same way as blocking. For example, a retrospective study of music education and grades might match each student who studies an instrument with someone of the same sex who is similar in family income but didn’t study an instrument. We could then compare grades of music students with those of non-music students. The matching would reduce the variation due to income and sex differences.

Blocking for experiments is the same idea as stratifying is for sampling. Both methods group together subjects that are similar and randomize within those groups as a way to remove unwanted variation. (But be careful to keep the terms straight. Don’t say that we “stratify” an experiment or “block” a sample.) We use blocks to reduce variability so we can see the effects of the factors; we’re not usually interested in studying the effects of the blocks themselves.

**EXAMPLE 11.5**

**Blocking**

**RECAP:** In 2007, pet food contamination put cats at risk, as well as dogs. Our experiment should probably test the safety of the new food on both animals.

**QUESTIONS:** Why shouldn’t we randomly assign a mix of cats and dogs to the two treatment groups? What would you recommend instead?

**ANSWERS:** Dogs and cats might respond differently to the foods, and that variability could obscure my results. Blocking by species can remove that superfluous variation. I’d randomize cats to the two treatments (test food and safe food) separately from the dogs. I’d measure their responses separately and look at the results afterward.

---

**Confounding**

Professor Stephen Ceci of Cornell University performed an experiment to investigate the effect of a teacher’s classroom style on student evaluations. He taught a class in developmental psychology during two successive terms to a total of 472 students in two very similar classes. He kept everything about his teaching identical (same text, same syllabus, same office hours, etc.) and modified only his style in class. During the fall term, he maintained a subdued demeanor. During the spring term, he used expansive gestures and lectured with more enthusiasm, varying his vocal pitch and using more hand gestures. He administered a standard student evaluation form at the end of each term.

The students in the fall term class rated him only an average teacher. Those in the spring term class rated him an excellent teacher, praising his knowledge and accessibility, and even the quality of the textbook. On the question “How much did you learn in the course?,” the average response changed from 2.93 to 4.05 on a 5-point scale.9

How much of the difference he observed was due to his difference in manner, and how much might have been due to the season of the year? Fall term in Ithaca, New York (home of Cornell University), starts out colorful and pleasantly warm but ends cold and bleak. Spring term starts out bitter and snowy and ends with blooming flowers and singing birds. Might students’ overall happiness have been affected by the season and reflected in their evaluations?

Unfortunately, there’s no way to tell. Nothing in the data enables us to tease apart these two effects, because all the students who experienced the subdued manner did so during the fall term and all who experienced the expansive manner did so during the spring. When the levels of one factor are associated with the levels of another factor, we say that these two factors are **confounded**.

---

9But the two classes performed almost identically well on the final exam.
In some experiments, such as this one, and some observational studies as well, it’s just not possible to avoid some confounding. Professor Ceci could have randomly assigned students to one of two classes during the same term, but then we might question whether mornings or afternoons were better, or whether he really delivered the same class the second time (after practicing on the first class). Or he could have had another professor deliver the second class, but that would have raised more serious issues about differences in the two professors and concern over more serious confounding.

**EXEMPLARY 11.6**

**Confounding**

**RECAP:** After many dogs and cats suffered health problems caused by contaminated foods, we’re trying to find out whether a newly formulated pet food is safe. Our experiment will feed some animals the new food and others a food known to be safe, and a veterinarian will check the response.

**QUESTION:** Why would it be a bad design to feed the test food to some dogs and the safe food to cats?

**ANSWER:** This would create confounding. We would not be able to tell whether any differences in animals’ health were attributable to the food they had eaten or to differences in how the two species responded.

**A Two-Factor Example**

Confounding can also arise from a badly designed multifactor experiment. Here’s a classic. A credit card bank wanted to test the sensitivity of the market to two factors: the annual fee charged for a card and the annual percentage rate charged. Not wanting to scrimp on sample size, the bank selected 100,000 people at random from a mailing list. It sent out 50,000 offers with a low rate and no fee and 50,000 offers with a higher rate and a $50 annual fee. Not surprising, people preferred the low-rate, no-fee card. In fact, they signed up for that card at over twice the rate as the other offer. And because of the large sample size, the bank was able to estimate the difference precisely. But the question the bank really wanted to answer was “how much of the change was due to the rate, and how much was due to the fee?” Unfortunately, there’s simply no way to separate out the two effects. If the bank had tested all four treatments—low rate with no fee, low rate with $50 fee, high rate with no fee, and high rate with $50 fee—each to 25,000 people, it could have learned about both factors and could have also seen what happens when the two factors occur in combination.

**Lurking and Confounding?**

A lurking variable creates an association between two other variables that tempts you to think that one may cause the other. Recall from the example in Chapter 8, that people’s countries with more TV sets per capita tend to have longer lives. You shouldn’t conclude it’s the TVs “causing” longer life. It’s more likely that a generally higher standard of living allows people to afford more TVs and get better health care, too. Our data revealed an association between TVs and life expectancy, but economic conditions were a likely lurking variable. A lurking variable, then, is usually thought of as a variable associated with both y and x that makes it appear that x may be causing y.
Confounding and lurking variables are very similar. Imagine an observational study hoping to understand the relationship between herbal supplements and patient health finds that patients who take the supplements report fewer colds. However, if they find from their survey that the patients who take the herbal supplements also tend to take larger doses of Vitamin C, we would say that taking Vitamin C is a confounder of herbal supplements. Had we not asked the question at all, and we later found that taking Vitamin C was more effective in preventing colds than the herbal supplement, we might call Vitamin C a lurking variable in the original study.

Both confounding and lurking variables are outside influences that make it harder to understand the relationship we are modeling. It’s important to realize that in any observational study or even in a carefully designed experiment, there may be variables that influence the relationship between that variable and the response other than the ones being studied. You should always be alert for the possible effects of other variables on the coefficients you care about. Be especially wary of variables that you might not have considered.

---

**WHAT CAN GO WRONG?**

- **Don’t give up just because you can’t run an experiment.** Sometimes we can’t run an experiment because we can’t identify or control the factors. Sometimes it would simply be unethical to run the experiment. (Consider randomly assigning students to take—or be graded in—a statistics course deliberately taught to be boring and difficult or one that had an unlimited budget to use multimedia, real-world examples, and field trips.) If we can’t perform an experiment, an observational study may be a good choice.

- **Beware of confounding.** Be aware of variables that may be confounded. In a prospective study, it may be possible to stratify the subjects by levels of one variable. In an experiment, unmeasured confounders will be balanced (on average) by randomization. To include a variable that may be a confounder, it is a good idea to block by the potential confounder to ensure that the levels are balanced. And always think about possible lurking variables that may be influencing the response that aren’t in your study as well.

- **Bad things can happen even to good experiments.** Protect yourself by recording additional information. An experiment in which the air conditioning failed for 2 weeks, affecting the results, was saved by recording the temperature (although that was not originally one of the factors) and estimating the effect the higher temperature had on the response.10

  It’s generally good practice to collect as much information as possible about your experimental units and the circumstances of the experiment. For example, in a nail polish experiment, it would be wise to record details (temperature, humidity) that might affect the durability of the polish on the acrylic nails. Sometimes we can use this extra information during the analysis to reduce biases.

- **Don’t spend your entire budget on the first run.** Just as it’s a good idea to pretest a survey, it’s always wise to try a small pilot experiment before running the full-scale experiment. You may learn, for example, how to choose factor levels more effectively, about effects you forgot to control, and about unanticipated confoundings.

---

Datasets Index

BE = Boxed Example; E = Exercise; JC = Just Checking; RM = Random Matters; SBS = Step-by-Step examples.

A
Accidents, (E): 156–157
Acid Rain, (E): 61
Adoptions, (E): 54
Age of a Tree, (E): 312–313
All Births (1998), (RM): 325–326
Antidepressants, (E): 190
A-Rod, (E): 60
Attendance 2016, (E): 192–193, 223, 224, 302, 583

B
Babybamp, (IE): 430
Ballplayer Births, (BE): 611
Baseball 2016, (E): 119–120, 580
Baseball Salaries, (E): 273
Being Successful, (E): 86
Bike Safety, (E): 161–162
Bird Species, (E): 57, 60
Birth Order, (E): 160
Birthrates, (E): 228–229
Blood Pressure, (E): 91
Bookstore Sales, (E): 220, 221
Boomtowns, (E): 62
Boyle, (E): 273
Brain Size, (E): 679
Brain Waves, (E): 606
Breaks, (E): 273
Burgers, (E): 192, 227, 228
Buy From a Friend, (IE): 551

C
Candy Bars per Serving, (E): 302, 305
Car Origins, (E): 639
Cars, (E): 311
Cars and Trucks, (E): 602
Chips, (E): 316
Chips Ahoy!, (E): 469, 505
Cholesterol and Smoking, (E): 119
Cigarettes, (E): 223, 224
City Climate, (E): 680
City Temperatures, (E): 605
Climate Change, (E): 229, 678, 679
Cloud Seeding, (E): 119, 120, 603
COIL 2016, (E): 114
COLAll 2016, (E): 114
College Values, (E): 93
Commuter Sample, (E): 369
Commuter Sample, (E): 369, 312–313
Companies, (E): 120, 316–317
Computer Labs Fees, (E): 466, 504–505
Cost of Living, (E): 121, 228
Couples, (IE): 592
CPI Worldwide, (BE): 29–30
Cramming, (E): 156, 310
Cranberry Juice, (E): 638–639
Craters, (BE): 647–648
Crawling, (E): 464, 681
Crocodile Lengths, (E): 312
Crowdedness, (E): 271, 272
Cups, (SBS): 100–101

D
Dexterity, (IE): 585
Dirt Bikes, (BE): 244–245, 291
Disk Drives, (E): 187, 220
Doctors and Life Expectancy, (E): 263, 318, (IE): 243–244
Doritos, (E): 469, 505
Dow Jones, (E): 309
Down the Drain, (E): 316
Drivers Licenses, (E): 92
Drug Abuse, (E): 192, 227
Drug Use, (E): 674

E
Ears to Live, (E): 275
Education and Mortality, (E): 683
Education by Age, (E): 641
Egyptians, (E): 579
E-mails, (E): 53, 54

F
Farmed Salmon, (BE): 448, (E): 466, 534
Fertility and Life Expectancy, (E): 270
Fingers and Heights, (IE): 447
Fish Diet, (E): 639, (SBS): 71–73
Flights, (E): 195
Flights On Time, (E): 466, (JC): 101
Floods, (E): 58
Freshman 15, (E): 605–609
Friday the 13th Accidents, (E): 604
Friday the 13th Traffic, (E): 603
Fritos, (E): 541
F-stops, (IE): 178–179
Fuel Economy, (E): 221, 231, 466, 678, 679
Full Moon, (E): 640

G
Gasoline, (E): 606–607
Gas Prices 2016, (E): 116
Gators, (E): 231
Gestation, (E): 268, 269
Global 500 2014, (E): 162
Golf Drives, (E): 59, 470
Grades, (E): 305
Graduate Earnings, (E): 672, 673, 674

H
Hard Water, (E): 157, 190, 230, 580, 583
HDII 2015, (E): 264
Heart Attack Stays, (E): 57
Heights and Weights, (IE): 170, 246
Historical Oil Prices, (IE): 237–238
Hopkins Forest, (E): 162, 271, 300, (IE): 168
Hot Dogs, (E): 675, 676
How Old Is That Tree?, (E): 312–313
Hurricane Frequencies, (E): 636

I
Iliad Injuries, (E): 635
Iliad Weapons, (E): 635
Income and Housing, (E): 191, 224
Industrial Experiment, (E): 119
Indy 500, (E): 15, 16
Inflation, (E): 270
Interest Rates and Mortgages, (E): 191, 224
IQ Brain, (E): 188

J
Job Satisfaction, (E): 580, 606

A-51
Census, 322, 322-323, (BE): 323
Center (of distribution), 31, 31-34
mean and, 32-34
median and, 32, 33-34
Central Limit Theorem (CLT), 442-445, 443
Normal model and, 474
Chi-square models, 613
Chi-square P-values, 613-614
Chi-square statistic, 613
calculating, 616, (BE): 617
Chi-square test(s)
Components, 622
for goodness-of-fit. See Goodness-of-fit
tests
residuals and, 621-622, (BE): 622
tech support for, 631-633
Chi-square test of homogeneity, 618,
618-620, (SBS): 620-621
calculations and, 619-620
Counted Data Condition and, (SBS): 620
Expected Cell Frequency Condition and,
(SBS): 620
Independence Assumption and,
(SBS): 620
Chi-square test of independence, 622,
623-624, (SBS): 625-626
assumptions and conditions for, 624
causation and, 628-629
conclusions for, (BE): 628
Counted Data Condition and, (SBS): 625
Expected Cell Frequency Condition and,
(SBS): 626
Independence Assumption and,
(SBS): 625
mechanics of, (BE): 626-627
residuals and, 627-628
Cluster(s), 327
Cluster sampling, 327, 327-328
Coefficients
correlation, 171, 174
multiple regression and, interpreting,
279-281, (BE): 280
regression, confidence interval for, 654
Collinearity, 637
multiple regression inference and,
657-658
Column percents, 65, 67
Complement Rule, probability and, 379,
(BE): 379
Confidencen(s), 131, 201
bootstrap confidence intervals and,
(SBS): 456
chi-square test of homogeneity and,
(SBS): 620
for chi-square test of independence, 624,
(SBS): 625, (SBS): 626
for confidence intervals for means,
449-450, (BE): 450
for correlation, 172, (SBS): 173
Counted Data Condition, 612, (BE): 612,
for counts, 611-612, (BE): 612
Does the Plot Thicken? Condition. See
Does the Plot Thicken? Condition
Expected Cell Frequency Condition, 612,
Expected Frequency Condition,
(SBS): 615
for goodness-of-fit tests, (SBS): 615
for groups, 544, (BE): 545, (SBS): 546
for hypothesis testing, (BE): 479, (BE):
515, (SBS): 481, (SBS): 490-491
for inference, 478, (BE): 479
for linear regression, 201, 207, 210-211,
(SBS): 212
for multiple regression, 281-283, (SBS):
284
Nearly Normal Condition. See Nearly
Normal Condition
Normal models and, 131, (SBS): 132,
(SBS): 136, (SBS): 137, (SBS): 138,
(SBS): 139, (SBS): 140
for one-sample t-test for the mean,
(SBS): 485
Outlier Condition, 211, 646, (SBS): 649
P-values and, (SBS): 509
for paired data, 587-589, (BE): 588-589
for paired t-intervals, (SBS): 593
for paired t-test, 587-589, (BE):
588-589, (SBS): 590
Quantitative Data Condition, 645
Quantitative Variables Condition, 210,
211
Randomization Condition. See Randomi-
 zation Condition
for regression, 645-646, (BE): 647-648,
(SBS): 649
Sample Size Condition, 612
sampling distribution and, 415, 417, 426,
427
Straight Enough Condition, 210, 211,
for Student’s t, 449-450, 553, (BE): 450,
(BE): 554
Success/Failure Condition. See Success/
Failure Condition
10% Condition, 415, 417, 426, 427,
for two-proportion z-intervals, (SBS):
545, (SBS): 546
for two-proportion z-test, (BE): 550
for two-sample t-test, (BE): 559, (SBS):
557
Conditional distributions, 68, 68-69, 389,
(BE): 69, (BE): 70, (RM): 73-74,
(SBS): 71-73
Conditional probability, 389, 389-390, 508,
(BE): 390
Bayes’ Rule and, 398, 399
General Multiplication Rule and,
395-396
P-values as, 479
reversing, 396-397, (BE): 399,
(SBS): 397-398
Confidence interval(s), 420
for the difference between two means,
(BE): 554-555
hypothesis testing and, 486-489,
(SBS): 489-491
for matched pairs, 592-593,
(SBS): 593-594
for the mean prediction, 659
for means. See Confidence intervals for
means
for proportions. See Confidence intervals
for proportions
regression and, 652, 658-660, (BE): 653,
(BE): 660-661
for the regression coefficient, 654
standard error of, 659-660
tech support for, 460-461
two-sample t-intervals, 552
Confidence interval for the mean predic-
tion, 659
Confidence interval for the regression
coefficient, 654
Confidence intervals for means, 441-471
assumptions and conditions and,
449-450, (BE): 450
bootstrap, 453-455, (BE): 455-456
Central Limit Theorem and, 442-445
constructing, 447-448
degrees of freedom and, 446-447
interpreting, 452-453
one-sample t-interval for the mean and,
(BE): 448, (SBS): 451-452
sample size and, (BE): 444
standard deviation and, (BE): 448
Confidence intervals for proportions,
418-440
critical values and, 420-421
interpreting, 422-423, (BE): 423
one-proportion z-interval and, 421,
(BE): 420
sample size and, 427-429, (BE): 428,
(BE): 429
tech support for, 432-434
Confounding, 355, 355-357, (BE): 356
lurking and, 356-357
Contingency tables, 65, 65-68, 389
displaying, 75-76, (BE): 76-78
statistical software packages, 83-84
tech support for, 83-84
Control, 346
Control groups, 351, 351-354
Control treatments, 351
Convenience samples, 333, (BE): 333
Correlation, 169-176, (RM): 175-176
assumptions and conditions for, 172
causation and, 176-177
changing scales and, (BE): 175
No Outliers Condition, 172, (SBS): 173
T

t-intervals, two-sample, 552

t-test, 654

paired, 587–591, (BE): 591–592,
(SBS): 589–591

Table(s)

ANOVA, 665
cells in, 612
contingency. See Contingency tables
data, 3, 4
frequency, 19
regression, 654
relative frequency, 19
two-way, 618

Table percents, 66, 67
tails (of distribution), 29

10% Condition, (BE): 479
hypothesis testing and, (BE): 479, (SBS):
481, (SBS): 490
Normal model for sampling distribution
of a proportion and, 415, 417, 426, 427

Theoretical probability, 376

TI-83/84 Plus

ci-square tests, 633
comparing distributions, 111
correlation, 434, 461
displaying and summarizing variables, 51
group hypothesis testing, 500
Normal probability plots, 146
paired t-test, 601
re-expression, 260
regression, 219, 259
regression inference, 671
scatterplots, 186
two-sample methods for proportions, 569
timeplots, 169

Transforming data, 105–108, 106, 179–181
for regression, 246–250, (SBS): 251–254
Travelling, 345
control, 351

Tree diagrams, 395, 395–396

Trials, 374
as hypothesis tests, 474–475
Tukey, John W., 26, 421, 454, (BE): 99
Two- and one-tailed alternatives, 477

Two-proportion z-intervals, 545
(SBS): 545–546
Independent Groups Assumption and,
(SBS): 545
Randomization Condition and,
(SBS): 545
Success/Failure Condition and,
(SBS): 546

Two-proportion z-test, 547–548, (BE): 550,
(SBS): 548–549
Independent Groups Assumption and,
(BE): 550
Randomization Condition and, (BE): 550
Success/Failure Condition and, (BE): 550

Two-sample methods for proportions. See
also specific methods
tech support for, 567–569
Two-sample t-intervals, 552
Two-sample t-test, 555, 555–556,
(BE): 559, (SBS): 556–558
Independent Groups Assumption and,
(BE): 559, (SBS): 557
Nearly Normal Condition and, (BE): 559,
(SBS): 557
Randomization Condition and, (BE): 559,
(SBS): 557

Two-sided alternative, 476, 476–477

Two-way tables, 618
Type I errors, 516–517, 521–522
sample size and, (BE): 522
Type II errors, 516–517, 521–522
sample size and, (BE): 522

U

undercoverage, 334
Uniform histograms, 28

Unimodal histograms, 28

Units, 6

V

Valid surveys, 331–332

Variability, sampling, 414

Variables, 5, 7–
associations between, (SBS): 70,
(BE): 173–174
categorical (qualitative; nominal), 5
displaying and summarizing, tech
support for, 47–51
explanatory (predictor), 167
identifier, 6
independent, 69
lurking, 177
ordinal, 6–7
quantitative. See Quantitative variables
response, 167, 345
roles in scatterplots, 167
statistical software packages for
displaying and summarizing, 47–51

Variance, 36

Variance Inflation Factor (VIF), 658

Variation, 37

Venn diagrams, 378, 381, (BE): 382

Voluntary response bias, 332
Voluntary response samples, 332,
(BE): 333

Z

z, Student’s t vs., (BE): 448
z-interval, one-proportion, 421,
(BE): 420
z-scores, 123
combining, (BE): 124
Normal model of, 129