

International Student Edition

# Statistics

FOURTH EDITION



David Freedman  
Robert Pisani  
Roger Purves

NOT FOR SALE IN THE UNITED STATES OR CANADA

This is trial version  
[www.adultpdf.com](http://www.adultpdf.com)

This is trial version  
[www.adultpdf.com](http://www.adultpdf.com)

# Statistics

Fourth Edition

This is trial version

DAVID FREEDMAN  
ROBERT PISANI  
ROGER PURVES

# Statistics

## Fourth Edition

W • W • NORTON & COMPANY

NEW YORK • LONDON



Copyright © 2007, 1998, 1991, 1978 by W. W. Norton & Company, Inc.  
All rights reserved.  
Printed in the United States of America.

*Cartoons by Dana Fradon and Leo Cullum*

The text of this book is composed in Times Roman.  
Composition by Integre Technical Publishing Company, Inc.  
Manufacturing by R. R. Donnelley.

**Library of Congress Cataloging-in-Publication Data**

Freedman, David, 1938—

Statistics. — 4th ed. / David Freedman, Robert Pisani, Roger Purves.

p. cm.

Rev. ed. of: Statistics / David Freedman . . . [et al.], 3rd ed.

©1998.

Includes bibliographical references and index.

**ISBN 0-393-92972-8**

**ISBN 13-978-0-393-92972-0**

1. Mathematical statistics. I. Pisani, Robert. II. Purves, Roger. III. Statistics. IV. Title.

QA276.F683

519.5—dc21

W.W. Norton & Company, Inc., 500 Fifth Avenue, New York, N.Y. 10110  
<http://www.wwnorton.com>

W.W. Norton & Company Ltd., Castle House, 75/76 Wells Street, London W1T 3QT

To Jerzy Neyman (1894–1981)

*Born in Russia, Neyman worked in Poland and England before coming to the United States in 1938. He was one of the great statisticians of our time.*

This is trial version

# Contents

## Preface

xv

## PART I. DESIGN OF EXPERIMENTS

<b>Chapter 1. Controlled Experiments</b>	<b>3</b>
1. The Salk Vaccine Field Trial	3
2. The Portacaval Shunt	7
3. Historical Controls	8
4. Summary	10
<b>Chapter 2. Observational Studies</b>	<b>12</b>
1. Introduction	12
2. The Clofibrate Trial	13
3. More Examples	15
4. Sex Bias in Graduate Admissions	17
5. Confounding	20
6. Review Exercises	24
7. Summary and Overview	27

## PART II. DESCRIPTIVE STATISTICS

<b>Chapter 3. The Histogram</b>	<b>31</b>
1. Introduction	31
2. Drawing a Histogram	35
3. The Density Scale	38
4. Variables	42
5. Controlling for a Variable	45
6. Cross-Tabulation	47
7. Selective Breeding	48
8. Review Exercises	50
9. Summary	56
<b>Chapter 4. The Average and the Standard Deviation</b>	<b>57</b>
1. Introduction	57
2. The Average	58
3. The Average and the Histogram	61
4. The Root-Mean-Square	66
5. The Standard Deviation	67
6. Computing the Standard Deviation	71
7. Using a Statistical Calculator	74
8. Review Exercises	74
9. Summary	76

<b>Chapter 5. The Normal Approximation for Data</b>	<b>78</b>
1. The Normal Curve	78
2. Finding Areas under the Normal Curve	82
3. The Normal Approximation for Data	85
4. Percentiles	88
5. Percentiles and the Normal Curve	90
6. Change of Scale	92
7. Review Exercises	93
8. Summary	96
<b>Chapter 6. Measurement Error</b>	<b>97</b>
1. Introduction	97
2. Chance Error	97
3. Outliers	102
4. Bias	103
5. Review Exercises	104
6. Special Review Exercises	105
7. Summary and Overview	108
<b>Chapter 7. Plotting Points and Lines</b>	<b>110</b>
1. Reading Points off a Graph	110
2. Plotting Points	112
3. Slope and Intercept	113
4. Plotting Lines	114
5. The Algebraic Equation for a Line	115
 <b>PART III. CORRELATION AND REGRESSION</b>	
<b>Chapter 8. Correlation</b>	<b>119</b>
1. The Scatter Diagram	119
2. The Correlation Coefficient	125
3. The SD Line	130
4. Computing the Correlation Coefficient	132
5. Review Exercises	134
6. Summary	139
<b>Chapter 9. More about Correlation</b>	<b>141</b>
1. Features of the Correlation Coefficient	141
2. Changing SDs	144
3. Some Exceptional Cases	147
4. Ecological Correlations	148
5. Association is Not Causation	150
6. Review Exercises	153
7. Summary	157
<b>Chapter 10. Regression</b>	<b>158</b>
1. Introduction	158
2. The Graph of Averages	162

3. The Regression Method for Individuals	165
4. The Regression Fallacy	169
5. There Are Two Regression Lines	174
6. Review Exercises	176
7. Summary	178
<b>Chapter 11. The R.M.S. Error for Regression</b>	<b>180</b>
1. Introduction	180
2. Computing the R.M.S. Error	185
3. Plotting the Residuals	187
4. Looking at Vertical Strips	190
5. Using the Normal Curve Inside a Vertical Strip	195
6. Review Exercises	198
7. Summary	201
<b>Chapter 12. The Regression Line</b>	<b>202</b>
1. Slope and Intercept	202
2. The Method of Least Squares	208
3. Does the Regression Make Sense?	211
4. Review Exercises	213
5. Summary and Overview	216
<b>PART IV. PROBABILITY</b>	
<b>Chapter 13. What Are the Chances?</b>	<b>221</b>
1. Introduction	221
2. Conditional Probabilities	226
3. The Multiplication Rule	228
4. Independence	230
5. The Collins Case	233
6. Review Exercises	234
7. Summary	236
<b>Chapter 14. More about Chance</b>	<b>237</b>
1. Listing the Ways	237
2. The Addition Rule	241
3. Two FAQs (Frequently Asked Questions)	243
4. The Paradox of the Chevalier De Méré	248
5. Are Real Dice Fair?	252
6. Review Exercises	252
7. Summary	254
<b>Chapter 15. The Binomial Formula</b>	<b>255</b>
1. Introduction	255
2. The Binomial Formula	259
3. Review Exercises	261
4. Special Review Exercises	263
5. Summary and Overview	268

**PART V. CHANCE VARIABILITY**

<b>Chapter 16. The Law of Averages</b>	<b>273</b>
1. What Does the Law of Averages Say?	273
2. Chance Processes	278
3. The Sum of Draws	279
4. Making a Box Model	281
5. Review Exercises	285
6. Summary	287
<b>Chapter 17. The Expected Value and Standard Error</b>	<b>288</b>
1. The Expected Value	288
2. The Standard Error	290
3. Using the Normal Curve	294
4. A Short-Cut	298
5. Classifying and Counting	299
6. Review Exercises	304
7. Postscript	307
8. Summary	307
<b>Chapter 18. The Normal Approximation for Probability Histograms</b>	<b>308</b>
1. Introduction	308
2. Probability Histograms	310
3. Probability Histograms and the Normal Curve	315
4. The Normal Approximation	317
5. The Scope of the Normal Approximation	319
6. Conclusion	325
7. Review Exercises	327
8. Summary	329

**PART VI. SAMPLING**

<b>Chapter 19. Sample Surveys</b>	<b>333</b>
1. Introduction	333
2. The <i>Literary Digest</i> Poll	334
3. The Year the Polls Elected Dewey	337
4. Using Chance in Survey Work	339
5. How Well Do Probability Methods Work?	342
6. A Closer Look at the Gallup Poll	343
7. Telephone Surveys	346
8. Chance Error and Bias	348
9. Review Exercises	351
10. Summary	353
<b>Chapter 20. Chance Errors in Sampling</b>	<b>355</b>
1. Introduction	355
2. The Expected Value and Standard Error	359
3. Using the Normal Curve	362

4. The Correction Factor	367
5. The Gallup Poll	370
6. Review Exercises	371
7. Summary	373
<b>Chapter 21. The Accuracy of Percentages</b>	<b>375</b>
1. Introduction	375
2. Confidence Intervals	381
3. Interpreting a Confidence Interval	383
4. <i>Caveat Emptor</i>	387
5. The Gallup Poll	389
6. Review Exercises	391
7. Summary	394
<b>Chapter 22. Measuring Employment and Unemployment</b>	<b>395</b>
1. Introduction	395
2. The Design of the Current Population Survey	396
3. Carrying out the Survey	398
4. Weighting the Sample	401
5. Standard Errors	402
6. The Quality of the Data	404
7. Bias	404
8. Review Exercises	405
9. Summary	407
<b>Chapter 23. The Accuracy of Averages</b>	<b>409</b>
1. Introduction	409
2. The Sample Average	415
3. Which SE?	422
4. A Reminder	424
5. Review Exercises	425
6. Special Review Exercises	428
7. Summary and Overview	436
<b>PART VII. CHANCE MODELS</b>	
<b>Chapter 24. A Model for Measurement Error</b>	<b>441</b>
1. Estimating the Accuracy of an Average	441
2. Chance Models	445
3. The Gauss Model	450
4. Conclusion	454
5. Review Exercises	455
6. Summary	457
<b>Chapter 25. Chance Models in Genetics</b>	<b>458</b>
1. How Mendel Discovered Genes	458
2. Did Mendel's Facts Fit His Model?	463
3. The Law of Regression	465

4. An Appreciation of the Model	468
5. Review Exercises	470
6. Summary and Overview	471

## PART VIII. TESTS OF SIGNIFICANCE

<b>Chapter 26. Tests of Significance</b>	<b>475</b>
1. Introduction	475
2. The Null and the Alternative	477
3. Test Statistics and Significance Levels	478
4. Making a Test of Significance	482
5. Zero-One Boxes	483
6. The <i>t</i> -Test	488
7. Review Exercises	495
8. Summary	500
<b>Chapter 27. More Tests for Averages</b>	<b>501</b>
1. The Standard Error for a Difference	501
2. Comparing Two Sample Averages	503
3. Experiments	508
4. More on Experiments	512
5. When Does the <i>z</i> -Test Apply?	517
6. Review Exercises	518
7. Summary	521
<b>Chapter 28. The Chi-Square Test</b>	<b>523</b>
1. Introduction	523
2. The Structure of the $\chi^2$ -Test	530
3. How Fisher Used the $\chi^2$ -Test	533
4. Testing Independence	535
5. Review Exercises	540
6. Summary	544
<b>Chapter 29. A Closer Look at Tests of Significance</b>	<b>545</b>
1. Was the Result Significant?	545
2. Data Snooping	547
3. Was the Result Important?	552
4. The Role of the Model	555
5. Does the Difference Prove the Point?	560
6. Conclusion	562
7. Review Exercises	563
8. Special Review Exercises	565
9. Summary and Overview	576
<b>Notes</b>	<b>A3</b>
<b>Answers to Exercises</b>	<b>A43</b>
<b>Tables</b>	<b>A104</b>
<b>Index</b>	<b>A107</b>

# Preface

*What song the Sirens sang, or what name Achilles assumed when he hid among women, though puzzling questions, are not beyond all conjecture.*

—SIR THOMAS BROWNE (ENGLAND, 1605–1682)

## TO THE READER

We are going to tell you about some interesting problems which have been studied with the help of statistical methods, and show you how to use these methods yourself. We will try to explain why the methods work, and what to watch out for when others use them. Mathematical notation only seems to confuse things for many people, so this book relies on words, charts, and tables; there are hardly any  $x$ 's or  $y$ 's. As a matter of fact, even when professional mathematicians read technical books, their eyes tend to skip over the equations. What they really want is a sympathetic friend who will explain the ideas and draw the pictures behind the equations. We will try to be that friend, for those who read our book.

## WHAT IS STATISTICS?

Statistics is the art of making numerical conjectures about puzzling questions.

- What are the effects of new medical treatments?
- What causes the resemblance between parents and children, and how strong is that force?
- Why does the casino make a profit at roulette?
- Who is going to win the next election? by how much?
- How many people are employed? unemployed?

These are difficult issues, and statistical methods help a lot if you want to think about them. The methods were developed over several hundred years by people who were looking for answers to their questions. Some of these people will be introduced later.

## AN OUTLINE

Part I is about designing experiments. With a good design, reliable conclusions can be drawn from the data. Some badly-designed studies are discussed too—so you can see the pitfalls, and learn what questions to ask when reading about a study. Study design is perhaps our most important topic; that is why we start there. The ideas look simple, but appearances may be deceptive: part I has a lot of depth.

Studies typically produce so many numbers that summaries are needed. Descriptive statistics—the art of summarizing data—is introduced in part II. Histograms, the average, the standard deviation, and the normal curve are all considered. The discussion continues in part III, where the focus is on analyzing relationships, for instance, the dependence of income on education. Here, correlation and regression are the main topics.

Much statistical reasoning depends on the theory of probability, discussed in part IV; the connection is through chance models, which are developed in part V. Coins, dice, and roulette wheels are the main examples in parts IV and V. The expected value and standard error are introduced; probability histograms are developed, and convergence to the normal curve is discussed.

Statistical inference—making valid generalizations from samples—is the topic of parts VI–VIII. Part VI is about estimation. For instance, how does the Gallup Poll predict the vote? Why are some methods for drawing samples better than others? Part VII uses chance models to analyze measurement error, and to develop genetic theory. Part VIII introduces tests of significance, to judge whether samples are consistent with hypotheses about the population. As parts VI–VIII show, statistical inferences depend on chance models. If the model is wrong, the resulting inference may be quite shaky.

Nowadays, inference is the branch of statistics most interesting to professionals. However, non-statisticians often find descriptive statistics a more useful branch, and the one that is easier to understand. That is why we take up descriptive statistics before inference. The bare bones of our subject are presented in chapters 1 to 6, 13, 16 to 21, 23, and 26. After that, the reader can browse anywhere. The next chapters to read might be 8, 10, 27, and 29.

## EXERCISES

The sections in each chapter usually have a set of exercises, with answers at the back of the book. If you work these exercises as they come along and check the answers, you will get practice in your new skills—and find out the extent to which you have mastered them. Every chapter (except 1 and 7) ends with a set of review exercises. The book does not give answers for those exercises. Chapters 6, 15, 23, and 29 also have “special review exercises,” covering all previous material. Such exercises must be answered without the clues provided by context.

When working exercises, you might be tempted to flip backward through the pages until the relevant formula materializes. However, reading the book backward will prove very frustrating. Review exercises demand much more than formulas. They call for rough guesses and qualitative judgments. In other words, they require a good intuitive understanding of what is going on. The way to develop that understanding is to read the book forward.

Why does the book include so many exercises that cannot be solved by plugging into a formula? The reason is that few real-life statistical problems can be solved that way. Blindly plugging into statistical formulas has caused a lot of confusion. So this book teaches a different approach: thinking.

## GRAPHICS

As in previous editions, extensive use is made of computer graphics to display the data. Working drawings, however, are done freehand; the reader is encouraged to make similar sketches, rather than being intimidated by too much precision. The book still features cartoons by Dana Fradon of *The New Yorker*.

## What's New in the Fourth Edition?

*Of the making of books, there is no end.*

—Ecclesiastes

The principal change is to the data. Statistics, like people, show wear and tear from aging. Fortunately or unfortunately, data are easier to rejuvenate. We started the first edition in 1971, and completed the fourth in 2006. These past 35 years were years of rapid change, as commentators have doubtless observed since prehistoric times.

There was explosive growth in computer use. Other technical developments include email (+), the world wide web (+), Windows ( $\pm$ ), cell phones ( $\pm$ ), and call centers with voice-activated menus (−). SAT scores bottomed out around 1990, and have since been slowly going up (chapter 5). Educational levels have been steadily increasing (chapter 4), but reading skills may—or may not—be in decline (chapter 27).

The population of the United States increased from 200 million to 300 million (chapter 24). There was corresponding growth in higher education. Over the period 1976 to 1999, the number of colleges and universities increased from about 3,000 to 4,000 (chapter 23). Student enrollments increased by about 40%, while the professoriate grew by 60%. The number of male faculty increased from 450,000 to 600,000; for women, the increase was 175,000 to 425,000. Student enrollments shifted from 53% male to 43% male.

There were remarkable changes in student attitudes (chapters 27, 29). In 1970, 60% of first-year students thought that capital punishment should be abolished; by 2000, only 30% favored abolition. In 1970, 36% of them thought that “being very well off financially” was “very important or essential”; by 2000, the figure was 73%.

The American public gained a fraction of an inch in height, and 20 pounds in weight (chapter 4). Despite the huge increase in obesity, there were steady gains in life expectancy—about 7 years over the 35-year period. Gain in life expectancy is a process (“the demographic transition”) that started in Europe around 1800. The trend toward longer lives has major societal implications, as well as ripple effects on our exercises.

Family incomes went up by a factor of four, although much of the change represents a loss of purchasing power in the dollar (chapter 3). Crime rates peaked somewhere around 1990, and have fallen precipitously since (chapters 2, 29). Jury awards in civil cases once seemed out of control, but have declined since the 1990s

along with crime rates. (See chapter 29; is this correlation or causation?) Our last topic is a perennial favorite: the weather. We have no significant changes to report (chapters 9, 24).\*

#### ACKNOWLEDGMENTS FOR THE FOURTH EDITION

Technical drawings are by Dale Johnson and Laura Southworth. Type was set in TeX by Integre. Nick Cox (Durham), Russ Lyons (Indiana), and Sam Rose (Berkeley) gave us detailed and useful feedback. Máire Ní Bhrolcháin (Southampton), David Card (Berkeley), Rob Hollister (Swarthmore), Josh Palmer (Berkeley), Diana Petitti (Kaiser Permanente), and Philip Stark (Berkeley) helped us navigate the treacherous currents of the scholarly literature, and the even more treacherous currents of the world wide web.

#### ACKNOWLEDGMENTS FOR PREVIOUS EDITIONS

Helpful comments came from many sources. For the third edition, we thank Mike Anderson (Berkeley), Dick Berk (Pennsylvania), Jeff Fehmi (Arizona), David Kaye (Arizona), Steve Klein (Los Angeles), Russ Lyons (Indiana), Mike Ostland (Berkeley), Erol Pekoz (Boston), Diana Petitti (Kaiser Permanente), Juliet Shaffer (Berkeley), Bill Simpson (Winnipeg), Terry Speed (Berkeley), Philip Stark (Berkeley), and Allan Stewart-Oaten (Santa Barbara). Ani Adhikari (Berkeley) participated in the second edition, and had many good comments on the third edition.

The writing of the first edition was supported by the Ford Foundation (1973–1974) and by the Regents of the University of California (1974–75). Earl Cheit and Sanford Elberg (Berkeley) provided help and encouragement at critical times. Special thanks go to our editor, Donald Lamm, who somehow turned a permanently evolving manuscript into a book. Finally, we record our gratitude to our students, and other readers of our several editions and innumerable drafts.

---

\*Most of the data cited here come from the *Statistical Abstract of the United States*, various editions. See chapter notes for details. On trends in life expectancy, see Dudley Kirk, "Demographic transition theory," *Population Studies* vol. 50 (1996) pp. 361–87.

## PART I

# Design of Experiments

— — — — —

# 1

## Controlled Experiments

*Always do right. This will gratify some people, and astonish the rest.*

—MARK TWAIN (UNITED STATES, 1835–1910)

### 1. THE SALK VACCINE FIELD TRIAL

A new drug is introduced. How should an experiment be designed to test its effectiveness? The basic method is *comparison*.<sup>1</sup> The drug is given to subjects in a *treatment group*, but other subjects are used as *controls*—they aren’t treated. Then the responses of the two groups are compared. Subjects should be assigned to treatment or control *at random*, and the experiment should be run *double-blind*: neither the subjects nor the doctors who measure the responses should know who was in the treatment group and who was in the control group. These ideas will be developed in the context of an actual field trial.<sup>2</sup>

The first polio epidemic hit the United States in 1916, and during the next forty years polio claimed many hundreds of thousands of victims, especially children. By the 1950s, several vaccines against this disease had been discovered. The one developed by Jonas Salk seemed the most promising. In laboratory trials, it had proved safe and had caused the production of antibodies against polio. By 1954, the Public Health Service and the National Foundation for Infantile Paralysis (NFIP) were ready to try the vaccine in the real world—outside the laboratory.

Suppose the NFIP had just given the vaccine to large numbers of children. If the incidence of polio in 1954 dropped sharply from 1953, that would seem to

prove the effectiveness of the vaccine. However, polio was an epidemic disease whose incidence varied from year to year. In 1952, there were about 60,000 cases; in 1953, there were only half as many. Low incidence in 1954 could have meant that the vaccine was effective—or that 1954 was not an epidemic year.

The only way to find out whether the vaccine worked was to deliberately leave some children unvaccinated, and use them as controls. This raises a troublesome question of medical ethics, because withholding treatment seems cruel. However, even after extensive laboratory testing, it is often unclear whether the benefits of a new drug outweigh the risks.<sup>3</sup> Only a well-controlled experiment can settle this question.

In fact, the NFIP ran a controlled experiment to show the vaccine was effective. The subjects were children in the age groups most vulnerable to polio—grades 1, 2, and 3. The field trial was carried out in selected school districts throughout the country, where the risk of polio was high. Two million children were involved, and half a million were vaccinated. A million were deliberately left unvaccinated, as controls; half a million refused vaccination.

This illustrates the method of comparison. Only the subjects in the treatment group were vaccinated: the controls did not get the vaccine. The responses of the two groups could then be compared to see if the treatment made any difference. In the Salk vaccine field trial, the treatment and control groups were of different sizes, but that did not matter. The investigators compared the rates at which children got polio in the two groups—cases per hundred thousand. Looking at rates instead of absolute numbers adjusts for the difference in the sizes of the groups.

Children could be vaccinated only with their parents' permission. So one possible design—which also seems to solve the ethical problem—was this. The children whose parents consented would go into the treatment group and get the vaccine; the other children would be the controls. However, it was known that higher-income parents would more likely consent to treatment than lower-income parents. This design is biased against the vaccine, because children of higher-income parents are more vulnerable to polio.

That may seem paradoxical at first, because most diseases fall more heavily on the poor. But polio is a disease of hygiene. A child who lives in less hygienic surroundings is more likely to contract a mild case of polio early in childhood, while still protected by antibodies from its mother. After being infected, these children generate their own antibodies, which protect them against more severe infection later. Children who live in more hygienic surroundings do not develop such antibodies.

Comparing volunteers to non-volunteers biases the experiment. The statistical lesson: the treatment and control groups should be as similar as possible, except for the treatment. Then, any difference in response between the two groups is due to the treatment rather than something else. If the two groups differ with respect to some factor other than the treatment, the effect of this other factor might be *confounded* (mixed up) with the effect of treatment. Separating these effects can be difficult, and confounding is a major source of bias.

For the Salk vaccine field trial, several designs were proposed. The NFIP had originally wanted to vaccinate all grade 2 children whose parents would consent,

leaving the children in grades 1 and 3 as controls. And this design was used in many school districts. However, polio is a contagious disease, spreading through contact. So the incidence could have been higher in grade 2 than in grades 1 or 3. This would have biased the study against the vaccine. Or the incidence could have been lower in grade 2, biasing the study in favor of the vaccine. Moreover, children in the treatment group, where parental consent was needed, were likely to have different family backgrounds from those in the control group, where parental consent was not required. With the NFIP design, the treatment group would include too many children from higher-income families. The treatment group would be more vulnerable to polio than the control group. Here was a definite bias against the vaccine.

Many public health experts saw these flaws in the NFIP design, and suggested a different design. The control group had to be chosen from the same population as the treatment group—children whose parents consented to vaccination. Otherwise, the effect of family background would be confounded with the effect of the vaccine. The next problem was assigning the children to treatment or control. Human judgment seems necessary, to make the control group like the treatment group on the relevant variables—family income as well as the children's general health, personality, and social habits.

Experience shows, however, that human judgment often results in substantial bias: it is better to rely on impersonal chance. The Salk vaccine field trial used a chance procedure that was equivalent to tossing a coin for each child, with a 50–50 chance of assignment to the treatment group or the control group. Such a procedure is objective and impartial. The laws of chance guarantee that with enough subjects, the treatment group and the control group will resemble each other very closely with respect to all the important variables, whether or not these have been identified. When an impartial chance procedure is used to assign the subjects to treatment or control, the experiment is said to be *randomized controlled*.<sup>4</sup>

Another basic precaution was the use of a *placebo*: children in the control group were given an injection of salt dissolved in water. During the experiment the subjects did not know whether they were in treatment or in control, so their response was to the vaccine, not the idea of treatment. It may seem unlikely that subjects could be protected from polio just by the strength of an idea. However, hospital patients suffering from severe post-operative pain have been given a “pain killer” which was made of a completely neutral substance: about one-third of the patients experienced prompt relief.<sup>5</sup>

Still another precaution: diagnosticians had to decide whether the children contracted polio during the experiment. Many forms of polio are hard to diagnose, and in borderline cases the diagnosticians could have been affected by knowing whether the child was vaccinated. So the doctors were not told which group the child belonged to. This was *double blinding*: the subjects did not know whether they got the treatment or the placebo, and neither did those who evaluated the responses. This randomized controlled double-blind experiment—which is about the best design there is—was done in many school districts.

How did it all turn out? Table 1 shows the rate of polio cases (per hundred thousand subjects) in the randomized controlled experiment, for the treatment

group and the control group. The rate is much lower for the treatment group, decisive proof of the effectiveness of the Salk vaccine.

Table 1. The results of the Salk vaccine trial of 1954. Size of groups and rate of polio cases per 100,000 in each group. The numbers are rounded.

<i>The randomized controlled double-blind experiment</i>			<i>The NFIP study</i>		
	<i>Size</i>	<i>Rate</i>		<i>Size</i>	<i>Rate</i>
Treatment	200,000	28	Grade 2 (vaccine)	225,000	25
Control	200,000	71	Grades 1 and 3 (control)	725,000	54
No consent	350,000	46	Grade 2 (no consent)	125,000	44

Source: Thomas Francis, Jr., "An evaluation of the 1954 poliomyelitis vaccine trials—summary report," *American Journal of Public Health* vol. 45 (1955) pp. 1–63.

Table 1 also shows how the NFIP study was biased against the vaccine. In the randomized controlled experiment, the vaccine cut the polio rate from 71 to 28 per hundred thousand. The reduction in the NFIP study, from 54 to 25 per hundred thousand, is quite a bit less. The main source of the bias was confounding. The NFIP treatment group included only children whose parents consented to vaccination. However, the control group also included children whose parents would not have consented. The control group was not comparable to the treatment group.

The randomized controlled double-blind design reduces bias to a minimum—the main reason for using it whenever possible. But this design also has an important technical advantage. To see why, let us play devil's advocate and assume that the Salk vaccine had no effect. Then the difference between the polio rates for the treatment and control groups is just due to chance. How likely is that?

With the NFIP design, the results are affected by many factors that seem random: which families volunteer, which children are in grade 2, and so on. However, the investigators do not have enough information to figure the chances for the outcomes. They cannot figure the odds against a big difference in polio rates being due to accidental factors. With a randomized controlled experiment, on the other hand, chance enters in a planned and simple way—when the assignment is made to treatment or control.

The devil's-advocate hypothesis says that the vaccine has no effect. On this hypothesis, a few children are fated to contract polio. Assignment to treatment or control has nothing to do with it. Each child has a 50–50 chance to be in treatment or control, just depending on the toss of a coin. Each polio case has a 50–50 chance to turn up in the treatment group or the control group.

Therefore, the number of polio cases in the two groups must be about the same. Any difference is due to the chance variability in coin tossing. Statisticians understand this kind of variability. They can figure the odds against a difference as large as the observed one. The calculation will be done in chapter 27, and the odds are astronomical—a billion to one against.

## 2. THE PORTACAVAL SHUNT

In some cases of cirrhosis of the liver, the patient may start to hemorrhage and bleed to death. One treatment involves surgery to redirect the flow of blood through a *portacaval shunt*. The operation to create the shunt is long and hazardous. Do the benefits outweigh the risks? Over 50 studies have been done to assess the effect of this surgery.<sup>6</sup> Results are summarized in table 2 below.

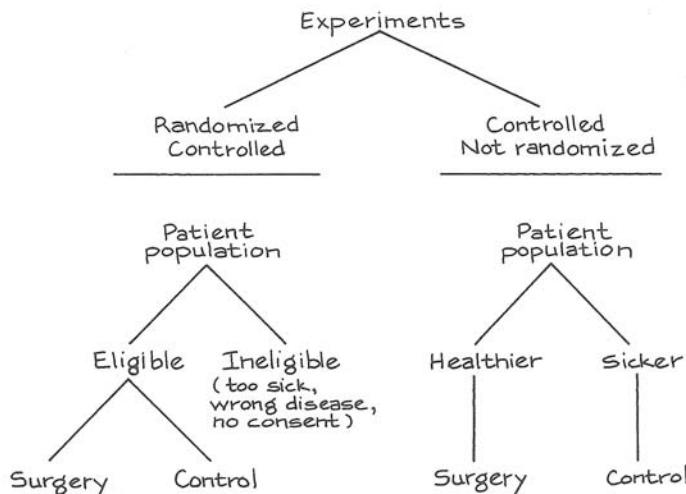
Table 2. A study of 51 studies on the portacaval shunt. The well-designed studies show the surgery to have little or no value. The poorly-designed studies exaggerate the value of the surgery.

Design	Degree of enthusiasm		
	Marked	Moderate	None
No controls	24	7	1
Controls, but not randomized	10	3	2
Randomized controlled	0	1	3

Source: N. D. Grace, H. Muench, and T. C. Chalmers, "The present status of shunts for portal hypertension in cirrhosis," *Gastroenterology* vol. 50 (1966) pp. 684-91.

There were 32 studies without controls (first line in the table): 24/32 of these studies, or 75%, were markedly enthusiastic about the shunt, concluding that the benefits definitely outweighed the risks. In 15 studies there were controls, but assignment to treatment or control was not randomized. Only 10/15, or 67%, were markedly enthusiastic about the shunt. But the 4 studies that were randomized controlled showed the surgery to be of little or no value. The badly designed studies exaggerated the value of this risky surgery.

A randomized controlled experiment begins with a well-defined patient population. Some are eligible for the trial. Others are ineligible: they may be too sick



to undergo the treatment, or they may have the wrong kind of disease, or they may not consent to participate (see the flow chart at the bottom of the previous page). Eligibility is determined first; then the eligible patients are randomized to treatment or control. That way, the comparison is made only among patients who could have received the therapy. The bottom line: the control group is like the treatment group. By contrast, with poorly-controlled studies, ineligible patients may be used as controls. Moreover, even if controls are selected among those eligible for surgery, the surgeon may choose to operate only on the healthier patients while sicker patients are put in the control group.

This sort of bias seems to have been at work in the poorly-controlled studies of the portacaval shunt. In both the well-controlled and the poorly-controlled studies, about 60% of the surgery patients were still alive 3 years after the operation (table 3). In the randomized controlled experiments, the percentage of controls who survived the experiment by 3 years was also about 60%. But only 45% of the controls in the nonrandomized experiments survived for 3 years.

In both types of studies, the surgeons seem to have used similar criteria to select patients eligible for surgery. Indeed, the survival rates for the surgery group are about the same in both kinds of studies. So, what was the crucial difference? With the randomized controlled experiments, the controls were similar in general health to the surgery patients. With the poorly controlled studies, there was a tendency to exclude sicker patients from the surgery group and use them as controls. That explains the bias in favor of surgery.

Table 3. Randomized controlled experiments vs. controlled experiments that are not randomized. Three-year survival rates in studies of the portacaval shunt. (Percentages are rounded.)

	<i>Randomized</i>	<i>Not randomized</i>
Surgery	60%	60%
Controls	60%	45%

### 3. HISTORICAL CONTROLS

Randomized controlled experiments are hard to do. As a result, doctors often use other designs which are not as good. For example, a new treatment can be tried out on one group of patients, who are compared to "historical controls:" patients treated the old way in the past. The problem is that the treatment group and the historical control group may differ in important ways besides the treatment. In a controlled experiment, there is a group of patients eligible for treatment at the beginning of the study. Some of these are assigned to the treatment group, the others are used as controls: assignment to treatment or control is done "contemporaneously," that is, in the same time period. Good studies use contemporaneous controls.

The poorly-controlled trials on the portacaval shunt (section 2) included some with historical controls. Others had contemporaneous controls, but assign-

ment to the control group was not randomized. Section 2 showed that the design of a study matters. This section continues the story. Coronary bypass surgery is a widely used—and very expensive—operation for coronary artery disease. Chalmers and associates identified 29 trials of this surgery (first line of table 4). There were 8 randomized controlled trials, and 7 were quite negative about the value of the operation. By comparison, there were 21 trials with historical controls, and 16 were positive. The badly-designed studies were more enthusiastic about the value of the surgery. (The other lines in the table can be read the same way, and lead to similar conclusions about other therapies.)

Table 4. A study of studies. Four therapies were evaluated both by randomized controlled trials and by trials using historical controls. Conclusions of trials were summarized as positive (+) about the value of the therapy, or negative (-).

<i>Therapy</i>	<i>Randomized controlled</i>		<i>Historically controlled</i>	
	+	-	+	-
Coronary bypass surgery	1	7	16	5
5-FU	0	5	2	0
BCG	2	2	4	0
DES	0	3	5	0

Note: 5-FU is used in chemotherapy for colon cancer; BCG is used to treat melanoma; DES, to prevent miscarriage.

Source: H. Sacks, T. C. Chalmers, and H. Smith, "Randomized versus historical controls for clinical trials," *American Journal of Medicine* vol. 72 (1982) pp. 233-40.<sup>7</sup>

Why are well-designed studies less enthusiastic than poorly-designed studies? In 6 of the randomized controlled experiments on coronary bypass surgery and 9 of the studies with historical controls, 3-year survival rates for the surgery group and the control group were reported (table 5). In the randomized controlled experiments, survival was quite similar in the surgery group and the control group. That is why the investigators were not enthusiastic about the operation—it did not save lives.

Table 5. Randomized controlled experiments vs. studies with historical controls. Three-year survival rates for surgery patients and controls in trials of coronary bypass surgery. Randomized controlled experiments differ from trials with historical controls.

	<i>Randomized</i>	<i>Historical</i>
Surgery	87.6%	90.9%
Controls	83.2%	71.1%

Note: There were 6 randomized controlled experiments enrolling 9,290 patients; and 9 studies with historical controls, enrolling 18,861 patients.

Source: See table 4.

Now look at the studies with historical controls. Survival in the surgery group is about the same as before. However, the controls have much poorer survival



rates. They were not as healthy to start with as the patients chosen for surgery. Trials with historical controls are biased in favor of surgery. Randomized trials avoid that kind of bias. That explains why the design of the study matters. Tables 2 and 3 made the point for the portacaval shunt; tables 4 and 5 make the same point for other therapies.

The last line in table 4 is worth more discussion. DES (diethylstibestrol) is an artificial hormone, used to prevent spontaneous abortion. Chalmers and associates found 8 trials evaluating DES. Three were randomized controlled, and all were negative: the drug did not help. There were 5 studies with historical controls, and all were positive. These poorly-designed studies were biased in favor of the therapy.

Doctors paid little attention to the randomized controlled experiments. Even in the late 1960s, they were giving the drug to 50,000 women each year. This was a medical tragedy, as later studies showed. If administered to the mother during pregnancy, DES can have a disastrous side-effect 20 years later, causing her daughter to develop an otherwise extremely rare form of cancer (clear-cell adenocarcinoma of the vagina). DES was banned for use on pregnant women in 1971.<sup>8</sup>

#### 4. SUMMARY

1. Statisticians use the *method of comparison*. They want to know the effect of a *treatment* (like the Salk vaccine) on a *response* (like getting polio). To find

out, they compare the responses of a *treatment group* with a *control group*. Usually, it is hard to judge the effect of a treatment without comparing it to something else.

2. If the control group is comparable to the treatment group, apart from the treatment, then a difference in the responses of the two groups is likely to be due to the effect of the treatment.

3. However, if the treatment group is different from the control group with respect to other factors, the effects of these other factors are likely to be *confounded* with the effect of the treatment.

4. To make sure that the treatment group is like the control group, investigators put subjects into treatment or control at random. This is done in *randomized controlled experiments*.

5. Whenever possible, the control group is given a *placebo*, which is neutral but resembles the treatment. The response should be to the treatment itself rather than to the idea of treatment.

6. In a *double-blind* experiment, the subjects do not know whether they are in treatment or in control; neither do those who evaluate the responses. This guards against bias, either in the responses or in the evaluations.

# 2

## Observational Studies

*That's not an experiment you have there, that's an experience.*

—SIR R. A. FISHER (ENGLAND, 1890–1962)

### 1. INTRODUCTION

Controlled experiments are different from *observational studies*. In a controlled experiment, the investigators decide who will be in the treatment group and who will be in the control group. By contrast, in an observational study it is the subjects who assign themselves to the different groups: the investigators just watch what happens.

The jargon is a little confusing, because the word *control* has two senses.

- A *control* is a subject who did not get the treatment.
- A *controlled experiment* is a study where the investigators decide who will be in the treatment group and who will not.

Studies on the effects of smoking, for instance, are necessarily observational: nobody is going to smoke for ten years just to please a statistician. However, the treatment-control idea is still used. The investigators compare smokers (the treatment or “exposed” group) with non-smokers (the control group) to determine the effect of smoking.

The smokers come off badly in this comparison. Heart attacks, lung cancer, and many other diseases are more common among smokers than non-smokers. So there is a strong *association* between smoking and disease. If cigarettes cause

disease, that explains the association: death rates are higher for smokers because cigarettes kill. Thus, association is circumstantial evidence for causation. However, the proof is incomplete. There may be some hidden confounding factor which makes people smoke and also makes them get sick. If so, there is no point in quitting; that will not change the hidden factor. Association is not the same as causation.

Statisticians like Joseph Berkson and Sir R. A. Fisher did not believe the evidence against cigarettes, and suggested possible confounding variables. Epidemiologists (including Sir Richard Doll in England, and E. C. Hammond, D. Horn, H. A. Kahn in the United States) ran careful observational studies to show these alternative explanations were not plausible. Taken together, the studies make a powerful case that smoking causes heart attacks, lung cancer, and other diseases. If you give up smoking, you will live longer.<sup>1</sup>

Observational studies are a powerful tool, as the smoking example shows. But they can also be quite misleading. To see if confounding is a problem, it may help to find out how the controls were selected. The main issue: was the control group really similar to the treatment group—apart from the exposure of interest? If there is confounding, something has to be done about it, although perfection cannot be expected. Statisticians talk about *controlling for* confounding factors in an observational study. This is a third use of the word *control*.

One technique is to make comparisons separately for smaller and more homogeneous groups. For example, a crude comparison of death rates among smokers and non-smokers could be misleading, because smokers are disproportionately male and men are more likely than women to have heart disease anyway. The difference between smokers and non-smokers might be due to the sex difference. To eliminate that possibility, epidemiologists compare male smokers to male non-smokers, and females to females.

Age is another confounding variable. Older people have different smoking habits, and are more at risk for lung cancer. So the comparison between smokers and non-smokers is done separately by age as well as by sex. For example, male smokers age 55–59 are compared to male non-smokers age 55–59. This controls for age and sex. Good observational studies control for confounding variables. In the end, however, most observational studies are less successful than the ones on smoking. The studies may be designed by experts, but experts make mistakes too. Finding the weak points is more an art than a science, and often depends on information outside the study.

## 2. THE CLOFIBRATE TRIAL

The Coronary Drug Project was a randomized, controlled double-blind experiment, whose objective was to evaluate five drugs for the prevention of heart attacks. The subjects were middle-aged men with heart trouble. Of the 8,341 subjects, 5,552 were assigned at random to the drug groups and 2,789 to the control group. The drugs and the placebo (lactose) were administered in identical capsules. The patients were followed for 5 years.

One of the drugs on test was clofibrate, which reduces the levels of cholesterol in the blood. Unfortunately, this treatment did not save any lives. About 20% of the clofibrate group died over the period of followup, compared to 21% of the control group. A possible reason for this failure was suggested—many subjects in the clofibrate group did not take their medicine.

Subjects who took more than 80% of their prescribed medicine (or placebo) were called “adherers” to the protocol. For the clofibrate group, the 5-year mortality rate among the adherers was only 15%, compared to 25% among the non-adherers (table 1). This looks like strong evidence for the effectiveness of the drug. However, caution is in order. This particular comparison is observational not experimental—even though the data were collected while an experiment was going on. After all, the investigators did not decide who would adhere to protocol and who would not. The subjects decided.

Table 1. The clofibrate trial. Numbers of subjects, and percentages who died during 5 years of followup. Adherers take 80% or more of prescription.

	<i>Clofibrate</i>		<i>Placebo</i>	
	<i>Number</i>	<i>Deaths</i>	<i>Number</i>	<i>Deaths</i>
Adherers	708	15%	1,813	15%
Non-adherers	357	25%	882	28%
Total group	1,103	20%	2,789	21%

Note: Data on adherence missing for 38 subjects in the clofibrate group and 94 in the placebo group.  
Deaths from all causes.

Source: The Coronary Drug Project Research Group, “Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project,” *New England Journal of Medicine* vol. 303 (1980) pp. 1038–41.

Maybe adherers were different from non-adherers in other ways, besides the amount of the drug they took. To find out, the investigators compared adherers and non-adherers in the control group. Remember, the experiment was double-blind. The controls did not know whether they were taking an active drug or the placebo; neither did the subjects in the clofibrate group. The psychological basis for adherence was the same in both groups.

In the control group too, the adherers did better. Only 15% of them died during the 5-year period, compared to 28% among the non-adherers. The conclusions:

- (i) Clofibrate does not have an effect.
- (ii) Adherers are different from non-adherers.

Probably, adherers are more concerned with their health and take better care of themselves in general. That would explain why they took their capsules and why they lived longer. Observational comparisons can be quite misleading. The investigators in the clofibrate trial were unusually careful, and they found out what was wrong with comparing adherers to non-adherers.<sup>2</sup>



"TO ADHERE OR NOT TO ADHERE,  
THAT IS THE QUESTION."

### 3. MORE EXAMPLES

*Example 1.* "Pellagra was first observed in Europe in the eighteenth century by a Spanish physician, Gaspar Casal, who found that it was an important cause of ill-health, disability, and premature death among the very poor inhabitants of the Asturias. In the ensuing years, numerous . . . authors described the same condition in northern Italian peasants, particularly those from the plain of Lombardy. By the beginning of the nineteenth century, pellagra had spread across Europe, like a belt, causing the progressive physical and mental deterioration of thousands of people in southwestern France, in Austria, in Rumania, and in the domains of the Turkish Empire. Outside Europe, pellagra was recognized in Egypt and South Africa, and by the first decade of the twentieth century it was rampant in the United States, especially in the south . . ."<sup>3</sup>

Pellagra seemed to hit some villages much more than others. Even within affected villages, many households were spared; but some had pellagra cases year after year. Sanitary conditions in diseased households were primitive; flies were everywhere. One blood-sucking fly (*Simulium*) had the same geographical range as pellagra, at least in Europe; and the fly was most active in the spring, just when most pellagra cases developed. Many epidemiologists concluded the disease was infectious, and—like malaria, yellow fever, or typhus—was transmitted from one person to another by insects. Was this conclusion justified?

*Discussion.* Starting around 1914, the American epidemiologist Joseph Goldberger showed by a series of observational studies and experiments that pellagra is caused by a bad diet, and is not infectious. The disease can be prevented or cured by foods rich in what Goldberger called the P-P (pellagra-preventive) factor. Since 1940, most of the flour sold in the United States is enriched with the P-P factor, among other vitamins; the P-P factor is called “niacin” on the label.

Niacin occurs naturally in meat, milk, eggs, some vegetables, and certain grains. Corn, however, contains relatively little niacin. In the pellagra areas, the poor ate corn—and not much else. Some villages and some households were poorer than others, and had even more restricted diets. That is why they were harder hit by the disease. The flies were a marker of poverty, not a cause of pellagra. Association is not the same as causation.

*Example 2. Cervical cancer and circumcision.* For many years, cervical cancer was one of the most common cancers among women. Many epidemiologists worked on identifying the causes of this disease. They found that in several different countries, cervical cancer was quite rare among Jews. They also found the disease to be very unusual among Moslems. In the 1950s, several investigators wrote papers concluding that circumcision of the males was the protective factor. Was this conclusion justified?

*Discussion.* There are differences between Jews or Moslems and members of other communities, besides circumcision. It turns out that cervical cancer is a sexually transmitted disease, spread by contact. Current research suggests that certain strains of HPV (human papilloma virus) are the causal agents. Some women are more active sexually than others, and have more partners; they are more likely to be exposed to the viruses causing the disease. That seems to be what makes the rate of cervical cancer higher for some groups of women. Early studies did not pay attention to this confounding variable, and reached the wrong conclusions.<sup>4</sup> (Cancer takes a long time to develop; sexual behavior in the 1930s or 1940s was the issue.)

*Example 3. Ultrasound and low birthweight.* Human babies can now be examined in the womb using ultrasound. Several experiments on lab animals have shown that ultrasound examinations can cause low birthweight. If this is true for humans, there are grounds for concern. Investigators ran an observational study to find out, at the Johns Hopkins hospital in Baltimore.

Of course, babies exposed to ultrasound differed from unexposed babies in many ways besides exposure; this was an observational study. The investigators found a number of confounding variables and adjusted for them. Even so, there was an association. Babies exposed to ultrasound in the womb had lower birthweight, on average, than babies who were not exposed. Is this evidence that ultrasound causes lower birthweight?

*Discussion.* Obstetricians suggest ultrasound examinations when something seems to be wrong. The investigators concluded that the ultrasound exams and low birthweights had a common cause—problem pregnancies. Later, a randomized controlled experiment was done to get more definite evidence. If anything, ultrasound was protective.<sup>5</sup>

*Example 4. The Samaritans and suicide.* Over the period 1964–70, the suicide rate in England fell by about one-third. During this period, a volunteer welfare organization called “The Samaritans” was expanding rapidly. One investigator thought that the Samaritans were responsible for the decline in suicides. He did an observational study to prove it. This study was based on 15 pairs of towns. To control for confounding, the towns in a pair were matched on the variables regarded as important. One town in each pair had a branch of the Samaritans; the other did not. On the whole, the towns with the Samaritans had lower suicide rates. So the Samaritans prevented suicides. Or did they?

*Discussion.* A second investigator replicated the study, with a bigger sample and more careful matching. He found no effect. Furthermore, the suicide rate was stable in the 1970s (after the first investigator had published his paper) although the Samaritans continued to expand. The decline in suicide rates in the 1960s is better explained by a shift from coal gas to natural gas for heating and cooking. Natural gas is less toxic. In fact, about one-third of suicides in the early 1960s were by gas. At the end of the decade, there were practically no such cases, explaining the decline in suicides. The switch to natural gas was complete, so the suicide rate by gas couldn’t decline much further. Finally, the suicide rate by methods other than gas was nearly constant over the 1960s—despite the Samaritans. The Samaritans were a good organization, but they do not seem to have had much effect on the suicide rate. And observational studies, no matter how carefully done, are not experiments.<sup>6</sup>

#### 4. SEX BIAS IN GRADUATE ADMISSIONS

To review briefly, one source of trouble in observational studies is that subjects differ among themselves in crucial ways besides the treatment. Sometimes these differences can be adjusted for, by comparing smaller and more homogeneous subgroups. Statisticians call this technique *controlling for* the confounding factor—the third sense of the word *control*.

An observational study on sex bias in admissions was done by the Graduate Division at the University of California, Berkeley.<sup>7</sup> During the study period, there were 8,442 men who applied for admission to graduate school and 4,321 women. About 44% of the men and 35% of the women were admitted. Taking percents adjusts for the difference in numbers of male and female applicants: 44 out of every 100 men were admitted, and 35 out of every 100 women.

Assuming that the men and women were on the whole equally well qualified (and there is no evidence to the contrary), the difference in admission rates looks like a strong piece of evidence to show that men and women are treated differently in the admissions procedure. The university seems to prefer men, 44 to 35.

Each major did its own admissions to graduate work. By looking at them separately, the university should have been able to identify the ones which discriminated against the women. At that point, a puzzle appeared. Major by major, there did not seem to be any bias against women. Some majors favored men, but others favored women. On the whole, if there was any bias, it ran against the men. What was going on?



"YES, ON THE SURFACE IT WOULD APPEAR TO BE SEX-BIAS  
BUT LET US ASK THE FOLLOWING QUESTIONS..."

Over a hundred majors were involved. However, the six largest majors together accounted for over one-third of the total number of applicants to the campus. And the pattern for these majors was typical of the whole campus. Table 2 shows the number of male and female applicants, and the percentage admitted, for each of these majors.

Table 2. Admissions data for the graduate programs in the six largest majors at University of California, Berkeley.

<i>Major</i>	<i>Men</i>		<i>Women</i>	
	<i>Number of applicants</i>	<i>Percent admitted</i>	<i>Number of applicants</i>	<i>Percent admitted</i>
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Note: University policy does not allow these majors to be identified by name.  
Source: The Graduate Division, University of California, Berkeley.

In each major, the percentage of female applicants who were admitted is roughly equal to the percentage for male applicants. The only exception is major A, which appears to discriminate against men. It admitted 82% of the women but only 62% of the men. The department that looks most biased against women is E. It admitted 28% of the men and 24% of the women. This difference only amounts to 4 percentage points. However, when all six majors are taken together, they admitted 44% of the male applicants, and only 30% of the females. The difference is 14 percentage points.

This seems paradoxical, but here is the explanation.

- The first two majors were easy to get into. Over 50% of the men applied to these two majors.
- The other four majors were much harder to get into. Over 90% of the women applied to these four majors.

The men were applying to the easy majors, the women to the harder ones. There was an effect due to the choice of major, confounded with the effect due to sex. When the choice of major is controlled for, as in table 2, there is little difference in the admissions rates for men or women. The statistical lesson: relationships between percentages in subgroups (for instance, admissions rates for men and women in each department separately) can be reversed when the subgroups are combined. This is called *Simpson's paradox*.<sup>8</sup>

*Technical note.* Table 2 is hard to read because it compares twelve admissions rates. A statistician might summarize table 2 by computing one overall admissions rate for men and another for women, but adjusting for the sex difference in application rates. The procedure would be to take some kind of average admission rate separately for the men and women. An ordinary average ignores the differences in size among the departments. Instead, a *weighted average* of the admission rates could be used, the weights being the total number of applicants (male and female) to each department; see table 3.

Table 3. Total number of applicants, from table 2.

Major	Total number of applicants
A	933
B	585
C	918
D	792
E	584
F	714
	4,526

The weighted average admission rate for men is

$$\frac{.62 \times 933 + .63 \times 585 + .37 \times 918 + .33 \times 792 + .28 \times 584 + .06 \times 714}{4,526}$$

This works out to 39%. Similarly, the weighted average admission rate for the women is

$$\frac{.82 \times 933 + .68 \times 585 + .34 \times 918 + .35 \times 792 + .24 \times 584 + .07 \times 714}{4,526}$$

This works out to 43%. In these formulas, the weights are the same for the men and women; they are the totals from table 3. The admission rates are different for men and women; they are the rates from table 2. The final comparison: the weighted average admission rate for men is 39%, while the weighted average admission rate for women is 43%. The weighted averages control for the confounding factor—choice of major. These averages suggest that if anything, the admissions process is biased against the men.

## 5. CONFOUNDING

Hidden confounders are a major problem in observational studies. As discussed in section 1, epidemiologists found an association between exposure (smoking) and disease (lung cancer): heavy smokers get lung cancer at higher rates than light smokers; light smokers get the disease at higher rates than non-smokers. According to the epidemiologists, the association comes about because smoking causes lung cancer. However, some statisticians—including Sir R. A. Fisher—thought the association could be explained by confounding.

Confounders have to be associated with (i) the disease and (ii) the exposure. For example, suppose there is a gene which increases the risk of lung cancer. Now, if the gene also gets people to smoke, it meets both the tests for a confounder. This gene would create an association between smoking and lung cancer. The idea is a bit subtle: a gene that causes cancer but is unrelated to smoking is not a confounder and is sideways to the argument, because it does not account for the facts—the association between smoking and cancer.<sup>9</sup> Fisher's “constitutional hypothesis” explained the association on the basis of genetic confounding; nowadays, there is evidence from twin studies to refute this hypothesis (review exercise 11, chapter 15).

Confounding means a difference between the treatment and control groups—other than the treatment—which affects the responses being studied. A confounder is a third variable, associated with exposure and with disease.

### Exercise Set A

1. In the U.S. in 2000, there were 2.4 million deaths from all causes, compared to 1.9 million in 1970—a 25% increase.<sup>10</sup> True or false, and explain: the data show that the public's health got worse over the period 1970–2000.

2. Data from the Salk vaccine field trial suggest that in 1954, the school districts in the NFIP trial and in the randomized controlled experiment had similar exposures to the polio virus.
  - (a) The data also show that children in the two vaccine groups (for the randomized controlled experiment and the NFIP design) came from families with similar incomes and educational backgrounds. Which two numbers in table 1 (p. 6) confirm this finding?
  - (b) The data show that children in the two no-consent groups had similar family backgrounds. Which pair of numbers in the table confirm this finding?
  - (c) The data show that children in the two control groups had different family backgrounds. Which pair of numbers in the table confirm this finding?
  - (d) In the NFIP study, neither the control group nor the no-consent group got the vaccine. Yet the no-consent group had a lower rate of polio. Why?
  - (e) To show that the vaccine works, someone wants to compare the 44/100,000 in the NFIP study with the 25/100,000 in the vaccine group. What's wrong with this idea?
3. Polio is an infectious disease; for example, it seemed to spread when children went swimming together. The NFIP study was not done blind: could that bias the results? Discuss briefly.
4. The Salk vaccine field trials were conducted only in certain experimental areas (school districts), selected by the Public Health Service in consultation with local officials.<sup>11</sup> In these areas, there were about 3 million children in grades 1, 2, or 3; and there were about 11 million children in those grades in the United States. In the experimental areas, the incidence of polio was about 25% higher than in the rest of the country. Did the Salk vaccine field trials cause children to get polio instead of preventing it? Answer yes or no, and explain briefly.
5. Linus Pauling thought that vitamin C prevents colds, and cures them too. Thomas Chalmers and associates did a randomized controlled double-blind experiment to find out.<sup>12</sup> The subjects were 311 volunteers at the National Institutes of Health. These subjects were assigned at random to 1 of 4 groups:

<i>Group</i>	<i>Prevention</i>	<i>Therapy</i>
1	placebo	placebo
2	vitamin C	placebo
3	placebo	vitamin C
4	vitamin C	vitamin C

All subjects were given six capsules a day for prevention, and an additional six capsules a day for therapy if they came down with a cold. However, in group 1 both sets of capsules just contained the placebo (lactose). In group 2, the prevention capsules had vitamin C while the therapy capsules were filled with the placebo. Group 3 was the reverse. In group 4, all the capsules were filled with vitamin C.

There was quite a high dropout rate during the trial. And this rate was significantly higher in the first 3 groups than in the 4th. The investigators noticed this, and found the reason. As it turned out, many of the subjects broke the blind. (That

is quite easy to do; you just open a capsule and taste the contents; vitamin C—ascorbic acid—is sour, lactose is not.) Subjects who were getting the placebo were more likely to drop out.

The investigators analyzed the data for the subjects who remained blinded, and vitamin C had no effect. Among those who broke the blind, groups 2 and 4 had the fewest colds; groups 3 and 4 had the shortest colds. How do you interpret these results?

6. (Hypothetical.) One of the other drugs in the Coronary Drug Project (section 2) was nicotinic acid.<sup>13</sup> Suppose the results on nicotinic acid were as reported below. Something looks wrong. What, and why?

	<i>Nicotinic acid</i>		<i>Placebo</i>	
	<i>Number</i>	<i>Deaths</i>	<i>Number</i>	<i>Deaths</i>
Adherers	558	13%	1,813	15%
Non-adherers	487	26%	882	28%
Total group	1,045	19%	2,695	19%

7. (Hypothetical.) In a clinical trial, data collection usually starts at “baseline,” when the subjects are recruited into the trial but before they are assigned to treatment or control. Data collection continues until the end of followup. Two clinical trials on prevention of heart attacks report baseline data on smoking, shown below. In one of these trials, the randomization did not work. Which one, and why?

	<i>Number of persons</i>	<i>Percent who smoked</i>
(i) {Treatment Control	1,012	49.3%
	997	69.0%
(ii) {Treatment Control	995	59.3%
	1,017	59.0%

8. Some studies find an association between liver cancer and smoking. However, alcohol consumption is a confounding variable. This means—  
 (i) Alcohol causes liver cancer.  
 (ii) Drinking is associated with smoking, and alcohol causes liver cancer.

Choose one option, and explain briefly.

9. Breast cancer is one of the most common malignancies among women in the U.S. If it is detected early enough—before the cancer spreads—chances of successful treatment are much better. Do screening programs speed up detection by enough to matter?

The first large-scale trial was run by the Health Insurance Plan of Greater New York, starting in 1963. The subjects (all members of the plan) were 62,000 women age 40 to 64. These women were divided at random into two equal groups. In the treatment group, women were encouraged to come in for annual screening, including examination by a doctor and X-rays. About 20,200 women in the treatment group did come in for the screening; but 10,800 refused. The control group was offered usual health care. All the women were followed for many years.

Results for the first 5 years are shown in the table below.<sup>14</sup> ("HIP" is the usual abbreviation for the Health Insurance Plan.)

*Deaths in the first five years of the HIP screening trial, by cause. Rates per 1,000 women.*

	<i>Cause of Death</i>				
	<i>Breast cancer</i>		<i>All other</i>		
	<i>Number</i>	<i>Rate</i>	<i>Number</i>	<i>Rate</i>	
Treatment group					
Examined	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control group	31,000	63	2.0	879	28

Epidemiologists who worked on the study found that (i) screening had little impact on diseases other than breast cancer; (ii) poorer women were less likely to accept screening than richer ones; and (iii) most diseases fall more heavily on the poor than the rich.

- (a) Does screening save lives? Which numbers in the table prove your point?
  - (b) Why is the death rate from all other causes in the whole treatment group ("examined" and "refused" combined) about the same as the rate in the control group?
  - (c) Breast cancer (like polio, but unlike most other diseases) affects the rich more than the poor. Which numbers in the table confirm this association between breast cancer and income?
  - (d) The death rate (from all causes) among women who accepted screening is about half the death rate among women who refused. Did screening cut the death rate in half? If not, what explains the difference in death rates?
10. (This continues exercise 9.)
- (a) To show that screening reduces the risk from breast cancer, someone wants to compare 1.1 and 1.5. Is this a good comparison? Is it biased against screening? For screening?
  - (b) Someone claims that encouraging women to come in for breast cancer screening increases their health consciousness, so these women take better care of themselves and live longer for that reason. Is the table consistent or inconsistent with the claim?
  - (c) In the first year of the HIP trial, 67 breast cancers were detected in the "examined" group, 12 in the "refused" group, and 58 in the control group. True or false, and explain briefly: screening causes breast cancer.
11. Cervical cancer is more common among women who have been exposed to the herpes virus, according to many observational studies.<sup>15</sup> Is it fair to conclude that the virus causes cervical cancer?
12. Physical exercise is considered to increase the risk of spontaneous abortion. Furthermore, women who have had a spontaneous abortion are more likely to have another. One observational study finds that women who exercise regularly have fewer spontaneous abortions than other women.<sup>16</sup> Can you explain the findings of this study?

13. A hypothetical university has two departments, A and B. There are 2,000 male applicants, of whom half apply to each department. There are 1,100 female applicants: 100 apply to department A and 1,000 to department B. Department A admits 60% of the men who apply and 60% of the women. Department B admits 30% of the men who apply and 30% of the women. “For each department, the percentage of men admitted equals the percentage of women admitted; this must be so for both departments together.” True or false, and explain briefly.

*Exercises 14 and 15 are designed as warm-ups for the next chapter. Do not use a calculator when working them. Just remember that “%” means “per hundred.” For example, 41 people out of 398 is just about 10%. The reason: 41 out of 398 is like 40 out of 400, that’s 10 out of 100, and that’s 10%.*

14. Say whether each of the following is about 1%, 10%, 25%, or 50%—
- (a) 39 out of 398
  - (b) 99 out of 407
  - (c) 57 out of 209
  - (d) 99 out of 197
15. Among beginning statistics students in one university, 46 students out of 446 reported family incomes ranging from \$40,000 to \$50,000 a year.
- (a) About what percentage had family incomes in the range \$40,000 to \$50,000 a year?
  - (b) Guess the percentage that had family incomes in the range \$45,000 to \$46,000 a year.
  - (c) Guess the percentage that had family incomes in the range \$46,000 to \$47,000 a year.
  - (d) Guess the percentage that had family incomes in the range \$47,000 to \$49,000 a year.

*The answers to these exercises are on pp. A43–45.*

## 6. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. The Federal Bureau of Investigation reports state-level and national data on crimes.<sup>17</sup>
  - (a) An investigator compares the incidence of crime in Minnesota and in Michigan. In 2001, there were 3,584 crimes in Minnesota, compared to 4,082 in Michigan. He concludes that Minnesotans are more law-abiding. After all, Michigan includes the big bad city of Detroit. What do you say?
  - (b) An investigator compares the incidence of crime in the U.S. in 1991 and 2001. In 1991, there were 28,000 crimes, compared to 22,000 in 2001. She concludes that the U.S. became more law-abiding over that time period. What do you say?
2. The National Highway and Traffic Safety Administration analyzed thefts of new cars in 2002, as well as sales figures for that year.<sup>18</sup>
  - (a) There were 99 Corvettes stolen, and 26 Infiniti Q45 sedans. Should you conclude that American thieves prefer American cars? Or is something missing from the equation?

- (b) There were 50 BMW 7-series cars stolen, compared to 146 in the 3-series. Should you conclude that thieves prefer smaller cars, which are more economical to run and easier to park? Or is something missing from the equation?
- (c) There were 429 Liberty Jeeps stolen, compared to 207,991 sold, for a rate of 2 per 100,000. True or false and explain: the rate is low because the denominator is large.
3. From table 1 in chapter 1 (p. 6), those children whose parents refused to participate in the randomized controlled Salk trial got polio at the rate of 46 per 100,000. On the other hand, those children whose parents consented to participation got polio at the slightly higher rate of 49 per 100,000 in the treatment group and control group taken together. Suppose that this field trial was repeated the following year. On the basis of the figures, some parents refused to allow their children to participate in the experiment and be exposed to this higher risk of polio. Were they right? Answer yes or no, and explain briefly.
4. The Public Health Service studied the effects of smoking on health, in a large sample of representative households.<sup>19</sup> For men and for women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking.
- (a) Why did they study men and women and the different age groups separately?
- (b) The lesson seems to be that you shouldn't start smoking, but once you've started, don't stop. Comment briefly.
5. There is a rare neurological disease (idiopathic hypoguesia) that makes food taste bad. It is sometimes treated with zinc sulfate. One group of investigators did two randomized controlled experiments to test this treatment. In the first trial, the subjects did not know whether they were being given the zinc sulfate or a placebo. However, the doctors doing the evaluations did know. In this trial, patients on zinc sulfate improved significantly; the placebo group showed little improvement. The second trial was run double-blind: neither the subjects nor the doctors doing the evaluation were told who had been given the drug or the placebo. In the second trial, zinc sulfate had no effect.<sup>20</sup> Should zinc sulfate be given to treat the disease? Answer yes or no, and explain briefly.
6. (Continues the previous exercise.) The second trial used what is called a "crossover" design. The subjects were assigned at random to one of four groups:

placebo	placebo
placebo	zinc
zinc	placebo
zinc	zinc

In the first group, the subjects stayed on the placebo through the whole experiment. In the second group, subjects began with the placebo, but halfway

through the experiment they were switched to zinc sulfate. Similarly, in the third group, subjects began on zinc sulfate but were switched to placebo. In the last group, they stayed on zinc sulfate. Subjects knew the design of the study, but were not told the group to which they were assigned.

Some subjects did not improve during the first half of the experiment. In each of the four groups, these subjects showed some improvement (on average) during the second half of the experiment. How can this be explained?

7. According to a study done at Kaiser Permanente in Walnut Creek, California, users of oral contraceptives have a higher rate of cervical cancer than non-users, even after adjusting for age, education, and marital status. Investigators concluded that the pill causes cervical cancer.<sup>21</sup>
  - (a) Is this a controlled experiment or an observational study?
  - (b) Why did the investigators adjust for age? education? marital status?
  - (c) Women using the pill were likely to differ from non-users on another factor which affects the risk of cervical cancer. What factor is that?
  - (d) Were the conclusions of the study justified by the data? Answer yes or no, and explain briefly.
8. Ads for ADT Security Systems claim<sup>22</sup>

When you go on vacation, burglars go to work.... According to FBI statistics, over 25% of home burglaries occur between Memorial Day and Labor Day.

Do the statistics prove that burglars go to work when other people go on vacation? Answer yes or no, and explain briefly.

9. People who get lots of vitamins by eating five or more servings of fresh fruit and vegetables each day (especially “cruciferous” vegetables like broccoli) have much lower death rates from colon cancer and lung cancer, according to many observational studies. These studies were so encouraging that two randomized controlled experiments were done. The treatment groups were given large doses of vitamin supplements, while people in the control groups just ate their usual diet. One experiment looked at colon cancer; the other, at lung cancer.

The first experiment found no difference in the death rate from colon cancer between the treatment group and the control group. The second experiment found that beta carotene (as a diet supplement) increased the death rate from lung cancer.<sup>23</sup> True or false, and explain:

- (a) The experiments confirmed the results of the observational studies.
- (b) The observational studies could easily have reached the wrong conclusions, due to confounding—people who eat lots of fruit and vegetables have lifestyles that are different in many other ways too.
- (c) The experiments could easily have reached the wrong conclusions, due to confounding—people who eat lots of fruit and vegetables have lifestyles that are different in many other ways too.

10. A study of young children found that those with more body fat tended to have more “controlling” mothers; the *San Francisco Chronicle* concluded that “Parents of Fat Kids Should Lighten Up.”<sup>24</sup>

- (a) Was this an observational study or a randomized controlled experiment?
- (b) Did the study find an association between mother’s behavior and her child’s level of body fat?
- (c) If controlling behavior by the mother causes children to eat more, would that explain an association between controlling behavior by the mother and her child’s level of body fat?
- (d) Suppose there is a gene which causes obesity. Would that explain the association?
- (e) Can you think of another way to explain the association?
- (f) Do the data support the *Chronicle’s* advice on child-rearing?

Discuss briefly.

11. California is evaluating a new program to rehabilitate prisoners before their release; the object is to reduce the recidivism rate—the percentage who will be back in prison within two years of release. The program involves several months of “boot camp”—military-style basic training with very strict discipline. Admission to the program is voluntary. According to a prison spokesman, “Those who complete boot camp are less likely to return to prison than other inmates.”<sup>25</sup>

- (a) What is the treatment group in the prison spokesman’s comparison? the control group?
- (b) Is the prison spokesman’s comparison based on an observational study or a randomized controlled experiment?
- (c) True or false: the data show that boot camp worked.

Explain your answers.

12. (Hypothetical.) A study is carried out to determine the effect of party affiliation on voting behavior in a certain city. The city is divided up into wards. In each ward, the percentage of registered Democrats who vote is higher than the percentage of registered Republicans who vote. True or false: for the city as a whole, the percentage of registered Democrats who vote must be higher than the percentage of registered Republicans who vote. If true, why? If false, give an example.

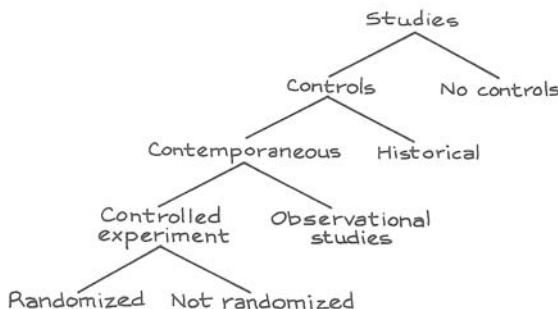
## 7. SUMMARY AND OVERVIEW

1. In an *observational study*, the investigators do not assign the subjects to treatment or control. Some of the subjects have the condition whose effects are being studied; this is the treatment group. The other subjects are the controls. For example, in a study on smoking, the smokers form the treatment group and the non-smokers are the controls.

2. Observational studies can establish *association*: one thing is linked to another. Association may point to causation: if exposure causes disease, then people who are exposed should be sicker than similar people who are not exposed. But association does not prove causation.

3. In an observational study, the effects of treatment may be confounded with the effects of factors that got the subjects into treatment or control in the first place. Observational studies can be quite misleading about cause-and-effect relationships, because of confounding. A *confounder* is a third variable, associated with exposure and with disease.

4. When looking at a study, ask the following questions. Was there any control group at all? Were historical controls used, or contemporaneous controls? How were subjects assigned to treatment—through a process under the control of the investigator (a controlled experiment), or a process outside the control of the investigator (an observational study)? If a controlled experiment, was the assignment made using a chance mechanism (randomized controlled), or did assignment depend on the judgment of the investigator?



5. With observational studies, and with nonrandomized controlled experiments, try to find out how the subjects came to be in treatment or in control. Are the groups comparable? different? What factors are confounded with treatment? What adjustments were made to take care of confounding? Were they sensible?

6. In an observational study, a confounding factor can sometimes be *controlled for*, by comparing smaller groups which are relatively homogeneous with respect to the factor.

7. Study design is a central issue in applied statistics. Chapter 1 introduced the idea of randomized experiments, and chapter 2 draws the contrast with observational studies. The great weakness of observational studies is confounding; randomized experiments minimize this problem. Statistical inference from randomized experiments will be discussed in chapter 27.

## PART II

# Descriptive Statistics

— — — — —

# 3

## The Histogram

*Grown-ups love figures. When you tell them that you have made a new friend, they never ask you any questions about essential matters. They never say to you, “What does his voice sound like? What games does he love best? Does he collect butterflies?” Instead, they demand: “How old is he? How many brothers has he? How much does he weigh? How much money does his father make?” Only from these figures do they think they have learned anything about him.*

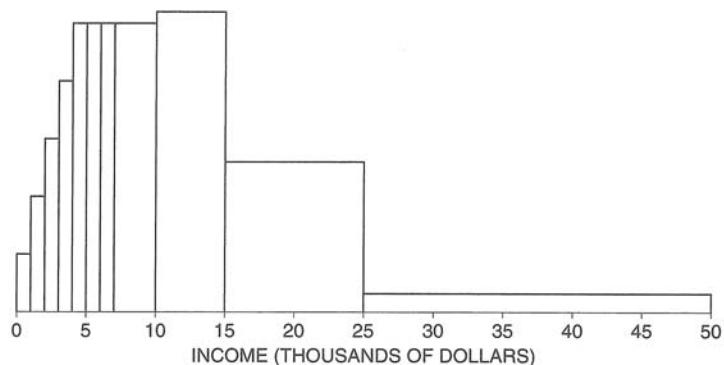
—*The Little Prince*<sup>1</sup>

### 1. INTRODUCTION

In the U.S., how are incomes distributed? How much worse off are minority groups? Some information is provided by government statistics, obtained from the Current Population Survey. Each month, interviewers talk to a representative cross section of about 50,000 American families (for details, see part VI). In March, these families are asked to report their incomes for the previous year. We are going to look at the results for 1973. These data have to be summarized—nobody wants to look at 50,000 numbers. To summarize data, statisticians often use a graph called a *histogram* (figure 1 on the next page).

This section explains how to read histograms. First of all, there is no vertical scale: unlike most other graphs, a histogram does not need a vertical scale. Now look at the horizontal scale. This shows income in thousands of dollars. The graph itself is just a set of blocks. The bottom edge of the first block covers the range from \$0 to \$1,000, the bottom edge of the second goes from \$1,000 to \$2,000;

Figure 1. A histogram. This graph shows the distribution of families by income in the U.S. in 1973.



Source: Current Population Survey.<sup>2</sup>

and so on until the last block, which covers the range from \$25,000 to \$50,000. These ranges are called *class intervals*. The graph is drawn so the area of a block is proportional to the number of families with incomes in the corresponding class interval.

To see how the blocks work, look more closely at figure 1. About what percentage of the families earned between \$10,000 and \$15,000? The block over this interval amounts to something like one-fourth of the total area. So about one-fourth, or 25%, of the families had incomes in that range.

Take another example. Were there more families with incomes between \$10,000 and \$15,000, or with incomes between \$15,000 and \$25,000? The block over the first interval is taller, but the block over the second interval is wider. The areas of the two blocks are about the same, so the percentage of families earning \$10,000 to \$15,000 is about the same as the percentage earning \$15,000 to \$25,000.

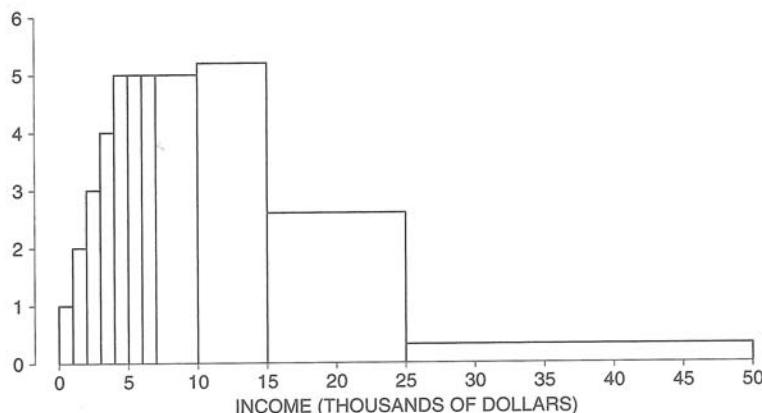
For a last example, take the percentage of families with incomes under \$7,000. Is this closest to 10%, 25%, or 50%? By eye, the area under the histogram between \$0 and \$7,000 is about a quarter of the total area, so the percentage is closest to 25%.

In a histogram, the areas of the blocks represent percentages.

The horizontal axis in figure 1 stops at \$50,000. What about the families earning more than that? The histogram simply ignores them. In 1973, only 1% of American families had incomes above that level: most are represented in the figure.

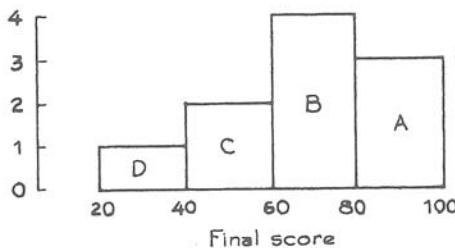
At this point, a good way to learn more about histograms is to do some exercises. Figure 2 shows the same histogram as figure 1, but with a vertical scale supplied. This scale will be useful in working exercise 1. Exercise 8 compares the income data for 1973 and 2004.

Figure 2. The histogram from figure 1, with a vertical scale supplied.

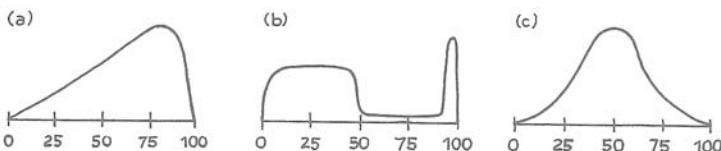


### Exercise Set A

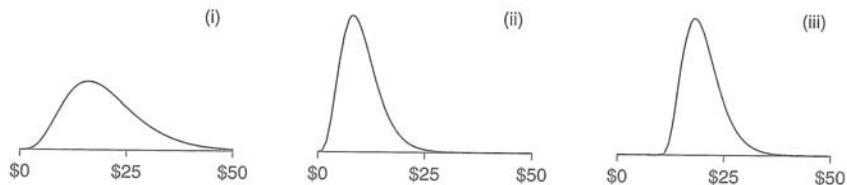
1. About 1% of the families in figure 2 had incomes between \$0 and \$1,000. Estimate the percentage who had incomes—
  - (a) between \$1,000 and \$2,000
  - (b) between \$2,000 and \$3,000
  - (c) between \$3,000 and \$4,000
  - (d) between \$4,000 and \$5,000
  - (e) between \$4,000 and \$7,000
  - (f) between \$7,000 and \$10,000
2. In figure 2, were there more families earning between \$10,000 and \$11,000 or between \$15,000 and \$16,000? Or were the numbers about the same? Make your best guess.
3. The histogram below shows the distribution of final scores in a certain class.
  - (a) Which block represents the people who scored between 60 and 80?
  - (b) Ten percent scored between 20 and 40. About what percentage scored between 40 and 60?
  - (c) About what percentage scored over 60?



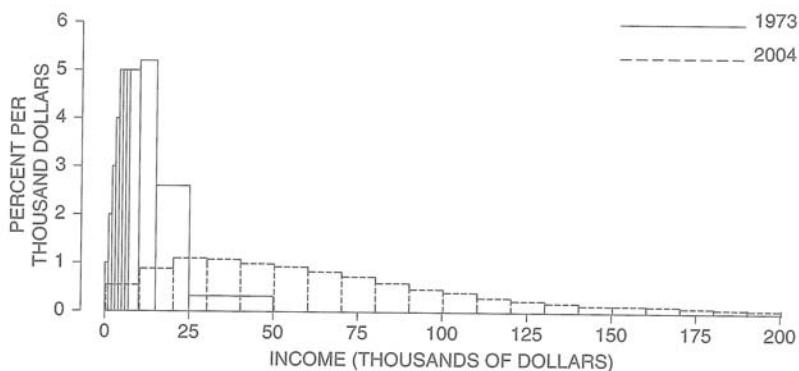
4. Below are sketches of histograms for test scores in three different classes. The scores range from 0 to 100; a passing score was 50. For each class, was the percent who passed about 50%, well over 50%, or well under 50%?



5. One class in exercise 4 had two quite distinct groups of students, with one group doing rather poorly on the test, and the other group doing very well. Which class was it?
6. In class (b) of exercise 4, were there more people with scores in the range 40–50 or 90–100?
7. An investigator collects data on hourly wage rates for three groups of people. Those in group B earn about twice as much as those in group A. Those in group C earn about \$10 an hour more than those in group A. Which histogram belongs to which group? (The histograms don't show wages above \$50 an hour.)



8. The figure below compares the histograms for family incomes in the U.S. in 1973 and in 2004. It looks as if family income went up by a factor of 4 over 30 years. Or did it? Discuss briefly.



Source: Current Population Survey.<sup>3</sup>

*The answers to these exercises are on pp. A45–46.*

## 2. DRAWING A HISTOGRAM

This section explains how to draw a histogram. The method is not difficult, but there are a couple of wrong turns to avoid. The starting point in drawing a histogram is a *distribution table*, which shows the percentage of families with incomes in each class interval (table 1). These percentages are found by going back to the original data—on the 50,000 families—and counting. Nowadays this sort of work is done by computer, and in fact table 1 was drawn up with the help of a computer at the Bureau of the Census.

The computer has to be told what to do with families that fall right on the boundary between two class intervals. This is called an *endpoint convention*. The convention followed in table 1 is indicated by the caption. The left endpoint is included in the class interval, the right endpoint is excluded. In the first line of the table, for example, \$0 is included and \$1,000 is excluded. This interval has the families that earn \$0 or more, but less than \$1,000. A family that earns \$1,000 exactly goes in the next interval.

Table 1. Distribution of families by income in the U.S. in 1973. Class intervals include the left endpoint, but not the right endpoint.

Income level	Percent
\$0-\$1,000	1
\$1,000-\$2,000	2
\$2,000-\$3,000	3
\$3,000-\$4,000	4
\$4,000-\$5,000	5
\$5,000-\$6,000	5
\$6,000-\$7,000	5
\$7,000-\$10,000	15
\$10,000-\$15,000	26
\$15,000-\$25,000	26
\$25,000-\$50,000	8
\$50,000 and over	1

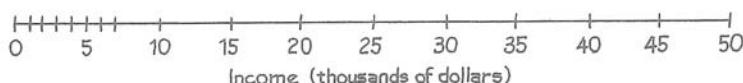
Note: Percents do not add to 100%, due to rounding.

Source: Current Population Survey.<sup>4</sup>

The first step in drawing a histogram is to put down a horizontal axis. For the income histogram, some people get

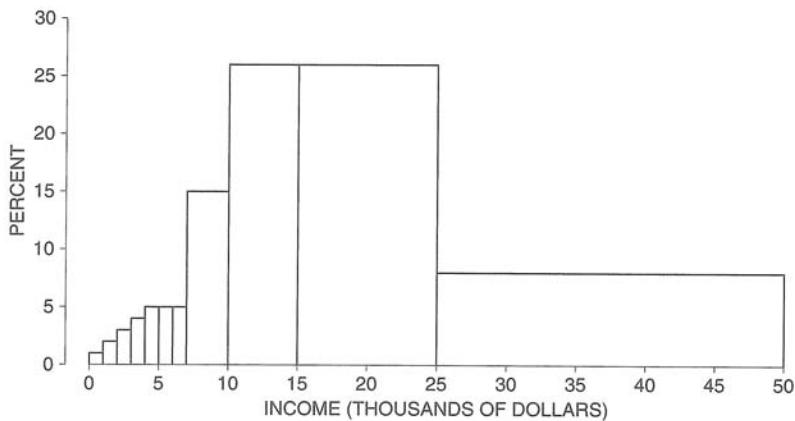


That is a mistake. The interval from \$7,000 to \$10,000 is three times as long as the interval from \$6,000 to \$7,000. So the horizontal axis should look like this:



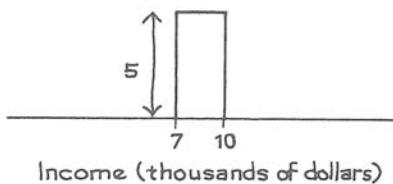
The next step is to draw the blocks. It's tempting to make their heights equal to the percents in the table. Figure 3 shows what happens if you make that mistake. The graph gives much too rosy a picture of the income distribution. For example, figure 3 says there were many more families with incomes over \$25,000 than under \$7,000. The U.S. was a rich country in 1973, but not that rich.

Figure 3. Don't plot the percents.

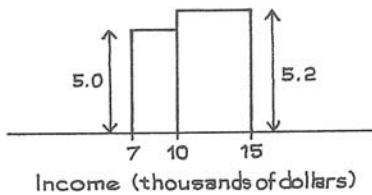


The source of the trouble is that some class intervals are longer than others, so the percents in table 1 are not on a par with one another. The 8% who earn \$25,000 to \$50,000, for instance, are spread over a much larger range of incomes than the 15% who earn \$7,000 to \$10,000. Plotting percents directly ignores this, and makes the blocks over the longer class intervals too big.

There is a simple way to compensate for the different lengths of the class intervals—use thousand-dollar intervals as a common unit. For example, the class interval from \$7,000 to \$10,000 contains three of these intervals: \$7,000 to \$8,000, \$8,000 to \$9,000, and \$9,000 to \$10,000. From table 1, 15% of the families had incomes in the whole interval. Within each of the thousand-dollar sub-intervals, there will only be about 5% of the families. This 5, not the 15, is what should be plotted above the interval \$7,000 to \$10,000.



For a second example, take the interval from \$10,000 to \$15,000. This contains 5 of the thousand-dollar intervals. According to table 1, 26% of the families had incomes in the whole interval. Within each of the 5 smaller intervals there will be about 5.2% of the families:  $26/5 = 5.2$ . The height of the block over the interval \$10,000 to \$15,000 is 5.2.

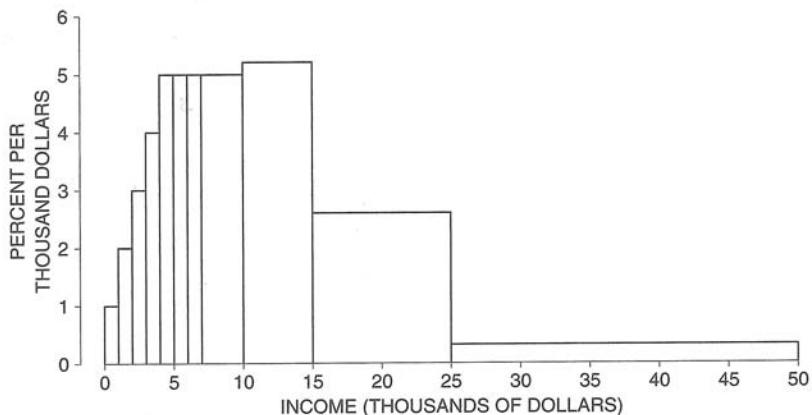


The work is done for two of the lines in table 1. To complete the histogram, do the same thing for the rest of the class intervals. Figure 4 (below) is the result.

To figure out the height of a block over a class interval, divide the percentage by the length of the interval.

That way, the area of the block equals the percentage of families in the class interval. The histogram represents the distribution as if the percent is spread evenly over the class interval. Often, this is a good first approximation.

Figure 4. Distribution of families by income in the U.S. in 1973.



The procedure is straightforward, but the units on the vertical scale are a little complicated. For instance, to get the height of the block over the interval \$7,000 to \$10,000, you divide 15 percent by 3 thousand dollars. So the units for the answer are percent per thousand dollars. Think about the “per” just as you would when reading that there are 50,000 people per square mile in Tokyo: in each square mile of the city, there are about 50,000 people. It is the same with histograms. The height of the block over the interval \$7,000 to \$10,000 is 5% per thousand dollars: in each thousand-dollar interval between \$7,000 and \$10,000, there are about 5% of the families. Figure 4 shows the complete histogram with these units on the vertical scale.

### Exercise Set B

- The table below gives the distribution of educational level for persons age 25 and over in the U.S. in 1960, 1970, and 1991. (“Educational level” means the number of years of schooling completed.) The class intervals include the left endpoint, but not the right; for example, from the second line of the table, in 1960 about 14% of the people had completed 5–8 years of schooling, 8 not included; in 1991, about 4% of the people were in this category. Draw a histogram for the 1991 data. You can interpret “16 or more” as 16–17 years of schooling; not many people completed more than 16 years of school, especially in 1960 and 1970. Why does your histogram have spikes at 8, 12, and 16 years of schooling?

<i>Educational level (years of schooling)</i>	1960	1970	1991
0–5	8	6	2
5–8	14	10	4
8–9	18	13	4
9–12	19	19	11
12–13	25	31	39
13–16	9	11	18
16 or more	8	11	21

Source: Statistical Abstract, 1988, Table 202; 1992, Table 220.

- Redraw the histogram for the 1991 data, combining the first two class intervals into one (0–8 years, with 6% of the people). Does this change the histogram much?
- Draw the histogram for the 1970 data, and compare it to the 1991 histogram. What happened to the educational level of the population between 1970 and 1991—did it go up, go down, or stay about the same?
- What happened to the educational level from 1960 to 1970?

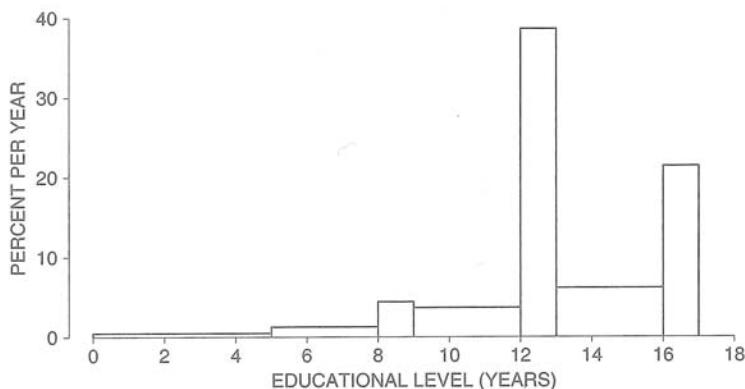
*The answers to these exercises are on p. A46.*

### 3. THE DENSITY SCALE

When reading areas off a histogram, it is convenient to have a vertical scale. The income histogram in the previous section was drawn using the *density scale*.<sup>5</sup> The unit on the horizontal axis was \$1,000 of family income, and the vertical axis showed the percentage of families per \$1,000 of income. Figure 5 is another example of a histogram with a density scale. This is a histogram for educational level of persons age 25 and over in the U.S. in 1991. “Educational level” means years of schooling completed; kindergarten doesn’t count.

The endpoint convention followed in this histogram is a bit fussy. The block over the interval 8–9 years, for example, represents all the people who finished eighth grade, but not ninth grade; people who dropped out part way through ninth

Figure 5. Distribution of persons age 25 and over in the U.S. in 1991 by educational level.



Source: *Statistical Abstract, 1992*, Table 220.

grade are included. The units on the horizontal axis of the histogram are years, so the units on the vertical axis are percent per year. For instance, the height of the histogram over the interval 13–16 years is 6% per year. In other words, about 6% of the population finished the first year of college, another 6% finished the second year, and another 6% finished the third year.

Section 1 described how area in a histogram represents percent. If one block covers a larger area than another, it represents a larger percent of the cases. What does the height of a block represent? Look at the horizontal axis in figure 5. Imagine the people lined up on this axis, with each person stationed at his or her educational level. Some parts of the axis—years—will be more crowded than others. The height of the histogram shows the crowding.

The histogram is highest over the interval 12–13 years, so the crowding is greatest there. This interval has all the people with high-school degrees. (Some people in this interval may have gone on to college, but they did not even finish the first year.) There are two other peaks, a small one at 8–9 years (finishing middle school) and a big one at 16–17 years—finishing college. The peaks show how people tend to stop their schooling at one of the three possible graduations rather than dropping out in between.

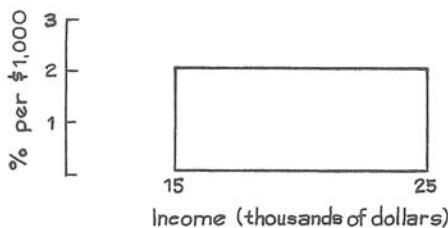
At first, it may be difficult to keep apart the notion of the crowding in an interval, represented by the height of the block, and the number in an interval, represented by the area of the block. An example will help. Look at the blocks over the intervals 8–9 years and 9–12 years in figure 5. The first block is a little taller, so this interval is a little more crowded. However, the block over 9–12 years has a much larger area, so this interval has many more people. Of course, there is more room in the second interval—it's 3 times as long. The two intervals are like the Netherlands and the U.S. The Netherlands is more crowded, but the U.S. has more people.

In a histogram, the height of a block represents crowding—percentage per horizontal unit.

By contrast, the area of the block represents the percentage of cases in the corresponding class interval (section 1).

Once you learn how to use it, the density scale can be quite helpful. For example, take the interval from 9 to 12 years in figure 5—the people who got through their first year of high school but didn't graduate. The height of the block over this interval is nearly 4% per year. In other words, each of the three one-year intervals 9–10, 10–11, and 11–12 holds nearly 4% of the people. So the whole three-year interval must hold nearly  $3 \times 4\% = 12\%$  of the people. Nearly 12% of the population age 25 and over got through at least one year of high school, but failed to graduate.

*Example 1.* The sketch below shows one block of the family-income histogram for a certain city. About what percent of the families in the city had incomes between \$15,000 and \$25,000?

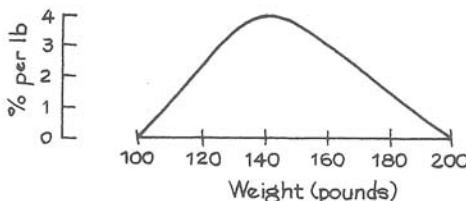


*Solution.* The height of the block is 2% per thousand dollars. Each thousand-dollar interval between \$15,000 and \$25,000 contains about 2% of the families in the city. There are 10 of these thousand-dollar intervals between \$15,000 and \$25,000. The answer is  $10 \times 2\% = 20\%$ . About 20% of the families in the city had incomes between \$15,000 and \$25,000.

The example shows that with the density scale, the areas of the blocks come out in percent. The horizontal units—thousands of dollars—cancel:

$$2\% \text{ per thousand dollars} \times 10 \text{ thousand dollars} = 20\%.$$

*Example 2.* Someone has sketched a histogram for the weights of some people, using the density scale. What's wrong?



*Solution.* The total area is 200%, and should only be 100%. The area can be calculated as follows. The histogram is almost a triangle, whose height is 4% per pound and whose base is 200 lb – 100 lb = 100 lb. The area is

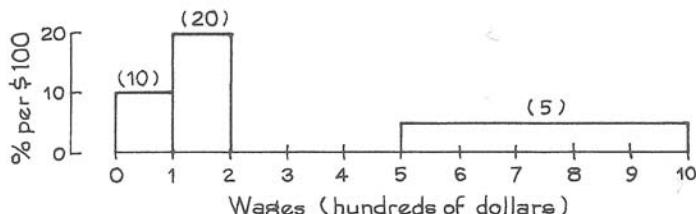
$$\frac{1}{2} \times \text{base} \times \text{height} = \frac{1}{2} \times 100 \text{ lb} \times 4\% \text{ per lb} = 200\%.$$

With the density scale on the vertical axis, the areas of the blocks come out in percent. The area under the histogram over an interval equals the percentage of cases in that interval.<sup>6</sup> The total area under the histogram is 100%.

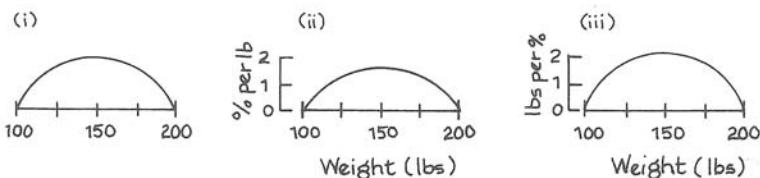
Since 1991, the educational level in the U.S. has continued to increase. Then, 21% of the population had a bachelor's degree or better (the "population" means people age 25 and over). In 2005, the corresponding figure was 28%.

### Exercise Set C

1. A histogram of monthly wages for part-time employees is shown below (densities are marked in parentheses). Nobody earned more than \$1,000 a month. The block over the class interval from \$200 to \$500 is missing. How tall must it be?



2. Three people plot histograms for the weights of subjects in a study, using the density scale. Only one is right. Which one, and why?



3. An investigator draws a histogram for some height data, using the metric system. She is working in centimeters (cm). The vertical axis shows density, and the top of the vertical axis is 10 percent per cm. Now she wants to convert to millimeters (mm). There are 10 millimeters to the centimeter. On the horizontal axis, she has to change 175 cm to \_\_\_\_\_ mm, and 200 cm to \_\_\_\_\_ mm. On the vertical axis, she has to change 10 percent per cm to \_\_\_\_\_ percent per mm, and 5 percent per cm to \_\_\_\_\_ percent per mm.

4. In a Public Health Service study, a histogram was plotted showing the number of cigarettes per day smoked by each subject (male current smokers), as shown below.<sup>7</sup> The density is marked in parentheses. The class intervals include the right endpoint, not the left.

(a) The percentage who smoked 10 cigarettes or less per day is around

1.5%      15%      30%      50%

(b) The percentage who smoked more than a pack a day, but not more than 2 packs, is around

1.5%      15%      30%      50%

(There are 20 cigarettes in a pack.)

(c) The percent who smoked more than a pack a day is around

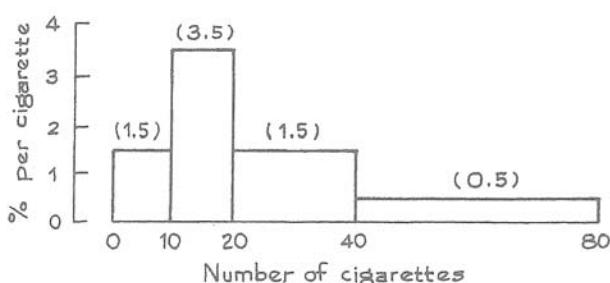
1.5%      15%      30%      50%

(d) The percent who smoked more than 3 packs a day is around

0.25 of 1%      0.5 of 1%      10%

(e) The percent who smoked 15 cigarettes per day is around

0.35 of 1%      0.5 of 1%      1.5%      3.5%      10%



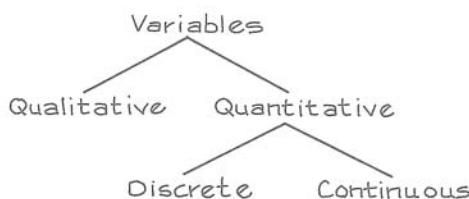
The answers to these exercises are on p. A46.

#### 4. VARIABLES

The Current Population Survey covers many other variables besides income. A *variable* is a characteristic which changes from person to person in a study. Interviewers for the survey use a battery of questions: How old are you? How many people are there in your family? What is your family's total income? Are you married? Do you have a job? The corresponding variables would be: age, family size, family income, marital status, and employment status. Some questions are answered by giving a number: the corresponding variables are *quantitative*. Age, family size, and family income are examples of quantitative variables. Some questions are answered with a descriptive word or phrase, and the corresponding variables are *qualitative*: examples are marital status (single, married, widowed,

divorced, separated) and employment status (employed, unemployed, not in the labor force).

Quantitative variables may be *discrete* or *continuous*. This is not a hard-and-fast distinction, but it is a useful one.<sup>8</sup> For a discrete variable, the values can only differ by fixed amounts. Family size is discrete. Two families can differ in size by 0 or 1 or 2, and so on. Nothing in between is possible. Age, on the other hand, is a continuous variable. This doesn't refer to the fact that a person is continuously getting older; it just means that the difference in age between two people can be arbitrarily small—a year, a month, a day, an hour, ... Finally, the terms *qualitative*, *quantitative*, *discrete*, and *continuous* are also used to describe data—qualitative data are collected on a qualitative variable, and so on.



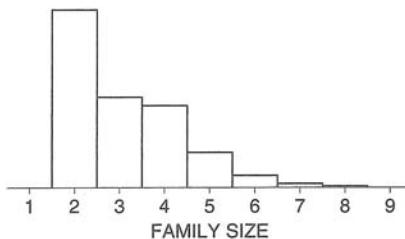
Section 2 showed how to plot a histogram starting with a distribution table. Often the starting point is the raw data—a list of cases (individuals, families, schools, etc.) and the corresponding values of the variable. In order to draw the histogram, a distribution table must be prepared. The first step is to choose the class intervals. With too many or too few, the histogram will not be informative. There is no rule, it is a matter of judgment or trial and error. It is common to start with ten or fifteen class intervals and work from there. In this book, the class intervals will always be given.<sup>9</sup>

When plotting a histogram for a continuous variable, investigators also have to decide on the endpoint convention—what to do with cases that fall right on the boundary. With a discrete variable, there is a convention which gets around this nuisance: center the class intervals at the possible values. For instance, family size can be 2 or 3 or 4, and so on. (The Census does not recognize one person as a family.) The corresponding class intervals in the distribution table would be

<i>Center</i>	<i>Class interval</i>
2	1.5 to 2.5
3	2.5 to 3.5
4	3.5 to 4.5
.	.
.	.

Since a family cannot have 2.5 members, there is no problem with endpoints. Figure 6 (on the next page) shows the histogram for family size. The bars seem to stop at 8; that is because there are so few families with 9 or more people.

Figure 6. Histogram showing distribution of families by size in 2005. With a discrete variable, the class intervals are centered at the possible values.



Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census.

### Exercise Set D

1. Classify each of the following variables as qualitative or quantitative; if quantitative, as discrete or continuous.
  - (a) occupation
  - (b) region of residence
  - (c) weight
  - (d) height
  - (e) number of automobiles owned
2. In the March Current Population Survey, women are asked how many children they have. Results are shown below for women age 25–39, by educational level.
  - (a) Is the number of children discrete or continuous?
  - (b) Draw histograms for these data. (You may take “5 or more” as 5—very few women had more than 5 children.)
  - (c) What do you conclude?

*Distribution of women age 25–39 by educational level and number of children (percent).*

Number of children	Women who are high-school graduates	Women with college degrees
0	30.2	47.9
1	21.8	19.4
2	28.4	22.7
3	13.7	8.0
4	4.4	1.5
5 or more	1.5	0.5

Note: High-school graduates with no further education. College degrees at the level of a B.A. or B.Sc. Own, never-married children under the age of 18. Percents may not add to 100%, due to rounding.

Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census.

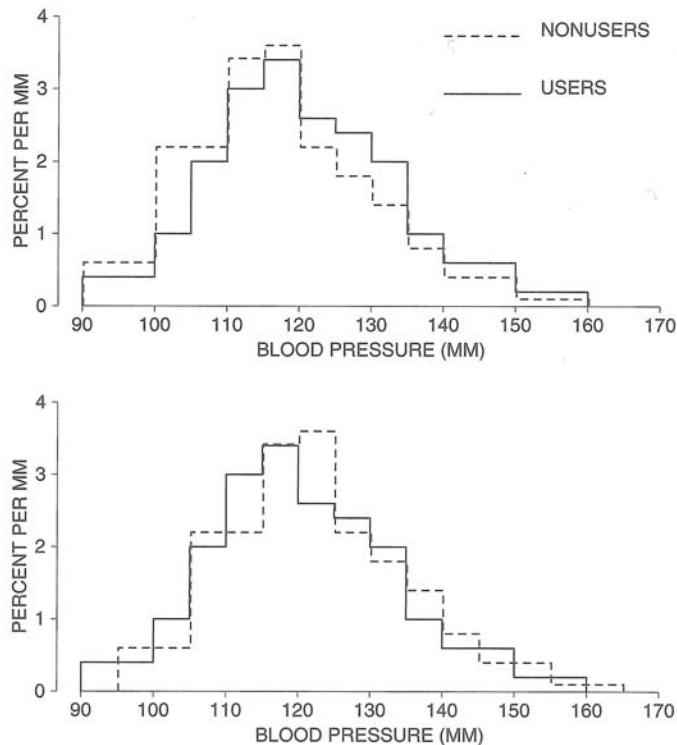
*The answers to these exercises are on p. A47.*

## 5. CONTROLLING FOR A VARIABLE

In the 1960s, many women began using oral contraceptives, “the pill.” Since the pill alters the body’s hormone balance, it is important to see what the side effects are. Research on this question is carried out by the Contraceptive Drug Study at the Kaiser Clinic in Walnut Creek, California. Over 20,000 women in the Walnut Creek area belong to the Kaiser Foundation Health Plan, paying a monthly insurance fee and getting medical services from Kaiser. One of these services is a routine checkup called the “multiphasic.” During the period 1969–1971, about 17,500 women age 17–58 took the multiphasic and became subjects for the Drug Study. Investigators compared the multiphasic results for two different groups of women:

- “users” who take the pill (the treatment group);
- “non-users” who don’t take the pill (the control group);

Figure 7. The effect of the pill. The top panel shows histograms for the systolic blood pressures of the 1,747 users and the 3,040 non-users age 25–34 in the Contraceptive Drug Study. The bottom panel shows the histogram for the non-users shifted to the right by 5 mm.



This is an observational study. It is the women who decided whether to take the pill or not. The investigators just watched what happens.

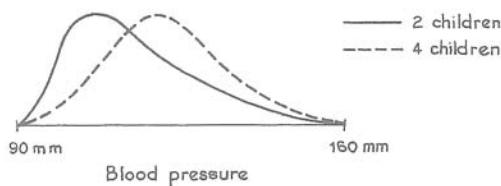
One issue was the effect of the pill on blood pressure. It might seem natural to compare the blood pressures for the users and non-users. However, this could be misleading. Blood pressure tends to go up with age, and the non-users were on the whole older than the users. For example, about 70% of the non-users were over 30, compared to 50% of the users. The effect of age is confounded with the effect of the pill. To make the full effect of the pill visible, it is necessary to make a separate comparison for each age group: this controls for age.<sup>10</sup> We will look only at the women age 25–34. Figure 7 shows the histograms for the users and non-users in this age group. (Blood pressure is measured relative to the length of a column of mercury; the units are “mm,” that is, millimeters.)

The two histograms in the top panel of figure 7 have very similar shapes. However, the user histogram is higher to the right of 120 mm, lower to the left. High blood pressure (above 120 mm) is more prevalent among users, low blood pressure less prevalent. Now imagine that 5 mm were added to the blood pressure of each non-user. That would shift their histogram 5 mm to the right, as shown in the bottom panel of figure 7. In the bottom panel, the two histograms match up quite well. As far as the histograms are concerned, it is as if using the pill adds about 5 mm to the blood pressure of each woman.

This conclusion must be treated with caution. The results of the Contraceptive Drug Study suggest that if a woman goes on the pill, her blood pressure will go up by around 5 mm. But the proof is not complete. It cannot be, because of the design. The Drug Study is an observational study, not a controlled experiment. Part I showed that observational studies can be misleading about cause-and-effect relationships. There could be some factor other than the pill or age, as yet unidentified, which is affecting the blood pressures. For the Drug Study, this is a bit farfetched. The physiological mechanism by which the pill affects blood pressure is well established. The Drug Study data show the size of the effect.

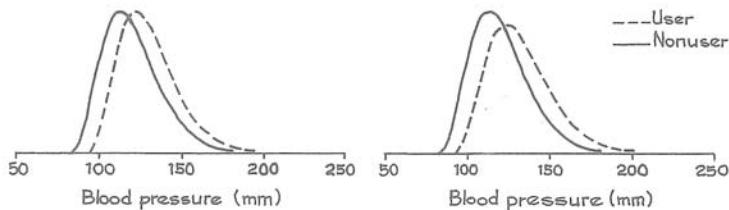
### Exercise Set E

- As a sideline, the Drug Study compared blood pressures for women having different numbers of children. Below are sketches of the histograms for women with 2 or 4 children. Which group has higher blood pressure? Does having children cause the blood pressures of the mothers to change? Or could the change be due to some other factor, whose effects are confounded with the effect of having children?



- (Hypothetical.) The sketches on the next page show results from two other studies

of the pill, for women age 25–29. In one study, the pill adds about 10 mm to blood pressures; in the other, the pill adds about 10%. Which is which, and why?



The answers to these exercises are on p. A47.

## 6. CROSS-TABULATION

The previous section explained how to control for the effect of age: it was a matter of doing the comparison separately for each age group. The comparison was made graphically, through the histograms in figure 7. Some investigators prefer to make the comparison in tabular form, using what is called a *cross-tab* (short for *cross-tabulation*). A cross-tab for blood pressure by age and pill use is shown in table 2. Such tables are a bit imposing, and the eye naturally tends to skip over

Table 2. Systolic blood pressure by age and pill use, for women in the Contraceptive Drug Study, excluding those who were pregnant or taking hormonal medication other than the pill. Class intervals include the left endpoint, but not the right. – means negligible. Table entries are in percent; columns may not add to 100 due to rounding.

Blood pressure (millimeters)	Age 17–24		Age 25–34		Age 35–44		Age 45–58	
	Non-users		Non-users		Non-users		Non-users	
	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
under 90	—	1	1	—	1	1	1	—
90–95	1	—	1	—	2	1	1	1
95–100	3	1	5	4	5	4	4	2
100–105	10	6	11	5	9	5	6	4
105–110	11	9	11	10	11	7	7	7
110–115	15	12	17	15	15	12	11	10
115–120	20	16	18	17	16	14	12	9
120–125	13	14	11	13	9	11	9	8
125–130	10	14	9	12	10	11	11	11
130–135	8	12	7	10	8	10	10	9
135–140	4	6	4	5	5	7	8	8
140–145	3	4	2	4	4	6	7	9
145–150	2	2	2	2	2	5	7	9
150–155	—	1	1	1	1	3	2	4
155–160	—	—	—	1	1	1	1	3
160 and over	—	—	—	—	1	2	2	5
Total percent	100	98	100	99	100	100	99	99
Total number	1,206	1,024	3,040	1,747	3,494	1,028	2,172	437

them until some of the numbers are needed. However, all the cross-tab amounts to is a distribution table for blood pressures, made separately for users and non-users in each age group.

Look at the columns for the age group 17–24. There were 1,206 non-users and 1,024 users. About 1% of the users had blood pressure below 90 mm; the corresponding percentage of non-users was negligible—that is what the dash means. To see the effect of the pill on the blood pressures of women age 17–24, it is a matter of looking at the percents in the columns for non-users and users in the age group 17–24. To see the effect of age, look first at the non-users column in each age group and see how the percents shift toward the high blood pressures as age goes up. Then do the same thing for the users.

### Exercise Set F

1. Use table 2 to answer the following questions.
  - (a) What percentage of users age 17–24 have blood pressures of 140 mm or more?
  - (b) What percentage of non-users age 17–24 have blood pressures of 140 mm or more?
  - (c) What do you conclude?
2. Draw histograms for the blood pressures of the users and non-users age 17–24. What do you conclude?
3. Compare the histograms of blood pressures for non-users age 17–24 and for non-users age 25–34. What do you conclude?

*The answers to these exercises are on p. A47.*

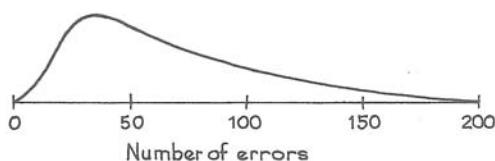
### 7. SELECTIVE BREEDING

In 1927, the psychologist Charles Spearman published *The Abilities of Man*, his theory of human intelligence. Briefly, Spearman held that test scores of intellectual abilities (like reading comprehension, arithmetic, or spatial perception) were weighted sums of two independent components: a general intelligence factor which Spearman called “g,” and an ability factor specific to each test. This theory attracted a great deal of attention.

As part of his Ph.D. research in the psychology department at Berkeley, Robert Tryon decided to check the theory on an animal population, where it is simpler to control extraneous variables.<sup>11</sup> Tryon used rats, which are easy to breed in the laboratory. To test their intelligence, he put the rats into a maze. When they ran the maze, the rats made errors by going into blind alleys. The test consisted of 19 runs through the maze; the animal’s “intelligence score” was the total number

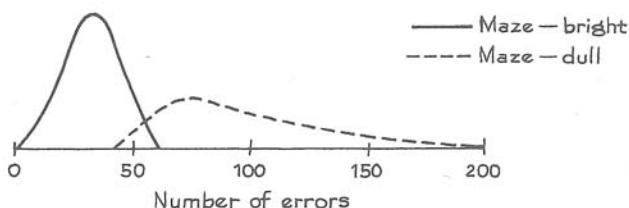
of errors it made. So the bright rats are the ones with low scores, the dulls are the ones with high scores. Tryon started out with 142 rats, and the distribution of their intelligence scores is sketched in figure 8.

Figure 8. Tryon's experiment. Distribution of intelligence in the original population.



The next step in the experiment was to breed for intelligence. In each generation, the "maze-bright" rats (the ones making only a small number of errors) were bred with each other. Similarly, the "maze-dull" animals (with high scores) were bred together. Seven generations later, Tryon had 85 rats in the maze-bright strain, and 68 in the maze-dull strain. There was a clear separation in scores. Figure 9 shows the distribution of intelligence for the two groups, and the histograms barely overlap. (In fact, Tryon went on with selective breeding past the seventh generation, but didn't get much more separation in scores.)

Figure 9. Tryon's experiment. After seven generations of selective breeding, there is a clear separation into "maze-bright" and "maze-dull" strains.

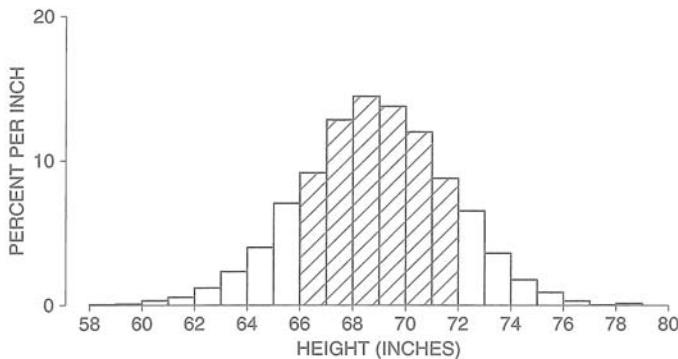


The two strains created by Tryon were used for many other experiments by the Berkeley psychology department. Generations later, rats from the maze-bright population continued to outperform the dulls at maze-running. So Tryon managed to breed for a mental ability—evidence that some mental abilities are at least in part genetically determined. What did the experiment say about Spearman's theory? Tryon found that the maze-bright rats did no better than the maze-dulls on other tests of animal intelligence, such as discriminating between geometric shapes, or between intensities of light. This was evidence against Spearman's theory of a general intelligence factor (at least for rats). On the other hand, Tryon did find intriguing psychological differences between the two rat populations. The "brights" seemed to be unsociable introverts, well adjusted to life in the maze, but neurotic in their relationships with other rats. The "dulls" were quite the opposite.

## 8. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. The figure below shows a histogram for the heights of a representative sample of men. The shaded area represents the percentage of men whose heights were between \_\_\_\_\_ and \_\_\_\_\_. Fill in the blanks.



Source: Data tape supplied by the Inter-University Consortium for Political and Social Research.

2. The age distribution of people in the U.S. in 2004 is shown below. Draw the histogram. (The class intervals include the left endpoint, not the right; for instance, on the second line of the table, 14% of the people were age 5 years or more but had not yet turned 15. The interval for “75 and over” can be ended at 85. Men and women are combined in the data.) Use your histogram to answer the following questions.
- Are there more children age 1, or elders age 71?
  - Are there more 21-year-olds, or 61-year-olds?
  - Are there more people age 0–4, or 65–69?
  - The percentage of people age 35 and over is around 25%, 50%, or 75%?

Age	Percent of population	Age	Percent of population
0–5	7	35–45	15
5–15	14	45–55	14
15–20	7	55–65	10
20–25	7	65–75	6
25–30	7	75 and over	6
30–35	7		

Source: *Statistical Abstract*, 2006, Table 11.

3. The American Housing Survey is done every year by the Bureau of the Census. Data from the 2003 survey can be used to find the distribution of occupied housing units (this includes apartments) by number of rooms. Results for the whole U.S. are shown below, separately for “owner-occupied” and “renter-

occupied” units. Draw a histogram for each of the two distributions. (You may assume that “10 or more” means 10 or 11; very few units have more than 11 rooms.)

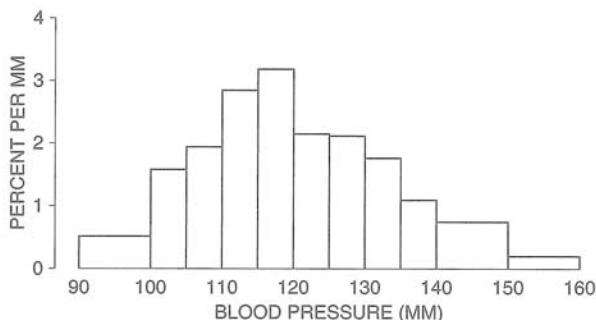
- The owner-occupied percents add up to 100.2% while the renter-occupied percents add up to 100.0%. Why?
- The percentage of one-room units is much smaller for owner-occupied housing. Is that because there are so many more owner-occupied units in total? Answer yes or no, and explain briefly.
- Which are larger, on the whole: the owner-occupied units or the renter-occupied units?

Number of rooms in unit	Owner-occupied (percent)	Renter-occupied (percent)
1	0.0	1.0
2	0.1	2.8
3	1.4	22.7
4	9.7	34.5
5	23.3	22.6
6	26.4	10.4
7	17.5	3.6
8	10.4	1.2
9	5.0	0.5
10 or more	6.4	0.7
Total	100.2	100.0
Number	72.2 million	33.6 million

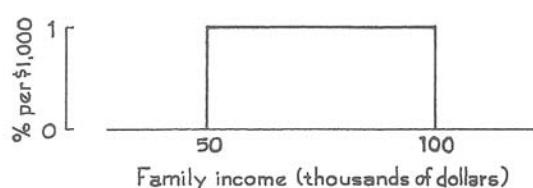
Source: [www.census.gov/hhres/www/housing/ahs/nationaldata.html](http://www.census.gov/hhres/www/housing/ahs/nationaldata.html)

- The figure below is a histogram showing the distribution of blood pressure for all 14,148 women in the Drug Study (section 5). Use the histogram to answer the following questions:

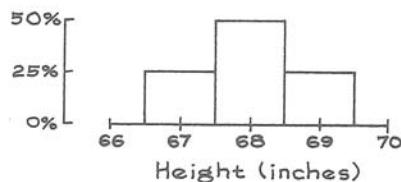
- Is the percentage of women with blood pressures above 130 mm around 25%, 50%, or 75%?
- Is the percentage of women with blood pressures between 90 mm and 160 mm around 1%, 50%, or 99%?
- In which interval are there more women: 135–140 mm or 140–150 mm?



- (d) Which interval is more crowded: 135–140 mm or 140–150 mm?  
 (e) On the interval 125–130 mm, the height of the histogram is about 2.1% per mm. What percentage of the women had blood pressures in this class interval?  
 (f) Which interval has more women: 97–98 mm or 102–103 mm?  
 (g) Which is the most crowded millimeter of all?
5. Someone has sketched one block of a family-income histogram for a wealthy suburb. About what percentage of the families in this suburb had incomes between \$90,000 and \$100,000 a year?



6. (Hypothetical.) In one study, 100 people had their heights measured to the nearest eighth of an inch. A histogram for the results is shown below. Two of the following lists have this histogram. Which ones, and why?
- (i) 25 people, 67 inches tall; 50 people, 68 inches tall; 25 people, 69 inches tall.
  - (ii) 10 people,  $66\frac{3}{4}$  inches tall; 15 people,  $67\frac{1}{4}$  inches tall; 50 people, 68 inches tall; 25 people, 69 inches tall.
  - (iii) 30 people, 67 inches tall; 40 people, 68 inches tall; 30 people, 69 inches tall.



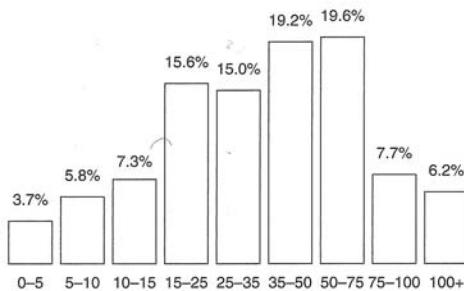
7. Two histograms are sketched below. One shows the distribution of age at death from natural causes (heart disease, cancer, and so forth). The other shows age at death from trauma (accident, murder, suicide). Which is which, and why?



8. The figure on the next page (adapted from the *San Francisco Chronicle*, May 18, 1992) shows the distribution of American families by income. Ranges include

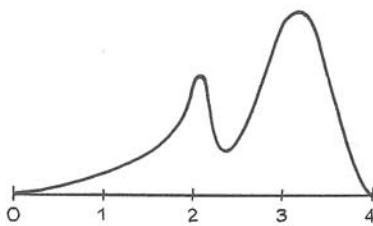
the left endpoint but not the right. For example, 3.7% of the families had incomes in the range \$0–\$4,999, 5.8% had incomes in the range \$5,000–\$9,999, and so forth. True or false, and explain:

- Although American families are not spread evenly over the whole income range, the families that earn between \$10,000 and \$35,000 are spread fairly evenly over that range.
- The families that earn between \$35,000 and \$75,000 are spread fairly evenly over that range.
- The graph is a histogram.



9. In a survey carried out at the University of California, Berkeley, a sample of students were interviewed and asked what their grade-point average was. A histogram of the results is shown below. (GPA ranges from 0 to 4, and 2 is a bare pass.)

- True or false: more students reported a GPA in the range 2.0 to 2.1 than in the range 1.5 to 1.6.
- True or false: more students reported a GPA in the range 2.0 to 2.1 than in the range 2.5 to 2.6.
- What accounts for the spike at 2?



10. The table on the next page shows the distribution of adults by the last digit of their age, as reported in the Census of 1880 and the Census of 1970.<sup>12</sup> You might expect each of the ten possible digits to turn up for 10% of the people, but this is not the case. For example, in 1880, 16.8% of all persons reported an age ending in 0—like 30 or 40 or 50. In 1970, this percentage was only 10.6%.

- Draw histograms for these two distributions.

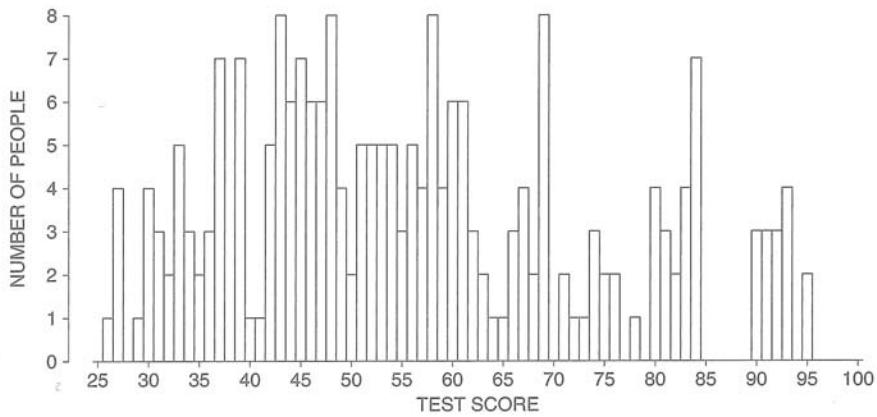
- (b) In 1880, there was a strong preference for the digits 0 and 5. How can this be explained?
- (c) In 1970, the preference was much weaker. How can this be explained?
- (d) Are even digits more popular, or odd ones, in 1880? 1970?

<i>Digit</i>	<i>1880</i>	<i>1970</i>
0	16.8	10.6
1	6.7	9.9
2	9.4	10.0
3	8.6	9.6
4	8.8	9.8
5	13.4	10.0
6	9.4	9.9
7	8.5	10.2
8	10.2	10.0
9	8.2	10.1

Source: United States Census.

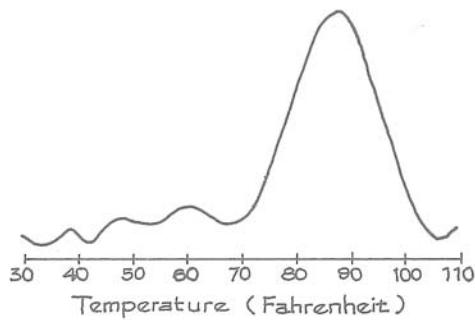
11. In the Sanitary District of Chicago, operating engineers are hired on the basis of a competitive civil-service examination. In 1966, there were 223 applicants for 15 jobs. The exam was held on March 12; the test scores are shown below, arranged in increasing order. The height of each bar in the histogram (top of next page) shows the number of people with the corresponding score. The examiners were charged with rigging the exam.<sup>13</sup> Why?

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



12. The late 1960s and early 1970s were years of turmoil in the U.S. Psychologists thought that rioting was related (among other things) to temperature, with hotter weather making people more aggressive.<sup>14</sup> Two investigators, however, argued that “the frequency of riots should increase with temperature through the mid-80s but then go down sharply with increases in temperature beyond this level.”

To support their theory, they collected data on 102 riots over the period 1967–71, including the temperature in the city where the riot took place. They plotted a histogram for the distribution of riots by temperature (a sketch is shown below). There is a definite peak around 85°. True or false, and explain: the histogram shows that higher temperatures prevent riots.



## 9. SUMMARY

1. A *histogram* represents percents by area. It consists of a set of blocks. The area of each block represents the percentage of cases in the corresponding *class interval*.
2. With the *density scale*, the height of each block equals the percentage of cases in the corresponding class interval, divided by the length of that interval.
3. With the density scale, area comes out in percent, and the total area is 100%. The area under the histogram between two values gives the percentage of cases falling in that interval.
4. A *variable* is a characteristic of the subjects in a study. It can be either *qualitative* or *quantitative*. A quantitative variable can be either *discrete* or *continuous*.
5. A confounding factor is sometimes controlled for by *cross-tabulation*.

# 4

## The Average and the Standard Deviation

*It is difficult to understand why statisticians commonly limit their enquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once.*

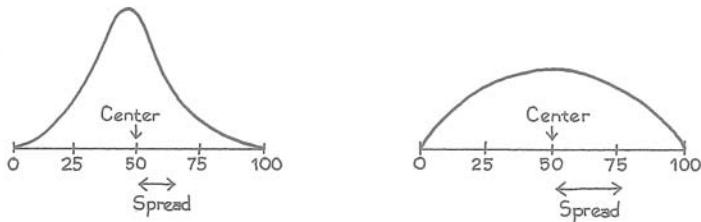
—SIR FRANCIS GALTON (ENGLAND, 1822–1911)<sup>1</sup>

### 1. INTRODUCTION

A histogram can be used to summarize large amounts of data. Often, an even more drastic summary is possible, giving just the center of the histogram and the spread around the center. (“Center” and “spread” are ordinary words here, without any special technical meaning.) Two histograms are sketched in figure 1 on the next page. The center and spread are shown. Both histograms have the same center, but the second one is more spread out—there is more area farther away from the center. For statistical work, precise definitions have to be given, and there are several ways to go about this. The *average* is often used to find the center, and so is the *median*.<sup>2</sup> The *standard deviation* measures spread around the average; the *interquartile range* is another measure of spread.

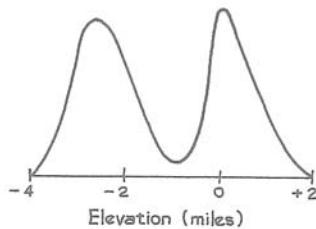
The histograms in figure 1 can be summarized by the center and the spread. However, things do not always work out so well. For instance, figure 2 gives the distribution of elevation over the earth’s surface. Elevation is shown along the

Figure 1. Center and spread. The centers of the two histograms are the same, but the second histogram is more spread out.



horizontal axis, in miles above (+) or below (−) sea level. The area under the histogram between two elevations gives the percentage of the earth's surface area between those elevations. There are clear peaks in this histogram. Most of the surface area is taken up by the sea floors, around 3 miles below sea level; or the continental plains, around sea level. Reporting only the center and spread of this histogram would miss the two peaks.<sup>3</sup>

Figure 2. Distribution of the surface area of the earth by elevation above (+) or below (−) sea level.



## 2. THE AVERAGE

The object of this section is to review the average; the difference between *cross-sectional* and *longitudinal* surveys will also be discussed. The context is HANES—the Health and Nutrition Examination Survey, in which the Public Health Service examines a representative cross section of Americans. This survey has been done at irregular intervals since 1959 (when it was called the Health Examination Survey). The objective is to get baseline data about—

- demographic variables, like age, education, and income;
- physiological variables like height, weight, blood pressure, and serum cholesterol levels;
- dietary habits;
- prevalence of diseases.

Subsequent analysis focuses on the interrelationships among the variables, and has some impact on health policy.<sup>4</sup>

The HANES2 sample was taken during the period 1976–80. Before looking at the data, let's make a quick review of averages.

The average of a list of numbers equals their sum, divided by how many there are.

For instance, the list 9, 1, 2, 2, 0 has 5 entries, the first being 9. The average of the list is

$$\frac{9 + 1 + 2 + 2 + 0}{5} = \frac{14}{5} = 2.8$$

Let's get back to HANES. What did the men and women in the sample (age 18–74) look like?

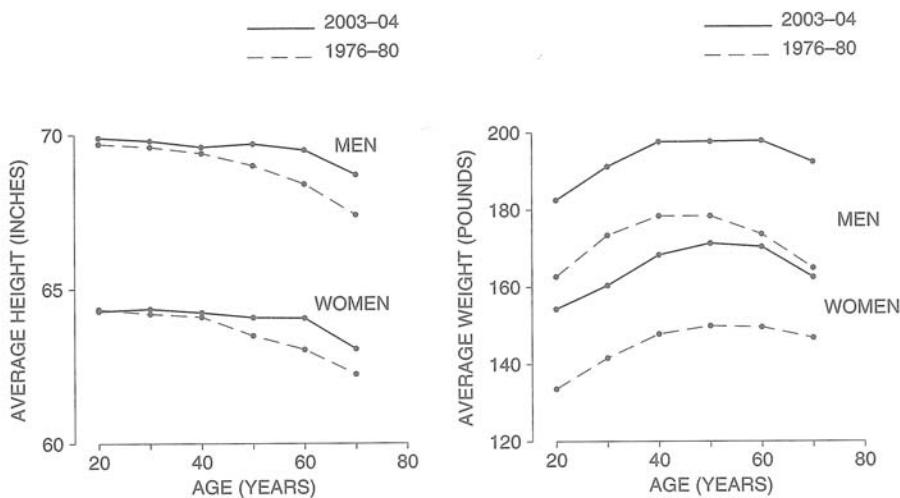
- The average height of the men was 5 feet 9 inches, and their average weight was 171 pounds.
- The average height of the women was 5 feet 3.5 inches, and their average weight was 146 pounds.

They're pretty chubby.

What's happened since 1980? The survey was done again in 2003–04 (HANES5). Average heights went up by a fraction of an inch, while weights went up by nearly 20 pounds—both for men and for women.

Figure 3 shows the averages for men and women, and for each age group; averages are joined by straight lines. From HANES2 to HANES5, average heights went up a little in each group—but average weights went up a lot. This could become a serious public-health problem, because excess weight is associated with many diseases, including heart disease, cancer, and diabetes.

Figure 3. Age-specific average heights and weights for men and women 18–74 in the HANES sample. The panel on the left shows height, the panel on the right shows weight.



Source: [www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm)

The average is a powerful way of summarizing data—many histograms are compressed into the four curves. But this compression is achieved only by smoothing away individual differences. For instance, in 2003–04, the average height of the men age 18–24 was 5 feet 10 inches. But 15% of them were taller than 6 feet 1 inch; another 15% were shorter than 5 feet 6 inches. This diversity is hidden by the average.

For a moment, we return to design issues (chapter 2). In the 1976–80 data, the average height of men appears to decrease after age 20, dropping about two inches in 50 years. Similarly for women. Should you conclude that an average person got shorter at this rate? Not really. HANES is *cross-sectional*, not *longitudinal*. In a cross-sectional study, different subjects are compared to each other at one point in time. In a longitudinal study, subjects are followed over time, and compared with themselves at different points in time. The people age 18–24 in figure 3 are completely different from those age 65–74. The first group was born a lot later than the second.

There is evidence to suggest that, over time, Americans have been getting taller. This is called the *secular trend* in height, and its effect is confounded with the effect of age in figure 3. Most of the two-inch drop in height seems to be due to the secular trend. The people age 65–74 were born around 50 years before those age 18–24, and are an inch or two shorter for that reason.<sup>5</sup> On the other hand, the secular trend has slowed down. (Reasons are unclear.) Average heights only increased a little from 1976–80 to 2003–04. The slowing also explains why the height curves for 2003–04 are flatter than the curves for 1976–80.

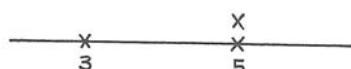
If a study draws conclusions about the effects of age, find out whether the data are cross-sectional or longitudinal.

### Exercise Set A

1. (a) The numbers 3 and 5 are marked by crosses on the horizontal line below. Find the average of these two numbers and mark it by an arrow.



- (b) Repeat (a) for the list 3, 5, 5.

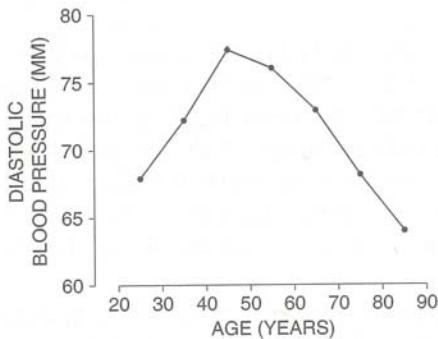


- (c) Two numbers are shown below by crosses on a horizontal axis. Draw an arrow pointing to their average.



2. A list has 10 entries. Each entry is either 1 or 2 or 3. What must the list be if the average is 1? If the average is 3? Can the average be 4?

3. Which of the following two lists has a bigger average? Or are they the same? Try to answer without doing any arithmetic.
  - (i) 10, 7, 8, 3, 5, 9
  - (ii) 10, 7, 8, 3, 5, 9, 11
4. Ten people in a room have an average height of 5 feet 6 inches. An 11th person, who is 6 feet 5 inches tall, enters the room. Find the average height of all 11 people.
5. Twenty-one people in a room have an average height of 5 feet 6 inches. A 22nd person, who is 6 feet 5 inches tall, enters the room. Find the average height of all 22 people. Compare with exercise 4.
6. Twenty-one people in a room have an average height of 5 feet 6 inches. A 22nd person enters the room. How tall would he have to be to raise the average height by 1 inch?
7. In figure 2, are the Rocky Mountains plotted near the left end of the axis, the middle, or the right end? What about Kansas? What about the trenches in the sea floor, like the Marianas trench?
8. Diastolic blood pressure is considered a better indicator of heart trouble than systolic pressure. The figure below shows age-specific average diastolic blood pressure for the men age 20 and over in HANES5 (2003–04).<sup>6</sup> True or false: the data show that as men age, their diastolic blood pressure increases until age 45 or so, and then decreases. If false, how do you explain the pattern in the graph? (Blood pressure is measured in “mm,” that is, millimeters of mercury.)



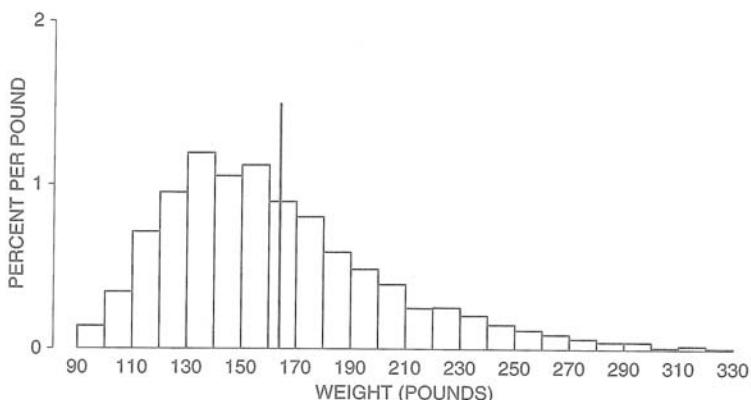
9. Average hourly earnings are computed each month by the Bureau of Labor Statistics using payroll data from commercial establishments. The Bureau figures the total wages paid out (to nonsupervisory personnel), and divides by the total hours worked. During recessions, average hourly earnings typically go up. When the recession ends, average hourly earnings often start going down. How can this be?

*The answers to these exercises are on pp. A47–48.*

### 3. THE AVERAGE AND THE HISTOGRAM

This section will indicate how the average and the median are related to histograms. To begin with an example, there were 2,696 women age 18 and over in HANES5 (2003–04). Their average weight was 164 pounds. It is natural to guess

Figure 4. Histogram for the weights of the 2,696 women in the HANES5 sample. The average is marked by a vertical line. Only 41% of the women were above average in weight.

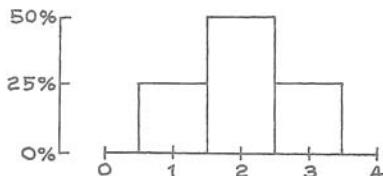


Source: [www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm).

that 50% of them were above average in weight, and 50% were below average. However, this guess is somewhat off. In fact, only 41% were above average, and 59% were below average. Figure 4 shows a histogram for the data: the average is marked by a vertical line. In other situations, the percentages can be even farther from 50%.

How is this possible? To find out, it is easiest to start with some hypothetical data—the list 1, 2, 2, 3. The histogram for this list (figure 5) is symmetric about the value 2. The average equals 2. If the histogram is symmetric around a value, that value equals the average. Furthermore, half the area under the histogram lies to the left of that value, and half to the right. (What does symmetry mean? Imagine drawing a vertical line through the center of the histogram and folding the histogram in half around that line: the two halves should match up.)

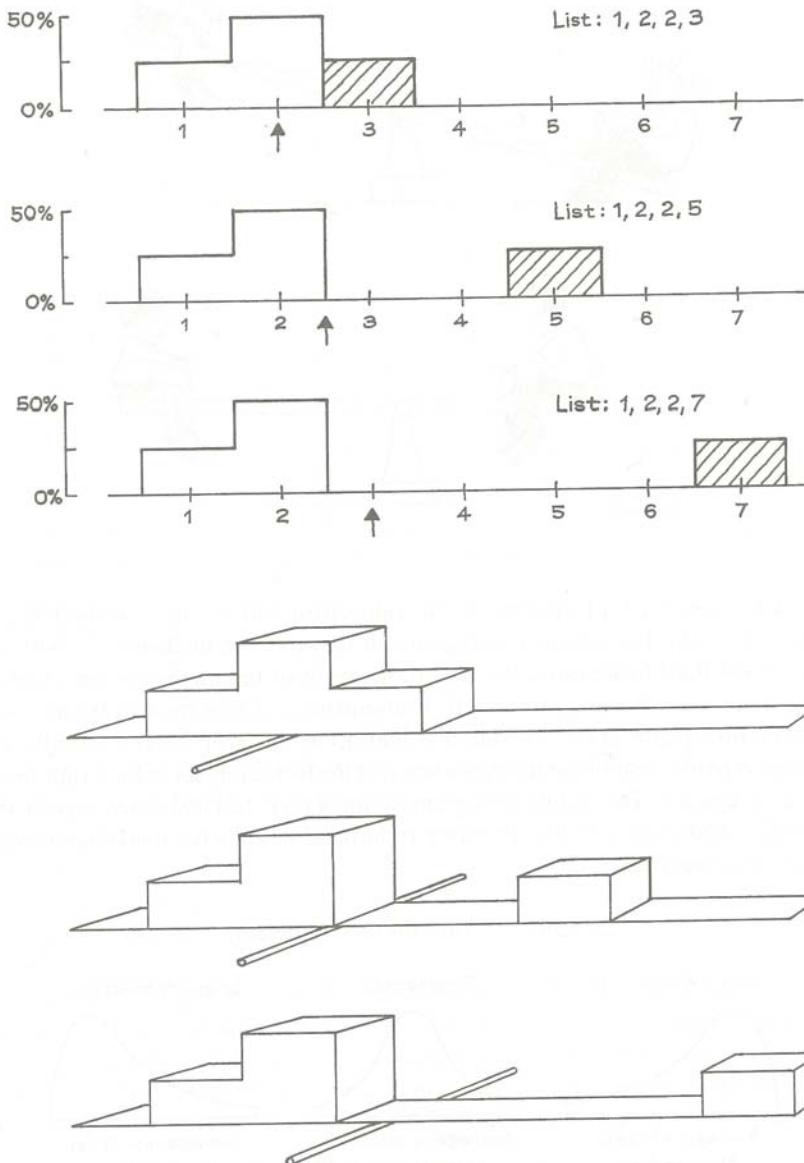
Figure 5. Histogram for the list 1, 2, 2, 3. The histogram is symmetric around 2, the average: 50% of the area is to the left of 2, and 50% is to the right.



What happens when the value 3 on the list 1, 2, 2, 3 is increased, say to 5 or 7? As shown in figure 6, the rectangle over that value moves off to the right, destroying the symmetry. The average for each histogram is marked with an arrow, and the arrow shifts to the right following the rectangle. To see why, imagine the histogram is made out of wooden blocks attached to a stiff, weightless board. Put the histogram across a taut wire, as illustrated in the bottom panel of figure 6. The

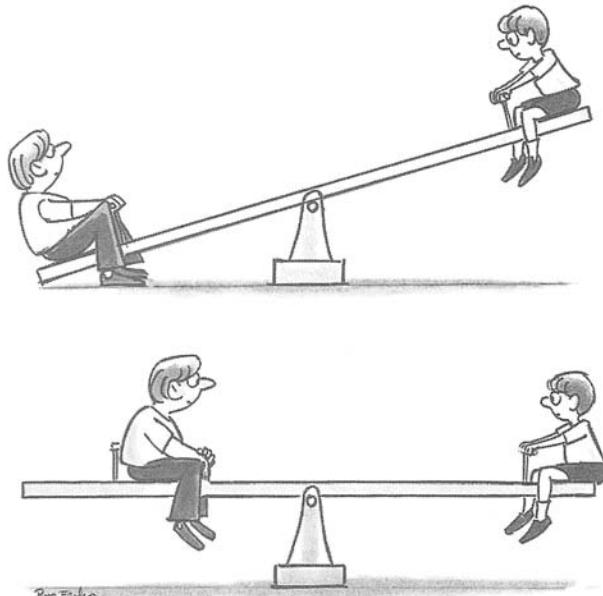
histogram will balance at the average.<sup>7</sup> A small area far away from the average can balance a large area close to the average, because areas are weighted by their distance from the balance point.

Figure 6. The average. The top panel shows three histograms; the averages are marked by arrows. As the shaded box moves to the right, it pulls the average along with it. The area to the left of the average gets up to 75%. The bottom panel shows the same three histograms made out of wooden blocks attached to a stiff, weightless board. The histograms balance when supported at the average.



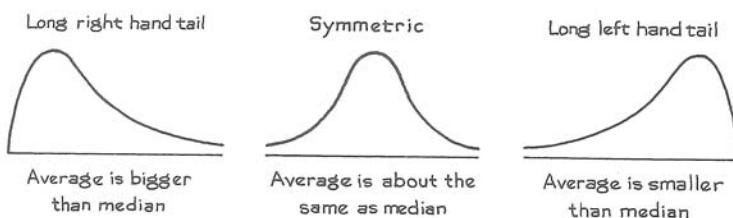
A histogram balances when supported at the average.

A small child sits farther away from the center of a seesaw in order to balance a large child sitting closer to the center. Blocks in a histogram work the same way. That is why the percentage of cases on either side of the average can differ from 50%.



The *median* of a histogram is the value with half the area to the left and half to the right. For all three histograms in figure 6, the median is 2. With the second and third histograms, the area to the right of the median is far away by comparison with the area to the left. Consequently, if you tried to balance one of those histograms at the median, it would tip to the right. More generally, the average is to the right of the median whenever the histogram has a long right-hand tail, as in figure 7. The weight histogram (figure 4 on p. 62) had an average of 164 lbs and a median of 155 lbs. The long right-hand tail is what made the average bigger than the median.

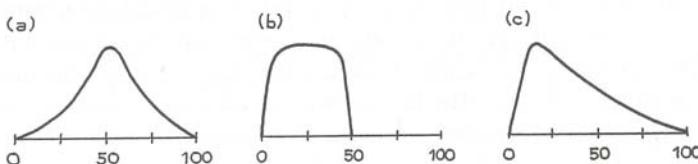
Figure 7. The tails of a histogram.



For another example, median family income in the U.S. in 2004 was about \$54,000. The income histogram has a long right-hand tail, and the average was higher—\$60,000.<sup>8</sup> When dealing with long-tailed distributions, statisticians might use the median rather than the average, if the average pays too much attention to the extreme tail of the distribution. We return to this point in the next chapter.

### Exercise Set B

1. Below are sketches of histograms for three lists. Fill in the blank for each list: the average is around \_\_\_\_\_. Options: 25, 40, 50, 60, 75.



2. For each histogram in exercise 1, is the median equal to the average? or is it to the left? to the right?
3. Look back at the cigarette histogram on p. 42. The median is around \_\_\_\_\_. Fill in the blank. Options: 10, 20, 30, 40
4. For this cigarette histogram, is the average around 15, 20, or 25?
5. For registered students at universities in the U.S., which is larger: average age or median age?
6. For each of the following lists of numbers, say whether the entries are on the whole around 1, 5, or 10 in size. No arithmetic is needed.
- |                        |                  |
|------------------------|------------------|
| (a) 1.3, 0.9, 1.2, 0.8 | (b) 13, 9, 12, 8 |
| (c) 7, 3, 6, 4         | (d) 7, -3, -6, 4 |

*The answers to these exercises are on pp. A48–49.*

*Technical note.* The median of a list is defined so that half or more of the entries are at the median or bigger, and half or more are at the median or smaller. This will be illustrated on 4 lists—

- (a) 1, 5, 7
- (b) 1, 2, 5, 7
- (c) 1, 2, 2, 7, 8
- (d) 8, -3, 5, 0, 1, 4, -1

For list (a), the median is 5: two entries out of the three are 5 or more, and two are 5 or less. For list (b), any value between 2 and 5 is a median; if pressed, most statisticians would choose 3.5 (which is halfway between 2 and 5) as “the” median. For list (c), the median is 2: four entries out of five are 2 or more, and three are 2 or less. To find the median of list (d), arrange it in increasing order:

$$-3, -1, 0, 1, 4, 5, 8$$

There are seven entries on this list: four are 1 or more, and four are 1 or less. So, 1 is the median.

#### 4. THE ROOT-MEAN-SQUARE

The next main topic in the chapter is the *standard deviation*, which is used to measure spread. This section presents a mathematical preliminary, illustrated on the list

$$0, \quad 5, \quad -8, \quad 7, \quad -3$$

How big are these five numbers? The average is 0.2, but this is a poor measure of size. It only means that to a large extent, the positives cancel the negatives. The simplest way around the problem would be to wipe out the signs and then take the average. However, statisticians do something else: they apply the *root-mean-square* operation to the list. The phrase “root-mean-square” says how to do the arithmetic, provided you remember to read it backwards:

- SQUARE all the entries, getting rid of the signs.
- Take the MEAN (average) of the squares.
- Take the square ROOT of the mean.

This can be expressed as an equation, with root-mean-square abbreviated to r.m.s.

$$\text{r.m.s. size of a list} = \sqrt{\text{average of (entries}^2\text{)}}.$$

*Example 1.* Find the average, the average neglecting signs, and the r.m.s. size of the list 0, 5, −8, 7, −3.

*Solution.*

$$\text{average} = \frac{0 + 5 - 8 + 7 - 3}{5} = 0.2$$

$$\text{average neglecting signs} = \frac{0 + 5 + 8 + 7 + 3}{5} = 4.6$$

$$\text{r.m.s. size} = \sqrt{\frac{0^2 + 5^2 + (-8)^2 + 7^2 + (-3)^2}{5}} = \sqrt{29.4} \approx 5.4$$

The r.m.s. size is a little bigger than the average neglecting signs. It always turns out like that—except in the trivial case when all the entries are the same size. The root and the square do not cancel, due to the intervening operation of taking the mean. (The “≈” means “nearly equal;” some rounding has been done.)

There doesn’t seem to be much to choose between the 5.4 and the 4.6 as a measure of the overall size for the list in the example. Statisticians use the r.m.s. size because it fits in better with the algebra that they have to do.<sup>9</sup> Whether this explanation is appealing or not, don’t worry. Everyone is suspicious of the r.m.s. at first, and gets used to it very quickly.

### Exercise Set C

1. (a) Find the average and the r.m.s. size of the numbers on the list  
1, -3, 5, -6, 3.  
(b) Do the same for the list -11, 8, -9, -3, 15.
2. Guess whether the r.m.s. size of each of the following lists of numbers is around 1, 10, or 20. No arithmetic is required.
  - (a) 1, 5, -7, 8, -10, 9, -6, 5, 12, -17
  - (b) 22, -18, -33, 7, 31, -12, 1, 24, -6, -16
  - (c) 1, 2, 0, 0, -1, 0, 0, -3, 0, 1
3. (a) Find the r.m.s. size of the list 7, 7, 7, 7.  
(b) Repeat, for the list 7, -7, 7, -7.
4. Each of the numbers 103, 96, 101, 104 is almost 100 but is off by some amount.  
Find the r.m.s. size of the amounts off.
5. The list 103, 96, 101, 104 has an average. Find it. Each number in the list is off the average by some amount. Find the r.m.s. size of the amounts off.
6. A computer is programmed to predict test scores, compare them with actual scores, and find the r.m.s. size of the prediction errors. Glancing at the printout, you see the r.m.s. size of the prediction errors is 3.6, and the following results for the first ten students:

predicted score:	90	90	87	80	42	70	67	60	83	94
actual score:	88	70	81	85	63	77	66	49	71	69

Does the printout seem reasonable, or is something wrong with the computer?

*The answers to these exercises are on p. A49.*

### 5. THE STANDARD DEVIATION

As the quote at the beginning of the chapter suggests, it is often helpful to think of the way a list of numbers spreads out around the average. This spread is usually measured by a quantity called the *standard deviation*, or SD. The SD measures the size of deviations from the average: it is a sort of average deviation. The program is to interpret the SD in the context of real data, and then see how to calculate it.

There were 2,696 women age 18 and over in the HANES5 sample. The average height of these women was about 63.5 inches, and the SD was close to 3 inches. The average tells us that most of the women were somewhere around 63.5 inches tall. But there were deviations from the average. Some of the women were taller than average, some shorter. How big were these deviations? That is where the SD comes in.

The SD says how far away numbers on a list are from their average. Most entries on the list will be somewhere around one SD away from the average. Very few will be more than two or three SDs away.

The SD of 3 inches says that many of the women differed from the average height by 1 or 2 or 3 inches: 1 inch is a third of an SD, and 3 inches is an SD. Few women differed from the average height by more than 6 inches (two SDs).

There is a rule of thumb which makes this idea more quantitative, and which applies to many data sets.

Roughly 68% of the entries on a list (two in three) are within one SD of the average, the other 32% are further away. Roughly 95% (19 in 20) are within two SDs of the average, the other 5% are further away. This is so for many lists, but not all.

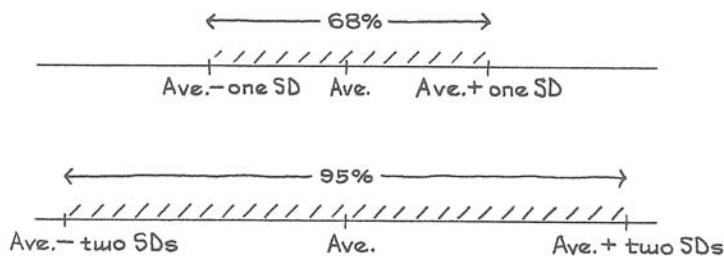


Figure 8 shows the histogram for the heights of women age 18 and over in HANES5. The average is marked by a vertical line, and the region within one SD of the average is shaded. This shaded area represents the women who differed from average height by one SD or less. The area is about 72%. About 72% of the women differed from the average height by one SD or less.

Figure 8. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within one SD of the average is shaded: 72% of the women differed from average by one SD (3 inches) or less.

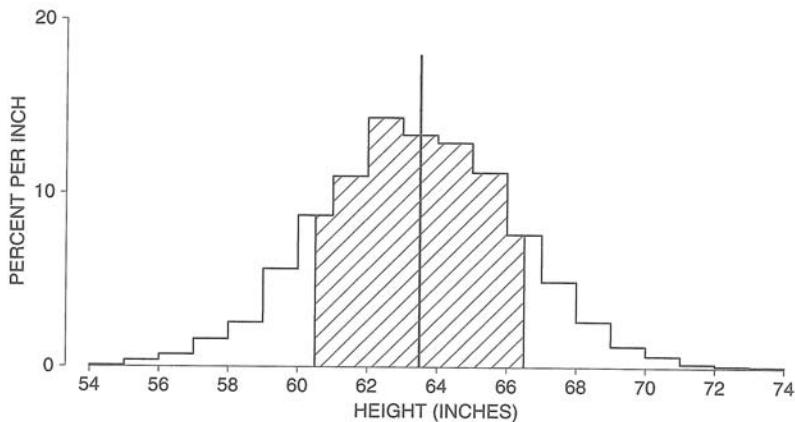
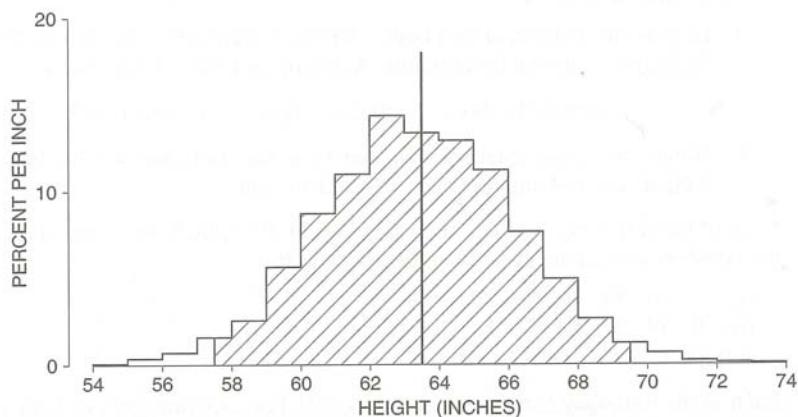


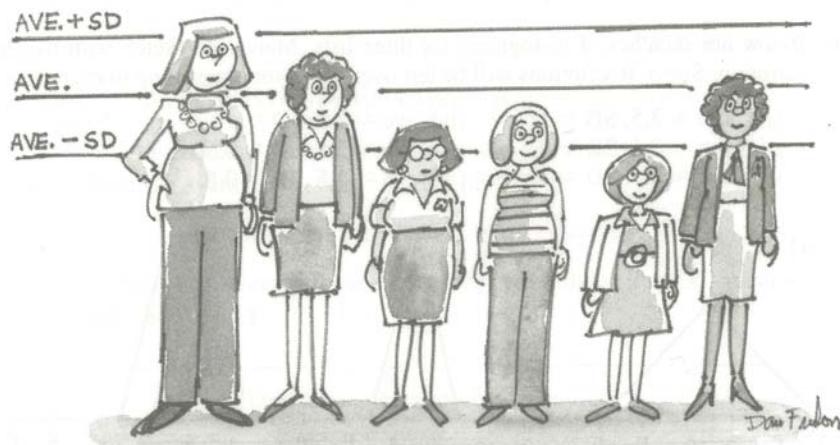
Figure 9 shows the same histogram. Now the area within two SDs of average is shaded. This shaded area represents the women who differed from average height by two SDs or less. The area is about 97%. About 97% of the women differed from the average height by two SDs or less.

Figure 9. The SD and the histogram. Heights of 2,696 women age 18 and over in HANES5. The average of 63.5 inches is marked by a vertical line. The region within two SDs of the average is shaded: 97% of the women differed from average by two SDs (6 inches) or less.



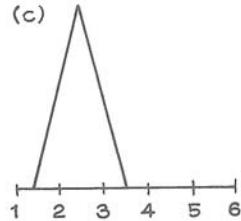
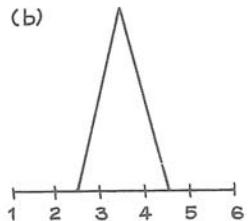
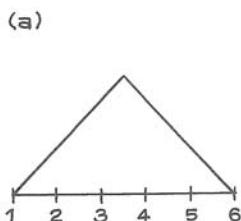
To sum up, about 72% of the women differed from average by one SD or less, and 97% differed from average by two SDs or less. There was only one woman in the sample who was more than three SDs away from the average, and none were more than four SDs away. For this data set, the 68%-95% rule works quite well. Where do the 68% and 95% come from? See chapter 5.<sup>10</sup>

*About two-thirds of the HANES women differed from the average by less than one SD.*



## Exercise Set D

1. The Public Health Service found that for boys age 11 in HANES2, the average height was 146 cm and the SD was 8 cm. Fill in the blanks.
- One boy was 170 cm tall. He was above average, by \_\_\_\_\_ SDs.
  - Another boy was 148 cm tall. He was above average, by \_\_\_\_\_ SDs.
  - A third boy was 1.5 SDs below average height. He was \_\_\_\_\_ cm tall.
  - If a boy was within 2.25 SDs of average height, the shortest he could have been is \_\_\_\_\_ cm and the tallest is \_\_\_\_\_ cm.
2. This continues exercise 1.
- Here are the heights of four boys: 150 cm, 130 cm, 165 cm, 140 cm. Match the heights with the descriptions. A description may be used twice.  
unusually short      about average      unusually tall
  - About what percentage of boys age 11 in the study had heights between 138 cm and 154 cm? Between 130 and 162 cm?
3. Each of the following lists has an average of 50. For which one is the spread of the numbers around the average biggest? smallest?
- 0, 20, 40, 50, 60, 80, 100
  - 0, 48, 49, 50, 51, 52, 100
  - 0, 1, 2, 50, 98, 99, 100
4. Each of the following lists has an average of 50. For each one, guess whether the SD is around 1, 2, or 10. (This does not require any arithmetic.)
- 49, 51, 49, 51, 49, 51, 49, 51, 49, 51
  - 48, 52, 48, 52, 48, 52, 48, 52, 48, 52
  - 48, 51, 49, 52, 47, 52, 46, 51, 53, 51
  - 54, 49, 46, 49, 51, 53, 50, 50, 49, 49
  - 60, 36, 31, 50, 48, 50, 54, 56, 62, 53
5. The SD for the ages of the people in the HANES5 sample is around \_\_\_\_\_. Fill in the blank, using one of the options below. Explain briefly. (This survey was discussed in section 2; the age range was 0–85 years.)
- 5 years      25 years      50 years
6. Below are sketches of histograms for three lists. Match the sketch with the description. Some descriptions will be left over. Give your reasoning in each case.
- |                      |                     |
|----------------------|---------------------|
| (i)<br>(ii)<br>(iii) | (iv)<br>(v)<br>(vi) |
|----------------------|---------------------|



7. (Hypothetical). In a clinical trial, data collection usually starts at “baseline,” when the subjects are recruited into the trial but before they are randomized to treatment or control. Data collection continues until the end of followup. Two clinical trials on prevention of heart attacks report baseline data on weight, shown below. In one of these trials, the randomization did not work. Which one, and why?

		<i>Number of persons</i>	<i>Average weight</i>	<i>SD</i>
(i)	Treatment	1,012	185 lb	25 lb
	Control	997	143 lb	26 lb
(ii)	Treatment	995	166 lb	27 lb
	Control	1,017	163 lb	25 lb

8. One investigator takes a sample of 100 men age 18–24 in a certain town. Another takes a sample of 1,000 such men.
- Which investigator will get a bigger average for the heights of the men in his sample? or should the averages be about the same?
  - Which investigator will get a bigger SD for the heights of the men in his sample? or should the SDs be about the same?
  - Which investigator is likely to get the tallest of the sample men? or are the chances about the same for both investigators?
  - Which investigator is likely to get the shortest of the sample men? or are the chances about the same for both investigators?
9. The men in the HANES5 sample had an average height of 69 inches, and the SD was 3 inches. Tomorrow, one of these men will be chosen at random. You have to guess his height. What should you guess? You have about 1 chance in 3 to be off by more than \_\_\_\_\_. Fill in the blank. Options: 1/2 inch, 3 inches, 5 inches.
10. As in exercise 9, but tomorrow a whole series of men will be chosen at random. After each man appears, his actual height will be compared with your guess to see how far off you were. The r.m.s. size of the amounts off should be \_\_\_\_\_. Fill in the blank. (Hint: Look at the bottom of this page.)

*The answers to these exercises are on pp. A49–50.*

## 6. COMPUTING THE STANDARD DEVIATION

To find the standard deviation of a list, take the entries one at a time. Each deviates from the average by some amount, perhaps 0:

$$\text{deviation from average} = \text{entry} - \text{average}.$$

The SD is the r.m.s. size of these deviations. (Reminder: “r.m.s.” means root-mean-square. See p. 66.)

$$\text{SD} = \text{r.m.s. deviation from average.}$$

*Example 2.* Find the SD of the list 20, 10, 15, 15.

*Solution.* The first step is to find the average:

$$\text{average} = \frac{20 + 10 + 15 + 15}{4} = 15.$$

The second step is to find the deviations from the average: just subtract the average from each entry. The deviations are

$$5 \quad -5 \quad 0 \quad 0$$

The last step is to find the r.m.s. size of the deviations:

$$\begin{aligned} \text{SD} &= \sqrt{\frac{5^2 + (-5)^2 + 0^2 + 0^2}{4}} \\ &= \sqrt{\frac{25 + 25 + 0 + 0}{4}} \\ &= \sqrt{\frac{50}{4}} = \sqrt{12.5} \approx 3.5 \end{aligned}$$

This completes the calculation.

The SD comes out in the same units as the data. For example, suppose heights are measured in inches. The intermediate squaring step in the procedure changes the units to inches squared, but the square root returns the answer to the original units.<sup>11</sup> Do not confuse the SD of a list with its r.m.s. size. The SD is the r.m.s., not of the original numbers on the list, but of their deviations from average.

### Exercise Set E

1. Guess which of the following two lists has the larger SD. Check your guess by computing the SD for both lists.
  - (i) 9, 9, 10, 10, 10, 12
  - (ii) 7, 8, 10, 11, 11, 13
2. Someone is telling you how to calculate the SD of the list 1, 2, 3, 4, 5:  
The average is 3, so the deviations from average are

$$-2 \quad -1 \quad 0 \quad 1 \quad 2$$

Drop the signs. The average deviation is

$$\frac{2 + 1 + 0 + 1 + 2}{5} = 1.2$$

And that's the SD.

Is this right? Answer yes or no, and explain briefly.

3. Someone is telling you how to calculate the SD of the list 1, 2, 3, 4, 5:

The average is 3, so the deviations from average are

$$-2 \quad -1 \quad 0 \quad 1 \quad 2$$

The 0 doesn't count, so the r.m.s. deviation is

$$\sqrt{\frac{4 + 1 + 1 + 4}{4}} = 1.6$$

And that's the SD.

Is this right? Answer yes or no, and explain briefly.

4. Three instructors are comparing scores on their finals; each had 99 students. In class A, one student got 1 point, another got 99 points, and the rest got 50 points. In class B, 49 students got a score of 1, one student got a score of 50, and 49 students got a score of 99. In class C, one student got a score of 1, one student got a score of 2, one student got a score of 3, and so forth, all the way through 99.
- (a) Which class had the biggest average? or are they the same?
  - (b) Which class had the biggest SD? or are they the same?
  - (c) Which class had the biggest range? or are they the same?
5. (a) For each list below, work out the average, the deviations from average, and the SD.
- (i) 1, 3, 4, 5, 7
  - (ii) 6, 8, 9, 10, 12
- (b) How is list (ii) related to list (i)? How does this relationship carry over to the average? the deviations from the average? the SD?
6. Repeat exercise 5 for the following two lists:
- (i) 1, 3, 4, 5, 7
  - (ii) 3, 9, 12, 15, 21
7. Repeat exercise 5 for the following two lists:
- (i) 5, -4, 3, -1, 7
  - (ii) -5, 4, -3, 1, -7
8. (a) The Governor of California proposes to give all state employees a flat raise of \$250 a month. What would this do to the average monthly salary of state employees? to the SD?
- (b) What would a 5% increase in the salaries, across the board, do to the average monthly salary? to the SD?
9. What is the r.m.s. size of the list 17, 17, 17, 17, 17? the SD?
10. For the list 107, 98, 93, 101, 104, which is smaller—the r.m.s. size or the SD? No arithmetic is needed.
11. Can the SD ever be negative?
12. For a list of positive numbers, can the SD ever be larger than the average?

*The answers to these exercises are on pp. A50–51.*

*Technical note.* There is an alternative way to compute the SD, which is more efficient in some cases.<sup>12</sup>

$$\text{SD} = \sqrt{\text{average of (entries}^2) - (\text{average of entries})^2}.$$

## 7. USING A STATISTICAL CALCULATOR

Most statistical calculators produce not the SD, but the slightly larger number  $\text{SD}^+$ . (The distinction between SD and  $\text{SD}^+$  will be explained more carefully in section 6 of chapter 26.) To find out what your machine is doing, put in the list  $-1, 1$ . If the machine gives you 1, it's working out the SD. If it gives you  $1.41\dots$ , it's working out the  $\text{SD}^+$ . If you're getting the  $\text{SD}^+$  and you want the SD, you have to multiply by a conversion factor. This depends on the number of entries on the list. With 10 entries, the conversion factor is  $\sqrt{9/10}$ . With 20 entries, it is  $\sqrt{19/20}$ . In general,

$$\text{SD} = \sqrt{\frac{\text{number of entries} - 1}{\text{number of entries}}} \times \text{SD}^+$$

## 8. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. (a) Find the average and SD of the list 41, 48, 50, 50, 54, 57.  
 (b) Which numbers on the list are within 0.5 SDs of average? within 1.5 SDs of average?
2. (a) Both of the following lists have the same average of 50. Which one has the smaller SD, and why? No computations are necessary.  
 (i) 50, 40, 60, 30, 70, 25, 75  
 (ii) 50, 40, 60, 30, 70, 25, 75, 50, 50, 50  
 (b) Repeat, for the following two lists.  
 (i) 50, 40, 60, 30, 70, 25, 75  
 (ii) 50, 40, 60, 30, 70, 25, 75, 99, 1
3. Here is a list of numbers:  

$$\begin{array}{cccccccccc} 0.7 & 1.6 & 9.8 & 3.2 & 5.4 & 0.8 & 7.7 & 6.3 & 2.2 & 4.1 \\ 8.1 & 6.5 & 3.7 & 0.6 & 6.9 & 9.9 & 8.8 & 3.1 & 5.7 & 9.1 \end{array}$$
  
 (a) Without doing any arithmetic, guess whether the average is around 1, 5, or 10.  
 (b) Without doing any arithmetic, guess whether the SD is around 1, 3, or 6.
4. For persons age 25 and over in the U.S., would the average or the median be higher for income? for years of schooling completed?

5. For the men age 18–24 in HANES5, the average systolic blood pressure was 116 mm and the SD was 11 mm.<sup>13</sup> Say whether each of the following blood pressures is unusually high, unusually low, or about average:

80 mm

115 mm

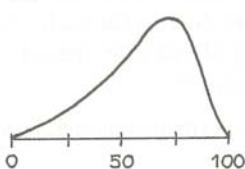
120 mm

210 mm

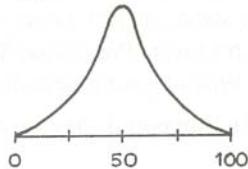
6. Below are sketches of histograms for three lists.

- In scrambled order, the averages are 40, 50, 60. Match the histograms with the averages.
- Match the histogram with the description:
  - the median is less than the average
  - the median is about equal to the average
  - the median is bigger than the average
- Is the SD of histogram (iii) around 5, 15, or 50?
- True or false, and explain: the SD for histogram (i) is a lot smaller than that for histogram (iii).

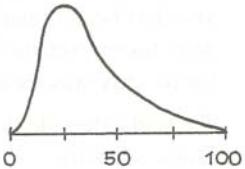
(i)



(ii)



(iii)



7. A study on college students found that the men had an average weight of about 66 kg and an SD of about 9 kg. The women had an average weight of about 55 kg and an SD of 9 kg.

- Find the averages and SDs, in pounds (1 kg = 2.2 lb).
- Just roughly, what percentage of the men weighed between 57 kg and 75 kg?
- If you took the men and women together, would the SD of their weights be smaller than 9 kg, just about 9 kg, or bigger than 9 kg? Why?

8. In the HANES5 sample, the average height of the boys was 137 cm at age 9 and 151 cm at age 11. At age 11, the average height of all the children was 151 cm.<sup>14</sup>

- On the average, are boys taller than girls at age 11?
- Guess the average height of the 10-year-old boys.

9. An investigator has a computer file showing family incomes for 1,000 subjects in a certain study. These range from \$5,800 a year to \$98,600 a year. By accident, the highest income in the file gets changed to \$986,000.

- Does this affect the average? If so, by how much?
- Does this affect the median? If so, by how much?

10. Incoming students at a certain law school have an average LSAT (Law School Aptitude Test) score of 163 and an SD of 8. Tomorrow, one of these students

will be picked at random. You have to guess the score now; the guess will be compared with the actual score, to see how far off it is. Each point off will cost a dollar. (For example, if the guess is 158 and the score is really 151, you will have to pay \$7.)

- (a) Is the best guess 150, 163, or 170?
- (b) You have about 1 chance in 3 to lose more than \_\_\_\_\_. Fill in the blank. Options: \$1, \$8, \$20.

(LSAT scores range from 120 to 180; the average across all test-takers is about 150 and the SD is about 9. The test is re-normed from time to time, the data are for 2005.)

11. As in exercise 10, but a whole series of students are chosen. The r.m.s. size of your losses should be around \_\_\_\_\_. Fill in the blank.
12. Many observers think there is a permanent underclass in American society—most of those in poverty typically remain poor from year to year. Over the period 1970–2000, the percentage of the American population in poverty each year has been remarkably stable, at 12% or so. Income figures for each year were taken from the March Current Population Survey of that year; the cutoff for poverty was based on official government definitions.<sup>15</sup>

To what extent do these data support the theory of the permanent underclass? Discuss briefly.

## 9. SUMMARY

1. A typical list of numbers can be summarized by its *average* and *standard deviation* (SD).
2. Average of a list = 
$$\frac{\text{sum of entries}}{\text{number of entries}}$$
.
3. The average locates the center of a histogram, in the sense that the histogram balances when supported at the average.



Drawing by Dana Fradon; © 1976 The New Yorker Magazine, Inc.

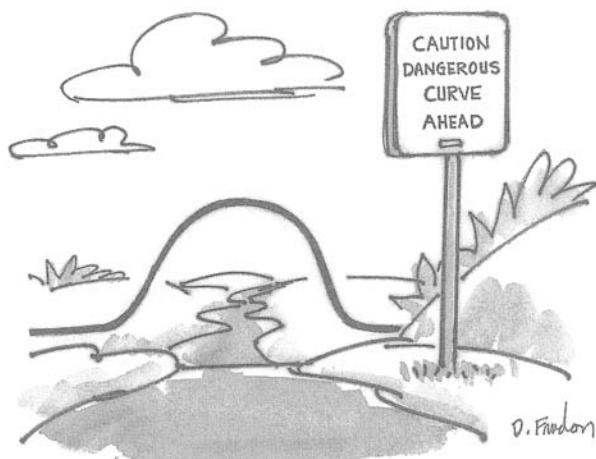
4. Half the area under a histogram lies to the left of the *median*, and half to the right. The median is another way to locate the center of a histogram.
5. The *r.m.s. size* of a list measures how big the entries are, neglecting signs.
$$6. \text{ r.m.s. size of a list} = \sqrt{\text{average of } (\text{entries}^2)}.$$
7. The SD measures distance from the average. Each number on a list is off the average by some amount. The SD is a sort of average size for these amounts off. More technically, the SD is the r.m.s. size of the deviations from the average.
8. Roughly 68% of the entries on a list of numbers are within one SD of the average, and about 95% are within two SDs of the average. This is so for many lists, but not all.
9. If a study draws conclusions about the effects of age, find out whether the data are cross-sectional or longitudinal.

# 5

## The Normal Approximation for Data

### 1. THE NORMAL CURVE

The normal curve was discovered around 1720 by Abraham de Moivre, while he was developing the mathematics of chance. (His work will be discussed again in parts IV and V.) Around 1870, the Belgian mathematician Adolph Quetelet had the idea of using the curve as an ideal histogram, to which histograms for data could be compared.

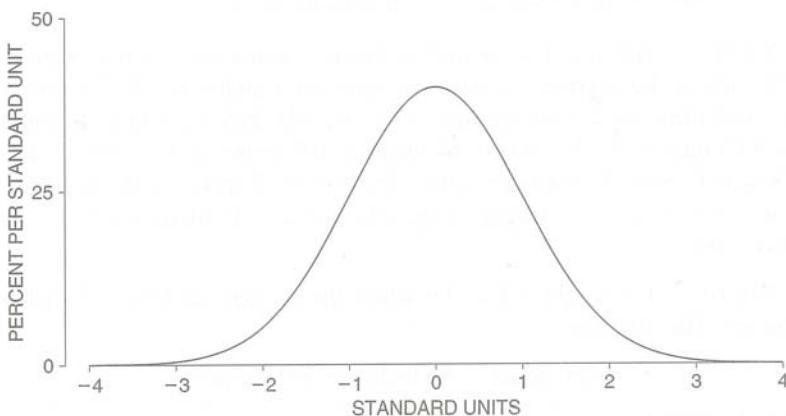


The normal curve has a formidable-looking equation:

$$y = \frac{100\%}{\sqrt{2\pi}} e^{-x^2/2}, \text{ where } e = 2.71828 \dots$$

This equation involves three of the most famous numbers in the history of mathematics:  $\sqrt{2}$ ,  $\pi$ , and  $e$ . This is just to show off a little. You will find it is easy to work with the normal curve through diagrams and tables, without ever using the equation. A graph of the curve is shown in figure 1.

Figure 1. The normal curve.



Several features of this graph will be important. First, the graph is symmetric about 0: the part of the curve to the right of 0 is a mirror image of the part to the left. Next, the total area under the curve equals 100%. (Areas come out in percent, because the vertical axis uses the density scale.) Finally, the curve is always above the horizontal axis. It appears to stop between 3 and 4, but that's only because the curve gets so low there. Only about 6/100,000 of the area is outside the interval from -4 to 4.

It will be helpful to find areas under the normal curve between specified values. For instance,

- the area under the normal curve between  $-1$  and  $+1$  is about 68%;
- the area under the normal curve between  $-2$  and  $+2$  is about 95%;
- the area under the normal curve between  $-3$  and  $+3$  is about 99.7%.

Finding these areas is a matter of looking things up in a table, or pushing a button on the right kind of calculator; the table will be explained in section 2.

Many histograms for data are similar in shape to the normal curve, provided they are drawn to the same scale. Making the horizontal scales match up involves *standard units*.<sup>1</sup>

A value is converted to standard units by seeing how many SDs it is above or below the average.

Values above the average are given a plus sign; values below the average get a minus sign. The horizontal axis of figure 1 is in standard units.

For instance, take the women age 18 and over in the HANES5 sample. Their average height was 63.5 inches; the SD was 3 inches. One of these women was 69.5 inches tall. What was her height in standard units? Our subject was 6 inches taller than average, and 6 inches is 2 SDs. In standard units, her height was +2.

*Example 1.* For women age 18 and over in the HANES5 sample—

- (a) Convert the following to standard units:
  - (i) 66.5 inches (ii) 57.5 inches (iii) 64 inches (iv) 63.5 inches
- (b) Find the height which is -1.2 in standard units.

*Solution.* Part (a). For (i), 66.5 inches is 3 inches above the average. That is 1 SD above the average. In standard units, 66.5 inches is +1. For (ii), 57.5 inches is 6 inches below the average. That is 2 SDs below average. In standard units, 57.5 inches is -2. For (iii), 64 inches is 0.5 inches above average. That is  $0.5/3 \approx 0.17$  SDs. The answer is 0.17. For (iv), 63.5 inches is the average. So, 63.5 inches is 0 SDs away from average. The answer is 0. (Reminder: “ $\approx$ ” means “nearly equal.”)

Part (b). The height is 1.2 SDs below the average, and  $1.2 \times 3$  inches = 3.6 inches. The height is

$$63.5 \text{ inches} - 3.6 \text{ inches} = 59.9 \text{ inches.}$$

That is the answer.

Standard units are used in figure 2. In this figure, the histogram for the heights of the women age 18 and over in the HANES5 sample is compared to the normal curve. The horizontal axis for the histogram is in inches; the horizontal axis for the normal curve is in standard units. The two match up as indicated in example 1. For instance, 66.5 inches is directly above +1, and 57.5 inches is directly above -2.

There are also two vertical axes in figure 2. The histogram is drawn relative to the inside one, in percent per inch. The normal curve is drawn relative to the outside one, in percent per standard unit. To see how the scales match up, take the top value on each axis: 60% per standard unit matches 20% per inch because there are 3 inches to the standard unit. Spreading 60% over an SD is the same as spreading 60% over 3 inches, and that comes to 20% per inch—

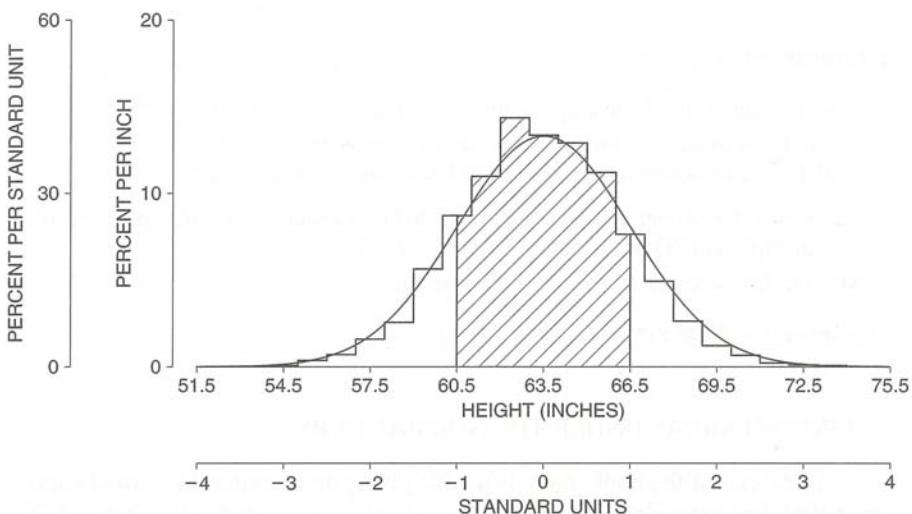
$$\begin{aligned} 60\% \text{ per standard unit} &= 60\% \text{ per 3 inches} \\ &= 60\% \div 3 \text{ inches} = 20\% \text{ per inch.} \end{aligned}$$

Similarly, 30% per standard unit matches 10% per inch. Any other pair of values can be dealt with in the same way.

The last chapter said that for many lists, roughly 68% of the entries are within one SD of average. This is the range

$$\text{average} - \text{SD} \text{ to } \text{average} + \text{SD}.$$

Figure 2. A histogram for heights of women compared to the normal curve. The area under the histogram between 60.5 inches and 66.5 inches (the percentage of women within one SD of average with respect to height) is about equal to the area between  $-1$  and  $+1$  under the curve—68%.

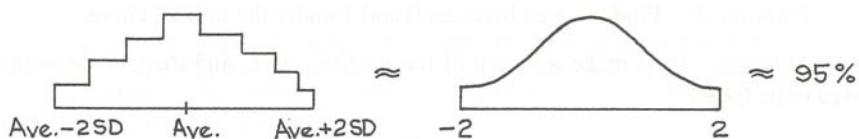


To see where the 68% comes from, look at figure 2. The percentage of women whose heights are within one SD of average equals the area under the histogram within one SD of average. This area is shaded in figure 2. The histogram follows the normal curve fairly well. Parts of it are higher than the curve, and parts of it are lower. But the highs balance out the lows. And the shaded area under the histogram is about the same as the area under the curve. The area under the normal curve between  $-1$  and  $+1$  is 68%. That is where the 68% comes from.

For many lists, roughly 95% of the entries are within 2 SDs of average. This is the range

$$\text{average} - 2\text{SDs} \text{ to } \text{average} + 2\text{SDs}.$$

The reasoning is similar. If the histogram follows the normal curve, the area under the histogram will be about the same as the area under the curve. And the area under the curve between  $-2$  and  $+2$  is 95%:



The normal curve can be used to estimate the percentage of entries in an interval, as follows.<sup>2</sup> First, convert the interval to standard units; second, find the

corresponding area under the normal curve. The method for getting areas will be explained in section 2. Finally, section 3 will put the two steps together. The whole procedure is called the *normal approximation*. The approximation consists in replacing the original histogram by the normal curve before finding the area.

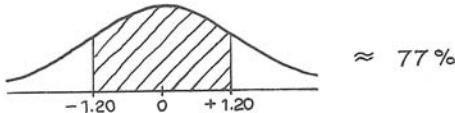
### Exercise Set A

1. On a certain exam, the average of the scores was 50 and the SD was 10.
  - (a) Convert each of the following scores to standard units: 60, 45, 75.
  - (b) Find the scores which in standard units are: 0, +1.5, -2.8.
2. (a) Convert each entry on the following list to standard units (that is, using the average and SD of the list): 13, 9, 11, 7, 10.  
 (b) Find the average and SD of the converted list.

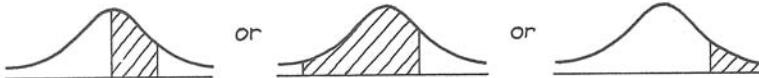
*The answers to these exercises are on p. A51.*

### 2. FINDING AREAS UNDER THE NORMAL CURVE

At the end of the book, there is a table giving areas under the normal curve (p. A104). For example, to find the area under the normal curve between  $-1.20$  and  $1.20$ , go to  $1.20$  in the column marked  $z$  and read off the entry in the column marked *Area*. This is about 77%, so the area under the normal curve between  $-1.20$  and  $1.20$  is about 77%.



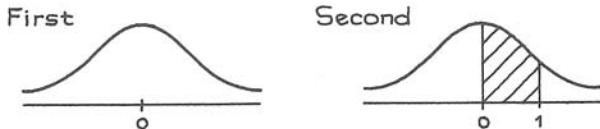
But you are also going to want to find other areas:



The method for finding such areas is indicated by example.

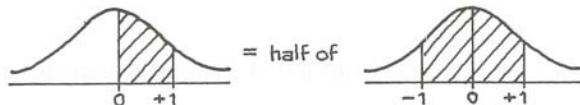
*Example 2.* Find the area between 0 and 1 under the normal curve.

*Solution.* First make a sketch of the normal curve, and then shade in the area to be found.



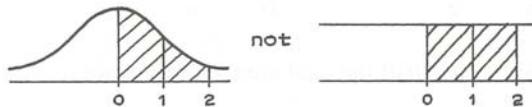
The table will give you the area between  $-1$  and  $+1$ . This is about 68%. By symmetry, the area between  $0$  and  $1$  is half the area between  $-1$  and  $+1$ , that is,

$$\frac{1}{2} \times 68\% = 34\%$$



*Example 3.* Find the area between  $0$  and  $2$  under the normal curve.

*Solution.* This isn't double the area between  $0$  and  $1$  because the normal curve isn't a rectangle.

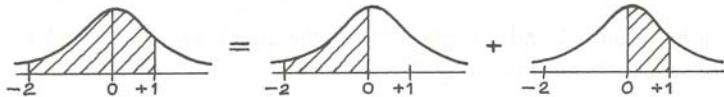


The procedure is the same as in example 2. The area between  $-2$  and  $2$  can be found from the table. It is about 95%. The area between  $0$  and  $2$  is half that, by symmetry:

$$\frac{1}{2} \times 95\% \approx 48\%.$$

*Example 4.* Find the area between  $-2$  and  $1$  under the normal curve.

*Solution.* The area between  $-2$  and  $1$  can be broken down into two other areas—

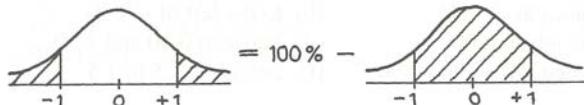


The area between  $-2$  and  $0$  is the same as the area between  $0$  and  $2$ , by symmetry, and is about 48% (example 3). The area between  $0$  and  $1$  is about 34% (example 2). The area between  $-2$  and  $1$  is about

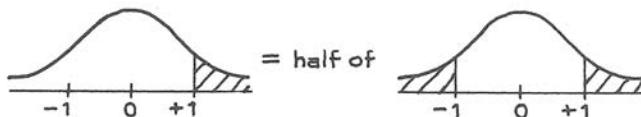
$$48\% + 34\% = 82\%.$$

*Example 5.* Find the area to the right of  $1$  under the normal curve.

*Solution.* The table gives the area between  $-1$  and  $1$ , which is 68%. The area outside this interval is 32%.

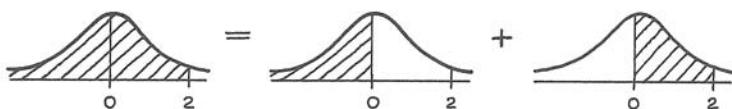


By symmetry, the area to the right of 1 is half this, or 16%.



*Example 6.* Find the area to the left of 2 under the normal curve.

*Solution.* The area to the left of 2 is the sum of the area to the left of 0, and the area between 0 and 2.



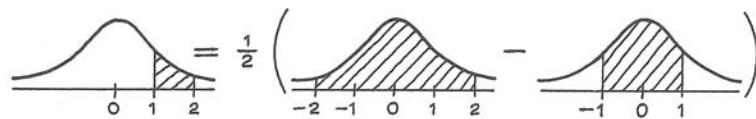
The area to the left of 0 is half the total area, by symmetry:

$$\frac{1}{2} \times 100\% = 50\%$$

The area between 0 and 2 is about 48%. The sum is  $50\% + 48\% = 98\%$ .

*Example 7.* Find the area between 1 and 2 under the normal curve.

*Solution.*



The area between  $-2$  and  $2$  is about 95%; the area between  $-1$  and  $1$  is about 68%. Half the difference is

$$\frac{1}{2} \times (95\% - 68\%) = \frac{1}{2} \times 27\% \approx 14\%.$$

There is no set procedure to use in solving this sort of problem. It is a matter of drawing pictures which relate the area you want to areas that can be read from the table.

### Exercise Set B

1. Find the area under the normal curve—

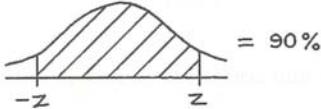
- |                              |                            |
|------------------------------|----------------------------|
| (a) to the right of 1.25     | (b) to the left of $-0.40$ |
| (c) to the left of 0.80      | (d) between 0.40 and 1.30  |
| (e) between $-0.30$ and 0.90 | (f) outside $-1.5$ to 1.5  |

2. Fill in the blanks:

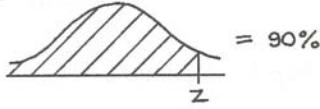
- (a) The area between  $\pm \underline{\hspace{1cm}}$  under the normal curve equals 68%.
- (b) The area between  $\pm \underline{\hspace{1cm}}$  under the normal curve equals 75%.

3. The normal curve is sketched below; solve for  $z$ .

(a)

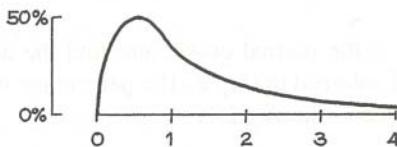


(b)



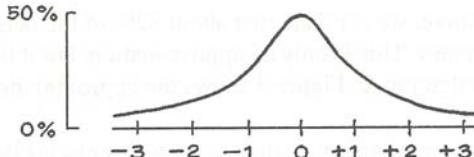
4. A certain curve (not the normal) is sketched below. The total area under it is 100%, and the area between 0 and 1 is 39%.

- (a) If possible, find the area to the right of 1.
- (b) If possible, find the area between 0 and 0.5.



5. A certain curve (not the normal) is sketched below. It is symmetric around 0, and the total area under it is 100%. The area between  $-1$  and  $1$  is 58%.

- (a) If possible, find the area between 0 and 1.
- (b) If possible, find the area to the right of 1.
- (c) If possible, find the area to the right of 2.



*The answers to these exercises are on p. A51.*

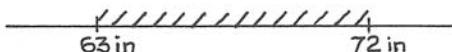
### 3. THE NORMAL APPROXIMATION FOR DATA

The method for the normal approximation will be explained here by example. The diagrams look so simple that you may not think they are worth drawing. However, it is easy to lose track of the area that is wanted. Please draw the diagrams.

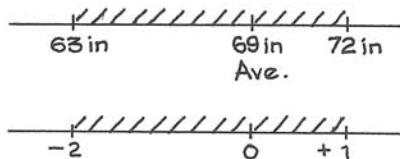
*Example 8.* The heights of the men age 18 and over in HANES5 averaged 69 inches; the SD was 3 inches. Use the normal curve to estimate the percentage of these men with heights between 63 inches and 72 inches.

*Solution.* The percentage is given by the area under the height histogram, between 63 inches and 72 inches.

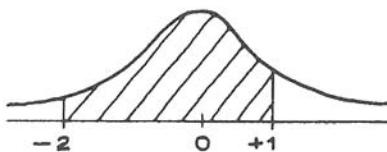
*Step 1.* Draw a number line and shade the interval.



*Step 2.* Mark the average on the line and convert to standard units.

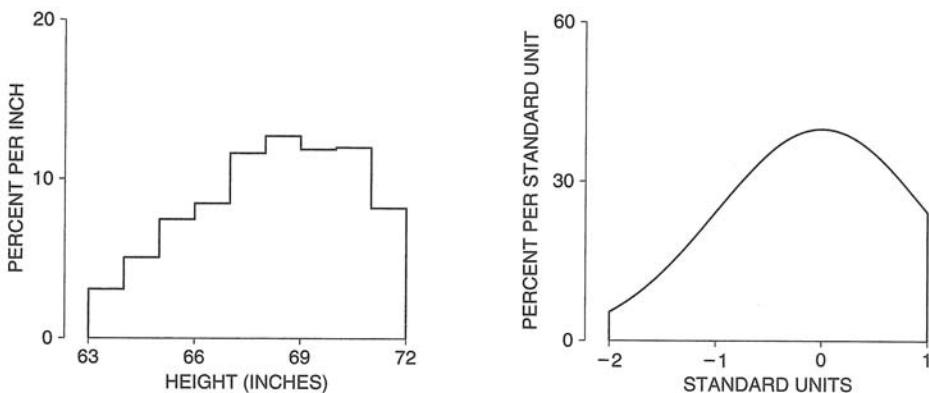


*Step 3.* Sketch in the normal curve, and find the area above the shaded standard-units interval obtained in step 2. The percentage is approximately equal to the shaded area, which is almost 82%.



Using the normal curve, we estimate that about 82% of the heights were between 63 inches and 72 inches. This is only an approximation, but it is pretty good: 81% of the men were in that range. Figure 3 shows the approximation.

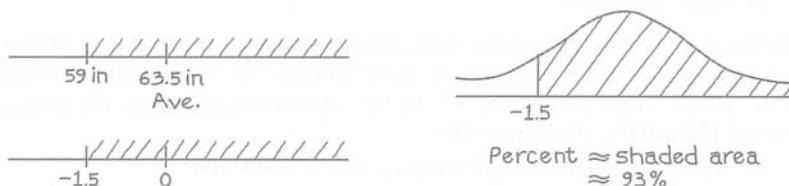
Figure 3. The normal approximation consists in replacing the original histogram by the normal curve before computing areas.



*Example 9.* The heights of the women age 18 and over in HANES5 averaged 63.5 inches; the SD was 3 inches. Use the normal curve to estimate the percentage with heights above 59 inches.

*Solution.* A height of 59 inches is 1.5 SDs below average:

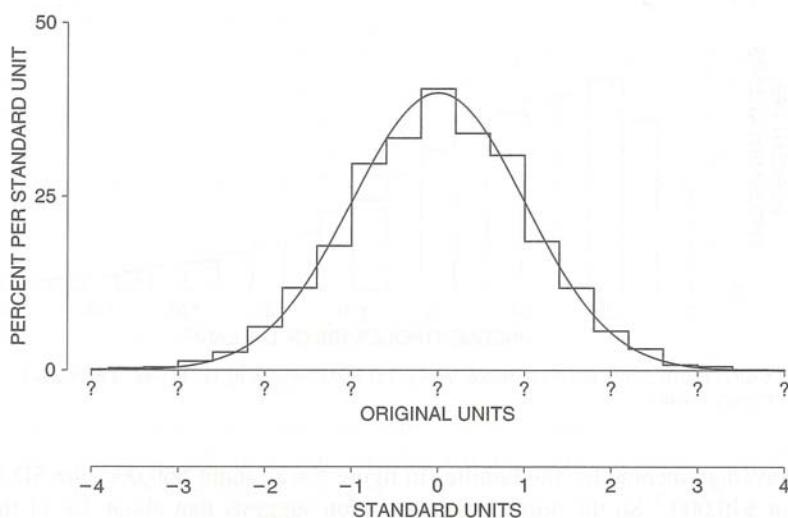
$$(59 - 63.5)/3 = -1.5.$$



Using the normal curve, we estimate that 93% of the women were more than 59 inches in height. This estimate is about right: 96% of the women were taller than 59 inches.

It is a remarkable fact that many histograms follow the normal curve. (The story continues in part V.) For such histograms, the average and SD are good summary statistics. If a histogram follows the normal curve, it looks something like the sketch in figure 4. The average pins down the center, and the SD gives the spread. That is nearly all there is to say about the histogram—if its shape is like the normal curve. Many other histograms, however, do not follow the normal curve. In such cases, the average and SD are poor summary statistics. More about this in the next section.

Figure 4. The average and SD. By locating the center and measuring the spread around the center, the average and SD summarize a histogram which follows the normal curve.



### Exercise Set C

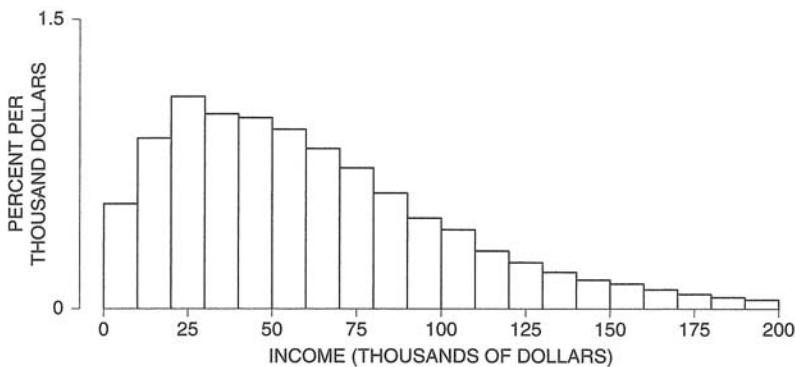
1. For the women age 18–24 in HANES2, the average height was about 64.3 inches; the SD was about 2.6 inches. Using the normal curve, estimate the percentage of women with heights—
  - (a) below 66 inches.
  - (b) between 60 inches and 66 inches.
  - (c) above 72 inches.
2. In a law school class, the entering students averaged about 160 on the LSAT; the SD was about 8. The histogram of LSAT scores followed the normal curve reasonably well. (LSAT scores range from 120 to 180; among all test-takers, the average is around 150 and the SD is around 9.)
  - (a) About what percentage of the class scored below 166?
  - (b) One student was 0.5 SDs above average on the LSAT. About what percentage of the students had lower scores than he did?
3. In figure 2 (p. 81), the percentage of women with heights between 61 inches and 66 inches is exactly equal to the area between 61 inches and 66 inches under the \_\_\_\_\_ and approximately equal to the area under the \_\_\_\_\_. Options: normal curve, histogram.

*The answers to these exercises are on pp. A51–52.*

### 4. PERCENTILES

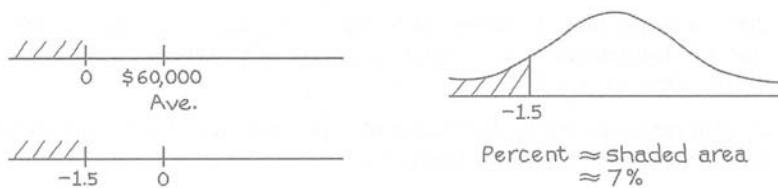
The average and SD can be used to summarize data following the normal curve. They are less satisfactory for other kinds of data. Take the distribution of family income in the U.S. in 2004, shown in figure 5.

Figure 5. Distribution of families by income: the U.S. in 2004.



Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census.  
Primary families.

The average income for the families in figure 5 was about \$60,000; the SD was about \$40,000.<sup>3</sup> So the normal approximation suggests that about 7% of these families had negative incomes:



The reason for this blunder: the histogram in figure 5 does not follow the normal curve at all well, it has a long right-hand tail. To summarize such histograms, statisticians often use *percentiles* (table 1).

Table 1. Selected percentiles for family income in the U.S. in 2004.

1	\$0
10	\$15,000
25	\$29,000
50	\$54,000
75	\$90,000
90	\$135,000
99	\$430,000

Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census. Primary families.

The 1st percentile of the income distribution was \$0, meaning that about 1% of the families had incomes of \$0 or less, and about 99% had incomes above that level. (Mainly, the families with no income were retired or not working for some other reason.) The 10th percentile was \$15,000: about 10% of the families had incomes below that level, and 90% were above. The 50th percentile is just the median (chapter 4).

By definition, the *interquartile range* equals

$$75\text{th percentile} - 25\text{th percentile}.$$

This is sometimes used as a measure of spread, when the distribution has a long tail. For table 1, the interquartile range is \$61,000.

For reasons of their own, statisticians call de Moivre's curve "normal." This gives the impression that other curves are abnormal. Not so. Many histograms follow the normal curve very well, and many others—like the income histogram—do not. Later in the book, we will present a mathematical theory which helps explain when histograms should follow the normal curve.

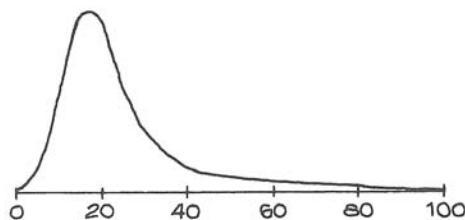
### Exercise Set D

1. Fill in the blanks, using the options below.

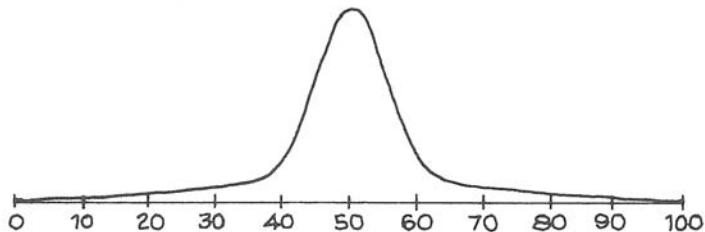
- (a) The percentage of families in table 1 with incomes below \$90,000 was about \_\_\_\_\_.
- (b) About 25% of the families in table 1 had incomes below \_\_\_\_\_.
- (c) The percentage of families in table 1 with incomes between \$15,000 and \$125,000 was about \_\_\_\_\_.

5%    10%    25%    60%    75%    95%    \$29,000    \$90,000

2. In 2004, a family with an income of \$9,000 was at the \_\_\_\_\_th percentile of the income distribution, while a family that made \$174,000 was at the \_\_\_\_\_th percentile. Options: 5, 95.
3. Is the 25th percentile for the distribution of family income in 1973 around \$7,000, \$10,000, or \$25,000? (See table 1 on p. 35.)
4. Skinfold thickness is used to measure body fat. A histogram for skinfold thickness is shown below; the units on the horizontal axis are millimeters (mm). The 25th percentile of skinfold thickness is \_\_\_\_\_ 25 mm. Fill in the blank, using one of the phrases below. Or can this be determined from the figure?
- quite a bit smaller than  
around  
quite a bit bigger than



5. A histogram is sketched below.
- How is it different from the normal curve?
  - Is the interquartile range around 15, 25, or 50?



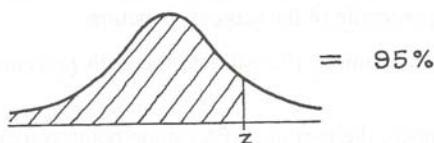
*The answers to these exercises are on p. A52.*

## 5. PERCENTILES AND THE NORMAL CURVE

When a histogram does follow the normal curve, the table can be used to estimate its percentiles. The method is indicated by example.

*Example 10.* Among all applicants to a certain university one year, the Math SAT scores averaged 535, the SD was 100, and the scores followed the normal curve. Estimate the 95th percentile of the score distribution.

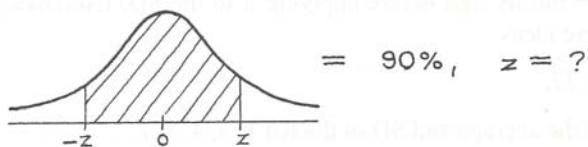
*Solution.* This score is above average, by some number of SDs. We need to find that number, call it  $z$ . There is an equation for  $z$ :



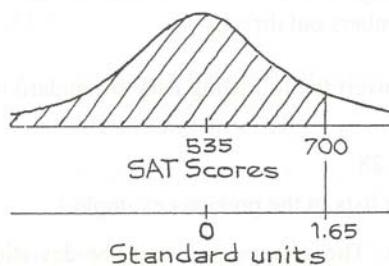
The normal table cannot be used directly, because it gives the area between  $-z$  and  $z$  rather than the area to the left of  $z$ .



The area to the right of our  $z$  is 5%, so the area to the left of  $-z$  is 5% too. Then the area between  $-z$  and  $z$  must be  $100\% - 5\% - 5\% = 90\%$ .



From the table,  $z \approx 1.65$ . You have to score 1.65 SDs above average to be in the 95th percentile of the Math SAT. Translated back to points, this score is above average by  $1.65 \times 100 = 165$  points. The 95th percentile of the score distribution is  $535 + 165 = 700$ .



The terminology is a little confusing. A *percentile* is a score: in example 10, the 95th percentile is a score of 700. A *percentile rank*, however, is a percent: if you score 700, your percentile rank is 95%. There is even a third way to say the same thing: a score of 700 puts you at the 95th percentile of the score distribution.

### Exercise Set E

- At the university in example 10, one applicant scored 750 on the Math SAT. She was at the \_\_\_\_\_ percentile of the score distribution.
- For the university in example 10, estimate the 80th percentile of the Math SAT scores.
- For Berkeley freshmen, the average GPA (grade point average) is around 3.0; the SD is about 0.5. The histogram follows the normal curve. Estimate the 30th percentile of the GPA distribution.

*The answers to these exercises are on p. A52.*

### 6. CHANGE OF SCALE

If you add the same number to every entry on a list, that number just gets added to the average; the SD does not change. (The deviations from the average do not change, because the added constant just cancels.) Furthermore, if you multiply every entry on a list by the same number, the average and the SD simply get multiplied by that number. There is one exception: if that constant multiplier is negative, wipe out its sign before applying it to the SD. Exercises 5–8 on p. 73 illustrated these ideas.

*Example 11.*

- Find the average and SD of the list 1, 3, 4, 5, 7.
- Take the list in part (a), multiply each entry by 3 and then add 7, to get the list 10, 16, 19, 22, 28. Find the average and SD of this new list.

*Solution. Part (a).* The average is 4. So the deviations from average are  $-3, -1, 0, 1, 3$ . The SD is 2.

*Part (b).* The average is  $3 \times 4 + 7 = 19$ , the SD is  $3 \times 2 = 6$ . (Of course, you can work these numbers out directly.)

*Example 12.* Convert the following lists to standard units:

- 1, 3, 4, 5, 7
- 10, 16, 19, 22, 28

(These are the two lists in the previous example.)

*Solution. Part (a).* The average is 4, and the deviations from average are  $-3, -1, 0, 1, 3$ . The SD is 2. Divide by 2 to get the list in standard units:

$$-1.5 \quad -0.5 \quad 0 \quad 0.5 \quad 1.5$$

*Part (b).* Now the average is 19, and the deviations from average are  $-9, -3, 0, 3, 9$ . The SD is 6. Divide by 6 to get the list in standard units:

$$-1.5 \quad -0.5 \quad 0 \quad 0.5 \quad 1.5$$

The two lists are the same in standard units.

List (b) comes from list (a) by changing the scale: multiply by 3, add 7. The 7 washes out when computing the deviations from average. The 3 washes out when dividing by the SD—because the SD got multiplied by 3 along with all the deviations. That is why the lists are the same in standard units. To summarize:

- (i) Adding the same number to every entry on a list adds that constant to the average; the SD does not change.
- (ii) Multiplying every entry on a list by the same positive number multiplies the average and the SD by that constant.
- (iii) These changes of scale do not change the standard units.

Conversion of temperature from Fahrenheit to Celsius is a practical example:

$$C^\circ = \frac{5}{9}(F^\circ - 32^\circ)$$

Statisticians call this a *change of scale*, because it is only the units that change. (What happens if you multiply all the numbers on a list by the same negative constant? In standard units, that just reverses all the signs.)

### Exercise Set F

1. A group of people have an average temperature of 98.6 degrees Fahrenheit, with an SD of 0.3 degrees.
  - (a) Translate these results into degrees Celsius.
  - (b) Someone's temperature is 1.5 SDs above average on the Fahrenheit scale. Convert this temperature to standard units, for an investigator who is using the Celsius scale.

*The answers to these exercises are on p. A52.*

## 7. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. The following list of test scores has an average of 50 and an SD of 10:

39	41	47	58	65	37	37	49	56	59	62	36	48
52	64	29	44	47	49	52	53	54	72	50	50	

- (a) Use the normal approximation to estimate the number of scores within 1.25 SDs of the average.
  - (b) How many scores really were within 1.25 SDs of the average?
2. You are looking at a computer printout of 100 test scores, which have been converted to standard units. The first 10 entries are

−6.2 3.5 1.2 −0.13 4.3 −5.1 −7.2 −11.3 1.8 6.3

Does the printout look reasonable, or is something wrong with the computer?

3. From the mid-1960s to the early 1990s, there was a slow but steady decline in SAT scores. For example, take the Verbal SAT. The average in 1967 was about 543; by 1994, the average was down to about 499. However, the SD stayed close to 110. The drop in averages has a large effect on the tails of the distribution.
- Estimate the percentage of students scoring over 700 in 1967.
  - Estimate the percentage of students scoring over 700 in 1994.

You may assume that the histograms follow the normal curve.

*Comments.* SAT scores range from 200 to 800. It does not seem that the SAT was getting harder. Most of the decline in the 1960s is thought to result from changes in the population of students taking the test. The decline in the 1970s cannot be explained that way. From 1994 to 2005, scores generally increased. The test was re-normalized in 1996, which complicates the interpretation; the averages mentioned above were converted to the new scale.<sup>4</sup>

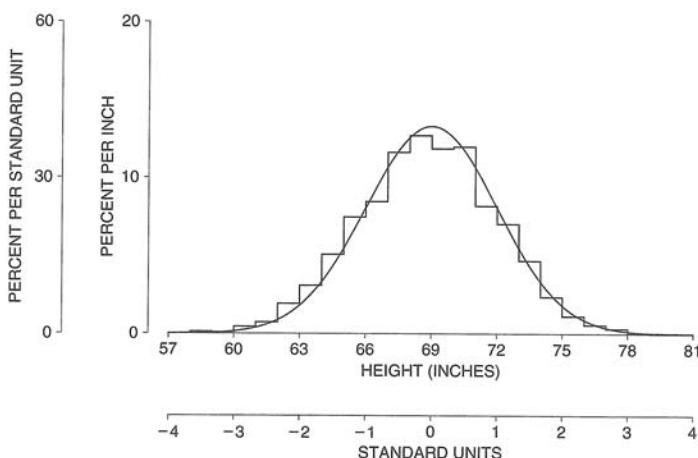
4. On the Math SAT, men have a distinct edge. In 2005, for instance, the men averaged about 538, and the women averaged about 504.
- Estimate the percentage of men getting over 700 on this test in 2005.
  - Estimate the percentage of women getting over 700 on this test in 2005.

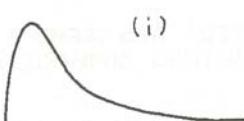
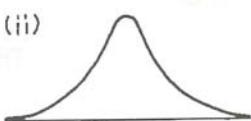
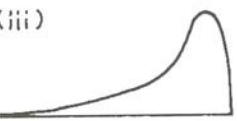
You may assume (i) the histograms followed the normal curve, and (ii) both SDs were about 120.<sup>4</sup>

5. In HANES5, the men age 18 and over had an average height of 69 inches and an SD of 3 inches. The histogram is shown below, with a normal curve. The percentage of men with heights between 66 inches and 72 inches is exactly equal to the area between (a) and (b) under the (c). This percentage is approximately equal to the area between (d) and (e) under the (f). Fill in the blanks. For (a), (b), (d) and (e), your options are

66 inches      72 inches       $-1$        $+1$

For (c) and (f), your options are: normal curve, histogram



6. Among applicants to one law school, the average LSAT score was about 169, the SD was about 9, and the highest score was 178. Did the LSAT scores follow the normal curve?
7. Among freshmen at a certain university, scores on the Math SAT followed the normal curve, with an average of 550 and an SD of 100. Fill in the blanks; explain briefly.
- A student who scored 400 on the Math SAT was at the \_\_\_\_\_ th percentile of the score distribution.
  - To be at the 75th percentile of the distribution, a student needed a score of about \_\_\_\_\_ points on the Math SAT.
8. True or false, and explain briefly—
- If you add 7 to each entry on a list, that adds 7 to the average.
  - If you add 7 to each entry on a list, that adds 7 to the SD.
  - If you double each entry on a list, that doubles the average.
  - If you double each entry on a list, that doubles the SD.
  - If you change the sign of each entry on a list, that changes the sign of the average.
  - If you change the sign of each entry on a list, that changes the sign of the SD.
9. Which of the following are true? false? Explain or give examples.
- The median and the average of any list are always close together.
  - Half of a list is always below average.
  - With a large, representative sample, the histogram is bound to follow the normal curve quite closely.
  - If two lists of numbers have exactly the same average of 50 and the same SD of 10, then the percentage of entries between 40 and 60 must be exactly the same for both lists.
10. For women age 25–34 with full time jobs, the average income in 2004 was \$32,000. The SD was \$26,000, and 1/4 of 1% had incomes above \$150,000. Was the percentage with incomes in the range from \$32,000 to \$150,000 about 40%, 50%, or 60%? Choose one option and explain briefly.<sup>5</sup>
11. One term, about 700 Statistics 2 students at the University of California, Berkeley, were asked how many college mathematics courses they had taken, other than Statistics 2. The average number of courses was about 1.1; the SD was about 1.5. Would the histogram for the data look like (i), (ii), or (iii)? Why?
- (i) 
- (ii) 
- (iii) 
12. In 2005, the average score on the Math SAT was about 520. However, among students who took a subject-matter test, the average score on the Math SAT was about 624.<sup>6</sup> What accounts for the difference?

## 8. SUMMARY

1. The *normal curve* is symmetric about 0, and the total area under it is 100%.
2. *Standard units* say how many SDs a value is, above (+) or below (-) the average.
3. Many histograms have roughly the same shape as the normal curve.
4. If a list of numbers follows the normal curve, the percentage of entries falling in a given interval can be estimated by converting the interval to standard units, and then finding the corresponding area under the normal curve. This procedure is called the *normal approximation*.
5. A histogram which follows the normal curve can be reconstructed fairly well from its average and SD. In such cases, the average and SD are good summary statistics.
6. All histograms, whether or not they follow the normal curve, can be summarized using *percentiles*.
7. If you add the same number to every entry on a list, that constant just gets added to the average; the SD does not change. If you multiply every entry on a list by the same positive number, the average and the SD just get multiplied by that constant. (If the constant is negative, wipe out the sign before multiplying the SD.)



# 6

## Measurement Error

*Jesus: I am come to bear witness unto the truth.  
Pilate: What is truth?*

### 1. INTRODUCTION

In an ideal world, if the same thing is measured several times, the same result would be obtained each time. In practice, there are differences. Each result is thrown off by chance error, and the error changes from measurement to measurement. One of the earliest scientists to deal with this problem was Tycho Brahe (1546–1601), the Danish astronomer. But it was probably noticed first in the market place, as merchants weighed out spices and measured off lengths of silk.

There are several questions about chance errors. Where do they come from? How big are they likely to be? How much is likely to cancel out in the average? The first question has a short answer: in most cases, nobody knows. The second question will be dealt with later in this chapter, and the third will be answered in part VII.

### 2. CHANCE ERROR

This section will discuss chance errors in precision weighing done at the National Bureau of Standards.<sup>1</sup> First, a brief explanation of standard weights. Stores weigh merchandise on scales. The scales are checked periodically by county

weights-and-measures officials, using county standard weights. The county standards too must be *calibrated* (checked against external standards) periodically. This is done at the state level. And state standards are calibrated against national standards, by the National Bureau of Standards in Washington, D.C.

This chain of comparisons ends at the International Prototype Kilogram (for short, The Kilogram), a platinum-iridium weight held at the International Bureau of Weights and Measures near Paris. By international treaty—The Treaty of the Meter, 1875—“one kilogram” was defined to be the weight of this object under standard conditions.<sup>2</sup> All other weights are determined relative to The Kilogram. For instance, something weighs a pound if it weighs just a bit less than half as much as The Kilogram. More precisely,

$$\text{The Pound} = 0.4539237 \text{ of The Kilogram.}$$

To say that a package of butter weighs a pound means that it has been connected by some long and complicated series of comparisons to The Kilogram in Paris, and weighs 0.4539237 times as much.

Each country that signed the Treaty of the Meter got a national prototype kilogram, whose exact weight had been determined as accurately as possible relative to The Kilogram. These prototypes were distributed by lot, and the United States got Kilogram #20. The values of all the U.S. national standards are determined relative to K<sub>20</sub>.

In the U.S., accuracy in weighing at the supermarket ultimately depends on the accuracy of the calibration work done at the Bureau. One basic issue is reproducibility: if a measurement is repeated, how much will it change? The Bureau gets at this issue by making repeated measurements on some of their own weights. We will discuss the results for one such weight, called NB 10 because it is owned by the National Bureau and its nominal value is 10 grams—the weight of two nickels. (A package of butter has a “nominal” weight of 1 pound; the exact weight will be a little different—chance error in butter; similarly, the people who manufactured NB 10 tried to make it weigh 10 grams, and missed by a little.)

NB 10 was acquired by the Bureau around 1940, and they’ve weighed it many times since then. We are going to look at 100 of these weighings. These measurements were made in the same room, on the same apparatus, by the same technicians. Every effort was made to follow the same procedure each time. All the factors known to affect the results, like air pressure or temperature, were kept as constant as possible.

The first five weighings in the series were

- 9.999591 grams
- 9.999600 grams
- 9.999594 grams
- 9.999601 grams
- 9.999598 grams

At first glance, these numbers all seem to be the same. But look more closely. It is only the first 4 digits that are solid, at 9.999. The last 3 digits are shaky, they change from measurement to measurement. This is chance error at work.<sup>3</sup>

NB 10 does weigh a bit less than 10 grams. Instead of writing out the 9.999 each time, the Bureau just reports the amount by which NB10 fell short of 10 grams. For the first weighing, this was

0.000409 grams.

The 0's are distracting, so the Bureau works not in grams but in micrograms: a *microgram* is the millionth part of a gram. In these units, the first five measurements on NB 10 are easier to read. They are

409      400      406      399      402.

All 100 measurements are shown in table 1. Look down the table. You can see that the results run around 400 micrograms, but some are more, some are less. The smallest is 375 micrograms (#94); the largest is 437 micrograms (#86). And there is a lot of variability in between. To keep things in perspective, one microgram is the weight of a large speck of dust; 400 micrograms is the weight of a grain or two of salt. This really is precision weighing!

Even so, the different measurements can't all be right. The exact amount by which NB 10 falls short of 10 grams is very unlikely to equal the first number

Table 1. One hundred measurements on NB 10. Almer and Jones, National Bureau of Standards. Units are micrograms below 10 grams.

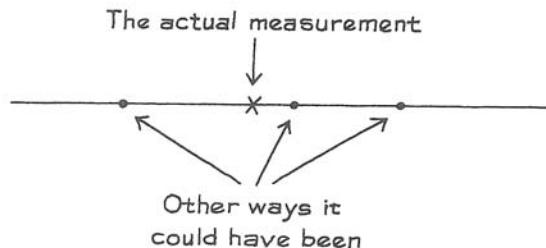
No.	Result	No.	Result	No.	Result	No.	Result
1	409	26	397	51	404	76	404
2	400	27	407	52	406	77	401
3	406	28	401	53	407	78	404
4	399	29	399	54	405	79	408
5	402	30	401	55	411	80	406
6	406	31	403	56	410	81	408
7	401	32	400	57	410	82	406
8	403	33	410	58	410	83	401
9	401	34	401	59	401	84	412
10	403	35	407	60	402	85	393
11	398	36	423	61	404	86	437
12	403	37	406	62	405	87	418
13	407	38	406	63	392	88	415
14	402	39	402	64	407	89	404
15	401	40	405	65	406	90	401
16	399	41	405	66	404	91	401
17	400	42	409	67	403	92	407
18	401	43	399	68	408	93	412
19	405	44	402	69	404	94	375
20	402	45	407	70	407	95	409
21	408	46	406	71	412	96	406
22	399	47	413	72	406	97	398
23	399	48	409	73	409	98	406
24	402	49	404	74	400	99	403
25	399	50	402	75	408	100	404

in the table, or the second, or any of them. Despite the effort of making these 100 measurements, the exact weight of NB 10 remains unknown and perhaps unknowable.

Why does the Bureau bother to weigh the same weight over and over again? One of the objectives is quality control. If the measurements on NB 10 jump from 400 micrograms below 10 grams to 500 micrograms above 10 grams, something has gone wrong and needs to be fixed. (For this reason, NB 10 is called a *check weight*; it is used to check the weighing process.)

To see another use for repeated measurements, imagine that a scientific laboratory sends a nominal 10-gram weight off to the Bureau for calibration. One measurement can't be the last word, because of chance error. The lab will want to know how big this chance error is likely to be. There is a direct way to find out: send the same weight back for a second weighing. If the two results differ by a few micrograms, the chance error in each one is only likely to be a few micrograms in size. On the other hand, if the two results differ by several hundred micrograms, each measurement is likely to be off by several hundred micrograms. The repeated weighings on NB 10 save everybody the bother of sending in weights more than once. There is no need to ask for replicate calibrations because the Bureau has already done the work.

No matter how carefully it was made, a measurement could have come out a bit differently. If the measurement is repeated, it will come out a bit differently. By how much? The best way to answer this question is to replicate the measurement.



The SD of the 100 measurements in table 1 is just over 6 micrograms. The SD tells you that each measurement on NB 10 was thrown off by a chance error something like 6 micrograms in size. Chance errors around 2 or 5 or 10 micrograms in size were fairly common. Chance errors around 50 or 100 micrograms must have been extremely rare. The conclusion: in calibrating other 10-gram weights by the same process, the chance errors should be something like 6 micrograms in size.

The SD of a series of repeated measurements estimates the likely size of the chance error in a single measurement.

There is an equation which helps explain the idea:

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

The chance error throws each individual measurement off the exact value by an amount which changes from measurement to measurement. The variability in repeated measurements reflects the variability in the chance errors, and both are gauged by the SD of the data. Mathematically, the SD of the chance errors must equal the SD of the measurements: adding the exact value is just a change of scale (pp. 92–93).

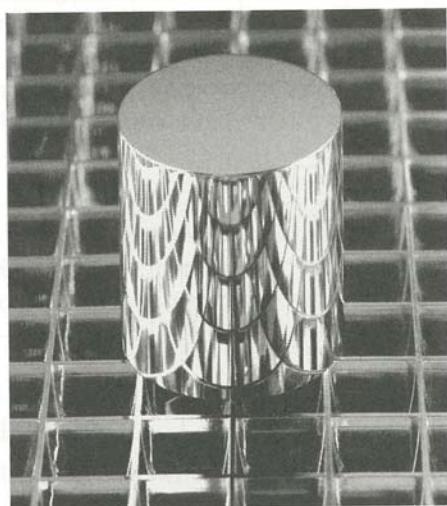
To go at this more slowly, the average of all 100 measurements reported in table 1 was 405 micrograms below 10 grams. This is very likely to be close to the exact weight of NB 10. The first measurement in table 1 differed from the average by 4 micrograms:

$$409 - 405 = 4.$$

This measurement must have differed from the exact weight by nearly 4 micrograms. The chance error was nearly 4 micrograms. The second measurement was below average by 5 micrograms; the chance error must have been around –5 micrograms. The typical deviation from average was around 6 micrograms in size, because the SD was 6 micrograms. Therefore, the typical chance error must have been something like 6 micrograms in size.

Of course, the average of all 100 measurements (405 micrograms below 10 grams) is itself only an estimate for the exact weight of NB10. This estimate too must be off by some infinitesimal chance error. Chapter 24 will explain how to figure the likely size of the chance error in this sort of average.

Figure 1. The U.S. national prototype kilogram, K<sub>20</sub>.



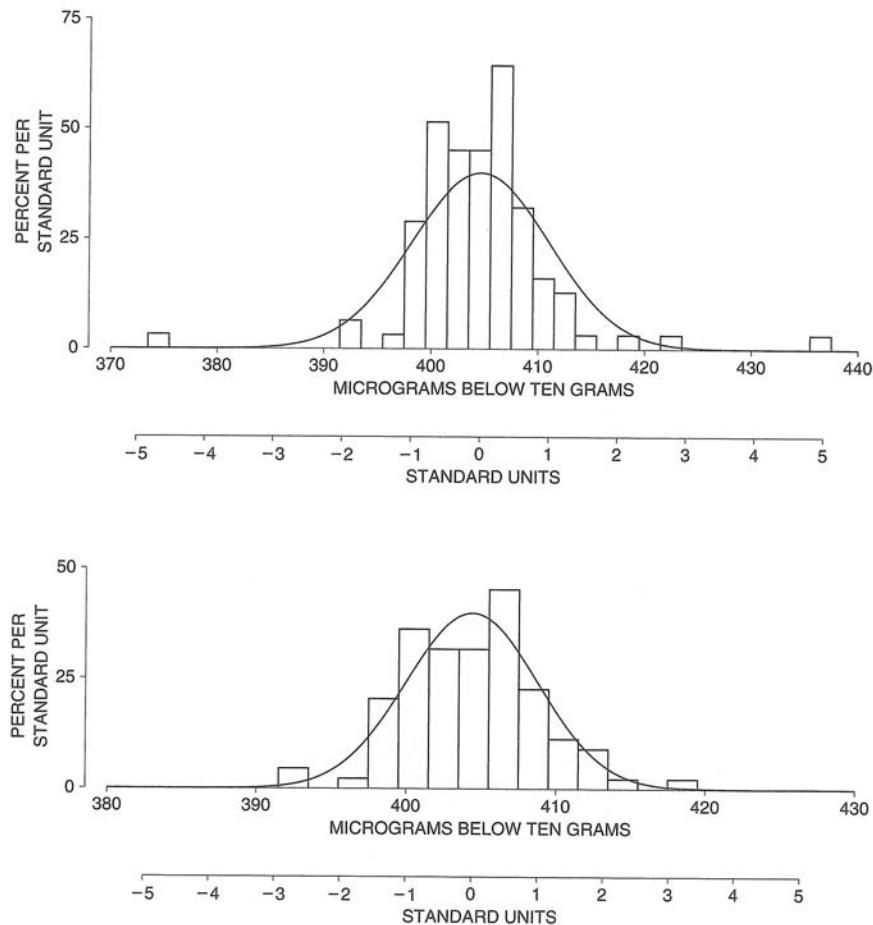
Source: National Institute of Science and Technology.

### 3. OUTLIERS

How well do the measurements reported in table 1 fit the normal curve? The answer is, not very well. Measurement #36 is 3 SDs away from the average; #86 and #94 are 5 SDs away—minor miracles. Such extreme measurements are called *outliers*. They do not result from blunders. As far as the Bureau could tell, nothing went wrong when these 3 observations were made. However, the 3 outliers inflate the SD. Consequently, the percentage of results falling closer to the average than one SD is 86%—quite a bit larger than the 68% predicted by the normal curve.

When the 3 outliers are discarded, the remaining 97 measurements average out to 404 micrograms below 10 grams, with an SD of only 4 micrograms. The average doesn't change much, but the SD drops by about 30%. As figure 2 shows,

Figure 2. Outliers. The top panel shows the histogram for all 100 measurements on NB 10; a normal curve is drawn for comparison. The curve does not fit well. The second panel shows the data with 3 outliers removed. The curve fits better. Most of the data follow the normal curve, but a few measurements are much further away from average than the curve suggests.



the remaining 97 measurements come closer to the normal curve. In sum, most of the data have an SD of about 4 micrograms. But a few of the measurements are quite a bit further away from the average than the SD would suggest. The overall SD of 6 micrograms is a compromise between the SD of the main part of the histogram—4 micrograms—and the outliers.

In careful measurement work, a small percentage of outliers is expected. The only unusual aspect of the NB 10 data is that the outliers are reported. Here is what the Bureau has to say about *not* reporting outliers.<sup>4</sup> For official prose, the tone is quite stern.

A major difficulty in the application of statistical methods to the analysis of measurement data is that of obtaining suitable collections of data. The problem is more often associated with conscious, or perhaps unconscious, attempts to make a particular process perform as one would like it to perform rather than accepting the actual performance . . . . Rejection of data on the basis of arbitrary performance limits severely distorts the estimate of real process variability. Such procedures defeat the purpose of the . . . program. Realistic performance parameters require the acceptance of all data that cannot be rejected for cause.

There is a hard choice to make when investigators see an outlier. Either they ignore it, or they have to concede that their measurements don't follow the normal curve. The prestige of the curve is so high that the first choice is the usual one—a triumph of theory over experience.

#### 4. BIAS

Suppose a butcher weighs a steak with his thumb on the scale. That causes an error in the measurement, but little has been left to chance. Take another example. Suppose a fabric store uses a cloth tape measure which has stretched from 36 inches to 37 inches in length. Every “yard” of cloth they sell to a customer has an extra inch tacked onto it. This isn't a chance error, because it always works for the customer. The butcher's thumb and the stretched tape are two examples of *bias*, or *systematic error*.

Bias affects all measurements the same way, pushing them in the same direction. Chance errors change from measurement to measurement, sometimes up and sometimes down.

The basic equation has to be modified when each measurement is thrown off by bias as well as chance error:

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

If there is no bias in a measurement procedure, the long-run average of repeated measurements should give the exact value of the thing being measured: the chance

errors should cancel out. However, when bias is present, the long-run average will itself be either too high or too low.

Usually, bias cannot be detected just by looking at the measurements themselves. Instead, the measurements have to be compared to an external standard or to theoretical predictions. In the U.S., all weight measurements depend on the connection between K<sub>20</sub> and The Kilogram. These two weights have been compared a number of times, and it is estimated that K<sub>20</sub> is a tiny bit lighter than The Kilogram—by 19 parts in a billion. All weight calculations at the Bureau are revised upward by 19 parts in a billion, to compensate. However, this factor itself is likely to be just a shade off: it too was the result of some measurement process. All weights measured in the U.S. are systematically off, by the same (tiny) percentage. This is another example of bias, but not one to worry about.

## 5. REVIEW EXERCISES

1. True or false, and explain: “An experienced scientist who is using the best equipment available only needs to measure things once—provided he doesn’t make a mistake. After all, if he measures the same thing twice, he’ll get the same results both times.”
2. A carpenter is using a tape measure to get the length of a board.
  - (a) What are some possible sources of bias?
  - (b) Which is more subject to bias, a steel tape or a cloth tape?
  - (c) Would the bias in a cloth tape change over time?
3. True or false, and explain.
  - (a) Bias is a kind of chance error.
  - (b) Chance error is a kind of bias.
  - (c) Measurements are usually affected by both bias and chance error.
4. You send a yardstick to a local laboratory for calibration, asking that the procedure be repeated three times. They report the following values:

35.96 inches      36.01 inches      36.03 inches

If you send the yardstick back for a fourth calibration, you would expect to get 36 inches give or take—

.01 inches or so      .03 inches or so      .06 inches or so

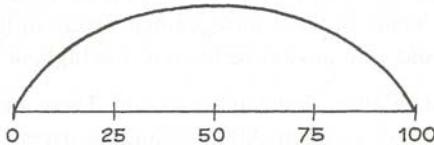
5. Nineteen students in a beginning statistics course were asked to measure the thickness of a table top, using a vernier gauge reading to 0.001 of an inch. Each person made two measurements, shown at the top of the next page. (The units are inches; for instance, the first person got 1.317 and 1.320 for the two measurements.)
  - (a) Did the students work independently of one another?
  - (b) Some friends of yours do not believe in chance error. How could you use these data to convince them?

<i>Person</i>	<i>Measurements (inches)</i>		<i>Person</i>	<i>Measurements (inches)</i>	
	<i>1st</i>	<i>2nd</i>		<i>1st</i>	<i>2nd</i>
1	1.317	1.320	11	1.333	1.334
2	13.26	13.25	12	1.315	1.317
3	1.316	1.335	13	1.316	1.318
4	1.316	1.328	14	1.321	1.319
5	1.318	1.324	15	1.337	1.343
6	1.329	1.326	16	1.349	1.336
7	1.332	1.334	17	1.320	1.336
8	1.342	1.328	18	1.342	1.340
9	1.337	1.342	19	1.317	1.318
10	13.26	13.25			

## 6. SPECIAL REVIEW EXERCISES

*These exercises cover all of parts I and II.*

1. In one course, a histogram for the scores on the final looked like the sketch below. True or false: because this isn't like the normal curve, there must have been something wrong with the test. Explain.



2. Fill in the blanks, using the options below, and give examples to show that you picked the right answers.
- The SD of a list is 0. This means \_\_\_\_\_.
  - The r.m.s. size of a list is 0. This means \_\_\_\_\_.

Options:

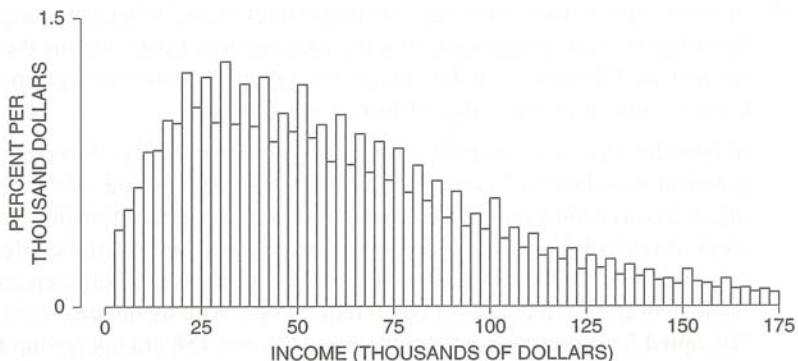
- there are no numbers on the list
- all the numbers on the list are the same
- all the numbers on the list are 0
- the average of the list is 0

3. A personality test is administered to a large group of subjects. Five scores are shown below, in original units and in standard units. Fill in the blanks.

79	64	52	72	_____
1.8	0.8	—	—	-1.4

4. Among first-year students at a certain university, scores on the Verbal SAT follow the normal curve; the average is around 550 and the SD is about 100.
- What percentage of these students have scores in the range 400 to 700?

- (b) There were about 1,000 students with scores in the range 450–650 on the Verbal SAT. About \_\_\_\_\_ of them had scores in the range 500 to 600. Fill in the blank; explain briefly.
5. In Cycle III of the Health Examination Survey (like HANES, but done in 1966–70), there were 6,672 subjects. The sex of each subject was recorded at two different stages of the survey. In 17 cases, there was a discrepancy: the subject was recorded as male at one interview, female at the other. How would you account for this?
6. Among entering students at a certain college, the men averaged 650 on the Math SAT, and their SD was 125. The women averaged 600, but had the same SD of 125. There were 500 men in the class, and 500 women.
- For the men and the women together, the average Math SAT score was \_\_\_\_\_.
  - For the men and the women together, was the SD of Math SAT scores less than 125, just about 125, or more than 125?
7. Repeat exercise 6, when there are 600 men in the class, and 400 women. (The separate averages and SDs for the men and women stay the same.)
8. Table 1 on p. 99 reported 100 measurements on the weight of NB 10; the top panel in figure 2 on p. 102 shows the histogram. The average was 405 micrograms, and the SD was 6 micrograms. If you used the normal approximation to estimate how many of these measurements were in the range 400 to 406 micrograms, would your answer be too low, too high, or about right? Why?
9. A teaching assistant gives a quiz to his section. There are 10 questions on the quiz and no part credit is given. After grading the papers, the TA writes down for each student the number of questions the student got right and the number wrong. The average number of right answers is 6.4 with an SD of 2.0. The average number of wrong answers is \_\_\_\_\_ with an SD of \_\_\_\_\_. Fill in the blanks—or do you need the data? Explain briefly.
10. A large, representative sample of Americans was studied by the Public Health Service, in the Health and Nutrition Examination Survey (HANES2).<sup>5</sup> The percentage of respondents who were left-handed decreased steadily with age, from 10% at 20 years to 4% at 70. “The data show that many people change from left-handed to right-handed as they get older.” True or false? Why? If false, how do you explain the pattern in the data?
11. For a certain group of women, the 25th percentile of height is 62.2 inches and the 75th percentile is 65.8 inches. The histogram follows the normal curve. Find the 90th percentile of the height distribution.
12. In March, the Current Population Survey asks a large, representative sample of Americans to say what their incomes were during the previous year.<sup>6</sup> A histogram for family income in 2004 is shown at the top of the next page. (Class intervals include the left endpoint but not the right.) From \$15,000 and on to the right, the blocks alternate regularly from high to low. Why is that?



13. To measure the effect of exercise on the risk of heart disease, investigators compared the incidence of this disease for two large groups of London Transport Authority busmen—drivers and conductors. The conductors got a lot more exercise as they walked around all day collecting fares.

The age distributions for the two groups were very similar, and all the subjects had been on the same job for 10 years or more. The incidence of heart disease was substantially lower among the conductors, and the investigators concluded that exercise prevents heart disease.

Other investigators were skeptical. They went back and found that London Transport Authority had issued uniforms to drivers and conductors at the time of hire; a record had been kept of the sizes.<sup>7</sup>

- (a) Why does it matter that the age distributions of the two groups were similar?
- (b) Why does it matter that all the subjects had been on the job for 10 years or more?
- (c) Why did the first group of investigators compare the conductors to drivers, not to London Transport Authority executive staff?
- (d) Why might the second group of investigators have been skeptical?
- (e) What would you do with the sizes of the uniforms?

14. Breast cancer is one of the most common malignancies among women in Canada and the U.S. If it is detected early enough—before the cancer spreads—chances of successful treatment are much better. Do screening programs speed up detection by enough to matter? Many studies have examined this question.

The Canadian National Breast Cancer Study was a randomized controlled experiment on mammography, that is, x-ray screening for breast cancer. The study found no benefit from screening. (The benefit was measured by comparing death rates from breast cancer in the treatment and control groups.)

Dr. Daniel Kopans argued that the randomization was not done properly: instead of following instructions, nurses assigned high risk women to the treatment group.<sup>8</sup> Would this bias the study? If so, would the bias make the benefit from screening look bigger or smaller than it really is? Explain your answer.

15. In some jurisdictions, there are “pretrial conferences,” where the judge confers with the opposing lawyers to settle the case or at least to define the issues before trial. Observational data suggest that pretrial conferences promote settlements and speed up trials, but there were doubts.

In New Jersey courts, pretrial conferences were mandatory. However, an experiment was done in 7 counties. During a six-month period, 2,954 personal injury cases (mainly automobile accidents) were assigned at random to treatment or control. For the 1,495 control cases (group A), pretrial conferences remained mandatory. For the 1,459 treatment cases, the conferences were made optional—either lawyer could request one. Among the treatment cases, 701 opted for a pretrial conference (group C), and 758 did not (group B).

The investigator who analyzed the data looked to see whether pretrial conferences encouraged cases to settle before reaching trial; or, if they went to trial, whether the conferences shortened the amount of trial time. (This matters, because trial time is very expensive.)

The investigator reported the main results as follows; tabular material is quoted from his report.<sup>9</sup>

- (i) Pretrial conferences had no impact on settlement; the same percentage go to trial in group B as in group A + C.

*Percentage of cases reaching trial*

	<i>Group B</i>	<i>Group A + C</i>
Reached trial	22%	23%
Number of cases	701	2,079

- (ii) Pretrial conferences do not shorten trial time; the percentage of short trials is highest in cases that refused pretrial conferences.

*Distribution of trial time among cases that go to trial*

	<i>Group B</i>	<i>Group A</i>	<i>Group C</i>
<i>Trial time (in hours)</i>			
1. 5 or less	43%	34%	28%
2. Over 5 to 10	35%	41%	39%
3. Over 10	22%	26%	33%
Number of cases	63	176	70

Comment briefly on the analysis.

## 7. SUMMARY AND OVERVIEW

1. No matter how carefully it was made, a measurement could have turned out a bit differently. This reflects *chance error*. Before investigators rely on a measurement, they should estimate the likely size of the chance error. The best way to do that: *replicate* the measurement.

2. The likely size of the chance error in a single measurement can be estimated by the SD of a sequence of repeated measurements made under the same conditions.

3. *Bias*, or *systematic error*, causes measurements to be systematically too high or systematically too low. The equation is

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

The chance error changes from measurement to measurement, but the bias stays the same. Bias cannot be estimated just by repeating the measurements.

4. Even in careful measurement work, a small percentage of *outliers* can be expected.

5. The average and SD can be strongly influenced by outliers. Then the histogram will not follow the normal curve at all well.

6. This part of the book introduced two basic descriptive statistics, the average and the standard deviation; histograms were used to summarize data. For many data sets, the histogram follows the normal curve. Chapter 6 illustrates these ideas on measurement data. Later in the book, histograms will be used for probability distributions, and statistical inference will be based on the normal curve. This is legitimate when the probability histograms follow the curve—the topic of chapter 18.

# 7

## Plotting Points and Lines

*Q. What did the dot say to the line?*

*A. Come to the point.*

### 1. READING POINTS OFF A GRAPH

This chapter reviews some of the ideas about plotting points and lines which will be used in part III. You can either read this chapter now, or return to it if you run into difficulty in part III. If you read the chapter now, the first four sections are the most important; the last section is more difficult.

Figure 1 shows a horizontal axis (the  $x$ -axis) and a vertical axis (the  $y$ -axis). The point shown in the figure has an  $x$ -coordinate of 3, because it is in line with 3 on the  $x$ -axis. It has a  $y$ -coordinate of 2, because it is in line with 2 on the  $y$ -axis. This point is written  $x = 3, y = 2$ . Sometimes, it is abbreviated even more, to  $(3, 2)$ . The point shown in figure 2 is  $(-2, -1)$ : it is directly below  $-2$  on the  $x$ -axis, and directly to the left of  $-1$  on the  $y$ -axis.

Figure 1.

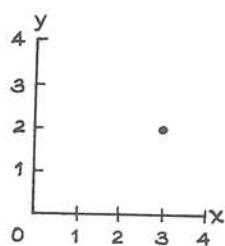
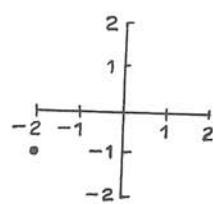


Figure 2.

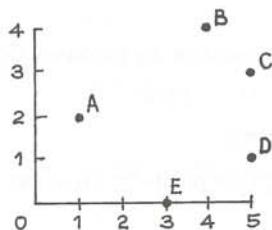


The idea of representing points by pairs of numbers is due to René Descartes (France, 1596–1650). In his honor, the  $x$ - and  $y$ -coordinates are often called “cartesian coordinates.”

### Exercise Set A

- Figure 3 shows five points. Write down the  $x$ -coordinate and  $y$ -coordinate for each point.
- As you move from point A to point B in figure 3, your  $x$ -coordinate goes up by \_\_\_\_\_; your  $y$ -coordinate goes up by \_\_\_\_\_.
- One point in figure 3 has a  $y$ -coordinate 1 bigger than the  $y$ -coordinate of point E. Which point is that?

Figure 3.



*The answers to these exercises are on p. A52.*



René Descartes (France, 1596–1650)

Wolff-Leavenworth Collection, courtesy of the Syracuse University Art Collection.

## 2. PLOTTING POINTS

Figure 4 shows a pair of axes. To plot the point  $(2, 1)$ , find the 2 on the  $x$ -axis. The point will be directly above this, as in figure 5. Find the 1 on the  $y$ -axis, the point will be directly to the right of this, as in figure 6.

Figure 4.

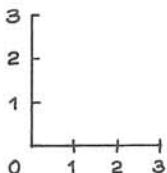


Figure 5.

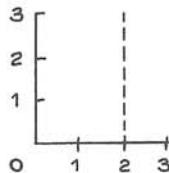
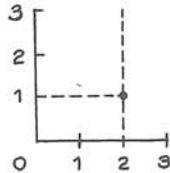


Figure 6.



## Exercise Set B

1. Draw a pair of axes and plot each of the following points:

$$(1, 1) \quad (2, 2) \quad (3, 3) \quad (4, 4)$$

What can you say about them?

2. Three out of the following four points lie on a line. Which is the maverick? Is it above or below the line?

$$(0, 0) \quad (0.5, 0.5) \quad (1, 2) \quad (2.5, 2.5)$$

3. The table below shows four points. In each case, the  $y$ -coordinate is computed from the  $x$ -coordinate by the rule  $y = 2x + 1$ . Fill in the blanks, then plot the four points. What can you say about them?

$x$	$y$
1	3
2	5
3	—
4	—

4. Figure 7 below shows a shaded region. Which of the following two points is in the region:  $(1, 2)$  or  $(2, 1)$ ?

5. Do the same for figure 8.

6. Do the same for figure 9.

Figure 7.

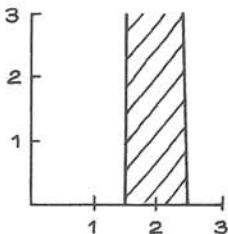


Figure 8.

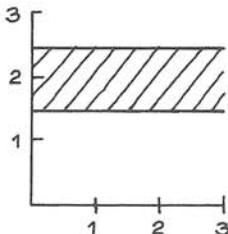
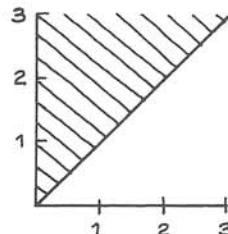


Figure 9.



The answers to these exercises are on p. A53.

### 3. SLOPE AND INTERCEPT

Figure 10 shows a line. Take any point on the line—for instance, point A. Now move up the line to any other point—for instance, point B. Your  $x$ -coordinate has increased by some amount, called the *run*. In this case, the run was 2. At the same time your  $y$ -coordinate has increased by some other amount, called the *rise*. In this case, the rise was 1. Notice that in this case, the rise was half the run. Whatever two points you take on this line, the rise will be half the run. The ratio *rise/run* is called the *slope* of the line:

$$\text{slope} = \text{rise/run}.$$

The slope is the rate at which  $y$  increases with  $x$ , along the line. To interpret it another way, imagine the line as a road going up a hill. The slope measures the steepness of the grade. For the line in figure 10, the grade is 1 in 2—quite steep for a road. In figure 11, the slope of the line is 0. In figure 12, the slope is  $-1$ . If the slope is positive, the line is going uphill. If the slope is 0, the line is horizontal. If the slope is negative, the line is going downhill.

Figure 10. Slope is  $1/2$ .

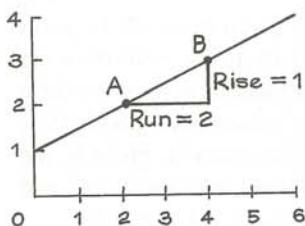


Figure 11. Slope is 0.

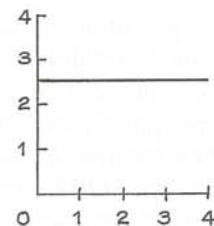
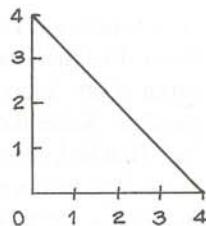


Figure 12. Slope is  $-1$ .



The *intercept* of a line is its height at  $x = 0$ . Usually, the axes cross at 0. Then, the intercept is where the line crosses the  $y$ -axis. In figure 13, the intercept is 2. Sometimes, the axes don't cross at 0, and then you have to be a little bit careful. In figure 14, the axes cross at  $(1, 1)$ . The intercept of the line in figure 14 is 0—that would be its height at  $x = 0$ .

Often, the axes of a graph show units. For example, in figure 15 the units for the  $x$ -axis are inches, the units for the  $y$ -axis are degrees celsius. Then the slope and intercept have units too. In figure 15, the slope of the line is 2.5 degrees per inch; the intercept is  $-5$  degrees.

Figure 13.

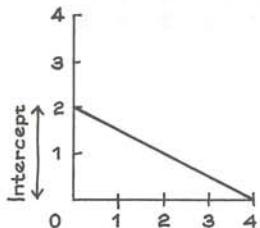


Figure 14.

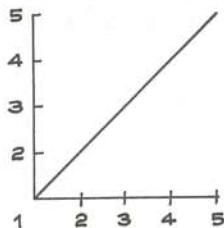
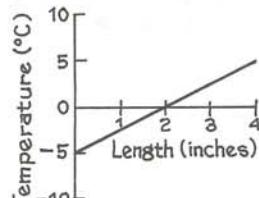


Figure 15.



## Exercise Set C

1. Figures 16 to 18 show lines. For each line, find the slope and intercept. Note: the axes do not cross at 0 in each case.

Figure 16.

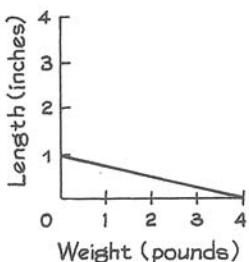


Figure 17.

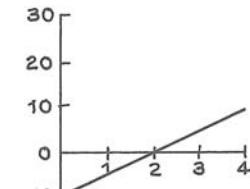
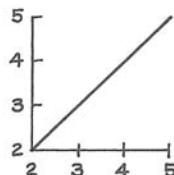


Figure 18.



The answers to these exercises are on p. A53.

## 4. PLOTTING LINES

*Example 1.* Plot the line which passes through the point  $(2, 1)$  and has slope  $1/2$ .

*Solution.* First draw a pair of axes and plot the given point  $(2, 1)$ , as in figure 19. Then move any convenient distance off directly to the right from the given point: figure 20 shows a run of 3. Make a construction point at this new location. Since the line slopes up, it passes above the construction point. How far? That is, how much will the line rise in a run of 3? The answer is given by the slope. The line is rising at the rate of half a vertical unit per horizontal unit, and in this case there is a run of 3 horizontal units, so the rise is  $3 \times 1/2 = 1.5$ :

$$\text{rise} = \text{run} \times \text{slope}.$$

Make a vertical move of 1.5 from the construction point, and mark a point at this third location, as in figure 21. This third point is on the line. Put a ruler down and join it to the given point  $(2, 1)$ .

Figure 19.

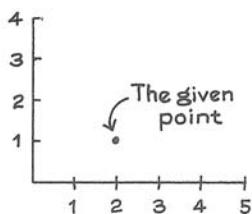


Figure 20.

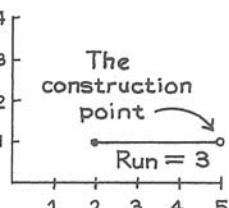
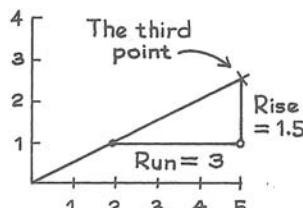


Figure 21.



## Exercise Set D

1. Draw lines through the point  $(2, 1)$  with the following slopes:
  - (a)  $+1$
  - (b)  $-1$
  - (c)  $0$
2. Start at the point  $(2, 1)$  in figure 21. If you move over 2 and up 1, will you be on the line, above the line, or below the line?
3. The same, but move over 4 and up 2.
4. The same but move over 6 and up 5.
5. Draw the line with intercept 2 and slope  $-1$ . Hint: this line goes through the point  $(0, 2)$ .
6. Draw the line with intercept 2 and slope 1.

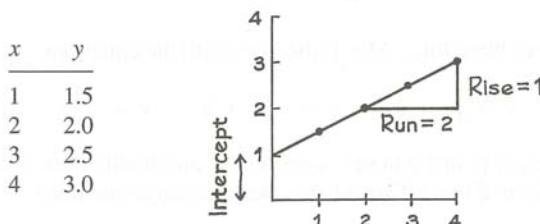
*The answers to these exercises are on p. A54.*

## 5. THE ALGEBRAIC EQUATION FOR A LINE

*Example 2.* Here is a rule for computing the  $y$ -coordinate of a point from its  $x$ -coordinate:  $y = \frac{1}{2}x + 1$ . The table below shows the points with  $x$ -coordinates of 1, 2, 3, 4. Plot the points. Do they fall on a line? If so, find the slope and intercept of this line.

*Solution.* The points are plotted in figure 22. They do fall on a line. Any point whose  $y$ -coordinate is related to its  $x$ -coordinate by the same equation  $y = \frac{1}{2}x + 1$  will fall on the same line. This line is called the *graph* of the equation. The slope of the line is  $\frac{1}{2}$ , the coefficient of  $x$  in the equation. The intercept is 1, the constant term in the equation.

Figure 22.



The graph of the equation  $y = mx + b$  is a straight line, with slope  $m$  and intercept  $b$ .

*Example 3.* Figure 23 shows a line. What is the equation of this line? What is the height of this line at  $x = 1$ ?

*Solution.* This line has slope  $-1$  and intercept  $4$ . Therefore, its equation is  $y = -x + 4$ . Substituting  $x = 1$  gives  $y = 3$ ; so the height of the line is  $3$  when  $x$  is  $1$ .

Figure 23.

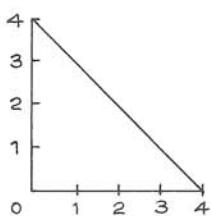


Figure 24.

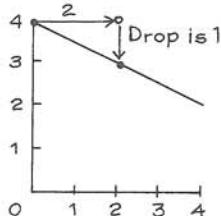
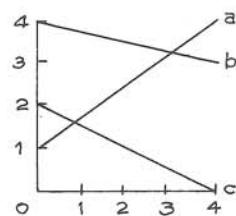


Figure 25.



*Example 4.* Plot the line whose equation is  $y = -\frac{1}{2}x + 4$ .

*Solution.* The intercept of this line is  $4$ . Plot the point  $(0, 4)$  as in figure 24. The line must go through this point. Make any convenient horizontal move—say  $2$ . The slope is  $-\frac{1}{2}$ , so the line must drop  $1$ . Mark the point which is  $2$  over and  $1$  down from the first point in figure 24. Then join these two points by a straight line.

### Exercise Set E

1. Plot the graphs of the following equations:

$$(a) \ y = 2x + 1 \quad (b) \ y = \frac{1}{2}x + 2$$

In each case, say what the slope and intercept are, and give the height of the line at  $x = 2$ .

2. Figure 25 shows three lines. Match the lines with the equations:

$$y = \frac{3}{4}x + 1 \quad y = -\frac{1}{4}x + 4 \quad y = -\frac{1}{2}x + 2$$

3. Plot four different points whose  $y$ -coordinates are double their  $x$ -coordinates. Do these points lie on a line? If so, what is the equation of the line?

4. Plot the points  $(1, 1)$ ,  $(2, 2)$ ,  $(3, 3)$ , and  $(4, 4)$  on the same graph. These points all lie on a line. What is the equation of this line?

5. For each of the following points, say whether it is on the line of exercise 4, or above, or below:

$$(a) (0, 0) \quad (b) (1.5, 2.5) \quad (c) (2.5, 1.5)$$

6. True or false:

- (a) If  $y$  is bigger than  $x$ , then the point  $(x, y)$  is above the line of exercise 4.
- (b) If  $y = x$ , then the point  $(x, y)$  is on the line of exercise 4.
- (c) If  $y$  is smaller than  $x$ , then the point  $(x, y)$  is below the line of exercise 4.

The answers to these exercises are on pp. A54–55.

PART III

# Correlation and Regression

— — — — —

# 8

## Correlation

*Like father, like son.*

### 1. THE SCATTER DIAGRAM

The methods discussed in part II are good for dealing with one variable at a time. Other methods are needed for studying the relationship between two variables.<sup>1</sup> Sir Francis Galton (England, 1822–1911) made some progress on this front while he was thinking about the degree to which children resemble their parents. Statisticians in Victorian England were fascinated by the idea of quantifying hereditary influences and gathered huge amounts of data in pursuit of this goal. We are going to look at the results of a study carried out by Galton's disciple Karl Pearson (England, 1857–1936).<sup>2</sup>

As part of the study, Pearson measured the heights of 1,078 fathers, and their sons at maturity. A list of 1,078 pairs of heights would be hard to grasp. But the relationship between the two variables—father's height and son's height—can be brought out in a *scatter diagram* (figure 1 on the next page). Each dot on the diagram represents one father-son pair. The  $x$ -coordinate of the dot, measured along the horizontal axis, gives the height of the father. The  $y$ -coordinate of the dot, along the vertical axis, gives the height of the son.

<sup>1</sup> See the section on “Measuring the strength of a correlation” on page 110.

<sup>2</sup> Pearson's work on heredity has been called “the most important single contribution to the science of statistics.”

Figure 1. Scatter diagram for heights of 1,078 fathers and sons. Shows positive association between son's height and father's height. Families where the height of the son equals the height of the father are plotted along the 45-degree line  $y = x$ . Families where the father is 72 inches tall (to the nearest inch) are plotted in the vertical strip.

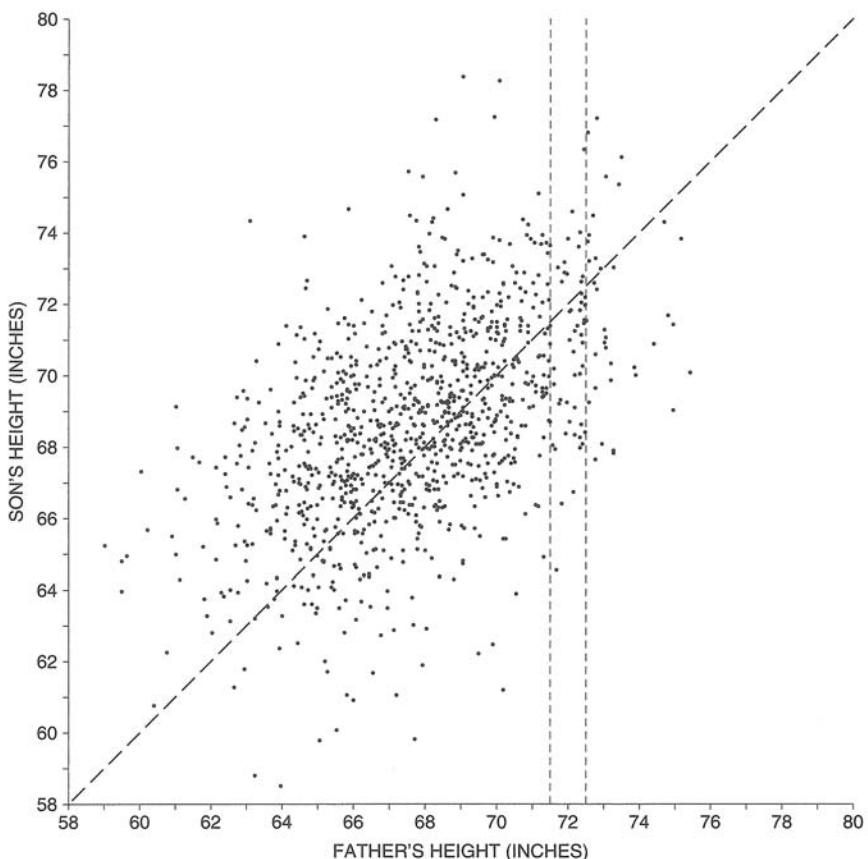


Figure 2a illustrates the mechanics of plotting scatter diagrams. (Chapter 7 has details.) The scatter diagram in figure 1 is a cloud shaped something like a football, with points straggling off the edges. When making a rough sketch of such a scatter diagram, it is only necessary to show the main oval portion—figure 2b.

The swarm of points in figure 1 slopes upward to the right, the  $y$ -coordinates of the points tending to increase with their  $x$ -coordinates. A statistician might say there is a *positive association* between the heights of fathers and sons. As a rule, the taller fathers have taller sons. This confirms the obvious. Now look at the 45-degree line in figure 1. This line corresponds to the families where son's height equals father's height. Along the line, for example, if the father is 72 inches tall then the son is 72 inches tall; if the father is 64 inches tall, the son is too; and so

Figure 2a. A point on a scatter diagram.

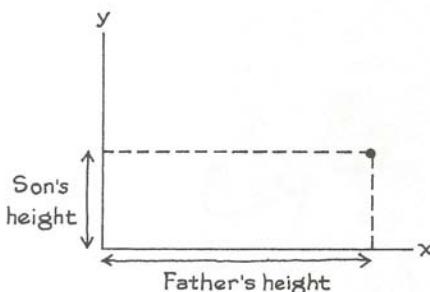
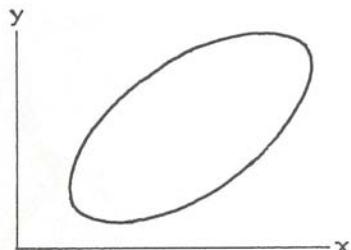


Figure 2b. Rough sketch.

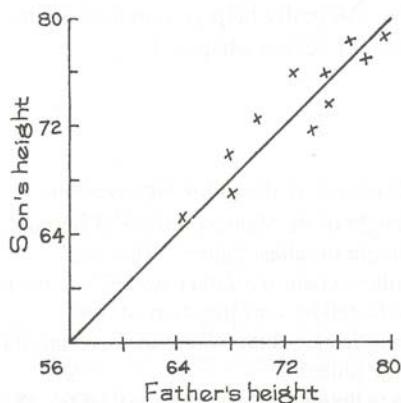


forth. Similarly, if a son's height is close to his father's height, then their point on the scatter diagram will be close to the line, like the points in figure 3.

There is a lot more spread around the 45-degree line in the actual scatter diagram than in figure 3. This spread shows the weakness of the relationship between father's height and son's height. For instance, suppose you have to guess the height of a son. How much help does the father's height give you? In figure 1, the dots in the chimney represent all the father-son pairs where the father is 72 inches tall to the nearest inch (father's height between 71.5 inches and 72.5 inches, where the dashed vertical lines cross the  $x$ -axis). There is still a lot of variability in the heights of the sons, as indicated by the vertical scatter in the chimney. Even if you know the father's height, there is still a lot of room for error in trying to guess the height of his son.

If there is a strong association between two variables, then knowing one helps a lot in predicting the other. But when there is a weak association, information about one variable does not help much in guessing the other.

Figure 3. Son's height close to father's height.





Sir Francis Galton (England, 1822–1911)

Source: *Biometrika* (November, 1903).

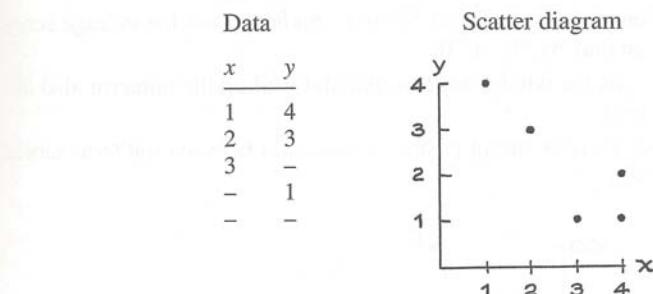
In social science studies of the relationship between two variables, it is usual to label one as *independent* and the other as *dependent*. Ordinarily, the independent variable is thought to influence the dependent variable, rather than the other way around. In figure 1, father's height is taken as the independent variable and plotted along the  $x$ -axis: father's height influences son's height. However, there is nothing to stop an investigator from using son's height as the independent variable. This choice might be appropriate, for example, if the problem is to guess a father's height from his son's height.

Before going on, it would be a good idea to work the exercises of this section. They are easy, and they will really help you understand the rest of this chapter. If you have trouble with them, review chapter 7.

### Exercise Set A

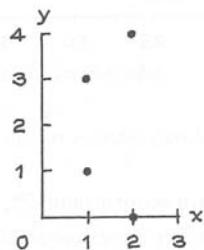
1. Use figure 1 (p. 120) to answer the following questions:
  - (a) What is the height of the shortest father? of his son?
  - (b) What is the height of tallest father? of his son?
  - (c) Take the families where the father was 72 inches tall, to the nearest inch. How tall was the tallest son? the shortest son?
  - (d) How many families are there where the sons are more than 78 inches tall? How tall are the fathers?
  - (e) Was the average height of the fathers around 64, 68, or 72 inches?
  - (f) Was the SD of the fathers' heights around 3, 6, or 9 inches?

2. Below is the scatter diagram for a certain data set. Fill in the blanks.

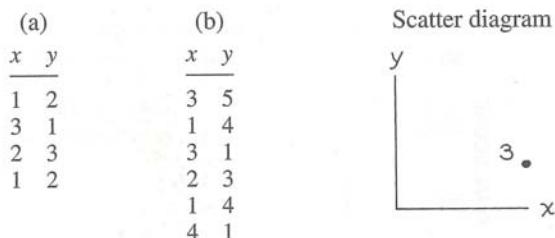


3. Below is a scatter diagram for some hypothetical data.

- Is the average of the  $x$ -values around 1, 1.5, or 2?
- Is the SD of the  $x$ -values around 0.1, 0.5, or 1?
- Is the average of the  $y$ -values around 1, 1.5, or 2?
- Is the SD of the  $y$ -values around 0.5, 1.5, or 3?



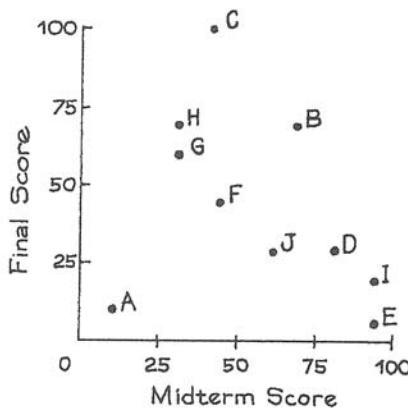
4. Draw the scatter diagram for each of the following hypothetical data sets. The variable labeled "x" should be plotted along the  $x$ -axis, the one labeled "y" along the  $y$ -axis. Mark each axis fully. In some cases, you will have to plot the same point more than once. The number of times such a multiple point appears can be indicated next to the point, as in the diagram below; please follow this convention.



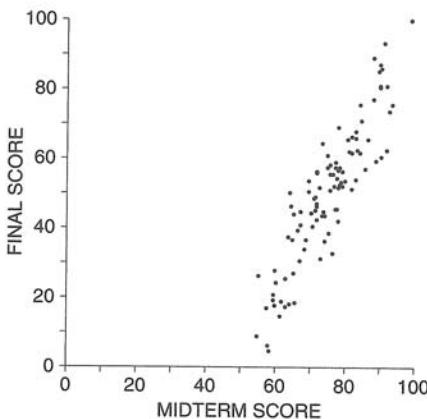
5. Students named A, B, C, D, E, F, G, H, I, and J took a midterm and a final in a certain course. A scatter diagram for the scores is shown on the next page.

- Which students scored the same on the midterm as on the final?
- Which students scored higher on the final?

- (c) Was the average score on the final around 25, 50, or 75?
- (d) Was the SD of the scores on the final around 10, 25, or 50?
- (e) For the students who scored over 50 on the midterm, was the average score on the final around 30, 50, or 70?
- (f) True or false: on the whole, students who did well on the midterm also did well on the final.
- (g) True or false: there is strong positive association between midterm scores and final scores.



6. The scatter diagram below shows scores on the midterm and final in a certain course.
- (a) Was the average midterm score around 25, 50, or 75?
  - (b) Was the SD of the midterm scores around 5, 10, or 20?
  - (c) Was the SD of the final scores around 5, 10, or 20?
  - (d) Which exam was harder—the midterm or the final?
  - (e) Was there more spread in the midterm scores, or the final scores?
  - (f) True or false: there was a strong positive association between midterm scores and final scores.



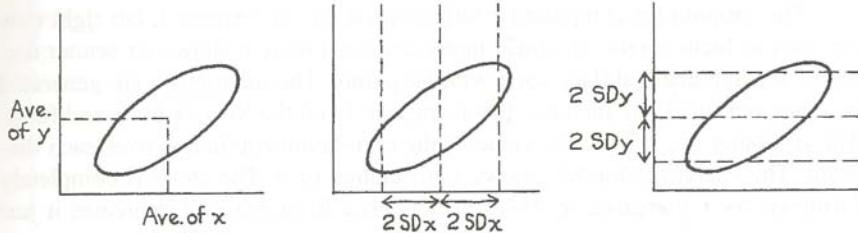
*The answers to these exercises are on pp. A55–56.*

## 2. THE CORRELATION COEFFICIENT

Suppose you are looking at the relationship between two variables, and have already plotted the scatter diagram. The graph is a football-shaped cloud of points. How can it be summarized? The first step would be to mark a point showing the average of the  $x$ -values and the average of the  $y$ -values (figure 4a). This is the *point of averages*, which locates the center of the cloud.<sup>3</sup> The next step would be to measure the spread of the cloud from side to side. This can be done using the SD of the  $x$ -values—the horizontal SD. Most of the points will be within 2 horizontal SDs on either side of the point of averages (figure 4b). In the same way, the SD of the  $y$ -values—the vertical SD—can be used to measure the spread of the cloud from top to bottom. Most of the points will be within 2 vertical SDs above or below the point of averages (figure 4c).

Figure 4. Summarizing a scatter diagram.

(a) The point of averages      (b) The horizontal SD      (c) The vertical SD



So far, the summary statistics are

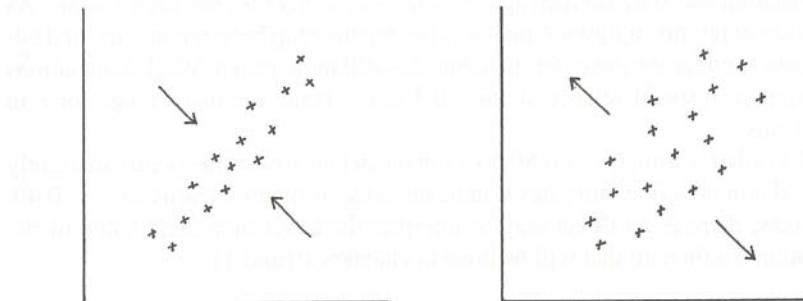
- average of  $x$ -values, SD of  $x$ -values,
- average of  $y$ -values, SD of  $y$ -values.

These statistics tell us the center of the cloud, and how spread out it is, both horizontally and vertically. But there is still something missing—the strength of the association between the two variables. Look at the scatter diagrams in figure 5.

Figure 5. Summarizing a scatter diagram. The correlation coefficient measures clustering around a line.

(a) Correlation near 1 means tight clustering.

(b) Correlation near 0 means loose clustering.



Both clouds have the same center and show the same spread, horizontally and vertically. However, the points in the first cloud are tightly clustered around a line: there is a strong linear association between the two variables. In the second cloud, the clustering is much looser. The strength of the association is different in the two diagrams. To measure the association, one more summary statistic is needed—the *correlation coefficient*. This coefficient is usually abbreviated as  $r$ , for no good reason (although there are two  $r$ 's in “correlation”).

The correlation coefficient is a measure of linear association, or clustering around a line. The relationship between two variables can be summarized by

- the average of the  $x$ -values, the SD of the  $x$ -values,
- the average of the  $y$ -values, the SD of the  $y$ -values,
- the correlation coefficient  $r$ .

The formula for computing  $r$  will be presented in section 4, but right now we want to focus on the graphical interpretation. Figure 6 shows six scatter diagrams for hypothetical data, each with 50 points. The diagrams were generated by computer. In all six pictures, the average is 3 and the SD is 1 for  $x$  and for  $y$ . The computer has printed the value of the correlation coefficient over each diagram. The one at the top left shows a correlation of 0. The cloud is completely formless. As  $x$  increases,  $y$  shows no tendency to increase or decrease: it just straggles around.

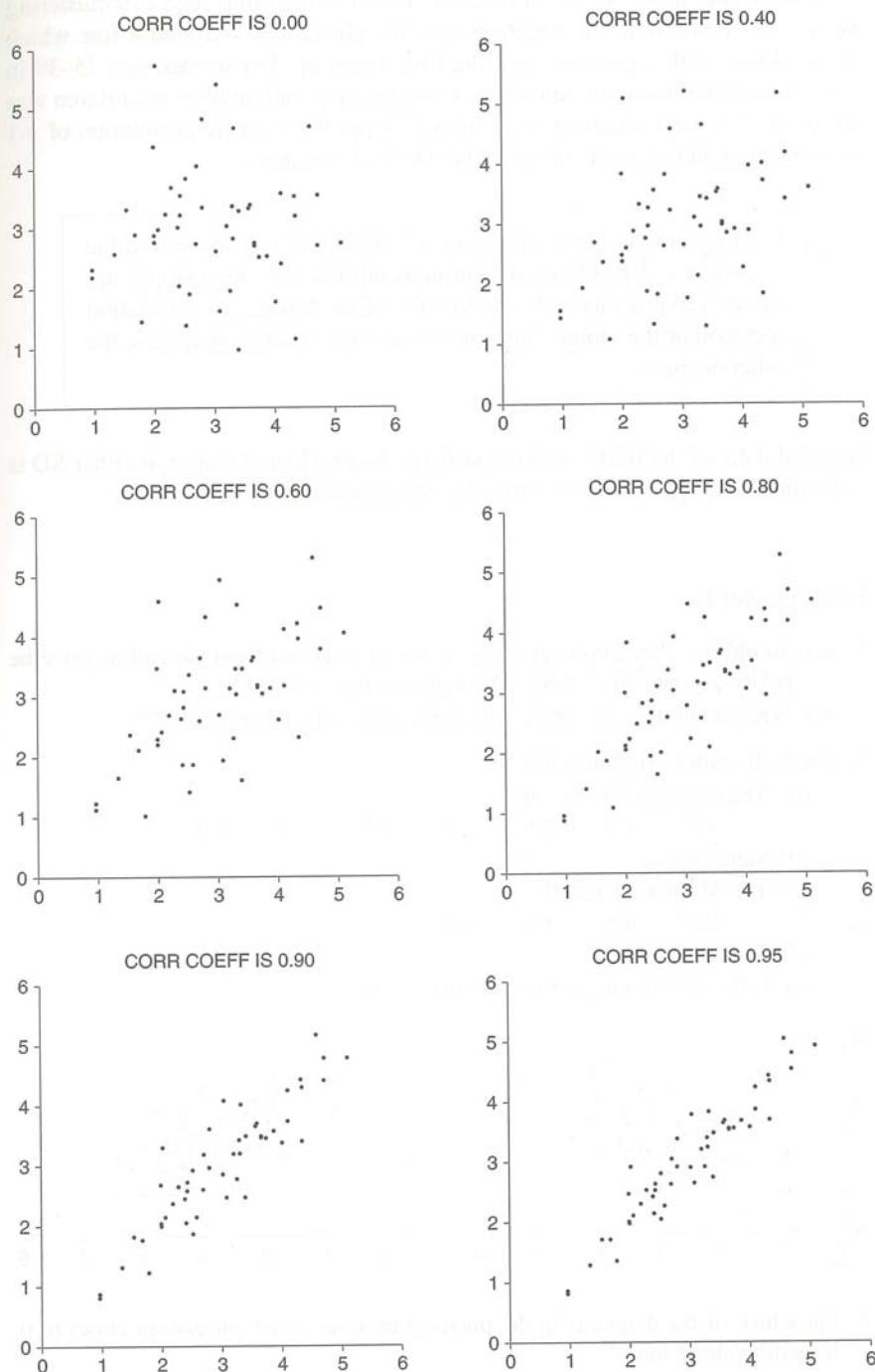
The next scatter diagram has  $r = 0.40$ ; a linear pattern is beginning to emerge. The next one has  $r = 0.60$ , with a stronger linear pattern. And so on, through the last one. The closer  $r$  is to 1, the stronger is the linear association between the variables, and the more tightly clustered are the points around a line. A correlation of 1, which does not appear in the figure, is often referred to as a *perfect correlation*—all the points lie exactly on a line, so there is a perfect linear relationship between the variables. Correlations are always 1 or less.

The correlation between the heights of identical twins is around 0.95.<sup>4</sup> The lower right scatter diagram in figure 6 has a correlation coefficient of 0.95. A scatter diagram for the twins would look about the same. Identical twins are like each other in height, and their points on a scatter diagram are fairly close to the line  $y = x$ . However, such twins do not have exactly the same height. That is what the scatter around the 45-degree line shows.

For another example, in the U.S. in 2005, the correlation between income and education was 0.07 for men age 18–24, rising to 0.43 for men age 55–64.<sup>5</sup> As the scatter diagrams in figure 6 indicate, the relationship between income and education is stronger for the older men, but it is still quite rough. Weak associations are common in social science studies, 0.3 to 0.7 being the usual range for  $r$  in many fields.

A word of warning:  $r = 0.80$  does not mean that 80% of the points are tightly clustered around a line, nor does it indicate twice as much linearity as  $r = 0.40$ . Right now, there is no direct way to interpret the exact numerical value of the correlation coefficient; that will be done in chapters 10 and 11.

Figure 6. The correlation coefficient—six positive values. The diagrams are scaled so that the average equals 3 and the SD equals 1, horizontally and vertically; there are 50 points in each diagram. Clustering is measured by the correlation coefficient.



So far, only positive association has been discussed. Negative association is indicated by a negative sign in the correlation coefficient. Figure 7 shows six more scatter diagrams for hypothetical data, each with 50 points. They are scaled just like figure 6, each variable having an average of 3 and an SD of 1.

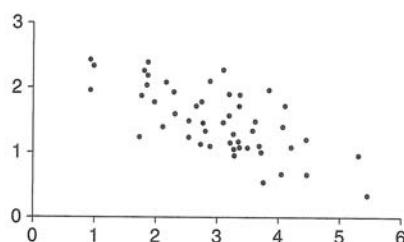
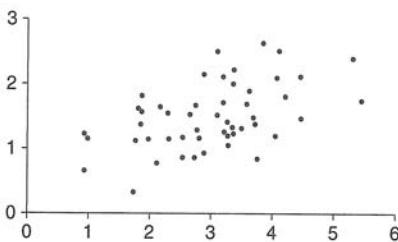
A correlation of  $-0.90$ , for instance, indicates the same degree of clustering as one of  $+0.90$ . With the negative sign, the clustering is around a line which slopes down; with a positive sign, the line slopes up. For women age 25–39 in the U.S. in 2005, the correlation between education and number of children was about  $-0.2$ , a weak negative association.<sup>6</sup> A perfect negative correlation of  $-1$  indicates that all the points lie on a line which slopes down.

Correlations are always between  $-1$  and  $1$ , but can take any value in between. A positive correlation means that the cloud slopes up; as one variable increases, so does the other. A negative correlation means that the cloud slopes down; as one variable increases, the other decreases.

In a real data set, both SDs will be positive. As a technical matter, if either SD is zero, there is no good way to define the correlation coefficient.

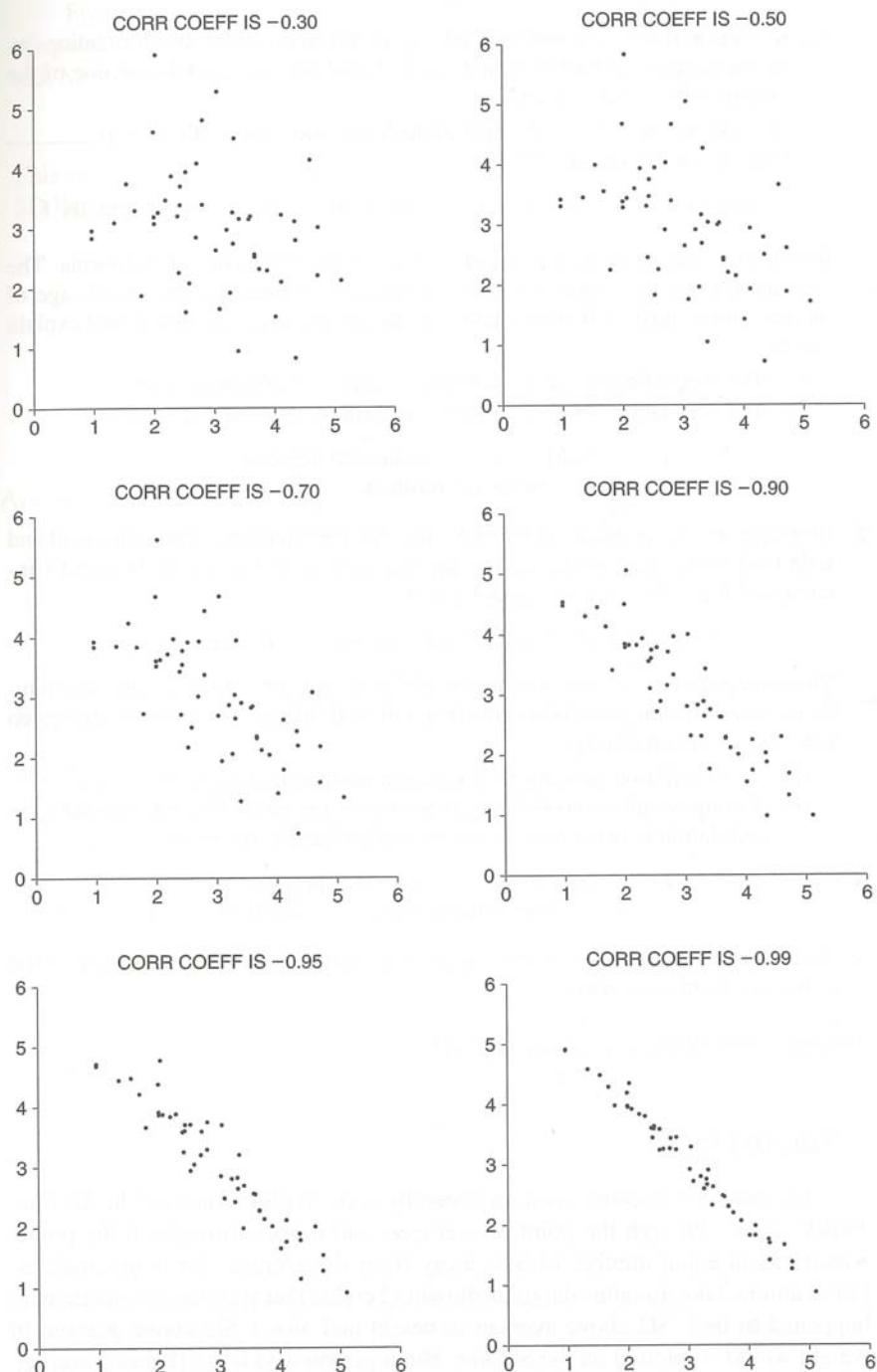
### Exercise Set B

1. (a) Would the correlation between the age of a second-hand car and its price be positive or negative? Why? (Antiques are not included.)  
 (b) What about the correlation between weight and miles per gallon?
2. For each scatter diagram below:
  - (a) The average of  $x$  is around  
 1.0    1.5    2.0    2.5    3.0    3.5    4.0
  - (b) Same, for  $y$ .
  - (c) The SD of  $x$  is around  
 0.25    0.5    1.0    1.5
  - (d) Same, for  $y$ .
  - (e) Is the correlation positive, negative, or 0?



3. For which of the diagrams in the previous exercise is the correlation closer to 0, forgetting about signs?

Figure 7. The correlation coefficient—six negative values. The diagrams are scaled so the average equals 3 and the SD equals 1, horizontally and vertically; there are 50 points in each diagram. Clustering is measured by the correlation coefficient.



4. In figure 1, is the correlation between the heights of the fathers and sons around  $-0.3$ ,  $0$ ,  $0.5$ , or  $0.8$ ?
5. In figure 1, if you took only the fathers who were taller than 6 feet, and their sons, would the correlation between the heights be around  $-0.3$ ,  $0$ ,  $0.5$  or  $0.8$ ?
6. (a) If women always married men who were five years older, the correlation between the ages of husbands and wives would be \_\_\_\_\_. Choose one of the options below, and explain.  
 (b) The correlation between the ages of husbands and wives in the U.S. is \_\_\_\_\_. Choose one option, and explain.

exactly  $-1$     close to  $-1$     close to  $0$     close to  $1$     exactly  $1$

7. Investigators are studying registered students at the University of California. The students fill out questionnaires giving their year of birth, age (in years), age of mother, and so forth. Fill in the blanks, using the options given below, and explain briefly.
- (a) The correlation between student's age and year of birth is \_\_\_\_\_.  
 (b) The correlation between student's age and mother's age is \_\_\_\_\_.  
 \_\_\_\_\_

\_\_\_\_\_    nearly  $-1$     somewhat negative  
 \_\_\_\_\_    somewhat positive    nearly  $1$     \_\_\_\_\_

8. Investigators take a sample of DINKS (dual-income families—where husband and wife both work—and no kids). The investigators have data on the husband's income and the wife's income. By definition,

$$\text{family income} = \text{husband's income} + \text{wife's income}.$$

The average family income was around \$85,000, and 10% of the couples had family income in the range \$80,000–\$90,000. Fill in the blanks, using the options given below, and explain briefly.

- (a) The correlation between wife's income and family income is \_\_\_\_\_.  
 (b) Among couples whose family income is in the range \$80,000–\$90,000, the correlation between wife's income and husband's income is \_\_\_\_\_.  
 \_\_\_\_\_

\_\_\_\_\_    nearly  $-1$     somewhat negative  
 \_\_\_\_\_    somewhat positive    nearly  $1$     \_\_\_\_\_

9. True or false, and explain: if the correlation coefficient is 0.90, then 90% of the points are highly correlated.

*The answers to these exercises are on p. A56.*

### 3. THE SD LINE

The points in a scatter diagram generally seem to cluster around the *SD line*. This line goes through the point of averages; and it goes through all the points which are an equal number of SDs away from the average, for both variables. For example, take a scatter diagram showing heights and weights. Someone who happened to be 1 SD above average in height and also 1 SD above average in weight would be plotted on the SD line. But a person who is 1 SD above average

in height and 0.5 SDs above average in weight would be off the line. Similarly, a person who is 2 SDs below average in height and also 2 SDs below average in weight would be on the line. Someone who is 2 SDs below average in height and 2.5 SDs below average in weight would be off the line.

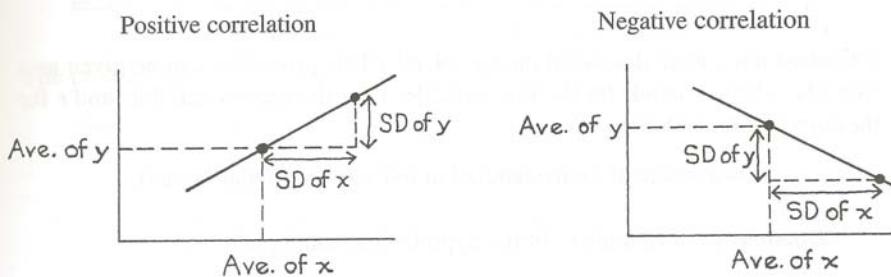
Figure 8 shows how to plot the SD line on a graph. The line goes through the point of averages, and climbs at the rate of one vertical SD for each horizontal SD. More technically, the slope is the ratio

$$(\text{SD of } y)/(\text{SD of } x).$$

This is for positive correlations. When the correlation coefficient is negative, the SD line goes down; the slope is<sup>7</sup>

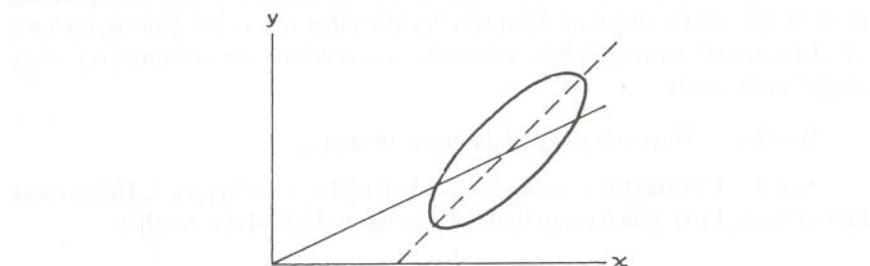
$$-(\text{SD of } y)/(\text{SD of } x).$$

Figure 8. Plotting the SD line.



### Exercise Set C

- True or false:
  - The SD line always goes through the point of averages.
  - The SD line always goes through the point (0, 0).
- For the scatter diagram shown below, say whether it is the solid line or the dashed line which is the SD line.



- One study on male college students found their average height to be 69 inches, with an SD of 3 inches. Their average weight was 140 pounds, with an SD of 20 pounds. And the correlation was 0.60. If one of these people is 72 inches tall, how heavy would he have to be to fall on the SD line?

4. Using the same data as in exercise 3, say whether each of the following students was on the SD line:
- height 75 inches, weight 180 pounds
  - height 66 inches, weight 130 pounds
  - height 66 inches, weight 120 pounds

*The answers to these exercises are on p. A57.*

#### 4. COMPUTING THE CORRELATION COEFFICIENT

Here is the procedure for computing the correlation coefficient.

Convert each variable to standard units. The average of the products gives the correlation coefficient.

(Standard units were discussed on pp. 79–80.) This procedure can be given as a formula, where  $x$  stands for the first variable,  $y$  for the second variable, and  $r$  for the correlation coefficient:

$$r = \text{average of } (x \text{ in standard units}) \times (y \text{ in standard units}).$$

*Example 1.* Compute  $r$  for the hypothetical data in table 1.

Table 1. Data.

$x$	$y$
1	5
3	9
4	7
5	1
7	13

*Note.* The first row of table 1 represents two measurements on one subject in the study; the two numbers are the  $x$ - and  $y$ -coordinates of the corresponding point on the scatter diagram. Similarly for the other rows. The pairing matters:  $r$  is defined only when you have two variables, and both are measured for every subject in the study.

*Solution.* The work can be laid out as in table 2.

*Step 1.* Convert the  $x$ -values to standard units, as in chapter 5. This is quite a lot of work. First, you have to find the average and SD of the  $x$ -values:

$$\text{average of } x\text{-values} = 4, \quad \text{SD} = 2.$$

Then, you have to subtract the average from each  $x$ -value, and divide by the SD:

$$\frac{1 - 4}{2} = -1.5 \quad \frac{3 - 4}{2} = -0.5 \quad \frac{4 - 4}{2} = 0 \quad \frac{5 - 4}{2} = 0.5 \quad \frac{7 - 4}{2} = 1.5$$

Table 2. Computing  $r$ .

$x$	$y$	$x$ in standard units	$y$ in standard units	Product
1	5	-1.5	-0.5	0.75
3	9	-0.5	0.5	-0.25
4	7	0.0	0.0	0.00
5	1	0.5	-1.5	-0.75
7	13	1.5	1.5	2.25

The results go into the third column of table 2. The numbers tell you how far above or below average the  $x$ -values are, in terms of the SD. For instance, the value 1 is 1.5 SDs below average.

*Step 2.* Convert the  $y$ -values to standard units; the results go into the fourth column of the table. That finishes the worst of the arithmetic.

*Step 3.* For each row of the table, work out the product

$$(x \text{ in standard units}) \times (y \text{ in standard units})$$

The products go into the last column of the table.

*Step 4.* Take the average of the products:

$$r = \text{average of } (x \text{ in standard units}) \times (y \text{ in standard units})$$

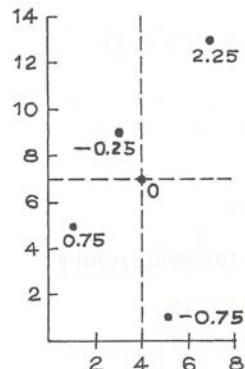
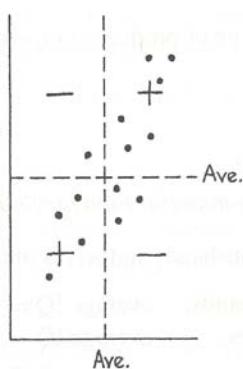
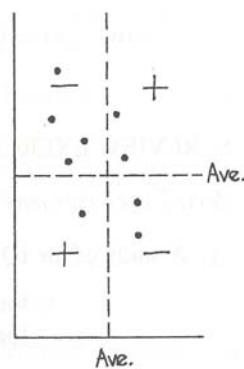
$$= \frac{0.75 - 0.25 + 0.00 - 0.75 + 2.25}{5} = 0.40$$

This completes the solution. If you plot a scatter diagram for the data (figure 9a), the points slope up but are only loosely clustered.

Why does  $r$  work as a measure of association? In figure 9a, the products are marked at the corresponding dots. Horizontal and vertical lines are drawn through the point of averages, dividing the scatter diagram into four quadrants. If a point is in the lower left quadrant, both variables are below average and are negative in

Figure 9. How the correlation coefficient works.

(a) Scatter diagram from Table 1

(b) Positive  $r$ (c) Negative  $r$ 

standard units; the product of two negatives is positive. In the upper right quadrant, the product of two positives is positive. In the remaining two quadrants, the product of a positive and a negative is negative. The average of all these products is the correlation coefficient. If  $r$  is positive, then points in the two positive quadrants will predominate, as in figure 9b. If  $r$  is negative, points in the two negative quadrants will predominate, as in figure 9c.

### Exercise Set D

- For each of the data sets shown below, calculate  $r$ .

(a)		(b)		(c)	
$x$	$y$	$x$	$y$	$x$	$y$
1	6	1	2	1	7
2	7	2	1	2	6
3	5	3	4	3	5
4	4	4	3	4	4
5	3	5	7	5	3
6	1	6	5	6	2
7	2	7	6	7	1

- Find the scatter diagram in figure 6 (p. 127) with a correlation of 0.95. In this diagram, the percentage of points where both variables are simultaneously above average is around

5%      25%      50%      75%      95%.

- Repeat exercise 2, for a correlation of 0.00.
- Using figure 7, repeat exercise 2 for a correlation of  $-0.95$ .

*The answers to these exercises are on p. A57.*

*Technical note.* There is another way to compute  $r$ , which is sometimes useful:<sup>8</sup>

$$r = \frac{\text{cov}(x, y)}{(\text{SD of } x) \times (\text{SD of } y)}$$

where

$$\text{cov}(x, y) = (\text{average of products } xy) - (\text{ave of } x) \times (\text{ave of } y).$$

### 5. REVIEW EXERCISES

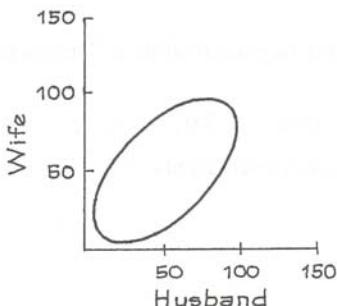
*Review exercises may cover material from previous chapters.*

- A study of the IQs of husbands and wives obtained the following results:

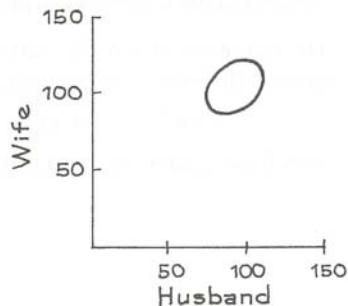
$$\begin{aligned} \text{for husbands, } & \text{average IQ} = 100, \quad \text{SD} = 15 \\ \text{for wives, } & \text{average IQ} = 100, \quad \text{SD} = 15 \\ & r = 0.6 \end{aligned}$$

One of the following is a scatter diagram for the data. Which one? Say briefly why you reject the others.

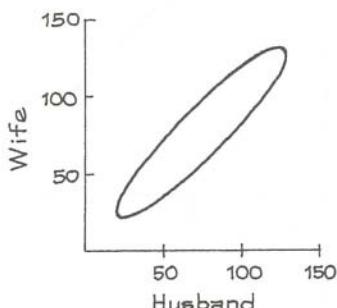
(a)



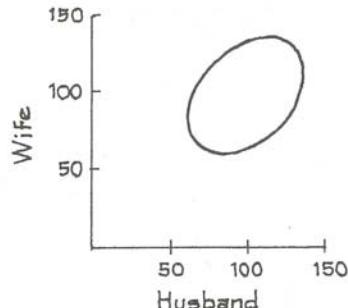
(b)



(c)



(d)



2. (a) For a representative sample of cars, would the correlation between the age of the car and its gasoline economy (miles per gallon) be positive or negative?  
 (b) The correlation between gasoline economy and income of owner turns out to be positive.<sup>9</sup> How do you account for this positive association?
3. Suppose men always married women who were exactly 8% shorter. What would the correlation between their heights be?
4. Is the correlation between the heights of husbands and wives in the U.S. around  $-0.9$ ,  $-0.3$ ,  $0.3$ , or  $0.9$ ? Explain briefly.
5. Three data sets are collected, and the correlation coefficient is computed in each case. The variables are
  - (i) grade point average in freshman year and in sophomore year
  - (ii) grade point average in freshman year and in senior year
  - (iii) length and weight of two-by-four boards

Possible values for correlation coefficients are

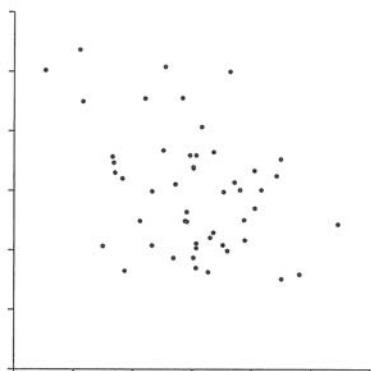
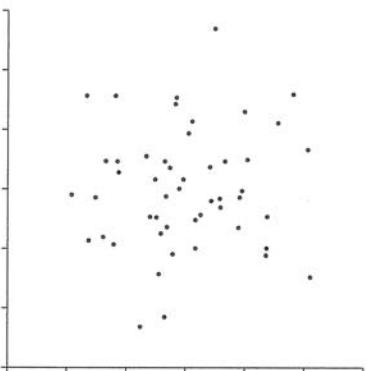
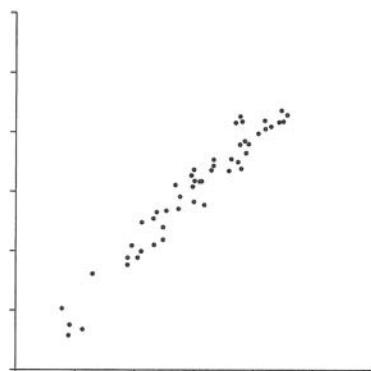
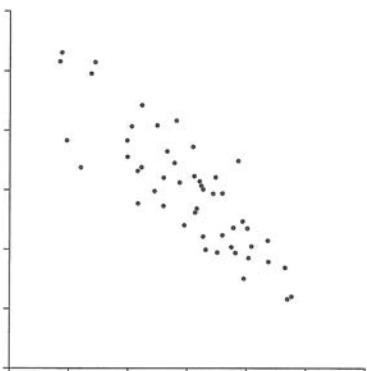
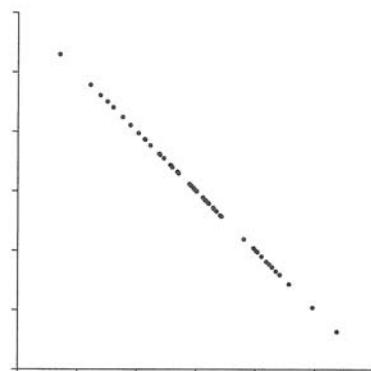
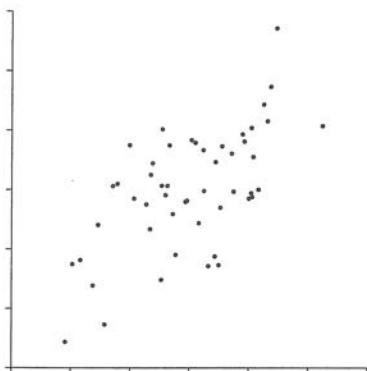
$-0.50 \quad 0.0 \quad 0.30 \quad 0.60 \quad 0.95$

Match the correlations with the data sets; two will be left over. Explain your choices.

6. In one class, the correlation between scores on the final and the midterm was 0.50, while the correlation between the scores on the final and the homework was 0.25. True or false, and explain: the relationship between the final scores and the midterm scores is twice as linear as the relationship between the final scores and the homework scores.
7. The figure below has six scatter diagrams for hypothetical data. The correlation coefficients, in scrambled order, are:

-0.85      -0.38      -1.00      0.06      0.97      0.62

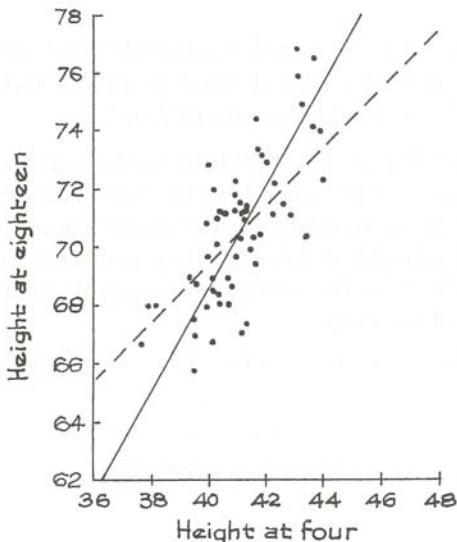
Match the scatter diagrams with the correlation coefficients.



8. A longitudinal study of human growth was started in 1929 at the Berkeley Institute of Human Development.<sup>10</sup> The scatter diagram below shows the heights of 64 boys, measured at ages 4 and 18.

- (a) The average height at age 4 is around  
38 inches    42 inches    44 inches
- (b) The SD of height at age 18 is around  
0.5 inches    1.0 inches    2.5 inches
- (c) The correlation coefficient is around  
0.50    0.80    0.95
- (d) Which is the SD line—solid or dashed?

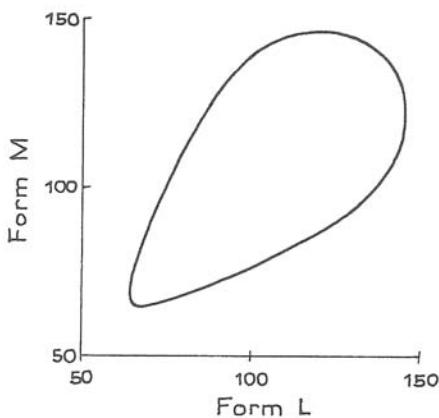
Explain your answers.



9. Find the correlation coefficient for each of the three data sets shown below.

(a)		(b)		(c)	
$x$	$y$	$x$	$y$	$x$	$y$
1	5	1	1	1	2
1	3	1	2	1	2
1	5	1	1	1	2
1	7	1	3	1	2
2	3	2	1	2	4
2	3	2	4	2	4
2	1	2	1	2	4
3	1	3	2	3	6
3	1	3	2	3	6
4	1	4	3	4	8

10. In a large psychology study, each subject took two IQ tests (form L and form M of the Stanford-Binet). A scatter diagram for the test scores is sketched at the top of the next page. You are trying to predict the score on



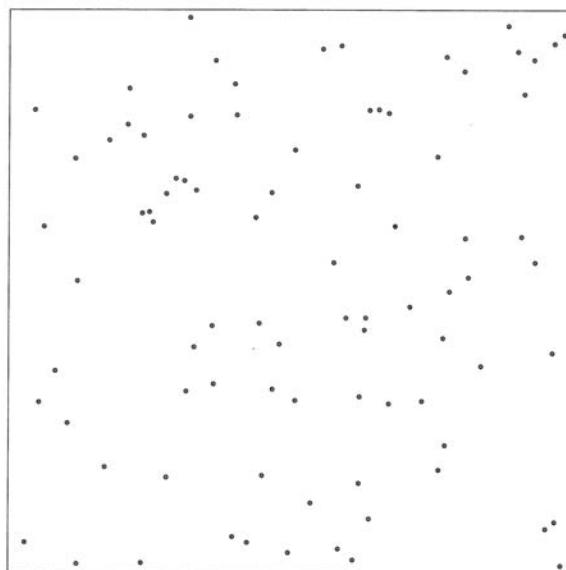
form M from the score on form L. Each prediction is off by some amount. On the whole, will these prediction errors be smaller when the score on form L is 75, or 125? or is it about the same for both?

11. A teaching assistant gives a quiz with 10 questions and no part credit. After grading the papers, the TA writes down for each student the number of questions the student got right and the number wrong. The average number of right answers is 6.4 with an SD of 2.0; the average number of wrong answers is 3.6 with the same SD of 2.0. The correlation between the number of right answers and the number of wrongs is

0      -0.50      +0.50      -1      +1      can't tell without the data

Explain.

*Figure for exercise 12*



12. Fifteen students in an elementary statistics course at U.C. Berkeley were asked to count the dots in a figure like the one at the bottom of the previous page; there were 85 dots in the figure. The counts are shown in the table below. Make a scatter diagram for the counts. Represent each student by one point on your diagram, showing the first and second count. Label both your axes fully. Choose the scale so you can see the pattern in the points. Use your scatter diagram to answer the following questions:
- Did the students work independently?
  - True or false: those students who counted high the first time also tended to be high the second time.

*The two counts*

<i>1st</i>	<i>2nd</i>
91	85
81	83
86	85
83	84
85	85
85	84
85	89
84	83
91	82
91	82
91	82
85	85
85	85
87	85
90	85

## 6. SUMMARY

- The relationship between two variables can be represented by a *scatter diagram*. When the scatter diagram is tightly clustered around a line, there is a strong *linear association* between the variables.
- A scatter diagram can be summarized by means of five statistics:
  - the average of the  $x$ -values, the SD of the  $x$ -values,
  - the average of the  $y$ -values, the SD of the  $y$ -values,
  - the *correlation coefficient*  $r$ .
- Positive association (a cloud which slopes up) is indicated by a plus-sign in the correlation coefficient. Negative association (a cloud which slopes down) is indicated by a minus-sign.
- In a series of scatter diagrams with the same SDs, as  $r$  gets closer to  $\pm 1$ , the points cluster more tightly around a line.

5. The correlation coefficient ranges from  $-1$  (when all the points lie on a line which slopes down), to  $+1$  (when all the points lie on a line which slopes up).

6. The *SD line* goes through the point of averages. When  $r$  is positive, the slope of the line is

$$(\text{SD of } y)/(\text{SD of } x).$$

When  $r$  is negative, the slope is

$$-(\text{SD of } y)/(\text{SD of } x).$$

7. To calculate the correlation coefficient, convert each variable to standard units, and then take the average product.

# 9

## More about Correlation

*“Very true,” said the Duchess: “flamingoes and mustard both bite. And the moral of that is—‘Birds of a feather flock together.’”*

*“Only mustard isn’t a bird,” Alice remarked.*

*“Right, as usual,” said the Duchess: “what a clear way you have of putting things!”*

*—Alice in Wonderland*

### 1. FEATURES OF THE CORRELATION COEFFICIENT

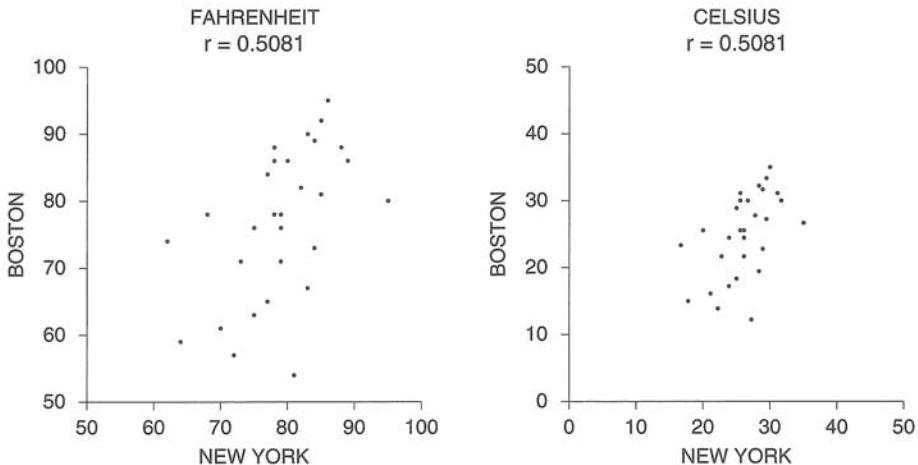
The correlation coefficient is a pure number. Why? Because the first step in computing  $r$  is a conversion to standard units. Original units—like inches for height data or degrees for temperature data—cancel out. In a similar way,  $r$  is not affected if you multiply all the values of one variable by the same positive number, or if you add the same number to all the values of one variable. (As a statistician might say,  $r$  is not affected by *changes of scale*; see pp. 92–93.)

For example, if you multiply each value of  $x$  by 3, then the average gets multiplied by 3. All the deviations from average get multiplied by 3 as well, and so does the SD. This common factor cancels in the conversion to standard units. So  $r$  stays the same. For another example, suppose you add 7 to each value of  $x$ . Then the average of  $x$  goes up by 7 too. However, the deviations from average do not change. Neither does  $r$ .

Figure 1 (on the next page) shows the correlation between daily maximum temperatures at New York and Boston. There is a dot in the diagram for each day of June 2005. The temperature in New York that day is plotted on the horizontal axis; the Boston temperature, on the vertical. The left hand panel does it in degrees

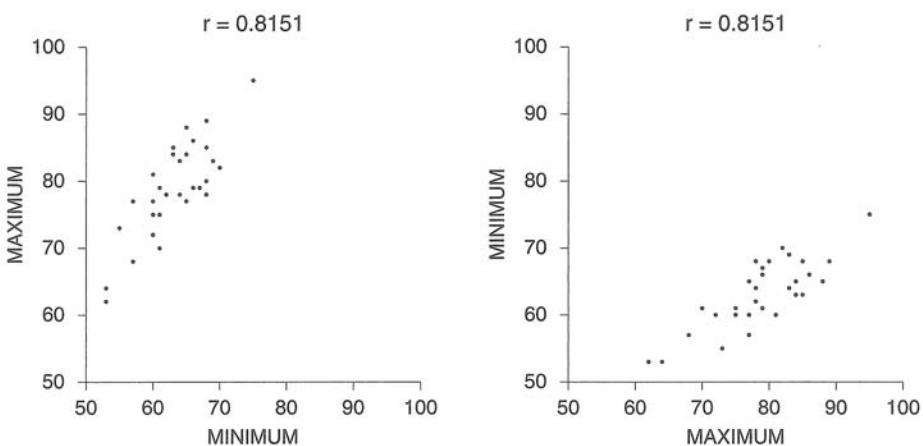
Fahrenheit, and  $r = 0.5081$ . The right hand panel does it in degrees Celsius, and  $r$  stays the same.<sup>1</sup> The conversion from Fahrenheit to Celsius is just a change of scale, which does not affect the correlation.

Figure 1. Daily maximum temperatures. New York and Boston, June 2005. The left hand panel plots the data in degrees Fahrenheit; the right hand panel, in degrees Celsius. This does not change  $r$ .



Another feature: The correlation between  $x$  and  $y$  is the same as the correlation between  $y$  and  $x$ . For example, the left hand panel in figure 2 is a scatter diagram for temperature data at New York in June 2005. The minimum tempera-

Figure 2. Daily temperatures. New York, June 2005.



ture each day is plotted on the horizontal axis; the maximum, on the vertical. The correlation between the minimum and the maximum temperature is 0.8151. The right hand panel shows exactly the same data. This time, the minimum is plotted on the vertical instead of the horizontal. The pictures look different because the points are reflected around the diagonal. But  $r$  stays the same. Switching the order of the variables does not affect  $r$ . Why? Remember,  $r$  is the average of the products after conversion to standard units. Products do not depend on the order of the factors ( $a \times b = b \times a$ ). It may be surprising that the correlation is only 0.8151, but the weather is full of surprises.

The correlation coefficient is a pure number, without units. It is not affected by

- interchanging the two variables,
- adding the same number to all the values of one variable,
- multiplying all the values of one variable by the same positive number.

### Exercise Set A

1. (a) In June 2005, which city was warmer—Boston or New York? Or were they about the same?  
 (b) In the left hand panel of figure 2, all the dots are above the 45-degree line. Why?
2. A small data set is shown below;  $r \approx 0.76$ . If you switch the two columns, does this change  $r$ ? Explain or calculate.

$x$	$y$
1	2
2	3
3	1
4	5
5	6

3. As in exercise 2, but you add 3 to each value of  $y$  instead of interchanging the columns.
4. As in exercise 2, but you double each value of  $x$ .
5. As in exercise 2, but you interchange the last two values (5 and 6) for  $y$ .
6. Suppose the correlation between  $x$  and  $y$  is 0.73.
  - (a) Does the scatter diagram slope up or down?
  - (b) If you multiply all the values of  $y$  by  $-1$ , would the new scatter diagram slope up or down?
  - (c) If you multiply all the values of  $y$  by  $-1$ , what happens to the correlation?

7. Two different investigators are working on a growth study. The first measures the heights of 100 children, in inches. The second prefers the metric system, and changes the results to centimeters (multiplying by the conversion factor 2.54 centimeters per inch). A scatter diagram is plotted, showing for each child its height in inches on the horizontal axis, and height in centimeters on the vertical axis.
- If no mistakes are made in the conversion, what is the correlation?
  - What happens to  $r$  if mistakes are made in the arithmetic?
  - What happens to  $r$  if the second investigator goes out and measures the same children again, using metric equipment?
8. In figure 1 on p. 120, the correlation is 0.5. Suppose we plot on the horizontal axis the height of the paternal grandfather (not the father); the height of the son is still plotted on the vertical axis. Would the correlation be more or less than 0.5?
9. Two weathermen compute the correlation between daily maximum temperatures for Washington and Boston. One does it for June; the other does it for the whole year. Who gets the bigger correlation? ("Washington" is the city, not the state.)
10. Six data sets are shown below. In (i), the correlation is 0.8571, and in (ii) the correlation is 0.7857. Find the correlations for the remaining data sets. No arithmetic is necessary.

(i)		(ii)		(iii)		(iv)		(v)		(vi)	
$x$	$y$	$x$	$y$	$x$	$y$	$x$	$y$	$x$	$y$	$x$	$y$
1	2	1	2	2	1	2	2	1	4	0	6
2	3	2	3	3	2	3	3	2	6	1	9
3	1	3	1	1	3	4	1	3	2	2	3
4	4	4	4	4	4	5	4	4	8	3	12
5	6	5	6	6	5	6	6	5	12	4	18
6	5	6	7	7	6	7	5	6	10	5	21
7	7	7	5	5	7	8	7	7	14	6	15

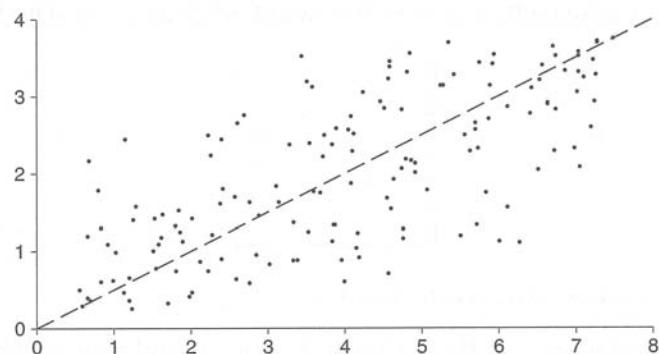
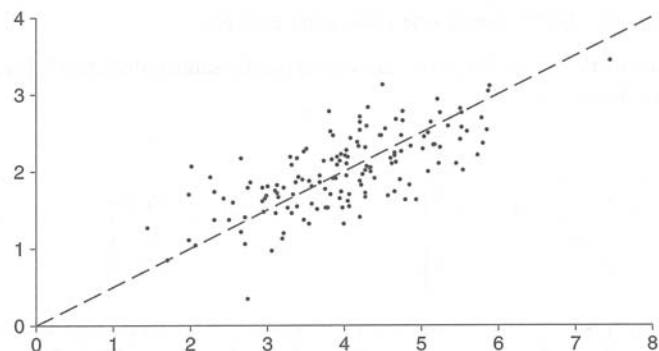
The answers to these exercises are on pp. A57–58.

## 2. CHANGING SDs

The appearance of a scatter diagram depends on the SDs. For instance, look at figure 3. In both diagrams,  $r$  is 0.70. However, the top one looks more tightly clustered around the SD line. That is because its SDs are smaller. The formula for  $r$  involves converting the variables to standard units: deviations from average are divided by the SD. So,  $r$  measures clustering not in absolute terms but in relative terms—relative to the SDs.

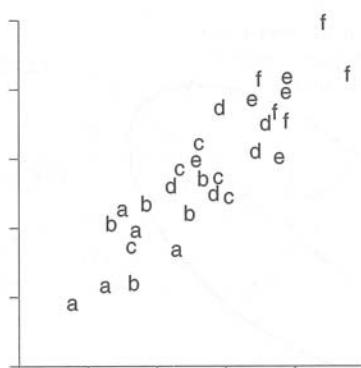
To interpret a correlation coefficient graphically, draw the scatter diagram in your mind's eye so the vertical SD covers the same distance on the page as the vertical SDs in figure 6 on p. 127; and likewise for the horizontal SD. If  $r$  for your scatter diagram is 0.40, it will probably show about the same amount of clustering around the diagonal as the one with an  $r$  of 0.40 in the figure at the top right. If  $r$  is 0.90, it will look like the diagram in the figure at the bottom left. In general, your scatter diagram will match the one that has a similar value for  $r$ .

Figure 3. The effect of changing SDs. The two scatter diagrams have the same correlation coefficient of 0.70. The top diagram looks more tightly clustered around the SD line because its SDs are smaller.

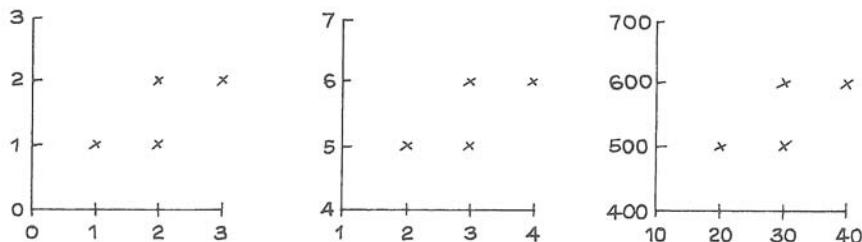


### Exercise Set B

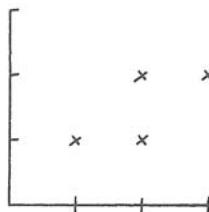
1. In the figure below, 6 scatter diagrams are plotted on the same pair of axes; in the first, the points are marked "a"; in the second, "b"; and so forth. For each of the 6 diagrams taken on its own, the correlation is around 0.6. Now take all the points together. For the combined diagram, is the correlation around 0.0, 0.6, or 0.9?



2. The National Health and Nutrition Examination Survey (p. 58) also covers children. In HANES2, at each age from 6 to 11, the correlation between height and weight was just about 0.67. For all the children together, would the correlation between height and weight be just about 0.67, somewhat more than 0.67, or somewhat less than 0.67? Choose one option and explain.
3. Below are three scatter diagrams. Do they have the same correlation? Try to answer without calculating.



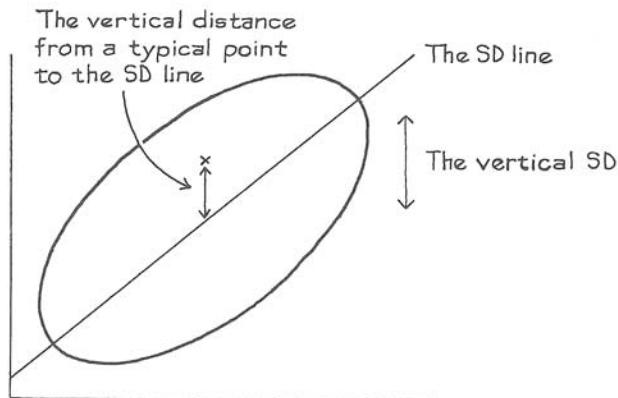
4. Someone hands you the scatter diagram shown below, but has forgotten to label the axes. Can you still calculate  $r$ ? If so, what is it? Or do you need the labels?



*The answers to these exercises are on p. A58.*

*Technical notes.* (i) If  $r$  is close to 1, then a typical point is only a small fraction of a vertical SD above or below the SD line. If  $r$  is close to 0, then a typical point is above or below the line by an amount roughly comparable in size to the vertical SD: see figure 4. (The “vertical SD” is the SD of the variable plotted on the  $y$ -axis.)

Figure 4. The correlation coefficient. As  $r$  gets close to 1, the distance of a typical point above or below the SD line becomes a small fraction of the vertical SD.



(ii) The connection between the correlation coefficient and the typical distance above or below the SD line can be expressed mathematically, as follows. The r.m.s. vertical distance to the SD line equals

$$\sqrt{2(1 - |r|)} \times \text{the vertical SD}$$

Take, for example, a correlation of 0.95. Then

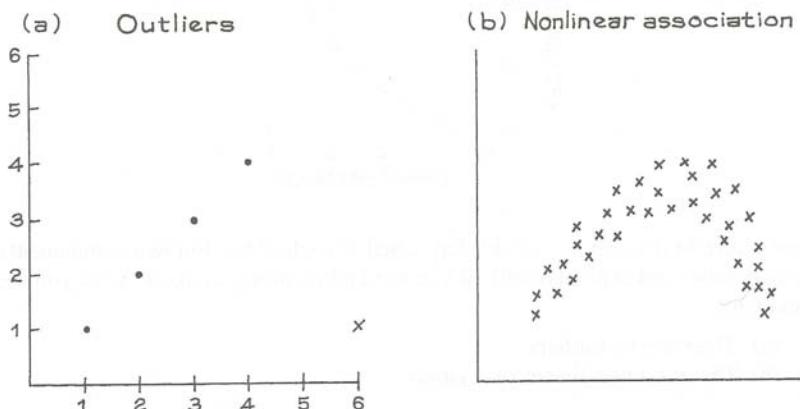
$$\sqrt{2(1 - |r|)} = \sqrt{0.1} \approx 0.3$$

So the spread around the SD line is about 30% of a vertical SD. That is why a scatter diagram with  $r = 0.95$  shows a fair amount of spread around the line (figure 6 on p. 127). There are similar formulas for the horizontal direction.

### 3. SOME EXCEPTIONAL CASES

The correlation coefficient is useful for football-shaped scatter diagrams. For other diagrams,  $r$  can be misleading. Outliers and non-linearity are problem cases. In figure 5a, the dots show a perfect correlation of 1. The outlier, marked by a cross, brings the correlation down almost to 0. Figure 5a should not be summarized using  $r$ . Some people get carried away in pursuit of outliers. However, in any scatter diagram there will be some points more or less detached from the main part of the cloud. These points should be rejected only if there is good reason to do so.

Figure 5. The correlation coefficient can be misleading in the presence of outliers or non-linear association.

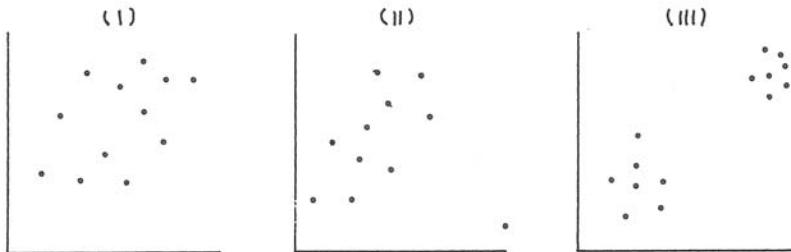


In figure 5b, the correlation coefficient is close to 0, even though the points show a strong association. The reason is that the graph does not look at all like a straight line: as  $x$  increases,  $y$  rises then falls. This pattern is shown by the association between weight and age for adult men (figure 3 on p. 59). Again, such data should not be summarized using  $r$ —the pattern gets lost.

$r$  measures linear association, not association in general.

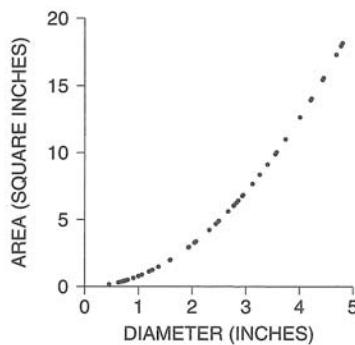
## Exercise Set C

1. Which of the following three scatter diagrams should be summarized by  $r$ ?



2. A class of 15 students happens to include 5 basketball players. True or false, and explain: the relationship between heights and weights for this class should be summarized using  $r$ .
3. A circle of diameter  $d$  has area  $\frac{1}{4}\pi d^2$ . An investigator plots a scatter diagram of area against diameter for a sample of circles with different diameters. (The diagram is shown below.) The correlation coefficient is \_\_\_\_\_. Fill in the blank, and explain. Options:

-1      nearly -1      nearly 0      nearly 1      1



4. For a certain data set,  $r = 0.57$ . Say whether each of the following statements is true or false, and explain briefly; if you need more information, say what you need, and why.
- There are no outliers.
  - There is a non-linear association.

*The answers to these exercises are on p. A58.*

#### 4. ECOLOGICAL CORRELATIONS

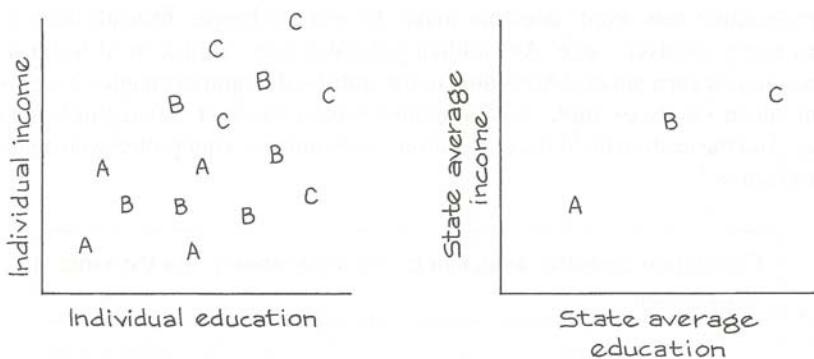
In 1955, Sir Richard Doll published a landmark article on the relationship between cigarette smoking and lung cancer.<sup>2</sup> One piece of evidence was a scatter diagram showing the relationship between the rate of cigarette smoking (per capita) and the rate of deaths from lung cancer in eleven countries. The correla-

tion between these eleven pairs of rates was 0.7, and this was taken as showing the strength of the relationship between smoking and cancer. However, it is not countries which smoke and get cancer, but people. To measure the strength of the relationship for people, it is necessary to have data relating smoking and cancer for individuals rather than countries. Such studies are available, and show that smoking does indeed cause cancer.

The statistical point: correlations based on rates or averages can be misleading. Here is another example. From Current Population Survey data for 2005, you can compute the correlation between income and education for men age 25–64 in the United States:  $r \approx 0.42$ . For each state (and D.C.), you can compute average educational level and average income. Finally, you can compute the correlation between the 51 pairs of averages:  $r \approx 0.70$ . If you used the correlation for the states to estimate the correlation for the individuals, you would be way off. The reason is that within each state, there is a lot of spread around the averages. Replacing the states by their averages eliminates the spread, and gives a misleading impression of tight clustering. Figure 6 shows the effect for three states.<sup>3</sup>

*Ecological* correlations are based on rates or averages. They are often used in political science and sociology. And they tend to overstate the strength of an association. So watch out.

Figure 6. Ecological correlations (based on rates or averages) are usually too big. The panel on the left represents income and education for individuals in three states, labeled A, B, C. Each individual is marked by a letter showing state of residence. The correlation is moderate. The panel on the right shows the averages for each state. The correlation between the averages is almost 1.



#### Exercise Set D

1. The table at the top of the next page is adapted from Doll and shows per capita consumption of cigarettes in various countries in 1930, and the death rates from lung cancer for men in 1950. (In 1930, hardly any women smoked; and a long period of time is needed for the effects of smoking to show up.)

<i>Country</i>	<i>Cigarette consumption</i>	<i>Deaths per million</i>
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1,100	350
Great Britain	1,100	460
Iceland	230	60
Netherlands	490	240
Norway	250	90
Sweden	300	110
Switzerland	510	250
U.S.	1,300	200

- (a) Plot a scatter diagram for these data.
- (b) True or false: the higher cigarette consumption was in 1930 in one of these countries, on the whole the higher the death rate from lung cancer in 1950. Or can this be determined from the data?
- (c) True or false: death rates from lung cancer tend to be higher among those persons who smoke more. Or can this be determined from the data?
2. A sociologist is studying the relationship between suicide and literacy in nineteenth-century Italy.<sup>4</sup> He has data for each province, showing the percentage of literates and the suicide rate in that province. The correlation is 0.6. Does this give a fair estimate of the strength of the association between literacy and suicide?

*The answers to these exercises are on p. A59.*

## 5. ASSOCIATION IS NOT CAUSATION

For school children, shoe size is strongly correlated with reading skills. However, learning new words does not make the feet get bigger. Instead, there is a third factor involved—age. As children get older, they learn to read better and they outgrow their shoes. (According to the statistical jargon of chapter 2, age is a confounder.) In the example, the confounder was easy to spot. Often, this is not so easy. And the arithmetic of the correlation coefficient does not protect you against third factors.<sup>5</sup>

Correlation measures association. But association is not the same as causation.

*Example 1. Education and unemployment.* During the Great Depression of 1929–1933, better-educated people tended to have shorter spells of unemployment. Does education protect you against unemployment?

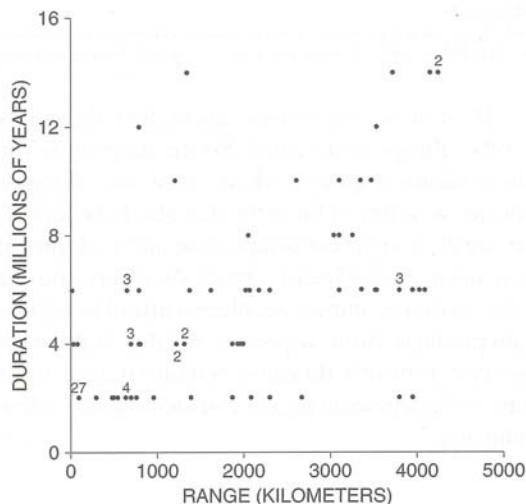
*Discussion.* Perhaps, but the data were observational. As it turned out, age was a confounding variable. The younger people were better educated, because

the educational level had been going up over time. (It still is.) Given a choice in hiring, employers seemed to prefer younger job-seekers. Controlling for age made the effect of education on unemployment much weaker.<sup>6</sup>

*Example 2. Range and duration of species.* Does natural selection operate at the level of species? This is a question of some interest for paleontologists. David Jablonski argues that geographical range is a heritable characteristic of species: a species with a wide range survives longer, because if a disaster strikes in one place, the species stays alive at other places.

One piece of evidence is a scatter diagram (figure 7). Ninety-nine species of gastropods (slugs, snails, etc.) are represented in the diagram. The duration of the species—its lifetime, in millions of years—is plotted on the vertical axis; its range is on the horizontal, in kilometers. Both variables are determined from the fossil record. There is a good positive association:  $r$  is about 0.64. (The cloud looks formless, but that is because of a few straggling points at the bottom right and the top left.) Does a wide geographical range promote survival of the species?

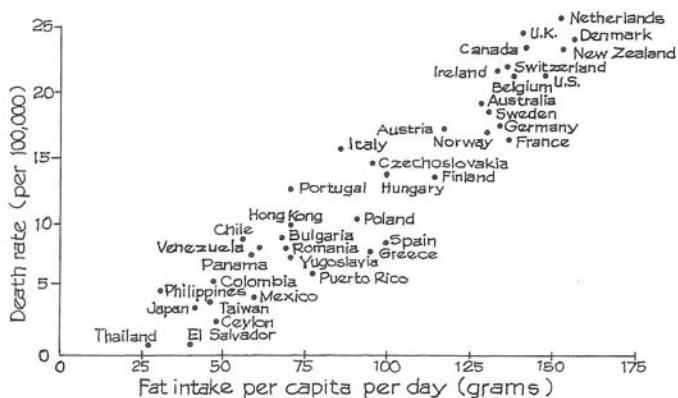
Figure 7. Duration of species in millions of years plotted against geographical range in kilometers, for 99 species of gastropods. Several species can be plotted at the same point; the number of such species is indicated next to the point.



*Discussion.* A wide range may cause a long lifetime. Or, a long lifetime may cause a wide range. Or, there may be something else going on. Jablonski had his eye on the first possibility. The second one is unlikely, because other evidence suggests that species achieve their ranges very soon after they emerge. But what about the third explanation? Michael Russell and David Lindberg point out that species with a wide geographical range have more chances to be preserved in the fossil record, which can create the appearance of a long lifetime. If so, figure 7 is a statistical artifact.<sup>7</sup> Association is not causation.

*Example 3. Fat in the diet and cancer.* In countries where people eat lots of fat—like the U.S.—rates of breast cancer and colon cancer are high. See figure 8 for data on breast cancer. This correlation is often used to argue that fat in the diet causes cancer. How good is the evidence?

Figure 8. Death rates from breast cancer plotted against fat in the diet, for a sample of countries.



Note: Age standardized.

Source: K. Carroll, "Experimental evidence of dietary factors and hormone-dependent cancers," *Cancer Research* vol. 35 (1975) p. 3379. Copyright by *Cancer Research*. Reproduced by permission.

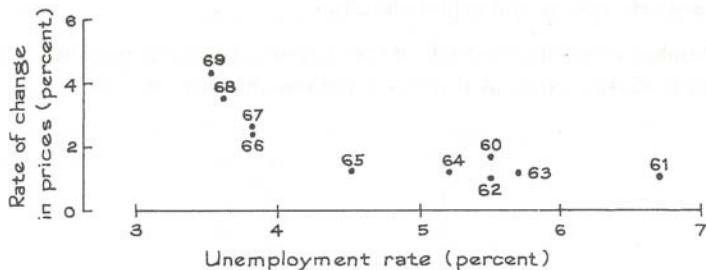
*Discussion.* If fat in the diet causes cancer, then the points in the diagram should slope up, other things being equal. So the diagram is some evidence for the theory. But the evidence is quite weak, because other things aren't equal. For example, the countries with lots of fat in the diet also have lots of sugar. A plot of breast cancer rates against sugar consumption would look just like figure 8, and nobody thinks that sugar causes breast cancer. As it turns out, fat and sugar are relatively expensive. In rich countries, people can afford to eat fat and sugar rather than starchier grain products. Some aspects of the diet in these countries, or other factors in the life-style, probably do cause certain kinds of cancer—and protect against other kinds. So far, epidemiologists can identify only a few of these factors with any real confidence.<sup>8</sup>

### Exercise Set E

1. The scatter diagram in figure 7 shows stripes. Why?
2. Is the correlation in figure 8 ecological? How is that relevant to the argument?
3. The correlation between height and weight among men age 18–74 in the U.S. is about 0.40. Say whether each conclusion below follows from the data; explain your answer.
  - (a) Taller men tend to be heavier.
  - (b) The correlation between weight and height for men age 18–74 is about 0.40.
  - (c) Heavier men tend to be taller.

- (d) If someone eats more and puts on 10 pounds, he is likely to get somewhat taller.
4. Studies find a negative correlation between hours spent watching television and scores on reading tests.<sup>9</sup> Does watching television make people less able to read? Discuss briefly.
5. Many studies have found an association between cigarette smoking and heart disease. One study found an association between coffee drinking and heart disease.<sup>10</sup> Should you conclude that coffee drinking causes heart disease? Or can you explain the association between coffee drinking and heart disease in some other way?
6. Many economists believe that there is trade-off between unemployment and inflation: low rates of unemployment will cause high rates of inflation, while higher rates of unemployment will reduce the rate of inflation. The relationship between the two variables is shown below for the U.S. in the decade 1960–69. There is one point for each year, with the rate of unemployment that year shown on the  $x$ -axis, and the rate of inflation shown on the  $y$ -axis. The points fall very close to a smooth curve known as the *Phillips Curve*. Is this an observational study or a controlled experiment? If you plotted the points for the 1970s or the 1950s, would you expect them to fall along the curve?

The Phillips curve for the 1960s:  
*Economic Report of the President (1975)*



The answers to these exercises are on p. A59.

## 6. REVIEW EXERCISES

Review exercises may cover material from previous chapters.

- When studying one variable, you can use a graph called a \_\_\_\_\_. When studying the relationship between two variables, you can use a graph called a \_\_\_\_\_.
- True or false, and explain briefly:
  - If the correlation coefficient is  $-0.80$ , below-average values of the dependent variable are associated with below-average values of the independent variable.
  - If  $y$  is usually less than  $x$ , the correlation coefficient between  $x$  and  $y$  will be negative.

3. In each case, say which correlation is higher, and explain briefly. (Data are from a longitudinal study of growth.)
- Height at age 4 and height at age 18, height at age 16 and height at age 18.
  - Height at age 4 and height at age 18, weight at age 4 and weight at age 18.
  - Height and weight at age 4, height and weight at age 18.
4. An investigator collected data on heights and weights of college students; results can be summarized as follows.

	Average	SD
Men's height	70 inches	3 inches
Men's weight	144 pounds	21 pounds
Women's height	64 inches	3 inches
Women's weight	120 pounds	21 pounds

The correlation coefficient between height and weight for the men was about 0.60; for the women, it was about the same. If you take the men and women together, the correlation between height and weight would be \_\_\_\_\_.

just about 0.60      somewhat lower      somewhat higher

Choose one option, and explain briefly.

5. A number is missing in each of the data sets below. If possible, fill in the blank to make  $r$  equal to 1. If this is not possible, say why not.

(a)		(b)	
$x$	$y$	$x$	$y$
1	1	1	1
2	3	2	3
2	3	3	4
4	—	4	—

6. A computer program prints out  $r$  for the two data sets shown below. Is the program working correctly? Answer yes or no, and explain briefly.

(i)		(ii)	
$x$	$y$	$x$	$y$
1	2	1	5
2	1	2	4
3	4	3	7
4	3	4	6
5	7	5	10
6	5	6	8
7	6	7	9

$r = 0.8214$        $r = 0.7619$

7. In 1910, Hiram Johnson entered the California gubernatorial primaries. For each county, data are available to show the percentage of native-born Americans in that county, as well as the percentage of the vote for Johnson. A

political scientist calculated the correlation between these percentages.<sup>11</sup> It is 0.5. Is this a fair measure of the extent to which "Johnson received native, as opposed to immigrant, support?" Answer yes or no, and explain briefly.

8. For women age 25 and over in the U.S. in 2005, the relationship between age and educational level (years of schooling completed) can be summarized as follows.<sup>12</sup>

$$\text{average age} \approx 50 \text{ years}, \quad \text{SD} \approx 16 \text{ years}$$

$$\text{average ed. level} \approx 13.2 \text{ years}, \quad \text{SD} \approx 3.0 \text{ years}, \quad r \approx -0.20$$

True or false, and explain: as you get older, you become less educated. If this statement is false, what accounts for the negative correlation?

9. At the University of California, Berkeley, Statistics 2 is a large lecture course with small discussion sections led by teaching assistants. As part of a study, at the second-to-last lecture one term, the students were asked to fill out anonymous questionnaires rating the effectiveness of their teaching assistants (by name), and the course, on the scale

1	2	3	4	5
poor	fair	good	very good	excellent

The following statistics were computed.

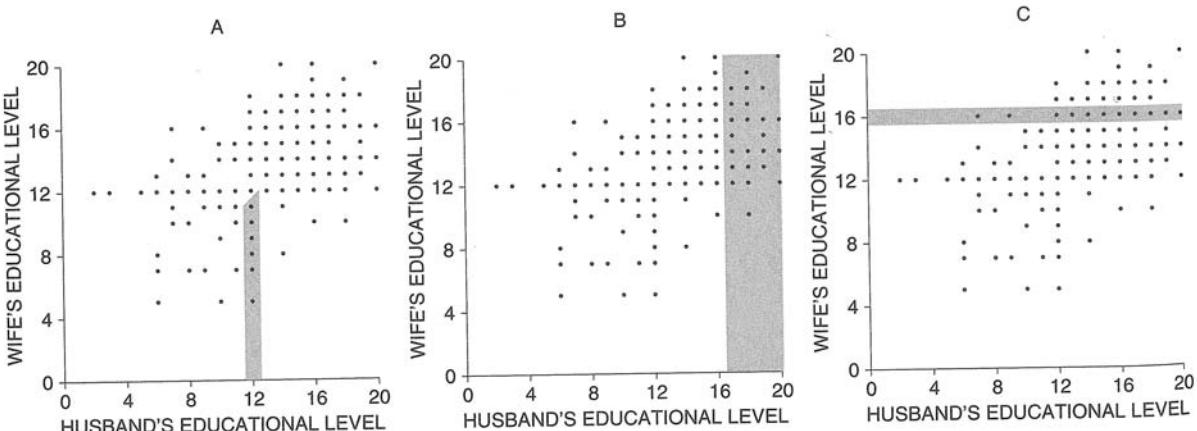
- The average rating of the assistant by the students in each section.
- The average rating of the course by the students in each section.
- The average score on the final for the students in each section.

Results are shown below (sections are identified by letter). Draw a scatter diagram for each pair of variables—there are three pairs—and find the correlations.

Section	Ave. rating of assistant	Ave. rating of course	Ave. score on final
A	3.3	3.5	70
B	2.9	3.2	64
C	4.1	3.1	47
D	3.3	3.3	63
E	2.7	2.8	69
F	3.4	3.5	69
G	2.8	3.6	69
H	2.1	2.8	63
I	3.7	2.8	53
J	3.2	3.3	65
K	2.4	3.3	64

The data are section averages. Since the questionnaires were anonymous, it was not possible to link up student ratings with scores on an individual basis. Student ability may be a confounding factor. However, controlling for pre-test results turned out to make no difference in the analysis.<sup>13</sup> Each assistant taught one section. True or false, and explain:

- (a) On the average, those sections that liked their TA more did better on the final.
- (b) There was almost no relationship between the section's average rating of the assistant and the section's average rating of the course.
- (c) There was almost no relationship between the section's average rating of the course and the section's average score on the final.
10. In a study of 2005 Math SAT scores, the Educational Testing Service computed the average score for each of the 51 states, and the percentage of the high-school seniors in that state who took the test.<sup>14</sup> (For these purposes, D.C. counts as a state.) The correlation between these two variables was equal to  $-0.84$ .
- (a) True or false: test scores tend to be lower in the states where a higher percentage of the students take the test. If true, how do you explain this? If false, what accounts for the negative correlation?
- (b) In Connecticut, the average score was only 517. But in Iowa, the average was 608. True or false, and explain: the data show that on average, the schools in Iowa are doing a better job at teaching math than the schools in Connecticut.
11. As part of the study described in exercise 10, the Educational Testing Service computed the average Verbal SAT score for each state, as well as the average Math SAT score for each state. (Again, D.C. counts as a state.) The correlation between these 51 pairs of averages was 0.97. Would the correlation between the Math SAT and the Verbal SAT—computed from the data on all the individuals who took the tests—be larger than 0.97, about 0.97, or less than 0.97? Explain briefly.
12. Shown below is a scatter diagram for educational levels (years of schooling completed) of husbands and wives in South Carolina, from the March 2005 Current Population Survey.
- (a) The points make vertical and horizontal stripes. Why?



- (b) There were 530 couples in the sample, and there is a dot for each couple. But if you count, there are only 104 dots in the scatter diagram. How can that be? Explain briefly.
- (c) Three areas are shaded. Match the area with the description. (One description will be left over.)
- (i) Wife completed 16 years of schooling.
  - (ii) Wife completed more years of schooling than husband.
  - (iii) Husband completed more than 16 years of schooling.
  - (iv) Husband completed 12 years of schooling and wife completed fewer years of schooling than husband.

## 7. SUMMARY

1. The correlation coefficient is a pure number, without units. It is not affected by
  - interchanging the two variables,
  - adding the same number to all the values of one variable,
  - multiplying all the values of one variable by the same positive number.
2. The correlation coefficient measures clustering around a line, relative to the SDs.
3. The correlation coefficient can be misleading in the presence of outliers or non-linear association. Whenever possible, look at the scatter diagram to check for these problems.
4. *Ecological* correlations, which are based on rates or averages, tend to overstate the strength of associations for individuals.
5. Correlation measures association. But association does not necessarily show causation. It may only show that both variables are simultaneously influenced by some third variable.

# 10

## Regression

*You've got to draw the line somewhere.*

### 1. INTRODUCTION

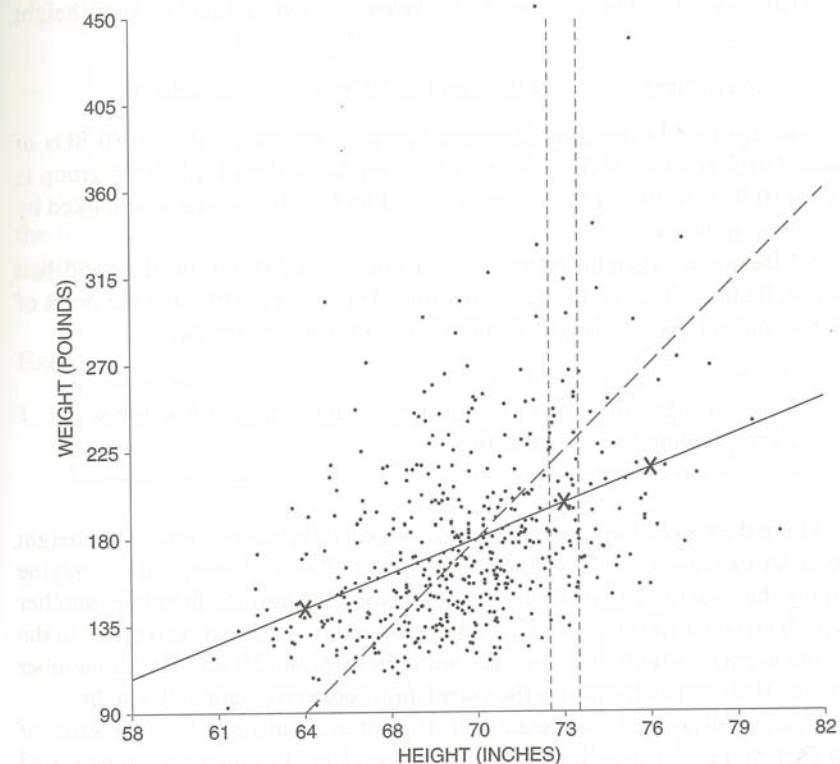
The regression method describes how one variable depends on another. For example, take height and weight. We have data for 471 men age 18–24 (from the Health and Nutrition Examination Survey—HANES5; see p. 58). In round numbers the average height of these men was 70 inches, and their overall average weight was 180 pounds. Naturally, the taller men weighed more. How much of an increase in weight is associated with a unit increase in height? To get started, look at the scatter diagram (figure 1 on the next page). Height is plotted on the horizontal axis, and weight on the vertical. The summary statistics are<sup>1</sup>

$$\begin{aligned}\text{average height} &\approx 70 \text{ inches}, & \text{SD} &\approx 3 \text{ inches} \\ \text{average weight} &\approx 180 \text{ pounds}, & \text{SD} &\approx 45 \text{ pounds}, & r &\approx 0.40\end{aligned}$$

The scales on the vertical and horizontal axes have been chosen so that one SD of height and one SD of weight cover the same distance on the page. This makes the SD line (dashed) rise at 45 degrees across the page. There is a fair amount of scatter around the line:  $r$  is only 0.40.

The vertical strip in figure 1 shows the men who were one SD above average in height (to the nearest inch). The men who were also one SD above average in weight would be plotted on the SD line. However, most of the points in the strip are well below the SD line. In other words, most of the men who were one SD above average in height were quite a bit less than one SD above average in

Figure 1. Scatter diagram. Each point shows the height and weight for one of the 471 men age 18–24 in HANES5. The vertical strip represents men who are about one SD above average in height. Those who are also one SD above average in weight would be plotted along the dashed SD line. Most of the men in the strip are below the SD line: they are only part of an SD above average in weight. The solid regression line estimates average weight at each height.



weight. The average weight of these men is only part of an SD above the overall average weight. This is where the correlation of 0.40 comes in. Associated with an increase of one SD in height there is an increase of only 0.40 SDs in weight, on the average.

To be more specific, take the men who are one SD above average in height:

$$\text{average height} + \text{SD of height} = 70 \text{ in} + 3 \text{ in} = 73 \text{ in.}$$

Their average weight will be above the overall average by 0.40 SDs of weight. Translated back to pounds, that's

$$0.40 \times 45 \text{ lb} = 18 \text{ lb.}$$

So, the average weight of these men is around

$$180 \text{ lb} + 18 \text{ lb} = 198 \text{ lb.}$$

The point (73 inches, 198 pounds) is marked by a cross in figure 1.

What about the men who are 2 SDs above average in height? Now

$$\text{average height} + 2 \text{ SD of height} = 70 \text{ in} + 2 \times 3 \text{ in} = 76 \text{ in.}$$

The average weight of this second group of men should be above the overall average by  $0.40 \times 2 = 0.80$  SDs of weight. That's  $0.80 \times 45 \text{ lb} = 36 \text{ lb}$ . So their average is around  $180 \text{ lb} + 36 \text{ lb} = 216 \text{ lb}$ . The point (76 inches, 216 pounds) is also marked by a cross in figure 1.

What about the men who are 2 SDs below average in height? Their height equals

$$\text{average height} - 2 \text{ SD of height} = 70 \text{ in} - 2 \times 3 \text{ in} = 64 \text{ in.}$$

Their average weight is below the overall average by  $0.40 \times 2 = 0.80$  SDs of weight. That's  $0.80 \times 45 \text{ lb} = 36 \text{ lb}$ . The average weight of this third group is around  $180 \text{ lb} - 36 \text{ lb} = 144 \text{ lb}$ . The point (64 inches, 144 pounds) is marked by a third cross in figure 1.

All the points (height, estimate for average weight) fall on the solid line shown in figure 1. This is the *regression line*. The line goes through the point of averages: men of average height should also be of average weight.

The regression line for  $y$  on  $x$  estimates the average value for  $y$  corresponding to each value of  $x$ .

Along the regression line, associated with each increase of one SD in height there is an increase of only 0.40 SDs in weight. To be more specific, imagine grouping the men by height. There is a group which is average in height, another group which is one SD above average in height, and so on. From each group to the next, the average weight also goes up, but only by around 0.40 SDs. Remember where the 0.40 comes from. It is the correlation between height and weight.

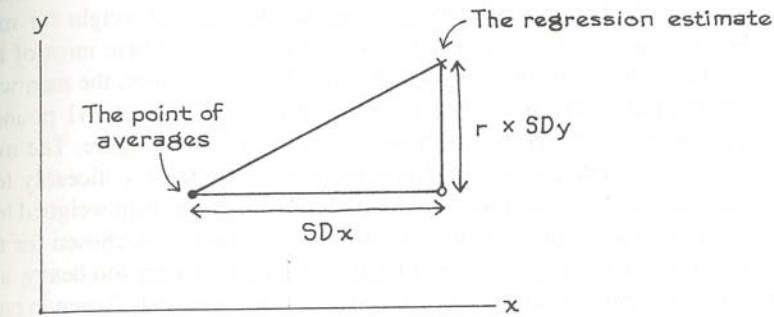
This way of using the correlation coefficient to estimate the average value of  $y$  for each value of  $x$  is called the *regression method*. The method can be stated as follows.

Associated with each increase of one SD in  $x$  there is an increase of only  $r$  SDs in  $y$ , on the average.

Two different SDs are involved here: the SD of  $x$ , to gauge changes in  $x$ ; and the SD of  $y$ , to gauge changes in  $y$ . It is easy to get carried away by the rhythm: if  $x$  goes up by one SD, so does  $y$ . But that's wrong. On the average,  $y$  only goes up by  $r$  SDs (figure 2, next page).

Why is  $r$  the right factor? Three cases are easy to see directly. First, suppose  $r$  is 0. Then there is no association between  $x$  and  $y$ . So a one-SD increase in  $x$  is accompanied by a zero-SD increase in  $y$ , on the average. Second, suppose  $r$  is 1. Then all the points lie on the SD line: a one-SD increase in  $x$  is accompanied by a one-SD increase in  $y$ . Third, suppose  $r$  is  $-1$ . The argument is the same, except

Figure 2. Regression method. When  $x$  goes up by one SD, the average value of  $y$  only goes up by  $r$  SDs.



the line slopes down. With in-between values of  $r$ , a complicated mathematical argument is needed—but  $r$  is the factor to use.

### Exercise Set A

1. In a certain class, midterm scores average out to 60 with an SD of 15, as do scores on the final. The correlation between midterm scores and final scores is about 0.50. Estimate the average final score for the students whose midterm scores were

(a) 75    (b) 30    (c) 60

Plot your regression estimates, as in figure 1.

2. For the men age 18 and over in HANES5,

average height  $\approx$  69 inches,  $SD \approx$  3 inches

average weight  $\approx$  190 pounds,  $SD \approx$  42 pounds,  $r \approx 0.41$

Estimate the average weight of the men whose heights were

(a) 69 inches    (b) 66 inches    (c) 24 inches    (d) 0 inches

Comment on your answers to (c) and (d).

3. The men age 45–74 in HANES5 had an average height of 69 inches, equal to the overall average height (exercise 2). True or false, and explain: their average weight should be around 190 pounds, that being the overall average weight.

4. For women age 25–34 in the U.S. in 2005, with full-time jobs, the relationship between education (years of schooling completed) and personal income can be summarized as follows:<sup>2</sup>

average education  $\approx$  14 years,  $SD \approx$  2.4 years

average income  $\approx$  \$32,000,  $SD \approx$  \$26,000,  $r \approx 0.34$

Estimate the average income of those women who have finished high school but have not gone on to college (so they have 12 years of education).

5. Suppose  $r = -1$ . Can you explain why a one-SD increase in  $x$  is matched by a one-SD decrease in  $y$ ?

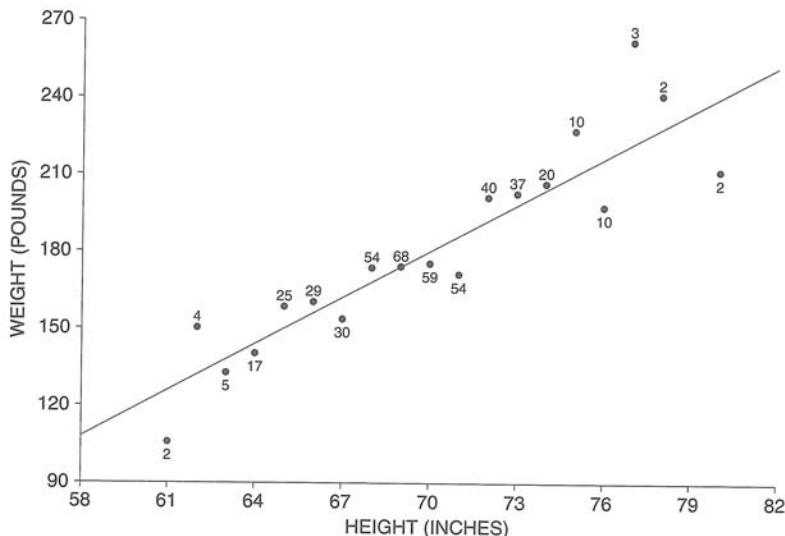
*The answers to these exercises are on pp. A59–60.*

## 2. THE GRAPH OF AVERAGES

Figure 3 is the *graph of averages* for the heights and weights of the men age 18–24 in the HANES5 sample.<sup>3</sup> The graph shows the average weight for men at each height, and is close to a straight line in the middle—where most of the people are. But at the ends, the graph is quite bumpy. For instance, the men who were 78 inches tall (to the nearest inch) had an average weight of 241 pounds. This is represented by the point (78 inches, 241 pounds) in the figure. The men who were 80 inches tall averaged 211 pounds in weight. This is noticeably less than the average for the men who were 78 inches tall. The taller men weighed less than the shorter men. Chance variation is at work. The men were chosen for the sample at random. By the luck of the draw, the 78-inch men were too heavy, and the 80-inch men weren't heavy enough. Of course, there were only 2 men in each group, as indicated by the little numbers above or below the dots. The regression line smooths away this kind of chance variation.

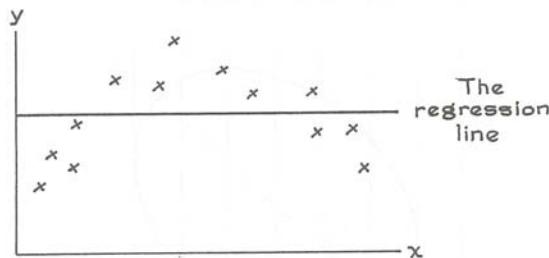
The regression line is a smoothed version of the graph of averages. If the graph of averages follows a straight line, that line is the regression line.

Figure 3. The graph of averages. Shows average weight at each height for the 471 men age 18–24 in the HANES5 sample. The regression line smooths this graph.



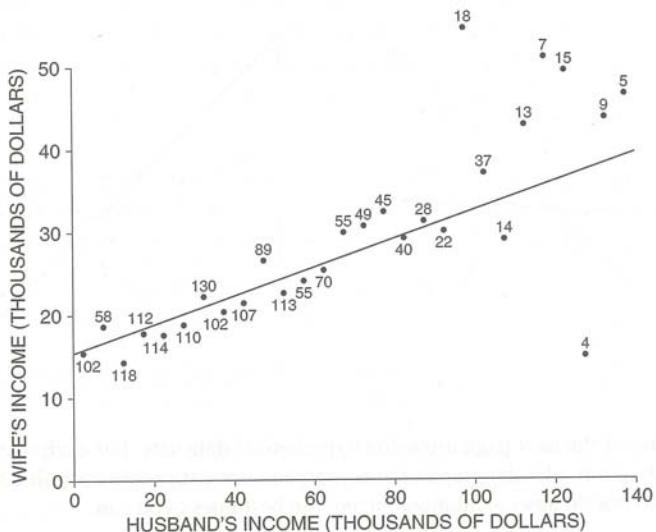
In some situations, the regression line smooths away too much. If there is a non-linear association between the two variables, as in figure 4 on the next page, the regression line will pass it by. Then, it is better to use the graph of averages. (Non-linearity came up for the correlation coefficient, section 3 of chapter 9; also see pp. 59 and 61 for data where the graph of averages is non-linear.)

Figure 4. Non-linear association. Regression lines should not be used when there is a non-linear association between the variables.



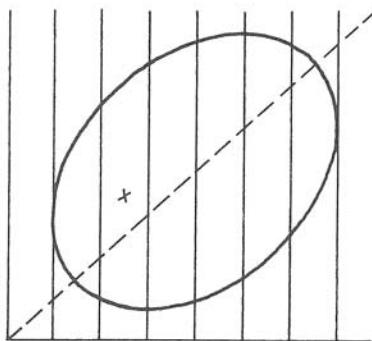
### Exercise Set B

1. The figure below is based on a representative sample of married couples in New York. The graph shows the average income of the wives, given their husband's income. With 102 couples, the husband's income was in the range \$1–\$5,000; for those couples, the wife's income averaged \$15,390, as indicated by the point (\$2,500, \$15,390). With 58 couples, the husband's income was in the range \$5,001–\$10,000; for those couples, the wife's income averaged \$18,645, as indicated by the point (\$7,500, \$18,645). And so forth. The regression line is plotted too.<sup>4</sup>
  - (a) True or false: there is a positive association between husband's income and wife's income. If true, how would you explain the association?
  - (b) Why is the dot at \$127,500 so far below the regression line?
  - (c) If you use the regression line to estimate wife's income from husband's income, would your estimates generally be a little too high, a little too low, or just about right—for the couples in the sample with husband's income in the range \$65,000–\$80,000?

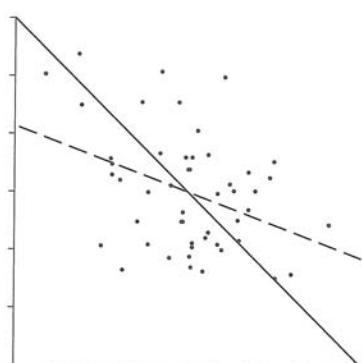
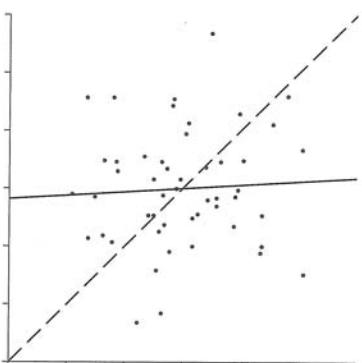
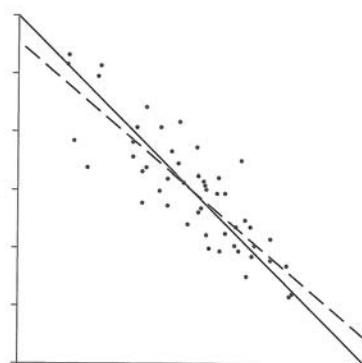
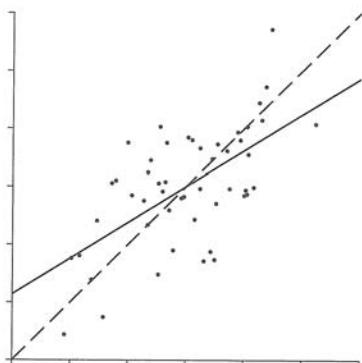


Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census.

2. Trace the diagram below on a piece of paper, and make a cross at the average for each of the vertical strips; one of them has already been done. Then draw the regression line for  $y$  on  $x$ . (The SD line is dashed.)



3. Below are four scatter diagrams, each with a solid line and a dashed line. For each diagram, say which is the SD line and which is the regression line for  $y$  on  $x$ .



4. At the top of the next page are some hypothetical data sets. For each one, draw the scatter diagram, plot the graph of averages, and draw the regression line for  $y$  on  $x$ . Please do not do any calculations: make the best guess you can.

(a)		(b)		(c)		(d)	
$x$	$y$	$x$	$y$	$x$	$y$	$x$	$y$
1	0	0	0	0	0	0	2
1	6	0	2	1	1	1	3
2	5	1	2	2	4	2	0
3	6					2	4
3	8					3	1
						4	2

The answers to these exercises are on pp. A61–62.

*Technical note.* In general, the regression line fitted to the graph of averages, with each point weighted according to the number of cases it represents, coincides with the regression line fitted to the original scatter diagram. This is exact when points with different  $x$ -coordinates are kept separate in the graph of averages; otherwise, it is a good approximation.

### 3. THE REGRESSION METHOD FOR INDIVIDUALS

For the men age 18–24 in HANES5, the relationship between height and weight can be summarized as follows:

$$\begin{aligned} \text{average height} &\approx 70 \text{ inches}, & \text{SD} &\approx 3 \text{ inches} \\ \text{average weight} &\approx 180 \text{ pounds}, & \text{SD} &\approx 45 \text{ pounds}, & r &\approx 0.40 \end{aligned}$$

Suppose one of these men is picked at random, and you have to guess his weight without being told anything about him. The best guess is the overall average weight, 180 pounds. Next, you are told the man's height: 73 inches, for example. This man is tall, and likely to be heavier than average. Your best guess for his weight is the average for all the 73-inch men in the study. This new average can be estimated by the regression method, as 198 pounds (p. 159). The rule: if you have to predict one variable from another, use the new average. In many cases, the regression method gives a sensible way of estimating the new average. Of course, if there is a non-linear association between the variables, the regression method would not apply.

*Example 1.* A university has made a statistical analysis of the relationship between Math SAT scores (ranging from 200 to 800) and first-year GPAs (ranging from 0 to 4.0), for students who complete the first year. The results:

$$\begin{aligned} \text{average SAT score} &= 550, & \text{SD} &= 80 \\ \text{average first-year GPA} &= 2.6, & \text{SD} &= 0.6, & r &= 0.4 \end{aligned}$$

The scatter diagram is football-shaped. A student is chosen at random, and has an SAT of 650. Predict this individual's first-year GPA.

*Solution.* This student is  $100/80 = 1.25$  SDs above average on the SAT. The regression estimate for first-year GPA is, above average by  $0.4 \times 1.25 = 0.5$  SDs. That's  $0.5 \times 0.6 = 0.3$  GPA points. The predicted GPA is  $2.6 + 0.3 = 2.9$ .

The logic: for all students with an SAT of around 650, the average first-year GPA is about 2.9, by the regression method. That is why we predict a first-year GPA of 2.9 for this individual.

Usually, investigators work out regression estimates from a study, and then extrapolate: they use the estimates on new subjects. In many cases this makes sense, provided the subjects in the survey are representative of the people about whom the inferences are going to be made. But you have to think about the issue each time. The mathematics of the regression method will not protect you. In example 1, the university only has experience with the students it admits. There could be a problem in using the regression procedure on students who are quite different from that group. (Admissions officers typically do extrapolate, from admitted students to students who are denied admission.)

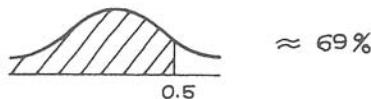
Now, another use for the regression method—to predict *percentile ranks*. If your percentile rank on a test is 90%, you did very well: only 10% of the class scored higher, the other 90% scored lower. A percentile rank of 25% is not so good: 75% of the class scored higher, the other 25% scored lower (p. 91).

*Example 2.* (This continues example 1.) Suppose the percentile rank of one student on the SAT is 90%, among the first-year students. Predict his percentile rank on first-year GPA. The scatter diagram is football-shaped. In particular, the SAT scores and GPAs follow the normal curve.

*Solution.* We are going to use the regression method. This student is above the average on the SAT. By how many SDs? Because SAT scores follow the normal curve, his percentile rank has this information—in disguise (section 5 of chapter 5):



This student scored 1.3 SDs above average on the SAT. The regression method predicts he will be  $0.4 \times 1.3 \approx 0.5$  SDs above average on first-year GPA. Finally, this can be translated back into a percentile rank:



That is the answer. The percentile rank on first-year GPA is predicted as 69%.

In solving this problem, the averages and SDs of the two variables were never used. All that mattered was  $r$ . Basically, this is because the whole problem was worked in standard units. The percentile ranks give you the standard units.

The student in example 2 was compared with his class in two different com-

petitions, the SAT and the first-year exams. He did very well on the SAT, scoring at the 90th percentile. But the regression estimate only puts him at the 69th percentile on the first year exams; still above average, but not as much. On the other hand, for poor students—say at the 10th percentile of the SAT—the regression method predicts an improvement. It will put them at the 31st percentile on the first-year tests. This is still below average, but closer.

To go at this more carefully, take all the people at the 90th percentile on the SAT—good students. Some of them will move up on the first-year tests, some will move down. On the average, however, this group moves down. For comparison, take all the people at the 10th percentile of the SAT—poor students. Again, some will do better on the first-year tests, others worse. On the average, however, this group moves up. That is what the regression method is telling us.

Initially, many people would predict a first-year rank equal to the SAT rank. This is not a good strategy. To see why, imagine that you had to predict a student's rank in a mathematics class. In the absence of other information, the safest guess is to put her at the median. However, if you knew that this student was very good in physics, you would probably put her well above the median in mathematics. After all, there is a strong correlation between physics and mathematics. On the other hand, if all you knew was her rank in a pottery class, that would not help very much in guessing the mathematics rank. The median looks good: there is not much correlation between pottery and mathematics.

Now, back to the problem of predicting first-year rank from SAT rank. If the two sets of scores are perfectly correlated, first-year rank will be equal to SAT rank. At the other extreme, if the correlation is zero, SAT rank does not help at all in predicting first-year rank. The correlation is somewhere between the two extremes, so we have to predict a rank on the first-year tests somewhere between the SAT rank and the median. The regression method tells us where.

### Exercise Set C

1. In a certain class, midterm scores average out to 60 with an SD of 15, as do scores on the final. The correlation between midterm scores and final scores is about 0.50. The scatter diagram is football-shaped. Predict the final score for a student whose midterm score is

(a) 75      (b) 30      (c) 60      (d) unknown

Compare your answers to exercise 1 on p. 161.

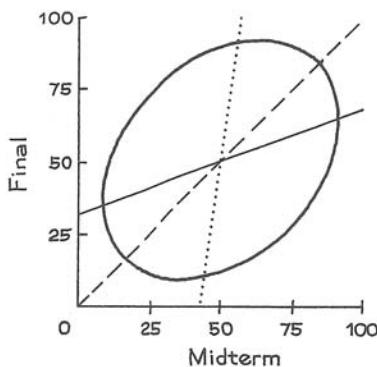
2. For the first-year students at a certain university, the correlation between SAT scores and first-year GPA was 0.60. The scatter diagram is football-shaped. Predict the percentile rank on the first-year GPA for a student whose percentile rank on the SAT was

(a) 90%      (b) 30%      (c) 50%      (d) unknown

Compare your answer to (a) with example 2.

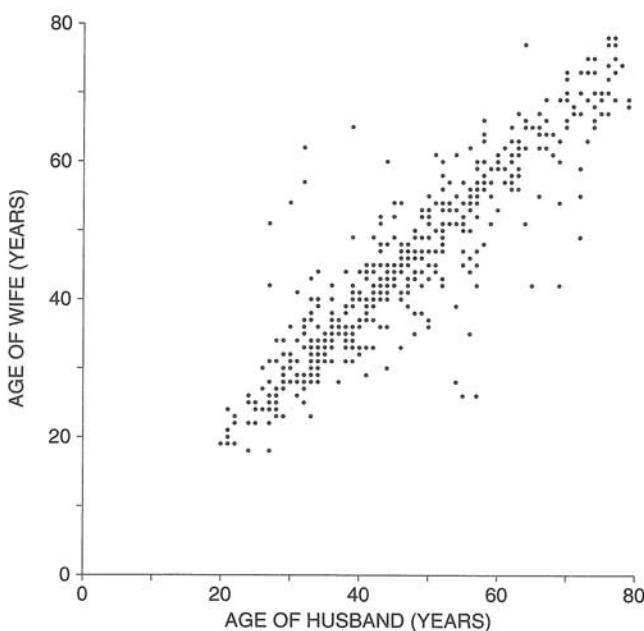
3. The scatter diagram below shows the scores on the midterm and final in a certain course. Three lines are drawn across the diagram.

- (a) People who have the same percentile rank on both tests are plotted along one of these lines. Which one, and why?
- (b) One of these lines would be used to predict final score from midterm score. Which one, and why?



4. The scatter diagram below shows ages of husbands and wives in Tennessee. (Data are from the March 2005 Current Population Survey.)

- (a) Why are there no dots in the lower left hand corner of the diagram?
- (b) Why does the diagram show vertical and horizontal stripes?



5. For the men age 18 and over in the HANES5 sample, the correlation between height and weight was 0.41; the SD of height was about 3 inches and the SD of weight was about 42 pounds. The men age 55–64 averaged about half an inch shorter than the men age 18–24. True or false, and explain: since half an inch is  $1/6 \approx 0.17$  SDs of height, the men age 55–64 must have averaged about  $0.41 \times 0.17 \times 42 \approx 3$  pounds lighter than the men age 18–24.

*The answers to these exercises are on p. A62.*

*Technical note.* The method discussed in example 2 is for median ranks. To see why, assume normality and  $r = 0.4$ . Of students at the 90th percentile on the SAT (relative to their classmates), about half will rank above the 69th percentile on first-year GPA, and half will rank below. The procedure for estimating average ranks is harder.

#### 4. THE REGRESSION FALLACY

A preschool program tries to boost children's IQs. Children are tested when they enter the program (the pre-test), and again when they leave (the post-test). On both occasions, the scores average out to nearly 100, and the SD is about 15. The program seems to have no effect. A closer look at the data, however, shows something very surprising. The children who were below average on the pre-test had an average gain of about 5 IQ points at the post-test. Conversely, those children who were above average on the pre-test had an average loss of about 5 points. What does this prove? Does the program operate to equalize intelligence? Perhaps when the brighter children play with the duller ones, the difference between the two groups tends to be diminished. Is this desirable or undesirable?

These speculations may be interesting, but the sad fact is that nothing much is going on, good or bad. Here is why. The children cannot be expected to score exactly the same on the two tests. There will be differences between the two scores. Nobody would think these differences mattered, or needed any explanation. But they make the scatter diagram for the test scores spread out around the SD line into that familiar football-shaped cloud. The spread around the line makes the bottom group come up and the top group come down. There is nothing else to it.

In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test—and the top group will on average fall back. This is the *regression effect*.

Thinking that the regression effect must be due to something important, not just the spread around the line, is the *regression fallacy*.

We are now going to see why the regression effect appears whenever there is spread around the SD line. This effect was first noticed by Galton in his study of family resemblances, so that is the context for the discussion. But the reasoning is general. Figure 5 shows a scatter diagram for the heights of 1,078 pairs of fathers and sons, as discussed in chapter 8. The summary statistics are<sup>5</sup>

$$\begin{aligned}\text{average height of fathers} &\approx 68 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches} \\ \text{average height of sons} &\approx 69 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches}, \quad r \approx 0.5\end{aligned}$$

The sons average 1 inch taller than the fathers. On this basis, it is natural to guess that a 72-inch father should have a 73-inch son; similarly, a 64-inch father should have a 65-inch son; and so on. Such fathers and sons are plotted along the dashed line in figure 5. Of course, not many families are going to be right on the line. In fact, there is a lot of spread around the line. Some of the sons are taller than their fathers; others are shorter.

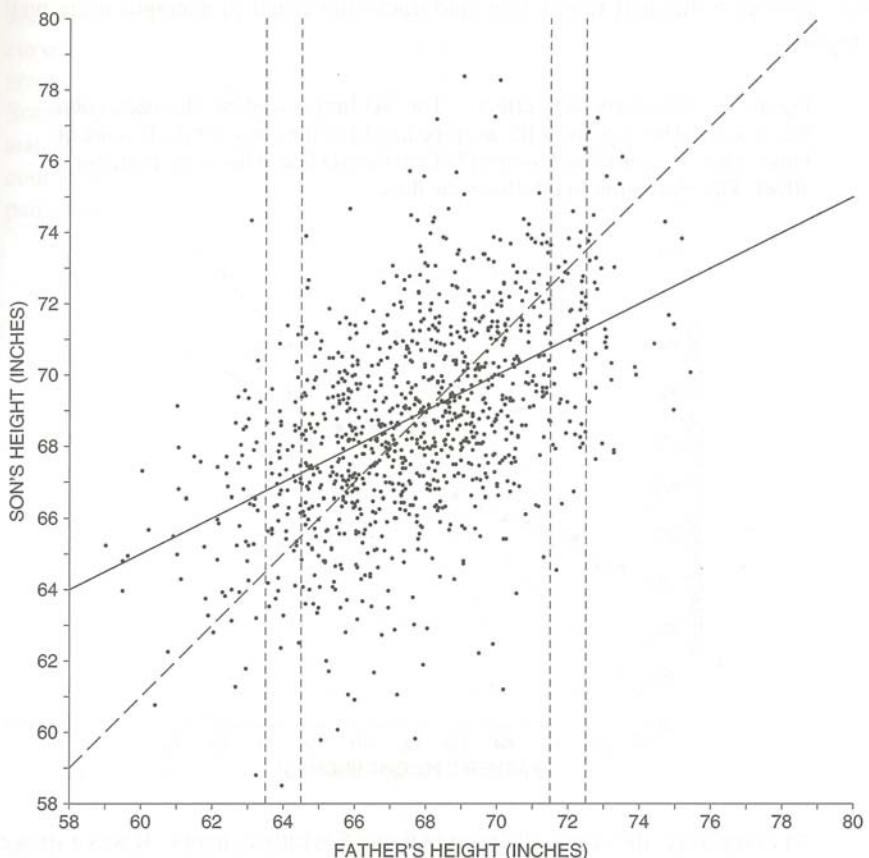
Take the fathers who are 72 inches tall, to the nearest inch. The corresponding families are plotted in the vertical strip over 72 inches in figure 5, and there is quite a range in the sons' heights. Some of the points are above the dashed line: the son is taller than 73 inches. But most of the points are below the dashed line: the son is shorter than 73 inches. All in all, the sons of the 72-inch fathers only average 71 inches in height. With tall fathers (high score on first test), on the average the sons are shorter (score on second test drops).

Now look at the points in the vertical strip over 64 inches, representing the families where the father is 64 inches tall, to the nearest inch. The height of the dashed line there is 65 inches, representing a son who is 1 inch taller than his 64-inch father. Some of the points fall below the dashed line, but most are above, and the sons of the 64-inch fathers average 67 inches in height. With short fathers (low score on first test), on the average the sons are taller (score on second test goes up). The aristocratic Galton termed this “regression to mediocrity.”

The dashed line in figure 5 goes through the point corresponding to an average father of height 68 inches, and his average son of height 69 inches. Along the dashed line, each one-SD increase in father's height is matched by a one-SD increase in son's height. These two facts make it the SD line. The cloud is symmetric around the SD line, but the strip at 72 inches is not. The strip only contains points with unusually big  $x$ -coordinates. And most of the points in this strip fall below the SD line. Conversely, the strip at 64 inches only contains points with unusually small  $x$ -coordinates. Most of the points in this strip fall above the SD line. The hidden imbalance is always there in football-shaped clouds. The graphical explanation for the regression effect may not seem very romantic. But then, statistics isn't known as a romantic subject.

Figure 5 also shows the regression line for the son's height on father's height. This solid line rises less steeply than the dashed SD line, and it picks off the center of each vertical strip of dots—the average  $y$ -value in the strip. For instance, take the fathers who are 72 inches tall. They are 4 inches above average in height:

Figure 5. The regression effect. If a son is 1 inch taller than his father, the family is plotted along the dashed line. The points in the strip over 72 inches correspond to the families where the father is 72 inches tall, to the nearest inch; most of these points are below the dashed line. The points in the strip over 64 inches correspond to families where the father is 64 inches tall, to the nearest inch; most of these points are above the dashed line. The solid regression line picks off the centers of all the vertical strips, and is flatter than the dashed line.



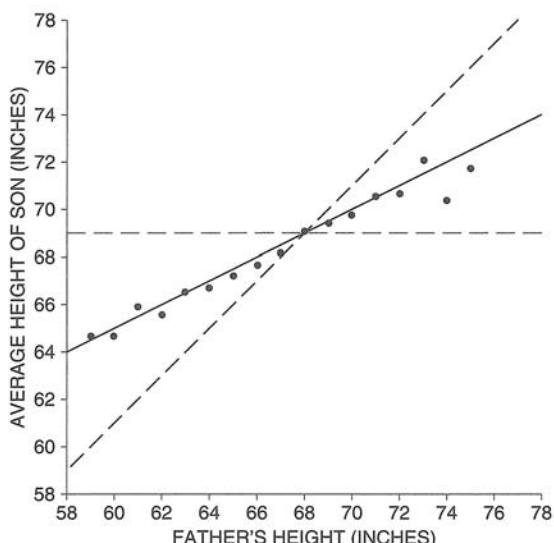
4 inches/2.7 inches  $\approx$  1.5 SDs. The regression line says their sons should be taller than average, by about

$$r \times 1.5 \text{ SDs} = 0.75 \text{ SDs} \approx 2 \text{ inches.}$$

The overall average height for sons is 69 inches, so the regression estimate for the average height of these sons is 71 inches—dead on.

Figure 6 shows the regression effect at its steepest, without the cloud. The dashed SD line rises at a 45 degree angle. The dots show the average height of the sons corresponding to each value of father's height. These dots are the centers of the vertical strips in figure 5. The dots rise less steeply than the SD line—the regression effect. On the whole, the dots are halfway between the SD line and the horizontal line through the point of averages. That is because the correlation coefficient is one half. Each one-SD increase in father's height is accompanied by a half-SD increase in son's height, not a one-SD increase. The solid regression line goes up at the half-to-one rate, and tracks the graph of averages quite well indeed.

Figure 6. The regression effect. The SD line is dashed, the regression line is solid. The dots show the average height of the sons, for each value of father's height. They rise less steeply than the SD line. This is the regression effect. The regression line follows the dots.



At first glance, the scatter diagram in figure 5 is rather chaotic. It was a stroke of genius on Galton's part to see a straight line in the chaos. Since Galton's time, many other investigators have found that the averages in their scatter diagrams followed straight lines too. That is why the regression line is so useful.

Now, a look behind the scenes: the regression effect can be understood a little better in some cases, for instance, in the context of a repeated IQ test. The basic fact is that the two scores are apt to be different. The difference can be explained in terms of chance variability. Each person may be lucky or unlucky on the first test. But if the score on the first test is very high, that suggests the person was

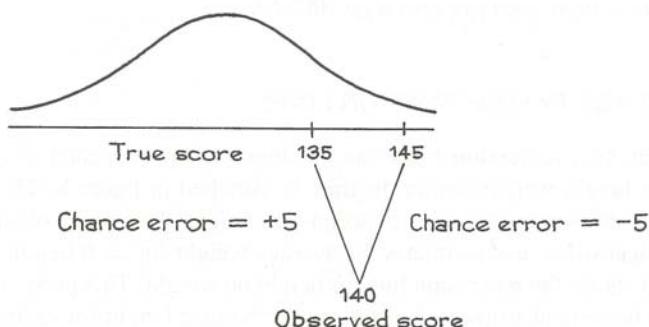
lucky on that occasion, implying that the score on the second test will probably be lower. (You wouldn't say, "He scored very high, must have had bad luck that day.") On the other hand, if the score on the first test was very low, the person was probably unlucky to some extent on that occasion and will do better next time.

Here is a crude model for the test-retest situation, which brings the explanation into sharper focus. The basic equation is

$$\text{observed test score} = \text{true score} + \text{chance error}.$$

Assume that the distribution of true scores in the population follows the normal curve, with an average of 100 and an SD of 15. Suppose too that the chance error is as likely to be positive as negative, and tends to be about 5 points in size. Someone who has a true score of 135 is just as likely to score 130 as 140 on the test. Someone with a true score of 145 is just as likely to score 140 as 150. Of course, the chance error could also be  $\pm 4$ , or  $\pm 6$ , and so forth: any symmetric pair of values can be dealt with in a similar way.

Figure 7. A model for the regression effect.



Take the people who scored 140 on the first test. There are two alternative explanations for this observed score:

- true score below 140, with a positive chance error;
- true score above 140, with a negative chance error.

The first explanation is more likely. For instance, more people have true scores of 135 than 145, as figure 7 shows.

The model accounts for the regression effect. If someone scores above average on the first test, the true score is probably a bit lower than the observed score. If this person takes the test again, we predict that the second score will be a bit lower than the first score. On the other hand, if a person scores below average on the first test, we estimate that the true score is a bit higher than the observed score, and our prediction for the second score will be a bit higher than the first score.

### Exercise Set D

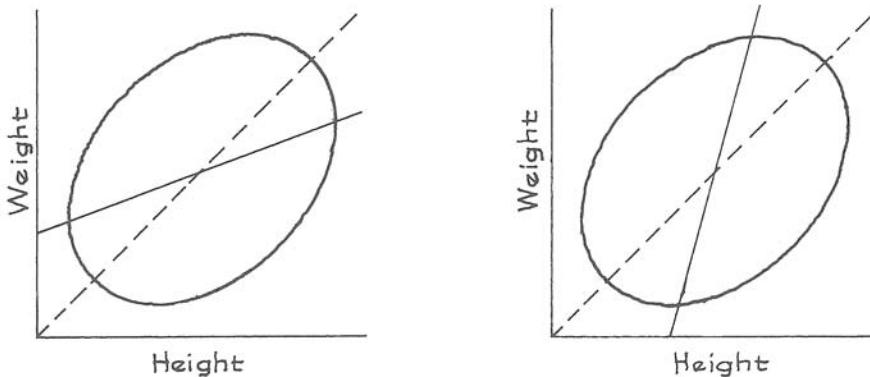
- As part of their training, air force pilots make two practice landings with instructors, and are rated on performance. The instructors discuss the ratings with the pilots after each landing. Statistical analysis shows that pilots who make poor landings the first time tend to do better the second time. Conversely, pilots who make good landings the first time tend to do worse the second time. The conclusion: criticism helps the pilots while praise makes them do worse. As a result, instructors were ordered to criticize all landings, good or bad. Was this warranted by the facts? Answer yes or no, and explain briefly.<sup>6</sup>
- An instructor standardizes her midterm and final each semester so the class average is 50 and the SD is 10 on both tests. The correlation between the tests is around 0.50. One semester, she took all the students who scored below 30 at the midterm, and gave them special tutoring. They all scored above 50 on the final. Can this be explained by the regression effect? Answer yes or no, and explain briefly.
- In the data set of figures 5 and 6, are the sons of the 61-inch fathers taller on the average than the sons of the 62-inch fathers, or shorter? What is the explanation?

*The answers to these exercises are on pp. A62–63.*

### 5. THERE ARE TWO REGRESSION LINES

In fact, two regression lines can be drawn across a scatter diagram. For example, a height-weight scatter diagram is sketched in figure 8. The left hand panel shows the regression line for weight on height. This picks off the centers of the vertical strips, and estimates the average weight for each height. The right hand panel shows the regression line for height on weight. This picks off the centers of the horizontal strips, and estimates the average height for each weight. In both panels, the regression line is solid and the SD line is dashed. The regression of weight on height seems more natural for most purposes, but the other line may come in handy too.

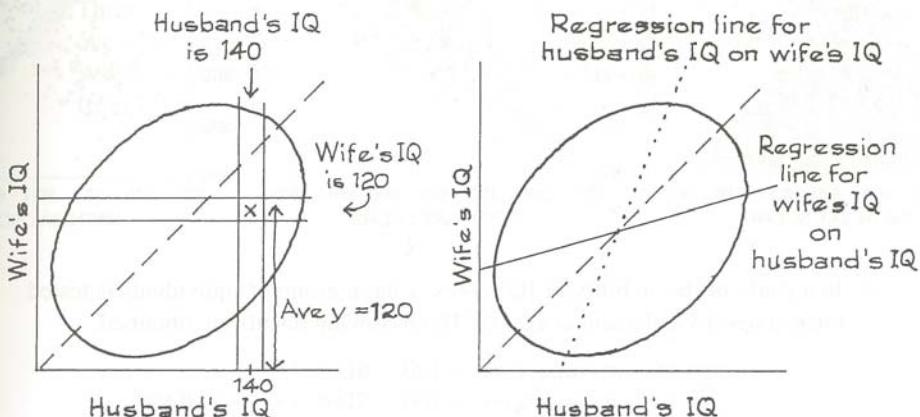
Figure 8. The left hand panel shows the regression of weight on height; the right hand panel, height on weight. The SD line is dashed.



*Example 3.* IQ scores are scaled to have an average of about 100, and an SD of about 15, both for men and for women. The correlation between the IQs of husbands and wives is about 0.50. A large study of families found that the men whose IQ was 140 had wives whose IQ averaged 120. Look at the wives in the study whose IQ was 120. Should the average IQ of their husbands be greater than 120? Answer yes or no, and explain briefly.

*Solution.* No, the average IQ of their husbands will be around 110. See figure 9. The families where the husband has an IQ of 140 are shown in the vertical strip. The average  $y$ -coordinate in this strip is 120. The families where the wife has an IQ of 120 are shown in the horizontal strip. This is a completely different set of families. The average  $x$ -coordinate for points in the horizontal strip is about 110. Remember, there are two regression lines. One line is for predicting the wife's IQ from her husband's IQ. The other line is for predicting the husband's IQ from his wife's.

Figure 9. The two regression lines.



### Exercise Set E

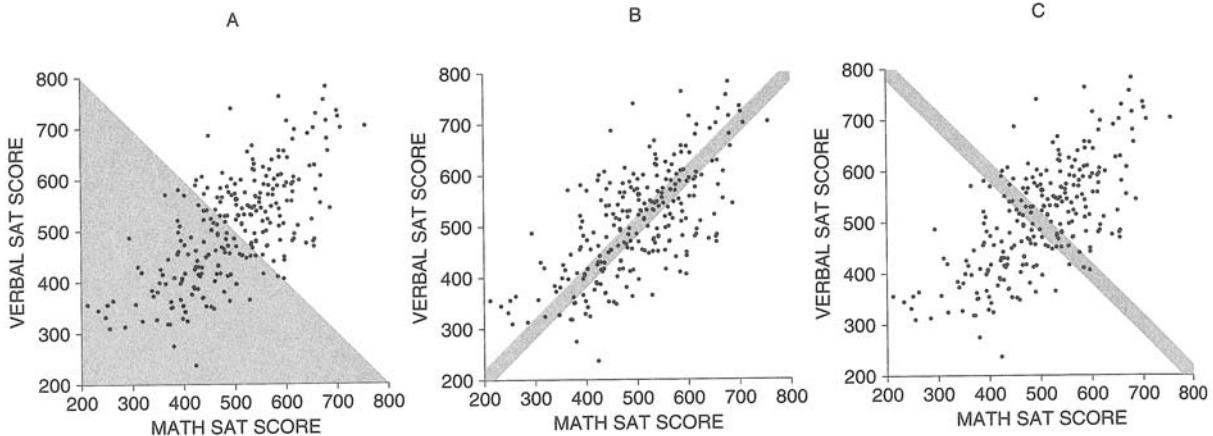
1. For the men age 18–24 in the HANES5 sample, the ones who were 63 inches tall averaged 138 pounds in weight. True or false, and explain: the ones who weighed 138 pounds must have averaged 63 inches in height.
2. In Pearson's study, the sons of the 72-inch fathers only averaged 71 inches in height. True or false: if you take the 71-inch sons, their fathers will average about 72 inches in height. Explain briefly.
3. In example 2 (p. 166), the regression method predicted that a student at the 90th percentile on the SAT would only be at the 69th percentile on first-year GPA. True or false, and explain: a student at the 69th percentile on first-year GPA should be at the 90th percentile on the SAT.

The answers to these exercises are on p. A63.

## 6. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. Shown below is a scatter diagram for Math and Verbal SAT scores for graduating seniors at a certain high school. Three areas are shaded. Match the area with the description. (One description will be left over.)
  - (i) Total score (Math + Verbal) is below 800.
  - (ii) Total score (Math + Verbal) is around 800.
  - (iii) Math score is about equal to Verbal score.
  - (iv) Math score is less than Verbal score.



2. In a study of the stability of IQ scores, a large group of individuals is tested once at age 18 and again at age 35. The following results are obtained.

$$\begin{aligned} \text{age 18: } & \text{average score} \approx 100, \quad \text{SD} \approx 15 \\ \text{age 35: } & \text{average score} \approx 100, \quad \text{SD} \approx 15, \quad r \approx 0.80 \end{aligned}$$

- (a) Estimate the average score at age 35 for all the individuals who scored 115 at age 18.
- (b) Predict the score at age 35 for an individual who scored 115 at age 18.

3. Pearson and Lee obtained the following results in a study of about 1,000 families:

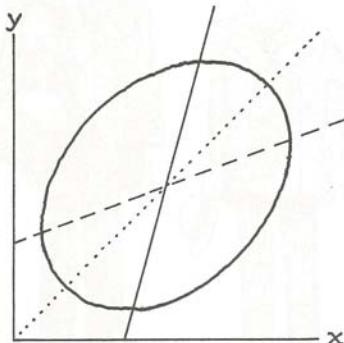
$$\begin{aligned} \text{average height of husband} &\approx 68 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches} \\ \text{average height of wife} &\approx 63 \text{ inches}, \quad \text{SD} \approx 2.5 \text{ inches}, \quad r \approx 0.25 \end{aligned}$$

Predict the height of a wife when the height of her husband is

- (a) 72 inches
- (b) 64 inches
- (c) 68 inches
- (d) unknown

4. In one study, the correlation between the educational level of husbands and wives in a certain town was about 0.50; both averaged 12 years of schooling completed, with an SD of 3 years.<sup>7</sup>

- (a) Predict the educational level of a woman whose husband has completed 18 years of schooling.
- (b) Predict the educational level of a man whose wife has completed 15 years of schooling.
- (c) Apparently, well-educated men marry women who are less well educated than themselves. But the women marry men with even less education. How is this possible?
5. An investigator measuring various characteristics of a large group of athletes found that the correlation between the weight of an athlete and the amount of weight that athlete could lift was 0.60. True or false, and explain:
- On the average, an athlete can lift 60% of his body weight.
  - If an athlete gains 10 pounds, he can expect to lift an additional 6 pounds.
  - The more an athlete weighs, on the average the more he can lift.
  - The more an athlete can lift, on the average the more he weighs.
  - 60% of an athlete's lifting ability can be attributed to his weight alone.
6. Three lines are drawn across the scatter diagram below. One is the SD line, one is the regression line for  $y$  on  $x$ , and one is the regression line for  $x$  on  $y$ . Which is which? Why? (The "regression line for  $y$  on  $x$ " is used to predict  $y$  from  $x$ .)



7. A doctor is in the habit of measuring blood pressures twice. She notices that patients who are unusually high on the first reading tend to have somewhat lower second readings. She concludes that patients are more relaxed on the second reading. A colleague disagrees, pointing out that the patients who are unusually low on the first reading tend to have somewhat higher second readings, suggesting they get more nervous. Which doctor is right? Or perhaps both are wrong? Explain briefly.
8. A large study was made on the blood-pressure problem discussed in the previous exercise. It found that first readings average 130 mm, and second readings average 120 mm; both SDs were about 15 mm. Does this support either doctor's argument? Or is it the regression effect? Explain.

9. In a large statistics class, the correlation between midterm scores and final scores is found to be nearly 0.50, every term. The scatter diagrams are football-shaped. Predict the percentile rank on the final for a student whose percentile rank on the midterm is
- (a) 5%    (b) 80%    (c) 50%    (d) unknown
10. True or false: A student who is at the 40th percentile of first-year GPAs is also likely to be at the 40th percentile of second-year GPAs. Explain briefly. (The scatter diagram is football-shaped.)

## 7. SUMMARY

1. Associated with an increase of one SD in  $x$ , there is an increase of only  $r$  SDs in  $y$ , on the average. Plotting these *regression estimates* gives the *regression line* for  $y$  on  $x$ .



2. The *graph of averages* is often close to a straight line, but may be a little bumpy. The regression line smooths out the bumps. If the graph of averages is a straight line, then it coincides with the regression line. If the graph of averages has a strong non-linear pattern, regression may be inappropriate.
3. The regression line can be used to make predictions for individuals. But if you have to extrapolate far from the data, or to a different group of subjects, be careful.

4. In a typical test-retest situation, the subjects get different scores on the two tests. Take the bottom group on the first test. Some improve on the second test, others do worse. On average, the bottom group shows an improvement. Now, the top group: some do better the second time, others fall back. On average, the top group does worse the second time. This is the *regression effect*, and it happens whenever the scatter diagram spreads out around the SD line into a football-shaped cloud of points.

5. The *regression fallacy* consists in thinking that the regression effect must be due to something other than spread around the SD line.

6. There are two regression lines that can be drawn on a scatter diagram. One predicts  $y$  from  $x$ ; the other predicts  $x$  from  $y$ .

# 11

## The R.M.S. Error for Regression

*Such are the formal mathematical consequences of normal correlation. Much biometric material certainly shows a general agreement with the features to be expected on this assumption: although I am not aware that the question has been subjected to any sufficiently critical enquiry. Approximate agreement is perhaps all that is needed to justify the use of the correlation as a quantity descriptive of the population; its efficacy in this respect is undoubted, and it is not improbable that in some cases it affords, in conjunction with the means and variances, a complete description of the simultaneous variation of the variates.*

—SIR R. A. FISHER (ENGLAND, 1890–1962)<sup>1</sup>

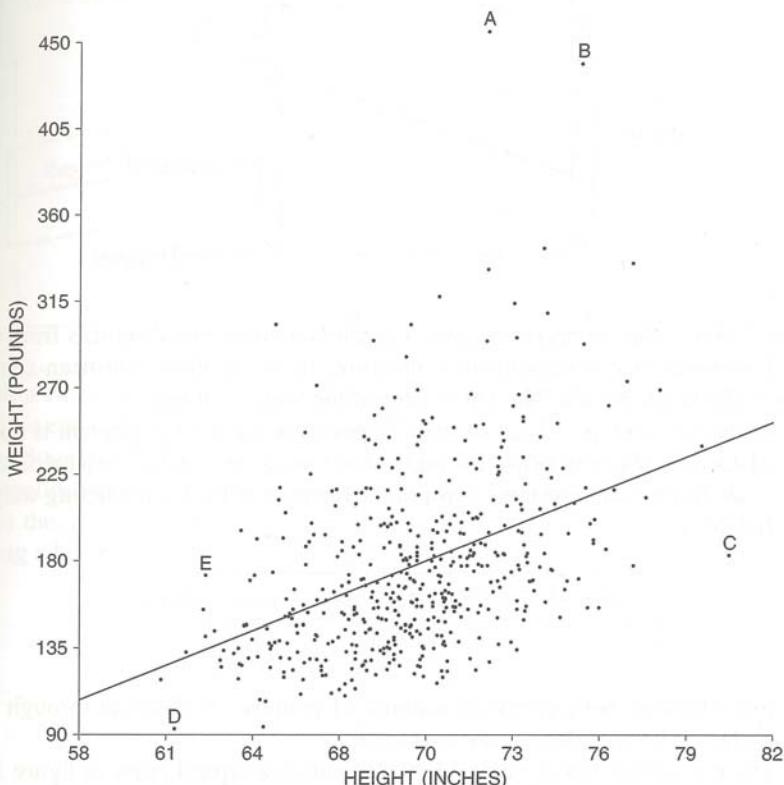
### 1. INTRODUCTION

The regression method can be used to predict  $y$  from  $x$ . However, actual values differ from predictions. By how much? The object of this section is to measure the overall size of the differences using the r.m.s. error. For example, take the heights and weights of the 471 men age 18–24 in the HANES5 sample (section 1 of chapter 10). The summary statistics:

$$\begin{aligned} \text{average height} &\approx 70 \text{ inches}, & \text{SD} &\approx 3 \text{ inches} \\ \text{average weight} &\approx 180 \text{ pounds}, & \text{SD} &\approx 45 \text{ pounds}, & r &\approx 0.40 \end{aligned}$$

To review briefly, given a man's height, his weight is predicted by the average weight for all the men with that height. The average can be estimated by the regression method. Figure 1 shows the regression line. Person A on the diagram is about 72 inches tall. The regression estimate for average weight at this height is

Figure 1. Prediction errors. The error is the distance above (+) or below (-) the regression line. The scatter diagram shows heights and weights for the 471 men age 18–24 in the HANES5 sample.



192 pounds (section 1 of chapter 10). However, A's actual weight is 456 pounds. The prediction is off, by 264 pounds:

$$\begin{aligned} \text{error} &= \text{actual weight} - \text{predicted weight} \\ &= 456 \text{ lb} - 192 \text{ lb} = 264 \text{ lb}. \end{aligned}$$

In the diagram, the prediction error is the vertical distance of A above the regression line.

Person C on the diagram is 80.5 inches tall and weighs 183 pounds. The regression line predicts his weight as 243 pounds. So there is a prediction error of  $183 \text{ lb} - 243 \text{ lb} = -60 \text{ lb}$ . In the diagram, this error is represented by the vertical distance of C below the regression line.

The distance of a point above (+) or below (-) the regression line is

$$\text{error} = \text{actual} - \text{predicted}.$$

Figure 2. Prediction error equals vertical distance from the line.

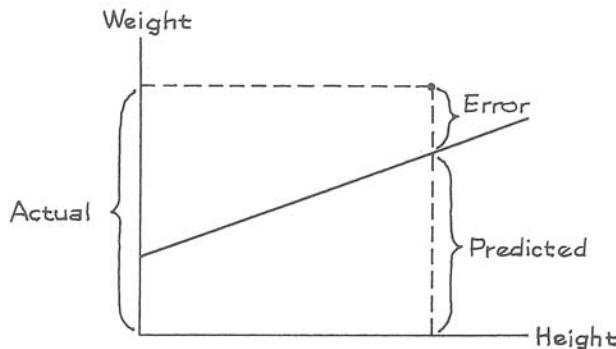


Figure 2 shows the connection between prediction errors and distances from the line. The overall size of these errors is measured by taking their root-mean-square (p. 66). The result is called the *r.m.s. error of the regression line*.

Go back to figure 1. Each of the 471 points in the scatter diagram is some vertical distance above or below the regression line, corresponding to a prediction error made by the line. The r.m.s. error of the regression line for predicting weight from height is

$$\sqrt{\frac{(\text{error } \#1)^2 + (\text{error } \#2)^2 + \dots + (\text{error } \#471)^2}{471}}$$

This looks painful, but the answer is about 41 pounds. (A short-cut through the arithmetic will be presented in the next section.)

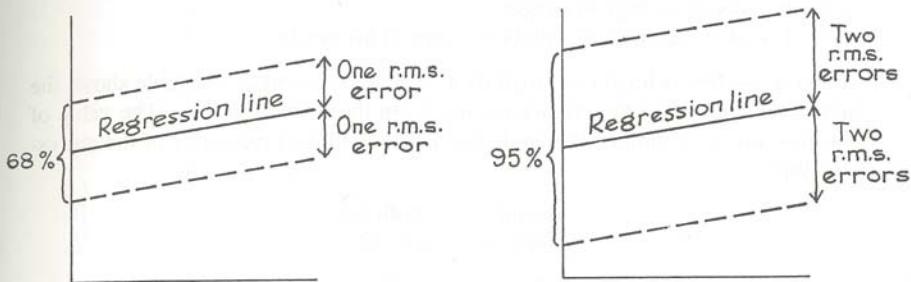
The r.m.s. error has a graphical interpretation: a typical point in figure 1 is above or below the regression line by something like 41 pounds. Since the line is predicting weight from height, we conclude that for typical men in the study, actual weight differs from predicted weight by around 41 pounds or so.

The r.m.s. error for regression says how far typical points are above or below the regression line.

The r.m.s. error is to the regression line as the SD is to the average. For instance, about 68% of the points on a scatter diagram will be within one r.m.s. error of the regression line; about 95% of them will be within two r.m.s. errors. This rule of thumb holds for many data sets, but not all; it is illustrated in figure 3.

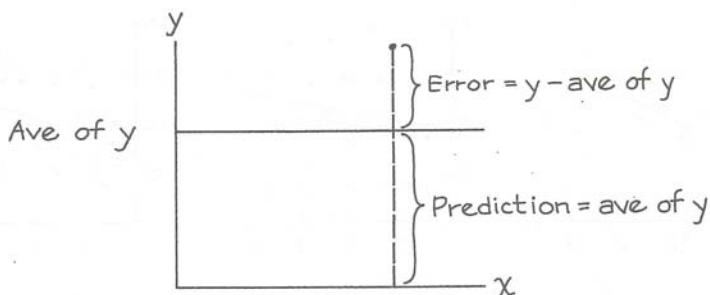
What about the height-weight data? The computer found that the predictions were right to within one r.m.s. error (41 pounds) for 340 out of 471 men, or 72% of them. The rule of thumb doesn't look bad at all. The predictions were right to

Figure 3. Rule of thumb. About 68% of the points on a scatter diagram fall inside the strip whose edges are parallel to the regression line, and one r.m.s. error away (up or down). About 95% of the points are in the wider strip whose edges are parallel to the regression line, and twice the r.m.s. error away.



within two r.m.s. errors (82 pounds) for 451 out of the 471 men, which is 96%. This is even better for the rule of thumb.

Soon, we will compare the r.m.s. error for regression to the r.m.s. error for a baseline prediction method. The baseline method just ignores the  $x$ -values and uses the average value of  $y$  to predict  $y$ . With this method, the predictions fall along a horizontal line through the average of  $y$ .



Graphically, the prediction errors for the second method are the vertical distances above and below this horizontal line, as shown by the sketch. Numerically, the errors are the deviations from the average of  $y$ . So the r.m.s. error for the second method is the SD of  $y$ : remember, the SD is the r.m.s. of the deviations from average.

The SD of  $y$  says how far typical points are above or below a horizontal line through the average of  $y$ . In other words, the SD of  $y$  is the r.m.s. error for the baseline method—predicting  $y$  by its average, just ignoring the  $x$ -values.

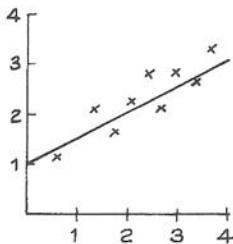
## Exercise Set A

- Look at figure 1, then fill in the blanks: person B is \_\_\_\_\_ and \_\_\_\_\_, while D is \_\_\_\_\_ and \_\_\_\_\_. Options: short, tall, skinny, chubby.
- Look at figure 1, then say whether each statement is true or false:
  - E is above average in weight.
  - E is above average in weight, for men of his height.
- A regression line is fitted to a small data set. For each subject, the table shows the actual value of  $y$  and the predicted value from the regression line. (The value of  $x$  is not shown.) Compute the prediction errors, and the r.m.s. error of the regression line.

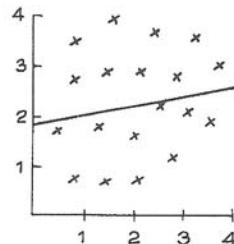
<i>Actual value of <math>y</math></i>	<i>Predicted value of <math>y</math></i>
57	64
63	62
43	40
51	52
49	45

- Below are three scatter diagrams. The regression line has been drawn across each one, by eye. In each case, guess whether the r.m.s. error is 0.2, or 1, or 5.

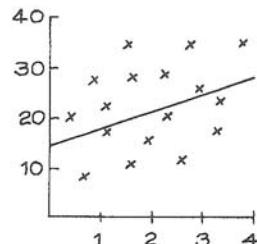
(a)



(b)



(c)

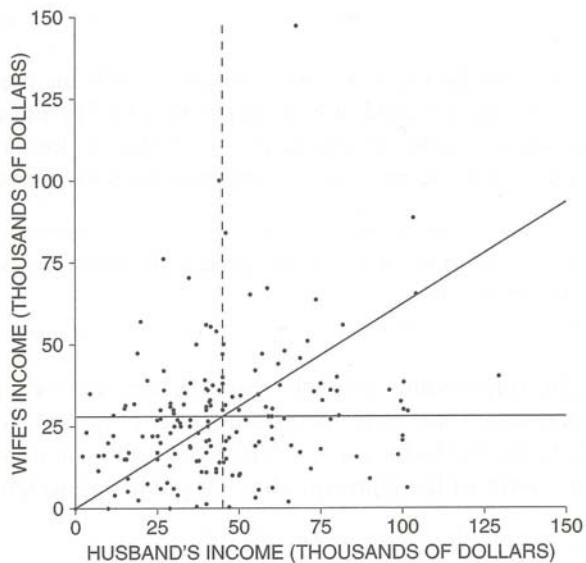


- A regression line for predicting income has an r.m.s. error of \$2,000. It predicts someone's income as \$20,000. This is likely to be right give or take: a few hundred dollars, a few thousand dollars, ten or twenty thousand dollars.
- An admissions officer is trying to choose between two methods of predicting first-year scores. One method has an r.m.s. error of 12. The other has an r.m.s. error of 7. Other things being equal, which should he choose? Why?
- A regression line for predicting test scores has an r.m.s. error of 8 points.
  - About 68% of the time, the predictions will be right to within \_\_\_\_\_ points.
  - About 95% of the time, the predictions will be right to within \_\_\_\_\_ points.
- The scatter diagram on the next page shows incomes for a sample of 168 working couples in Louisiana. Summary statistics are as follows:
 

average husband's income = \$45,000, SD = \$25,000  
   average wife's income = \$28,000, SD = \$20,000

  - If you predict wife's income as \$28,000, ignoring husband's income, your r.m.s. error will be \_\_\_\_\_.

- (b) All the predictions are on one of the lines in the diagram. Which one? Explain your answer.

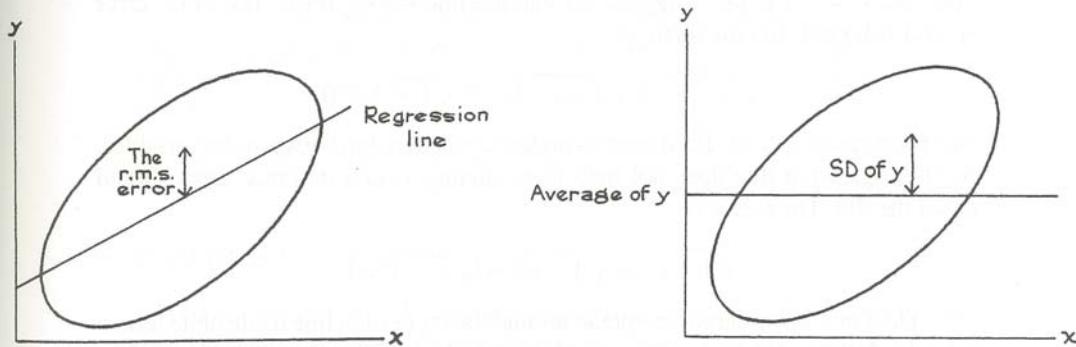


The answers to these exercises are on pp. A63–64.

## 2. COMPUTING THE R.M.S. ERROR

The r.m.s. error for the regression line measures distances above or below the regression line (left-hand panel of figure 4). The right-hand panel of figure 4 shows another line, namely, the horizontal line through the average of  $y$ . The r.m.s. error for that line is just the SD of  $y$ , as discussed on p. 183.

Figure 4. The r.m.s. error of the regression line, and the SD of  $y$ .



The r.m.s. error for the regression line will be smaller than the SD of  $y$ , because the regression line gets closer to the points than the horizontal line. The r.m.s. will be smaller by the factor  $\sqrt{1 - r^2}$ .

The r.m.s. error for the regression line of  $y$  on  $x$  can be figured as

$$\sqrt{1 - r^2} \times \text{the SD of } y.$$

Which SD goes into the formula? The SD of the variable being predicted. If you are predicting weight from height, use the SD of weight. The r.m.s. error has to come out in pounds, not inches. If you are predicting income from education, use the SD of income. The r.m.s. error has to come out in dollars, not years.

The units for the r.m.s. error are the same as the units for the variable being predicted.

In the height-weight scatter diagram (figure 1), there were 471 prediction errors, one for each man. Finding the root-mean-square of these 471 errors looked like a lot of work. But the factor  $\sqrt{1 - r^2}$  gives you a shortcut through the arithmetic. The r.m.s. error of the regression line for predicting weight from height equals

$$\sqrt{1 - r^2} \times \text{SD of weight} = \sqrt{1 - 0.40^2} \times 45 \text{ lb} \approx 41 \text{ lb}.$$

The r.m.s. error isn't much smaller than the SD of weight, because weight is not that well correlated with height:  $r \approx 0.40$ . Knowing a man's height does not help so much in predicting his weight.

The formula is hard to prove without algebra. But three special cases are easy to see. First, suppose  $r = 1$ . Then all the points lie on a straight line which slopes up. The regression line goes through all the points on the scatter diagram, and all the prediction errors are 0. So the r.m.s. error should be 0. And that is what the formula says. The factor works out to

$$\sqrt{1 - r^2} = \sqrt{1 - 1^2} = \sqrt{1 - 1} = 0.$$

The case  $r = -1$  is the same, except that the line slopes down. The r.m.s. error should still be 0, and the factor is

$$\sqrt{1 - r^2} = \sqrt{1 - (-1)^2} = \sqrt{1 - 1} = 0.$$

The third case is  $r = 0$ . Then there is no linear relationship between the variables. So the regression line does not help in predicting  $y$ , and its r.m.s. error should equal the SD. The factor is

$$\sqrt{1 - r^2} = \sqrt{1 - 0^2} = \sqrt{1 - 0} = 1.$$

The r.m.s. error measures spread around the regression line in absolute terms: pounds, dollars, and so on. The correlation coefficient, on the other hand, measures spread relative to the SD, and has no units. The r.m.s. error is connected to the SD through the correlation coefficient. This is the third time that  $r$  comes into the story.

- $r$  describes the clustering of the points around a line, relative to the SDs (chapter 8).
- $r$  says how the average value of  $y$  depends on  $x$ —associated with each one-SD increase in  $x$  there is an increase of only  $r$  SDs in  $y$ , on the average (chapter 10).
- $r$  determines the accuracy of the regression predictions, through the formula for r.m.s. error.

*A cautionary note.* If you extrapolate beyond the data, or use the line to make estimates for people who are different from the subjects in the study, the r.m.s. error cannot tell you how far off you are likely to be. That is beyond the power of mathematics.

### Exercise Set B

1. A law school finds the following relationship between LSAT scores and first-year scores:

$$\begin{array}{ll} \text{average LSAT score} = 165, & \text{SD} = 5 \\ \text{average first-year score} = 65, & \text{SD} = 10, \quad r = 0.6 \end{array}$$

The admissions officer uses the regression line to predict first-year scores from LSAT scores. The r.m.s. error of the line is \_\_\_\_\_. Options:

$$5 \quad 10 \quad \sqrt{1 - 0.6^2} \times 5 \quad \sqrt{1 - 0.6^2} \times 10$$

2. (This continues exercise 1.)

- (a) One of these students is chosen at random; you have to guess his first-year score, without being told his LSAT score. How would you do this?  
 (b) Your r.m.s. error would be \_\_\_\_\_. Options:

$$5 \quad 10 \quad \sqrt{1 - 0.6^2} \times 5 \quad \sqrt{1 - 0.6^2} \times 10$$

- (c) Repeat parts (a) and (b), if you are allowed to use his LSAT score.

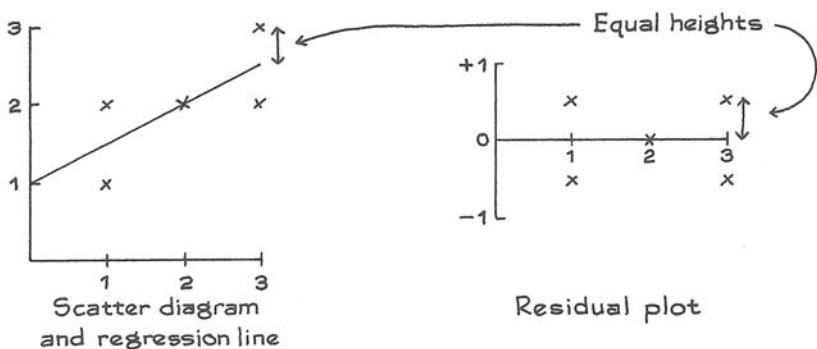
3. At a certain college, first-year GPAs average about 3.0, with an SD of about 0.5; they are correlated about 0.6 with high-school GPA. Person A predicts first-year GPAs just using the average. Person B predicts first-year GPAs by regression, using the high-school GPAs. Which person makes the smaller r.m.s. error? Smaller by what factor?

*The answers to these exercises are on p. A64.*

### 3. PLOTTING THE RESIDUALS

Prediction errors are often called *residuals*. Statisticians recommend graphing the residuals. The method is indicated by figure 5 on the next page. Each point on the scatter diagram is transferred to a second diagram, called the *residual plot*, in the following way. The  $x$ -coordinate is left alone. But the  $y$ -coordinate is replaced by the residual at the point—the distance above (+) or below (−)

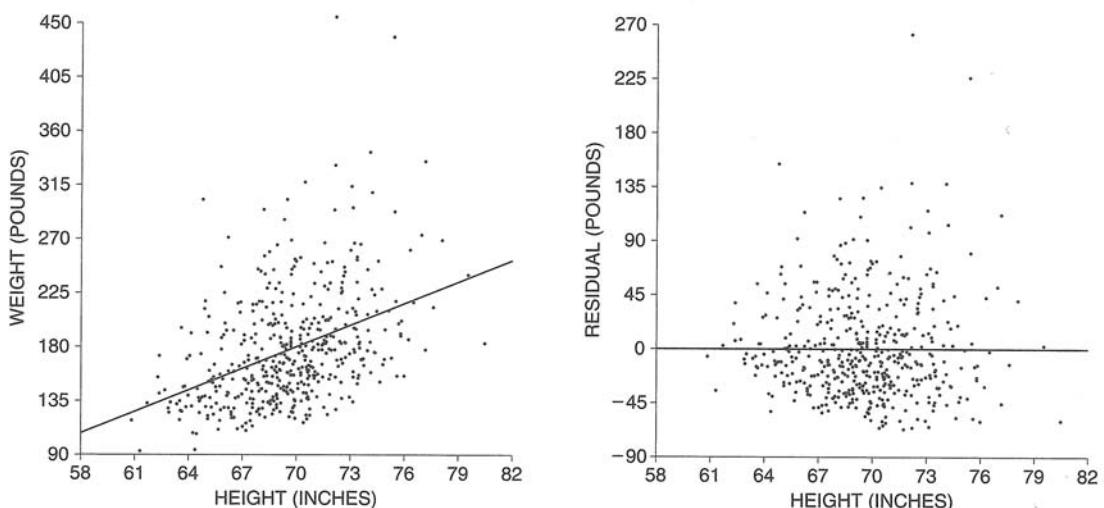
Figure 5. Plotting the residuals.



the regression line. Figure 6 shows the residual plot for the height-weight scatter diagram of figure 1. Figures 5 and 6 suggest that the positive residuals balance out the negative ones. Mathematically, the residuals from the regression line must average out to 0. The figures show something else too. As you look across the residual plot, there is no systematic tendency for the points to drift up (or down). Basically, the reason is that all the trend up or down has been taken out of the residuals, and has been absorbed into the regression line.

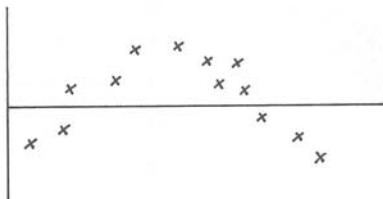
The residuals average out to 0; and the regression line for the residual plot is horizontal.

Figure 6. A residual plot. The scatter diagram at the left shows the heights and weights of the 471 men age 18–24 in the HANES5 sample, with the regression line. The residual plot is shown at the right. There is no trend or pattern in the residuals.



The residual plot in figure 6 shows no pattern. By comparison, figure 7 shows a residual plot (for hypothetical data) with a strong pattern. With this kind of pattern, it is probably a mistake to use a regression line. Often, you can spot non-linearities by looking at the scatter diagram. However, the residual plot may give a more sensitive test—because the vertical scale can be made big enough so things can be examined carefully. Residual plots are useful diagnostics in *multiple regression*; for example, in predicting first-year GPA from SAT scores and high-school GPA.<sup>2</sup> (Multiple regression is discussed in section 3 of chapter 12.)

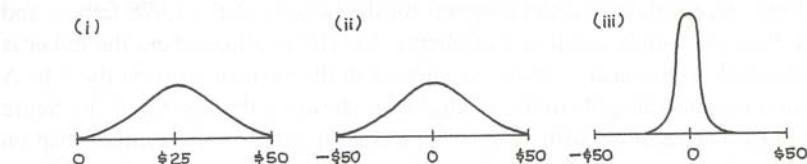
Figure 7. A residual plot with a strong pattern. It may have been a mistake to fit the regression line.



### Exercise Set C

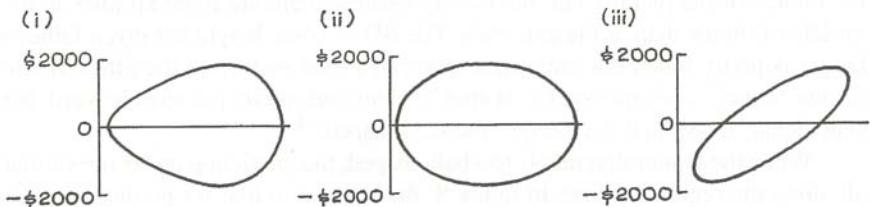
1. Several different regression lines are used to predict the price of a stock (from different independent variables). Histograms for the residuals from each line are sketched below. Match the description with the histogram:

- (a) r.m.s. error = \$5      (b) r.m.s. error = \$15      (c) something's wrong



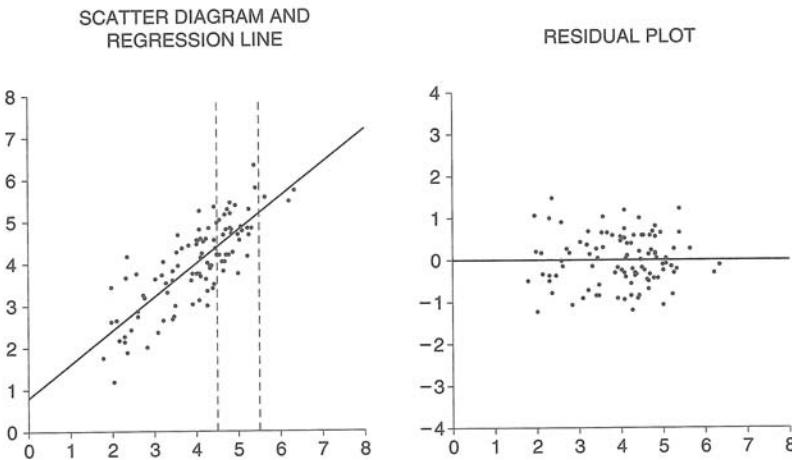
2. Several regression lines are used to predict the monthly salaries in a certain company, from different independent variables. Residual plots from each regression are shown below. Match the description with the plot. Explain. (You may use the same description more than once.)

- (a) r.m.s. error = \$1,000    (b) r.m.s. error = \$5,000    (c) something's wrong



3. Look at the figure below.

- (a) Is the SD of  $y$  about 0.6, 1.0, or 2.0?
- (b) Is the SD of the residuals about 0.6, 1.0, or 2.0?
- (c) Take the points in the scatter diagram whose  $x$ -coordinates are between 4.5 and 5.5. Is the SD of their  $y$ -coordinates about 0.6, 1.0, or 2.0?



*The answers to these exercises are on p. A64.*

#### 4. LOOKING AT VERTICAL STRIPS

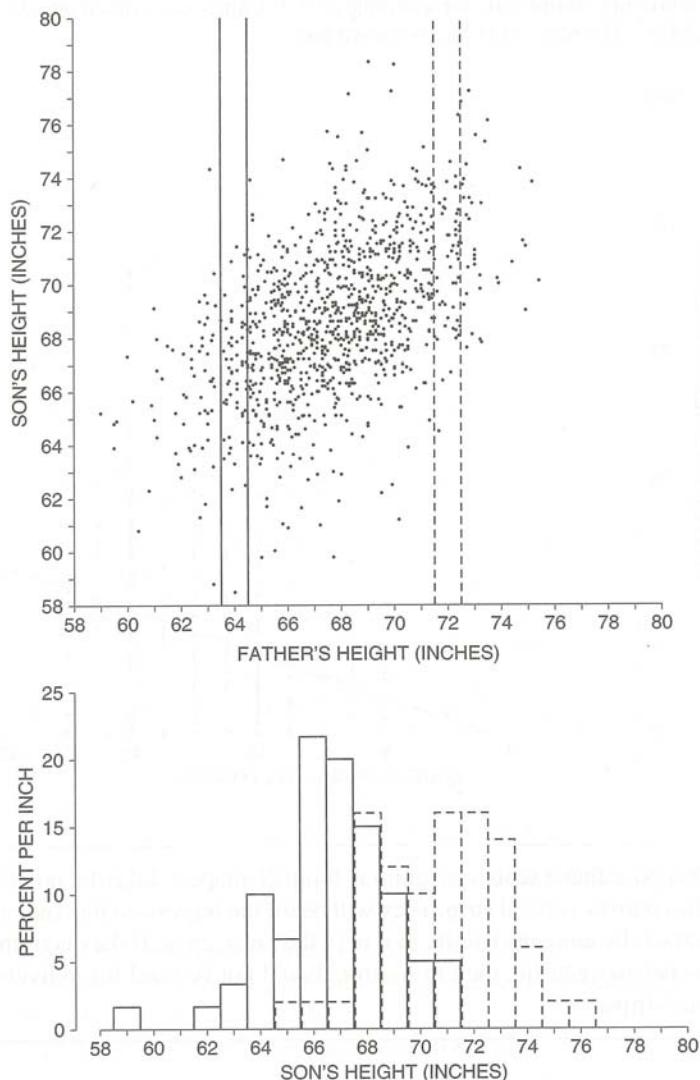
Figure 8 repeats the scatter diagram for the heights of the 1,078 fathers and sons in Pearson's study (section 1 of chapter 8). The families where the father is 64 inches tall, to the nearest inch, are plotted in the vertical strip on the left. A histogram for son's heights in these families is shown at the bottom of the figure (solid line). The families with 72-inch fathers are plotted in the vertical strip on the right. A histogram for the heights of those sons is shown too (dashed line). The dashed histogram is farther to the right than the solid one: on the average, the taller fathers do have taller sons. However, both histograms have similar shapes, and just about the same amount of spread.<sup>3</sup>

When all the vertical strips in a scatter diagram show similar amounts of spread, the diagram is said to be *homoscedastic*. The scatter diagram in figure 8 is homoscedastic. The range of sons' heights for given father's height is greater in the middle of the picture, but that is only because there are more families in the middle of things than at the extremes. The SD of sons' height for given father's height is pretty much the same from one end of the picture to the other. *Homo* means "same," *scedastic* means "scatter." *Homoscedasticity* is a terrible word, but statisticians insist on it: we prefer "football-shaped."<sup>4</sup>

When the scatter diagram is football-shaped, the prediction errors are similar all along the regression line. In figure 8, the regression line for predicting son's

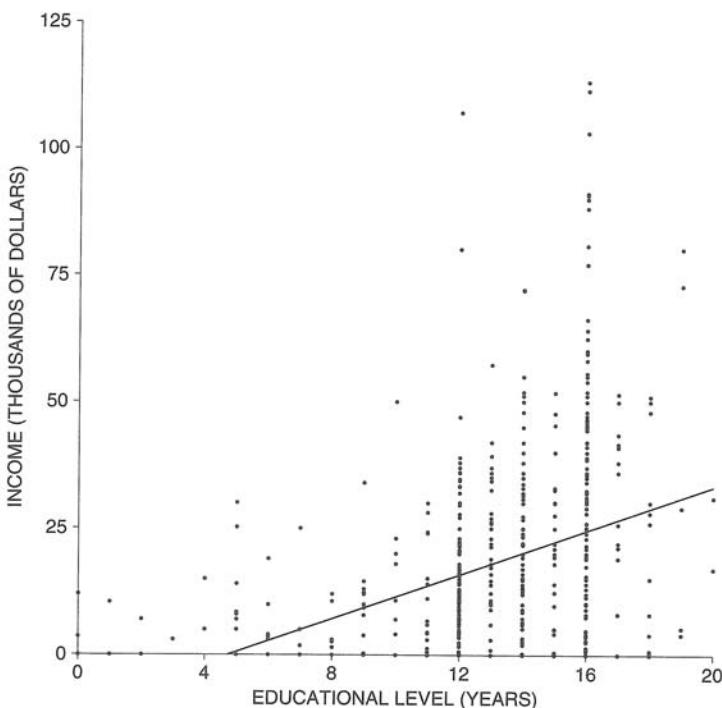
height from father's height had an r.m.s. error of 2.3 inches. If the father is 64 inches tall, the prediction for the son's height is 67 inches, and this is likely to be off by 2.3 inches or so. If the father is 72 inches tall, the prediction for the son's height is 71 inches, and this is likely to be off by the same amount, 2.3 inches or so.<sup>5</sup>

Figure 8. Homoscedastic scatter diagram. Heights of fathers and sons. Families with 64-inch fathers are plotted in the solid vertical strip; the solid histogram is for the heights of those sons. Families with 72-inch fathers are plotted in the dashed vertical strip; the dashed histogram is for the heights of those sons. The two histograms have similar shapes, and their SDs are nearly the same.



By comparison, figure 9 shows the *heteroscedastic* scatter diagram of income against education (*hetero* means “different”). As education goes up, average income goes up, and so does the spread in income. When the scatter diagram is heteroscedastic, the regression method is off by different amounts in different parts of the scatter diagram. In figure 9, the r.m.s. error of the regression line is about \$19,000. However, it is quite a bit harder to predict the incomes of the highly educated people. With 8 years of schooling, the prediction errors are something like \$6,000. At 12 years, the errors go up to \$15,000 or so. At 16 years, the errors go up even more, to \$27,000 or so. In this case, the r.m.s. error of the regression line gives a sort of average error—across all the different  $x$ -values.

Figure 9. Heteroscedastic scatter diagram. Income and education (years of schooling completed) for a sample of 570 California women age 25–29 in 2005.<sup>6</sup> The regression line is shown too.



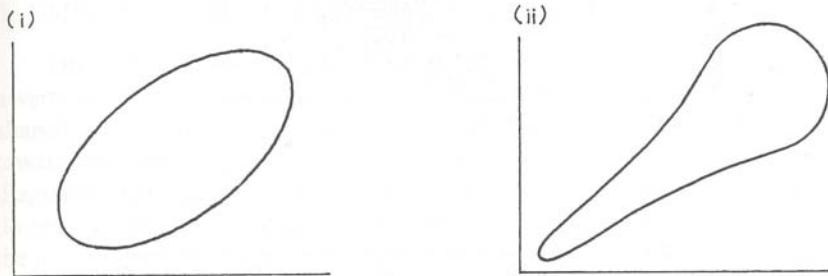
Suppose that a scatter diagram is football-shaped. Take the points in a narrow vertical strip. They will be off the regression line (up or down) by amounts similar in size to the r.m.s. error. If the diagram is heteroscedastic, the r.m.s. error should not be used for individual strips.

## Exercise Set D

1. In 1937, the Stanford-Binet IQ test was restandardized with two forms (L and M). A large number of subjects took both tests. The results can be summarized as follows:

$$\begin{aligned} \text{Form L average } &\approx 100, \quad \text{SD } \approx 15 \\ \text{Form M average } &\approx 100, \quad \text{SD } \approx 15, \quad r \approx 0.80 \end{aligned}$$

- (a) True or false, and explain: the regression line for predicting the score on form M from the score on form L has an r.m.s. error of about 9 points.
- (b) Suppose the scatter diagram looks like (i) below. If someone scores 130 on form L, the regression method predicts 124 for the score on form M. True or false, and explain: this prediction is likely to be off by 9 points or so.
- (c) Repeat, if the scatter diagram looks like (ii).



2. The data in figure 8 can be summarized as follows:

$$\begin{aligned} \text{average height of fathers } &\approx 68 \text{ inches}, \quad \text{SD } \approx 2.7 \text{ inches} \\ \text{average height of sons } &\approx 69 \text{ inches}, \quad \text{SD } \approx 2.7 \text{ inches}, \quad r \approx 0.5 \end{aligned}$$

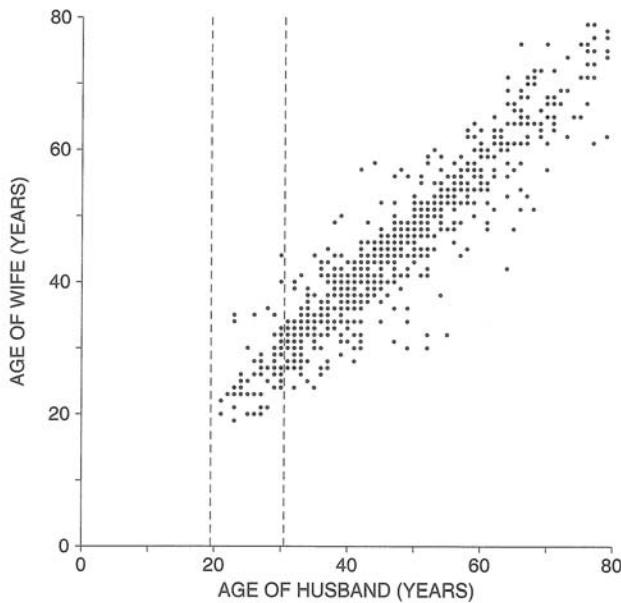
- (a) Find the r.m.s. error of the regression line for predicting son's height from father's height.
- (b) If a father is 72 inches tall, predict his son's height.
- (c) This prediction is likely to be off by \_\_\_\_\_ inches or so. If more information is needed, say what it is, and why.
- (d) Repeat parts (b) and (c), if the father is 66 inches tall.

3. The data in figure 9 can be summarized as follows:

$$\begin{aligned} \text{average education } &\approx 13.0 \text{ years}, \quad \text{SD } \approx 3.4 \text{ years} \\ \text{average income } &\approx \$18,000, \quad \text{SD } \approx \$20,000, \quad r \approx 0.37 \end{aligned}$$

- (a) Find the r.m.s. error of the regression line for predicting income from education.
- (b) Predict the income of a woman with 16 years of education.
- (c) This prediction is likely to be off by \$\_\_\_\_\_ or so. If more information is needed, say what it is, and why.
- (d) Repeat parts (b) and (c), for a woman with 8 years of education.

4. The figure below is a scatter diagram for the ages of husbands and wives in Indiana. Data are from the March 2005 Current Population Survey.<sup>7</sup> The vertical strip represents the families where the \_\_\_\_\_ is between \_\_\_\_\_ and \_\_\_\_\_ years of age.



5. (Continues exercise 4.) Fill in the blanks, using the options given below.

.25      .5      .95      1      5      15      25      50

- (a) The average age for all the husbands is about \_\_\_\_\_; the SD is about \_\_\_\_\_.
- (b) The average age for all the wives is about \_\_\_\_\_; the SD is about \_\_\_\_\_.
- (c) The correlation between the ages of all the husbands and wives is about \_\_\_\_\_.
- (d) Among families plotted in the vertical strip, the average age for the wives is about \_\_\_\_\_; the SD is about \_\_\_\_\_.
- (e) Among families plotted in the vertical strip, the correlation between the ages of the husbands and wives is about \_\_\_\_\_.

6. (Continues exercises 4 and 5.)

- (a) The SD is computed for the ages of—
  - (i) all the wives, and
  - (ii) the wives whose husbands are 20–30 years old.

Which SD is bigger? Or are the SDs about the same?

- (b) The SD is computed for the ages of—
  - (i) all the wives, and
  - (ii) the wives whose husbands were born in March.

Which SD is bigger? Or are the SDs about the same?

7. In one study of identical male twins, the average height was found to be about 68 inches, with an SD of about 3 inches. The correlation between the heights of the twins was about 0.95, and the scatter diagram was football-shaped.
- You have to guess the height of one of these twins, without any further information. What method would you use?
  - Find the r.m.s. error for the method in (a).
  - One twin of the pair is standing in front of you. You have to guess the height of the other twin. What method would you use? (For instance, suppose the twin you see is 6 feet 6 inches.)
  - Find the r.m.s. error for the method in (c).

*The answers to these exercises are on pp. A64–65.*

## 5. USING THE NORMAL CURVE INSIDE A VERTICAL STRIP

Often, it is possible to use the normal approximation when working inside a vertical strip. For this to be legitimate, the scatter diagram has to be football-shaped, with the dots thickly scattered in the center of the picture and fading off toward the edges. Figure 8 is a good example. On the other hand, if the scatter diagram is heteroscedastic (figure 9), or shows a non-linear pattern (figure 7), do not use the method of this section. With the height-weight data in figure 6, the normal curve would not work especially well either: the cloud isn't football-shaped, it is stretched out on top and squeezed in at the bottom.

*Example 1.* A law school finds the following relationship between LSAT scores and first-year scores (for students who finish the first year):

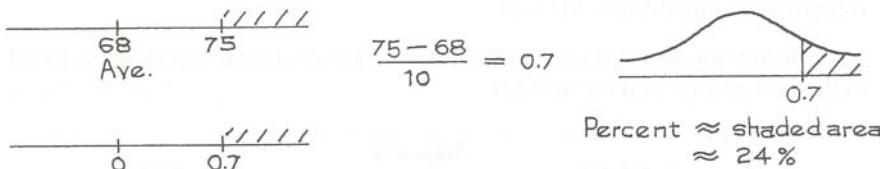
$$\text{average LSAT score} = 162, \quad \text{SD} = 6$$

$$\text{average first-year score} = 68, \quad \text{SD} = 10, \quad r = 0.60$$

The scatter diagram is football-shaped.

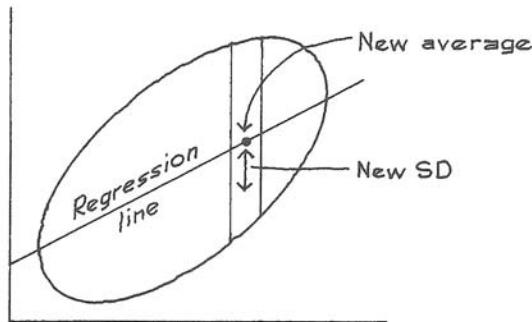
- About what percentage of the students had first-year scores over 75?
- Of the students who scored 165 on the LSAT, about what percentage had first-year scores over 75?

*Solution. Part (a).* This is a straightforward normal approximation problem. The LSAT results and  $r$  have nothing to do with it.



*Part (b).* This is a new problem. It is about a special group of students—those who scored 165 on the LSAT. These students are all in the same vertical

Figure 10. A football-shaped scatter diagram. Take the points inside a narrow vertical strip. Their  $y$ -values are a new data set. The new average is given by the regression method. The new SD is given by the r.m.s. error of the regression line. Inside the strip, a typical  $y$ -value is around the new average—give or take the new SD.



strip (figure 10). Their first-year scores are a new data set. To do the normal approximation, you need the average and the SD of this new data set.

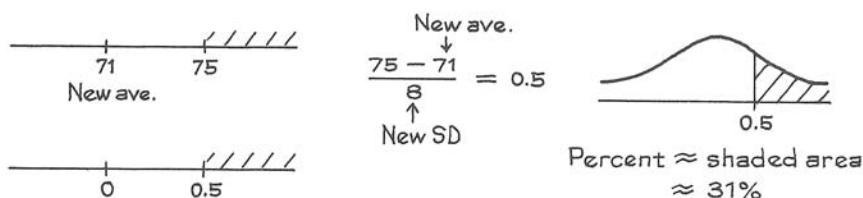
*The new average.* The students who scored 165 on the LSAT are better than average. As a group, they will do better than average in the first year of law school—although there is a fair amount of spread (vertical scatter inside the strip). The group average can be estimated by the regression method: 165 is 0.5 SDs above average, so the group will score above average in the first year, by about  $r \times 0.5 = 0.6 \times 0.5 = 0.3$  SDs. This is  $0.3 \times 10 = 3$  points. The new average is  $68 + 3 = 71$ .

*The new SD.* The students who scored 165 on the LSAT are a smaller and more homogeneous group. So the SD of their first-year scores is less than 10 points. How much less? Since the diagram is football-shaped, the scatter around the regression line is about the same in each vertical strip, and is given by the r.m.s. error for the regression line (section 4). The new SD is

$$\sqrt{1 - r^2} \times \text{SD of } y = \sqrt{1 - 0.6^2} \times 10 = 8 \text{ points.}$$

(We are predicting first-year scores from LSAT scores, so the error is in first-year points: 10 goes into the formula, not 6.) A typical student who scored around 165 on the LSAT will have a first-year score of about 71, give or take 8 or so. The new average is 71, and the new SD is 8.

*The normal approximation* is the last step. This is done as usual, but is based on the new average and the new SD.



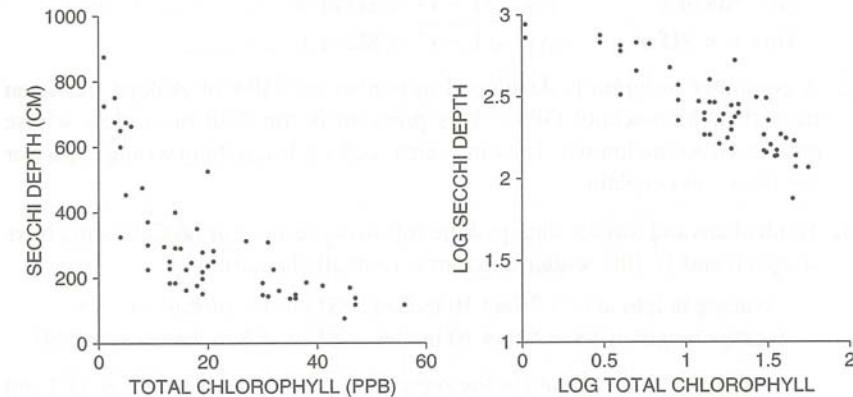
Why is the new SD smaller? Look at figure 10: there is less vertical scatter in the strip than in the whole diagram. Also see exercises 4–6 on p. 194.

Suppose that a scatter diagram is football-shaped. Take the points in a narrow vertical strip. Their  $y$ -values are a new data set. The new average is estimated by the regression method. The new SD is about equal to the r.m.s. error for the regression line.

The normal approximation can be done as usual, based on the new average and the new SD.

*Technical note.* What can you do with non-linear or heteroscedastic data? Often a transformation will help—for example, taking logarithms. The left hand panel in figure 11 shows a scatter diagram for Secchi depth (a measure of water clarity) versus total chlorophyll concentration (a measure of algae in the water).<sup>8</sup> The data are non-linear and heteroscedastic. The right hand panel shows the same data, after taking logs: the diagram is more like a football.

Figure 11. Left-hand panel: scatter diagram for Secchi depth versus total chlorophyll concentration. (Units for chlorophyll concentration are ppb, or parts per billion in the water.) Right-hand panel: data have been transformed by taking logs to base 10.



### Exercise Set E

1. Pearson and Lee obtained the following results for about 1,000 families:

$$\begin{aligned} \text{average height of husband} &\approx 68 \text{ inches}, & \text{SD} &\approx 2.7 \text{ inches} \\ \text{average height of wife} &\approx 63 \text{ inches}, & \text{SD} &\approx 2.5 \text{ inches}, & r &\approx 0.25 \end{aligned}$$

- (a) What percentage of the women were over 5 feet 8 inches?
- (b) Of the women who were married to men of height 6 feet, what percentage were over 5 feet 8 inches?

2. From the same study:

$$\begin{aligned}\text{average height of father} &\approx 68 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches} \\ \text{average height of son} &\approx 69 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches}, \quad r \approx 0.50\end{aligned}$$

- (a) What percentage of the sons were over 6 feet tall?
- (b) What percentage of the 6-foot fathers had sons over 6 feet tall?

3. From the same study:

$$\begin{aligned}\text{average height of men} &\approx 68 \text{ inches}, \quad \text{SD} \approx 2.7 \text{ inches} \\ \text{average forearm length} &\approx 18 \text{ inches}, \quad \text{SD} \approx 1 \text{ inch}, \quad r \approx 0.80\end{aligned}$$

- (a) What percentage of men have forearms which are 18 inches long, to the nearest inch?
- (b) Of the men who are 68 inches tall, what percentage have forearms which are 18 inches long, to the nearest inch?

*The answers to these exercises are on p. A65.*

## 6. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. The r.m.s. error of the regression line for predicting  $y$  from  $x$  is \_\_\_\_\_.

- |       |                            |      |   |
|-------|----------------------------|------|---|
| (i)   | SD of $y$                  | (iv) | $r \times \text{SD of } x$              |
| (ii)  | SD of $x$                  | (v)  | $\sqrt{1 - r^2} \times \text{SD of } y$ |
| (iii) | $r \times \text{SD of } y$ | (vi) | $\sqrt{1 - r^2} \times \text{SD of } x$ |

2. A computer program is developed to predict the GPA of college freshmen from their high-school GPAs. This program is tried out on a class whose college GPAs are known. The r.m.s. error is 3.12. Is anything wrong? Answer yes or no, and explain.

3. Tuddenham and Snyder obtained the following results for 66 California boys at ages 6 and 18 (the scatter diagram is football-shaped):<sup>9</sup>

$$\begin{aligned}\text{average height at 6} &\approx 3 \text{ feet 10 inches}, \quad \text{SD} \approx 1.7 \text{ inches}, \\ \text{average height at 18} &\approx 5 \text{ feet 10 inches}, \quad \text{SD} \approx 2.5 \text{ inches}, \quad r \approx 0.80\end{aligned}$$

- (a) Find the r.m.s. error for the regression prediction of height at 18 from height at 6.
- (b) Find the r.m.s. error for the regression prediction of height at 6 from height at 18.

4. A statistical analysis was made of the midterm and final scores in a large course, with the following results:

$$\begin{aligned}\text{average midterm score} &\approx 50, \quad \text{SD} \approx 25 \\ \text{average final score} &\approx 55, \quad \text{SD} \approx 15, \quad r \approx 0.60\end{aligned}$$

The scatter diagram was football-shaped. For each student, the final score was predicted from the midterm score using the regression line.

- (a) For about 1/3 of the students, the prediction for the final score was off by more than \_\_\_\_\_ points. Options: 6, 9, 12, 15, 25.  
(b) Predict the final score for a student whose midterm score was 80.  
(c) This prediction is likely to be off by \_\_\_\_\_ points or so. Options: 6, 9, 12, 15, 25.

Explain your answers.

5. Use the data in exercise 4 to answer the following questions.

- (a) About what percentage of students scored over 80 on the final?  
(b) Of the students who scored 80 on the midterm, about what percentage scored over 80 on the final?

Explain your answers.

6. In a study of high-school students, a positive correlation was found between hours spent per week doing homework, and scores on standardized achievement tests. The investigators concluded that doing homework helps prepare students for these tests. Does the conclusion follow from the data? Answer yes or no, and explain briefly.

7. The freshmen at a large university are required to take a battery of aptitude tests. Students who score high on the mathematics test also tend to score high on the physics test. On both tests, the average score is 60; the SDs are the same too. The scatter diagram is football-shaped. Of the students who scored about 75 on the mathematics test:

- (i) just about half scored over 75 on the physics test.  
(ii) more than half scored over 75 on the physics test.  
(iii) less than half scored over 75 on the physics test.

Choose one option and explain.

8. The bends are caused by rapid changes in air pressure, resulting in the formation of nitrogen bubbles in the blood. The symptoms are acute pain, and sometimes paralysis leading to death. In World War II, pilots got the bends during certain battle maneuvers. It was feasible to simulate these conditions in a pressure chamber. As a result, pilot trainees were tested under these conditions once, at the beginning of their training. If they got the bends (only mild cases were induced), they were excluded from the training on the grounds that they were more likely to get the bends under battle conditions. This procedure was severely criticized by the statistician Joe Berkson, and he persuaded the Air Force to replicate the test—that is, repeat it several times for each trainee.

- (a) Why might Berkson have suggested this?  
(b) Give another example where replication is helpful.

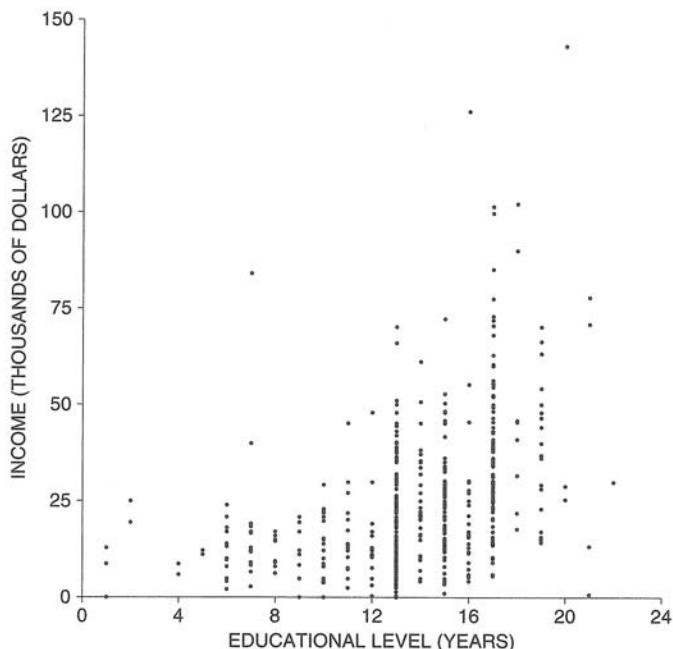
9. Every year, baseball's major leagues honor their outstanding first-year players with the title "Rookie of the Year." The overall batting average for the Rookies of the Year is around .290, far above the major league batting average of .260. However, Rookies of the Year don't do so well in their second year—their

overall second-season batting average is only .275. Baseball writers call this “sophomore slump,” the idea being that star players get distracted by outside activities like product endorsements and television appearances. Do the data support the idea of the sophomore slump? Answer yes or no, and explain briefly.<sup>10</sup>

10. A study was made of the relationship between stock prices on the last trading day of 2005 and the last trading day of 2006. A formula was developed to predict the 2006 price from the 2005 price, using data on 100 stocks. An analyst is now reviewing the results. Data are shown below for five out of the 100 stocks; prices are in dollars. Was the regression method used to predict the 2006 price from the 2005 price? Answer yes or no and explain. If you need more information, explain why.

<i>Stock</i>	<i>2005 price actual</i>	<i>2006 price predicted</i>	<i>2006 price actual</i>
A	10	8	8
B	10	8	3
C	12	13	17
D	14	12	6
E	15	20	27

11. The figure below is a scatter plot of income against education, for a representative sample of men age 25–29 in Texas. Or is something wrong? Explain briefly. (“Educational level” means years of schooling completed, not counting kindergarten.)



12. For the men age 25–34 in HANES5, the relationship between education (years of schooling completed) and systolic blood pressure can be summarized as follows.<sup>11</sup>

$$\begin{aligned}\text{average education} &\approx 13 \text{ years}, \quad \text{SD} \approx 3 \text{ years} \\ \text{average blood pressure} &\approx 119 \text{ mm}, \quad \text{SD} \approx 11 \text{ mm}, \quad r \approx -0.1\end{aligned}$$

One man in the sample had 20 years of education, and his blood pressure was 118 mm. True or false, and explain: compared to other men at his educational level, his blood pressure was a bit on the high side.

## 7. SUMMARY

- When the regression line is used to predict  $y$  from  $x$ , the difference between the actual value and the predicted value is a *residual*, or prediction error.
- In a scatter diagram, the vertical distance of a point above or below the regression line is the graphical counterpart of the prediction error made by the regression method.
- The *r.m.s. error* of the regression line is the root-mean-square of the residuals. This measures the accuracy of the regression predictions. The predictions are off by amounts similar in size to the r.m.s. error. For many scatter diagrams, about 68% of the predictions will be right to within one r.m.s. error. About 95% will be right to within two r.m.s. errors.
- The SD of  $y$  is equal to the r.m.s. error of a horizontal line through the average of  $y$ . The r.m.s. error of the regression line is smaller, by the factor  $\sqrt{1 - r^2}$ . Therefore, the r.m.s. error for the regression line of  $y$  on  $x$  can be figured as

$$\sqrt{1 - r^2} \times \text{the SD of } y.$$

- After carrying out a regression, statisticians often graph the residuals. If the *residual plot* shows a pattern, the regression may not have been appropriate.
- When all the vertical strips in a scatter diagram show similar amounts of spread, the diagram is *homoscedastic*: the prediction errors are similar in size all along the regression line. When the scatter diagram is *heteroscedastic*, the prediction errors are different in different parts of the scatter diagram. Football-shaped diagrams are homoscedastic.
- Suppose that a scatter diagram is football-shaped. Take the points inside a narrow vertical strip. Their  $y$ -values are a new data set. The new average is estimated by the regression method. The new SD is about equal to the r.m.s. error for the regression line. And the normal approximation can be done as usual, based on the new average and the new SD.

# 12

## The Regression Line

*The estimation of a magnitude using an observation subject to a larger or smaller error can be compared not inappropriately to a game of chance in which one can only lose and never win and in which each possible error corresponds to a loss .... However, what specific loss we should ascribe to any specific error is by no means clear of itself. In fact, the determination of this loss depends at least in part on our judgment .... Among the infinite variety of possible functions the one that is simplest seems to have the advantage and this is unquestionably the square .... Laplace treated the problem in a similar fashion, but he chose the size of the error as the measure of loss. However, unless we are mistaken this choice is surely not less arbitrary than ours.*

—C. F. GAUSS (GERMANY, 1777–1855)<sup>1</sup>

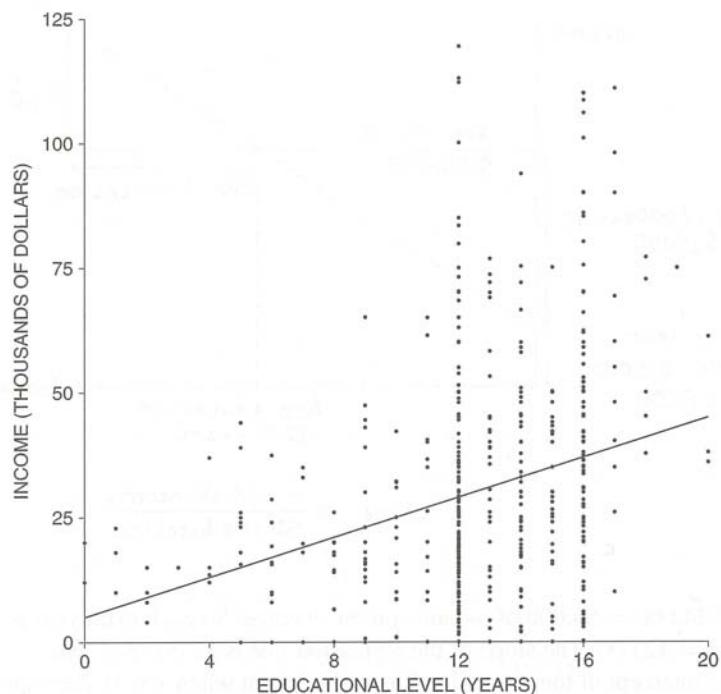
### 1. SLOPE AND INTERCEPT

Does education pay? Figure 1 shows the relationship between income and education, for a sample of 562 California men age 25–29 in 2005. The summary statistics:<sup>2</sup>

$$\begin{aligned} \text{average education} &\approx 12.5 \text{ years}, & \text{SD} &\approx 3 \text{ years} \\ \text{average income} &\approx \$30,000, & \text{SD} &\approx \$24,000, & r &\approx 0.25 \end{aligned}$$

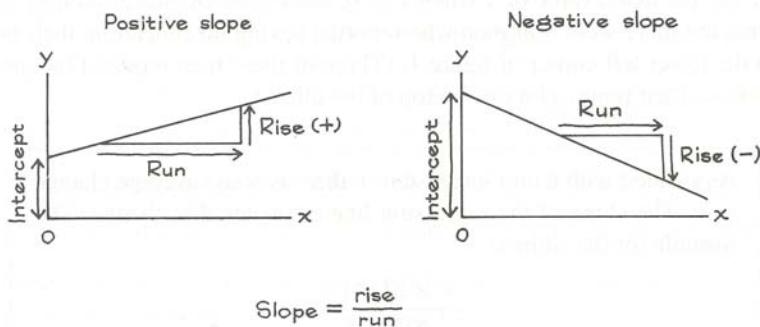
The regression estimates for average income at each educational level fall along the regression line shown in the figure. The line slopes up, showing that on the average, income does go up with education.

Figure 1. The regression line. The scatter diagram shows income and education, for a sample of 562 California men age 25–29 in 2005.



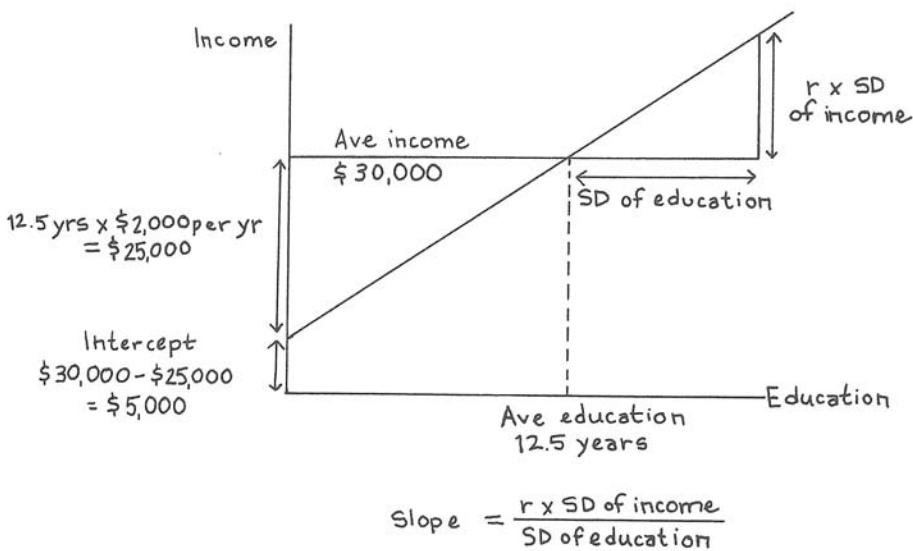
Any line can be described in terms of its slope and intercept (chapter 7). The  $y$ -intercept is the height of the line when  $x$  is 0. And the slope is the rate at which  $y$  increases, per unit increase in  $x$ . Slope and intercept are illustrated in figure 2.

Figure 2. Slope and intercept.



How do you get the slope of the regression line? Take the income-education example. Associated with an increase of one SD in education, there is an increase of  $r$  SDs in income. On this basis, 3 extra years of education are worth an extra

Figure 3. Finding the slope and intercept of the regression line.



$$\text{Slope} = \frac{r \times \text{SD of income}}{\text{SD of education}}$$

$0.25 \times \$24,000 = \$6,000$  of income, on the average. So each extra year is worth  $\$6,000/3 = \$2,000$ . The slope of the regression line is  $\$2,000$  per year.

The intercept of the regression line is the height when  $x = 0$ , corresponding to men with 0 years of education. These men are 12.5 years below average in education. Each year costs \$2,000—that is what the slope says. A man with no education should have an income which is below average by

$$12.5 \text{ years} \times \$2,000 \text{ per year} = \$25,000.$$

His income should be  $\$30,000 - \$25,000 = \$5,000$ . That is the intercept (figure 3): the predicted value of  $y$  when  $x = 0$ . Zero years of education may sound extreme, but there were four men who reported having no education; their points are in the lower left corner of figure 1. (Three of these men reported incomes of \$20,000, so their points plot one on top of the other.)

Associated with a unit increase in  $x$  there is some average change in  $y$ . The slope of the regression line estimates this change. The formula for the slope is

$$\frac{r \times \text{SD of } y}{\text{SD of } x}$$

The intercept of the regression line is just the predicted value for  $y$  when  $x$  is 0.

The equation of a line can be written in terms of the slope and intercept:

$$y = \text{slope} \times x + \text{intercept}.$$

The equation for the regression line is called (not surprisingly) the *regression equation*. In figure 3, the regression equation is

$$\text{predicted income} = \$2,000 \text{ per year} \times \text{education} + \$5,000.$$

There is nothing new here. The regression equation is just a way of predicting  $y$  from  $x$  by the regression method. However, social scientists often report the regression equation; the slope and intercept can be interesting in their own right.

*Example 1.* Education and income for 570 California women age 25–29 are shown in figure 9 on p. 192. The summary statistics are:<sup>3</sup>

$$\begin{aligned}\text{average education} &\approx 13.0 \text{ years}, & \text{SD} &\approx 3.4 \text{ years} \\ \text{average income} &\approx \$18,000, & \text{SD} &\approx \$20,000, & r &\approx 0.37\end{aligned}$$

- (a) Find the regression equation for predicting income from education.
- (b) Use the equation to predict the income of a woman whose educational level is: 8 years, 12 years, 16 years.

*Solution. Part (a).* The first step is to find the slope. In a run of one SD of education, the regression line rises  $r$  SDs of income. So

$$\text{slope} = \frac{0.37 \times \$20,000}{3.4 \text{ years}} \approx \$2,176 \text{ per year.}$$

On the average, each extra year of schooling is worth an extra \$2,176 of income; each year less of schooling costs \$2,176 of income. (Income has such a large SD because the distribution has a long right hand tail.)

The next step is to find the intercept. That is the height of the regression line at  $x = 0$ —in other words, the predicted income for a woman with no education. She is 13 years below average. Her income should be below average by

$$13 \text{ years} \times \$2,176 \text{ per year} = \$28,288.$$

Her predicted income is

$$\$18,000 - \$28,288 = -\$10,288.$$

That is the intercept: the prediction for  $y$  when  $x = 0$ . The regression equation is

$$\text{predicted income} = \$2,176 \text{ per year} \times \text{education} - \$10,288.$$

The regression line becomes unreliable when you are far from the center of the data, so a negative intercept is not too disturbing.

*Part (b).* Substitute 8 years for education, to get

$$\$2,176 \text{ per year} \times 8 \text{ years} - \$10,288 = \$7,120.$$

Substitute 12 years for education:

$$\$2,176 \text{ per year} \times 12 \text{ years} - \$10,288 = \$15,824.$$

Substitute 16 years:

$$\$2,176 \text{ per year} \times 16 \text{ years} - \$10,288 = \$24,528.$$

This completes the solution. Despite the negative intercept, the predictions are quite reasonable for most of the women.

In example 1, the slope is \$2,176 per year. Associated with each extra year of education, there is an increase of \$2,176 in income, on the average. The phrase “associated with” sounds like it is talking around some difficulty, and here is the issue. Are income differences caused by differences in educational level, or do both reflect the common influence of some third variable? The phrase “associated with” was invented to let statisticians talk about regressions without having to commit themselves on this sort of point.

The slope is often used to predict how  $y$  will respond, if someone intervenes and changes  $x$ . This is legitimate when the data come from a controlled experiment. With observational studies, the inference is often shaky—because of confounding. Look at example 1. On the average, the women who finished high school (12 years of education) earned about \$9,000 more than women who just finished middle school (8 years).

Now, take a representative group of women with 8 years of education. If the government intervened and sent them on to get high-school degrees, the slope suggests that their incomes would go up by an average of  $4 \times \$2,176 \approx \$9,000$ . However, example 1 is based on survey data rather than a controlled experiment. One group of women in the survey had 8 years of education. Another, separate, group had 12 years. The two groups were different with respect to many factors besides education—like intelligence, ambition, and family background.

The effects of these factors are confounded with the effect of education, and go into the slope. Sending people off to high school probably would make their incomes go up, but not by the full \$9,000. To measure the impact on incomes, it might be necessary to run a controlled experiment. (Many investigators would use a technique called *multiple regression*; more about this in section 3.<sup>4</sup>)

With an observational study, the slope and intercept of the regression line are only descriptive statistics. They say how the average value of one variable is related to values of another variable in the population being observed. The slope cannot be relied on to predict how  $y$  would respond if you intervene to change the value of  $x$ .

If you run an observational study, the regression line only describes the data that you see. The line cannot be relied on for predicting the results of interventions.

There is another assumption that we have been making throughout this section: that the average of  $y$  depends linearly on  $x$ . If the relationship is non-linear, the

regression line may be quite misleading—whether the data come from an experiment or an observational study.<sup>5</sup>

### Exercise Set A

- For the men in figure 1, the regression equation for predicting average income from education is

$$\text{predicted income} = \$2,000 \text{ per year} \times \text{education} + \$5,000.$$

Predict the income for one of these men who has

- (a) 8 years of schooling—elementary education
- (b) 12 years of schooling—a high-school diploma
- (c) 16 years of schooling—a college degree.

- The International Rice Research Institute in the Philippines developed the hybrid rice IR 8, setting off “the green revolution” in tropical agriculture. Among other things, they made a thorough study of the effects of fertilizer on rice yields. These experiments involved a number of experimental plots (of about 20 square yards in size). Each plot was planted with IR 8, and fertilized with some amount of nitrogen chosen by the investigators. (The amounts ranged from 0 to about a pound.) When the rice was harvested, the yield was measured and related to the amount of nitrogen applied. In one such experiment, the correlation between rice yield and nitrogen was 0.95, and the regression equation was<sup>6</sup>

$$\text{predicted rice yield} = (20 \text{ oz rice per oz nitrogen}) \times (\text{nitrogen}) + 240 \text{ oz.}$$

- (a) An unfertilized plot can be expected to produce around \_\_\_\_\_ of rice.
  - (b) Each extra ounce of nitrogen fertilizer can be expected to increase the rice yield by \_\_\_\_\_.
  - (c) Predict the rice yield when the amount of fertilizer is  

$$\begin{array}{ll} 3 \text{ ounces of nitrogen} & 4 \text{ ounces of nitrogen} \end{array}$$
  - (d) Was this an observational study or a controlled experiment?
  - (e) In fact, fertilizer was applied only in the following amounts: 0 ounces, 4 ounces, 8 ounces, 12 ounces, 16 ounces. Would you trust the prediction for 3 ounces of nitrogen, even though this particular amount was never applied?
  - (f) Would you trust the prediction for 100 ounces of nitrogen?
- Summary statistics for heights of fathers and sons are on p. 170.
    - Find the regression equation for predicting the height of a son from the height of his father.
    - Find the regression equation for predicting the height of a father from the height of his son.
  - An expert witness offers testimony that<sup>7</sup>  
 Regression is a substitute for controlled experiments. It provides a precise estimate of the effect of one variable on another.

Comment briefly.

*The answers to these exercises are on pp. A65–66.*

## 2. THE METHOD OF LEAST SQUARES

Chapter 10 discussed regression from one point of view, and section 1 went over the same ground using the regression equation. This section is a third pass at the same topic, from yet another perspective. (For statisticians, regression is an important technique.) Sometimes the points on a scatter diagram seem to be following a line. The problem discussed in this section is how to find the line which best fits the points. Usually, this involves a compromise: moving the line closer to some points will increase its distance from others. To resolve the conflict, two steps are necessary. First, define an average distance from the line to all the points. Second, move the line around until this average distance is as small as possible.

To be more specific, suppose the line will be used to predict  $y$  from  $x$ . Then, the error made at each point is the vertical distance from the point to the line. In statistics, the usual way to define the average distance is by taking the root-mean-square of the errors. This measure of average distance is called the *r.m.s. error of the line*. (It was first proposed by Gauss; see the chapter opening quote.)

The second problem, how to move the line around to minimize the r.m.s. error, was also solved by Gauss.

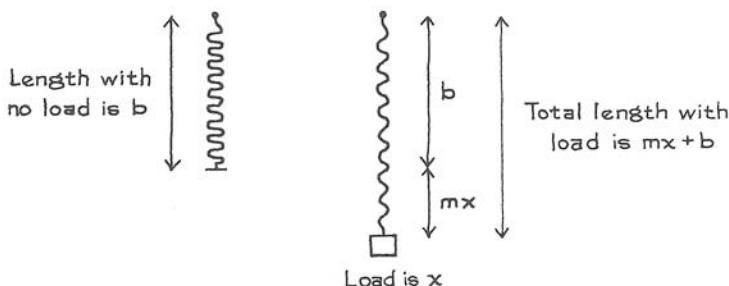
Among all lines, the one that makes the smallest r.m.s. error in predicting  $y$  from  $x$  is the regression line.

For this reason, the regression line is often called the *least squares line*: the errors are squared to compute the r.m.s. error, and the regression line makes the r.m.s. error as small as possible. (The r.m.s. error of the regression line was discussed in section 1 of chapter 11.)

Now, an example. Robert Hooke (England, 1653–1703) was able to determine the relationship between the length of a spring and the load placed on it. He just hung weights of different sizes on the end of a spring, and watched what happened. When he increased the load, the spring got longer. When he reduced the load, the spring got shorter. And the relationship was linear.

Let  $b$  be the length of the spring with no load. A weight of  $x$  kilograms is attached to the end of the spring. As illustrated in figure 4, the spring stretches to

Figure 4. Hooke's law: the stretch is proportional to the load.



a new length. According to Hooke's law, the amount of stretch is proportional to the weight  $x$ . The new length of the spring is

$$y = mx + b.$$

In this equation,  $m$  and  $b$  are constants which depend on the spring. Their values are unknown, and have to be estimated using experimental data.

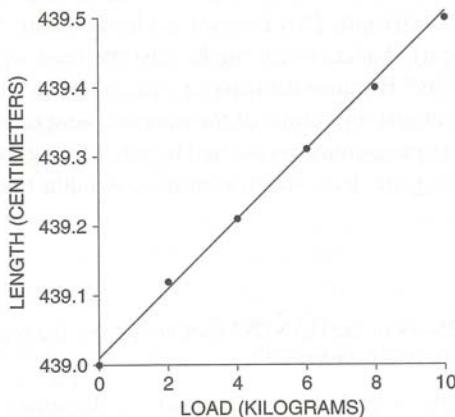
Table 1 shows the results of an experiment on Hooke's law, done in a physics class at the University of California, Berkeley. Different weights were hung on the end of a long piano wire.<sup>8</sup> The first column shows the load. The second column shows the measured length. With 20 pounds of load, this "spring" only stretched about 0.2 inch ( $10 \text{ kg} \approx 22 \text{ lb}$ ,  $0.5 \text{ cm} \approx 0.2 \text{ in}$ ). Piano wire is not very stretchy.

Table 1. Data on Hooke's law.

Weight (kg)	Length (cm)
0	439.00
2	439.12
4	439.21
6	439.31
8	439.40
10	439.50

The correlation coefficient for the data in table 1 is 0.999, very close to 1 indeed. So the points almost form a straight line (figure 5), just as Hooke's law predicts. The minor deviations from linearity are probably due to measurement error; neither the weights nor the lengths have been measured with perfect accuracy. (Nothing ever is.)

Figure 5. Scatter diagram for table 1.



Our goal is to estimate  $m$  and  $b$  in the equation of Hooke's law for the piano wire:

$$y = mx + b.$$

The graph of this equation is a perfect straight line. If the points in figure 5 happened to fall exactly on some line, the slope of that line would estimate  $m$ , and its intercept would estimate  $b$ . However, the points do not line up perfectly. Many different lines could be drawn across the scatter diagram, each having a slightly different slope and intercept.

Which line should be used? Hooke's equation predicts length from weight. As discussed above, it is natural to choose  $m$  and  $b$  so as to minimize the r.m.s. error: this is the *method of least squares*. The line  $y = mx + b$  which does the job is the regression line.<sup>9</sup> In other words,  $m$  in Hooke's law should be estimated as the slope of the regression line, and  $b$  as its intercept. These are called the *least squares estimates*, because they minimize root-mean-square error. If you do the arithmetic,

$$m \approx 0.05 \text{ cm per kg} \text{ and } b \approx 439.01 \text{ cm}$$

The length of the spring under no load is estimated as 439.01 cm. And each kilogram of load causes the spring to stretch by about 0.05 cm. There is no need to hedge, because the estimates are based on a controlled experiment. The investigator puts the weights on, and the wire stretches. Take the weights off, and the wire comes back to its original length. This can be repeated as often as you want. There is no question here about what is causing what; the language of "association" is not needed. Of course, even Hooke's law has its limits: beyond some point, the spring will break. *Extrapolating beyond the data is risky.*

The method of least squares and the regression method involve the same mathematics; but the contexts may be different. In some fields, investigators talk about "least squares" when they are estimating parameters—unknown constants of nature like  $m$  and  $b$  in Hooke's law. In other fields, investigators talk about regression when they are studying the relationship between two variables, like income and education, using non-experimental data.

A technical point: The least squares estimate for the length of the spring under no load was 439.01 cm. This is a tiny bit longer than the measured length at no load (439.00 cm). A statistician might trust the least squares estimate over the measurement. Why? Because the least squares estimate takes advantage of all six measurements, not just one: some of the measurement error is likely to cancel out. Of course, the six measurements are tied together by a good theory—Hooke's law. Without the theory, the least squares estimate wouldn't be worth much.

### Exercise Set B

- For the men age 25–34 in the HANES2 sample (p. 58), the regression equation for predicting height from education is<sup>10</sup>

$$\text{predicted height} = (0.25 \text{ inches per year}) \times (\text{education}) + 66.75 \text{ inches}$$

Predict the height of a man with 12 years of education; with 16 years of education. Does going to college increase a man's height? Explain.

2. For the data in table 1 (p. 209), the regression equation for predicting length from weight is

$$\text{predicted length} = (0.05 \text{ cm per kg}) \times (\text{weight}) + 439.01 \text{ cm}$$

Predict the length of the wire when the weight is 3 kg; 5 kg. Does putting more weight on the spring make it longer? Explain.

3. A study is made of Math and Verbal SAT scores for the entering class at a certain college. The summary statistics:

$$\text{average M-SAT} = 560, \quad \text{SD} = 120$$

$$\text{average V-SAT} = 540, \quad \text{SD} = 110, \quad r = 0.66$$

The investigator uses the SD line to predict V-SAT score from M-SAT score.

(a) If a student scores 680 on the M-SAT, the predicted V-SAT score is \_\_\_\_\_.

(b) If a student scores 560 on the M-SAT, the predicted V-SAT score is \_\_\_\_\_.

(c) The investigator's r.m.s. error is \_\_\_\_\_  $\sqrt{1 - 0.66^2} \times 110$ . Options:

greater than      equal to      less than

If more information is needed, say what you need, and why.

4. Repeat exercise 3, if the investigator always predicts a V-SAT of 540.

5. Exercise 3 describes one way to predict V-SAT from M-SAT; exercise 4 describes a second way; and regression is a third way. Which way will have the smallest r.m.s. error?

*The answers to these exercises are on p. A66.*

### 3. DOES THE REGRESSION MAKE SENSE?

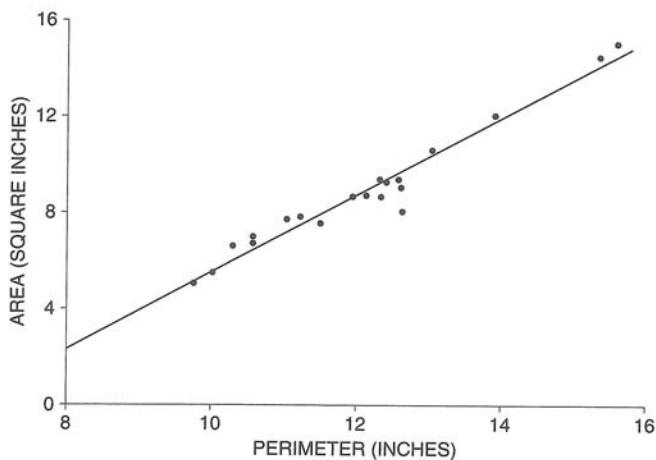
A regression line can be put down on any scatter diagram. However, there are two questions to ask: First, was there a non-linear association between the variables? If so, the regression line may be quite misleading (p. 163). Even if the association looks linear, there is a second question: Did the regression make sense? The second question is harder. Answering it requires some understanding of the mechanism which produced the data. If this mechanism is not understood, fitting a line can be intellectually disastrous.

To make up an example, suppose an investigator does not know the formula for the area of a rectangle. He thinks area ought to depend on perimeter. Taking an empirical approach, he draws 20 typical rectangles, measuring the area and the perimeter for each one. The correlation coefficient turns out to be 0.98—almost as good as Hooke's law. The investigator concludes that he is really on to something. His regression equation is

$$\text{area} = (1.60 \text{ inches}) \times (\text{perimeter}) - 10.51 \text{ square inches}$$

(Area is measured in square inches and perimeter in inches.)

Figure 6. Scatter diagram of area against perimeter for 20 rectangles; the regression line is shown too.



There is a scatter diagram in figure 6, with one dot for each rectangle; the regression line is plotted too. The rectangles themselves are shown in figure 7. The arithmetic is all in order, but the regression is silly. The investigator should have looked at two other variables, length and width. These two variables determine both area and perimeter:

$$\text{area} = \text{length} \times \text{width}, \quad \text{perimeter} = 2(\text{length} + \text{width})$$

Our straw-man investigator would never find this out by doing regressions.

When looking at a regression study, ask yourself whether it is more like Hooke's law, or more like area and perimeter. Of course, the area-perimeter example is hypothetical. But many investigators do fit lines to data without facing the issues. That can make a lot of trouble.<sup>11</sup>

*Technical note.* Example 1 in section 1 presented a regression equation for predicting income from education. This is a good way to describe the relationship between income and education. But it may not be legitimate to interpret the slope as the effect on income if you intervene to change education. The problem—the effects of other variables may be confounded with the effects of education.

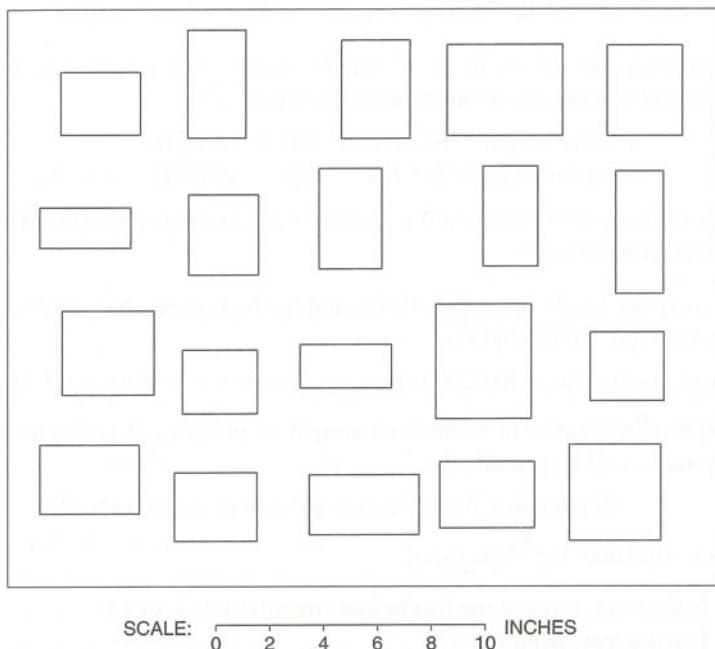
Many investigators would use multiple regression to control for other variables. For instance, they might develop some measure for the socioeconomic status of parents, and fit a multiple regression equation of the form

$$y = a + b \times E + c \times S,$$

where

$$\begin{aligned} y &= \text{predicted income}, \quad E = \text{educational level}, \\ S &= \text{measure of parental status}. \end{aligned}$$

Figure 7. The 20 rectangles themselves.



The coefficient  $b$  would be interpreted as showing the effect of education, controlling for the effect of parental status.

This might give sensible results. But it can equally well produce nonsense. Take the hypothetical investigator who was working on the area of rectangles. He could decide to control for the shape of the rectangles by multiple regression, using the length of the diagonal to measure shape. (Of course, this isn't a good measure of shape, but nobody knows how to measure status very well either.) The investigator would fit a multiple regression equation of the form

$$\text{predicted area} = a + b \times \text{perimeter} + c \times \text{diagonal}.$$

He might tell himself that  $b$  measures the effect of perimeter, controlling for the effect of shape. As a result, he would be even more confused than before. The perimeter and the diagonal do determine the area, but only by a non-linear formula. Multiple regression is a powerful technique, but it is not a substitute for understanding.

#### 4. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

- Find the regression equation for predicting final score from midterm score, based on the following information:

$$\begin{aligned}\text{average midterm score} &= 70, \quad \text{SD} = 10 \\ \text{average final score} &= 55, \quad \text{SD} = 20, \quad r = 0.60\end{aligned}$$

2. For women age 25–34 in the HANES5 sample, the relationship between height and income can be summarized as follows:<sup>12</sup>

$$\begin{aligned}\text{average height} &\approx 64 \text{ inches}, \quad \text{SD} \approx 2.5 \text{ inches} \\ \text{average income} &\approx \$21,000, \quad \text{SD} \approx \$20,000, \quad r \approx 0.2\end{aligned}$$

What is the regression equation for predicting income from height? What does the equation tell you?

3. For men age 18–24 in the HANES5 sample, the regression equation for predicting height from weight is

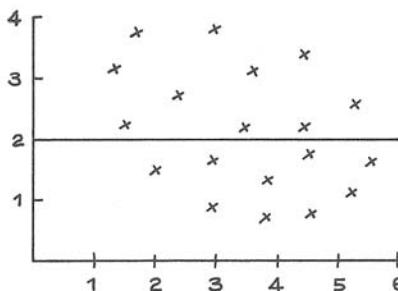
$$\text{predicted height} = (0.0267 \text{ inches per pound}) \times (\text{weight}) + 65.2 \text{ inches}$$

(Height is measured in inches and weight in pounds.) If someone puts on 20 pounds, will he get taller by

$$20 \text{ pounds} \times 0.0267 \text{ inches per pound} \approx 0.5 \text{ inches?}$$

If not, what does the slope mean?

4. (a) Is the r.m.s. error of the line below around 0.1, 0.3, or 1?  
 (b) Is it the regression line?



5. A study is made of working couples. The regression equation for predicting wife's income from husband's income is

$$\text{wife's income} = 0.1667 \times \text{husband's income} + \$24,000.$$

Another investigator solves this equation for husband's income, and gets

$$\text{husband's income} = 6 \times \text{wife's income} - \$144,000.$$

True or false, and explain: the second investigator has found the regression equation for predicting husband's income from wife's income. If you want to compute anything,

$$\begin{aligned}\text{husband's average income} &= \$54,000, \quad \text{SD} = \$39,000 \\ \text{wife's average income} &= \$33,000, \quad \text{SD} = \$26,000, \quad r = 0.25\end{aligned}$$

6. (Continues exercise 5.) The couples in the previous exercise are followed for a year. Suppose everyone's income goes up by 10%. Find the new regression line for predicting wife's income from husband's income.
7. A statistician is doing a study on a group of undergraduates. On average, these students drink 4 beers a month, with an SD of 8. They eat 4 pizzas a month, with an SD of 4. There is some positive association between beer and pizza, and the regression equation is<sup>13</sup>
- predicted number of beers = \_\_\_\_\_ × number of pizzas + 2.
- However, the statistician lost the data and forgot the slope of the equation. (Perhaps he had too much beer and pizza.) Can you help him remember the slope? Explain.
8. An investigator wants to use a straight line to predict IQ from lead levels in the blood, for a representative group of children aged 5–9.<sup>14</sup> There is a weak positive association in the data. True or false, and explain—
- He can use many different lines.
  - He has to use the regression line.
  - Only the regression line has an r.m.s. error.
  - Any line he uses will have an r.m.s. error.
  - Among all lines, the regression line has the smallest r.m.s. error.

9. In a large study (hypothetical) of the relationship between parental income and the IQs of their children, the following results were obtained:

$$\begin{array}{ll} \text{average income} \approx \$90,000, & \text{SD} \approx \$45,000 \\ \text{average IQ} \approx 100, & \text{SD} \approx 15, \quad r \approx 0.50 \end{array}$$

For each income group (\$0–\$9999, \$10,000–\$19,999, \$20,000–\$29,999, etc.), the average IQ of children with parental income in that group was calculated and then plotted above the midpoint of the group (\$5,000, \$15,000, \$25,000, etc.). It was found that the points on this graph followed a straight line very closely. The slope of this line (in IQ points per dollar) would be about:

$$\begin{array}{ccccccccc} 6,000 & 3,000 & 1,500 & 500 & 1/500 & 1/1,500 & 1/3,000 & 1/6,000 \\ & & & & & & & \\ & & & & & & & \text{can't say from the information given} \end{array}$$

Explain briefly.

10. One child in the study referred to in exercise 9 had an IQ of 110, but the information about his parents' income was lost. At \$150,000 the height of the line plotted in exercise 9 corresponds to an IQ of 110. Is \$150,000 a good estimate for the parents' income? Or is the estimate likely to be too high? too low? Explain.
11. (Hypothetical.) A congressional report is discussing the relationship between income of parents and educational attainment of their daughters. Data are

from a sample of families with daughters age 18–24. Average parental income is \$79,300; average educational attainment of the daughters is 12.7 years of schooling completed; the correlation is 0.37.

The regression line for predicting daughter's education from parental income is reported as  $y = mx + b$ , with  $x$  = parental income (dollars),  $y$  = predicted education (years),  $m = 0.00000925$  years per dollar, and  $b = 10.3$  years:

$$\text{predicted education} = 0.00000925 \times \text{income} + 10.3$$

Is anything wrong? Or do you need more information to decide? Explain briefly.

12. Many epidemiologists think that a high level of salt in the diet causes high blood pressure. INTERSALT is often cited to support this view. INTERSALT was a large study done at 52 centers in 32 countries.<sup>15</sup> Each center recruited 200 subjects in 8 age- and sex-groups. Salt intake was measured, as well as blood pressure and several possible confounding variables. After adjusting for age, sex, and the other confounding variables, the authors found a significant association between high salt intake and high blood pressure. However, a more detailed analysis showed that in 25 of the centers, there was a positive association between blood pressure and salt; in the other 27, the association was negative. Do the data support the theory that high levels of salt cause high blood pressure? Answer yes or no, and explain briefly.

## 5. SUMMARY AND OVERVIEW

1. The regression line can be specified by two descriptive statistics: the *slope* and the *intercept*.

2. The slope of the regression line for  $y$  on  $x$  is the average change in  $y$ , per unit change in  $x$ . This equals

$$r \times \text{SD of } y / \text{SD of } x.$$

3. The intercept of the regression line equals the regression estimate for  $y$ , when  $x$  is 0.

4. The equation of the regression line for  $y$  on  $x$  is

$$y = \text{slope} \times x + \text{intercept}.$$

5. The equation can be used to make all the regression predictions, by substitution.

6. Among all lines, the regression line for  $y$  on  $x$  makes the smallest r.m.s. error in predicting  $y$  from  $x$ . For that reason, the regression line is often called the *least squares line*.

7. Sometimes, two quantities are thought to be connected by a linear relationship (for example, length and weight in Hooke's law). The statistical problem is to estimate the slope and intercept of the line. The *least squares estimates* are the slope and intercept of the regression line.

8. In this part of the book, scatter diagrams are used to graph the association between two variables. If the scatter diagram is football-shaped, it can be summarized by the average and SD of the two variables, with  $r$  to measure the strength of the association.

9. How does the average of one variable depend on the values of the other variable? The regression line can be used to answer that question.

10. With a controlled experiment, the slope can tell you the average change in  $y$  that would be caused by a change in  $x$ . With an observational study, however, the slope cannot be relied on to predict the results of interventions. It takes a lot of hard work to draw causal inferences from observational data, with or without regression.

11. If the average of  $y$  depends on  $x$  in a non-linear way, the regression line can be quite misleading.

— — — — —

PART IV

# Probability

— — — — —

“The probability that a number selected at random from within a unit interval will be greater than  $\frac{1}{2}$  is  $\frac{1}{2}$ . The probability that it will be greater than  $\frac{1}{3}$  is  $\frac{2}{3}$ . The probability that it will be greater than  $\frac{1}{4}$  is  $\frac{3}{4}$ . The probability that it will be greater than  $\frac{1}{5}$  is  $\frac{4}{5}$ . As we increase the number of terms in the sequence, the probability that the sum of the first  $n$  terms will be greater than  $\frac{n}{2}$  increases and approaches 1. In fact, it can be shown that the probability that the sum of the first  $n$  terms will be greater than  $\frac{n}{2} + \epsilon$  for any  $\epsilon > 0$  approaches 1 as  $n$  increases without bound.”

## 13

# What Are the Chances?

*In the long run, we are all dead.*

—JOHN MAYNARD KEYNES (ENGLAND, 1883–1946)

### 1. INTRODUCTION

People talk loosely about chance all the time, without doing any harm. What are the chances of getting a job? of meeting someone? of rain tomorrow? But for scientific purposes, it is necessary to give the word *chance* a definite, clear interpretation. This turns out to be hard, and mathematicians have struggled with the job for centuries. They have developed some careful and rigorous theories, but these theories cover just a small range of the cases where people ordinarily speak of chance. This book will present the *frequency theory*, which works best for processes which can be repeated over and over again, independently and under the same conditions.<sup>1</sup> Many games fall into this category, and the frequency theory was originally developed to solve gambling problems. One of the great early masters was Abraham de Moivre, a French Protestant who fled to England to avoid religious persecution. Part of the dedication to his book, *The Doctrine of Chances*, is reproduced in figure 1 on the next page.<sup>2</sup>

Figure 1. De Moivre's dedication to *The Doctrine of Chances*.

To the Right Honorable the  
Lord CARPENTER.

My Lord,

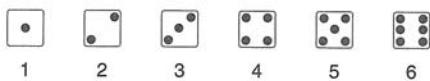
There are many people in the World who are prepossessed with an Opinion, that the Doctrine of Chances has a Tendency to promote Play; but they soon will be undeceived, if they think fit to look into the general Design of this Book; in the mean time it will not be improper to inform them, that your Lordship is pleased to espouse the Patronage of this second Edition; which your strict Probity, and the distinguished Character you bear in the World, would not have permitted, were not their Apprehensions altogether groundless.

Your Lordship does easily perceive, that this Doctrine is so far from encouraging Play, that it is rather a Guard against it, by setting in a clear light, the Advantages and Disadvantages of those Games wherein Chance is concerned . . . .

Another use to be made of this Doctrine of Chances is that it may serve in conjunction with the other parts of the Mathematicks, as a fit Introduction to the Art of Reasoning: it being known by experience that nothing can contribute more to the attaining of that Art, than the consideration of a long Train of Consequences, rightly deduced from undoubted Principles, of which this Book affords many Examples.

One simple game of chance involves betting on the toss of a coin. The process of tossing the coin can be repeated over and over again, independently and under the same conditions. The chance of getting heads is 50%: in the long run, heads will turn up about 50% of the time.

Take another example. A die (plural, “dice”) is a cube with six faces, labelled



When the die is rolled, the faces are equally likely to turn up. The chance of getting an ace— $\square$ —is 1 in 6, or  $16\frac{2}{3}\%$ . The interpretation: if the die is rolled over and over again, repeating the basic chance process under the same conditions, in the long run an ace will show about  $16\frac{2}{3}\%$  of the time.

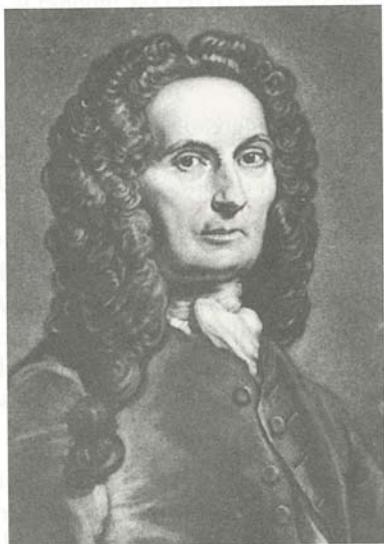
The chance of something gives the percentage of time it is expected to happen, when the basic process is done over and over again, independently and under the same conditions.

If something is impossible, it happens 0% of the time. At the other extreme, if something is sure to happen, then it happens 100% of the time. All chances are between these two extremes.

Chances are between 0% and 100%.

Here is another basic fact. Suppose you are playing a game, and have a 45% chance to win. In other words, you expect to win about 45% of the time. So you must expect to lose the other 55% of the time.

The chance of something equals 100% minus the chance of the opposite thing.



Abraham de Moivre (England, 1667–1754)  
Etching by Faber. Copyright © British Museum.

*Example 1.* A box contains red marbles and blue marbles. One marble is drawn at random from the box (each marble has an equal chance to be drawn). If it is red, you win \$1. If it is blue, you win nothing. You can choose between two boxes:

- box A contains 3 red marbles and 2 blue ones;
- box B contains 30 red marbles and 20 blue ones.

Which box offers a better chance of winning, or are they the same?

*Solution.* Some people prefer box A, because it has fewer blue marbles. Others prefer B, because it has more red marbles. Both views are wrong. The two boxes offer the same chance of winning, 3 in 5. To see why, imagine drawing many times at random from box A (replacing the marble after each draw, so as not to change the conditions of the experiment). In the long run each of the

5 marbles will appear about 1 time in 5. So the red marbles will turn up about  $\frac{3}{5}$  of the time. With box A, your chance of drawing a red marble is  $\frac{3}{5}$ , that is, 60%.

Now imagine drawing many times at random with replacement from box B. Each of the 50 marbles will turn up about 1 time in 50. But now there are 30 red marbles. With box B, your chance of winning is  $\frac{30}{50} = \frac{3}{5} = 60\%$ , just as for box A. What counts is the ratio

$$\frac{\text{number of red marbles}}{\text{total number of marbles}}.$$

The ratio is the same in both boxes. De Moivre's solution for this example is given in figure 2.

Figure 2. De Moivre's solution.

The Probability of an Event is greater or less, according to the number of Chances by which it may happen, compared with the whole number of Chances by which it may either happen or fail.

Wherefore, if we constitute a Fraction whereof the Numerator be the number of Chances whereby an Event may happen, and the Denominator the number of all the Chances whereby it may either happen or fail, that Fraction will be a proper designation of the Probability of it happening. Thus if an Event has 3 Chances to happen, and 2 to fail, the Fraction  $\frac{3}{5}$  will fitly represent the Probability of its happening, and may be taken as the measure of it.

The same things may be said of the Probability of failing, which will likewise be measured by a Fraction, whose Numerator is the number of Chances whereby it may fail, and the Denominator the whole number of Chances, both for its happening and failing; thus the Probability of the failing of that Event which has 2 Chances to fail and 3 to happen will be measured by the Fraction  $\frac{2}{5}$ .

The Fractions which represent the Probabilities of happening and failing, being added together, their Sum will always be equal to Unity, since the Sum of their Numerators will be equal to their common Denominator: now it being a certainty that an Event will either happen or fail, it follows that Certainty, which may be conceived under the notion of an infinitely great degree of Probability, is fitly represented by Unity. [By "Unity," de Moivre means the number 1.]

These things will easily be apprehended, if it be considered that the word Probability includes a double Idea: first, of the number of Chances whereby an Event may happen; secondly, of the number of Chances whereby it may either happen or fail.

Many problems, like example 1, take the form of drawing at random from a box. A typical instruction is,

Draw two tickets at random WITH replacement from the box



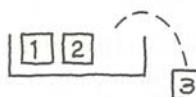
This asks you to imagine the following process: shake the box, draw out one ticket at random (equal chance for all three tickets), make a note of the number on it, put it back in the box, shake the box again, draw a second ticket at random (equal chance for all three tickets), make a note of the number on it, and put the second ticket back in the box. The contrast is with the instruction,

Draw two tickets at random WITHOUT replacement from the box



The second instruction asks you to imagine the following process: shake the box, draw out one ticket at random (equal chance for all three tickets), set it aside, draw out a second ticket at random (equal chance for the two tickets left in the box). See figure 3.

Figure 3. The difference between drawing with or without replacement. Two draws are made at random from the box Suppose the first draw is .



WITH replacement . . . the second draw is from



WITHOUT replacement . . . the second draw is from



When you draw at random, all the tickets in the box have the same chance to be picked.

### Exercise Set A

1. A computer is programmed to compute various chances. Match the numerical answers with the verbal descriptions (which may be used more than once).

*Numerical answer*

- (a) -50%  
 (b) 0%  
 (c) 10%  
 (d) 50%  
 (e) 90%  
 (f) 100%  
 (g) 200%

*Verbal description*

- (i) This is as likely to happen as not.  
 (ii) This is very likely to happen, but it's not certain.  
 (iii) This won't happen.  
 (iv) This may happen, but it's not likely.  
 (v) This will happen, for sure.  
 (vi) There's a bug in the program.

2. A coin will be tossed 1,000 times. About how many heads are expected?  
 3. A die will be rolled 6,000 times. About how many aces are expected?  
 4. In five-card draw poker, the chance of being dealt a full house (one pair and three of a kind) is 0.14 of 1%. If 10,000 hands are dealt, about how many will be a full house?  
 5. One hundred tickets will be drawn at random with replacement from one of the two boxes shown below. On each draw, you will be paid the amount shown on the ticket, in dollars. Which box is better and why?

(i) 

1	2
---	---

      (ii) 

1	3
---	---

*The answers to these exercises are on p. A66.*

## 2. CONDITIONAL PROBABILITIES

This section introduces conditional probabilities. The examples involve cards. A deck of cards has 4 suits: clubs, diamonds, hearts, spades. There are 13 cards in each suit: 2 through 10, jack, queen, king, ace. So there are  $4 \times 13 = 52$  cards in the deck.

*Example 2.* A deck of cards is shuffled and the top two cards are put on a table, face down. You win \$1 if the second card is the queen of hearts.

- (a) What is your chance of winning the dollar?  
 (b) You turn over the first card. It is the seven of clubs. Now what is your chance of winning?

*Solution. Part (a).* The bet is about the second card, not the first. Initially, this will seem a little strange. Some illustrations may help.

- If the first card is the two of spades and the second is the queen of hearts, you win.
- If the first card is the jack of clubs and the second is the queen of hearts, you win.
- If the first card is the seven of clubs and the second is the king of hearts, you lose.

The bet can be settled without even looking at the first card. The second card is all you need to know.

The chance of winning is 1/52. To see why, think about shuffling the deck. That brings the cards into random order. The queen of hearts has to wind up somewhere. There are 52 possible positions, and they are all equally likely. So there is 1 chance in 52 for her to wind up as the second card in the deck—and bring you the dollar.

*Part (b).* There are 51 cards left. They are in random order, and the queen of hearts is one of them. She has 1 chance in 51 to be on the table. Your chance goes up a little, to 1/51. That is the answer.

The 1/51 in part (b) is a *conditional* chance. The problem puts a condition on the first card: it has to be the seven of clubs. A mathematician might talk about the conditional probability that the second card is the queen of hearts *given* the first card is the seven of clubs. To emphasize the contrast, the 1/52 in part (a) is called an *unconditional* chance: the problem puts no conditions on the first card.

### Exercise Set B

1. Two tickets are drawn at random without replacement from the box 1 2 3 4.

  - (a) What is the chance that the second ticket is 4?
  - (b) What is the chance that the second ticket is 4, given the first is 2?

2. Repeat exercise 1, if the draws are made with replacement.
3. A penny is tossed 5 times.
  - (a) Find the chance that the 5th toss is a head.
  - (b) Find the chance that the 5th toss is a head, given the first 4 are tails.
4. Five cards are dealt off the top of a well-shuffled deck.
  - (a) Find the chance that the 5th card is the queen of spades.
  - (b) Find the chance that the 5th card is the queen of spades, given that the first 4 cards are hearts.

*The answers to these exercises are on pp. A66–67.*

*Technical notes.* (i) Mathematicians write the probability for the second card to be the queen of hearts as follows:

$$P(\text{2nd card is queen of hearts}).$$

The “P” is short for “probability.”

(ii) The conditional probability for the second card to be the queen of hearts, given the first was the seven of clubs, is written as follows:

$$P(\text{2nd card is queen of hearts} \mid \text{1st card is seven of clubs}).$$

The vertical bar is read “given.”

### 3. THE MULTIPLICATION RULE

This section will show how to figure the chance that two events happen, by multiplying probabilities.

*Example 3.* A box has three tickets, colored red, white and blue.



Two tickets will be drawn at random without replacement. What is the chance of drawing the red ticket and then the white?

*Solution.* Imagine a large group of people. Each of these people holds a box R W B and draws two tickets at random without replacement. About one third of the people get R on the first draw, and are left with



On the second draw, about half of these people will get W. The fraction who draw R W is therefore

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

The chance is 1 in 6, or  $16\frac{2}{3}\%$ .

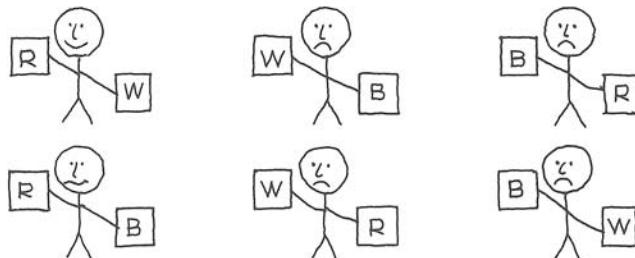
For instance, suppose you start with 600 people. About 200 of them will get R on the first draw. Of these 200 people, about 100 will get W on the second draw. So  $100/600 = 1/6$  of the people draw the red ticket first and then the white one. In figure 4, the people who draw R W are at the top left.

Statisticians usually multiply the chances in reverse order:

$$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}.$$

The reason:  $1/3$  refers to the first draw, and  $1/2$  to the second.

Figure 4. The multiplication rule. Each stick figure corresponds to 100 people.



The method in example 3 is called the multiplication rule.

*Multiplication Rule.* The chance that two things will both happen equals the chance that the first will happen, multiplied by the chance that the second will happen given the first has happened.

*Example 4.* Two cards will be dealt off the top of a well-shuffled deck. What is the chance that the first card will be the seven of clubs and the second card will be the queen of hearts?

*Solution.* This is like example 3, with a much bigger box. The chance that the first card will be the seven of clubs is  $1/52$ . Given that the first card was the seven of clubs, the chance that the second card will be the queen of hearts is  $1/51$ . The chance of getting both cards is

$$\frac{1}{52} \times \frac{1}{51} = \frac{1}{2,652}.$$

This is a small chance: about 4 in 10,000, or 0.04 of 1%.

*Example 5.* A deck of cards is shuffled, and two cards are dealt. What is the chance that both are aces?

*Solution.* The chance that the first card is an ace equals  $4/52$ . Given that the first card is an ace, there are 3 aces among the 51 remaining cards. So the chance of a second ace equals  $3/51$ . The chance that both cards are aces equals

$$\frac{4}{52} \times \frac{3}{51} = \frac{12}{2,652}.$$

This is about 1 in 200, or  $1/2$  of 1%.

*Example 6.* A coin is tossed twice. What is the chance of a head followed by a tail?

*Solution.* The chance of a head on the first toss equals  $1/2$ . No matter how the first toss turns out, the chance of tails on the second toss equals  $1/2$ . So the chance of heads followed by tails equals

$$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

### Exercise Set C

- A deck is shuffled and two cards are dealt.
  - Find the chance that the second card is a heart given the first card is a heart.
  - Find the chance that the first card is a heart and the second card is a heart.

2. A die is rolled three times.
  - (a) Find the chance that the first roll is an ace .
  - (b) Find the chance that the first roll is an ace , the second roll is a deuce , and the third roll is a trey .
3. A deck is shuffled and three cards are dealt.
  - (a) Find the chance that the first card is a king.
  - (b) Find the chance that the first card is a king, the second is a queen, and the third is a jack.
4. A die will be rolled six times. You have a choice—
  - (i) to win \$1 if at least one ace shows
  - (ii) to win \$1 if an ace shows on all the rolls

Which option offers the better chance of winning? Or are they the same? Explain.

5. Someone works example 2(a) on p. 226 this way:

For me to win, the queen can't be the first card dealt (51 chances in 52) and she must be the second card (1 chance in 51), so the answer is

$$\frac{51}{52} \times \frac{1}{51} = \frac{1}{52}.$$

Is the multiplication legitimate? Why?

6. “A cat-o’nine-\_\_\_\_\_ can be used to punish \_\_\_\_\_ of state, but this is seldom done.” A coin is tossed twice, to fill in the blanks. What is the chance of the coin getting it right?
7. A coin is tossed 3 times.
  - (a) What is the chance of getting 3 heads?
  - (b) What is the chance of not getting 3 heads?
  - (c) What is the chance of getting at least 1 tail?
  - (d) What is the chance of getting at least 1 head?

*The answers to these exercises are on p. A67.*

#### 4. INDEPENDENCE

This section introduces the idea of independence, which will be used many times in the rest of the book.

Two things are *independent* if the chances for the second given the first are the same, no matter how the first one turns out. Otherwise, the two things are *dependent*.

*Example 7.* Someone is going to toss a coin twice. If the coin lands heads on the second toss, you win a dollar.

- (a) If the first toss is heads, what is your chance of winning the dollar?
- (b) If the first toss is tails, what is your chance of winning the dollar?

(c) Are the tosses independent?

*Solution.* If the first toss is heads, there is a 50% chance to get heads the second time. If the first toss is tails, the chance is still 50%. The chances for the second toss stay the same, however the first toss turns out. That is independence.

*Example 8.* Two draws will be made at random with replacement from

1	1	2	2	3
---	---	---	---	---

- (a) Suppose the first draw is  $\boxed{1}$ . What is the chance of getting a  $\boxed{2}$  on the second draw?
- (b) Suppose the first draw is  $\boxed{2}$ . What is the chance of getting  $\boxed{2}$  on the second draw?
- (c) Are the draws independent?

*Solution.* Whether the first draw is  $\boxed{1}$  or  $\boxed{2}$  or anything else, the chance of getting  $\boxed{2}$  on the second draw stays the same—two in five, or 40%. The reason: the first ticket is replaced, so the second draw is always made from the same box  $\boxed{1 \ 1 \ 2 \ 2 \ 3}$ . The draws are independent.

*Example 9.* As in example 8, but the draws are made without replacement.

*Solution.* If the first draw turns out to be  $\boxed{1}$  then the second draw is from the box  $\boxed{1 \ 2 \ 2 \ 3}$ . The chance for the second draw to be  $\boxed{2}$  is 50%. On the other hand, if the first draw turns out to be  $\boxed{2}$ , then the second draw is from the box  $\boxed{1 \ 1 \ 2 \ 3}$ . Now there is only a 25% chance for the second to be  $\boxed{2}$ . The draws are dependent.

When drawing at random with replacement, the draws are independent. Without replacement, the draws are dependent.

What does independence of the draws mean? To answer this question, think about bets which can be settled on one draw: for instance, that the draw will be 3 or more. Then the conditional chance of winning the bet must stay the same, no matter how the other draws turn out.

*Example 10.* A box has three tickets, colored red, white, and blue.

R	W	B
---	---	---

Two tickets will be drawn at random with replacement. What is the chance of drawing the red ticket and then the white?

*Solution.* The draws are independent, so the chance is

$$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}.$$

Compare this with example 3. The answers are different. Independence matters. And it's easier this time: you don't need to work out conditional probabilities.

If two things are independent, the chance that both will happen equals the product of their unconditional probabilities. This is a special case of the multiplication rule (p. 229).

### Exercise Set D

1. For each of the following boxes, say whether color and number are dependent or independent.

(a) 

1	2	2	1	2	2
---	---	---	---	---	---

      (c) 

1	2	3	1	2	2
---	---	---	---	---	---

  
 (b) 

1	2	1	2	1	2
---	---	---	---	---	---

2. (a) In the box shown below, each ticket has two numbers.

1	2	1	3	4	2	4	3
---	---	---	---	---	---	---	---

(For instance, the first number on  $\boxed{4|2}$  is 4 and the second is 2.) A ticket is drawn at random. Are the two numbers dependent or independent?

- (b) Repeat, for the box

1	2	1	3	1	3	4	2	4	3	4	3
---	---	---	---	---	---	---	---	---	---	---	---

- (c) Repeat, for the box

1	2	1	3	1	3	4	2	4	2	4	3
---	---	---	---	---	---	---	---	---	---	---	---

3. Every week you buy a ticket in a lottery that offers one chance in a million of winning. What is the chance that you never win, even if you keep this up for ten years?
4. Two draws are made at random without replacement from the box  $\boxed{1 \ 2 \ 3 \ 4}$ . The first ticket is lost, and nobody knows what was written on it. True or false, and explain: the two draws are independent.
5. Suppose that in a certain class, there are
- 80% men and 20% women;
  - 15% freshmen and 85% sophomores.
- (a) The percentage of sophomore women in the class can be as small as \_\_\_\_\_.  
 (b) This percentage can be as large as \_\_\_\_\_.
6. One student is chosen at random from the class described in the previous exercise.
- (a) The chance of getting a sophomore woman can be as small as \_\_\_\_\_.  
 (b) This chance can be as large as \_\_\_\_\_.
7. In 2002, about 50.9% of the population of the United States was female. Also, 1.6% of the population was age 85 and over.<sup>3</sup> True or false, and explain: the percentage of the population consisting of women age 85 and over is

$$50.9\% \text{ of } 1.6\% = 0.509 \times 1.6\% \approx 0.8 \text{ of } 1\%$$

8. (Hard.) In a certain psychology experiment, each subject is presented with three ordinary playing cards, face down. The subject takes one of these cards. The subject also takes one card at random from a separate, full deck of playing cards. If the two cards are from the same suit, the subject wins a prize. What is the chance of winning? If more information is needed, explain what you need, and why.

*The answers to these exercises are on pp. A67–68.*

## 5. THE COLLINS CASE

*People v. Collins* is a law case in which there was a major statistical issue. A black man and a white woman were charged with robbery. The facts were described by the court as follows.<sup>4</sup>

On June 18, 1964, about 11:30 A.M. Mrs. Juanita Brooks, who had been shopping, was walking home along an alley in the San Pedro area of the City of Los Angeles. She was pulling behind her a wicker basket carryall containing groceries and had her purse on top of the packages. She was using a cane. As she stooped down to pick up an empty carton, she was suddenly pushed to the ground by a person whom she neither saw nor heard approach. She was stunned by the fall and felt some pain. She managed to look up and saw a young woman running from the scene. According to Mrs. Brooks the latter appeared to weigh about 145 pounds, was wearing “something dark,” and had hair “between a dark blond and a light blond,” but lighter than the color of defendant Janet Collins’ hair as it appeared at trial. Immediately after the incident, Mrs. Brooks discovered that her purse, containing between \$35 and \$40, was missing.

About the same time as the robbery, John Bass, who lived on the street at the end of the alley, was in front of his house watering his lawn. His attention was attracted by “a lot of crying and screaming” coming from the alley. As he looked in that direction, he saw a woman run out of the alley and enter a yellow automobile parked across the street from him. He was unable to give the make of the car. The car started off immediately and pulled wide around another parked vehicle so that in the narrow street it passed within six feet of Bass. The latter then saw that it was being driven by a male Negro, wearing a mustache and beard. At the trial Bass identified defendant as the driver of the yellow automobile. However, an attempt was made to impeach his identification by his admission that at the preliminary hearing he testified to an uncertain identification at the police lineup shortly after the attack on Mrs. Brooks, when defendant was beardless.

In his testimony Bass described the woman who ran from the alley as a Caucasian, slightly over five feet tall, of ordinary build, with her hair in a dark blond ponytail, and wearing dark clothing. He further testified that her ponytail was “just like” one which Janet had in a police photograph taken on June 22, 1964.

The prosecutor then had a mathematics instructor at a local state college explain the multiplication rule, without paying much attention to independence, or the distinction between conditional and unconditional probabilities. After this testimony, the prosecution assumed the following chances:

Yellow automobile	1/10	Woman with blond hair	1/3
Man with mustache	1/4	Black man with beard	1/10
Woman with ponytail	1/10	Interracial couple in car	1/1,000

When multiplied together, these come to 1 in 12,000,000. According to the prosecution, this procedure gave the chance “that any [other] couple possessed the distinctive characteristics of the defendants.” If no other couple possessed these characteristics, the defendants were guilty. The jury convicted. On appeal, the Supreme Court of California reversed the verdict. It found no evidence to support the assumed values for the six chances. Furthermore, these were presented as unconditional probabilities. The basis for multiplying them, as the mathematics instructor should have explained, was independence. And there was no evidence to support that assumption either. On the contrary, some factors were clearly dependent—like “Black man with beard” and “interracial couple in car.”

Blindly multiplying chances can make real trouble. Check for independence, or use conditional probabilities.

There is another objection to the prosecutor’s reasoning. Probability calculations like the multiplication rule were developed for dealing with games of chance, where the basic process can be repeated independently and under the same conditions. The prosecutor was trying to apply this theory to a unique event: something that either happened—or didn’t happen—on June 18, 1964, at 11:30 A.M. What does chance mean, in this new context? It was up to the prosecutor to answer this question, and to show that the theory applied to his situation.<sup>5</sup>

In the 1990s, DNA evidence began to be used for identification of criminals: the idea is to match a suspect’s DNA with DNA left at the scene of the crime—for instance, in bloodstains. Matching is done on a set of characteristics of DNA. The technical issues are similar to those raised by the Collins case: Can you estimate the fraction of the population with a given characteristic? Are those characteristics independent? Is the lab work accurate? Many experts believe that such questions have satisfactory answers, others are quite skeptical.<sup>6</sup>

## 6. REVIEW EXERCISES

*When a die is rolled, each of the six faces is equally likely to come up. A deck of cards has 4 suits (clubs, diamonds, hearts, spades) with 13 cards in each suit—2, 3, . . . , 10, jack, queen, king, ace. See pp. 222 and 226.*

- True or false, and explain:
  - If something has probability 1,000%, it is sure to happen.
  - If something has probability 90%, it can be expected to happen about nine times as often as its opposite.
- Two cards will be dealt off the top of a well-shuffled deck. You have a choice:
  - To win \$1 if the first is a king.

- (ii) To win \$1 if the first is a king and the second is a queen.  
 Which option is better? Or are they equivalent? Explain briefly.
3. Four cards will be dealt off the top of a well-shuffled deck. There are two options:
- To win \$1 if the first card is a club and the second is a diamond and the third is a heart and the fourth is a spade.
  - To win \$1 if the four cards are of four different suits.
- Which option is better? Or are they the same? Explain.
4. A poker hand is dealt. Find the chance that the first four cards are aces and the fifth is a king.
5. One ticket will be drawn at random from the box below. Are color and number independent? Explain.
- |   |   |   |   |   |   |
|---|---|---|---|---|---|
| 1 | 1 | 8 | 1 | 1 | 8 |
|---|---|---|---|---|---|
6. A deck of cards is shuffled and the top two cards are placed face down on a table. True or false, and explain:
- There is 1 chance in 52 for the first card to be the ace of clubs.
  - There is 1 chance in 52 for the second card to be the ace of diamonds.
  - The chance of getting the ace of clubs and then the ace of diamonds is  $1/52 \times 1/52$ .
7. A coin is tossed six times. Two possible sequences of results are
- H T T H T H
  - H H H H H H
- (The coin must land H or T in the order given; H = heads, T = tails.) Which of the following is correct? Explain.<sup>7</sup>
- Sequence (i) is more likely.
  - Sequence (ii) is more likely.
  - Both sequences are equally likely.
8. A die is rolled four times. What is the chance that—
- all the rolls show 3 or more spots?
  - none of the rolls show 3 or more spots?
  - not all the rolls show 3 or more spots?
9. A die is rolled 10 times. Find the chance of—
- getting 10 sixes.
  - not getting 10 sixes.
  - all the rolls showing 5 spots or less.
10. Which of the two options is better, or are they the same? Explain briefly.
- You toss a coin 100 times. On each toss, if the coin lands heads, you win \$1. If it lands tails, you lose \$1.
  - You draw 100 times at random with replacement from | 1 0 |. On each draw, you are paid (in dollars) the number on the ticket.

11. In the box shown below, each ticket should have two numbers:

<input type="text"/>								
----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------

A ticket will be drawn at random. Can you fill in the blanks so the two numbers are independent?

12. You are thinking about playing a lottery. The rules: you buy a ticket, choose 3 different numbers from 1 to 100, and write them on the ticket. The lottery has a box with 100 balls numbered from 1 through 100. Three balls are drawn at random without replacement. If the numbers on these balls are the same as the numbers on your ticket, you win. (Order doesn't matter.) If you decide to play, what is your chance of winning?

## 7. SUMMARY

1. The *frequency theory* of chance applies most directly to chance processes which can be repeated over and over again, independently and under the same conditions.
2. The chance of something gives the percentage of times the thing is expected to happen, when the basic process is repeated over and over again.
3. Chances are between 0% and 100%. Impossibility is represented by 0%, certainty by 100%.
4. The chance of something equals 100% minus the chance of the opposite thing.
5. The chance that two things will both happen equals the chance that the first will happen, multiplied by the *conditional* chance that the second will happen given that the first has happened. This is the *multiplication rule*.
6. Two things are *independent* if the chances for the second one stay the same no matter how the first one turns out.
7. If two things are independent, the chance that both will happen equals the product of their unconditional chances. This is a special case of the multiplication rule.
8. When you draw at random, all the tickets in the box have the same chance to be picked. Draws made at random with replacement are independent. Without replacement, the draws are dependent.
9. Blindly multiplying chances can make real trouble. Check for independence, or use conditional chances.
10. The mathematical theory of chance only applies in some situations. Using it elsewhere can lead to ridiculous results.

# 14

## More about Chance

*Some of the Problems about Chance having a great appearance of Simplicity, the Mind is easily drawn into a belief, that their Solution may be attained by the meer Strength of natural good Sense; which generally proving otherwise and the Mistakes occasioned thereby being not unfrequent, 'tis presumed that a Book of this Kind, which teaches to distinguish Truth from what seems so nearly to resemble it, will be looked upon as a help to good Reasoning.*

—ABRAHAM DE MOIVRE (ENGLAND, 1667–1754)<sup>1</sup>

### 1. LISTING THE WAYS

A *probabilist* is a mathematician who specializes in computing the probabilities of complex events. In the twentieth century, two of the leading probabilists were A. N. Kolmogorov (Russia, 1903–1987) and P. Lévy (France, 1886–1971). The techniques they developed are beyond the scope of this book, but we can look at more basic methods, developed by earlier mathematicians.

When trying to figure chances, it is sometimes very helpful to list all the possible ways that a chance process can turn out. If this is too hard, writing down a few typical ones is a good start.

*Example 1.* Two dice are thrown. What is the chance of getting a total of 2 spots?

*Solution.* The chance process here consists of throwing the two dice. What matters is the number of spots shown by each die. To keep the dice separate, imagine that one is white and the other black. One way for the dice to fall is



This means the white die showed 2 spots, and the black die showed 3. The total number of spots is 5.

How many ways are there for the two dice to fall? To begin with, the white die can fall in any one of 6 ways:



When the white die shows  $\square$ , say, there are still 6 possible ways for the black die to fall:



We now have 6 of the possible ways that the two dice can fall. These ways are shown in the first row of figure 1. Similarly, the second row shows another 6 ways for the dice to fall, with the white die showing  $\square$ . And so on. The figure shows there are  $6 \times 6 = 36$  possible ways for the dice to fall. They are all equally likely, so each has 1 chance in 36. There is only one way to get a total of 2 spots:  $\square \square$ . The chance is 1/36. That is the answer.

There may be several methods for answering questions about chance. In figure 1, for example, the chance for each of the 36 outcomes can also be worked out using the multiplication rule:  $1/6 \times 1/6 = 1/36$ .

*Example 2.* A pair of dice are thrown. What is the chance of getting a total of 4 spots?

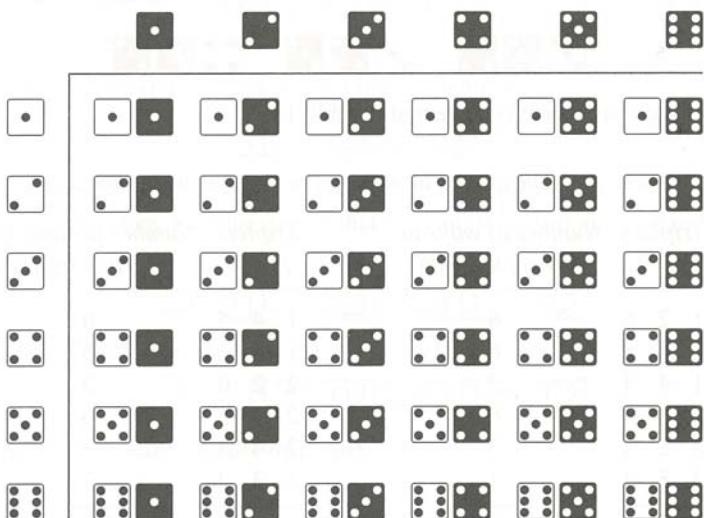
*Solution.* Look at figure 1. There are 3 ways to get a total of four spots:



The chance is 3 in 36. That is the answer.

What about three dice? A three-dimensional picture like figure 1 would be a bit much to absorb, but similar reasoning can be used. In the seventeenth century, Italian gamblers used to bet on the total number of spots rolled with three dice. They believed that the chance of rolling a total of 9 ought to equal the chance of

Figure 1. Throwing a pair of dice. There are 36 ways for the dice to fall, shown in the body of the diagram; all are equally likely.



rolling a total of 10. For instance, they said, one combination with a total of 9 spots is

1 spot on one die, 2 spots on another die, 6 spots on the third die.

This can be abbreviated as "1 2 6." There are altogether six combinations for 9:

1 2 6    1 3 5    1 4 4    2 3 4    2 2 5    3 3 3

Similarly, they found six combinations for 10:

1 4 5    1 3 6    2 2 6    2 3 5    2 4 4    3 3 4

Thus, argued the gamblers, 9 and 10 should by rights have the same chance. However, experience showed that 10 came up a bit more often than 9.

They asked Galileo for help, and he reasoned as follows. Color one of the dice white, another one grey, and another one black—so they can be kept apart. This won't affect the chances. How many ways can the three dice fall? The white die can land in 6 ways. Corresponding to each of them, the grey die can land in 6 ways, making  $6 \times 6$  possibilities. Corresponding to each of these possibilities, there are still 6 for the black die. Altogether, there are  $6 \times 6 \times 6 = 6^3$  ways for three dice to land. (With 4 dice, there would be  $6^4$ ; with 5 dice,  $6^5$  and so on.)

Now  $6^3 = 216$  is a lot of ways for three dice to fall. But Galileo sat down and listed them. Then he went through his list and counted the ones with a total of 9 spots. He found 25. And he found 27 ways to get a total of 10 spots. He concluded that the chance of rolling 9 is  $25/216 \approx 11.6\%$ , while the chance of rolling 10 is  $27/216 = 12.5\%$ .

The gamblers made a basic error: they didn't get down to the different ways for the dice to land. For instance, the triplet 3 3 3 for 9 only corresponds to one way for the dice to land:



But the triplet 3 3 4 for 10 corresponds to three ways for the dice to land:



The gamblers' argument is corrected in table 1.

Table 1. The chance of getting 9 or 10 spots with three dice.

<i>Triplets for 9</i>	<i>Number of ways to roll each triplet</i>	<i>Triplets for 10</i>	<i>Number of ways to roll each triplet</i>
1 2 6	6	1 4 5	6
1 3 5	6	1 3 6	6
1 4 4	3	2 2 6	3
2 3 4	6	2 3 5	6
2 2 5	3	2 4 4	3
3 3 3	1	3 3 4	3
Total	25	Total	27



Galileo (Italy, 1564–1642)

Wolff-Leavenworth Collection, courtesy of the Syracuse University Art Collection.

### Exercise Set A

1. Look at figure 1 and make a list of the ways to roll a total of 5 spots. What is the chance of throwing a total of 5 spots with two dice?
2. Two draws are made at random with replacement from the box 1 2 3 4 5.

- Draw a picture like figure 1 to represent all possible results. How many are there? What is the chance that the sum of the two draws turns out to equal 6?
3. A pair of dice is thrown 1,000 times. What total should appear most often? What totals should appear least often?
  4. (a) In the box shown below, each ticket has two numbers.

<span style="border: 1px solid black; padding: 2px;">1   2</span>	<span style="border: 1px solid black; padding: 2px;">1   3</span>	<span style="border: 1px solid black; padding: 2px;">3   1</span>	<span style="border: 1px solid black; padding: 2px;">3   2</span>
---	---	---	---

(For instance, the first number on 3 | 1 is 3 and the second is 1.) A ticket is drawn at random. Find the chance that the sum of the two numbers is 4.

- (b) Repeat, for the box

<span style="border: 1px solid black; padding: 2px;">1   2</span>	<span style="border: 1px solid black; padding: 2px;">1   3</span>	<span style="border: 1px solid black; padding: 2px;">1   3</span>	<span style="border: 1px solid black; padding: 2px;">3   2</span>	<span style="border: 1px solid black; padding: 2px;">3   3</span>	<span style="border: 1px solid black; padding: 2px;">3   3</span>
---	---	---	---	---	---

- (c) Repeat, for the box

<span style="border: 1px solid black; padding: 2px;">1   2</span>	<span style="border: 1px solid black; padding: 2px;">1   3</span>	<span style="border: 1px solid black; padding: 2px;">1   3</span>	<span style="border: 1px solid black; padding: 2px;">3   1</span>	<span style="border: 1px solid black; padding: 2px;">3   2</span>	<span style="border: 1px solid black; padding: 2px;">3   3</span>
---	---	---	---	---	---

The answers to these exercises are on p. A68.

## 2. THE ADDITION RULE

This section is about the chance that at least one of two specified things will happen: either the first happens, or the second, or both. The possibility of both happening turns out to be a complication, which can sometimes be ruled out.

Two things are *mutually exclusive* when the occurrence of one prevents the occurrence of the other: one excludes the other.

*Example 3.* A card is dealt off the top of a well-shuffled deck. The card might be a heart. Or, it might be a spade. Are these two possibilities mutually exclusive?

*Solution.* If the card is a heart, it can't be a spade. These two possibilities are mutually exclusive.

We can now state a general principle for figuring chances. It is called the addition rule.

*Addition Rule.* To find the chance that at least one of two things will happen, check to see if they are mutually exclusive. If they are, add the chances.

*Example 4.* A card is dealt off the top of a well-shuffled deck. There is 1 chance in 4 for it to be a heart. There is 1 chance in 4 for it to be a spade. What is the chance for it to be in a major suit (hearts or spades)?

*Solution.* The question asks for the chance that one of the following two things will happen:

- the card is a heart;
- the card is a spade.

As in example 3, if the card is a heart then it can't be a spade: these are mutually exclusive events. So it is legitimate to add the chances. The chance of getting a card in a major suit is  $1/4 + 1/4 = 1/2$ . (A check on the reasoning: there are 13 hearts and 13 spades, so  $26/52 = 1/2$  of the cards in the deck are in a major suit.)

*Example 5.* Someone throws a pair of dice. True or false: the chance of getting at least one ace is  $1/6 + 1/6 = 1/3$ .

*Solution.* This is false. Imagine one of the dice is white, the other black.



The question asks for the chance that one of the following two things will happen:

- the white die lands ace  $\square$ ;
- the black die lands ace  $\square$ .

A white ace does not prevent a black ace. These two events are not mutually exclusive, so the addition rule does not apply. Adding the chances gives the wrong answer.

Look at figure 1. There are 6 ways for the white die to show  $\square$ . There are 6 ways for the black die to show  $\square$ . But the number of ways to get at least one ace is not  $6 + 6$ . Addition double counts the outcome  $\square \square$  at the top left corner. The chance of getting at least one ace is

$$(6 + 6 - 1)/36 = 11/36, \text{ not } (6 + 6)/36 = 12/36 = 1/3.$$

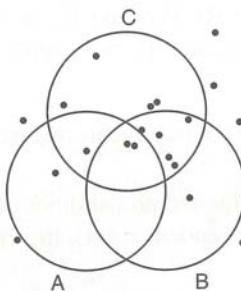
If you want to find the chance that at least one event occurs, and the events are not mutually exclusive, do not add the chances: the sum will be too big.

Blindly adding chances can give the wrong answer, by double counting the chance that two things happen. With mutually exclusive events, there is no double counting: that is why the addition rule works.

### Exercise Set B

1. Fifty children went to a party where cookies and ice cream were served: 12 children took cookies; 17 took ice cream. True or false: 29 children must have had cookies or ice cream. Explain briefly.

2. There are 20 dots in the diagram below, and 3 circles. The circles are labeled A, B, and C. One of the dots will be chosen at random.
- What is the probability that the dot falls inside circle A?
  - What is the probability that the dot falls inside circle B?
  - What is the probability that the dot falls inside circle C?
  - What is the probability that the dot falls inside at least one of the circles?



3. Two cards are dealt off the top of a well-shuffled deck. You have a choice:
- to win \$1 if the first card is an ace or the second card is an ace;
  - to win \$1 if at least one of the two cards is an ace.
- Which option is better? or are they the same? Explain briefly.
4. Two dice will be rolled. The chance that the first one lands  $\square$  is  $1/6$ . The chance that the second one lands  $\square$  is  $1/6$ . True or false: the chance that the first one lands  $\square$  or the second one lands  $\square$  equals  $1/6 + 1/6$ . Explain briefly.
5. A box contains 10 tickets numbered 1 through 10. Five draws will be made at random with replacement from this box. True or false: there are 5 chances in 10 of getting  $\square$  at least once. Explain briefly.
6. A number is drawn at random from a box. There is a 20% chance for it to be 10 or less. There is a 10% chance for it to be 50 or more. True or false: the chance of getting a number between 10 and 50 (exclusive) is 70%. Explain briefly.

*The answers to these exercises are on pp. A68–69.*

### 3. TWO FAQs (FREQUENTLY ASKED QUESTIONS)

- What's the difference between mutually exclusive and independent?
- When do I add and when do I multiply?

“Mutually exclusive” is one idea; independence is another. Both ideas apply to pairs of events, and say something about how the events are related. However, the relationships are quite different.

- Two events are mutually exclusive if the occurrence of one prevents the other from happening.
- Two events are independent if the occurrence of one does not change the chances for the other.

The addition rule, like the multiplication rule, is a way of combining chances. However, the two rules solve different problems (pp. 229 and 241).

- The addition rule finds the chance that at least one of two things happens.
- The multiplication rule finds the chance that two things both happen.

So, the first step in deciding whether to add or to multiply is to read the question: Do you want to know  $P(A \text{ or } B)$ ,  $P(A \text{ and } B)$ , or something else entirely? But there is also a second step—because the rules apply only if the events are related in the right way.

- Adding the probabilities of two events requires them to be mutually exclusive.<sup>2</sup>
- Multiplying the unconditional probabilities of two events requires them to be independent. (For dependent events, the multiplication rule uses conditional probabilities.)

*Example 6.* A die is rolled 6 times; a deck of cards is shuffled.

- The chance that the first roll is an ace or the last roll is an ace equals \_\_\_\_\_.
- The chance that the first roll is an ace and the last roll is an ace equals \_\_\_\_\_.
- The chance that the top card is the ace of spades or the bottom card is the ace of spades equals \_\_\_\_\_.
- The chance that the top card is the ace of spades and the bottom card is the ace of spades equals \_\_\_\_\_.

Options for parts (a) and (b):

(i)  $\frac{1}{6} + \frac{1}{6}$       (ii)  $\frac{1}{6} \times \frac{1}{6}$       (iii) neither of these

Options for parts (c) and (d):

(i)  $\frac{1}{52} + \frac{1}{52}$       (ii)  $\frac{1}{52} \times \frac{1}{52}$       (iii) neither of these

*Solution. Part (a).* You want the chance that at least one of the two things will happen, so the addition rule looks relevant. However, the two things are not mutually exclusive. Do not use the addition rule, it will give the wrong answer (example 5). If you can't add, maybe you can multiply? The two events are independent, but you do not want the chance that both happen. Do not use the multiplication rule either, it too will give the wrong answer. Choose option (iii).

*Part (b).* You want the chance that both events happen, and they are independent. Now is the time to multiply. Choose option (ii).



*Part (c).* The chance the top card is the ace of spades equals  $1/52$ . The chance that the bottom card is the ace of spades—computed before looking at any of the cards (example 2 on p. 226) also equals  $1/52$ . The two events are mutually exclusive; you want the chance that at least one of the two will occur. This is when the addition rule shines. Choose (i).

*Part (d).* The two events are mutually exclusive, but you do not want the chance that at least one of the two will occur. Therefore, do not use the addition rule, it will give the wrong answer. You want the chance that both things happen, so multiplication may be relevant. However, the events are dependent. Do not multiply the unconditional probabilities, you will get the wrong answer. Choose (iii). (The chance is 0: the ace of spades cannot turn up in both places.)

As example 6 indicates, you may not be able either to add or to multiply. Then more thinking is needed. (The cartoon is trying to tell you something.) The next section gives an example—The Paradox of the Chevalier de Méré.

*Technical notes.* The chance of two aces is  $1/36$ , so the chance in example 6(a) can be figured as

$$\frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}$$

However, if the die is rolled 3 times, the chance of getting at least one ace is not

$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} - \left(\frac{1}{6}\right)^3$$

Think about 12 rolls! This sort of problem will be solved in the next section.

In example 6(d), the multiplication rule can be used, with conditional probabilities—although this is very fussy. The chance that the top card is the ace of spades equals  $1/52$ . Given that the top card is the ace of spades, the conditional chance that the bottom card is the ace of spades equals 0. The chance that both things happen equals  $1/52 \times 0 = 0$ .

### Exercise Set C

1. A large group of people are competing for all-expense-paid weekends in Philadelphia. The Master of Ceremonies gives each contestant a well-shuffled deck of cards. The contestant deals two cards off the top of the deck, and wins a weekend in Philadelphia if the first card is the ace of hearts or the second card is the king of hearts.
  - (a) All the contestants whose first card was the ace of hearts are asked to step forward. What fraction of the contestants do so?
  - (b) The contestants return to their original places. Then, the ones who got the king of hearts for their second card are asked to step forward. What fraction of the contestants do so?
  - (c) Do any of the contestants step forward twice?
  - (d) True or false, and explain: the chance of winning a weekend in Philadelphia is  $1/52 + 1/52$ .



"PHILADELPHIA, PLEASE."

2. A large group of people are competing for all-expense-paid weekends in Philadelphia. The Master of Ceremonies gives each contestant a well-shuffled deck of cards. The contestant deals two cards off the top of the deck, and wins a weekend in Philadelphia if the first card is the ace of hearts or the second card is the ace of hearts. (This is like exercise 1, but the winning cards are a little different.)
- All the contestants whose first card was the ace of hearts are asked to step forward. What fraction of the contestants do so?
  - The contestants return to their original places. Then, the ones who got the ace of hearts for their second card are asked to step forward. What fraction of the contestants do so?
  - Do any of the contestants step forward twice?
  - True or false, and explain: the chance of winning a weekend in Philadelphia is  $1/52 + 1/52$ .
3. A deck of cards is shuffled. True or false, and explain briefly:
- The chance that the top card is the jack of clubs equals  $1/52$ .
  - The chance that the bottom card is the jack of diamonds equals  $1/52$ .
  - The chance that the top card is the jack of clubs or the bottom card is the jack of diamonds equals  $2/52$ .
  - The chance that the top card is the jack of clubs or the bottom card is the jack of clubs equals  $2/52$ .
  - The chance that the top card is the jack of clubs and the bottom card is the jack of diamonds equals  $1/52 \times 1/52$ .
  - The chance that the top card is the jack of clubs and the bottom card is the jack of clubs equals  $1/52 \times 1/52$ .
4. The unconditional probability of event A is  $1/2$ . The unconditional probability of event B is  $1/3$ . Say whether each of the following is true or false, and explain briefly.
- The chance that A and B both happen must be  $1/2 \times 1/3 = 1/6$ .
  - If A and B are independent, the chance that they both happen must be  $1/2 \times 1/3 = 1/6$ .
  - If A and B are mutually exclusive, the chance that they both happen must be  $1/2 \times 1/3 = 1/6$ .
  - The chance that at least one of A or B happens must be  $1/2 + 1/3 = 5/6$ .
  - If A and B are independent, the chance that at least one of them happens must be  $1/2 + 1/3 = 5/6$ .
  - If A and B are mutually exclusive, the chance that at least one of them happens must be  $1/2 + 1/3 = 5/6$ .
5. Two cards are dealt off the top of a well-shuffled deck.
- Find the chance that the second card is an ace.
  - Find the chance that the second card is an ace, given the first card is a king.
  - Find the chance that the first card is a king and the second card is an ace.

*The answers to these exercises are on pp. A69–70.*

#### 4. THE PARADOX OF THE CHEVALIER DE MÉRÉ

In the seventeenth century, French gamblers used to bet on the event that with 4 rolls of a die, at least one ace would turn up: an ace is  $\square \bullet$ . In another game, they bet on the event that with 24 rolls of a pair of dice, at least one double-ace would turn up: a double-ace is a pair of dice which show  $\square \square$ .

The Chevalier de Mérém, a French nobleman of the period, thought the two events were equally likely. He reasoned this way about the first game:

- In one roll of a die, I have  $1/6$  of a chance to get an ace.
- So in 4 rolls, I have  $4 \times 1/6 = 2/3$  of a chance to get at least one ace.

His reasoning for the second game was similar:

- In one roll of a pair of dice, I have  $1/36$  of a chance to get a double-ace.
- So in 24 rolls, I must have  $24 \times 1/36 = 2/3$  of a chance to get at least one double-ace.

By this argument, both chances were the same, namely  $2/3$ . However, the gamblers found that the first event was a bit more likely than the second. This contradiction became known as the *Paradox of the Chevalier de Mérém*.

De Mérém asked the philosopher Blaise Pascal about the problem, and Pascal solved it with the help of his friend, Pierre de Fermat. Fermat was a judge and a member of parliament, who is remembered today for the mathematical research he did after hours. Fermat saw that de Mérém was adding chances for events that were not mutually exclusive. In fact, pushing de Mérém's argument a little further, it shows the chance of getting an ace in 6 rolls of a die to be  $6/6$ , or 100%. Something had to be wrong.

The question is how to calculate the chances correctly. Pascal and Fermat solved this problem, with a typically indirect piece of mathematical reasoning—



Blaise Pascal (France, 1623–1662)

Wolff-Leavenworth Collection, courtesy of the  
Syracuse University Art Collection.



Pierre de Fermat (France, 1601–1665)

From the *Oeuvres Complètes*

the kind that always leaves non-mathematicians feeling a bit cheated. Of course, a direct attack like Galileo's (section 1) could easily bog down. With 4 rolls of a die, there are  $6^4 = 1,296$  outcomes to worry about. With 24 rolls of a pair of dice, there are  $36^{24} \approx 2.2 \times 10^{37}$  outcomes.

The conversation between Pascal and Fermat is lost to history, but here is a reconstruction.<sup>3</sup>

*Pascal.* Let's look at the first game first.

*Fermat.* Bon. The chance of winning is hard to compute, so let's work out the chance of the opposite event—losing. Then

$$\text{chance of winning} = 100\% - \text{chance of losing}.$$

*Pascal.* D'accord. The gambler loses when none of the four rolls shows an ace. But how do you work out the chances?

*Fermat.* It does look complicated. Let's start with one roll. What's the chance that the first roll doesn't show an ace?

*Pascal.* It has to show something from 2 through 6, so the chance is  $5/6$ .

*Fermat.* C'est ça. Now, what's the chance that the first two rolls don't show aces?

*Pascal.* We can use the multiplication rule. The chance that the first roll doesn't give an ace and the second doesn't give an ace equals  $5/6 \times 5/6 = (5/6)^2$ . After all, the rolls are independent, n'est-ce pas?

*Fermat.* What about 3 rolls?

*Pascal.* It looks like  $5/6 \times 5/6 \times 5/6 = (5/6)^3$ .

*Fermat.* Oui. Now what about 4 rolls?

*Pascal.* Must be  $(5/6)^4$ .

*Fermat.* Sans doute, and that's about 0.482, or 48.2%.

*Pascal.* So there is a 48.2% chance of losing. Now

$$\begin{aligned}\text{chance of winning} &= 100\% - \text{chance of losing} \\ &= 100\% - 48.2\% = 51.8\%.\end{aligned}$$

*Fermat* That settles the first game. The chance of winning is a little over 50%. Now what about the second?

*Pascal* Eh bien, in one roll of a pair of dice, there is 1 chance in 36 of getting a double-ace, and 35 chances in 36 of not getting a double-ace. By the multiplication rule, in 24 rolls of a pair of dice the chance of getting no double-aces must be

$$(35/36)^{24}.$$

*Fermat* Entendu. That's about 50.9%. So we have the chance of losing. Now

$$\begin{aligned}\text{chance of winning} &= 100\% - \text{chance of losing} \\ &= 100\% - 50.9\% = 49.1\%.\end{aligned}$$

*Pascal* Le résultat is a bit less than 50%. Voilà. That's why you win the second game a bit less frequently than the first. But you have to roll a lot of dice to see the difference.

If the chance of an event is hard to find, try to find the chance of the opposite event. Then subtract from 100%. (See p. 223.) This is useful when the chance of the opposite event is easier to compute.

### Exercise Set D

1. A die is rolled three times. You bet \$1 on some proposition. Below is a list of 6 bets, and then a list of 3 outcomes. For each bet, find all the outcomes where you win. For instance, with (a), you win on (i) only.

*Bets*

- (a) all aces
- (b) at least one ace
- (c) no aces
- (d) not all aces
- (e) 1st roll is an ace, or 2nd roll is an ace, or 3rd roll is an ace
- (f) 1st roll is an ace, and 2nd roll is an ace, and 3rd roll is an ace

*Outcomes*

- (i)  (ii)  (iii) 

2. In exercise 1, which is a better bet—(a) or (f)? Or are they same? What about (b) and (e)? What about (c) and (d)? (You do not need to compute the chances.)
3. A box contains four tickets, one marked with a star, and the other three blank:



Two draws are made at random with replacement from this box.

- (a) What is the chance of getting a blank ticket on the first draw?
- (b) What is the chance of getting a blank ticket on the second draw?
- (c) What is the chance of getting a blank ticket on the first draw and a blank ticket on the second draw?
- (d) What is the chance of not getting the star in the two draws?
- (e) What is the chance of getting the star at least once in the two draws?
4. (a) A die is rolled 3 times. What is the chance of getting at least one ace?
- (b) Same, with 6 rolls.
- (c) Same, with 12 rolls.
5. A pair of dice is rolled 36 times. What is the chance of getting at least one double-ace?
6. According to de Moivre, in eighteenth-century England people played a game similar to modern roulette. It was called “Royal Oak.” There were 32 “points” or num-

bered pockets on a table. A ball was thrown in such a way that it landed in each pocket with an equal chance, 1 in 32.

If you bet 1 pound on a point and it came up, you got your stake back, together with winnings of 27 pounds. If your point didn't come up, you lost your pound. The players (or "Adventurers," as de Moivre called them) complained that the game was unfair, and they should have won 31 pounds if their point came up. (They were right; section 1 of chapter 17.) De Moivre continues:

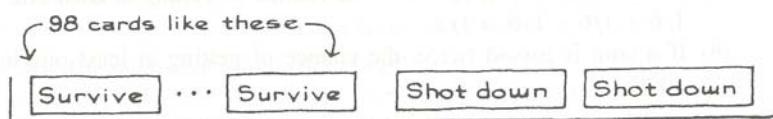
The Master of the Ball maintained they had no reason to complain; since he would undertake that any particular point of the Ball should come up in Two-and-Twenty Throws: of this he would offer to lay a Wager, and actually laid it when required. The seeming contradiction between the Odds of One-and-Thirty to One, and Twenty-two Throws for any [point] to come up, so perplexed the Adventurers, that they begun to think the Advantage was on their side: for which reason they played on and continued to lose. [Two-and-Twenty is 22, One-and-Thirty is 31.]

What is the chance that the point 17, say, will come up in Two-and-Twenty Throws? (The Master of the Ball laid this wager at even money, so if the chance is over 50%, he shows a profit here too.)



7. In his novel *Bomber*, Len Deighton argues that a World War II pilot had a 2% chance of being shot down on each mission. So in 50 missions he is "mathematically certain" to be shot down:  $50 \times 2\% = 100\%$ . Is this a good argument?

*Hint:* To make chance calculations, you have to see how the situation is like a game of chance. The analogy here is getting the card "survive" every time, if you draw 50 times at random with replacement from the box



*The answers to these exercises are on p. A70.*

## 5. ARE REAL DICE FAIR?

According to Galileo (section 1), when a die is rolled it is equally likely to show any of its 6 faces. Galileo was thinking of an ideal die which is perfectly symmetric. This is like ignoring friction in the study of physics: the results are only a first approximation. What does Galileo's calculation say about real dice?

- For real dice, the 216 possible ways three dice can land are close to being equally likely.
- If these ways were equally likely, the chance of rolling a total of 9 spots would be exactly 25 in 216.
- So for real dice, the chance of rolling a total of 9 spots is just about 25 in 216.

For loaded dice, the calculations would be badly off. But ordinary dice, coins, and the like are very close to fair—in the sense that all the outcomes are equally likely. Of course, you have to put some effort into shaking the dice or flipping the coins. And the games of chance based on these fair mechanisms may be quite unfair (chapter 17).

In a similar way, if you are told that a ticket is drawn at random, you should assume that each ticket in the box is equally likely to be drawn. If the tickets are close to the same size, shape, and texture, and the box is well shaken, this is quite a reasonable approximation.

## 6. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

*When a die is rolled, each of the 6 faces is equally likely to come up. A deck of cards has 4 suits (clubs, diamonds, hearts, spades) with 13 cards in each suit—2, 3, . . . , 10, jack, queen, king, ace. See pp. 222 and 226.*

1. A pair of dice are thrown.
  - (a) Find the chance that both dice show 3 spots.
  - (b) Find the chance that both dice show the same number of spots.
2. In the game of Monopoly, a player rolls two dice, counts the total number of spots, and moves that many squares. Find the chance that the player moves 11 squares (no more and no less).
3. True or false, and explain:
  - (a) If a die is rolled three times, the chance of getting at least one ace is  $1/6 + 1/6 + 1/6 = 1/2$ .
  - (b) If a coin is tossed twice, the chance of getting at least one head is 100%.
4. Two cards will be dealt off the top of a well-shuffled deck. You have a choice:

- (i) to win \$1 if at least one of the two cards is a queen.
- (ii) to win \$1 if the first is a queen.

Which option is better? Or are they equivalent? Explain.

5. The chance of A is  $1/3$ ; the chance of B is  $1/10$ . True or false, and explain:
  - (a) If A and B are independent, they must also be mutually exclusive.
  - (b) If A and B are mutually exclusive, they cannot be independent.
6. One event has chance  $1/2$ , another has chance  $1/3$ . Fill in the blanks, using one phrase from each pair below, to make up two true sentences. Write out both sentences.  
 “If you want to find the chance that (i) will happen, check to see if they are (ii). If so, you can (iii) the chances.”
  - (i) at least one of the two events, both events
  - (ii) independent, mutually exclusive
  - (iii) add, multiply
7. Four draws are going to be made at random with replacement from the box 

1	2	2	3	3
---	---	---	---	---

. Find the chance that 2 is drawn at least once.
8. Repeat exercise 7, if the draws are made at random without replacement.
9. One ticket will be drawn at random from each of the two boxes shown below:

(A) 

1	2	3
---	---	---

(B) 

1	2	3	4
---	---	---	---

Find the chance that:

- (a) The number drawn from A is larger than the one from B.
- (b) The number drawn from A equals the one from B.
- (c) The number drawn from A is smaller than the one from B.

10. There are two options:
  - (i) A die will be rolled 60 times. Each time it shows an ace or a six, you win \$1; on the other rolls, you win nothing.
  - (ii) Sixty draws will be made at random with replacement from the box 

1	1	1	0	0	0
---	---	---	---	---	---

. On each draw, you will be paid the amount shown on the ticket, in dollars.

Which option is better? or are they the same? Explain briefly.

11. Three cards are dealt from a well-shuffled deck.
  - (a) Find the chance that all of the cards are diamonds.
  - (b) Find the chance that none of the cards are diamonds.
  - (c) Find the chance that the cards are not all diamonds.
12. A coin is tossed 10 times. True or false, and explain:
  - (a) The chance of getting 10 heads in a row is  $1/1,024$ .
  - (b) Given that the first 9 tosses were heads, the chance of getting 10 heads in a row is  $1/2$ .

*Exercises 13 and 14 are more difficult.*

13. A box contains 2 red marbles and 98 blue ones. Draws are made at random with replacement. In \_\_\_\_\_ draws from the box, there is better than a 50% chance for a red marble to appear at least once. Fill in the blank with the smallest number that makes the statement true. (You will need a calculator.)
14. In Lotto 6-53, there is a box with 53 balls, numbered from 1 to 53. Six balls are drawn at random without replacement from the box. You win the grand prize if the numbers on your lottery ticket are the same as the numbers on the six balls; order does not matter.

Person A bought two tickets, with the following numbers:

Ticket #1	5	12	21	30	42	51
Ticket #2	5	12	23	30	42	49

Person B bought two tickets, with the following numbers:

Ticket #1	7	11	25	28	34	50
Ticket #2	9	14	20	22	37	45

Which person has the better chance of winning? Or are their chances the same? Explain briefly.

## 7. SUMMARY

1. When figuring chances, one helpful strategy is to write down a complete list of all the possible ways that the chance process can turn out. If this is too hard, at least write down a few typical ways, and count how many ways there are in total.
2. The chance that at least one of two things will happen equals the sum of the individual chances, provided the things are mutually exclusive. Otherwise, adding the chances will give the wrong answer—double counting.
3. If you are having trouble working out the chance of an event, try to figure out the chance of its opposite; then subtract from 100%.

# 15

## The Binomial Formula

*Man is a reed, but a reed that thinks.*

—BLAISE PASCAL (FRANCE, 1623–1662)

### 1. INTRODUCTION

This chapter explains how to answer questions like the following.

- A coin is tossed four times. What is the chance of getting exactly one head?
- A die is rolled ten times. What is the chance of getting exactly three aces?
- A box contains one red marble and nine green ones. Five draws are made at random with replacement. What is the chance that exactly two draws will be red?

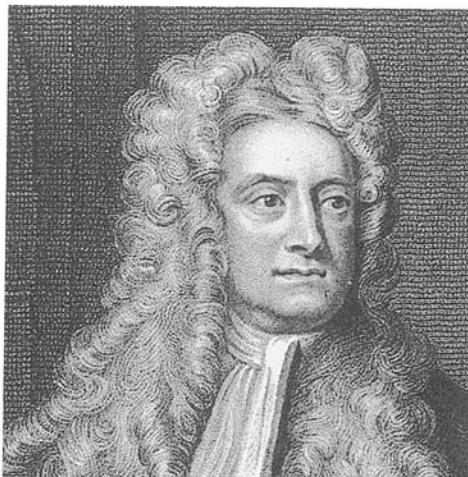
These problems are all similar, and can be solved using the *binomial coefficients*, discovered by Pascal and Newton.<sup>1</sup> The method will be illustrated on the marbles.

The problem is to find the chance of getting two reds (no more and no less) in five draws from the box; so the other three draws must be green. One way this can happen is that the first two draws are red and the final three are green. With R for red and G for green, this possibility can be written

R R G G G

Of course, there are many other ways to get two reds. For example, the second and the fifth draws might be red, while all the rest are green:

G R G G R



Isaac Newton (England, 1642–1727).

From the Warden Collection; engraved by W.T. Fry after a painting by G. Kneller.

To solve the problem, we must find all the possible ways, calculate the chance of each, and then use the addition rule to add up the chances. The first task seems formidable, so we postpone it for a moment and turn to the second one.

The chance of the pattern R R G G G is

$$\frac{1}{10} \times \frac{1}{10} \times \frac{9}{10} \times \frac{9}{10} \times \frac{9}{10} = \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^3$$

This follows from the multiplication rule: on each draw, the chance of red is 1/10, the chance of green is 9/10.

Similarly, the chance of the pattern G R G G R equals

$$\frac{9}{10} \times \frac{1}{10} \times \frac{9}{10} \times \frac{9}{10} \times \frac{1}{10} = \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^3$$

The pattern G R G G R has the same chance as the pattern R R G G G. In fact, each pattern with 2 reds and 3 greens has the same chance,  $(1/10)^2(9/10)^3$ , since the 2 reds will contribute  $(1/10)^2$  to the product and the 3 greens will contribute  $(9/10)^3$ . The sum of the chances of all the patterns, therefore, equals the number of patterns times the common chance.

How many patterns are there? Each pattern is specified by writing down in a row 2 R's and 3 G's, in some order. The number of patterns is given by the *binomial coefficient*,

$$\frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1) \times (3 \times 2 \times 1)} = 10$$

In other words, there are 10 different patterns with 2 R's and 3 G's. So the chance of drawing exactly 2 reds is

$$10 \times \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^3 \approx 7\%$$

Binomial coefficients look messy. Mathematicians get around this by introducing convenient notation. They use an exclamation mark (!) to indicate the result of multiplying together a number and all the numbers which come before it. For example,

$$\begin{aligned}1! &= 1 \\2! &= 2 \times 1 = 2 \\3! &= 3 \times 2 \times 1 = 6 \\4! &= 4 \times 3 \times 2 \times 1 = 24\end{aligned}$$

And so on. The exclamation mark is read "factorial," so that  $4! = 24$  is read "four-factorial equals twenty-four." Now the binomial coefficient is easier to read:

$$\frac{5!}{2! 3!}$$

Remember what the formula represents—the number of different ways of arranging 2 R's and 3 G's in a row.

The 5 in the numerator of the formula is the sum of 2 and 3 in the denominator. Binomial coefficients always take this form. For example, the number of ways to arrange four R's and one G in a row is

$$\frac{5!}{4! 1!} = 5$$

The patterns are

R R R R G      R R R G R      R R G R R      R G R R R      G R R R R

How many ways are there to arrange five R's and zero G's in a row? There is only one way, R R R R R. Applying the formula mechanically gives

$$\frac{5!}{5! 0!}$$

But we have not yet said what 0! means. It is a convention of mathematics that  $0! = 1$ . With this convention, the binomial coefficient does equal 1.

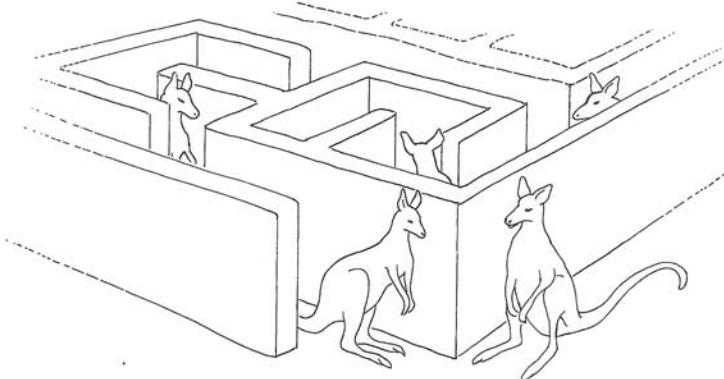
Binomial coefficients and factorials get very large very quickly. For instance, the number of ways to arrange 10 R's and 10 G's in a row is given by the binomial coefficient

$$\frac{20!}{10! 10!} = 184,756$$

However, there was a lot of cancellation going on:  $10! = 3,628,800$ ; and  $20! \approx 2 \times 10^{18}$ , or 2 followed by 18 zeros. (A trillion is 1 followed by 12 zeros.)

### Exercise Set A

1. Find the number of different ways of arranging one R and three G's in a row. Write out all the patterns.
2. Find the number of different ways of arranging two R's and two G's in a row. Write out all the patterns.
3. A box contains one red ball and five green ones. Four draws are made at random with replacement from the box. Find the chance that—
  - (a) a red ball is never drawn
  - (b) a red ball appears exactly once
  - (c) a red ball appears exactly twice
  - (d) a red ball appears exactly three times
  - (e) a red ball appears on all the draws
  - (f) a red ball appears at least twice
4. A die is rolled four times. Find the chance that—
  - (a) an ace (one dot) never appears
  - (b) an ace appears exactly once
  - (c) an ace appears exactly twice
5. A coin is tossed 10 times. Find the chance of getting exactly 5 heads. Find the chance of obtaining between 4 and 6 heads inclusive.
6. It is claimed that a vitamin supplement helps kangaroos learn to run a special maze with high walls. To test whether this is true, 20 kangaroos are divided up into 10 pairs. In each pair, one kangaroo is selected at random to receive the vitamin supplement; the other is fed a normal diet. The kangaroos are then timed as they learn to run the maze. In 7 of the 10 pairs, the treated kangaroo learns to run the maze more quickly than its untreated partner. If in fact the vitamin supplement has



"Did you bring the vitamins ?"

no effect, so that each animal of the pair is equally likely to be the quicker, what is the probability that 7 or more of the treated animals would learn the maze more quickly than their untreated partners, just by chance?

*The answers to these exercises are on pp. A70–71.*

## 2. THE BINOMIAL FORMULA

The reasoning of section 1 is summarized in the *binomial formula*. Suppose a chance process is carried out as a sequence of trials. An example would be rolling a die 10 times, where each roll counts as a trial. There is an event of interest which may or may not occur at each trial: the die may or may not land ace. The problem is to calculate the chance that the event will occur a specified number of times.

The chance that an event will occur exactly  $k$  times out of  $n$  is given by the binomial formula

$$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

In this formula,  $n$  is the number of trials,  $k$  is the number of times the event is to occur, and  $p$  is the probability that the event will occur on any particular trial. The assumptions:

- The value of  $n$  must be fixed in advance.
- $p$  must be the same from trial to trial.
- The trials must be independent.

The formula starts with the binomial coefficient (p. 256),

$$\frac{n!}{k!(n-k)!}$$

Remember, this is the number of ways to arrange  $n$  objects in a row, when  $k$  are alike of one kind and  $n - k$  are alike of another (for instance, red and green marbles).

*Example 1.* A die is rolled 10 times. What is the chance of getting exactly 2 aces?

*Solution.* The number of trials is fixed in advance. It is 10. So  $n = 10$ . The event of interest is rolling an ace. The probability of rolling an ace is the same from trial to trial. It is  $1/6$ . So  $p = 1/6$ . The trials are independent. The binomial formula can be used, and the answer is

$$\frac{10!}{2!8!} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^8 \approx 29\%$$

*Example 2.* A die is rolled until it first lands six. If this can be done using the binomial formula, find the chance of getting 2 aces. If not, why not?

*Solution.* The number of trials is not fixed in advance. It could be 1, if the die lands six right away. Or it could be 2, if the die lands five then six. Or it could be 3. And so forth. The binomial formula does not apply.

*Example 3.* Ten draws are made at random with replacement from the box  $\boxed{1} \boxed{1} \boxed{2} \boxed{3} \boxed{4} \boxed{5}$ . However, just before the last draw is made, whatever else has gone on, the ticket  $\boxed{5}$  is removed from the box. True or false: the chance of drawing exactly two  $\boxed{1}$ 's is

$$\frac{10!}{2! 8!} \left(\frac{2}{6}\right)^2 \left(\frac{4}{6}\right)^8$$

*Solution.* In this example,  $n$  is fixed in advance and the trials are independent. However,  $p$  changes at the last trial from  $2/6$  to  $2/5$ . So the binomial formula does not apply, and the statement is false.

*Example 4.* Four draws are made at random without replacement from the box in example 3. True or false: the chance of drawing exactly two  $\boxed{1}$ 's is

$$\frac{4!}{2! 2!} \left(\frac{2}{6}\right)^2 \left(\frac{4}{6}\right)^2$$

*Solution.* The trials are dependent, so the binomial formula does not apply.

*Technical notes.* (i) To work out the chance in example 4, take a pattern with exactly two 1's, like 1 1 N N, where N means "not 1." The chance of getting 1 1 N N equals

$$\frac{2}{6} \times \frac{1}{5} \times \frac{4}{4} \times \frac{3}{3} = \frac{1}{15}$$

Surprisingly, the chance is the same for all such patterns. How many patterns have exactly two 1's? The answer is

$$\frac{4!}{2! 2!} = 6$$

So the chance of getting exactly two 1's is

$$6 \times \frac{1}{15} = \frac{2}{5}$$

(ii) Mathematicians usually write  $\binom{n}{k}$  for the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

They read  $\binom{n}{k}$  as “ $n$  choose  $k$ ,” the idea being that the formula gives the number of ways to choose  $k$  things out of  $n$ . Older books write the binomial coefficient as  $_n C_k$  or  ${}^n C_k$ , the “number of combinations of  $n$  things taken  $k$  at a time.”

### 3. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. A die will be rolled 6 times. What is the chance of obtaining exactly 1 ace?
2. A die will be rolled 10 times. The chance it never lands six can be found by one of the following calculations. Which one, and why?

$$(i) \left(\frac{1}{6}\right)^{10} \quad (ii) 1 - \left(\frac{1}{6}\right)^{10} \quad (iii) \left(\frac{5}{6}\right)^{10} \quad (iv) 1 - \left(\frac{5}{6}\right)^{10}$$

3. Of families with 4 children, what proportion have more girls than boys? You may assume that the sex of a child is determined as if by drawing at random with replacement from<sup>2</sup>

$$\boxed{\text{M}} \boxed{\text{F}} \quad \text{M = male, F = female}$$

4. A box contains 8 red marbles and 3 green ones. Six draws are made at random without replacement. True or false: the chance that the 3 green marbles are drawn equals

$$\frac{6!}{3!3!} \left(\frac{8}{11}\right)^3 \left(\frac{3}{11}\right)^3$$

Explain briefly.

5. There are 8 people in a club.<sup>3</sup> One person makes up a list of all the possible committees with 2 members. Another person makes up a list of all the possible committees with 5 members. True or false: the second list is longer than the first. Explain briefly.
6. There are 8 people in a club. One person makes up a list of all the possible committees with 2 members. Another person makes up a list of all the possible committees with 6 members. True or false: the second list is longer than the first. Explain briefly.
7. A box contains one red marble and nine green ones. Five draws are made at random with replacement. The chance that exactly two draws will be red is

$$10 \times \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^3$$

Is the addition rule used in deriving this formula? Answer yes or no, and explain carefully.

8. A coin will be tossed 10 times. Find the chance that there will be exactly 2 heads among the first 5 tosses, and exactly 4 heads among the last 5 tosses.
9. For each question (a–e) below, choose one of the answers (i–viii); explain your choice.

*Questions*

A deck of cards is shuffled. What is the chance that—

- (a) the top card is the king of spades and the bottom card is the queen of spades?
- (b) the top card is the king of spades and the bottom card is the king of spades?
- (c) the top card is the king of spades or the bottom card is the king of spades?
- (d) the top card is the king of spades or the bottom card is the queen of spades?
- (e) of the top and bottom cards, one is the king of spades and the other is the queen of spades?

*Answers*

- (i)  $1/52 \times 1/51$
- (ii)  $1/52 + 1/51$
- (iii)  $1/52 \times 1/52$
- (iv)  $1/52 + 1/52$
- (v)  $1 - (1/52 \times 1/51)$
- (vi)  $1 - (1/52 \times 1/52)$
- (vii)  $2/52 \times 1/51$
- (viii) None of the above

10. A box contains 3 red tickets and 2 green ones. Five draws will be made at random. You win \$1 if 3 of the draws are red and 2 are green. Would you prefer the draws to be made with or without replacement? Why?
11. It is now generally accepted that cigarette smoking causes heart disease, lung cancer, and many other diseases. However, in the 1950s, this idea was controversial. There was a strong association between smoking and ill-health, but association is not causation. R. A. Fisher advanced the “constitutional hypothesis:” there is some genetic factor that disposes you both to smoke and to die.

To refute Fisher’s idea, the epidemiologists used twin studies. They identified sets of smoking-discordant monozygotic twin pairs. (“Monozygotic” twins come from one egg and have identical genetic makeup; “smoking-discordant” means that one twin smokes, the other doesn’t.) Now there is a race. Which twin dies first, the smoker or the non-smoker? Data from a Finnish twin study are shown at the top of the next page.<sup>4</sup>

*Data from the Finnish twin study*

	<i>Smokers</i>	<i>Non-smokers</i>
All causes	17	5
Coronary heart disease	9	0
Lung cancer	2	0

According to the first line of the table, there were 22 smoking-discordant monozygotic twin pairs where at least one twin of the pair died. In 17 cases, the smoker died first; in 5 cases, the non-smoker died first. According to the second line, there were 9 pairs where at least one twin died of coronary heart disease; in all 9 cases, the smoker died first. According to the last line, there were 2 pairs where at least one twin died of lung cancer, and in both pairs the smoker won the race to death. (Lung cancer is a rare disease, even among smokers.)

For parts (a–c), suppose that each twin in the pair is equally likely to die first, so the number of pairs in which the smoker dies first is like the number of heads in coin-tossing.

- (a) On this basis, what is the chance of having 17 or more pairs out of 22 where the smoker dies first?
- (b) Repeat the test in part (a), for the 9 deaths from coronary heart disease.
- (c) Repeat the test in part (a), for the 2 deaths from lung cancer.
- (d) Can the difference between the death rates for smoking and non-smoking twins be explained by
  - (i) chance?
  - (ii) genetics?
  - (iii) health effects of smoking?

#### 4. SPECIAL REVIEW EXERCISES

*These exercises cover all of parts I–IV.*

1. In the U.S. in 1990, 20,273 people were murdered, compared to 16,848 in 1970—nearly a 20% increase. “These figures show that the U.S. became a more violent society over the period 1970–1990.” True or false, and explain briefly.<sup>5</sup>
2. A leading cause of death in the U.S. is coronary artery disease—a breakdown of the main arteries to the heart. The disease can be treated with coronary bypass surgery (section 3 of chapter 1). In one of the first trials, Dr. Daniel Ulliyot and associates performed coronary bypass surgery on a test group of patients: 98% survived 3 years or more. The conventional treatment used drugs and special diets to reduce blood pressure and eliminate fatty deposits in the arteries. According to previous studies, only 68% of the patients getting the conventional treatment survived 3 years or more. [Exercise continues . . .]

A newspaper article described Ullyot's results as "spectacular," because the survival rate among Ullyot's patients was much higher than the survival rate in previous studies.<sup>6</sup>

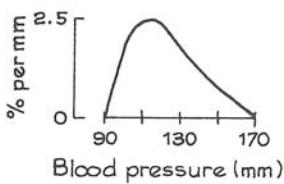
- (a) Did Ullyot's study have contemporaneous controls? If not, what patients were used as the comparison group?
- (b) Was the newspaper article's enthusiasm justified by the study? Discuss briefly.
3. Susan Bouman was denied a promotion to sergeant in the Los Angeles County Sheriff's Department after she took a competitive exam for the position. She filed suit in federal court in April 1980, claiming that the exam was discriminatory.<sup>7</sup> Data for 1975 and 1977 are shown below. In 1975, the pass rate for women was  $10/79 = 12.7\%$ ; for men, it was  $250/1,312 = 19.1\%$ . The "selection ratio" was  $12.7/19.1 = 66.5\%$ : in other words, the women's pass rate was only 66.5% of the men's pass rate. For 1977, the selection ratio was 67.1%. Selection ratios below 80% are generally regarded by the Equal Opportunity Employment Commission as showing "adverse impact" on a "protected group."

Results can also be analyzed by "pooling" data for the two years—just adding up the numbers. For the two years combined, there were  $102 + 79 = 181$  women applicants of whom  $10 + 18 = 28$  passed, and so forth. True or false, and explain: "The selection ratio was 66.5% in 1975 and 67.1% in 1977; therefore, the selection ratio for the pooled data must be between 66.5% and 67.1%."

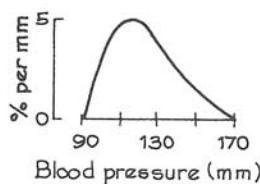
	<i>Women</i>	<i>Men</i>
1975		
Applicants	79	1,312
Passed the exam	10	250
1977		
Applicants	102	1,259
Passed the exam	18	331

4. Three people have tried to sketch the histogram for blood pressures of the subjects in a certain study, using the density scale. Only one is right. Which one, and why?

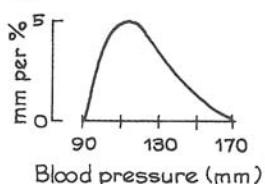
(a)



(b)

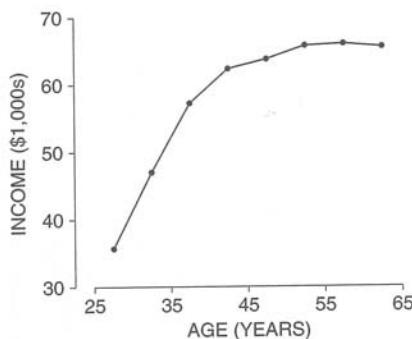


(c)



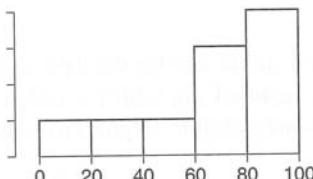
5. A study is made of the age at entrance of college freshmen.<sup>8</sup> Is the SD about 1 month, 1 year, or 5 years. Why?

6. A study is based on a representative sample of men age 25–64 in 2005, who were working full time. The figure below plots average income for each age group.<sup>9</sup> True or false, and explain: the data show that on average, if a man keeps working, his income will increase until age 50 or so, then stabilize. If false, how do you account for the pattern in the data?



Source: March 2005 Current Population Survey; CD-ROM supplied by the Bureau of the Census.

7. True or false, and explain: for the histogram below, the 60th percentile is equal to twice the 30th percentile. (You may assume the distribution is uniform on each class interval.)

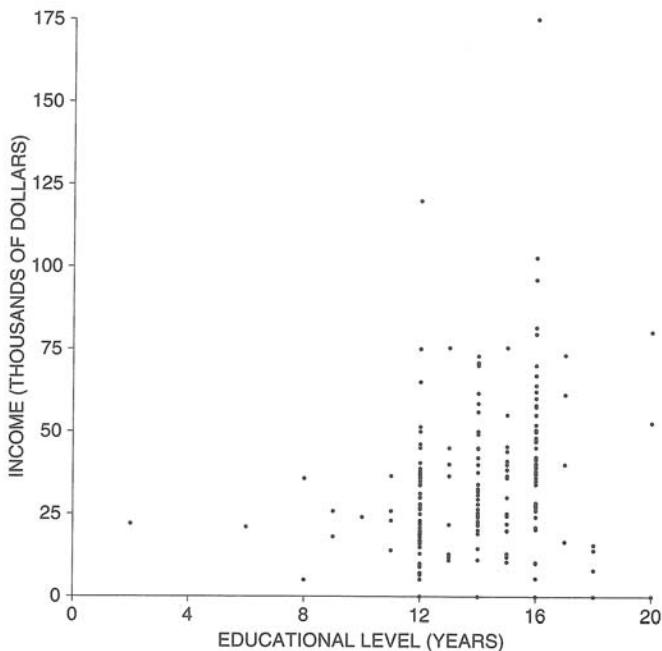


8. True or false, and explain. (You don't need to compute the average or the SD of the lists.)
- The following two lists are the same, when converted to standard units:
 

(i)	1	3	4	7	9	9	9	21	32
(ii)	3	7	9	15	19	19	19	43	65
  - The following two lists are the same, when converted to standard units:
 

(i)	1	3	4	7	9	9	9	21	32
(ii)	-1	-5	-7	-13	-17	-17	-17	-41	-63
9. In a large class, the average score on the final was 50 out of 100, and the SD was 20. The scores followed the normal curve.
- Two brothers took the final. One placed at the 70th percentile and the other was at the 80th percentile. How many points separated them?
  - Two sisters took the final. One placed at the 80th percentile and the other was at the 90th percentile. How many points separated them?

10. The figure below is a scatter plot of income against education (years of schooling completed) for a representative sample of men age 25–34 in Kansas. Or is something wrong? Explain briefly.

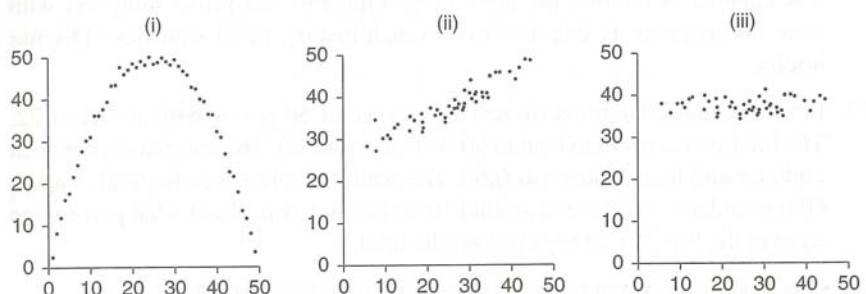


11. (a) Find the correlation coefficient for the data set in table (i) below.  
 (b) If possible, fill in the blanks in table (ii) below so the correlation coefficient is 1. If this is not possible, explain why not.

(i)		(ii)	
$x$	$y$	$x$	$y$
4	7	5	7
5	0	7	—
7	9	8	9
8	9	8	13
8	13	10	—
10	16		

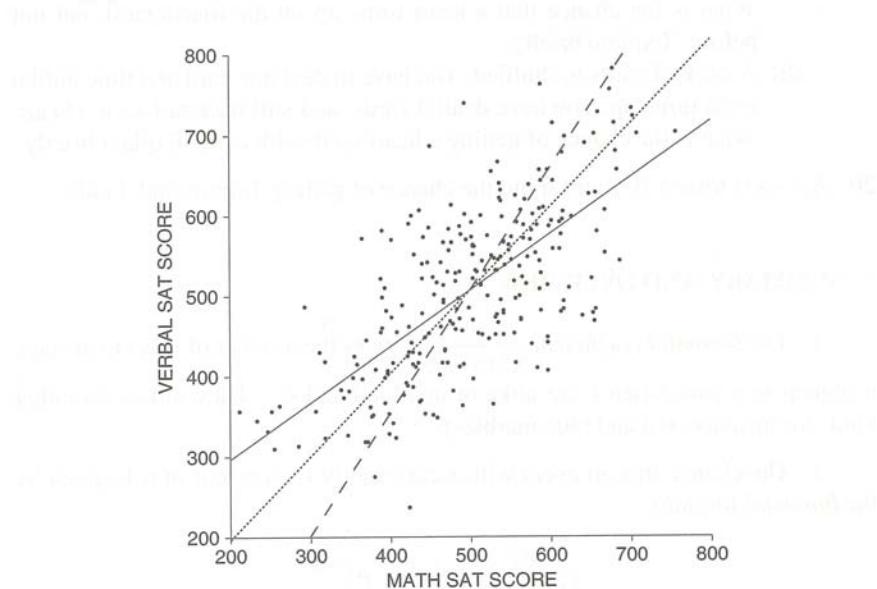
12. In a study of Danish draftees, T. W. Teasdale and associates found a positive correlation between near-sightedness and intelligence.<sup>10</sup> True or false, and explain:
- Draftees who were more near-sighted were also more intelligent, on average.
  - Draftees who were more intelligent were also more near-sighted, on average.
  - The data show that near-sightedness causes intelligence.
  - The data show that intelligence causes near-sightedness.

13. For each diagram below, say whether  $r$  is nearly  $-1$ ,  $0$ , or  $1$ . Explain briefly.



14. The figure below shows a scatter diagram for test scores. Verbal SAT is plotted on the vertical axis and Math SAT on the horizontal. Three lines are drawn across the diagram. Match the line with the description (one description will be left over). Explain briefly.

- (i) estimated average score on V-SAT for given score on M-SAT
- (ii) estimated average score on M-SAT for given score on V-SAT
- (iii) nearly equal percentile ranks on both tests
- (iv) total score on the two tests is about 1,100



15. At a certain law school, first-year scores average 65 and the SD is 12. The correlation between LSAT scores and first-year scores is 0.55. The scatter diagram is football-shaped. The dean's office uses regression to predict first-year scores from LSAT scores. About what percent of the students do better than predicted, by 10 points or more? Explain your answer. If you need more information, say what you need and why.

16. The great prime ministers of France generally served under kings who were less talented. Similarly, the great kings typically had prime ministers who were not as great. Is this a fact of French history, or of statistics? Discuss briefly.
17. In a large class, the midterm had an average of 50 points with an SD of 22. The final scores averaged out to 60 with an SD of 20. The correlation between midterm and final scores was 0.60. The scatter diagram was football-shaped. Of the students who scored around 50 on the midterm, about what percentage were in the top 25% of the class on the final?
18. One ticket is drawn at random from each of the two boxes below:
- |   |   |
|---|---|
| (A)   <span style="border: 1px solid black; padding: 2px;">1</span> <span style="border: 1px solid black; padding: 2px;">2</span> <span style="border: 1px solid black; padding: 2px;">3</span> <span style="border: 1px solid black; padding: 2px;">4</span> <span style="border: 1px solid black; padding: 2px;">5</span> | (B)   <span style="border: 1px solid black; padding: 2px;">1</span> <span style="border: 1px solid black; padding: 2px;">2</span> <span style="border: 1px solid black; padding: 2px;">3</span> <span style="border: 1px solid black; padding: 2px;">4</span> <span style="border: 1px solid black; padding: 2px;">5</span> <span style="border: 1px solid black; padding: 2px;">6</span> |
|---|---|
- Find the chance that—
- One of the numbers is 2 and the other is 5.
  - The sum of the numbers is 7.
  - One number is bigger than twice the other.
19. There are 52 cards in a deck, and 13 of them are hearts.
- Four cards are dealt, one at a time, off the top of a well-shuffled deck. What is the chance that a heart turns up on the fourth card, but not before? Explain briefly.
  - A deck of cards is shuffled. You have to deal one card at a time until a heart turns up. You have dealt 3 cards, and still have not seen a heart. What is the chance of getting a heart on the 4th card? Explain briefly.
20. A coin is tossed 10 times. Find the chance of getting 7 heads and 3 tails.

## 5. SUMMARY AND OVERVIEW

- The *binomial coefficient*  $\frac{n!}{k!(n-k)!}$  gives the number of ways to arrange  $n$  objects in a row, when  $k$  are alike of one kind and  $n - k$  are alike of another kind (for instance, red and blue marbles).
- The chance that an event will occur exactly  $k$  times out of  $n$  is given by the *binomial formula*

$$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

In this formula,  $n$  is the number of trials,  $k$  is the number of times the event is to occur, and  $p$  is the probability that the event will occur on any particular trial. The assumptions:

- The value of  $n$  must be fixed in advance.
- $p$  must be the same from trial to trial.
- The trials must be independent.

3. This part of the book defined conditional probabilities, independence, and the multiplication rule. The addition rule was introduced for mutually exclusive events.

4. The binomial formula is an application of the multiplication rule combined with the addition rule.

5. Independence is the basis for the statistical theory to be developed in part V, and the crucial assumption behind many of the procedures to be discussed in parts VI–VIII.

— — — — —

PART V

# Chance Variability

— — — — —

# 16

## The Law of Averages

*The roulette wheel has neither conscience nor memory.*

—JOSEPH BERTRAND (FRENCH MATHEMATICIAN, 1822–1900)

### 1. WHAT DOES THE LAW OF AVERAGES SAY?

A coin lands heads with chance 50%. After many tosses, the number of heads should equal the number of tails: isn't that the law of averages? John Kerrich, a South African mathematician, found out the hard way. He was visiting Copenhagen when World War II broke out. Two days before he was scheduled to fly to England, the Germans invaded Denmark. Kerrich spent the rest of the war interned at a camp in Jutland. To pass the time he carried out a series of experiments in probability theory.<sup>1</sup> One experiment involved tossing a coin 10,000 times. With his permission, some of the results are summarized in table 1 and figure 1 (pp. 274–275 below). What do these results say about the law of averages? To find out, let's pretend that at the end of World War II, Kerrich was invited to demonstrate the law of averages to the King of Denmark. He is discussing the invitation with his assistant.

*Assistant.* So you're going to tell the king about the law of averages.

*Kerrich.* Right.

*Assistant.* What's to tell? I mean, everyone knows about the law of averages, don't they?

*Kerrich.* OK. Tell me what the law of averages says.

*Assistant.* Well, suppose you're tossing a coin. If you get a lot of heads, then tails start coming up. Or if you get too many tails, the chance for heads goes up. In the long run, the number of heads and the number of tails even out.

*Kerrich.* It's not true.

*Assistant.* What do you mean, it's not true?

*Kerrich.* I mean, what you said is all wrong. First of all, with a fair coin the chance for heads stays at 50%, no matter what happens. Whether there are two heads in a row or twenty, the chance of getting a head next time is still 50%.

*Assistant.* I don't believe it.

*Kerrich.* All right. Take a run of four heads, for example. I went through the record of my first 2,000 tosses. In 130 cases, the coin landed heads four times in a row; 69 of these runs were followed by a head, and only 61 by a tail. A run of heads just doesn't make tails more likely next time.

*Assistant.* You're always telling me these things I don't believe. What are you going to tell the king?

*Kerrich.* Well, I tossed the coin 10,000 times, and I got about 5,000 heads. The exact number was 5,067. The difference of 67 is less than 1% of the number of tosses. I have the record here in table 1.

*Assistant.* Yes, but 67 heads is a lot of heads. The king won't be impressed, if that's the best the law of averages can do.

*Kerrich.* What do you suggest?

Table 1. John Kerrich's coin-tossing experiment. The first column shows the number of tosses. The second shows the number of heads. The third shows the difference

number of heads — half the number of tosses.

Number of tosses	Number of heads	Differ- ence	Number of tosses	Number of heads	Differ- ence
10	4	-1	600	312	12
20	10	0	700	368	18
30	17	2	800	413	13
40	21	1	900	458	8
50	25	0	1,000	502	2
60	29	-1	2,000	1,013	13
70	32	-3	3,000	1,510	10
80	35	-5	4,000	2,029	29
90	40	-5	5,000	2,533	33
100	44	-6	6,000	3,009	9
200	98	-2	7,000	3,516	16
300	146	-4	8,000	4,034	34
400	199	-1	9,000	4,538	38
500	255	5	10,000	5,067	67

*Assistant.* Toss the coin another 10,000 times. With 20,000 tosses, the number of heads should be quite a bit closer to the expected number. After all, eventually the number of heads and the number of tails have to even out, right?

*Kerrich.* You said that before, and it's wrong. Look at table 1. In 1,000 tosses, the difference between the number of heads and the expected number was 2. With 2,000 tosses, the difference went up to 13.

*Assistant.* That was just a fluke. By toss 3,000, the difference was only 10.

*Kerrich.* That's just another fluke. At toss 4,000, the difference was 29. At 5,000, it was 33. Sure, it dropped back to 9 at toss 6,000, but look at figure 1. The chance error is climbing pretty steadily from 1,000 to 10,000 tosses, and it's going straight up at the end.

*Assistant.* So where's the law of averages?

*Kerrich.* With a large number of tosses, the size of the difference between the number of heads and the expected number is likely to be quite large in absolute terms. But compared to the number of tosses, the difference is likely to be quite small. That's the law of averages. Just like I said, 67 is only a small fraction of 10,000.

*Assistant.* I don't understand.

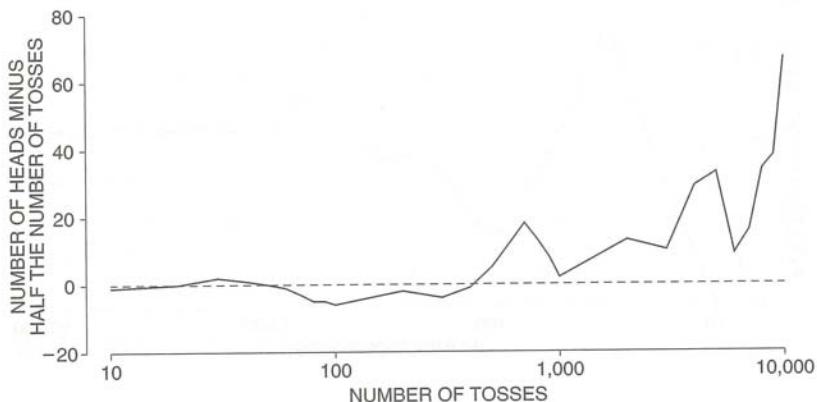
*Kerrich.* Look. In 10,000 tosses you expect to get 5,000 heads, right?

*Assistant.* Right.

*Kerrich.* But not exactly. You only expect to get around 5,000 heads. I mean, you could just as well get 5,001 or 4,998 or 5,007. The amount off 5,000 is what we call "chance error."

Figure 1. Kerrich's coin-tossing experiment. The "chance error" is  
 $\text{number of heads} - \frac{1}{2} \times \text{number of tosses}$ .

This difference is plotted against the number of tosses. As the number of tosses goes up, the size of the chance error tends to go up. The horizontal axis is not to scale and the curve is drawn by linear interpolation.



*Assistant.* Can you be more specific?

*Kerrich.* Let me write an equation:

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

This error is likely to be large in absolute terms, but small compared to the number of tosses. Look at figure 2. That's the law of averages, right there.

*Assistant.* Hmm. But what would happen if you tossed the coin another 10,000 times. Then you'd have 20,000 tosses to work with.

*Kerrich.* The chance error would go up, but not by a factor of two. In absolute terms, the chance error gets bigger.<sup>2</sup> But as a percentage of the number of tosses, it gets smaller.

*Assistant.* Tell me again what the law of averages says.

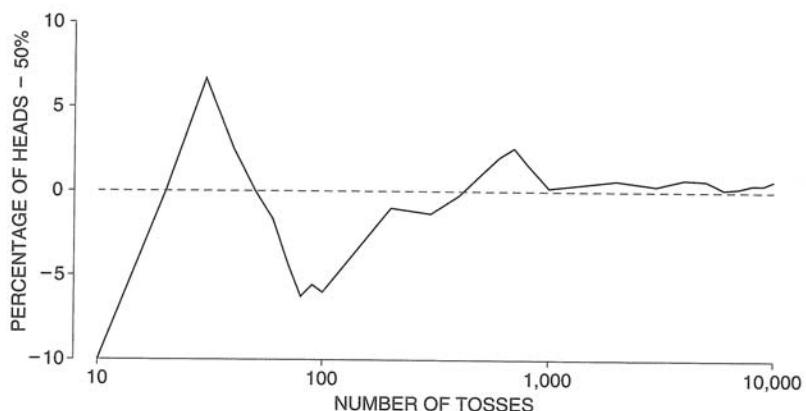
*Kerrich.* The number of heads will be around half the number of tosses, but it will be off by some amount—chance error. As the number of tosses goes up, the chance error gets bigger in absolute terms. Compared to the number of tosses, it gets smaller.

*Assistant.* Can you give me some idea of how big the chance error is likely to be?

*Kerrich.* Well, with 100 tosses, the chance error is likely to be around 5 in size. With 10,000 tosses, the chance error is likely to be around 50 in size. Multiplying the number of tosses by 100 only multiplies the likely size of the chance error by  $\sqrt{100} = 10$ .

*Assistant.* What you're saying is that as the number of tosses goes up, the difference between the number of heads and half the number of tosses gets

Figure 2. The chance error expressed as a percentage of the number of tosses. When the number of tosses goes up, this percentage goes down: the chance error gets smaller relative to the number of tosses. The horizontal axis is not to scale and the curve is drawn by linear interpolation.





bigger; but the difference between the percentage of heads and 50% gets smaller.

*Kerrich.* That's it.

### Exercise Set A

1. A machine has been designed to toss a coin automatically and keep track of the number of heads. After 1,000 tosses, it has 550 heads. Express the chance error both in absolute terms and as a percentage of the number of tosses.
2. After 1,000,000 tosses, the machine in exercise 1 has 501,000 heads. Express the chance error in the same two ways.
3. A coin is tossed 100 times, landing heads 53 times. However, the last seven tosses are all heads. True or false: the chance that the next toss will be heads is somewhat less than 50%. Explain.
4. (a) A coin is tossed, and you win a dollar if there are more than 60% heads. Which is better: 10 tosses or 100? Explain.  
 (b) As in (a), but you win the dollar if there are more than 40% heads.  
 (c) As in (a), but you win the dollar if there are between 40% and 60% heads.  
 (d) As in (a), but you win the dollar if there are exactly 50% heads.
5. With a Nevada roulette wheel, there are 18 chances in 38 that the ball will land in a red pocket. A wheel is going to be spun many times. There are two choices:
  - (i) 38 spins, and you win a dollar if the ball lands in a red pocket 20 or more times.
  - (ii) 76 spins, and you win a dollar if the ball lands in a red pocket 40 or more times.

Which is better? Or are they the same? Explain.

*The next three exercises involve drawing at random from a box. This was described in section 1 of chapter 13 and is reviewed in section 3 below.*

6. A box contains 20% red marbles and 80% blue marbles. A thousand marbles are drawn at random with replacement. One of the following statements is true. Which one, and why?
  - (i) Exactly 200 marbles are going to be red.
  - (ii) About 200 marbles are going to be red, give or take a dozen or so.
7. Repeat exercise 6, if the draws are made at random without replacement and the box contains 50,000 marbles.
8. One hundred tickets will be drawn at random with replacement from one of the two boxes shown below. On each draw, you will be paid the amount shown on the ticket, in dollars. (If a negative number is drawn, that amount will be taken away from you.) Which box is better? Or are they the same?
 

(i)   <span style="border: 1px solid black; padding: 2px;">-1</span>   <span style="border: 1px solid black; padding: 2px;">-1</span>   <span style="border: 1px solid black; padding: 2px;">1</span>   <span style="border: 1px solid black; padding: 2px;">1</span>	(ii)   <span style="border: 1px solid black; padding: 2px;">-1</span>   <span style="border: 1px solid black; padding: 2px;">1</span>
---	---
9. (Hard.) Look at figure 1. If Kerrich kept on tossing, would the graph ever get negative?

*The answers to these exercises are on pp. A71–72.*

## 2. CHANCE PROCESSES

Kerrich's assistant was struggling with the problem of chance variability. He came to see that when a coin is tossed a large number of times, the actual number of heads is likely to differ from the expected number. But he didn't know how big a difference to anticipate. A method for calculating the likely size of the difference will be presented in the next chapter. This method works in many different situations. For example, it can be used to see how much money the house should expect to win at roulette (chapter 17) or how accurate a sample survey is likely to be (chapter 21).

What is the common element? All these problems are about chance processes.<sup>3</sup> Take the number of heads in Kerrich's experiment. Chance comes in with each toss of the coin. If you repeat the experiment, the tosses turn out differently, and so does the number of heads. Second example: the amount of money won or lost at roulette. Spinning the wheel is a chance process, and the amounts won or lost depend on the outcome. Spin again, and winners become losers. A final example: the percentage of Democrats in a random sample of voters. A chance process is used to draw the sample. So the number of Democrats in the sample is determined by the luck of the draw. Take another sample, and the percentages would change.

To what extent are the numbers influenced by chance? This sort of question must be faced over and over again in statistics. A general strategy will be presented in the next few chapters. The two main ideas:

- Find an analogy between the process being studied (sampling voters in the poll example) and drawing numbers at random from a box.

- Connect the variability you want to know about (for example, in the estimate for the Democratic vote) with the chance variability in the sum of the numbers drawn from the box.

The analogy between a chance process and drawing from a box is called a *box model*. The point is that the chance variability in the sum of numbers drawn from a box will be easy to analyze mathematically. More complicated processes can then be dealt with through the analogy.

### 3. THE SUM OF DRAWS

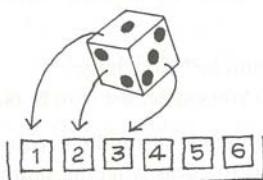
The object of this section is to illustrate the following process. There is a box of tickets. Each ticket has a number written on it. Then some tickets are drawn at random from the box, and the numbers on these tickets are added up. For example, take the box

<b>[1]</b>	<b>[2]</b>	<b>[3]</b>	<b>[4]</b>	<b>[5]</b>	<b>[6]</b>
------------	------------	------------	------------	------------	------------

Imagine drawing twice at random with replacement from this box. You shake the box to mix up the tickets, pick one ticket at random, make a note of the number on it, put it back in the box. Then you shake the box again, and make a second draw at random. The phrase “with replacement” reminds you to put the ticket back in the box before drawing again. Putting the tickets back enables you to draw over and over again, under the same conditions. (Drawing with and without replacement was discussed in section 1 of chapter 13.)

Having drawn twice at random with replacement, you add up the two numbers. For example, the first draw might be **[3]** and the second **[5]**. Then the sum of the draws is 8. Or the first draw might be **[3]** and the second **[3]** too, so the sum of the draws is 6. There are many other possibilities. The sum is subject to chance variability. If the draws turn out one way, the sum is one thing; if they turn out differently, the sum is different too.

At first, this example may seem artificial. But it is just like a turn at Monopoly—you roll a pair of dice, add up the two numbers, and move that many squares. Rolling a die is just like picking a number from the box.



Next, imagine taking 25 draws from the same box

<b>[1]</b>	<b>[2]</b>	<b>[3]</b>	<b>[4]</b>	<b>[5]</b>	<b>[6]</b>
------------	------------	------------	------------	------------	------------

Of course, the draws must be made with replacement. About how big is their sum going to be? The most direct way to find out is by experiment. We programmed

the computer to make the draws.<sup>4</sup> It got 3 on the first draw, 2 on the second, 4 on the third. Here they all are:

3 2 4 6 2    3 5 4 4 2    3 6 4 1 2    4 1 5 5 6    2 2 2 5 5

The sum of these 25 draws is 88.

Of course, if the draws had been different, their sum would have been different. So we had the computer repeat the whole process ten times. Each time, it made 25 draws at random with replacement from the box, and took their sum. The results:

88 84 80 90 83 78 95 94 80 89

Chance variability is easy to see. The first sum is 88, the second drops to 84, the third drops even more to 80. The values range from a low of 78 to a high of 95.

In principle, the sum could have been as small as  $25 \times 1 = 25$ , or as large as  $25 \times 6 = 150$ . But in fact, the ten observed values are all between 75 and 100. Would this keep up with more repetitions? Just what is the chance that the sum turns out to be between 75 and 100? That kind of problem will be solved in the next two chapters.

The *sum of the draws* from a box is shorthand for the process discussed in this section:

- Draw tickets at random from a box.
- Add up the numbers on the tickets.<sup>5</sup>

### Exercise Set B

1. One hundred draws are made at random with replacement from the box  $\boxed{1} \boxed{2}$ . Forty-seven draws turn out to be  $\boxed{1}$ , and the remaining 53 are  $\boxed{2}$ . How much is the sum?
2. One hundred draws are made at random with replacement from the box  $\boxed{1} \boxed{2}$ .
  - (a) How small can the sum be? How large?
  - (b) How many times do you expect the ticket  $\boxed{1}$  to turn up? The ticket  $\boxed{2}$ ?
  - (c) About how much do you expect the sum to be?
3. One hundred draws are made at random with replacement from the box  $\boxed{1} \boxed{2} \boxed{9}$ .
  - (a) How small can the sum be? How large?
  - (b) About how much do you expect the sum to be?
4. One hundred draws will be made at random with replacement from one of the following boxes. Your job is to guess what the sum will be, and you win \$1 if you are right to within 10. In each case, what would you guess? Which box is best? Worst?
  - (i)  $\boxed{1} \boxed{9}$
  - (ii)  $\boxed{4} \boxed{6}$
  - (iii)  $\boxed{5} \boxed{5}$
5. One ticket will be drawn at random from the box
 
$$\boxed{1} \boxed{2} \boxed{3} \boxed{4} \boxed{5} \boxed{6} \boxed{7} \boxed{8} \boxed{9} \boxed{10}$$

What is the chance that it will be 1? That it will be 3 or less? 4 or more?

6. Fifty draws will be made at random with replacement from one of the two boxes shown below. On each draw, you will be paid in dollars the amount shown on the ticket; if a negative number is drawn, that amount will be taken away from you. Which box is better? Or are they the same? Explain.

(i) 

-1	2
----	---

(ii) 

-1	-1	2
----	----	---

7. You gamble four times at a casino. You win \$4 on the first play, lose \$2 on the second, win \$5 on the third, lose \$3 on the fourth. Which of the following calculations tells how much you come out ahead? (More than one may be correct.)

- (i)  $\$4 + \$5 - (\$2 + \$3)$
- (ii)  $\$4 + (-\$2) + \$5 + (-\$3)$
- (iii)  $\$4 + \$2 + \$5 - \$3$
- (iv)  $-\$4 + \$2 + \$5 + \$3$

*The answers to these exercises are on p. A72.*

#### 4. MAKING A BOX MODEL

The object of this section is to make some box models, as practice for later. The sum of the draws from the box turns out to be the key ingredient for many statistical procedures, so keep your eye on the sum. There are three questions to answer when making a box model:

- What numbers go into the box?
- How many of each kind?
- How many draws?

The purpose of a box model is to analyze chance variability, which can be seen in its starkest form at any gambling casino. So this section will focus on box models for roulette. A Nevada roulette wheel has 38 pockets. One is numbered 0, another is numbered 00, and the rest are numbered from 1 through 36. The croupier spins the wheel, and throws a ball onto the wheel. The ball is equally likely to land in any one of the 38 pockets. Before it lands, bets can be placed on the table (figure 3 on the next page).

One bet is *red or black*. Except for 0 and 00, which are colored green, the numbers on the roulette wheel alternate red and black. If you bet a dollar on red, say, and a red number comes up, you get the dollar back together with another dollar in winnings. If a black or green number comes up, the croupier smiles and rakes in your dollar.

Suppose you are at the Golden Nugget in Las Vegas. You have just put a dollar on red, and the croupier spins the wheel. It may seem hard to figure your chances, but a box model will help. What numbers go into the box? You will either win a dollar or lose a dollar. So the tickets must show either  $+\$1$  or  $-\$1$ .

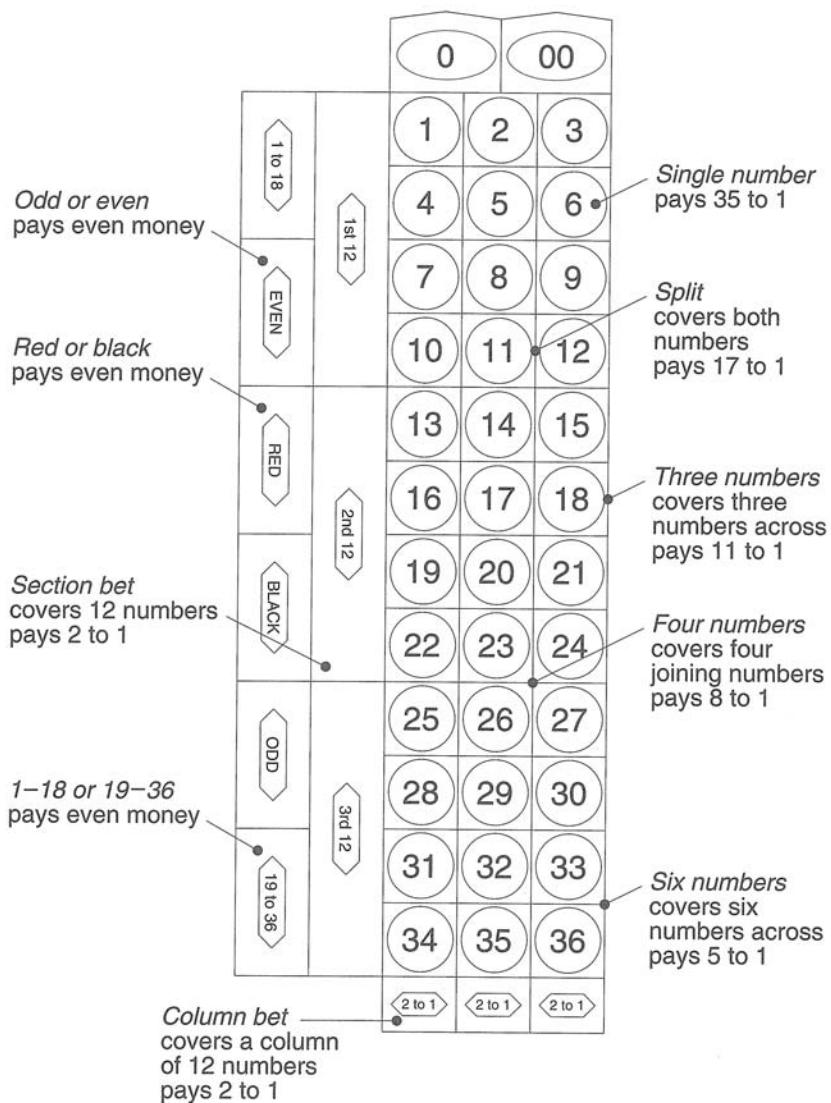
The second question is, how many of each kind? You win if one of the 18 red numbers comes up, and lose if one of the 18 black numbers comes up. But you also lose if 0 or 00 come up. And that is where the house gets its edge. Your

chance of winning is only 18 in 38, and the chance of losing is 20 in 38. So there are 18  $+\$1$ 's and 20  $-\$1$ 's. The box is

18 tickets $+\$1$	20 tickets $-\$1$
-------------------	-------------------

As far as the chances are concerned, betting a dollar on red is just like drawing a ticket at random from the box. The great advantage of the box model is that all

Figure 3. A Nevada roulette table.



*Roulette is a pleasant, relaxed, and highly comfortable way to lose your money.*

—JIMMY THE GREEK

the irrelevant details—the wheel, the table, and the croupier's smile—have been stripped away. And you can see the cruel reality: you have 18 tickets, they have 20.

That does one play. But suppose you play roulette ten times, betting a dollar on red each time. What is likely to happen then? You will end up ahead or behind by some amount. This amount is called your *net gain*. The net gain is positive if you come out ahead, negative if you come out behind.

To figure the chances, the net gain has to be connected to the box. On each play, you win or lose some amount. These ten win-lose numbers are like ten draws from the box, made at random with replacement. (Replacing the tickets keeps the chances on each draw the same as the chances for the wheel.) The net gain—the total amount won or lost—is just the sum of these ten win-lose numbers. Your net gain in ten plays is like the sum of ten draws made at random with replacement from the box

18 tickets	$+\$1$	20 tickets	$-\$1$
------------	--------	------------	--------

This is our first model, so it is a good idea to look at it more closely. Suppose, for instance, that the ten plays came out this way:

R R R B G    R R B B R

(R means red, B means black, and G means green—the house numbers 0 and 00). Table 2 below shows the ten corresponding win-lose numbers, and the net gain.

Table 2. The net gain. This is the cumulative sum of the win-lose numbers.

Plays	R	R	R	B	G	R	R	B	B	R
Win-lose numbers	+1	+1	+1	-1	-1	+1	+1	-1	-1	+1
Net gain	1	2	3	2	1	2	3	2	1	2

Follow the net gain along. When you get a red, the win-lose number is +1, and the net gain goes up by 1. When you get a black or a green, the win-lose number is -1, and the net gain goes down by 1. The net gain is just the sum of the win-lose numbers, and these are like the draws from the box. That is why the net gain is like the sum of draws from the box. This game had a happy ending: you came out ahead \$2. To see what would happen if you kept on playing, read the next chapter.

*Example 1.* If you bet a dollar on a single number at Nevada roulette, and that number comes up, you get the \$1 back together with winnings of \$35. If any other number comes up, you lose the dollar. Gamblers say that a single number *pays 35 to 1*. Suppose you play roulette 100 times, betting a dollar on the number 17 each time. Your net gain is like the sum of \_\_\_\_\_ draws made at random with replacement from the box \_\_\_\_\_. Fill in the blanks.

*Solution.* What numbers go into the box? To answer this question, think about one play of the game. You put a dollar chip on 17. If the ball drops into the pocket 17, you'll be up \$35. If it drops into any other pocket, you'll be down \$1. So the box has to contain the tickets  $\boxed{\$35}$  and  $\boxed{-\$1}$ .

The tickets in the box show the various amounts that can be won or lost on a single play.

How many tickets of each kind? Keep thinking about one play. You have only 1 chance in 38 of winning, so the chance of drawing  $[\$35]$  has to be 1 in 38. You have 37 chances in 38 of losing, so the chance of drawing  $[-\$1]$  has to be 37 in 38. The box is

1 ticket	$[\+$35]$	37 tickets	$[-\$1]$
----------	-----------	------------	----------

The chance of drawing any particular number from the box must equal the chance of winning that amount on a single play. (“Winning” a negative amount is the mathematical equivalent of what most people call losing.)

How many draws? You are playing 100 times. The number of draws has to be 100. Tickets must be replaced after each draw, so as not to change the odds.

The number of draws equals the number of plays.
---

So, the net gain in 100 plays is like the sum of 100 draws made at random with replacement from the box

1 ticket	$[\+$35]$	37 tickets	$[-\$1]$
----------	-----------	------------	----------

This completes the solution.

### Exercise Set C

- Consider the following three situations.
  - A box contains one ticket marked “0” and nine marked “1.” A ticket is drawn at random. If it shows “1” you win a panda bear.
  - A box contains ten tickets marked “0” and ninety marked “1.” One ticket is drawn at random. If it shows “1” you win the panda.
  - A box contains one ticket marked “0” and nine marked “1.” Ten draws are made at random with replacement. If the sum of the draws equals 10, you win the panda.

Assume you want the panda. Which is better—(i) or (ii)? Or are they the same? What about (i) and (iii)?

- A gambler is going to play roulette 25 times, putting a dollar on a *split* each time. (A split is two adjacent numbers, like 11 and 12 in figure 3 on p. 282.) If either

number comes up, the gambler gets the dollar back, together with winnings of \$17. If neither number comes up, he loses the dollar. So a split pays 17 to 1, and there are 2 chances in 38 to win. The gambler's net gain in the 25 plays is like the sum of 25 draws made from one of the following boxes. Which one, and why?

- (i) 

0	00
---	----

 36 tickets numbered 

1
---

 through 

36
----
- (ii) 

\$17	\$17
------	------

 34 tickets 

-\$1
------
- (iii) 

\$17	\$17
------	------

 36 tickets 

-\$1
------

3. In one version of chuck-a-luck, 3 dice are rolled out of a cage. You can bet that all 3 show six. The house pays 36 to 1, and the bettor has 1 chance in 216 to win. Suppose you make this bet 10 times, staking \$1 each time. Your net gain is like the sum of \_\_\_\_\_ draws made at random with replacement from the box \_\_\_\_\_. Fill in the blanks.

*The answers to these exercises are on p. A72.*

## 5. REVIEW EXERCISES

1. A box contains 10,000 tickets: 4,000 

0
---

's and 6,000 

1
---

's. And 10,000 draws will be made at random with replacement from this box. Which of the following best describes the situation, and why?
  - (i) The number of 1's will be 6,000 exactly.
  - (ii) The number of 1's is very likely to equal 6,000, but there is also some small chance that it will not be equal to 6,000.
  - (iii) The number of 1's is likely to be different from 6,000, but the difference is likely to be small compared to 10,000.
2. Repeat exercise 1 for 10,000 draws made at random without replacement from the box.
3. A gambler loses ten times running at roulette. He decides to continue playing because he is due for a win, by the law of averages. A bystander advises him to quit, on the grounds that his luck is cold. Who is right? Or are both of them wrong?
4. (a) A die will be rolled some number of times, and you win \$1 if it shows an ace (

•
---

) more than 20% of the time. Which is better: 60 rolls, or 600 rolls? Explain.
  - (b) As in (a), but you win the dollar if the percentage of aces is more than 15%.
  - (c) As in (a), but you win the dollar if the percentage of aces is between 15% and 20%.
  - (d) As in (a), but you win the dollar if the percentage of aces is exactly  $16\frac{2}{3}\%$ .
5. True or false: if a coin is tossed 100 times, it is not likely that the number of heads will be exactly 50, but it is likely that the percentage of heads will be exactly 50%. Explain.

6. According to genetic theory, there is very close to an even chance that both children in a two-child family will be of the same sex. Here are two possibilities.
- 15 couples have two children each. In 10 or more of these families, it will turn out that both children are of the same sex.
  - 30 couples have two children each. In 20 or more of these families, it will turn out that both children are of the same sex.

Which possibility is more likely, and why?

7. A quiz has 25 multiple choice questions. Each question has 5 possible answers, one of which is correct. A correct answer is worth 4 points, but a point is taken off for each incorrect answer. A student answers all the questions by guessing at random. The score will be like the sum of \_\_\_\_\_ draws from the box \_\_\_\_\_. Fill in the first blank with a number and the second with a box of tickets. Explain your answers.
8. A gambler will play roulette 50 times, betting a dollar on four joining numbers each time (like 23, 24, 26, 27 in figure 3, p. 282). If one of these four numbers comes up, she gets the dollar back, together with winnings of \$8. If any other number comes up, she loses the dollar. So this bet pays 8 to 1, and there are 4 chances in 38 of winning. Her net gain in 50 plays is like the sum of \_\_\_\_\_ draws from the box \_\_\_\_\_. Fill in the blanks; explain.
9. A box contains red and blue marbles; there are more red marbles than blue ones. Marbles are drawn one at a time from the box, at random with replacement. You win a dollar if a red marble is drawn more often than a blue one.<sup>6</sup> There are two choices:
- 100 draws are made from the box.
  - 200 draws are made from the box.
- Choose one of the four options below; explain your answer.
- A gives a better chance of winning.
  - B gives a better chance of winning.
  - A and B give the same chance of winning.
  - Can't tell without more information.
10. Two hundred draws will be made at random with replacement from the box [-3] [-2] [-1] [0] [1] [2] [3].
- If the sum of the 200 numbers drawn is 30, what is their average?
  - If the sum of the 200 numbers drawn is -20, what is their average?
  - In general, how can you figure the average of the 200 draws, if you are told their sum?
  - There are two alternatives:
    - winning \$1 if the sum of the 200 numbers drawn is between -5 and +5.
    - winning \$1 if the average of the 200 numbers drawn is between -0.025 and +0.025.

Which is better, or are they the same? Explain.

## 6. SUMMARY

- There is *chance error* in the number of heads:

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

The error is likely to be large in absolute terms, but small relative to the number of tosses. That is the *law of averages*.

2. The law of averages can be stated in percentage terms. With a large number of tosses, the percentage of heads is likely to be close to 50%, although it is not likely to be exactly equal to 50%.

3. The law of averages does not work by changing the chances. For example, after a run of heads in coin tossing, a head is still just as likely as a tail.

4. A complicated chance process for generating a number can often be modeled by drawing from a box. The sum of the draws is a key ingredient.

- The basic questions to ask when making a box model:

- Which numbers go into the box?
- How many of each kind?
- How many draws?

6. For gambling problems in which the same bet is made several times, a box model can be set up as follows:

- The tickets in the box show the amounts that can be won (+) or lost (-) on each play.
- The chance of drawing any particular value from the box equals the chance of winning that amount on a single play.
- The number of draws equals the number of plays.

Then, the *net gain* is like the sum of the draws from the box.



Drawing by Dana Fradon; © 1976 The New Yorker Magazine, Inc.

# 17

## The Expected Value and Standard Error

*If you believe in miracles, head for the Keno lounge.*

—JIMMY THE GREEK

### 1. THE EXPECTED VALUE

A chance process is running. It delivers a number. Then another. And another. You are about to drown in random output. But mathematicians have found a little order in this chaos. The numbers delivered by the process vary around the *expected value*, the amounts off being similar in size to the *standard error*. To be more specific, imagine generating a number through the following chance process: count the number of heads in 100 tosses of a coin. You might get 57 heads. This is 7 above the expected value of 50, so the chance error is +7. If you made another 100 tosses, you would get a different number of heads, perhaps 46. The chance error would be -4. A third repetition might generate still another number, say 47; and the chance error would be -3. Your numbers will be off 50 by chance amounts similar in size to the standard error, which is 5 (section 5 below).

The formulas for the expected value and standard error depend on the chance process which generates the number. This chapter deals with the sum of draws from a box, and the formula for the expected value will be introduced with an example: the sum of 100 draws made at random with replacement from the box

[1]	[1]	[1]	[5]
-----	-----	-----	-----

About how large should this sum be? To answer this question, think how the draws should turn out. There are four tickets in the box, so **5** should come up on around one-fourth of the draws, and **1** on three-fourths. With 100 draws, you can expect to get around twenty-five **5**'s, and seventy-five **1**'s. The sum of the draws should be around

$$25 \times 5 + 75 \times 1 = 200.$$

That is the expected value.

The formula for the expected value is a short-cut. It has two ingredients:

- the number of draws;
- the average of the numbers in the box, abbreviated to “average of box.”

The expected value for the sum of draws made at random with replacement from a box equals

$$\text{(number of draws)} \times \text{(average of box)}.$$

To see the logic behind the formula, go back to the example. The average of the box is

$$\frac{1 + 1 + 1 + 5}{4} = 2.$$

On the average, each draw adds around 2 to the sum. With 100 draws, the sum must be around  $100 \times 2 = 200$ .

*Example 1.* Suppose you are going to Las Vegas to play Keno. Your favorite bet is a dollar on a single number. When you win, they give you the dollar back and two dollars more. When you lose, they keep the dollar. There is 1 chance in 4 to win.<sup>1</sup> About how much should you expect to win (or lose) in 100 plays, if you make this bet on each play?

*Solution.* The first step is to write down a box model. On each play, your net gain either goes up by \$2 or goes down by \$1. There is 1 chance in 4 to go up; there are 3 chances in 4 to go down. So your net gain after 100 plays is like the sum of 100 draws at random with replacement from the box

\$2	-\$1	-\$1	-\$1	
-----	------	------	------	--

The average of this box is

$$\frac{\$2 - \$1 - \$1 - \$1}{4} = -\$0.25$$

On the average, each play costs you a quarter. In 100 plays, you can expect to lose around \$25. This is the answer. If you continued on, in 1,000 plays you should expect to lose around \$250. The more you play, the more you lose. Perhaps you should look for another game.

### Exercise Set A

1. Find the expected value for the sum of 100 draws at random with replacement from the box—
 

(a)	0	1	1	6	
-----	---	---	---	---	--

(b)	-2	-1	0	2	
-----	----	----	---	---	--

(c)	-2	-1	3	
-----	----	----	---	--

(d)	0	1	1	
-----	---	---	---	--
2. Find the expected number of squares moved on the first play in Monopoly (p. 279).
3. Someone is going to play roulette 100 times, betting a dollar on the number 17 each time. Find the expected value for the net gain. (See pp. 283–284.)
4. You are going to play roulette 100 times, staking \$1 on red-or-black each time. Find the expected value for your net gain. (This bet pays even money, and you have 18 chances in 38 of winning; figure 3 on p. 282.)
5. Repeat exercise 4 for 1,000 plays.
6. A game is *fair* if the expected value for the net gain equals 0: on the average, players neither win nor lose. A generous casino would offer a bit more than \$1 in winnings if a player staked \$1 on red-and-black in roulette and won. How much should they pay to make it a fair game? (Hint: Let  $x$  stand for what they should pay. The box has 18 tickets  $\boxed{x}$  and 20 tickets  $\boxed{-\$1}$ . Write down the formula for the expected value in terms of  $x$  and set it equal to 0.)
7. If an Adventurer at the Game of the Royal Oak staked 1 pound on a point and won, how much should the Master of the Ball have paid him, for the Game to be fair? (The rules are explained in exercise 6 on pp. 250–251.)

The answers to these exercises are on pp. A72–73.

### 2. THE STANDARD ERROR

Suppose 25 draws are made at random with replacement from the box

	0	2	3	4	6	
--	---	---	---	---	---	--

(There is nothing special about the numbers in the box; they were chosen to make later calculations come out evenly.) Each of the five tickets should appear on about one-fifth of the draws, that is, 5 times. So the sum should be around

$$5 \times 0 + 5 \times 2 + 5 \times 3 + 5 \times 4 + 5 \times 6 = 75.$$

That is the expected value for the sum. Of course, each ticket won't appear on exactly one-fifth of the draws, just as Kerrich didn't get heads on exactly half the tosses. The sum will be off the expected value by a chance error:

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

The chance error is the amount above (+) or below (−) the expected value. For example, if the sum is 70, the chance error is  $-5$ .

How big is the chance error likely to be? The answer is given by the *standard error*, usually abbreviated to SE.

A sum is likely to be around its expected value, but to be off by a chance error similar in size to the standard error.

There is a formula to use in computing the SE for a sum of draws made at random with replacement from a box. It is called the square root law, because it involves the square root of the number of draws. The statistical procedures in the rest of the book depend on this formula.<sup>2</sup>

*The square root law.* When drawing at random with replacement from a box of numbered tickets, the standard error for the sum of the draws is

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

The formula has two ingredients: the square root of the number of draws, and the SD of the list of numbers in the box (abbreviated to “SD of the box”). The SD measures the spread among the numbers in the box. If there is a lot of spread in the box, the SD is big, and it is hard to predict how the draws will turn out. So the standard error must be big too. Now for the number of draws. The sum of two draws is more variable than a single draw. The sum of 100 draws is still more variable. Each draw adds some extra variability to the sum, because you don’t know how it is going to turn out. As the number of draws goes up, the sum gets harder to predict, the chance errors get bigger, and so does the standard error. However, the standard error goes up slowly, by a factor equal to the square root of the number of draws. For instance, the sum of 100 draws is only  $\sqrt{100} = 10$  times as variable as a single draw.

The SD and the SE are different.<sup>3</sup> The SD applies to spread in lists of numbers. It is worked out using the method explained on p. 71. By contrast, the SE applies to chance variability—for instance, in the sum of the draws.

The SD is for a list

1 2 3 4 5 6

The SE is for a chance process



At the beginning of the section, we looked at the sum of 25 draws made at random with replacement from the box

0	2	3	4	6
---	---	---	---	---

The expected value for this sum is 75. The sum will be around 75, but will be off by a chance error. How big is the chance error likely to be? To find out, calculate the standard error. The average of the numbers in the box is 3. The deviations

from the average are

-3 -1 0 1 3

The SD of the box is

$$\begin{aligned}\sqrt{\frac{(-3)^2 + (-1)^2 + 0^2 + 1^2 + 3^2}{5}} &= \sqrt{\frac{9 + 1 + 0 + 1 + 9}{5}} \\ &= \sqrt{\frac{20}{5}} = 2.\end{aligned}$$

This measures the variability in the box. According to the square root law, the sum of 25 draws is more variable, by the factor  $\sqrt{25} = 5$ . The SE for the sum of 25 draws is  $5 \times 2 = 10$ . In other words, the likely size of the chance error is 10. And the sum of the draws should be around 75, give or take 10 or so. In general, the sum is likely to be around its expected value, give or take a standard error or so.

To show what this means empirically, we had the computer programmed to draw 25 times at random with replacement from the box 0 2 3 4 6. It got

0 0 4 4 0    4 3 2 6 2    2 0 2 6 2    6 4 2 6 3    0 3 6 4 0

The sum of these 25 draws is 71. This is 4 below the expected value, so the chance error is  $-4$ . The computer drew another 25 times and took the sum, getting 76. The chance error was  $+1$ . The third sum was 86, with a chance error of  $+11$ . In fact, we had the computer generate 100 sums, shown in table 1. These numbers are all around 75, the expected value. They are off by chance errors similar in size to 10, the standard error.

The sum of the draws is likely to be around \_\_\_\_\_, give or take \_\_\_\_\_ or so. The expected value for the sum fills in the first blank.  
The SE for the sum fills in the second blank.

Some terminology: the number 71 in table 1 is an *observed value* for the sum of the draws; the 76 is another observed value. All told, the table has 100 observed values for the sum. These observed values differ from the expected value of 75. The difference is chance error. For example, the chance error in 71 is  $-4$ , because  $71 - 75 = -4$ . The chance error in 76 is  $+1$ , because  $76 - 75 = 1$ . And so forth.

The observed values in table 1 show remarkably little spread around the expected value. In principle, they could be as small as 0, or as large as  $25 \times 6 = 150$ . However, all but one of them are between 50 and 100, that is, within 2.5 SEs of the expected value.

Observed values are rarely more than 2 or 3 SEs away from the expected value.

Table 1. Computer simulation: the sum of 25 draws made at random with replacement from the box  $\boxed{0} \boxed{2} \boxed{3} \boxed{4} \boxed{6}$ .

<i>Repe-tition</i>	<i>Sum</i>	<i>Repe-tition</i>	<i>Sum</i>	<i>Repe-tition</i>	<i>Sum</i>	<i>Repe-tition</i>	<i>Sum</i>
1	71	21	80	41	64	61	64
2	76	22	77	42	65	62	70
3	86	23	70	43	88	63	65
4	78	24	71	44	77	64	78
5	88	25	79	45	82	65	64
6	67	26	56	46	73	66	77
7	76	27	79	47	92	67	81
8	59	28	65	48	75	68	72
9	59	29	72	49	57	69	66
10	75	30	73	50	68	70	74
11	76	31	78	51	80	71	70
12	66	32	75	52	70	72	76
13	76	33	89	53	90	73	80
14	84	34	77	54	76	74	70
15	58	35	81	55	77	75	56
16	60	36	68	56	65	76	49
17	79	37	70	57	67	77	60
18	78	38	86	58	60	78	98
19	66	39	70	59	74	79	81
20	71	40	71	60	83	80	72
							100 62

### Exercise Set B

- One hundred draws are going to be made at random with replacement from the box  $\boxed{1} \boxed{2} \boxed{3} \boxed{4} \boxed{5} \boxed{6} \boxed{7}$ .
  - Find the expected value and standard error for the sum.
  - The sum of the draws will be around \_\_\_\_\_, give or take \_\_\_\_\_ or so.
  - Suppose you had to guess what the sum was going to be. What would you guess? Would you expect to be off by around 2, 4, or 20?
- You gamble 100 times on the toss of a coin. If it lands heads, you win \$1. If it lands tails, you lose \$1. Your net gain will be around \_\_\_\_\_, give or take \_\_\_\_\_ or so. Fill in the blanks, using the options  
 $-\$10 \quad -\$5 \quad \$0 \quad +\$5 \quad +\$10$
- The expected value for a sum is 50, with an SE of 5. The chance process generating the sum is repeated ten times. Which is the sequence of observed values?
  - 51, 57, 48, 52, 57, 61, 58, 41, 53, 48
  - 51, 49, 50, 52, 48, 47, 53, 50, 49, 47
  - 45, 50, 55, 45, 50, 55, 45, 50, 55, 45
- Fifty draws are made at random with replacement from the box  $\boxed{1} \boxed{2} \boxed{3} \boxed{4} \boxed{5}$ ; the sum of the draws turns out to be 157. The expected value for the sum is \_\_\_\_\_, the observed value is \_\_\_\_\_, the chance error is \_\_\_\_\_, and the standard error is \_\_\_\_\_. Fill in the blanks, and explain briefly.

5. Tickets are drawn at random with replacement from a box of numbered tickets. The sum of 25 draws has expected value equal to 50, and the SE is 10. If possible, find the expected value and SE for the sum of 100 draws. Or do you need more information?
6. One hundred draws are going to be made at random with replacement from the box  $\boxed{0 \ 2 \ 3 \ 4 \ 6}$ . True or false and explain.
- The expected value for the sum of the draws is 300.
  - The expected value for the sum of the draws is 300, give or take 20 or so.
  - The sum of the draws will be 300.
  - The sum of the draws will be around 300, give or take 20 or so.
7. In the simulation for table 1 (p. 293), if the computer kept on running, do you think it would eventually generate a sum more than 3 SEs away from the expected value? Explain.

*The answers to these exercises are on p. A73.*

### 3. USING THE NORMAL CURVE

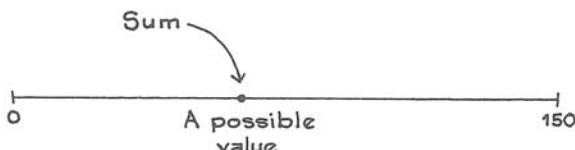
A large number of draws will be made at random with replacement from a box. What is the chance that the sum of the draws will be in a given range? Mathematicians discovered the normal curve while trying to solve problems of this kind. The logic behind the curve will be discussed in the next chapter. The object of this section is only to sketch the method, which applies whenever the number of draws is reasonably large. Basically, it is a matter of converting to standard units (using the expected value and standard error) and then working out areas under the curve, just as in chapter 5.

Now for an example. Suppose the computer is programmed to take the sum of 25 draws made at random with replacement from the magic box

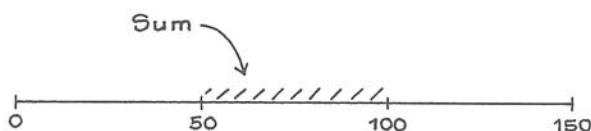
$\boxed{0 \ 2 \ 3 \ 4 \ 6}$

It prints out the result, repeating the process over and over again. About what percentage of the observed values should be between 50 and 100?

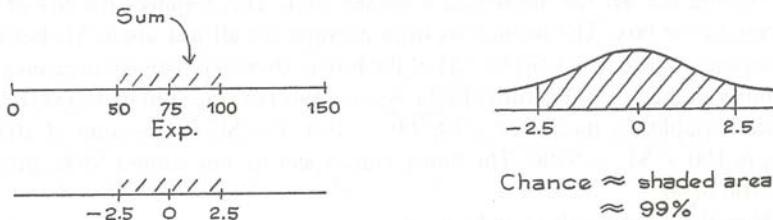
Each sum will be somewhere on the horizontal axis between 0 and  $25 \times 6 = 150$ .



The problem is asking for the chance that the sum will turn out to be between 50 and 100.



To find the chance, convert to standard units and use the normal curve. Standard units say how many SEs a number is away from the expected value.<sup>4</sup> In the example, 100 becomes 2.5 in standard units. The reason: the expected value for the sum is 75 and the SE is 10, so 100 is 2.5 SEs above the expected value. Similarly, 50 becomes  $-2.5$ .

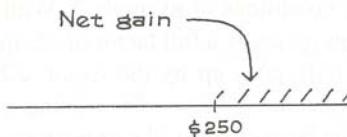


The interval from 50 to 100 is the interval within 2.5 SEs of the expected value, so the sum should be there about 99% of the time.

That finishes the calculation. Now for some data. Table 1 above reported 100 observed values for the sum: about 99 of them should be in the interval from 50 to 100, and in fact 99 of them are. To take some less extreme ranges, about 68% of the observed values should be in the interval from  $75 - 10$  to  $75 + 10$ . In fact, 73 are. Finally, about 95% of the observed values in table 1 should be in the range  $75 \pm 20$ , and 98 of them are. The theory looks pretty good. (Ranges include endpoints;  $\pm$  is read “plus-or-minus.”)

*Example 2.* In a month, there are 10,000 independent plays on a roulette wheel in a certain casino. To keep things simple, suppose the gamblers only stake \$1 on red at each play. Estimate the chance that the house will win more than \$250 from these plays.<sup>5</sup> (Red-or-black pays even money, and the house has 20 chances in 38 to win.)

*Solution.* The problem asks for the chance that the net gain of the house will be more than \$250.



The box model is the first thing. The box is

20 tickets	$+\$1$	18 tickets	$-\$1$
------------	--------	------------	--------

The net gain for the house is like the sum of 10,000 draws from this box.

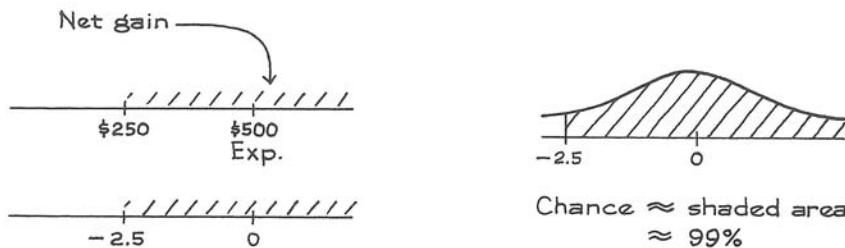
The expected value for the net gain is the number of draws times the average of the numbers in the box. The average is

$$\frac{\overbrace{\$1 + \dots + \$1}^{20 \text{ tickets}} - \overbrace{\$1 - \dots - \$1}^{18 \text{ tickets}}}{38} = \frac{\$20 - \$18}{38} = \frac{\$2}{38} \approx \$0.05$$

On the average, each draw adds around \$0.05 to the sum. The sum of 10,000 draws has an expected value of  $10,000 \times \$0.05 = \$500$ . The house averages about a nickel on each play, so in 10,000 plays it can expect to win around \$500. (The gambler and the house are on opposite sides of the box: 20 tickets are good for the house, and 18 are good for the gambler; see pp. 281–283.)

Finding the SE for the net gain comes next. This requires the SD of the numbers in the box. The deviations from average are all just about \$1, because the average is close to \$0. So the SD of the box is about \$1. This \$1 measures the variability in the box. According to the square root law, the sum of 10,000 draws is more variable, by the factor  $\sqrt{10,000} = 100$ . The SE for the sum of 10,000 draws is  $100 \times \$1 = \$100$ . The house can expect to win around \$500, give or take \$100 or so.

Now the normal curve can be used.



This completes the solution. The key idea: the net gain is like the sum of the draws from a box; that provided a logical basis for the square root law.

The house has about a 99% chance to win more than \$250. This may not seem like much, but you have to remember that the house owns many wheels, there often is a crowd of gamblers playing on each spin of each wheel, and a lot of bets are over a dollar. The house can expect to win about 5% of the money that crosses the table, and the square root law virtually eliminates the risk. For instance, suppose the house runs 25 wheels. To be very conservative, suppose each wheel operates under the conditions of example 2. With these assumptions, the casino's expected winnings go up by a full factor of 25, to  $25 \times \$500 = \$12,500$ . But their standard error only goes up by the factor  $\sqrt{25} = 5$ , to \$500. Now the casino can be virtually certain—99%—of winning at least \$11,000. For the casino, roulette is a volume business, just like groceries are for Safeway.

### Exercise Set C

- One hundred draws will be made at random with replacement from the box 

1	1	2	2	2	4
---	---	---	---	---	---

.  
 (a) The smallest the sum can be is \_\_\_\_\_, the largest is \_\_\_\_\_.  
 (b) The sum of the draws will be around \_\_\_\_\_, give or take \_\_\_\_\_ or so.  
 (c) The chance that the sum will be bigger than 250 is almost \_\_\_\_\_%.

2. One hundred draws will be made at random with replacement from the box  
 $\boxed{1} \boxed{3} \boxed{3} \boxed{9}$ .  
 (a) How large can the sum be? How small?  
 (b) How likely is the sum to be in the range from 370 to 430?
3. You can draw either 10 times or 100 times at random with replacement from the box  
 $\boxed{-1} \boxed{1}$ . How many times should you draw—  
 (a) To win \$1 when the sum is 5 or more, and nothing otherwise?  
 (b) To win \$1 when the sum is  $-5$  or less, and nothing otherwise?  
 (c) To win \$1 when the sum is between  $-5$  and 5, and nothing otherwise?

No calculations are needed, but explain your reasoning.

4. There are two options:
- One hundred draws will be made at random with replacement from the box  
 $\boxed{1} \boxed{1} \boxed{5} \boxed{7} \boxed{8} \boxed{8}$ .
  - Twenty-five draws will be made at random with replacement from the box  
 $\boxed{14} \boxed{17} \boxed{21} \boxed{23} \boxed{25}$ .

Which is better, if the payoff is—

- \$1 when the sum is 550 or more, and nothing otherwise?
  - \$1 when the sum is 450 or less, and nothing otherwise?
  - \$1 when the sum is between 450 and 550, and nothing otherwise?
5. Suppose that in one week at a certain casino, there are 25,000 independent plays at roulette. On each play, the gamblers stake \$1 on red. Is the chance that the casino will win more than \$1,000 from these 25,000 plays closest to 2%, 50%, or 98%? Explain briefly.
6. Suppose that one person stakes \$25,000 on one play at red-or-black in roulette. Is the chance that the casino will win more than \$1,000 from this play closest to 2%, 50%, or 98%? Explain briefly.
7. A gambler plays once at roulette, staking \$1,000 on each number (including 0 and 00). So this person has staked \$38,000 in all. What will happen? Explain briefly.
8. A box contains 10 tickets. Each ticket is marked with a whole number between  $-5$  and 5. The numbers are not all the same; their average equals 0. There are two choices:
- 100 draws are made from the box, and you win \$1 if the sum is between  $-15$  and 15.
  - 200 draws are made from the box, and you win \$1 if the sum is between  $-30$  and 30.

Choose one of the four options below; explain your answer.<sup>6</sup>

- A gives a better chance of winning.
- B gives a better chance of winning.
- A and B give the same chance of winning.
- Can't tell without the SD of the box.

*The answers to these exercises are on p. A74.*

#### 4. A SHORT-CUT

Finding SDs can be painful, but there is a short-cut for lists with only two different numbers, a big one and a small one.<sup>7</sup> (Each number can be repeated several times.)

When a list has only two different numbers (“big” and “small”), the SD equals

$$\left( \frac{\text{big number}}{\text{small number}} - 1 \right) \times \sqrt{\frac{\text{fraction with big number}}{\text{fraction with small number}}} \times \sqrt{\frac{1}{\text{fraction with big number}} \times \frac{3}{\text{fraction with small number}}}$$

For example, take the list 5, 1, 1, 1. The short-cut can be used because there are only two different numbers, 5 and 1. The SD is

$$(5 - 1) \times \sqrt{\frac{1}{4} \times \frac{3}{4}} \approx 1.73$$

The short-cut involves much less arithmetic than finding the root-mean-square of the deviations from average (p. 71), and gives exactly the same answer. The short-cut is helpful in many gambling problems (and in other contexts too).

*Example 3.* A gambler plays roulette 100 times, staking \$1 on the number 10 each time. The bet pays 35 to 1, and the gambler has 1 chance in 38 to win. Fill in the blanks: the gambler will win \$\_\_\_\_\_, give or take \$\_\_\_\_\_ or so.

*Solution.* The first thing to do is to make a box model for the net gain. (See example 1 on pp. 283–284.) The gambler’s net gain is like the sum of 100 draws made at random with replacement from

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

What is the expected net gain? This is 100 times the average of the box. The average of the numbers in the box is their total, divided by 38. The winning ticket contributes \$35 to the total, while the 37 losing tickets take away \$37 in all. So the average is

$$\frac{\$35 - \$37}{38} = \frac{-\$2}{38} \approx -\$0.05$$

In 100 plays, the expected net gain is

$$100 \times (-\$0.05) = -\$5$$

In other words, the gambler expects to lose about \$5 in 100 plays.

The next step is to find the SE for the sum of the draws: this is  $\sqrt{100}$  times the SD of the box. The short-cut can be used, and the SD of the box equals

$$[\$35 - (-\$1)] \times \sqrt{\frac{1}{38} \times \frac{37}{38}} \approx \$36 \times 0.16 \approx \$5.76$$

The SE for the sum of the draws is  $\sqrt{100} \times \$5.76 \approx \$58$ .

The gambler will lose about \$5, give or take \$58 or so. This completes the solution. The large SE gives the gambler a reasonable chance of winning, and that is the attraction. Of course, on average the gambler loses; and the SE also means that the gambler can lose a bundle.

### Exercise Set D

1. Does the formula give the SD of the list? Explain.

List	Formula
(a) 7, 7, 7, -2, -2	$5 \times \sqrt{3/5 \times 2/5}$
(b) 0, 0, 0, 0, 5	$5 \times \sqrt{1/5 \times 4/5}$
(c) 0, 0, 1	$\sqrt{2/3 \times 1/3}$
(d) 2, 2, 3, 4, 4, 4	$2 \times \sqrt{1/6 \times 2/6 \times 3/6}$

2. Suppose a gambler bets a dollar on a single number at Keno (example 1 on p. 289). In 100 plays, the gambler's net gain will be \$\_\_\_\_\_, give or take \$\_\_\_\_\_. or so.
3. At Nevada roulette tables, the "house special" is a bet on the numbers 0, 00, 1, 2, 3. The bet pays 6 to 1, and there are 5 chances in 38 to win.
- (a) For all other bets at Nevada roulette tables, the house expects to make about 5 cents out of every dollar put on the table. How much does it expect to make per dollar on the house special?
  - (b) Someone plays roulette 100 times, betting a dollar on the house special each time. Estimate the chance that this person comes out ahead.
4. A gambler plays roulette 100 times. There are two possibilities:
- (i) Betting \$1 on a section each time (see figure 3 on p. 282).
  - (ii) Betting \$1 on red each time.

A section bet pays 2 to 1, and there are 12 chances in 38 to win. Red pays even money, and there are 18 chances in 38 to win. True or false, and explain:

- (a) The chance of coming out ahead is the same with (i) and (ii).
- (b) The chance of winning more than \$10 is bigger with (i).
- (c) The chance of losing more than \$10 is bigger with (i).

*The answers to these exercises are on pp. A74–75.*

### 5. CLASSIFYING AND COUNTING

Some chance processes involve counting. The square root law can be used to get the standard error for a count, but the box model has to be set up correctly. The next example will show how to do this.

*Example 4.* A die is rolled 60 times.

(a) The total number of spots should be around \_\_\_\_\_, give or take \_\_\_\_\_ or so.

(b) The number of 6's should be around \_\_\_\_\_, give or take \_\_\_\_\_ or so.

By way of illustration, table 2 shows the results of throwing a die 60 times: the first throw was a 4, the second was a 5, and so on.

Table 2. Sixty throws of a die.

4	5	5	2	4	5	3	2	6	3	5	4	6	2	6	4	4	2	5	6
1	5	3	1	2	2	1	2	5	3	3	6	6	1	1	5	1	6	1	2
4	4	2	1	4	4	5	2	6	3	2	4	6	1	6	4	6	1	5	2

*Solution.* Part (a) is familiar. It involves adding. Each throw contributes some number of spots, and we add these numbers up. The total number of spots in 60 throws of the die is like the sum of 60 draws from the box

1	2	3	4	5	6
---	---	---	---	---	---

The average of this box is 3.5 and the SD is 1.71. The expected value for the sum is  $60 \times 3.5 = 210$ ; the SE for the sum is  $\sqrt{60} \times 1.71 \approx 13$ . The total number of spots will be around 210, give or take 13 or so. In fact, the sum of the numbers in table 2 is 212. The sum was off its expected value by around one-sixth of an SE.

Part (b). Filling in the first blank is easy. Each of the six faces should come up on about one-sixth of the throws, so the expected value for the number of 6's is  $60 \times 1/6 = 10$ . The second blank is harder. We need a new kind of box because the sum of the draws from | 1 2 3 4 5 6 | is no longer relevant. Instead of being added, each throw of the die is classified: is it a 6, or not? (There are only two classes here, 6's on one hand, everything else on the other.) Then, the number of 6's is counted up.

The point to notice is that on each throw, the number of 6's either goes up by 1, or stays the same:

- 1 is added to the count if the throw is 6;
- 0 is added to the count if the throw is anything else.

The count has 1 chance in 6 to go up by one, and 5 chances in 6 to stay the same. Therefore, on each draw, the sum must have 1 chance in 6 to go up by one, and 5 chances in 6 to stay the same. The right box to use is

x <sup>0</sup>	x <sup>1</sup>				
----------------	----------------	----------------	----------------	----------------	----------------

As far as the chances are concerned, the number of 6's in 60 throws of the die is just like the sum of 60 draws from the new box. This puts us in a position to use the square root law.

The new box has five  $\boxed{0}$ 's and a  $\boxed{1}$ . The SD is  $\sqrt{1/6 \times 5/6} \approx 0.37$ , by the short-cut method. And the SE for the sum of the draws is  $\sqrt{60} \times 0.37 \approx 3$ . In 60 throws of a die, the number of 6's will be around 10, give or take 3 or so. In fact, in table 2 there were eleven 6's. The observed number of 6's was off its expected value by a third of an SE. This completes the example. It's the old story, for a new box.

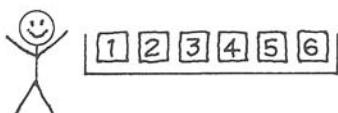
This example makes a general point. Although they may look quite different, many problems about chance processes can be solved in the same way. In these problems, some tickets are drawn at random from a box. An operation is performed on the draws, and the problem asks for the chance that the result will be in a given interval. In this chapter, there are two possible operations on the draws:

- adding,
- classifying and counting.

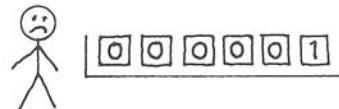
The message is that both operations can be treated the same way—provided you change the box.

If you have to classify and count the draws, put 0's and 1's on the tickets. Mark 1 on the tickets that count for you, 0 on the others.

For adding up the draws,  
the box is



For counting 6's,  
the box is



*Remember to change the tickets!*

*Example 5.* A coin will be tossed 100 times. Find the expected value and standard error for the number of heads. Estimate the chance of getting between 40 and 60 heads.

*Solution.* The first thing is to make a box model. The problem involves classifying the tosses as heads or tails, and then counting the number of heads. So there should be only 0's and 1's in the box. The chances are 50–50 for heads, so the box should be  $\boxed{0} \boxed{1}$ . The number of heads in 100 tosses of a coin is like the sum of 100 draws made at random with replacement from the box  $\boxed{0} \boxed{1}$ . (The coin is even simpler than the die in example 4: each toss either pushes the number of heads up by 1 or leaves it alone, with a 50–50 chance; likewise, each draw from the box either pushes the sum up by 1 or leaves it alone, with the same 50–50 chance.) This completes the model.

Since the number of heads is like the sum of the draws, the square root law can be used. The SD of the box is  $1/2$ . So the SE for the sum of 100 draws is  $\sqrt{100} \times 1/2 = 5$ . The number of heads will be around 50, give or take 5 or so.

The range from 40 to 60 heads represents the expected value, give or take 2 SEs. And the chance is around 95%. This completes the solution.

To interpret this 95% chance, imagine counting the number of heads in 100 tosses of a coin. You might get 44 heads. Toss again: you might get 54 heads. A third time, the number would change once more, perhaps to 48 heads. And so on. In the long run, about 95% of these counts would come out in the range from 40 to 60. John Kerrich actually did this experiment. Table 3 shows the results, with Kerrich's 10,000 tosses broken down into successive groups of one hundred. In fact, 95 out of 100 groups had 40 to 60 heads (inclusive). The theory looks good.

Table 3. Kerrich's coin tossing experiment, showing the number of heads he got in each successive group of 100 tosses.

<i>Group of 100 tosses</i>	<i>No. of heads</i>						
1–100	44	2,501–2,600	44	5,001–5,100	42	7,501–7,600	48
101–200	54	2,601–2,700	34	5,101–5,200	68	7,601–7,700	43
201–300	48	2,701–2,800	59	5,201–5,300	45	7,701–7,800	58
301–400	53	2,801–2,900	50	5,301–5,400	37	7,801–7,900	57
401–500	56	2,901–3,000	51	5,401–5,500	47	7,901–8,000	48
501–600	57	3,001–3,100	51	5,501–5,600	52	8,001–8,100	45
601–700	56	3,101–3,200	48	5,601–5,700	51	8,101–8,200	50
701–800	45	3,201–3,300	56	5,701–5,800	49	8,201–8,300	53
801–900	45	3,301–3,400	57	5,801–5,900	48	8,301–8,400	46
901–1,000	44	3,401–3,500	50	5,901–6,000	37	8,401–8,500	56
1,001–1,100	40	3,501–3,600	54	6,001–6,100	47	8,501–8,600	58
1,101–1,200	54	3,601–3,700	47	6,101–6,200	52	8,601–8,700	54
1,201–1,300	53	3,701–3,800	53	6,201–6,300	45	8,701–8,800	49
1,301–1,400	55	3,801–3,900	50	6,301–6,400	48	8,801–8,900	48
1,401–1,500	52	3,901–4,000	53	6,401–6,500	44	8,901–9,000	45
1,501–1,600	54	4,001–4,100	52	6,501–6,600	51	9,001–9,100	55
1,601–1,700	58	4,101–4,200	54	6,601–6,700	55	9,101–9,200	51
1,701–1,800	50	4,201–4,300	55	6,701–6,800	53	9,201–9,300	48
1,801–1,900	53	4,301–4,400	52	6,801–6,900	52	9,301–9,400	56
1,901–2,000	42	4,401–4,500	51	6,901–7,000	60	9,401–9,500	55
2,001–2,100	56	4,501–4,600	53	7,001–7,100	50	9,501–9,600	55
2,101–2,200	53	4,601–4,700	54	7,101–7,200	57	9,601–9,700	50
2,201–2,300	53	4,701–4,800	47	7,201–7,300	49	9,701–9,800	48
2,301–2,400	45	4,801–4,900	42	7,301–7,400	46	9,801–9,900	59
2,401–2,500	52	4,901–5,000	44	7,401–7,500	62	9,901–10,000	52

It is time to connect the square root law and the law of averages. Suppose a coin is tossed a large number of times. Then heads will come up on about half the tosses:

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

How big is the chance error likely to be? At first, Kerrich's assistant thought it would be very small. The record showed him to be wrong. As Kerrich kept tossing the coin, the chance error grew in absolute terms but shrank relative to the number of tosses, just as the mathematics predicts. (See figures 1 and 2, pp. 275–276.)

According to the square root law, the likely size of the chance error is  $\sqrt{\text{number of tosses}} \times 1/2$ . For instance, with 10,000 tosses the standard error is  $\sqrt{10,000} \times 1/2 = 50$ . When the number of tosses goes up to 1,000,000, the standard error goes up too, but only to 500—because of the square root. As the number of tosses goes up, the SE for the number of heads gets bigger and bigger in absolute terms, but smaller and smaller relative to the number of tosses. That is why the percentage of heads gets closer and closer to 50%. The square root law is the mathematical explanation for the law of averages.

### Exercise Set E

1. A coin is tossed 16 times.
  - (a) The number of heads is like the sum of 16 draws made at random with replacement from one of the following boxes. Which one and why?
    - (i)  head  tail
    - (ii)  0  1
    - (iii)  0  1  1
  - (b) The number of heads will be around \_\_\_\_\_, give or take \_\_\_\_\_ or so.
2. One hundred draws are made at random with replacement from  1  2  3  4  5. What is the chance of getting between 8 and 32 tickets marked "5"?
3. According to the simplest genetic model, the sex of a child is determined at random, as if by drawing a ticket at random from the box
 

male  female

What is the chance that of the next 2,500 births (not counting twins or other multiple births), more than 1,275 will be females?

4. This exercise and the next are based on Kerrich's coin-tossing experiment (table 3, p. 302). For example, in tosses 1–100, the observed number of heads was 44, the expected number was 50, so the chance error was  $44 - 50 = -6$ . Fill in the blanks.

<i>Group of 100 tosses</i>	<i>Observed value</i>	<i>Expected value</i>	<i>Chance error</i>	<i>Standard error</i>
1–100	44	50	-6	—
101–200	54	50	—	—
201–300	48	—	—	—
301–400	—	—	—	—

5. How many of the counts in table 3 on p. 302 should be in the range 45 to 55? How many are? (Endpoints included.)

6. (a) A coin is tossed 10,000 times. What is the chance that the number of heads will be in the range 4,850 to 5,150?  
 (b) A coin is tossed 1,000,000 times. What is the chance that the number of heads will be in the range 498,500 to 501,500?
7. Fifty draws are made at random with replacement from the box  $\boxed{0} \boxed{0} \boxed{1} \boxed{1} \boxed{1}$ ; there are 33  $\boxed{1}$ 's among the draws. The expected number of  $\boxed{1}$ 's is \_\_\_\_\_, the observed number is \_\_\_\_\_, the chance error is \_\_\_\_\_, and the SE is \_\_\_\_\_.
8. A computer program is written to do the following job. There is a box with ten blank tickets. You tell the program what numbers to write on the tickets, and how many draws to make. Then, the computer will draw that many tickets at random with replacement from the box, add them up, and print out the sum—but not the draws. This program does not know anything about coin tossing. Still, you can use it to simulate the number of heads in 1,000 tosses of a coin. How?
9. A die is rolled 100 times. Someone figures the expected number of aces as  $100 \times 1/6 = 16.67$ , and the SE as  $\sqrt{100} \times \sqrt{1/6 \times 5/6} \approx 3.73$ . (An ace is  $\boxed{\bullet}$ .) Is this right? Answer yes or no, and explain.

*The answers to these exercises are on p. A75.*

## 6. REVIEW EXERCISES

1. One hundred draws will be made at random with replacement from the box  $\boxed{1} \boxed{6} \boxed{7} \boxed{9} \boxed{9} \boxed{10}$ .
- (a) How small can the sum of the draws be? How large?  
 (b) The sum is between 650 and 750 with a chance of about  
 $1\%$      $10\%$      $50\%$      $90\%$      $99\%$
- Explain.
2. A gambler plays roulette 100 times, betting a dollar on a column each time. The bet pays 2 to 1, and there are 12 chances in 38 to win. Fill in the blanks; show work.
- (a) In 100 plays, the gambler's net gain will be around \$\_\_\_\_\_, give or take \$\_\_\_\_\_.  
 (b) In 100 plays, the gambler should win \_\_\_\_\_ times, give or take \_\_\_\_\_ or so.  
 (c) How does the column bet compare with betting on a single number at Keno (example 1 on p. 289)?
3. Match the lists with the SDs. Explain your reasoning
- |                   |  |
|-------------------|--|
| (a) 1, -2, -2     | (i) $\sqrt{1/3 \times 2/3}$            |
| (b) 15, 15, 16    | (ii) $2 \times \sqrt{1/3 \times 2/3}$  |
| (c) -1, -1, -1, 1 | (iii) $3 \times \sqrt{1/3 \times 2/3}$ |
| (d) 0, 0, 0, 1    | (iv) $\sqrt{1/4 \times 3/4}$           |
| (e) 0, 0, 2       | (v) $2 \times \sqrt{1/4 \times 3/4}$   |

4. A large group of people get together. Each one rolls a die 180 times, and counts the number of  $\boxed{\square}$ 's. About what percentage of these people should get counts in the range 15 to 45?
5. A die will be thrown some number of times, and the object is to guess the total number of spots. There is a one-dollar penalty for each spot that the guess is off. For instance, if you guess 200 and the total is 215, you lose \$15. Which do you prefer: 50 throws, or 100? Explain.
6. One hundred draws are made at random with replacement from the box  $\boxed{1} \boxed{1} \boxed{2} \boxed{3}$ . The draws come out as follows: 45  $\boxed{1}$ 's, 23  $\boxed{2}$ 's, 32  $\boxed{3}$ 's. For each number below, find the phrase which describes it.

<i>Number</i>	<i>Phrase</i>
12	observed value for the sum of the draws
45	observed value for the number of 3's
187	observed value for the number of 1's
25	expected value for the sum of the draws
50	expected value for the number of 3's
175	expected value for the number of 1's
5	chance error in the sum of the draws
32	standard error for the number of 1's

7. One hundred draws are made at random with replacement from the box  $\boxed{1} \boxed{2} \boxed{3} \boxed{4} \boxed{5} \boxed{6}$ .
- If the sum of the draws is 321, what is their average?
  - If the average of the draws is 3.78, what is the sum?
  - Estimate the chance that the average of the draws is between 3 and 4.
8. A coin is tossed 100 times.
- The difference "number of heads – number of tails" is like the sum of 100 draws from one of the following boxes. Which one, and why?
    - $\boxed{\text{heads}} \quad \boxed{\text{tails}}$
    - $\boxed{-1} \quad \boxed{1}$
    - $\boxed{-1} \quad \boxed{0}$
    - $\boxed{0} \quad \boxed{1}$
    - $\boxed{-1} \quad \boxed{0} \quad \boxed{1}$
  - Find the expected value and standard error for the difference.
9. A gambler plays roulette 1,000 times. There are two possibilities:
- Betting \$1 on a column each time.
  - Betting \$1 on a number each time.
- A column pays 2 to 1, and there are 12 chances in 38 to win; a number pays 35 to 1, and there is 1 chance in 38 to win. True or false and explain:

- (a) The chance of coming out ahead is the same with (i) and (ii).  
 (b) The chance of winning more than \$100 is bigger with (ii).  
 (c) The chance of losing more than \$100 is bigger with (ii).
10. A box contains numbered tickets. Draws are made at random with replacement from the box. Below are three statements about this particular box; (i) and (ii) are true. Is (iii) true or false? Explain.
- (i) For a certain number of draws, the expected value for the sum of the draws equals 400.
  - (ii) For that same number of draws, there is about a 75% chance that the sum will be between 350 and 450.
  - (iii) For twice that number of draws, there is about a 75% chance that the sum will be between 700 and 900.
11. One hundred draws are made at random with replacement from the box  $\boxed{-2} \boxed{-1} \boxed{0} \boxed{1} \boxed{3}$ . The sum of the positive numbers will be around \_\_\_\_\_, give or take \_\_\_\_\_ or so.
12. One hundred draws are made at random with replacement from the box  $\boxed{1} \boxed{2} \boxed{3} \boxed{4} \boxed{5} \boxed{6} \boxed{7}$ .
- (a) The sum of the draws is 431. The expected value for the sum of the draws is \_\_\_\_\_, the observed value is \_\_\_\_\_, the chance error is \_\_\_\_\_, and the standard error is \_\_\_\_\_.
  - (b) The sum of the draws is 386. The expected value for the sum of the draws is \_\_\_\_\_, the observed value is \_\_\_\_\_, the chance error is \_\_\_\_\_, and the standard error is \_\_\_\_\_.
  - (c) The sum of the draws is 417. The expected value for the sum of the draws is \_\_\_\_\_, the observed value is \_\_\_\_\_, the chance error is \_\_\_\_\_, and the standard error is \_\_\_\_\_.
13. A letter is drawn 1,000 times, at random, from the word A R A B I A. There are two offers.
- (A) You win a dollar if the number of A's among the draws is 10 or more above the expected number.
  - (B) You win a dollar if the number of B's among the draws is 10 or more above the expected number.
- Choose one option and explain.
- (i) A gives a better chance of winning than B.
  - (ii) A and B give the same chance of winning.
  - (iii) B gives better chance of winning than A.
  - (iv) There is not enough information to decide.
14. In roulette, once in a while, someone will bet \$1 on red; and, at the same time, someone else will bet \$1 on black (p. 282). Suppose this pair of bets is made 100 times in the course of an evening.
- (a) The house will make money on \_\_\_\_\_ of the 100 pairs of bets, give or take \_\_\_\_\_ or so.

- (b) The net gain for the house from the 100 pairs of bets will be around \_\_\_\_\_ give or take \_\_\_\_\_ or so.

## 7. POSTSCRIPT

The exercises of this chapter teach a melancholy lesson. The more you gamble, the more you lose. The basic reason is that all the bets are unfair, in the sense that your expected net gain is negative. So the law of averages works for the house, not for you. Of course, this chapter only discussed simple strategies, and gamblers have evolved complicated systems for betting on roulette, craps, and the like. But it is a theorem of mathematics that no system for compounding unfair bets can ever make your expected net gain positive. In proving this theorem, only two assumptions are needed: (i) you aren't clairvoyant, and (ii) your financial resources are finite. The game of blackjack is unusual. Under some circumstances there are bets with a positive expected net gain.<sup>8</sup> As a result, people have won a lot of money on blackjack. However, the casinos change the rules to make this harder and harder.

## 8. SUMMARY

1. An *observed value* should be somewhere around the *expected value*; the difference is chance error. The likely size of the chance error is given by the *standard error*. For instance, the sum of the draws from a box will be around the expected value, give or take a standard error or so.

2. When drawing at random with replacement from a box of numbered tickets, each draw adds to the sum an amount which is around the average of the box. So the expected value for the sum is

$$(\text{number of draws}) \times (\text{average of box}).$$

3. When drawing at random with replacement from a box of numbered tickets,

$$\text{SE for sum} = \sqrt{\text{number of draws}} \times (\text{SD of box}).$$

This is the *square root law*.

4. When a list has only two different numbers ("big" and "small"), the SD can be figured by a short-cut method:

$$\left( \frac{\text{big number} - \text{small number}}{\text{big number} + \text{small number}} \right)^2 \times \sqrt{\frac{\text{fraction with big number}}{\text{big number}} \times \frac{\text{fraction with small number}}{\text{small number}}}$$

5. If you have to classify and count the draws, remember to put 1 on the tickets that count for you, 0 on the others.

6. Provided the number of draws is sufficiently large, the normal curve can be used to figure chances for the sum of the draws.

# 18

## The Normal Approximation for Probability Histograms

*Everybody believes in the [normal approximation], the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact.*

—G. LIPPmann (FRENCH PHYSICIST, 1845–1921)

### 1. INTRODUCTION

According to the law of averages, when a coin is tossed a large number of times, the percentage of heads will be close to 50%. Around 1700, the Swiss mathematician James Bernoulli put this on a rigorous mathematical footing. Twenty years later, Abraham de Moivre made a substantial improvement on Bernoulli's work, by showing how to compute the chance that the percentage of heads will fall in any given interval around 50%. The computation is not exact, but the approximation gets better and better as the number of tosses goes up. (De Moivre's work was discussed before, in chapter 13.)

Bernoulli and de Moivre both made the same assumptions about the coin: the tosses are independent, and on each toss the coin is as likely to land heads as tails. From these assumptions, it follows that the coin is as likely to land in any specific pattern of heads and tails as in any other. What Bernoulli did was to show that for most patterns, about 50% of the entries are heads.

You can see this starting to happen even with 5 tosses. Imagine tossing the coin 5 times, and keeping a record of how it lands on each toss. There is one possible pattern with 5 heads: H H H H H. How many patterns are there with

four heads? The answer is 5:

T H H H H    H T H H H    H H T H H    H H H T H    H H H H T

The pattern T H H H H, for instance, means that the coin landed tails on the first toss, then gave four straight heads. Table 1 shows how many patterns there are, for any given number of heads. With 5 tosses, there are altogether  $2^5 = 32$  possible patterns in which the coin can land. And 20 patterns out of the 32 have nearly half heads (two or three out of five).

Table 1. The number of patterns corresponding to a given number of heads, in 5 tosses of a coin.

<i>Number of heads</i>	<i>Number of patterns</i>
zero	1
one	5
two	10
three	10
four	5
five	1

De Moivre managed to count, to within a small margin of error, the number of patterns having a given number of heads—for any number of tosses. With 100 tosses, the number of patterns he had to think about is  $2^{100}$ . This is quite a large number. If you tried to write all these patterns out, it might be possible to get a hundred of them on a page the size of this one. By the time you finished writing, you would have enough books to fill a shelf reaching from the earth to the farthest known star.

Still and all, mathematicians have a formula for the number of patterns with exactly 50 heads:

$$\frac{100!}{50! \times 50!} = \frac{100 \times 99 \times \dots \times 51}{50 \times 49 \times \dots \times 1}.$$

(Binomial coefficients are covered in chapter 15; they won't really matter here.)

The formula was of no immediate help to de Moivre, because the arithmetic is nearly impossible to do by hand. By calculator,<sup>1</sup>

$$\frac{100 \times 99 \times \dots \times 51}{50 \times 49 \times \dots \times 1} \approx 1.01 \times 10^{29}.$$

Similarly, the total number of patterns is  $2^{100} \approx 1.27 \times 10^{30}$ . So the chance of getting exactly 50 heads in 100 tosses of a coin is

$$\frac{\text{number of patterns with 50 heads}}{\text{total number of patterns}} \approx \frac{1.01 \times 10^{29}}{1.27 \times 10^{30}} \approx 0.08 = 8\%.$$

Of course, de Moivre did not have anything like a modern calculator available. He needed a mathematical way of estimating the binomial coefficients, without having to work the arithmetic out. And he found a way to do it (though the approximation is usually credited to another mathematician, James Stirling). De Moivre's procedure led him to the normal curve. For example, he found that the chance of getting exactly 50 heads in 100 tosses of a coin was about equal to

the area under the normal curve between  $-0.1$  and  $+0.1$ . In fact, he was able to prove that the whole *probability histogram* for the number of heads is close to the normal curve when the number of tosses is large. Modern researchers have extended this result to the sum of draws made at random from any box of tickets. The details of de Moivre's argument are too complicated to go into here—but we can present his idea graphically, using a computer to draw the pictures.<sup>2</sup>

## 2. PROBABILITY HISTOGRAMS

When a chance process generates a number, the expected value and standard error are a guide to where that number will be. But the probability histogram gives a complete picture.

A probability histogram is a new kind of graph. This graph represents chance, not data.

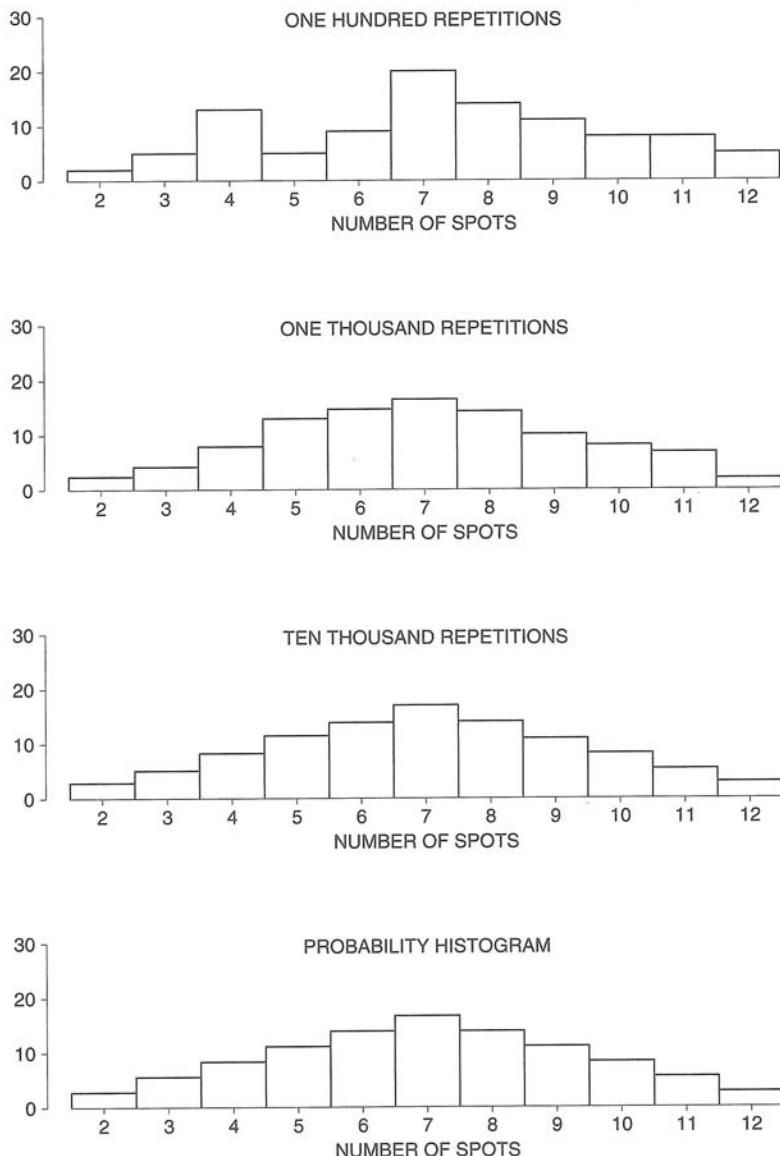
Here is an example. Gamblers playing craps bet on the total number of spots shown by a pair of dice. (The numbers range from 2 through 12.) So the odds depend on the chance of rolling each possible total. To find the chances, a casino might hire someone to throw a pair of dice. This experiment was simulated on the computer; results for the first 100 throws are shown in table 2.

Table 2. Rolling a pair of dice. The computer simulated rolling a pair of dice, and finding the total number of spots. It repeated this process 10,000 times. The first 100 repetitions are shown in the table.

<i>Repe-</i> <i>tition</i>	<i>Total</i>								
1	8	21	10	41	8	61	8	81	11
2	9	22	4	42	10	62	5	82	9
3	7	23	8	43	6	63	3	83	7
4	10	24	7	44	3	64	11	84	4
5	9	25	7	45	4	65	9	85	7
6	5	26	3	46	8	66	4	86	4
7	5	27	8	47	4	67	12	87	7
8	4	28	8	48	4	68	7	88	6
9	4	29	12	49	5	69	10	89	7
10	4	30	2	50	4	70	4	90	11
11	10	31	11	51	11	71	7	91	6
12	8	32	12	52	8	72	4	92	11
13	3	33	12	53	10	73	7	93	8
14	11	34	7	54	9	74	9	94	8
15	7	35	7	55	10	75	9	95	7
16	8	36	6	56	12	76	11	96	9
17	9	37	6	57	7	77	6	97	10
18	8	38	2	58	6	78	9	98	5
19	6	39	6	59	7	79	9	99	7
20	8	40	3	60	7	80	7	100	7

The top panel in figure 1 shows the histogram for the data in table 2. The total of 7 came up 20 times, so the rectangle over 7 has an area of 20%, and similarly for the other possible totals. The next panel shows the empirical histogram for the first 1,000 repetitions, and the third is for all 10,000. These empirical his-

Figure 1. Empirical histograms converging to a probability histogram. The computer simulated rolling a pair of dice and finding the total number of spots. It repeated the process 100 times, and made a histogram for the 100 numbers (top panel). This is an empirical histogram—based on observations. The second panel is for 1,000 repetitions, the third panel for 10,000. (Each repetition involves rolling a pair of dice.) The bottom panel is the ideal or probability histogram for the total number of spots when a pair of dice are rolled.



tograms converge to the ideal probability histogram shown in the bottom panel of the figure. (*Empirical* means “experimentally observed,” *converge* means “gets closer and closer to.”)

Of course, this probability histogram can be computed using a theoretical argument. As shown in chapter 14, there are 6 chances in 36 of rolling a 7. That’s  $16\frac{2}{3}\%$ . Consequently, the area of the rectangle over 7 in the probability histogram equals  $16\frac{2}{3}\%$ . Similarly for the other rectangles.

A probability histogram represents chance by area.

The probability histogram (bottom panel, figure 1) is made up of rectangles. The base of each rectangle is centered at a possible value for the sum of the draws, and the area of the rectangle equals the chance of getting that value.<sup>3</sup> The total area of the histogram is 100%.

For another example, look at the product of the numbers on a pair of dice, instead of the sum. The computer was programmed to repeat the following chance process over and over again: roll a pair of dice and take the product of the numbers. The top panel of figure 2 gives the empirical histogram for 100 repetitions. The product 10 came up 4 times, so the area of the rectangle over 10 equals 4%. Other values are done the same way. The second panel gives the empirical histogram for 1,000 repetitions; the third, for 10,000. (Each repetition involves rolling a pair of dice and taking the product.) The last panel shows the probability histogram. The empirical histogram for 10,000 repetitions looks almost exactly like the probability histogram.

Figure 2 is very different from figure 1: the new histograms have gaps. To see why, it helps to think about the possible values of the product. The smallest value is 1, if both dice show  $\square$ ; the biggest is 36, if both show  $\blacksquare$ . But there is no way to get the product 7. There is no rectangle over 7, because the chance is zero. For the same reason, there is no rectangle over 11. All the gaps can be explained in this way.

### Exercise Set A

1. The figure below is a probability histogram for the sum of 25 draws from the box  $\boxed{1} \boxed{2} \boxed{3} \boxed{4} \boxed{5}$ . The shaded area represents the chance that the sum will be between \_\_\_\_\_ and \_\_\_\_\_ (inclusive).

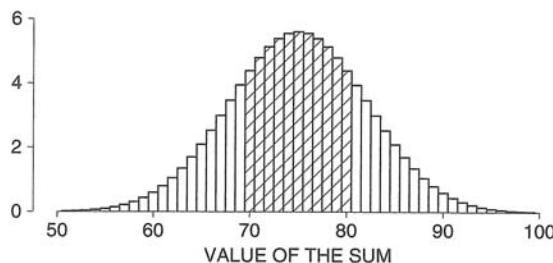
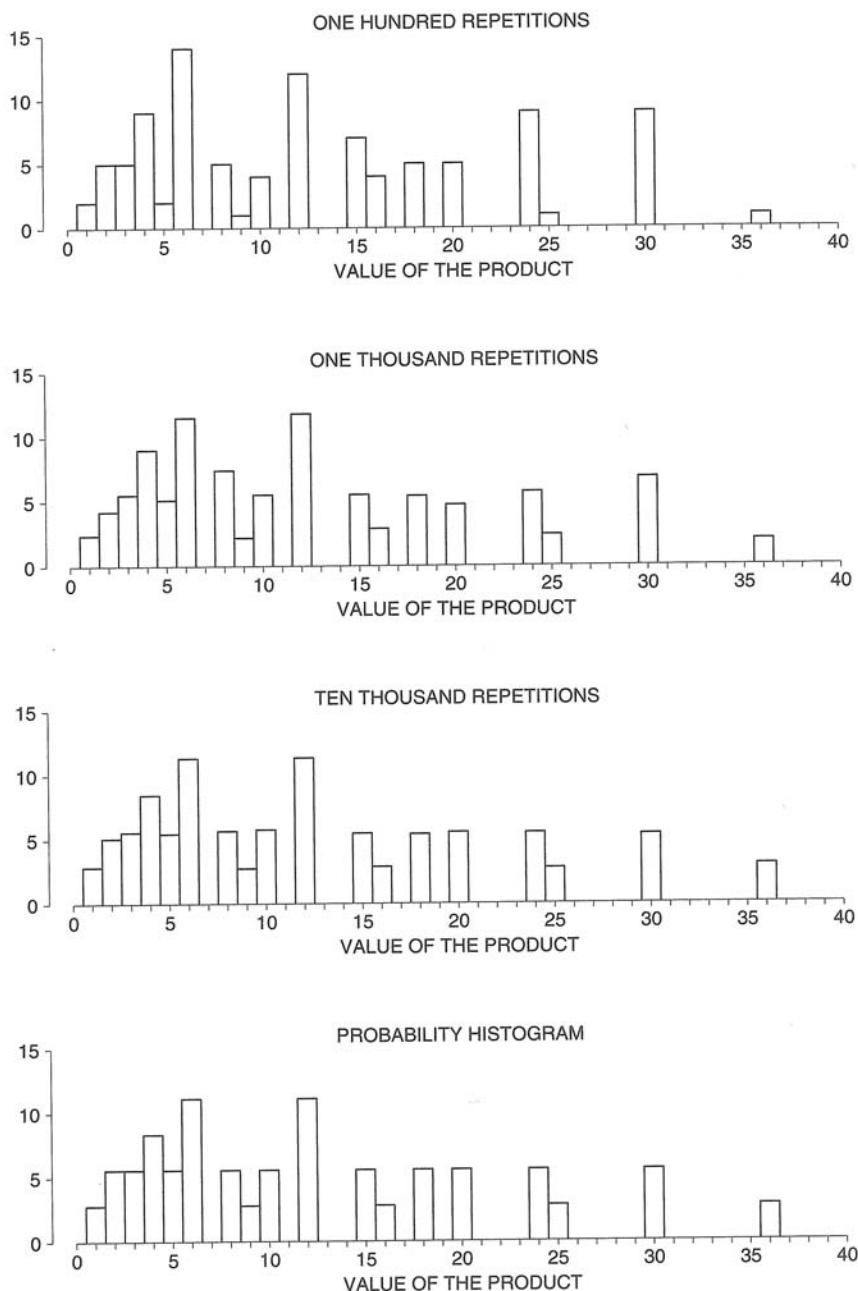


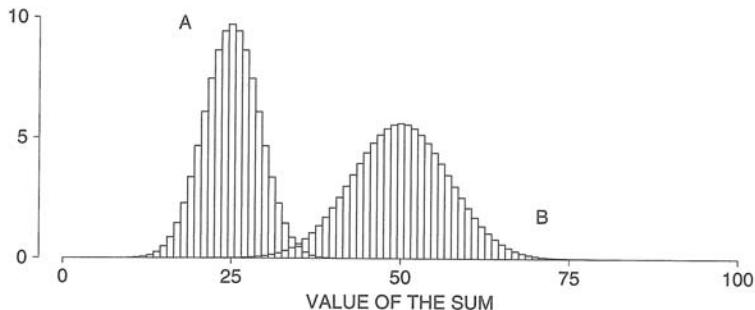
Figure 2. Empirical histograms converging to a probability histogram. The computer simulated rolling a pair of dice and taking the product of the two numbers. It repeated the process 100 times, and made a histogram for the 100 products (top panel). This is an empirical histogram—based on observations. The second panel is for 1,000 repetitions, the third panel for 10,000. (Each repetition involves rolling a pair of dice.) The bottom panel is the ideal or probability histogram for the product of the two numbers that come up when a pair of dice are rolled.



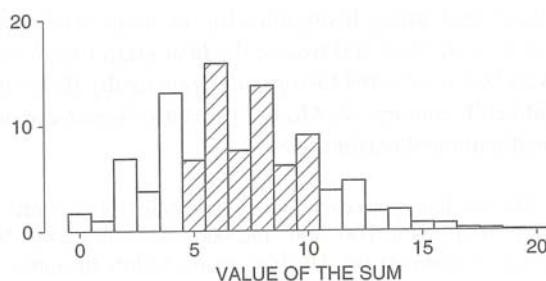
2. The bottom panel of figure 1 (p. 311) shows the probability histogram for the total number of spots when a pair of dice are rolled.
- The chance that the total number of spots will be between 7 and 10 (inclusive) equals the area under the histogram between \_\_\_\_\_ and \_\_\_\_\_.
  - The chance that the total number of spots will be 7 equals the area under the histogram between \_\_\_\_\_ and \_\_\_\_\_.
3. This exercise—like exercise 2—refers to figure 1 on p. 311.
- If a pair of dice are rolled, the total number of spots is most likely to be \_\_\_\_\_.
  - In 1,000 rolls of the pair of dice, which total came up most often?
  - In the top panel of figure 1, the rectangle over 4 is bigger than the rectangle over 5. Is this because 4 is more likely than 5? Explain.
  - Look at the top panel of the figure. The area of the rectangle above 8 represents—
    - the chance of getting a total of 8 spots when a pair of dice are rolled.
    - the chance of getting a total of 8 spots when 100 dice are rolled.
    - the percentage of times the total of 8 comes up in table 2.
- Choose one option, and explain.
4. Figure 2 on p. 313 is about the product of the numbers on a pair of dice.
- If the dice land  $\boxed{\bullet} \boxed{\circ}$ , what is the product? If they land  $\boxed{\circ} \boxed{\bullet}$ ?
  - “2 is as likely a value for the product as 3.” Which panel should you look at to check this statement? Is it true?
  - In 1,000 rolls, which value appeared more often for the product: 2 or 3? Explain.
  - None of the histograms has a rectangle above 14. Why?
  - In the bottom panel of figure 2, the area of the rectangle above 6 is 11.1%. What does this 11.1% represent?

5. The figure below shows the probability histograms for the sum of 25 draws made at random with replacement from boxes (i) and (ii). Which histogram goes with which box? Explain.

(i)  $\boxed{0} \boxed{1} \boxed{2}$       (ii)  $\boxed{0} \boxed{1} \boxed{2} \boxed{3} \boxed{4}$



6. The figure at the top of the next page is the probability histogram for the sum of 25 draws made at random with replacement from a box. True or false: the shaded area represents the percentage of times you draw a number between 5 and 10 inclusive.



The answers to these exercises are on p. A76.

### 3. PROBABILITY HISTOGRAMS AND THE NORMAL CURVE

The object of this section is to show how the probability histogram for the number of heads gets close to the normal curve when the number of tosses becomes large. For instance, suppose the coin is tossed 100 times. The probability histogram for the number of heads is a bit jagged, but follows the normal curve quite well (figure 3).

The figure has two horizontal axes. The probability histogram is drawn relative to the upper axis, showing the number of heads. The normal curve is drawn relative to the lower axis, showing standard units. The expected number of heads is 50, and the SE is 5. So 50 on the number-of-heads axis corresponds to 0 on the standard-units axis, 55 corresponds to +1, and so on.

There are also two vertical axes in the figure. The probability histogram is drawn relative to the inside one, showing percent per head. The normal curve is drawn relative to the outside one, showing percent per standard unit. To see how the scales match up, take the top value on each axis. Why does 50% per standard unit match up with 10% per head? The SE is 5, so there are 5 heads to the standard unit. And  $50/5 = 10$ . Any other pair of values can be dealt with in the same way. (Also see p. 80 on data histograms.)

Figure 3. The probability histogram for the number of heads in 100 tosses of a coin, compared to the normal curve. The curve is drawn on the standard-units scale for the histogram.

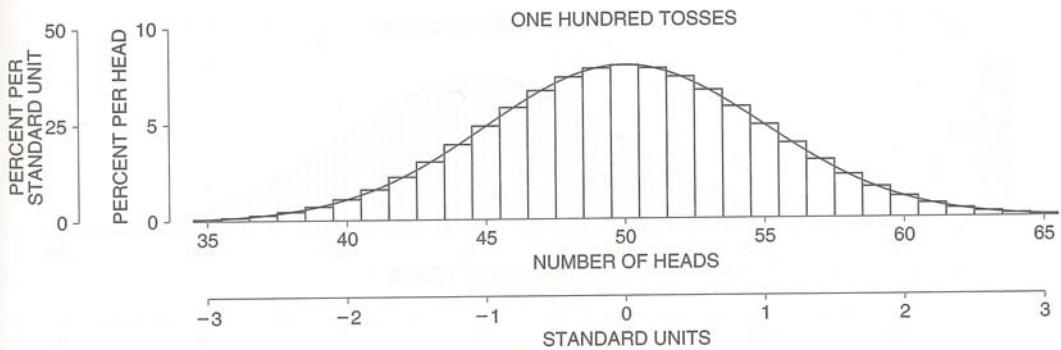
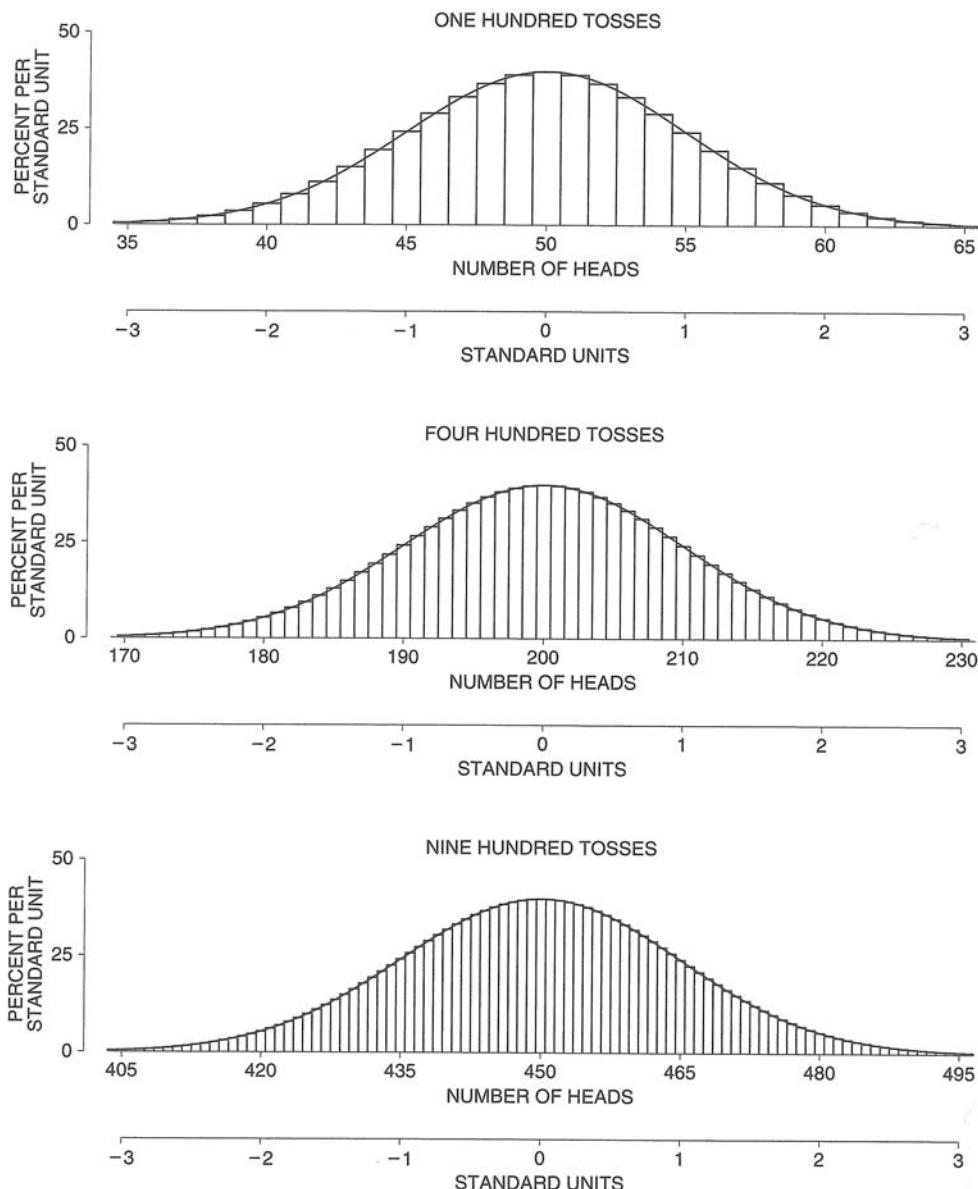


Figure 4 shows probability histograms for the number of heads in 100, 400, and 900 tosses of a coin. With 100 tosses, the histogram follows the curve but is more jagged. With 900 tosses, the histogram is practically the same as the curve. In the early eighteenth century, de Moivre proved this convergence had to take place, by pure mathematical reasoning.

Figure 4. The normal approximation. Probability histograms are shown for the number of heads in 100, 400, and 900 tosses of a coin. The normal curve is shown for comparison. The histograms follow the curve better and better as the number of tosses goes up.



#### 4. THE NORMAL APPROXIMATION

The normal curve has already been used in chapter 17 to figure chances. This section will explain the logic. It will also present a technique for taking care of endpoints, which should be used when the number of tosses is small or extra accuracy is wanted.

*Example 1.* A coin will be tossed 100 times. Estimate the chance of getting—

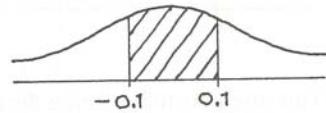
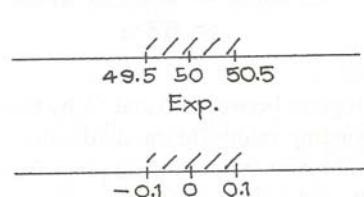
- (a) exactly 50 heads.
- (b) between 45 and 55 heads inclusive.
- (c) between 45 and 55 heads exclusive.

*Solution.* The expected number of heads is 50 and the standard error is 5, as shown on p. 301.

*Part (a).* Look at figure 3 (p. 315). The chance of getting exactly 50 heads equals the area of the rectangle over 50. The base of this rectangle goes from 49.5 to 50.5 on the number-of-heads scale. In standard units, the base of the rectangle goes from -0.1 to 0.1:

$$\frac{49.5 - 50}{5} = -0.1, \quad \frac{50.5 - 50}{5} = 0.1$$

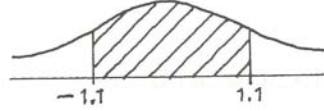
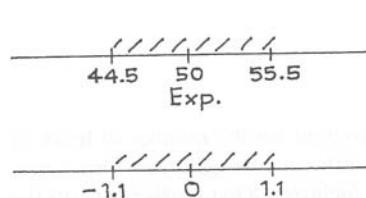
But the histogram and the normal curve almost coincide. So the area of the rectangle is nearly equal to the area between -0.1 and 0.1 under the curve.



Chance  $\approx$  shaded area  
 $\approx 7.97\%$

(The exact chance is 7.96%, to two decimals; the approximation is excellent.<sup>4</sup>)

*Part (b).* The chance of getting between 45 and 55 heads inclusive equals the area of the eleven rectangles over the values 45 through 55 in figure 3. That is the area under the histogram between 44.5 and 55.5 on the number-of-heads scale, which correspond to -1.1 and 1.1 on the standard-units scale. Because the histogram follows the normal curve so closely, this area is almost equal to the area under the curve.



Chance  $\approx$  shaded area  
 $\approx 72.87\%$

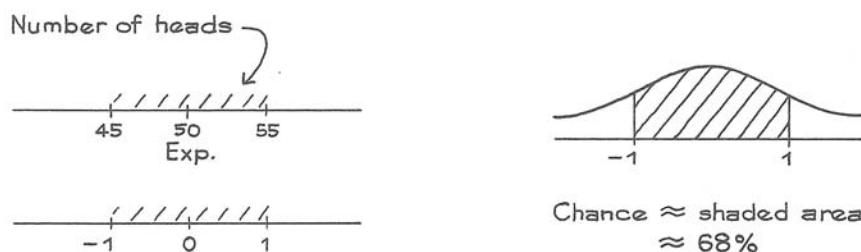
(The exact chance is 72.87%, to two decimals.)

*Part (c).* The chance of getting 45 to 55 heads exclusive equals the total area of the nine rectangles over the values 46 through 54. That is the area under the histogram between 45.5 and 54.5 on the number-of-heads scale, which correspond to  $-0.9$  and  $0.9$  on the standard-units scale.



(The exact chance is 63.18%, to two decimals.)

Often, the problem will only ask for the chance that (for instance) the number of heads is between 45 and 55, without specifying whether endpoints are included or excluded. Then, you can use the compromise procedure:

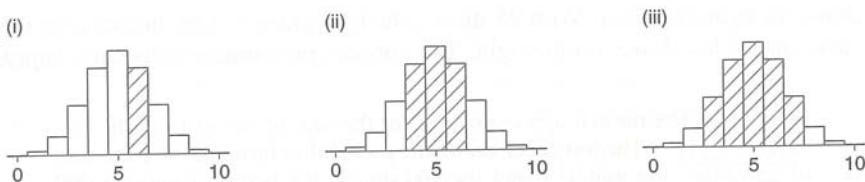


This amounts to replacing the area under the histogram between 45 and 55 by the area under the normal curve between the corresponding values (in standard units). It splits the two end rectangles in half, and does not give quite as much precision as the method used in example 1. Keeping track of the endpoints has an official name—"the continuity correction." The correction is worthwhile if the rectangles are big, or if a lot of precision is needed. Usually, the exercises in this book can be worked without the correction.

The normal approximation consists in replacing the actual probability histogram by the normal curve before computing areas. This is legitimate when the probability histogram follows the normal curve. Probability histograms are often hard to work out, while areas under the normal curve are easy to look up in the table.<sup>5</sup>

### Exercise Set B

1. A coin is tossed 10 times. The probability histogram for the number of heads is shown at the top of the next page, with three different shaded areas. One corresponds to the chance of getting 3 to 7 heads inclusive. One corresponds to the chance of getting 3 to 7 heads exclusive. And one corresponds to the chance of getting exactly 6 heads. Which is which, and why?



2. In figure 3 on p. 315, the chance of getting 52 heads is exactly equal to the area between \_\_\_\_\_ and \_\_\_\_\_ under the \_\_\_\_\_. Fill in the blanks. For the last one, your options are: normal curve, probability histogram. Explain your answers.
3. A coin is tossed 100 times. Estimate the chance of getting 60 heads.
4. Kerrich's data on 10,000 tosses of a coin can be taken in successive groups of 100 tosses (table 3 on p. 302). About how many groups should show exactly 60 heads? How many actually do?
5. A coin is tossed 10,000 times. Estimate the chance of getting—
  - (a) 4,900 to 5,050 heads
  - (b) 4,900 heads or fewer
  - (c) 5,050 heads or more
6. (a) Suppose you were going to estimate the chance of getting 50 heads or fewer in 100 tosses of a coin. Should you keep track of the edges of the rectangles?  
 (b) Same, for the chance of getting 450 heads or fewer in 900 tosses.

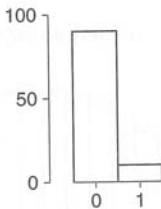
No calculations are needed, just look at figure 4 on p. 316.

*The answers to these exercises are on pp. A76–77.*

## 5. THE SCOPE OF THE NORMAL APPROXIMATION

In the preceding section, the discussion has been about a coin, which lands heads or tails with chance 50%. What about drawing from a box? Again, the normal approximation works perfectly well, so long as you remember one thing. The more the histogram of the numbers in the box differs from the normal curve, the more draws are needed before the approximation takes hold. Figure 5 shows the histogram for the tickets in the lopsided box [9 [0]'s [1]].

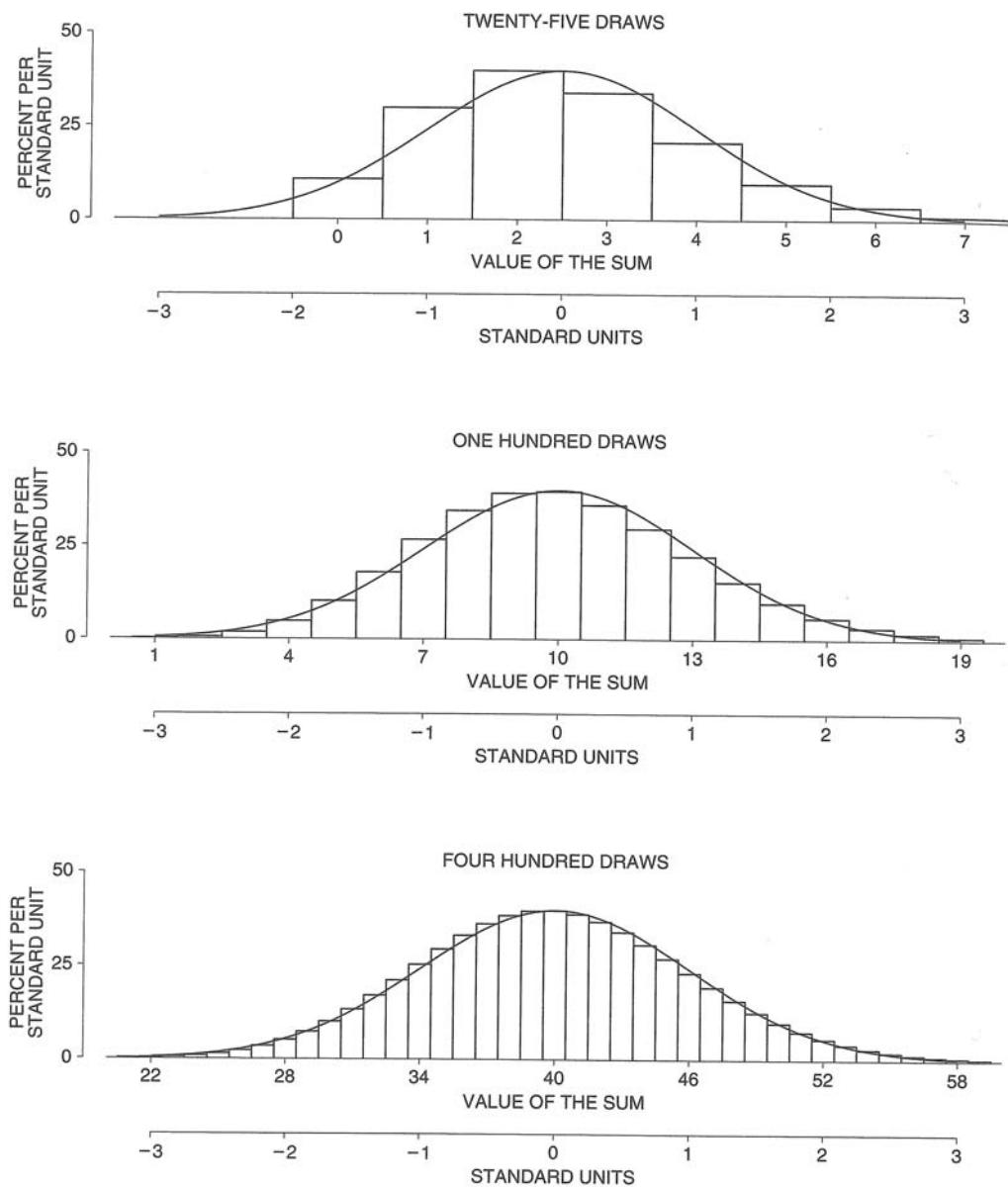
Figure 5. Histogram for the lopsided box [9 [0]'s [1]].



The probability histogram for the sum will be lopsided too, until the number of draws gets fairly large. The computer was programmed to work out the probability histogram for the sum of 25, 100, or 400 draws from the box. The histograms are

shown in figure 6 below. With 25 draws, the histogram is a lot higher than the curve on the left, lower on the right. The normal approximation does not apply.

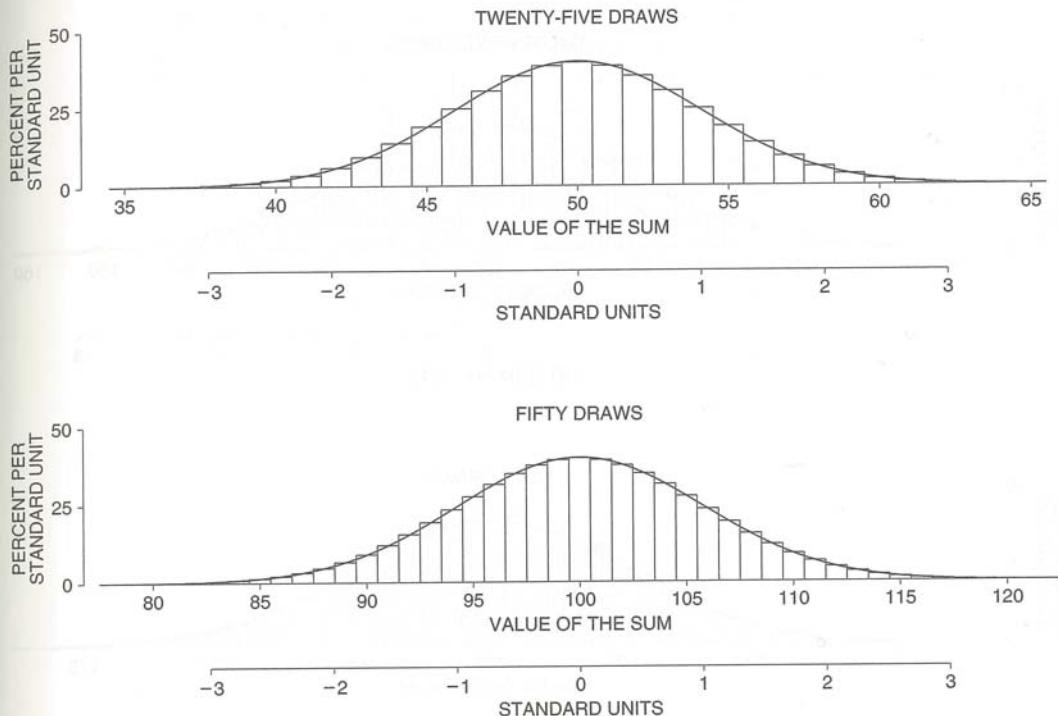
Figure 6. The normal approximation for the sum of draws from the box  $[9 \square 0]'s [1]$ . The top panel shows the probability histogram for the sum of 25 draws, the middle panel for 100 draws, the bottom panel for 400 draws. A normal curve is shown for comparison. The histograms are higher than the normal curve on the left and lower on the right, because the box is lopsided.<sup>6</sup> As the number of draws goes up, the histograms follow the curve more and more closely.



With 100 draws, the histogram follows the curve much better. At 400 draws, you have to look closely to see the difference.

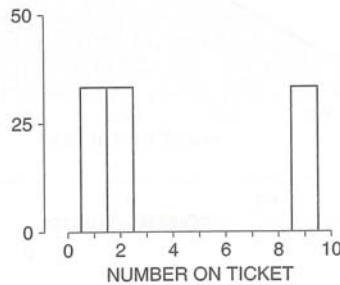
So far, there have only been 0's and 1's in the box. What about other numbers? Our next example is  $\boxed{1} \boxed{2} \boxed{3}$ . The probability histogram for the sum of 25 draws from this box is already close to the curve; with 50 draws, the histogram follows the curve very closely indeed (figure 7).

Figure 7. Probability histograms for the sum of 25 or 50 draws from the box  $\boxed{1} \boxed{2} \boxed{3}$ . These histograms follow the normal curve very well.



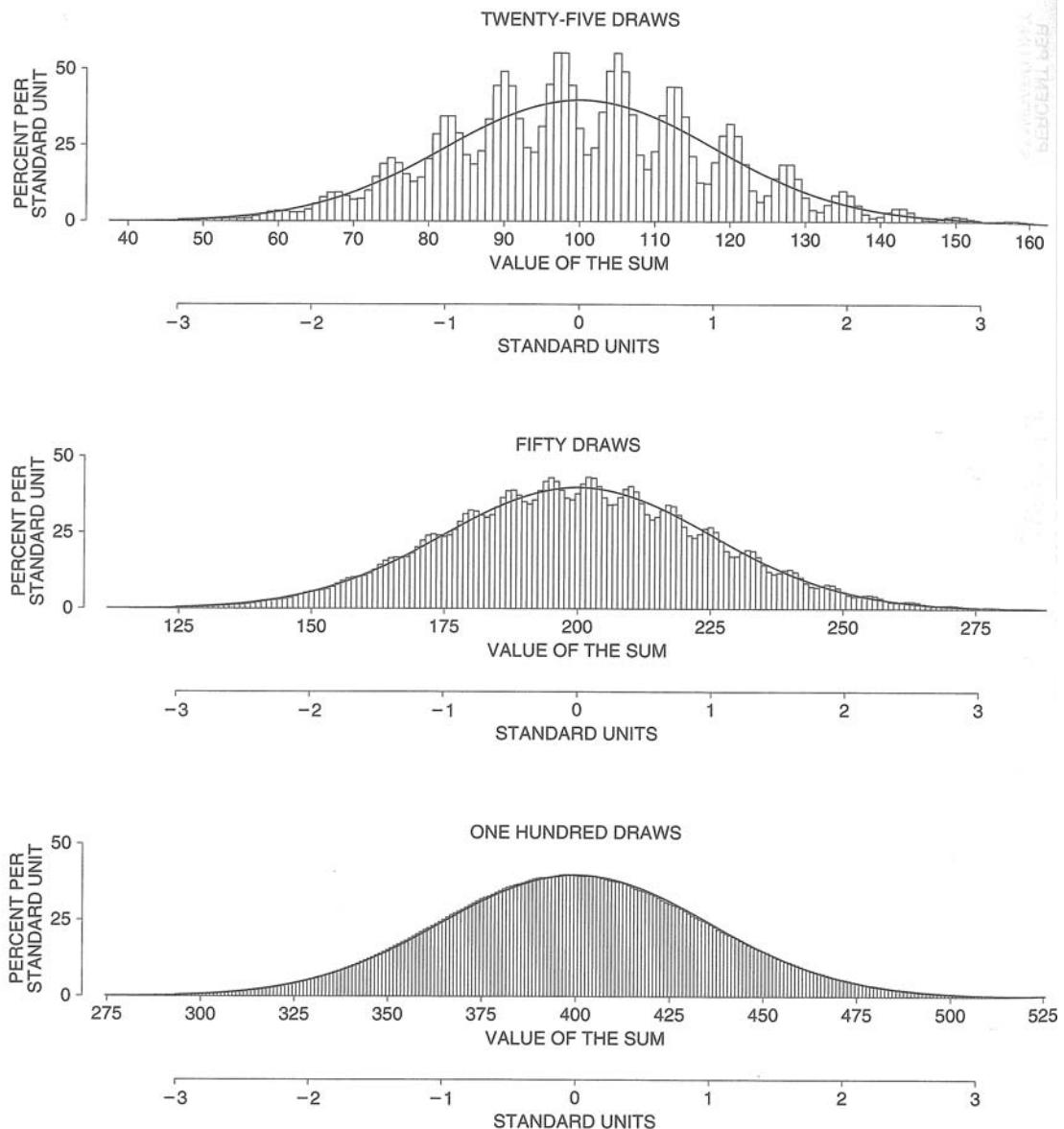
Our last example is the box  $\boxed{1} \boxed{2} \boxed{9}$ . A histogram for the numbers in the box is shown in figure 8. This histogram looks nothing like the normal curve.

Figure 8. Histogram for the box  $\boxed{1} \boxed{2} \boxed{9}$ . The histogram is nothing like the normal curve.



With 25 draws, the probability histogram for the sum is still quite different from the curve—it shows waves (figure 9). With 50 draws, the waves are still there, but much smaller. And by 100 draws, the probability histogram is indistinguishable from the curve.

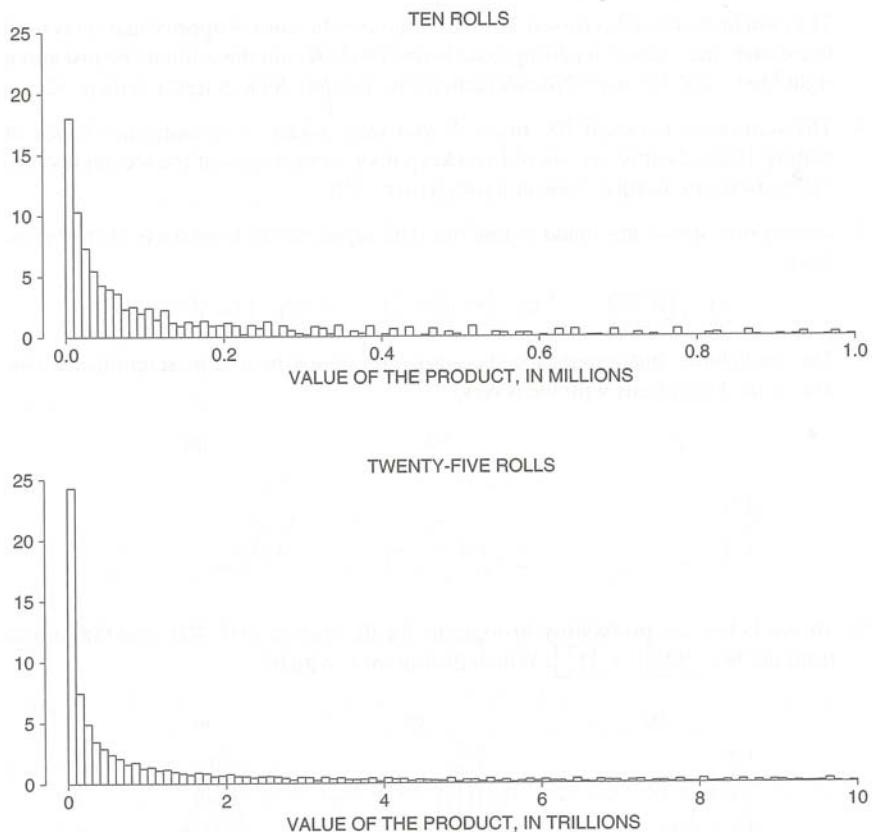
Figure 9. The normal approximation for a sum. Probability histograms are shown for the sum of draws from the box  $\boxed{1} \boxed{2} \boxed{9}$ . The top panel is for 25 draws, and does not follow the normal curve especially well. (Note the waves.<sup>7</sup>) The middle panel is for 50 draws. The bottom panel is for 100 draws. It follows the normal curve very well.



The normal curve is tied to sums. For instance, the probability histogram for a product will usually be quite different from normal. The top panel of figure 10 shows the probability histogram for the product of 10 rolls of a die. This is nothing like the normal curve. Making the number of rolls larger does not make the histogram more normal: the probability histogram for the product of 25 rolls is shown in the bottom panel, and is even worse.<sup>8</sup> Multiplication is different from addition. The normal approximation works for the sum of draws made at random from a box—not for the product.

With 10 rolls, the histogram for the product is shown out to a million; 6% of the area lies beyond that point and is not shown. A million looks like a big number, but products build up fast. The largest value for the product is 6 multiplied by itself 10 times:  $6^{10} = 60,466,176$ . On this scale, a million is not so big after all.

Figure 10. Probability histograms for the product of 10 and 25 rolls of a die. The histograms look nothing like the normal curve. The base of each rectangle covers a range of values of the product, and the area of the rectangle equals the chance of the product taking a value in that range. With 10 rolls, about 6% of the area is not shown; with 25 rolls, about 20% is not shown. In the top panel, the vertical scale is percent per 10,000; in the bottom panel, percent per  $10^{11}$ .

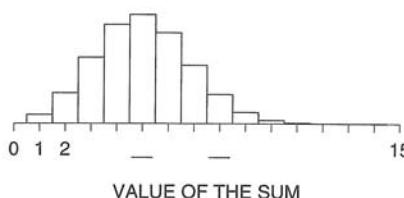


With 25 rolls, the largest possible value for the product really is a big number:  $6^{25} \approx 3 \times 10^{19}$ , or 3 followed by 19 zeros. (The U.S. federal debt was “only” \$8 trillion in 2006, that is, \$8 followed by 12 zeros.)

### Exercise Set C

1. Shown below is the probability histogram for the sum of 15 draws from the box  $\boxed{0} \boxed{0} \boxed{1}$ .

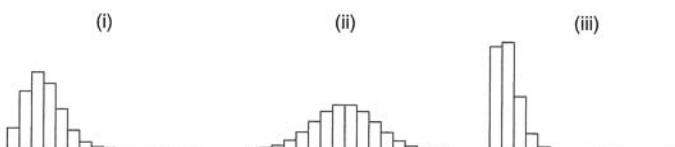
- (a) What numbers go into the blanks?
- (b) Which is a more likely value for the sum, 3 or 8? Explain.



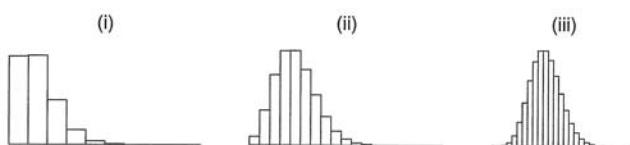
2. A biased coin has one chance in ten of landing heads. It is tossed 400 times. Estimate the chance of getting exactly 40 heads.
3. The coin in exercise 2 is tossed 25 times. Suppose the normal approximation is used to estimate the chance of getting exactly one head. Would the estimate be just about right? too high? too low? No calculations are needed; look at figure 6 on p. 320.
4. The same coin is tossed 100 times. If you were asked to estimate the chance of getting 10 heads or fewer, should you keep track of the edges of the rectangles? No calculations are needed; look at figure 6 on p. 320.
5. Twenty-five draws are made at random with replacement from each of the boxes below.

- A)  $\boxed{0} \boxed{1}$       B)  $\boxed{9} \boxed{0}'s \boxed{1}$       C)  $\boxed{24} \boxed{0}'s \boxed{1}$

The probability histograms for the sums are shown below, in scrambled order. Match the histograms with the boxes.



6. Shown below are probability histograms for the sum of 100, 400, and 900 draws from the box  $\boxed{99} \boxed{0}'s \boxed{1}$ . Which histogram is which?



7. This exercise refers to the top panel of figure 9 (p. 322), which shows the probability histogram for the sum of 25 draws from the box  $\boxed{1} \boxed{2} \boxed{9}$ . The chance that the sum is 100 equals (i) the area between 99.5 and 100.5 under the probability histogram? Or is it (ii) the area under the normal curve between 99.5 in standard units and 100.5 in standard units? Choose one option, and explain.
8. This exercise, like the previous one, can be worked using the top panel of figure 9. Among the options listed below, the sum of 25 draws from the box  $\boxed{1} \boxed{2} \boxed{9}$  is most likely to equal \_\_\_\_\_ and least likely to equal \_\_\_\_\_ even though its expected value is \_\_\_\_\_. Options:

100    101    102    103    104    105

9. This exercise refers to the top panel of figure 10 on p. 323.
- (a) The expected value for the product is nearly 276,000. The chance that the product will exceed this number is—  
 just about 50%    much bigger than 50%    much smaller than 50%
- Choose one option, and explain.
- (b) There are 100 rectangles in the histogram. The width of each one is  
 1    10    100    1,000    10,000    100,000    1,000,000
- (c) Which is a more likely range for the product?  
 390,000–400,000                  400,000–410,000

*The answers to these exercises are on pp. A77–78.*

## 6. CONCLUSION

We have looked at the sum of the draws from four different boxes:

$\boxed{0} \boxed{1}$      $\boxed{9} \boxed{0}'s \boxed{1}$      $\boxed{1} \boxed{2} \boxed{3}$      $\boxed{1} \boxed{2} \boxed{9}$

There are plenty more where those came from. But the pattern is always the same. With enough draws, the probability histogram for the sum will be close to the normal curve. Mathematicians have a name for this fact. They call it “the central limit theorem,” because it plays a central role in statistical theory.<sup>9</sup>

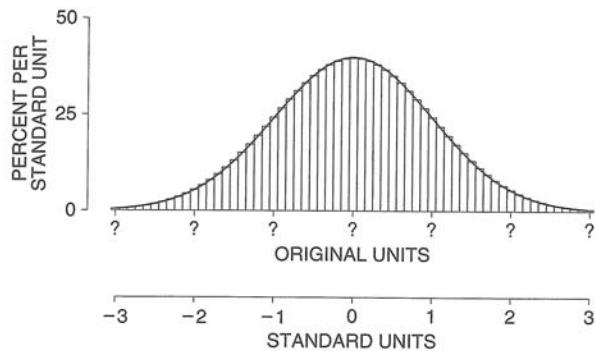
*The Central Limit Theorem.* When drawing at random with replacement from a box, the probability histogram for the sum will follow the normal curve, even if the contents of the box do not. The histogram must be put into standard units, and the number of draws must be reasonably large.

The central limit theorem applies to sums but not to other operations like products (figure 10). The theorem is the basis for many of the statistical procedures discussed in the rest of the book.

How many draws do you need? There is no set answer. Much depends on the contents of the box—remember the waves in figure 9. However, for many boxes,

the probability histogram for the sum of 100 draws will be close enough to the normal curve.

When the probability histogram does follow the normal curve, it can be summarized by the expected value and standard error. For instance, suppose you had to plot such a histogram without any further information. In standard units you can do it, at least to a first approximation:



To finish the picture, you have to translate the standard units back into original units by filling in the question marks. This is what the expected value and standard error do. They tell you almost all there is to know about this histogram, because it follows the normal curve.

The expected value pins the center of the probability histogram to the horizontal axis, and the standard error fixes its spread.

According to the square root law, the expected value and standard error for a sum can be computed from

- the number of draws,
- the average of the box,
- the SD of the box.

These three quantities just about determine the behavior of the sum. That is why the SD of the box is such an important measure of its spread.<sup>10</sup>

This chapter discussed two sorts of convergence for histograms, and it is important to separate them. In figure 1, the number of draws from the box [1] [2] [3] [4] [5] [6] was fixed. It was 2. The basic chance process was drawing from the box and taking the sum. This process was repeated a larger and larger number of times—100, 1,000, 10,000. The empirical histogram for the observed values of the sum (a histogram for data) converged to the probability histogram (a histogram for chances). In section 5, on the other hand, the number of draws

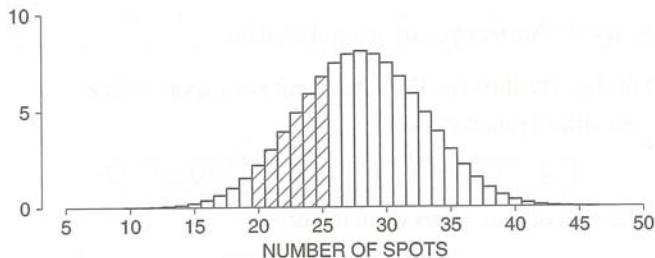
from the box got larger and larger. Then the probability histogram for the sum got smoother and smoother, and in the limit became the normal curve. Empirical histograms are one thing; probability histograms quite another.

In part II of the book, the normal curve was used for data. In some cases, this can be justified by a mathematical argument which uses the two types of convergence discussed in this chapter. When the number of repetitions is large, the empirical histogram will be close to the probability histogram. When the number of draws is large, the probability histogram for the sum will be close to the normal curve. Consequently, when the number of repetitions and the number of draws are both large, the empirical histogram for the sums will be close to the curve.<sup>11</sup> This is all a matter of pure logic: a mathematician can prove every step.

But there is still something missing. It has to be shown that the process generating the data is like drawing numbers from a box and taking the sum. This sort of argument will be discussed in part VII. More than mathematics is involved—there will be questions of fact to settle.

## 7. REVIEW EXERCISES

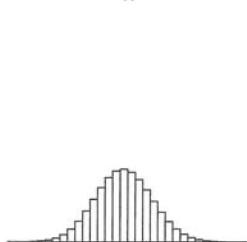
1. The figure below shows the probability histogram for the total number of spots when a die is rolled eight times. The shaded area represents the chance that the total will be between \_\_\_\_\_ and \_\_\_\_\_ (inclusive).



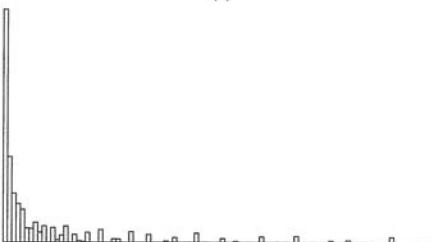
2. Four hundred draws will be made at random with replacement from the box  $\boxed{1} \boxed{3} \boxed{5} \boxed{7}$ .
  - (a) Estimate the chance that the sum of the draws will be more than 1,500.
  - (b) Estimate the chance that there will be fewer than 90  $\boxed{3}$ 's.
3. Ten draws are going to be made at random with replacement from the box  $\boxed{0} \boxed{1} \boxed{2} \boxed{3}$ . The chance that the sum will be in the interval from 10 to 20 inclusive equals the area under \_\_\_\_\_ between \_\_\_\_\_ and \_\_\_\_\_. Fill in the blanks. For the first one, your options are: the normal curve, the probability histogram for the sum. Explain your answers.
4. A coin is tossed 25 times. Estimate the chance of getting 12 heads and 13 tails.

5. Twenty-five draws are made at random with replacement from the box  $\boxed{1} \boxed{1} \boxed{2} \boxed{2} \boxed{3}$ . One of the graphs below is a histogram for the numbers drawn. One is the probability histogram for the sum. And one is the probability histogram for the product. Which is which? Why?

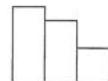
(i)



(ii)



(iii)



6. A programmer is working on a new program, COIN, to simulate tossing a coin. As a preliminary test, he sets up the program to do one million tosses. The program returns with a count of 502,015 heads. The programmer looks at this and thinks:

Hmmm. Two thousand and fifteen off. That's a lot. No, wait. Compare it to the million. Two thousand—forget the fifteen—out of a million is two out of a thousand. That's one in five hundred. One fifth of a percent. Very small. Good. COIN passes.

Do you agree? Answer yes or no, and explain.

7. A pair of dice are thrown. The total number of spots is like

- (i) one draw from the box

$\boxed{2}$	$\boxed{3}$	$\boxed{4}$	$\boxed{5}$	$\boxed{6}$	$\boxed{7}$	$\boxed{8}$	$\boxed{9}$	$\boxed{10}$	$\boxed{11}$	$\boxed{12}$
-------------	-------------	-------------	-------------	-------------	-------------	-------------	-------------	--------------	--------------	--------------

- (ii) the sum of two draws from the box

$\boxed{1}$	$\boxed{2}$	$\boxed{3}$	$\boxed{4}$	$\boxed{5}$	$\boxed{6}$
-------------	-------------	-------------	-------------	-------------	-------------

Explain.

8. A coin is tossed 100 times. True or false, and explain:

- (a) The expected value for the number of heads is 50.
- (b) The expected value for the number of heads is 50, give or take 5 or so.
- (c) The number of heads will be 50.
- (d) The number of heads will be around 50, give or take 5 or so.

9. One hundred draws are made at random with replacement from a box with ninety-nine tickets marked “0” and one ticket marked “1.” True or false, and explain:

- (a) The sum will be around 1, give or take 1 or so.
- (b) There is about a 68% chance that the sum will be in the range 0 to 2.

10. Ten thousand draws are made at random with replacement from a box with ninety-nine tickets marked “0” and one ticket marked “1.” True or false, and explain:
- The sum will be around 100, give or take 10 or so.
  - There is about a 68% chance that the sum will be in the range 90 to 110.
11. One hundred draws are made at random with replacement from the box  $\boxed{1} \boxed{2} \boxed{2} \boxed{5}$ . The draws come out as follows: 17  $\boxed{1}$ 's, 54  $\boxed{2}$ 's, 29  $\boxed{5}$ 's. Fill in the blanks, using the options below; show work.
- For the \_\_\_\_\_, the observed value is 0.8 SEs above the expected value.
  - For the \_\_\_\_\_, the observed value is 1.33 SEs above the expected value.

Options (one will be left over):

sum of the draws      number of 1's      number of 2's

12. A box contains ten tickets, four marked with a positive number and six with a negative number. All the numbers are between  $-10$  and  $10$ . One thousand draws will be made at random with replacement from the box. You are asked to estimate the chance that the sum will be positive.
- Can you do it on the basis of the information already given?
  - Can you do it if you are also told the average and SD of the numbers in the box, but are not told the numbers themselves?

Explain briefly.

13. Repeat exercise 12, if you are asked to estimate the chance of getting 100 or more  $\boxed{3}$ 's.
14. Repeat exercise 12, if you are asked to estimate the chance of getting 425 or more positive numbers.
15. A box contained 1,500 marbles; 600 were red and the others, blue. The following procedure was repeated many times.

One hundred draws were made at random with replacement from the box; the number of red marbles among the draws was counted.

The first 10 counts were 38, 35, 37, 31, 36, 39, 36, 33, 30, 34. Is anything fishy? Answer yes or no, and explain.

## 8. SUMMARY

- If the chance process for getting a sum is repeated many times, the empirical histogram for the observed values converges to the *probability histogram*.

2. A probability histogram represents chance by area.
3. When drawing at random with replacement from a box, the probability histogram for the sum will follow the normal curve, even if the contents of the box do not—the “central limit theorem.” The histogram must be put into standard units, and the number of draws must be reasonably large.
4. The normal approximation consists in replacing the actual probability histogram by the normal curve, before computing areas. Often, the accuracy of the approximation can be improved by keeping track of the edges of the rectangles—the “continuity correction.”
5. Probability histograms which follow the normal curve can be summarized quite well by the expected value and SE. The expected value locates the center of the probability histogram, and the SE measures the spread.
6. Chapter 16 developed box models for gambling. These models are basic to statistical inference (parts VI–VIII).
7. Chapter 17 introduced the SE for sums of draws from a box. The SE for counts, percents (chapter 20), or averages (chapter 23) are then easily computed. Confidence intervals are derived in chapter 21.
8. Chapter 18 showed that probability histograms for sums converge to the normal curve. That justifies “large-sample” statistical theory—reading confidence levels and  $P$ -values off the curve (chapters 21 and 26).
9. “Small-sample” statistical theory includes the  $t$ -test (section 26.6) and the sign test (section 27.5). In such cases, distributions other than the normal are used.

PART VI

# Sampling

— — — — —

# 19

## Sample Surveys

*"Data! data! data!" he cried impatiently. "I can't make bricks without clay."*

—Sherlock Holmes<sup>1</sup>

### 1. INTRODUCTION

An investigator usually wants to generalize about a class of individuals. This class is called the *population*. For example, in forecasting the results of a presidential election in the U.S., one relevant population consists of all eligible voters. Studying the whole population is usually impractical. Only part of it can be examined, and this part is called the *sample*. Investigators will make generalizations from the part to the whole. In more technical language, they make *inferences* from the sample to the population.<sup>2</sup>

Usually, there are some numerical facts about the population which the investigators want to know. Such numerical facts are called *parameters*. In forecasting a presidential election in the U.S., two relevant parameters are

- the average age of all eligible voters,
- the percentage of all eligible voters who are currently registered to vote.

Ordinarily, parameters like these cannot be determined exactly, but can only be estimated from a sample. Then a major issue is accuracy. How close are the estimates going to be?

Parameters are estimated by *statistics*, or numbers which can be computed from a sample. For instance, with a sample of 10,000 Americans, an investigator could calculate the following two statistics:

- the average age of the eligible voters in the sample,
- the percentage of the eligible voters in the sample who are currently registered to vote.

Statistics are what investigators know; parameters are what they want to know.

Estimating parameters from the sample is justified when the sample represents the population. This is impossible to check just by looking at the sample. The reason: to see whether the sample is like the population in the ways that matter, investigators would have to know the facts about the population that they are trying to estimate—a vicious circle. Instead, one has to look at how the sample was chosen. Some methods tend to do badly. Others are likely to give representative samples.

The two main lessons of this chapter:

- the method of choosing the sample matters a lot;
- the best methods involve the planned introduction of chance.

Similar issues come up when assigning subjects to treatment or control in experiments: see part I.

## 2. THE LITERARY DIGEST POLL

In 1936, Franklin Delano Roosevelt was completing his first term of office as president of the U.S. It was an election year, and the Republican candidate was Governor Alfred Landon of Kansas. The country was struggling to recover from the Great Depression. There were still nine million unemployed: real income had dropped by one-third in the period 1929–1933 and was just beginning to turn upward. But Landon was campaigning on a program of economy in government, and Roosevelt was defensive about his deficit financing.<sup>3</sup>

*Landon.* The spenders must go.

*Roosevelt.* We had to balance the budget of the American people before we could balance the budget of the national government. That makes common sense, doesn't it?

The Nazis were rearming Germany, and the Civil War in Spain was moving to its hopeless climax. These issues dominated the headlines in the *New York Times*, but were ignored by both candidates.

*Landon.* We must mind our own business.

Most observers thought Roosevelt would be an easy winner. Not so the *Literary Digest* magazine, which predicted an overwhelming victory for Landon, with Roosevelt getting only 43% of the popular vote. This prediction was based on the largest number of people ever replying to a poll—about 2.4 million individuals. It was backed by the enormous prestige of the *Digest*, which had called the winner in every presidential election since 1916. However, Roosevelt won the 1936 election by a landslide—62% to 38%. (The *Digest* went bankrupt soon after.)

The magnitude of the *Digest's* error is staggering. It is the largest ever made by a major poll. Where did it come from? The number of replies was more than big enough. In fact, George Gallup was just setting up his survey organization.<sup>4</sup> Using his own methods, he drew a sample of 3,000 people and predicted what the *Digest* predictions were going to be—well in advance of their publication—with an error of only one percentage point. Using another sample of about 50,000 people, he correctly forecast the Roosevelt victory, although his prediction of Roosevelt's share of the vote was off by quite a bit. Gallup forecast 56% for Roosevelt; the actual percentage was 62%, so the error was  $62\% - 56\% = 6$  percentage points. (Survey organizations use "percentage points" as the units for the difference between actual and predicted percents.) The results are summarized in table 1.

Table 1. The election of 1936.

	<i>Roosevelt's percentage</i>
The election result	62
The <i>Digest</i> prediction of the election result	43
Gallup's prediction of the <i>Digest</i> prediction	44
Gallup's prediction of the election result	56

Note: Percentages are of the major-party vote. In the election, about 2% of the ballots went to minor-party candidates.

Source: George Gallup, *The Sophisticated Poll-Watcher's Guide* (1972).

To find out where the *Digest* went wrong, you have to ask how they picked their sample. A sampling procedure should be fair, selecting people for inclusion in the sample in an impartial way, so as to get a representative cross section of the public. A systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample is called *selection bias*. The *Digest's* procedure was to mail questionnaires to 10 million people. The names and addresses of these 10 million people came from sources like telephone books and club membership lists. That tended to screen out the poor, who were unlikely to belong to clubs or have telephones. (At the time, for example, only one household in four had a telephone.) So there was a very strong bias against the poor in the *Digest's* sampling procedure. Prior to 1936, this bias may not have affected the predictions very much, because rich and poor voted along similar lines. But in 1936, the political split followed economic lines more closely. The poor voted overwhelmingly for Roosevelt, the rich were for Landon. One reason for the magnitude of the *Digest's* error was selection bias.

When a selection procedure is biased, taking a large sample does not help. This just repeats the basic mistake on a larger scale.

The *Digest* did very badly at the first step in sampling. But there is also a second step. After deciding which people ought to be in the sample, a survey

organization still has to get their opinions. This is harder than it looks. If a large number of those selected for the sample do not in fact respond to the questionnaire or the interview, *non-response bias* is likely.

The non-respondents differ from the respondents in one obvious way: they did not respond. Experience shows they tend to differ in other important ways as well.<sup>5</sup> For example, the *Digest* made a special survey in 1936, with questionnaires mailed to every third registered voter in Chicago. About 20% responded, and of those who responded over half favored Landon. But in the election Chicago went for Roosevelt, by a two-to-one margin.

Non-respondents can be very different from respondents. When there is a high non-response rate, look out for non-response bias.

In the main *Digest* poll, only 2.4 million people bothered to reply, out of the 10 million who got the questionnaire. These 2.4 million respondents do not even represent the 10 million people who were polled, let alone the population of all voters. The *Digest* poll was spoiled both by selection bias and non-response bias.<sup>6</sup>

Special surveys have been carried out to measure the difference between respondents and non-respondents. It turns out that lower-income and upper-income people tend not to respond to questionnaires, so the middle class is over-represented among respondents. For these reasons, modern survey organizations prefer to use personal interviews rather than mailed questionnaires. A typical response rate for personal interviews is 65%, compared to 25% for mailed questionnaires.<sup>7</sup> However, the problem of non-response bias still remains, even with personal interviews. Those who are not at home when the interviewer calls may be quite different from those who are at home, with respect to working hours, family ties, social background, and therefore with respect to attitudes. Good survey organizations keep this problem in mind, and have ingenious methods for dealing with it (section 6).

Some samples are really bad. To find out whether a sample is any good, ask how it was chosen. Was there selection bias? non-response bias? You may not be able to answer these questions just by looking at the data.

In the 1936 election, how did Gallup predict the *Digest* predictions? He just chose 3,000 people at random from the same lists the *Digest* was going to use, and mailed them all a postcard asking how they planned to vote. He knew that a random sample was likely to be quite representative, as will be explained in the next two chapters.

### 3. THE YEAR THE POLLS ELECTED DEWEY

Thomas Dewey rose to fame as a crusading D.A. in New York City, and went on to capture the governor's mansion in Albany. In 1948 he was the Republican candidate for president, challenging the incumbent Harry Truman. Truman began political life as a protégé of Boss Pendergast in Kansas City. After being elected to the Senate, Truman became FDR's vice president, succeeding to the presidency when Roosevelt died. Truman was one of the most effective presidents of the 20th century, as well as one of the most colorful. He kept a sign on his desk, "The buck stops here." Another of his favorite aphorisms became part of America's political vocabulary: "If you can't stand the heat, stay out of the kitchen." But Truman was the underdog in 1948, for it was a troubled time. World War II had barely ended, and the uneasy half-peace of the Cold War had just begun. There was disquiet at home, and complicated involvement abroad.

Three major polls covered the election campaign: Crossley, for the Hearst newspapers; Gallup, syndicated in about 100 independent newspapers across the country; and Roper, for *Fortune* magazine. By fall, all three had declared Dewey the winner, with a lead of around 5 percentage points. Gallup's prediction was based on 50,000 interviews; and Roper's on 15,000. As the *Scranton Tribune* put it,

#### DEWEY AS GOOD AS ELECTED, STATISTICS CONVINCE ROPER

The statistics didn't convince the American public. On Election Day, Truman scored an upset victory with just under 50% of the popular vote. Dewey got just over 45% (table 2).

Table 2. The election of 1948.

<i>The candidates</i>	<i>The predictions</i>			<i>The results</i>
	<i>Crossley</i>	<i>Gallup</i>	<i>Roper</i>	
Truman	45	44	38	50
Dewey	50	50	53	45
Thurmond	2	2	5	3
Wallace	3	4	4	2

Source: F. Mosteller and others, *The Pre-Election Polls of 1948* (New York: Social Science Research Council, 1949).

To find out what went wrong for the polls, it is necessary to find out how they chose their samples.<sup>8</sup> The method they all used is called *quota sampling*. With this procedure, each interviewer was assigned a fixed quota of subjects to interview. The numbers falling into certain categories (like residence, sex, age, race, and economic status) were also fixed. In other respects, the interviewers were free to select anybody they liked. For instance, a Gallup Poll interviewer in St. Louis was required to interview 13 subjects, of whom:<sup>9</sup>

- exactly 6 were to live in the suburbs, and 7 in the central city,
- exactly 7 were to be men, and 6 women.

Of the 7 men (and there were similar quotas for the women):

- exactly 3 were to be under forty years old, and 4 over forty,
- exactly 1 was to be black, and 6 white.

The monthly rentals to be paid by the 6 white men were specified also:

- 1 was to pay \$44.01 or more;
- 3 were to pay \$18.01 to \$44.00;
- 2 were to pay \$18.00 or less.

Remember, these are 1948 prices!

From a common-sense point of view, quota sampling looks good. It seems to guarantee that the sample will be like the voting population with respect to all the important characteristics that affect voting behavior. (Distributions of residence, sex, age, race, and rent can be estimated quite closely from Census data.) But the 1948 experience shows this procedure worked very badly. We are now going to see why.

The survey organizations want a sample which faithfully represents the nation's political opinions. However, no quotas can be set on Republican or Democratic votes. The distribution of political opinion is precisely what the survey organizations do not know and are trying to find out. The quotas for the other variables are an indirect effort to make the sample reflect the nation's politics. Fortunately or unfortunately, there are many factors which influence voting behavior besides the ones the survey organizations control for. There are rich white men in the suburbs who vote Democratic, and poor black women in the central cities who vote Republican. As a result, survey organizations may hand-pick a sample which is a perfect cross section of the nation on all the demographic variables, but find the sample voting one way while the nation goes the other. This possibility must have seemed quite theoretical—before 1948.

The next argument against quota sampling is the most important. It involves a crucial feature of the method, which is easy to miss the first time through. Within the assigned quotas, the interviewers are free to choose anybody they like. That leaves a lot of room for human choice. And human choice is always subject to bias. In 1948, the interviewers chose too many Republicans. On the whole, Republicans are wealthier and better educated than Democrats. They are more likely to own telephones, have permanent addresses, and live on nicer blocks. Within each demographic group, Republicans are marginally easier to interview. If you were an interviewer, you would probably end up with too many Republicans.

The interviewers chose too many Republicans in every presidential election from 1936 through 1948, as shown by the Gallup Poll results in table 3. Prior to 1948, the Democratic lead was so great that it swamped the Republican bias in the polls. The Democratic lead was much slimmer in 1948, and the Republican bias in quota sampling had real impact.

Table 3. The Republican bias in the Gallup Poll, 1936–1948.

Year	Gallup's prediction of Republican vote	Actual Republican vote	Error in favor of the Republicans
1936	44	38	6
1940	48	45	3
1944	48	46	2
1948	50	45	5

Note: Percentages are of the majority-party vote, except in 1948.

Source: F. Mosteller and others, *The Pre-Election Polls of 1948* (New York: Social Science Research Council, 1949).

In quota sampling, the sample is hand-picked to resemble the population with respect to some key characteristics. The method seems reasonable, but does not work very well. The reason is unintentional bias on the part of the interviewers.

The quotas in quota sampling are sensible enough, although they do not guarantee success—far from it. But the method of filling the quotas, free choice by the interviewers, is disastrous.<sup>10</sup> The alternative is to use objective and impartial chance mechanisms to select the sample. That will be the topic of the next section.

#### 4. USING CHANCE IN SURVEY WORK

Even in 1948, a few survey organizations used *probability methods* to draw their samples. Now, many organizations do. What is a probability method for drawing a sample? To get started, imagine carrying out a survey of 100 voters in a small town with a population of 1,000 eligible voters. Then, it is feasible to list all the eligible voters, write the name of each one on a ticket, put all 1,000 tickets in a box, and draw 100 tickets at random. Since there is no point interviewing the same person twice, the draws are made without replacement. In other words, the box is shaken to mix up the tickets. One is drawn out at random and set aside. That leaves 999 in the box. The box is shaken again, a second ticket is drawn out and set aside. The process is repeated until 100 tickets have been drawn. The people whose tickets have been drawn form the sample.

This process is called *simple random sampling*: tickets have simply been drawn at random without replacement. At each draw, every ticket in the box has an equal chance to be chosen. The interviewers have no discretion at all in whom they interview, and the procedure is impartial—everybody has the same chance to get into the sample. Consequently, the law of averages guarantees that the percentage of Democrats in the sample is likely to be close to the percentage in the population.

*Simple random sampling* means drawing at random without replacement.

What happens in a more realistic setting, when the Gallup Poll tries to predict a presidential election? A natural idea is to take a nationwide simple random sample of a few thousand eligible voters. However, this isn't as easy to do as it sounds. Drawing names at random, in the statistical sense, is hard work. It is not at all the same as choosing people haphazardly.

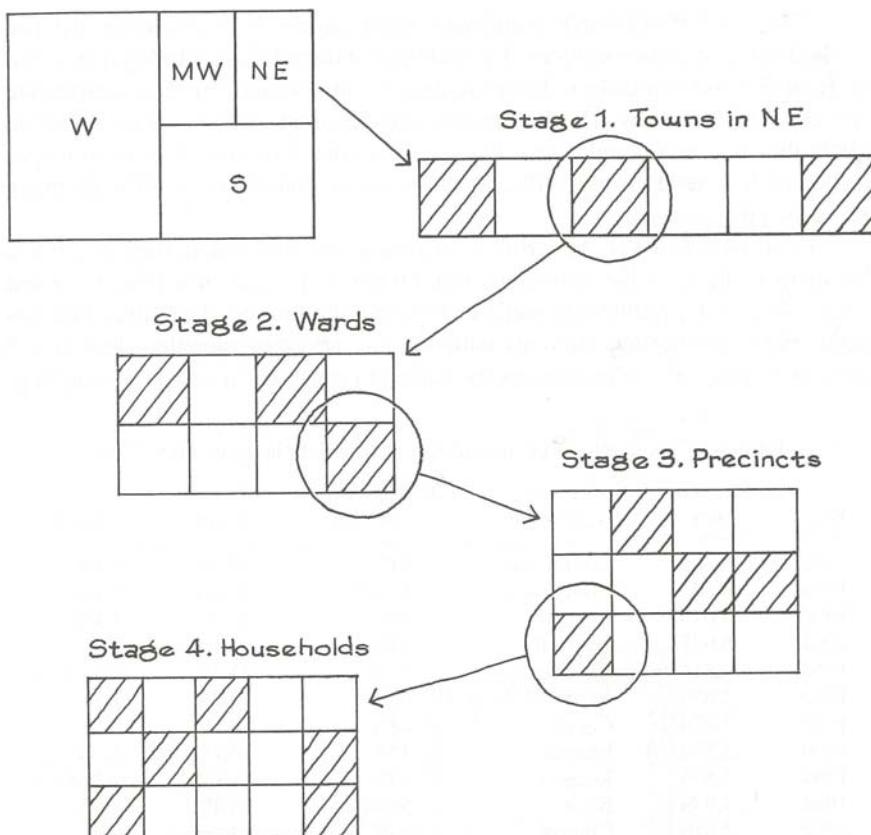
To begin drawing eligible voters at random, you would need a list of all of them—well over 200 million names. There is no such list.<sup>11</sup> Even if there were, drawing a few thousand names at random from 200 million is a job in itself. (Remember, on each draw every name in the box has to have an equal chance of being selected.) And even if you could draw a simple random sample, the people would be scattered all over the map. It would be prohibitively expensive to send interviewers around to find them all.

It just is not practical to take a simple random sample. Consequently, most survey organizations use a probability method called *multistage cluster sampling*. The name is complicated, and so are the details. But the idea is straightforward. It will be described in the context of the Gallup pre-election surveys during the period from 1952 through 1984; these surveys were all done using just about the same procedure. The Gallup Poll makes a separate study in each of the four geographical regions of the United States—Northeast, South, Midwest, and West (figure 1). Within each region, they group together all the population centers of similar sizes. One such grouping might be all towns in the Northeast with a population between 50 and 250 thousand. Then, a random sample of these towns is selected. Interviewers are stationed in the selected towns, and no interviews are conducted in the other towns of that group. Other groupings are handled the same way. This completes the first stage of sampling.<sup>12</sup>

For election purposes, each town is divided up into *wards*, and the wards are subdivided into *precincts*. At the second stage of sampling, some wards are selected—at random—from each town chosen in the stage before. At the third stage, some precincts are drawn at random from each of the previously selected wards. At the fourth stage, households are drawn at random from each selected precinct.<sup>13</sup> Finally, some members of the selected households are interviewed. Even here, no discretion is allowed. For instance, Gallup Poll interviewers are instructed to “speak to the youngest man 18 or older at home, or if no man is at home, the oldest woman 18 or older.”<sup>14</sup>

This design offers many of the advantages of quota sampling. For instance, it is set up so the distribution of the sample by residence is the same as the distribution for the nation. But each stage in the selection procedure uses an objective and impartial chance mechanism to select the sample units. This completely eliminates the worst feature of quota sampling: selection bias on the part of the interviewer.

Figure 1. Multistage cluster sampling.



Simple random sampling is the basic probability method. Other methods can be quite complicated. But all probability methods for sampling have two important features:

- the interviewers have no discretion at all as to whom they interview;
- there is a definite procedure for selecting the sample, and it involves the planned use of chance.

As a result, with a probability method it is possible to compute the chance that any particular individuals in the population will get into the sample.<sup>15</sup>

Quota sampling is not a probability method. It fails both tests. The interviewers have a lot of discretion in choosing subjects. And chance only enters in the most unplanned and haphazard way. What kinds of people does the interviewer like to approach? Who is going to be walking down a particular street at a particular time of day? No survey organization can put numbers on these kinds of chances.

## 5. HOW WELL DO PROBABILITY METHODS WORK?

Since 1948, the Gallup Poll and many other major polls have used probability methods to choose their samples. The Gallup Poll record in post-1948 presidential elections is shown in table 4. There are three points to notice. (i) The sample size has gone down sharply. The Gallup Poll used a sample of size about 50,000 in 1948; they now use samples less than a tenth of that size. (ii) There is no longer any consistent trend favoring either Republicans or Democrats. (iii) The accuracy has gone up appreciably.

From 1936 to 1948, the errors were around 5%. Since then, they are quite a bit smaller. (In 1992, the error went back up to 6%; the reason will be discussed on p. 346.) Using probability methods to select the sample, the Gallup Poll has been able to predict the elections with startling accuracy, sampling less than 5 persons in 100,000—which proves the value of probability methods in sampling.

Table 4. The Gallup Poll record in presidential elections after 1948.

Year	Sample size	Winning candidate	Gallup Poll prediction	Election result	Error
1952	5,385	Eisenhower	51%	55.1%	4.1%
1956	8,144	Eisenhower	59.5%	57.4%	2.1%
1960	8,015	Kennedy	51%	49.7%	1.3%
1964	6,625	Johnson	64%	61.1%	2.9%
1968	4,414	Nixon	43%	43.4%	0.4 of 1%
1972	3,689	Nixon	62%	60.7%	1.3%
1976	3,439	Carter	48%	50.1%	2.1%
1980	3,500	Reagan	47%	50.7%	3.7%
1984	3,456	Reagan	59%	58.8%	0.2 of 1%
1988	4,089	Bush	56%	53.4%	2.6%
1992	2,019	Clinton	49%	43.0%	6.0%
1996	2,895	Clinton	52%	49.2%	2.8%
2000	3,571	Bush	48%	47.9%	0.1 of 1%
2004	2,014	Bush	49%	50.6%	1.6%

Note: The percentages are of the popular vote. The error is the absolute difference “predicted – actual.”

Source: The Gallup Poll (American Institute of Public Opinion) for predictions; *Statistical Abstract*, 2006, Table 384 for actuals.

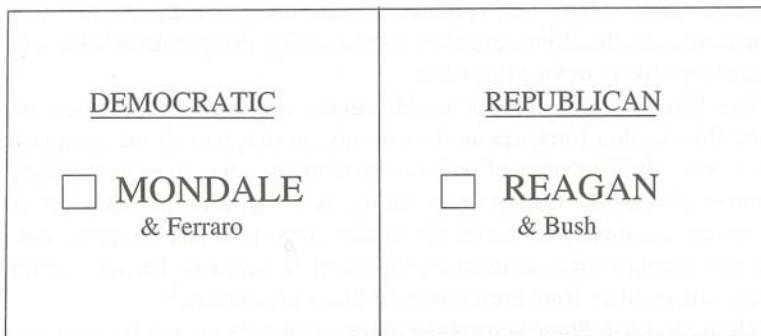
Why do probability methods work so well? At first, it may seem that judgment is needed to choose the sample. For instance, quota sampling guarantees that the percentage of men in the sample will be equal to the percentage of men in the population. With probability sampling, we can only say that the percentage of men in the sample is likely to be close to the percentage in the population: certainty is reduced to likelihood. But judgment and choice usually show bias, while chance is impartial. That is why probability methods work better than judgment.

To minimize bias, an impartial and objective probability method should be used to choose the sample.

## 6. A CLOSER LOOK AT THE GALLUP POLL

Some degree of bias is almost inevitable even when probability methods are used to select the sample, due to the many practical difficulties that survey organizations must overcome. The discussion here is organized around the questionnaire used by the Gallup Poll in the presidential election of 1984. See figures 2 and 3.

Figure 2. The Gallup Poll ballot, 1984. The interviewers use secret ballots, to minimize the number of undecided respondents.



"I'D SAY I'M ABOUT FORTY-TWO PERCENT FOR REAGAN, THIRTY-NINE PERCENT FOR MONDALE, AND NINETEEN PERCENT UNDECIDED."

*The nonvoters.* In a typical presidential election, between one-third and one-half of the eligible voters fail to vote. The job of the Gallup Poll is to predict what the voters will do; the non-voters are irrelevant and should be screened out of the sample as far as possible. That is not so easy. There is a stigma to non-voting, and many respondents say they will vote even if they know better. The problem of screening out non-voters is handled by questions 1–6, and some later questions too: this is a hot topic for pollsters. Question 3, for instance, asks where the respondent would go to vote (figure 3, p. 345). Respondents who know the answer are more likely to vote. Question 13 (p. 347) asks whether the respondent voted in the last election, and is phrased to make a negative answer easy to give—compensating for the stigma attached to non-voting. Respondents who voted last time are more likely to vote this time.

This battery of questions is used to decide whether the respondent is likely to vote; the election forecasts are based only on that part of the sample judged likely to vote. It is a matter of record who actually votes in each election. Post-election studies by the Gallup organization show that their judgments as to who will vote are reasonably accurate. The studies also show that screening out likely non-voters increases the accuracy of the election forecasts, because preferences of likely voters differ from preferences of likely non-voters.<sup>16</sup>

*The undecided.* Some percentage of the subjects being interviewed are undecided how they will vote. Question 7, which asks for the preferences, is designed to keep the percentage as small as possible. To begin with, it asks how the respondent would vote the day of the interview, rather than Election Day. Subjects who cannot decide are asked to indicate “the candidates toward whom you lean as of today.” A final device is the paper ballot (figure 2, p. 343). Instead of naming their preferences out loud, the respondents just mark the ballot and drop it into a box carried by the interviewer.<sup>17</sup>

These techniques have been found to minimize the percentage of undecided. But there are still some left, and if they are thought likely to vote, the Gallup Poll has to guess how. Some information about political attitudes is available from questions 12–14 (p. 347). This information might be used to predict how the undecided respondents are going to vote, but it is difficult to say how well the predictions work.

*Response bias.* The answers given by respondents are influenced to some extent by the phrasing of the questions, and even the tone or attitude of the interviewer. This kind of distortion is called *response bias*. There was a striking example in the 1948 election survey: changing the order of the candidates’ names was found to change the response by 5%, the advantage being with the candidate who was named first. To control response bias, all interviewers use the same questionnaire, and the interview procedure is standardized as far as possible. The ballot technique was found to reduce the effect of the political attitudes of the interviewer on the responses of the subjects.

*Non-response bias.* Even with personal interviews, many subjects are missed. Since they tend to be different from the subjects available for the interview, a non-response bias is created. To some extent, this bias can be adjusted out, by giving more weight to the subjects who were available but hard to get. This information is obtained by question 20 (p. 347), which asks whether the subject was at home on the previous days. This is done quite subtly, as you can tell by reading the question.

Figure 3. The Gallup Poll questionnaire for the 1984 election. Courtesy of the Gallup Poll News Service.

<p><b>SURVEY:</b> A1813 <b>DATE:</b> October 25, 1984</p> <p>No publication, reproduction, dissemination or other use of this questionnaire or any replies thereto, written or oral, is authorized by The Gallup Organization, Inc. Violators of this notice will be prosecuted to the fullest extent of the law.</p>	 <p><b>Sponsored by leading Newspapers, Corporations and Agencies.</b></p> <p>Copyright 1977 The Gallup Organization, Inc. Princeton, New Jersey 08540</p> <p><b>SUGGESTED INTRODUCTION:</b> I'm taking a GALLUP SURVEY. I'd like YOUR opinion on some topics of interest.</p>	<p>Time started: _____</p> <p>Time ended: _____</p> <p>Length: _____</p>
<p>1. How much thought have you given to the coming November elections—quite a lot, or only a little?</p> <p>1 <input type="checkbox"/> Quite a lot 2 <input type="checkbox"/> Some — (volunteered) 3 <input type="checkbox"/> Little y <input type="checkbox"/> None</p> <p>2. Have you ever voted in this precinct or district?</p> <p>1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No y <input type="checkbox"/> Don't know</p> <p>3. Where do people who live in this neighborhood go to vote?</p> <p>1 <input type="checkbox"/> Specify: _____ y <input type="checkbox"/> Don't know</p> <p>4a. Are you NOW registered so that you can vote in the election this November?</p> <p>1 <input type="checkbox"/> Yes — (GO TO Q. 5) 2 <input type="checkbox"/> No 3 <input type="checkbox"/> Don't have to register — (GO TO Q. 5) y <input type="checkbox"/> Don't know</p> <p>4b. Do you plan to register so that you can vote in the November election?</p> <p>1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No 3 <input type="checkbox"/> Other: _____</p> <p>5. Generally speaking, how much interest would you say you have in politics — a great deal, a fair amount, only a little, or no interest at all?</p> <p>1 <input type="checkbox"/> Great deal 2 <input type="checkbox"/> Fair amount 3 <input type="checkbox"/> Little y <input type="checkbox"/> None</p> <p>6. How often would you say you vote — always, nearly always, part of the time, or seldom?</p> <p>1 <input type="checkbox"/> Always 2 <input type="checkbox"/> Nearly always 3 <input type="checkbox"/> Part of the time 4 <input type="checkbox"/> Seldom 5 <input type="checkbox"/> Other: _____ y <input type="checkbox"/> Never vote</p> <p>7. Suppose you were voting TODAY for president and vice president of the United States. Here is a Gallup Poll secret ballot listing the candidates for these offices. (TEAR OFF ATTACHED BALLOT AND HAND TO RESPONDENT.) Will you please MARK that secret ballot for the candidates you favor today — and then drop the folded ballot into the box.</p> <p>INTERVIEWER: IF RESPONDENT HANDS BACK BALLOT AND SAYS HE HASN'T MADE UP HIS MIND OR REFUSES TO MARK IT SAY:</p>		
<p>Well, would you please mark the ballot for the candidates toward whom you lean as of today?</p> <p>IF RESPONDENT STILL CAN'T DECIDE OR REFUSES TO MARK THE BALLOT, PLEASE WRITE THAT ON THE BALLOT AND BE SURE TO DROP IT IN THE BOX.</p> <p>8. Right now, how strongly do you feel about your choice—very strongly, fairly strongly, or not strongly at all?</p> <p>1 <input type="checkbox"/> Very strongly 2 <input type="checkbox"/> Fairly strongly 3 <input type="checkbox"/> Not strongly at all 4 <input type="checkbox"/> Didn't make choice y <input type="checkbox"/> Don't know</p> <p>9a. Do you, yourself, plan to vote in the election this November, or not?</p> <p>1 <input type="checkbox"/> Yes 2 <input type="checkbox"/> No _____ (GO TO Q. 10a) y <input type="checkbox"/> Don't know _____</p> <p>9b. How certain are you that you will vote—ABSOLUTELY certain, FAIRLY certain, or NOT certain?</p> <p>1 <input type="checkbox"/> Absolutely 2 <input type="checkbox"/> Fairly y <input type="checkbox"/> Not certain</p> <p>10a. If the elections for Congress were being held TODAY, which party would you like to see win in this Congressional district, the Democratic Party or the Republican Party?</p> <p>1 <input type="checkbox"/> Democratic 2 <input type="checkbox"/> Republican _____ (GO TO Q. 11) 3 <input type="checkbox"/> Other _____ y <input type="checkbox"/> Undecided, refused</p> <p>10b. As of today, do you lean more to the Democratic Party or more to the Republican Party?</p> <p>1 <input type="checkbox"/> Democratic 2 <input type="checkbox"/> Republican 3 <input type="checkbox"/> Other: 4 <input type="checkbox"/> Undecided y <input type="checkbox"/> Refused</p> <p>11. Here is a picture of a ladder. (HAND RESPONDENT CARD 1.) Suppose we say the top of the ladder (POINT) marked 10 represents a person who definitely will vote in the election this November, and the bottom of the ladder (POINT) marked zero represents a person who definitely will not vote in the election. How far up or down the ladder would you place yourself? (INTERVIEWER: CIRCLE NUMBER.)</p> <p style="text-align: right;">10 9 8 7 6 5 4 3 2 1 0</p> <p>y <input type="checkbox"/> Don't know</p>		

*Check data.* The Gallup sample usually includes proportionately too many people with higher education. In a detailed analysis, less weight is put on the responses of those subjects (question 16). Other demographic data can be used in a similar way. This weighting technique is called "ratio estimation." Do not confuse ratio estimation with quota sampling. Ratio estimation is an objective, arithmetic technique applied to the sample after it is chosen, to compensate for various small biases in the sampling procedure. Quota sampling is a method for choosing the sample. It has a large, subjective component—when the interviewer chooses the subjects—and introduces large biases.

*Interviewer control.* In large-scale survey work, there is always the problem of making sure interviewers follow instructions. Some redundancy is built into the questionnaire, so the answers can be checked for consistency: inconsistencies suggest the interviewer may not be doing the job properly. A small percentage of the subjects are reinterviewed by administrative staff, as a further check on the quality of the work.

*Talk is cheap.* It is a little risky to predict what people will do on Election Day from what they tell the interviewer they are going to do. People may be unwilling to reveal their true preferences. Even if they do, they may change their minds later. Words and deeds are often different.

*The 1992 election.* In 1992, there was a fairly large percentage of undecided respondents, and Gallup allocated all of them to Clinton. That turned out to be a bad idea. Many of the undecided seem in the end to have voted for Perot, explaining Gallup's large error for the 1992 election (table 4, p. 342). Predicted and actual votes for Clinton, Bush, and Perot are shown below.

	Clinton	Bush	Perot
Gallup	49%	37%	14%
Actual	43.0%	37.4%	18.9%

## 7. TELEPHONE SURVEYS

Many surveys are now conducted by telephone. The savings in costs are dramatic, and—if the work is up to scratch—the results are good. The Gallup Poll changed over to the telephone in 1988, with 200 interviewers covering the whole country in a few days, from offices in Atlanta, Austin, Lincoln, Minneapolis, and Omaha.

How do they pick the sample? In 1988, the Gallup Poll used a multistage cluster sample based on area codes, "exchanges," and "banks:"

Area code	Exchange	Bank	Digits
415	767	26	76

In 1992, they switched to a simpler design. There are 4 time zones in the U.S. The Gallup Poll divided each zone into 3 types of areas, according to population density (heavy, medium, light). That gives  $4 \times 3 = 12$  strata. For example, one stratum consisted of heavily populated areas in the Eastern time zone; another consisted of lightly populated areas on Pacific time. Within each stratum, the Gallup Poll just drew a simple random sample of telephone numbers, using the computer to

Figure 3. The Gallup Poll questionnaire for the 1984 election, continued.  
Courtesy of the Gallup Poll News Service.

NOW, HERE ARE A FEW QUESTIONS SO THAT MY OFFICE CAN KEEP TRACK OF THE CROSS-SECTION OF PEOPLE I'VE TALKED TO:

12. In politics, as of TODAY, do you consider yourself a Republican, Democrat, or Independent?

- Republican
- Democrat
- Independent
- Other: \_\_\_\_\_

13. In the election in November 1980 — when Carter ran against Reagan and Anderson — did things come up which kept you from voting, or did you happen to vote? For whom?

- Carter
- Reagan
- Anderson
- Other
- Voted, don't remember for whom
- No, didn't vote
- Don't remember if voted

14. Are you, or is your (husband/wife) a member of a labor union?

- Yes, respondent is
- Yes, spouse is
- Yes, both are
- No, neither is

15. (HAND RESPONDENT CARD 2) Please tell me which of the categories on this card MOST NEARLY describes the kind of work the chief wage earner in your immediate family does. Just call off the number, please. (INTERVIEWER: IF THE CHIEF WAGE EARNER IS UNEMPLOYED, ASK WHAT TYPE OF WORK HE/SHE WOULD DO IF EMPLOYED.)

- |                             |  |
|-----------------------------|--|
| <input type="checkbox"/> 1  | <input type="checkbox"/> 11              |
| <input type="checkbox"/> 2  | <input type="checkbox"/> 12              |
| <input type="checkbox"/> 3  | <input type="checkbox"/> 13              |
| <input type="checkbox"/> 4  | <input type="checkbox"/> 14              |
| <input type="checkbox"/> 5  | <input type="checkbox"/> 15              |
| <input type="checkbox"/> 6  | <input type="checkbox"/> 16 Other: _____ |
| <input type="checkbox"/> 7  | <input type="checkbox"/> 17 Can't say    |
| <input type="checkbox"/> 8  |  |
| <input type="checkbox"/> 9  |  |
| <input type="checkbox"/> 10 |  |

16. What was the last grade or class you COMPLETED in school?

- None or Grades 1-4
- Grades 5, 6, 7
- Grade 8
- High school, incomplete (Grades 9-11)
- High school, graduated (Grade 12)
- Technical, trade, or business
- College, university, incomplete
- College, university, graduated

17. What is your religious preference — Protestant, Roman Catholic, Jewish, or an Orthodox church such as the Greek or Russian Orthodox Church?

- Protestant
- Roman Catholic
- Jewish
- Orthodox Church
- Other: \_\_\_\_\_
- None

18. How many persons 18 years and over are there now living in this household, including yourself? Include lodgers, servants, or other employees living in the household. (CIRCLE NUMBER)

1    2    3    4    5    6    7    8    9 or more

19. (HAND RESPONDENT CARD 3) From what nationality group or groups are you mainly descended? Just call off the number please.

- |                             |   |
|-----------------------------|---|
| <input type="checkbox"/> 1  | <input type="checkbox"/> 11                                     |
| <input type="checkbox"/> 2  | <input type="checkbox"/> 12                                     |
| <input type="checkbox"/> 3  | <input type="checkbox"/> 13                                     |
| <input type="checkbox"/> 4  | <input type="checkbox"/> 14 Don't know (VOLUNTEERED) or refused |
| <input type="checkbox"/> 5  |   |
| <input type="checkbox"/> 6  |   |
| <input type="checkbox"/> 7  |   |
| <input type="checkbox"/> 8  |   |
| <input type="checkbox"/> 9  |   |
| <input type="checkbox"/> 10 |   |

20a. We are interested in finding out how often people are at home to watch TV or listen to the radio. Would you mind telling me whether or not you happened to be at home yesterday (last night/last Saturday) at this particular time?

(INTERVIEWER: SEE INTERVIEWER'S BULLETIN FOR HANDLING THIS QUESTION.)

- Yes, at home
- No, not at home

20b. How about the day (night/Saturday) before at this time?

- Yes, at home
- No, not at home

20c. And how about the day (night/Saturday) before at this time? That was \_\_\_\_\_.

- Yes, at home
- No, not at home

21. And what is your age?

RECORD AGE: \_\_\_\_\_

22. CHECK WHETHER:

- White man
- White woman
- Black man
- Black woman
- Other man (SPECIFY) \_\_\_\_\_
- Other woman (SPECIFY) \_\_\_\_\_

So that my office can check my work in this interview if it wants to, may I have your name, address, and telephone number please?

NAME: \_\_\_\_\_

ADDRESS: \_\_\_\_\_ House No. or RFD Route, St. or Rd., Apt. No.)

CITY: \_\_\_\_\_ STATE: \_\_\_\_\_ ZIP \_\_\_\_\_

TELEPHONE: Area Code \_\_\_\_\_ Phone No. \_\_\_\_\_ y  No tel.

PLACE INTERVIEWER BADGE NUMBER HERE
--

I hereby attest that this is a true and honest interview.

(Interviewer's Signature) \_\_\_\_\_

Date of interview: \_\_\_\_\_

Time interview ended: \_\_\_\_\_

exclude businesses by checking the yellow pages. Choosing telephone numbers at random is called RDD, for *random digit dialing*.<sup>18</sup>

People who do not have phones must be different from the rest of us, and that does cause a bias in telephone surveys. The effect is small, because these days nearly everybody has a phone. On the other hand, about one-third of residential telephones are unlisted. Rich people and poor people are more likely to have unlisted numbers, so the telephone book tilts toward the middle class. Sampling from directories would create a real bias, but random digit dialing gets around this difficulty. In 2005, survey organizations were just beginning to work on the questions raised by cell phones. What about dropped calls? Who pays for air time? What to do with people who have land lines and cell phones?

Non-respondents create problems, as usual. So the Gallup Poll does most of its interviewing on evenings and weekends, when people are more likely to be at home. If there is no answer, the interviewer will call back up to 3 times.<sup>19</sup> (Some designs have up to 15 call-backs; that is better, but more expensive.) For many purposes, results are comparable to those from face-to-face interviews, and the cost is about one-third as much. That is why survey organizations are using the telephone.

## 8. CHANCE ERROR AND BIAS

The previous sections indicated the practical difficulties faced by real survey organizations. People are not at home, or they do not reveal their true preferences, or they change their minds. However, even if all these difficulties are assumed away, the sample is still likely to be off—due to chance error.

To focus the issue, imagine a box with a very large number of tickets, some marked 1 and the others marked 0. That is the population. A survey organization is hired to estimate the percentage of 1's in the box. That is the parameter. The organization draws 1,000 tickets at random without replacement. That is the sample. There is no problem about response—the tickets are all there in the box. Drawing them at random eliminates selection bias. And the tickets do not change back and forth between 0 and 1. As a result, the percentage of 1's in the sample is going to be a good estimate for the percentage of 1's in the box. But the estimate is still likely to be a bit off, because the sample is only part of the population. Since the sample is chosen at random, the amount off is governed by chance:

$$\text{percentage of 1's in sample} = \text{percentage of 1's in box} + \text{chance error.}$$

Now there are some questions to ask about chance errors—

- How big are they likely to be?
- How do they depend on the size of the sample? the size of the population?
- How big does the sample have to be in order to keep the chance errors under control?

These questions will be answered in the next two chapters.

In more complicated situations, the equation has to take bias into account:

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}.$$

Chance error is often called “sampling error;” the “error” comes from the fact that the sample is only part of the whole. Similarly, bias is called “non-sampling error”—the error from other sources, like non-response. Bias is often a more serious problem than chance error, but methods for assessing bias are not well developed. Usually, “bias” means prejudice. However, statistics is a dry subject. For a statistician, bias just means any kind of systematic error in an estimate. “Non-sampling error” is a more neutral term, and may be better for that reason.

### Exercise Set A

1. A survey is carried out at a university to estimate the percentage of undergraduates living at home during the current term. What is the population? the parameter?
2. The registrar keeps an alphabetical list of all undergraduates, with their current addresses. Suppose there are 10,000 undergraduates in the current term. Someone proposes to choose a number at random from 1 to 100, count that far down the list, taking that name and every 100th name after it for the sample.
  - (a) Is this a probability method?
  - (b) Is it the same as simple random sampling?
  - (c) Is there selection bias in this method of drawing a sample?
3. The monthly Gallup Poll opinion survey is based on a sample of about 1,500 persons, “scientifically chosen as a representative cross section of the American public.” The Gallup Poll thinks the sample is representative mainly because—
  - (i) it resembles the population with respect to such characteristics as race, sex, age, income, and education

or

  - (ii) it was chosen using a probability method.
4. In the Netherlands, all men take a military pre-induction exam at age 18. The exam includes an intelligence test known as “Raven’s progressive matrices,” and includes questions about demographic variables like family size. A study was done in 1968, relating the test scores of 18-year-old men to the number of their brothers and sisters.<sup>20</sup> The records of all the exams taken in 1968 were used.
  - (a) What is the population? the sample?
  - (b) Is there any sampling error? Explain briefly.
5. Polls often conduct pre-election surveys by telephone. Could this bias the results? How? What if the sample is drawn from the telephone book?
6. About 1930, a survey was conducted in New York on the attitude of former black slaves towards their owners and conditions of servitude.<sup>21</sup> Some of the interviewers were black, some white. Would you expect the two groups of interviewers to get similar results? Give your reasons.

7. One study on slavery estimated that "11.9% of slaves were skilled craftsmen." This estimate turns out to be based on the records of thirty plantations in Plaquemines Parish, Louisiana.<sup>22</sup> Is it trustworthy? Explain briefly.
8. In one study, the Educational Testing Service needed a representative sample of college students.<sup>23</sup> To draw the sample, they first divided up the population of all colleges and universities into relatively homogeneous groups. (One group consisted of all public universities with 25,000 or more students; another group consisted of all private four-year colleges with 1,000 or fewer students; and so on.) Then they used their judgment to choose one representative school from each group. That created a sample of schools. Each school in the sample was then asked to pick a sample of students. Was this a good way to get a representative sample of students? Answer yes or no, and explain briefly.
9. A study was done on the prevalence of chest diseases in a Welsh coal mining town; 600 volunteers had chest X-rays done.<sup>24</sup> At the time, the two main chest diseases in the town were pneumoconiosis (scarring of the lung tissue due to inhalation of dust) and tuberculosis. The data were analyzed by the order in which the volunteers presented themselves. The percentage with tuberculosis among the first 200 subjects to appear for the examination was probably \_\_\_\_\_ the percentage among the last 200. Fill in the blank, using one of the phrases

(i) about the same as      (ii) quite a bit different from

Explain your reasoning.

10. Television advertising sales are strongly influenced by the Nielsen ratings. In its annual report, the Nielsen organization does not describe how it takes samples. The report does say:<sup>25</sup>

Nielsen, today as in the past, is dedicated to using the newest, most reliable, and thoroughly tested research technologies. This is a commitment to those we serve through the television, cable, and advertising communities . . . .

The Nielsen data in this booklet are estimates of the audiences and other characteristics of television usage as derived from Nielsen Television Index and Nielsen Station Index measurements. The use of mathematical terms herein should not be regarded as a representation by Nielsen that such measurements are exact to precise mathematical values . . . .

Comment briefly.

11. The *San Francisco Examiner* ran a story headlined—

#### 3 IN 10 BIOLOGY TEACHERS BACK BIBLICAL CREATIONISM

*Arlington, Texas.* Thirty percent of high school biology teachers polled believe in the biblical creation and 19 percent incorrectly think that humans and dinosaurs lived at the same time, according to a nationwide survey published Saturday.

"We're doing something very, very, very wrong in biology education," said Dana Dunn, one of two sociologists at the University of Texas, Arlington.

Dunn and Raymond Eve sent questionnaires to 20,000 high school biology teachers selected at random from a list provided by the National Science Teachers Association and received 200 responses . . . .

The newspaper got it wrong. Dunn and Eve did not send out 20,000 question-

naires: they chose 400 teachers at random from the National Science Teachers association list, sent questionnaires to these 400 people, and received 200 replies.<sup>26</sup> Why do these corrections matter?

12. In any survey, a fair number of people who are in the original sample cannot be contacted by the survey organization, or are contacted but refuse to answer questions. A high non-response rate is a serious problem for survey organizations. True or false, and explain: this problem is serious because the investigators have to spend more time and money getting additional people to bring the sample back up to its planned size.

*The answers to these exercises are on pp. A78–79.*

## 9. REVIEW EXERCISES

*Review exercises may cover material from previous chapters.*

1. A survey organization is planning to do an opinion survey of 2,500 people of voting age in the U.S. True or false, and explain: the organization will choose people to interview by taking a simple random sample.
2. Two surveys are conducted to measure the effect of an advertising campaign for a certain brand of detergent.<sup>27</sup> In the first survey, interviewers ask housewives whether they use that brand of detergent. In the second, the interviewers ask to see what detergent is being used. Would you expect the two surveys to reach similar conclusions? Give your reasons.
3. One study on slavery estimated that a slave had only a 2% chance of being sold into the interstate trade each year. This estimate turns out to be based on auction records in Anne Arundel County, Maryland.<sup>28</sup> Is it trustworthy? Explain briefly.
4. In one study, it was necessary to draw a representative sample of Japanese-Americans resident in San Francisco.<sup>29</sup> The procedure was as follows. After consultation with representative figures in the Japanese community, the four most representative blocks in the Japanese area of the city were chosen. All persons resident in those four blocks were taken for the sample. However, a comparison with Census data shows that the sample did not include a high-enough proportion of Japanese with college degrees. How can this be explained?
5. (Hypothetical.) A survey is carried out by the finance department to determine the distribution of household size in a certain city. They draw a simple random sample of 1,000 households. After several visits, the interviewers find people at home in only 653 of the sample households. Rather than face such a high non-response rate, the department draws a second batch of households, and uses the first 347 completed interviews in the second batch to bring the sample up to its planned strength of 1,000 households. The department counts 3,087 people in these 1,000 households, and estimates the average household size in the city to be about 3.1 persons. Is this estimate likely to be too low, too high, or about right? Why?

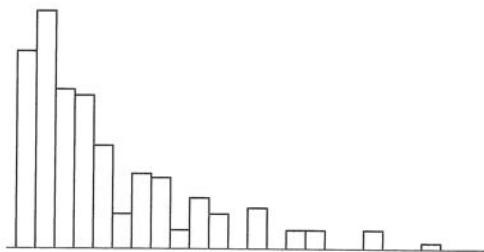
6. “Ecstasy” was a popular drug in the 1990s. It produced a sense of euphoria derisively called the “yuppie high.” One investigator made a careful sample survey to estimate the prevalence of drug use at Stanford University. Two assistants were stationed on the main campus plaza and instructed to interview all students who passed through at specified times. As it turned out, 39% of 369 students interviewed said they had used Ecstasy at least once.<sup>30</sup> Does the investigator’s procedure give a probability sample of Stanford students? Answer yes or no, and explain.

7. A coin is tossed 1,000 times. There are two options:

- (i) To win \$1 if the number of heads is between 490 and 510.
- (ii) To win \$1 if the percentage of heads is between 48% and 52%.

Which option is better? Or are they the same? Explain.

8. Can you tell whether the figure below is a probability histogram or a histogram for data? If so, which is it and why? If you can’t tell, why not?



9. One hospital has 218 live births during the month of January.<sup>31</sup> Another has 536. Which is likelier to have 55% or more male births? Or is it equally likely? Explain. (There is about a 52% chance for a live-born infant to be male.)

10. A coin will be tossed 100 times. You get to pick 11 numbers. If the number of heads turns out to equal one of your 11 numbers, you win a dollar. Which 11 numbers should you pick, and what is your chance (approximately) of winning? Explain.

11. A sorcerer has hidden a Porsche in one of an infinite row of boxes



The sorcerer will let you drive away with the car if you can find it. But you are only allowed to look in 11 boxes. He agrees to give you a hint, by tossing a coin 100 times and counting the number of heads. He will not tell you this number, or the number of the box in which he hid the car. But he will tell you the sum of the two numbers.

- (a) If the sum is 65, which 11 boxes would you look in?  
 (b) As in (a), except replace 65 by 95.  
 (c) What is the general rule?  
 (d) Following this rule, how likely are you to get the Porsche?
12. The *San Francisco Chronicle* reported on a survey of top high-school students in the U.S. According to the survey,

Cheating is pervasive. Nearly 80 percent admitted some dishonesty, such as copying someone's homework or cheating on an exam. The survey was sent last spring to 5,000 of the nearly 700,000 high achievers included in the 1993 edition of *Who's Who Among American High School Students*. The results were based on the 1,957 completed surveys that were returned. "The survey does not pretend to be representative of all teenagers," said *Who's Who* spokesman Andrew Weinstein. "Students are listed in *Who's Who* if they are nominated by their teachers or guidance counselors. Ninety-eight percent of them go on to college."

- (a) Why isn't the survey "representative of all teenagers"?  
 (b) Is the survey representative "of the nearly 700,000 high achievers included in the 1993 edition of *Who's Who Among American High School Students*"? Answer yes or no, and explain briefly.

## 10. SUMMARY

1. A *sample* is part of a *population*.
2. A *parameter* is a numerical fact about a population. Usually a parameter cannot be determined exactly, but can only be estimated.
3. A *statistic* can be computed from a sample, and used to estimate a parameter. A statistic is what the investigator knows. A parameter is what the investigator wants to know.
4. When estimating a parameter, one major issue is accuracy: how close is the estimate going to be?
5. Some methods for choosing samples are likely to produce accurate estimates. Others are spoiled by *selection bias* or *non-response bias*. When thinking about a sample survey, ask yourself:
  - What is the population? the parameter?
  - How was the sample chosen?
  - What was the response rate?
6. Large samples offer no protection against bias.
7. In *quota sampling*, the sample is hand picked by the interviewers to resemble the population in some key ways. This method seems logical, but often

gives bad results. The reason: unintentional bias on the part of the interviewers, when they choose subjects to interview.

8. *Probability methods* for sampling use an objective chance process to pick the sample, and leave no discretion to the interviewer. The hallmark of a probability method: the investigator can compute the chance that any particular individuals in the population will be selected for the sample. Probability methods guard against bias, because blind chance is impartial.

9. One probability method is *simple random sampling*. This means drawing subjects at random without replacement.

10. Even when using probability methods, bias may come in. Then the estimate differs from the parameter, due to bias and chance error:

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}.$$

Chance error is also called "sampling error," and bias is "non-sampling error."