Notes and Data Sources

"About This Book" Notes


2. This summary of the committee’s report was unanimously endorsed by the Board of Directors of the American Statistical Association. The full report is George Cobb, “Teaching statistics,” in L. A. Steen (ed.), *Heeding the Call for Change: Suggestions for Curricular Action*, Mathematical Association of America, 1990, pp. 3–43.

3. A summary of the GAISE “Introductory College Course Guidelines,” which has also been endorsed by the ASA board of directors, appears in *Amstat News*, June 2006, p. 31. See www.amstat.org/education/gaise for details.


"Statistical Thinking" Notes


2. Contributed by Marigene Arnold of Kalamazoo College.


5. The data in Figure 2 are based on a component of the Consumer Price Index, from the Bureau of Labor Statistics Web site: www.bls.gov. I converted the index number into cents per gallon using retail price information from the Energy Information Agency site: www.eia.doe.gov.


Chapter 1 Notes


6. Our eyes do respond to area, but not quite linearly. It appears that we perceive the ratio of two bars to be about the 0.7 power of the ratio of their actual areas. See W. S. Cleveland, *The Elements of Graphing Data*, Wadsworth, 1985, pp. 278–284.

7. See Note 5.

8. Data from Gary Community School Corporation, courtesy of Celeste Foster, Purdue University.

9. See Note 5.


13. See Note 3.


19. Monthly stock returns from the Web site of Professor Kenneth French of Dartmouth, mba.tuck.dartmouth.edu/pages/faculty/ken.french. A fine point: the data are actually the “excess returns” on stocks, the actual returns less the small monthly returns on Treasury bills.

20. National Climatic Data Center storm events data base, sciencepolicy.colorado.edu/sourcebook/tornadoes.html.


22. Found online at earthtrends.wri.org.


Notes and Data Sources


Chapter 2 Notes

1. From 2003 American Community Survey, at the Bureau of the Census Web site, www.census.gov. The data are a subsample of the 13,194 individuals in the ACS North Carolina sample who had travel times greater than zero.

2. This isn’t a mathematical theorem. The mean can be less than the median in right-skewed distributions that take only a few values, many of which lie exactly at the median. The rule almost never fails for distributions taking many values, and counterexamples don’t appear clearly skewed in graphs even though they may be slightly skewed according to technical measures of skewness. See Paul T. von Hippel, “Mean, median, and skew: correcting a textbook rule,” Journal of Statistics Education, 13, No. 2 (2005), online journal, www.amstat.org/publications/jse.


5. Figure 2.2 displays the daily returns for 301 market days starting on March 1, 2004, for the CREF Equity Index Fund and the TIAA Real Estate Fund. Daily price data for these funds are at www.tiaa-cref.org. Returns can be easily calculated because dividends are incorporated in the daily prices rather than given separately.


10. From the University of Miami athletics Web site, hurricanesports.collegesports.com.


**Chapter 3 Notes**

1. See Note 8 for Chapter 1.


4. Data provided by Darlene Gordon, Purdue University.


**Chapter 4 Notes**


3. Based on T. N. Lam, “Estimating fuel consumption from engine size,” *Journal of Transportation Engineering*, 111 (1985), pp. 339–357. The data for 10 to 50 km/h are measured; those for 60 and higher are calculated from a model given in the paper and are therefore smoothed.


5. A careful study of this phenomenon is W. S. Cleveland, P. Diaconis, and R. McGill, “Variables on scatterplots look more highly correlated when the scales are increased,” *Science*, 216 (1982), pp. 1138–1141.


8. Data provided by Robert Dale, Purdue University.


10. See Note 8.

11. Compiled from Fidelity data by the *Fidelity Insight* newsletter, 20 (2004), No. 1.


**Chapter 5 Notes**


2. See Note 1 for Chapter 4.


12. From a presentation by Charles Knauf, Monroe County (New York) Environmental Health Laboratory.


15. See Note 10 for Chapter 1.


Chapter 6 Notes
5. Siem Oppe and Frank De Charro, “The effect of medical care by a helicopter trauma team on the probability of survival and the quality of life of hospitalized victims,” Accident Analysis and Prevention, 33 (2001), pp. 129–138. The authors give the data in Example 6.6 as a “theoretical example” to illustrate the need for their more elaborate analysis of actual data using severity scores for each victim.

Chapter 7 Notes
Notes and Data Sources

20. Data provided by Samuel Phillips, Purdue University.

Chapter 8 Notes

1. See Note 3 for “Statistical Thinking.”
4. Jeffrey G. Johnson et al., “Television viewing and aggressive behavior during adolescence and adulthood,” Science, 295 (2002), pp. 2468–2471. The authors use statistical adjustments to control for the effects of a number of lurking variables. The association between TV viewing and aggression remains significant. Statistical adjustment had been used in
the observational studies that supported hormone replacement (Example 8.1) as well, a warning not to place too much trust in these methods.


7. The regulations that govern seat belt survey design can be found at www.nhtsa.dot.gov. Details on the Hawaii survey are in Karl Kim et al., Results of the 2002 Highway Seat Belt Use Survey, at www.state.hi.us/dot.


Chapter 9 Notes

1. Details of the Carolina Abecedarian Project, including references to published work, can be found online at www.epg.unc.edu/~abc.


10. Shailja V. Nigdikar et al., “Consumption of red wine polyphenols reduces the susceptibility of low-density lipoproteins to oxidation in vivo,” *American Journal of Clinical Nutrition*, 68 (1998), pp. 258–265. (There were in fact only 30 subjects, some of whom received more than one treatment with a four-week period intervening.)


Data Ethics Notes


5. Dr. Hennekens’s words are from an interview in the Annenberg/Corporation for Public Broadcasting video series *Against All Odds: Inside Statistics*.


Chapter 10 Notes


4. Information from [www.ncsu.edu/class/grades](http://www.ncsu.edu/class/grades).

5. See Note 3 for Chapter 1.


Chapter 11 Notes


2. Strictly speaking, the formula $\sigma/\sqrt{n}$ for the standard deviation of $\bar{x}$ assumes that we draw an SRS of size $n$ from an infinite population. If the population has finite size $N$, this standard deviation is multiplied by $\sqrt{1-(n-1)/(N-1)}$. This “finite population correction”
Notes and Data Sources

approaches 1 as $N$ increases. When the population is at least 20 times as large as the sample, the correction factor is between about 0.97 and 1. It is reasonable to use the simpler form $\sigma/\sqrt{n}$ in these settings.


4. Elroy Dimson, Paul Marsh, and Mike Staunton, _Triumph of the Optimists: 101 Years of Global Investment Returns_, Princeton University Press, 2002. Sophisticates will note that for compounding over several years we want the geometric mean return, which was 6.7%.

Chapter 12 Notes


4. Information about Internet users comes from sample surveys carried out by the Pew Internet and American Life Project, www.pewinternet.org.


8. The probabilities given are realistic, according to the fundraising firm SCM Associates, scmassoc.com.

Chapter 13 Notes


Chapter 14 Notes


2. This and similar results of Gallup polls are from the Gallup Organization Web site, www.gallup.com.

Chapter 15 Notes


2. See Note 6 for Chapter 14.

3. See Note 15 for Chapter 8.


6. See Note 4 for Chapter 7.


Chapter 16 Notes


2. See Note 16 for Chapter 8.


**Chapter 17 Notes**


3. See Note 24 for Chapter 1.


6. See Note 17 for Chapter 1.

7. Data simulated from a Normal distribution with $\mu = 98.2$ and $\sigma = 0.7$. These values are based on P. A. Mackowiak, S. S. Wasserman, and M. M. Levine, “A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich,” *Journal of the American Medical Association*, 268 (1992), pp. 1578–1580.


19. See Note 4 for Chapter 12.


**Chapter 18 Notes**

1. Note 2 for Chapter 11 explains the reason for this condition in the case of inference about a population mean.


3. See Note 3 for Chapter 14.


5. R. A. Berner and G. P. Landis, “Gas bubbles in fossil amber as possible indicators of the major gas composition of ancient air,” *Science*, 239 (1988), pp. 1406–1409. The 95% $t$ confidence interval is 54.78 to 64.40. A bootstrap BCa interval is 55.03 to 62.63. So $t$ is reasonably accurate despite the skew and the small sample.

6. See Note 3 for Chapter 14.

7. For a qualitative discussion explaining why skewness is the most serious violation of the Normal shape condition, see Dennis D. Boos and Jacqueline M. Hughes-Oliver, “How large does $n$ have to be for the $Z$ and $t$ intervals?” *American Statistician*, 54 (2000), pp. 121–128. Our recommendations are based on extensive computer work. See, for example, Harry O. Posten, “The robustness of the one-sample $t$-test over the Pearson system,” *Journal of Statistical Computation and Simulation*, 9 (1979), pp. 133–149; and E. S. Pearson and N. W. Please, “Relation between the shape of population distribution and the robustness of four simple test statistics,” *Biometrika*, 62 (1975), pp. 223–241.

8. For more advanced users, a good way to ascertain if the $t$ procedures are safe is to compare the 95% confidence interval produced by $t$ with the BCa interval from a bootstrap with at least 1000 resamples. For (b) the $t$ interval is 29.428 to 32.254 and a BCa interval is 29.106 to 31.894. For (c), on the other hand, $t$ gives 38.93 to 40.49 and BCa gives 38.97 to 40.44. These results confirm the judgment that $t$ is safe for (c) but not for (b).

9. TUDA results for 2003 from the National Center for Education Statistics, nces.ed.gov/nationsreportcard.


11. Data from the “wine” data base in the archive of machine learning data bases at the University of California, Irvine, ftp.ics.uci.edu/pub/machine-learning-databases.

12. Data provided by Chris Olsen, who found the information in scuba-diving magazines.

Notes and Data Sources


15. See Note 8 for Chapter 7.


17. Data provided by Jason Hamilton, University of Illinois. The study is reported in Evan H. DeLucia et al., “Net primary production of a forest ecosystem with experimental CO2 enhancement,” *Science*, 284 (1999), pp. 1177–1179. No method for inference can be trusted with \( n = 3 \). In this study, each observation is very costly, so the small \( n \) is inevitable.


19. See Note 4 for Chapter 15.


21. Harry B. Meyers, “Investigations of the life history of the velvetleaf seed beetle, *Althaea folkertsi* Kingsolver,” MS thesis, Purdue University, 1996. The 95% \( t \) interval is 1227.9 to 2507.6. A 95% bootstrap BCa interval is 1444 to 2718, confirming that \( t \) inference is inaccurate for these data.

22. Data provided by Timothy Sturm.

23. Lianng Yuh, “A biopharmaceutical example for undergraduate students,” manuscript, no date.

Chapter 19 Notes


6. See Note 14 for Chapter 2.

7. See Note 7 for Chapter 2.

8. See the extensive simulation studies in Harry O. Posten, “The robustness of the two-sample \( t \)-test over the Pearson system,” *Journal of Statistical Computation and Simulation*, 6 (1978),
Chapter 20 Notes


2. Strictly speaking, the formula $\sqrt{p(1-p)/n}$ for the standard deviation of $\hat{p}$ assumes that we draw an SRS of size $n$ from an infinite population. If the population has finite size $N$, this standard deviation is multiplied by $\sqrt{1-(n-1)/(N-1)}$. This “finite population correction” approaches 1 as $N$ increases. When the population is at least 20 times as large as the sample, the correction factor is between about 0.97 and 1. It is reasonable to use the simpler form $\sqrt{p(1-p)/n}$ in these settings. See also Note 2 for Chapter 11.
3. This rule of thumb is based on study of computational results in the papers cited in Note 6 and discussion with Alan Agresti. We strongly recommend using the plus four interval.

4. The quotation is from page 1104 of the article cited in Note 1.

5. See Note 19 for Chapter 19.


8. From Alan Agresti and Brian Caffo, “Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures,” The American Statistician, 45 (2000), pp. 280–288. When can the plus four interval be safely used? The answer depends on just how much accuracy you insist on. Brown and coauthors (see Note 6) recommend $n \geq 40$. Agresti and Coull demonstrate that performance is almost always satisfactory in their eyes when $n \geq 5$. Our rule of thumb $n \geq 10$ allows for confidence levels C other than 95% and fits our philosophy of not insisting on more exact results than practice requires. The big point is that plus four is very much more accurate than the standard interval for most values of $p$ and all but very large $n$.


10. See Note 3 for Chapter 19.


12. In fact, $P$-values for two-sided tests are more accurate than those for one-sided tests. Our rule of thumb is a compromise to avoid the confusion of too many rules.


17. See Note 1 for Chapter 15.


19. See Note 16 for Chapter 8.


**Chapter 21 Notes**

1. The National Longitudinal Study of Adolescent Health interviewed a stratified random sample of 27,000 adolescents, then reinterviewed many of the subjects six years later, when most were aged 19 to 25. These data are from the Wave III reinterviews in 2000 and 2001, found at the Web site of the Carolina Population Center, [www.cpc.unc.edu](http://www.cpc.unc.edu).

2. See Note 4 for Chapter 12.


4. The plus four method is due to Alan Agresti and Brian Caffo. See Note 8 for Chapter 20.


9. Data courtesy of Raymond Dumett, Purdue University.


12. Based on Alan G. Sanfey et al., “The neural basis of economic decision-making in the ultimatum game,” *Science*, 300 (2003), pp. 1755–1758. The paper reports a chi-square test (equivalent to a two-sided z test). This analysis is incorrect for the paper’s data, as there were in fact only 19 participants, each appearing twice in each row of the table given in the exercise. Exercise 21.17 therefore amends the data, assuming 76 participants, so that the elementary analysis is correct.

Notes and Data Sources

15. See Note 10 for Chapter 7.
16. See Note 18 for Chapter 20.

Chapter 22 Notes

1. See Note 16 for Chapter 19.
8. See Note 5 for Chapter 17. The exercises are simplified, in that the measures reported in this paper have been statistically adjusted for “sociodemographic status.”
11. These data were originally collected by L. M. Linde of UCLA but were first published by M. R. Mickey, O. J. Dunn, and V. Clark, “Note on the use of stepwise regression in detecting outliers,” Computers and Biomedical Research, 1 (1967), pp. 105–111. The data have been used by several authors. I found them in N. R. Draper and J. A. John, “Influential observations and outliers in regression,” Technometrics, 23 (1981), pp. 21–26.
12. Data provided by Charles Hicks, Purdue University.


14. Data provided by Marigene Arnold, Kalamazoo College.

15. Data provided by Corinne Lim, Purdue University, from a student project supervised by Professor Joseph Vanable.

16. See Note 6 for Chapter 21.


Chapter 23 Notes


3. See Note 10 for Chapter 6.

4. Data from www.gewinternet.org. The counts are not exact because the report gave only percents rounded to the nearest whole percent.

5. There are many computer studies of the accuracy of chi-square critical values for $X^2$. Our guideline goes back to W. G. Cochran (1954). Later work has shown that it is often conservative in the sense that if the expected cell counts are all similar and the degrees of freedom exceed 1, the chi-square approximation works well for an average expected count as small as 1 or 2. Our guideline protects against dissimilar expected counts. It has the added advantage that it is safe in the $2 \times 2$ case, where the chi-square approximation is least good. So our guideline is helpful for beginners—there is no single condition that is not conservative and applies to $2 \times 2$ and larger tables with similar and dissimilar expected cell counts. There are exact procedures that (with software) should be used for tables that do not satisfy our guideline. For a survey, see Alan Agresti, “A survey of exact inference for contingency tables,” Statistical Science, 7 (1992), pp. 131–177.

6. See Note 18 for Chapter 21.

7. See Note 4 for Chapter 6.


10. From the GSS data base at the University of Michigan, webapp.icpsr.umich.edu/GSS.


13. From the GSS data archive at the Survey Documentation and Analysis site at the University of California, Berkeley, sda.berkeley.edu.


15. See Note 18 for Chapter 20.


17. See Note 13.


19. See Note 17 for Chapter 8.

20. See Note 1 for Chapter 21.

21. See Note 9 for Chapter 7.

22. See Note 1 for Chapter 21.

23. See Note 10 for Chapter 7.


**Chapter 24 Notes**


2. See Note 6 for Chapter 4.

3. See Note 4 for Chapter 4.


8. See Note 16 for Chapter 5.
9. From Table S2 in the online supplement to Antonio Dell’Anno and Roberto Danovaro, “Extracellular DNA plays a key role in deep-sea ecosystem functioning,” Science, 309 (2005), p. 2179.

10. See Note 1 for Chapter 4.

11. See Note 8 for Chapter 7.

12. Based on Marion E. Dunshee, “A study of factors affecting the amount and kind of food eaten by nursery school children,” Child Development, 2 (1931), pp. 163–183. This article gives the means, standard deviations, and correlation for 37 children, from which the data in Exercise 24.41 are simulated.

13. See Note 20 for Chapter 7.


**Chapter 25 Notes**

1. See Note 6 for Chapter 2.

2. See Note 9 for Chapter 22.


4. See Note 7 for Chapter 2.


6. See Note 14 for Chapter 2.


10. See Note 18 for Chapter 18.


14. The data and the full story can be found in the Data and Story Library at lib.stat.cmu.edu. The original study is by Faith Loven, “A study of interlist
Notes and Data Sources


15. Data provided by Matthew Moore.
16. See Note 13 for Chapter 22.
17. See Note 3 for Chapter 14.
18. See Note 1 for Chapter 19.
19. See Note 3 for Chapter 11.