within each stratum is done objectively, using chance.

Review exercises 10 and 11 are designed to ease the students into confidence intervals; but the connection may need to be pointed out—later. Review exercise 12 is about non-response bias; so is exercise 12 on p.351. Students like chapter 19, and they have little trouble with the exercises. They do have trouble with the terminology: *sample percentage*, *population*, *population percentage*, and *parameter* are all a bit remote.

Chapter 20. Chance Errors in Sampling

Section 20.1 reviews the definition of simple random sampling, and drives home the idea that the sample percentage will differ from the population percentage. Section 20.2 presents our version of $\sqrt{pq/n}$, except that the formula doesn't appear. (Well, it does, but only in a technical note on p.362.) This may seem a bit idiosyncratic, and we would like to explain why we moved from the conventional formula to our version.

The students seemed to find \sqrt{pq} rather hard to swallow. So we taught them to make a model with 0's and 1's in the box. Since we were working in percents, the formula became

$$\frac{\text{SD of 0-1 box}}{\sqrt{n}} \times 100\%.$$

We presented it that way for several years, but there was still a hitch. The students were willing to compute an SE as $\text{SD} \times \sqrt{n}$ in part V. When they hit part VI, there was a tremendous shifting of gears needed to compute the SE as SD / \sqrt{n} . Once they changed over, they stopped being able to compute the SE for a sum as $\text{SD} \times \sqrt{n}$ —they insisted on dividing. We tried hard to explain that there was one formula to use with sums and another for averages, but they wouldn't buy it.

Eventually, we decided to have only one formula: the SE for a sum. Everything else is worked out from that. For instance, section 2 gives an example where 400 people are chosen at random from a population consisting of 3091 men and 3581 women; the problem is to compute the SE for the percentage of men in the sample. When presenting this problem in a lecture, we begin by writing on the blackboard:

Percent of men in sample will be around _____ give or take _____ or so.

Then we proceed as follows:

Step 1. Set up a box. First we write an empty box on the board:

We ask how many tickets there should be in the box. (Many students will answer 400.) Eventually, we arrive at

3581 0's 3091 1's

The number of men in the sample is like the sum of 400 draws from this box.

The last is a key sentence: it connects the box to the problem. (If students can be persuaded to write this sort of sentence on homework or tests, they will be in relatively good shape; also see exercise 1 on p.391.)

Step 2. Now the calculation can be made:

$$\begin{aligned}\text{expected value for sum of 400 draws} &= 400 \times \text{average of box} \\ &= 400 \times 0.46 = 184. \\ \text{SE for sum of 400 draws} &= \sqrt{400} \times \text{SD of box} \\ &= \sqrt{400} \times \sqrt{0.46 \times 0.54} \\ &\approx 20 \times 0.5 = 10.\end{aligned}$$

We pause to interpret the results: the number of men in the sample will be around 184, give or take 10 or so.

Step 3. Convert to percent: 184 out of 400 is 46%, and 10 out of 400 is 2.5%. So the percentage of men in the sample will be around 46%, give or take 2.5% or so.

This works reasonably well for many students. Others will just compute “the SE” using a formula, and have one chance in four of picking the right formula out of the tool box. There are some exercises to discourage random formulas, for instance, numbers 3 and 5 on p.361. Exercise 7 on p.362 brings back the SE for the sum of quantitative variables.

Students have a hard time connecting the normal approximation for percentages with the mathematics in chapter 18—percentages look quite different from numbers. Figure 3 on p.365 tries to make the connection, and seems to work reasonably well. (Also see figure 1 on p.411.)

So far, we have been a bit sloppy about whether the draws are to be made with or without replacement. When the sample is only a small part of the population, it makes little difference. Section 4 discusses this issue, and eventually comes up with the correction factor

$$\sqrt{\frac{\text{number of tickets in the box} - \text{number of draws}}{\text{number of tickets in the box} - \text{one}}}.$$

In our opinion, this formula is somewhat technical for elementary students, and pushing it too hard obscures the really interesting point. When estimating percentages, accuracy depends mainly on the absolute size of the sample, rather than size relative to the population. On the other hand, when estimating numbers, the game changes (see, e.g., note 5 to the chapter).

Notes on review exercises. Exercise 1 covers the procedure for calculating the SE for a percentage, connecting it to the SE for a number. Exercise 2 puts in a plug for box models. Exercises 5 and 12 require computing the SE for a sum. Likewise, exercises 9–11 are about numbers. (Such exercises help to prevent the students from forgetting about part V.) Exercise 6 tests the point that accuracy depends mainly on the absolute size of the sample rather than the relative size.

Chapter 21. The Accuracy of Percentages

This chapter contains the first technical treatment of inference from the sample to the population. Section 1 states the question to be answered: how accurate is an estimated percentage likely to be? (Before that, however, the section reminds the student of the basic problem—the estimate is apt to be a bit off.) The chapter explains the answer: (i) accuracy is determined by the SE; (ii) the estimate is likely to be about right, but off by an SE or so.

The procedure for estimating the standard error from the sample—substitution of estimates for parameters in the formula—is called “the bootstrap method.” (In our context, the procedure does happen to be a special case of the bootstrap; the samples are large, so the bootstrap works like a charm.) Many students will have trouble, because they do not distinguish between what is known and what is unknown. The point is somewhat delicate. After all, there is a substantial shift from the last chapter to this one. For instance, suppose there is a town with 10,000 residents of voting age and unknown political preferences. To estimate the percentage of Democrats in the town, a simple random sample of size 100 will be used. Consider two strategies:

- Determine the political leanings of every one of these 10,000 people, draw 100 at random and take the percentage of Democrats in the sample.
- Draw 100 at random, determine their political leanings and take the percentage of Democrats in the sample.

The first is zany, the second very practical. The usual standard-error calculation is made by thinking about the first process, the result being carried over to the second. Mathematically, that is fine—the probability distribution for the sample percentage of Democrats is the same in both setups. Students may feel the jolt.

We confront the distinction between the known and the unknown, at least to some degree (pp. 377–79 and 416). We even have some exercises where the students have to say what is given exactly and what must be estimated from the sample: see, for instance, exercise 1 on p. 379, exercise 9 on p. 380, or exercise 1 on p. 383. The distinction between “observed” and “expected” values comes in handy at this point.

Next, we discuss some problems in teaching the main worked example in the section (p. 378). The example is repeated here for ease of reference.

Example 1. In fall 2005, a city university had 25,000 registered students. To estimate the percentage who were living at home, a simple random sample of 400 students was drawn. It turned out that 317 of them were living at home. Estimate the percentage of students at the university who were living at home in fall 2005. Attach a standard error to the estimate.

In working such examples, teaching assistants often demonstrate a natural desire for mathematical efficiency:

$$\frac{\sqrt{400} \times \sqrt{0.79 \times 0.21}}{400} \times 100\% \approx 2\%.$$

We resist, because the parts lose their meaning for the students.

- $\sqrt{0.79 \times 0.21} \approx 0.41$ is the SD of the box, estimated by the bootstrap procedure.

- $\sqrt{400} \times 0.41 \approx 8$ is the SE for the number of students living at home. There were 317 such students in the sample, and the 8 measures the likely size of the chance error in the 317.
- $\frac{8}{400} \times 100\%$ is the SE for the percentage.

Truth to tell, we sometimes dodge the last step by saying, “8 out of 400 is 2 out of 100, or 2%.” The idea is to keep the interpretation as rates, rather than letting percents disappear into ritual formalism.

This chapter makes the transition from probability calculations to statistical inference, and here is one consequence. Students will not take us seriously if we tell them, in working the example, “the sample number will be around its expected value give or take an SE or so.” After all, the sample number is right there in front of them—it *is* 317. But the 317 is a little shaky, being based on a sample; the 8 tells us how shaky: and that is how we interpret the SE.

After dealing with standard errors, the chapter explains how to get confidence intervals for the population percentage at the 68%, 95%, and 99.7% levels by going 1, 2, or 3 SEs either way from the sample percentage. (The distinction between 1.96 SEs and 2 SEs, for instance, just didn’t seem worth pursuing—among other things, the normal approximation may not be right to 3 decimal places.)

The conventional frequency interpretation for confidence intervals is given in section 3. (Bayesian colleagues are asked to temper justice with mercy.) Even for a hard-bitten frequentist, this is a difficult passage to teach, because many students will want to say,

There is a 95% chance that the percentage of Democrats in the town is between. . . .

This is a natural human hope and we try not to deal with it too harshly. The section explains that the chance variability is in the sampling process not in the parameter. Exercises 1 and 2 on p. 386 reinforce the frequentist interpretation. Exercises 4–7 on pp. 386–87 are useful, but students may find the distinctions somewhat irritating.

Unfortunately, students find confidence intervals quite hard. In struggling with the complications, they are likely to lose track of the main point. So the section restates it, on p. 386: the SE tells you the likely size of the amount off. From our perspective, there is nothing wrong with omitting confidence intervals, and focusing on the SE as a measure of reliability. Just be careful about homework assignments.

As mentioned before, the Gallup poll uses a complex multistage cluster sample, and $\sqrt{pq/n}$ does not apply. This is hard. Students want to analyze the data, which is right there in front of them. They do not want to pay attention to the process generating the data, which is more remote. The point is tackled in section 4; also see exercise sets D and E. Many elementary statistics books do not face up to the issue, and perhaps that is one reason why investigators run around computing $\sqrt{pq/n}$ in situations where the results make little sense.

Notes on terminology. (i) We could not write the chapter without using the sample percentage-population percentage terminology, which is confusing to some students. The percentage of Democrats in the sample and the percentage of Democrats

in the town are much more tangible, and the students pick up the idea through the examples. (ii) We try to distinguish between the “true” standard error computed from the box, and the standard error estimated from the sample. The latter is a “standard error of estimate,” but this terminological elaboration would be too confusing. (Our use of SE rather than SD for random quantities is consistent with the standard-error-of-estimate language.)

Notes on review exercises. Exercises 1–2 cover the basics. Exercise 3 reminds the students that confidence intervals depend on the normal approximation (and see 3–4 in exercise set B on p.383). Review exercises 7 and 9 are meant to teach the students not to use the standard error formula where it does not apply. Number 8 tries to block a purely syntactic approach—answering questions on the basis of key words or phrases, or even layout. Exercise 10 on sums is designed to review techniques from part V, and keep quantitative variables alive. Exercise 13 distinguishes between the histogram for the data and the probability histogram. Exercise 14 makes the point that the expected value and standard error depend on the box, not on the draws. Exercise 15 distinguishes what is known from what is estimated, in the sampling context. Some exercises are worded to suggest that calculations may not be feasible; students will find this disturbing, but the idea is an important one.

Chapter 22. Measuring Employment and Unemployment

Government estimates for the unemployment rate are prepared from the monthly Current Population Survey. This sample survey is discussed from the ground up. Such detail is unusual in an elementary text, but it consolidates the understanding of the material presented in the previous chapters, and gives the students a flying start on understanding any other large-scale survey. We don't test the students on details of the design. Mainly, we want them to learn that real surveys do not use simple random samples, so $\sqrt{pq/n}$ does not apply. The standard errors have to be estimated differently, and the half-sample method is sketched in section 5. One conclusion is that the calculation for the standard error should depend on the sample design. If the design is unknown, or poorly defined, sensible calculations are hard to make.

Many professionals are surprised to find that the complex design used by the Current Population Survey gives somewhat less accuracy than a simple random sample. Although the stratification and the ratio estimation reduce sampling error, the clustering increases it (p.402). Of course, without the clustering nobody could afford to do the Survey. The real surprise, to us, is that the Current Population Survey is almost as accurate as a simple random sample. In complex designs, the effective sample size is often reduced by 15% to 50%. The Current Population Survey design is amazingly effective. Other statisticians ask why the ratio estimates are practically unbiased. Our explanation: the sample is very large, so the SEs are rather small, and the ratio estimates are almost linear in the data.

Notes on review exercises. Exercise 1(a) tests understanding of ratio estimates (section 4); part (b) does labor force definitions (section 3). Exercises 2–3–4 are about the half-sample method (section 5). Exercises 5 and 6 review definitions from

chapter 19. Exercise 6 also makes the point that the SE depends on the sampling method. Exercises 7 and 8 test the understanding of probability samples. Exercise 9 is about interviewer bias (chapter 19). Exercise 11 tries to stop the parameter from being the random variable, after the sample is drawn. (Bayesians are permitted a wry chuckle.) Exercise 12 makes the point that confidence levels depend on the normal approximation, which will break down if the distribution is sufficiently skewed.

Chapter 23. The Accuracy of Averages

Section 1 explains how to calculate the standard error for the average of draws made at random with replacement from a box, by working back to the sum (p.410 of the text). The interpretation is that the average of the draws will be around the average of the box, give or take an SE or so. Students handle this reasonably well, although by force of habit a few will go

$$\text{SE for average of draws} = (\text{SE for sum/number of draws}) \times 100\%.$$

Others will want to use the SE for the sum, with little sense that the order of magnitude is wrong. The formula “ σ/\sqrt{n} ” appears only in the technical note on p.415 of the text; we do not teach it for reasons given earlier (pp.27 of this manual).

The application to inference is in section 2. With a simple random sample, the SE of the average is estimated by substituting the SD of the sample for the unknown SD of the box. Then, confidence intervals are obtained by going the right number of SEs either way from the average of the sample. (In this chapter, the samples are large: small samples are dealt with, by Student’s t , in chapter 26.)

At this point, to mix a metaphor, a lot of very tough chickens may come home to roost. Many students are going (somehow) to want 0–1 boxes in section 2. Others will want to use the SE for the sum rather than the SE for the average. Survivors will mix up the probability histogram for the average of the sample with a histogram for the data. Another confusion is between the SD of the sample and the SE of the average, so confidence intervals get interpreted as follows:

95% of the population is within 2 SEs of the average of the sample.

Some students fly over chapter 18, because they see no new techniques presented there. But in chapter 23, they have to come to grips with the central limit theorem. After all, how does the normal curve fit into a problem on educational levels, if the data are so far from normal? Figure 1 on p.411 tries to explain why the probability histogram for the average of the draws follows the normal curve, making the connection to chapter 18 via the obvious (to us but not to them) change of scale.

The ideas in section 2 have all been introduced before, but they are difficult, and they interact in funny ways. Many students profit from studying figures 1 and 2. Others get things under control by working exercise set B. Section 3—and exercise set C—will also help: this exercise set reviews the mechanics and tests the distinctions between what is estimated and what is known. Exercise 4 in set C may seem primitive, but it forces the students to confront the concepts and pay attention to

the scale of a histogram. Despite our best efforts, many students see no relationship among the SEs for sums, averages, numbers, and percents. Section 3 tries once again for unity.

As discussed earlier, the standard-error calculations presuppose simple random sampling, and the students are reminded of this in section 4. The calculations for confidence levels also depend on the normal approximation. Exercises 2–3 in set D (p. 425) make the point. Exercises 4 and 6 reinforce the lesson that the SE depends on the design of the sample—and the SD.

Note on terminology. Students seem to find “sample average” a bit confusing: is it a sample of averages, or what? “The average of the sample” is better, and “the average of the draws” better yet. “Population” also tends to throw things off course. We find ourselves talking about “the box,” and being understood better.

Notes on review exercises. These exercises force the students to distinguish between the SE and the SD. They also make the students separate out the histogram for the sample and the probability histogram for the average of the sample. They teach that the calculations depend on the normal approximation, and on simple random sampling. So they are tough, but provide good diagnostics.

The special review exercises cover parts I–VI. We comment on some of the problems. Numbers 3 and 4 review some issues in study design, and try to sharpen the understanding of confounders. The material on handedness came up in special review exercise 10, chapter 6. The twist here is using average age at death. As epidemiologists know, average age at death is a rather tricky statistic. Exercise 4 is designed to bring out the difficulty.

Exercise 5 is on the mean vs. the median. Exercise 8 reverses number 2 on p. 174, defending against the syntactic approach. Exercise 10 reviews material from chapter 11, and tries to sharpen the connection between inequalities and regions in the scatter diagram. Exercise 11 is another version of 9–10 in chapter 12.

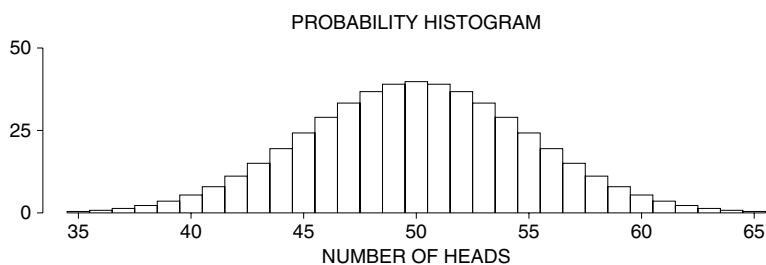
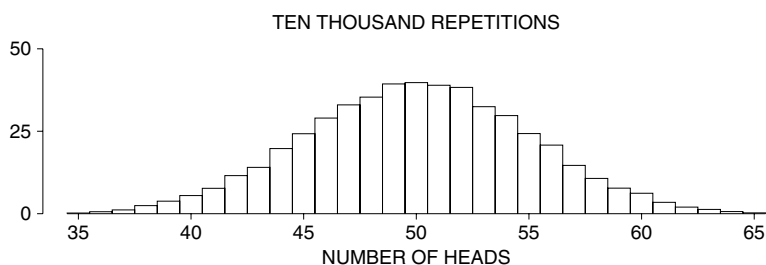
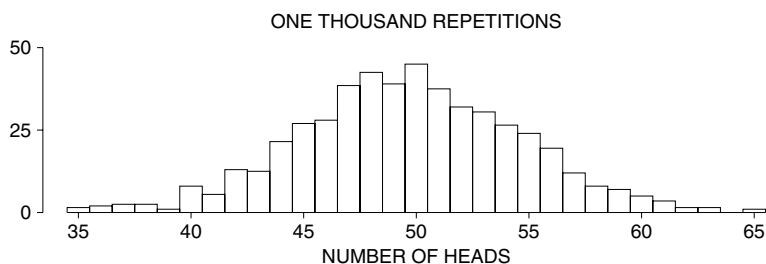
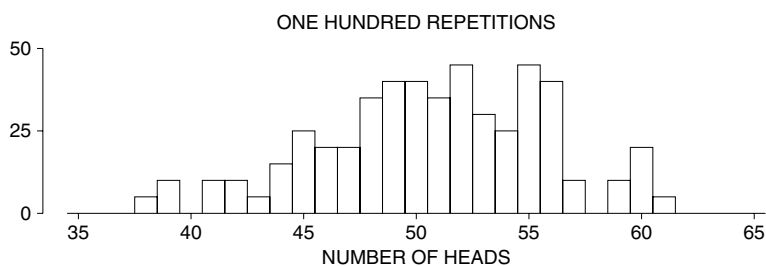
Exercises 13–15 cover part IV. Exercise 16 tries to separate the law of the averages from de Méré’s paradox. Exercise 17 is a hard modeling question. Exercise 22 recaps observed values. Exercise 23 makes them squint at histograms, to see the connection between sums and averages. Exercise 25 is on selection bias. Exercise 27 tests to see if they know what confidence intervals are for. Exercise 28 has some interesting data, and tests the idea of cluster samples. Exercise 30, with random digit dialling, is the flip side of number 6 on p. 372.

Exercise 19—an oldie-but-goodie—reviews probability histograms. There are two stumbling blocks:

- (i) seeing that the number of heads when 100 coins are tossed is like the number of heads when one coin is tossed 100 times;
- (ii) separating repetitions of tossing the coin within the group of 100 from repetitions of tossing the whole group.

The figure on the next page may help.

Special review exercise 23.19. A group of 100 coins are tossed over and over again. The top panel shows data on the number of heads with 100 repetitions, i.e., $100 \times 100 = 10,000$ individual tosses. The second panel is for 1000 repetitions, i.e., $1000 \times 100 = 100,000$ tosses; the third, for 10,000 repetitions, i.e., $10,000 \times 100 = 1,000,000$ tosses. The bottom panel is the probability histogram.



Part VII. Chance Models

In part VII, box models are used to study two topics: measurement error (chapter 24) and genetics (chapter 25). These topics are a bit unusual for an elementary statistics course; instructors who wish to skip them will find that part VIII was written with this possibility in mind. Part VII is designed to reinforce the lesson that to make a good statistical inference, the investigator has to get the box model right.¹

Chapter 24. A Model for Measurement Error

With a large number of measurements, the standard error for the average is estimated as in chapter 23. You start by finding the SE for the sum of the measurements—

$$\sqrt{\text{number of measurements}} \times \text{SD}.$$

Then, you divide by the number of measurements, to get from the sum to the average. As in the sampling context, there is room for confusion between the SE and the SD. The discussion on pp. 442–43 (and the cartoon) try to separate these two quantities.

Despite the familiarity of the arithmetic, there is an issue in this chapter, and it is dealt with in sections 2–3. The procedure for computing the standard error is based on the square root law. The justification depends on viewing the measurements as the observed values of a sequence of independent, identically distributed random variables.

In our experience, that formulation does not convey much to students. We state the idea this way: the data are like the results of drawing at random with replacement from a box of numbered tickets. In particular, if there is any trend or pattern in the data, the model does not apply (pp. 445–49). Dependence between the measurements also rules the model out. Students can use this principle as a heuristic, relying on the ordinary meaning of “dependence.”

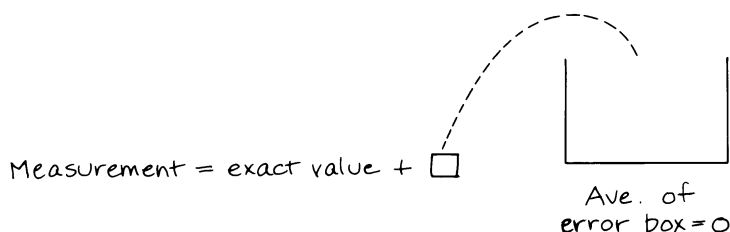
In many cases, the model fits measurement data rather badly. The investigator develops some notion of what the next measurement “ought” to be, based on the previous data, and tends to report this notion instead of the real measurement, destroying the independence. That kind of observer bias is eliminated by the weighing design used at the National Bureau of Standards. See note 8 to the chapter.

Usually, one objective of measurement error models is to make a clean separation between the parameter being estimated (the “exact value” of the thing being measured) and the chance errors. There is a practical reason for this separation. For example, if repeated measurements are made by a certain process on a check weight, the variability in the results can be used to judge the likely size of the chance error in a measurement on another weight (example 5 on p. 451).

We set the model up with this in mind. There is a box of tickets, called the *error box*. Each ticket in the box represents a possible chance error, and the average

¹ Box models look special, because the draws (when made with replacement) are independent. However, the boxes can be modified to handle dependence. Just for one example, a pair of dependent random variables can be modeled by drawing at random from a box of tickets, where each ticket shows a pair of numbers (chapter 27).

of the numbers in the box is assumed to be 0. Then, each measurement equals the exact value of the thing being measured, plus a draw with replacement from the box. This is the Gauss model for measurement error. (The name should not be taken to imply that the errors follow the normal curve.) In our somewhat primitive notation, the model looks like this:



More conventionally, the model would be stated as follows:

$$X_i = \mu + \epsilon_i$$

where the ϵ_i are independent, identically distributed, and have expectation 0.

The model is explained in section 3, and the procedure for calculating the SE is derived from the model. Bias—often a major problem—is taken up at the end of the section. (Up to this point, bias has been assumed to be negligible.) The role of the model in making inferences is summarized in section 4.

Notes on review exercises. Parts (a–b) of exercise 1 are the basic blurs; parts (c–f) try to ward off various misinterpretations of confidence intervals; part (f) is hard. Exercise 2 tries to isolate the role of the normal curve; also see exercise 10. Exercise 6 is about the role of the model. Exercises 8 and 9 bring the SE for the sum back into play; of course, for the students, the first issue is to see that sums are involved.

Chapter 25. Chance Models in Genetics

This chapter gives a brief account of Mendel's genetic theory, based on his experiments with peas. For statisticians, there is an interesting twist to the story: Fisher argued that Mendel's data were massaged to make the frequencies closer to their expected values (section 2). The geneticists do not agree, see note 7 to the chapter. Fisher also showed that Galton's law of regression could be explained by Mendelian theory. One version of the argument is presented in section 3, but it is out of reach for most students.

The physical source of the randomness in Mendelian genetics is described in section 4. This is a tough story, but worth telling. One of the great strengths of the model is the precise description of the physical sources of randomness. As we say in the text, this chapter is included for two reasons:

- Mendel's theory of genetics is beautiful science.
- The theory shows the power of simple chance models in action.

Part VIII. Tests of significance

Chapter 26. Tests of Significance


The basic idea of the z -test is easy. If an observed value is too many SEs away from its expected value, something is wrong. But students find the vocabulary bewildering, and the implicit double negative is hard to follow: investigators usually proceed by rejecting the opposite of what they want to prove. Our objective was to teach the basic idea, and some of the conventional language—null hypothesis, test statistic, P -value. A more reasonable objective, perhaps, is just to teach the idea and skip the language. (Section 26.1 is organized with this possibility in mind.)

We decided to focus on one test first, developing the ideas and the language in that case, and only then moving on to other tests. We chose to start with the z -test. One-tailed tests are used throughout this chapter and the next, as students find them more natural than the two-tailed variety. (There are enough other complications to justify postponing this one to section 29.2.)

Section 1 introduces the idea of the z -test. In the example, the null hypothesis says that the average of the box is 50. The alternative hypothesis says that the average of the box is less than 50. There is a difference between the observed sample average of 48 and the expected value of 50. The null hypothesis interprets this difference as chance variation. The alternative says the difference is real, i.e., reflects a fact about the box.

Section 2 recommends that you set up the null and alternative hypotheses as statements about a box. Few students will pay attention to this advice, but it is the key to all that follows. As we see it, a box model is needed to make the z -test, because the model is what defines the chances. This argument is taken up again in chapter 29.

Section 3 introduces the *test statistic* z and the *observed significance level* or P -value:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}, \quad p \approx \text{area under curve to the left of } z$$


When the P -value of a test is very small, we tend to quote it as a fraction rather than a percent (p.479). Some students will need help in seeing the connection.

The conventional frequentist interpretation of P is given (apologies to our Bayesian colleagues). If the null hypothesis is right, and the experiment is repeated many times, then P is the proportion of repetitions giving z 's more extreme than the observed one. The students are then taught that a test of significance is an argument by contradiction (not an easy pitch to make, because many of them don't know what an argument by contradiction is). Exercises 4–5 in set D try to help with the frequentist interpretation of P .

Section 4 reviews the steps involved in making a test, and introduces the 5% and 1% levels. As we tell the students, a result is *significant* if P is less than 5%, *highly significant* if P is less than 1%. (However, we suggest reporting P instead of just saying how it compares to 5% and 1%.) Many students jump to the conclusion

that P represents the chance of the null hypothesis being true. Measures are taken to prevent this mistake, in the text, in exercise 2 on p.481, and in other exercises. Some students will need to be told, more than once, that small P is bad for the null, big P is good for the null (e.g., exercise 2 on p.482).

Section 5 shows how to make the z -test for qualitative data. The lead example is an ESP experiment done by Charles Tart at U.C. Davis. In this example, and many others, we think there is no natural alternative hypothesis about the box. If a subject has ESP, there is no reason to suppose the successive guesses are independent, so $p > 1/2$ isn't a plausible hypothesis—there is no p . After the first edition of *Statistics* was published, Tart tried to replicate his ESP experiment, but found no effect—section 29.5. (This is association not causation.) He explained the failure to replicate by a change in student attitudes: “In the last year or two, students have become more serious, competitive and achievement-oriented. . . .”



Exercises 1–5 in set E (pp.486ff) go through testing, step by step. Number 9 reinforces the point that the argument is about the box (i.e., the parameters in the model) not the sample. Also see exercise 4 on p.478. Exercise 11 does the sign test. Exercise 10 is interesting, and there are two ways for students to go off the rails:

- (i) using the sample SD instead of the population SD, and
- (ii) making a two-sample test, using the two SDs.

Instructors will get to see the second mistake only by having the exercise on a quiz, after doing chapter 27.

Our version of the z -statistic is

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}.$$

Many students find this equation a bit cryptic, and do not see how get started using it. We ask, “Well, what is observed?” If the observed value is an average, for example,

then they need the expected value for the average, i.e., the average of the box—and the SE for the average of the draws. The discussion on p.485 may help. We motivate the equation this way: z puts the observed value into standard units.

Section 6 does the t -test. We consider this to be a fairly technical topic for an introductory course, and skip it when pressed for time.

Notes on review exercises. The exercises are designed to emphasize the logical steps involved in making a z -test: formulating hypotheses as statements about a box model, then computing z and P . In many of the exercises—for instance, numbers 8 or 10—students will have a very hard time setting up the box model. (The issue for them in working #8 is choosing the right SD.) Exercise 11 boils down to testing whether a coin is fair or biased. Exercise 12 explains methods for handling paired data (the sign test, the z -test on differences).

Note on coverage. We do not introduce the terms *size* or *level*, or use the symbol α . The concept of *power* is not introduced: there is enough to do as it is. The connection between tests and confidence intervals is not established: students rarely see the point of isomorphisms.

Chapter 27. More Tests for Averages

Section 1 explains how to calculate the standard error for the difference of two independent chance quantities. Example 2 and exercises in set A stress the assumption of independence. Section 2 presents the two-sample z -test. The context is the decrease in reading scores over the period 1990–2004, as measured by NAEP (National Assessment of Educational Progress). The section shows how to set up the model, with two boxes. Another example does the 0–1 coding. Our test statistic is the standard one, in disguise (note 3 to the chapter). Some students will get lost in scaling. For instance, they will figure the difference in percentage points, but its SE in decimals. Exercise 5 on p.507 helps. Ideally, of course, each SE should be seen as the margin of error in some estimate.

Section 3 applies the two-sample z -test to experimental data. We set up the model with two possible responses for each subject. One is observed if you put the subject into the treatment group, the other if you put the subject into the control group. But you cannot observe both. Suppose there are N subjects: n are chosen at random for the treatment group, and m for the control group, with $n + m \leq N$. If $n + m$ is much smaller than N , there are in effect two separate boxes, and the theory of section 27.1 applies directly (for a real example, see review exercise 8). Now there is a glitch. If $n + m$ is comparable to N —and $n + m = N$ is the usual case in clinical trials—the treatment and control averages are dependent. Furthermore, the difference between drawing with or without replacement matters. In principle, then, it is wrong to model the data as two independent samples drawn from two large boxes. Fortunately or otherwise, this fine point has no practical consequences. Ordinarily, treating the two samples as independent and drawn with replacement will give an excellent approximation to the SE for the difference between the averages. (For discussion, see notes 11 and 14 to the chapter, which also provide a brief review of the literature on the model.)

In example 4 on p. 508, the calculation is made blindly. The logic is discussed afterward, on pp. 509–10. This is a difficult passage. Students have to work hard to see that the sample averages are dependent. Some of them will be irritated to find that the dependence does not matter—for reasons which may also seem mysterious. Section 4 presents a real example with qualitative data—an experimental test of “rational” decision theory. Exercise 3 on p. 515 does some calculations for the HIP trial on mammography (pp. 22–23), and points to a design issue. Breast cancer is a rare disease. Even if screening cuts the death rate from breast cancer in half, the impact on the total death rate is unlikely to achieve statistical significance—unless sample sizes are incredibly large. That is why investigators look at cause-specific mortality rates. Section 5 tries to explain when the z -test applies.

Notes on review exercises. The point of exercise 1 is to make the students distinguish between one-sample and two-sample tests. Exercises 2–3 are straightforward two-sample problems; number 2(b) hints that averages may have better power. In exercise 4, the test cannot be done—dependence (see exercises 5–6 on pp. 515–16). Review exercises 5–7 are fairly straightforward experimental setups. Number 8, again on experiments, is much harder. Students either don’t see what is being compared to what, or find the comparisons too unnatural to make. In grading this one, we insist on a substantive conclusion—for instance, that people are poor predictors of their own behavior, but tend to live up to their predictions about themselves—as one character in the drawing understands. Question 11 is very theoretical. The hope is to persuade at least some of the students that a significance test is not a ritual, but an argument that has its own internal logic.

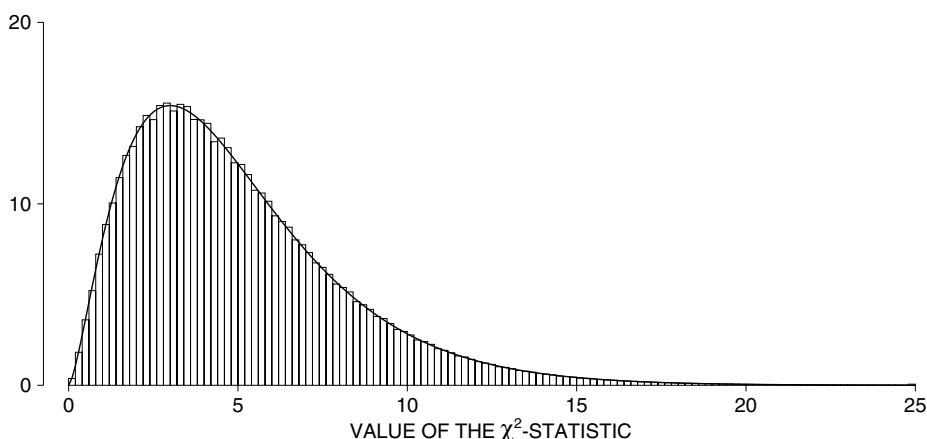


“I’m not asking for a raise, Sir. I just want to know how you would react if I did.”

Chapter 28. The Chi-Square Test

Section 1 presents the χ^2 -test for goodness of fit, when the model is completely specified. Students have a hard time deciding when to use the χ^2 -test and when to use the z -test; some help is given on p.523; also see exercises 3–6 on pp.539–40. The text explains how to read the χ^2 -table (p.527), and says that the χ^2 -distribution is only an approximation. The mathematical underpinnings for the approximation are discussed in section 1. The main one is a box model; this is emphasized in the text. Figure 2 (p.528) for 60 rolls may help. As the number of rolls goes up—60, 600, 6000—the probability histogram will get closer and closer to the smooth curve. The figure below plots the probability histogram for 600 rolls. The histogram is already very close to the curve.

Probability histogram for the null distribution of the χ^2 -statistic in 600 rolls of a fair die. (Continues figure 2 in chapter 28.)



Section 1 closes with a real example—testing the wheel of fortune. Section 2 describes χ^2 , in some degree of generality, as a goodness-of-fit test. Section 3 discusses the pooling of independent χ^2 's, and shows how Fisher used the χ^2 -test to check up on Mendel (but see note 7 to chapter 25, for the geneticists' counter-arguments). Fisher computed a left-hand tail area, rather than a right-hand tail (p.534). Students see a possible trap, so the issue will get air time. Section 4 shows how χ^2 is used to test for dependence in $m \times n$ tables. With the current exposition, this is fairly easy going.

Notes on review exercises. Exercise 1 confronts the issue of which test to use when. Exercises 2 and 7 are straightforward goodness-of-fit questions. Number 3 does independence in a 3×3 table, while number 9 does a 3×2 table. Exercises 4–5 are qualitative, and get the students to focus again on probability histograms and tail areas. Students may find exercise 6 a little ambiguous, but left-hand tail areas are called for. Exercise 10 is based on a court case—does the χ^2 -test show discrimination in the criminal justice system of Northern Ireland? (In a law case, finding a mistake by an opposing expert is powerful—and guessing how the mistake was made is dynamite.)

Chapter 29. A Closer Look at Tests of Significance

Many people find tests of significance both complicated and mysterious. Perhaps as a result, the limitations of the technique are often ignored. This often creates unnecessary confusion. So we think it is important to discuss what tests of significance don't do. That is the topic of chapter 29.

Section 1 is about fixed-level testing (a procedure we do not recommend). Section 2 covers data snooping. Students find it very hard to understand that significance levels are compromised by multiple looks at the data. Exercise 5 on p.483 and exercise 1 on p.550 should help, a little; exercises 2–5 on pp.551–52 give some practical examples. We see the “one-tail-or-two” issue as quite minor.¹ Many professionals will not agree with us, and the students like a definite rule for deciding whether to use a one-tailed or a two-tailed test (pp.547–50). The issue will get some attention.

Section 3 tries to explain that small differences can be statistically significant—or big differences insignificant—depending on the sample size. This point is hard, and irritating. Students have invested a lot of time learning how to operate the tool, they want it to be useful. Section 4 is about the role of the model in testing. Since the arithmetic of the test seems to generate the chances—the P -value—this section is quite subtle. Section 5 stresses the role of design, and section 6 is a reminder about the basic question being addressed by significance tests: is the difference too large to explain by chance?

Notes on review exercises. Exercises 1 and 2 are straightforward questions, which can be answered from the reading. Exercise 3 is a math question but a little tricky, the point being that P -values depend on sample size. Exercise 4 is about data snooping, among other things; hard. Exercises 5 and 7 are about not doing tests when you have all the data, an idea the students pick up. Exercises 8–9 are on sample design, and are hard. Exercises 6 and 10 are about real studies and raise real questions; very hard.

Finally, we comment on some of the special review exercises, which cover the whole book. Exercises 1–2 are on study design. Exercise 3 covers Simpson's paradox. Exercise 4 makes the point that histograms are different from bar graphs (also see exercise 8 on pp.52–53). Exercise 6 is on percentiles for skew distributions. Exercise 7 makes them look at scatter diagrams (to find the child brides and grooms at the lower left). Exercises 8–14 cover part III: number 8 involves a lot of work on a small data set. Exercise 9(a) does attenuation, while 9(b) covers ecological correlations. Exercise 10 is another variation on the regression effect. Exercise 11 involves percentile ranks, and will be quite a challenge. With exercise 14, students will have to work out a percentage from the normal approximation, then a number.

Exercises 15–17 cover parts IV and V. Exercise 15 requires careful reading; compare exercise 11 on p.253. Exercise 16 looks like a binomial problem, but it isn't—they will need to think about the ideas in order to work the problem. Exercise 17 requires a box model; not completely transparent. Sampling is the next topic.

¹ The data-snooping that goes into developing a typical regression model seems much more serious; of course, the application to cholesterol is far from minor (example 2 on p.550).

Number 18 is on selection bias; 19 is on evaluation of survey results. Exercise 21 puts reverse spin on #30, p. 436. Exercise 22 combines ideas from sampling with the continuity correction. Exercise 23 examines some design issues in sampling. Students often make “cluster sample” mean any kind of sample they don’t like, and we try to block that move. Exercises 25, 26, and 28 try to stop some misinterpretations of expected values and confidence levels; #28 is based on a court opinion which got the wrong answer. Exercise 31 is on measurement error, and 32 on genetics. Exercise 33 tries to make the students understand when to use a one-sample z -test or a two-sample test. Exercise 34 is to prevent misinterpretations of P . The data in exercise 36 are interesting, and the idea is not to make a two-sample z -test with correlated responses.