

# Appendix A

## Bibliography

- [1] Source: [www.stats4schools.gov.uk](http://www.stats4schools.gov.uk), November 10, 2009.
- [2] B. Ritz, F. Yu, G. Chapa, and S. Fruin, “Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993,” *Epidemiology*, vol. 11, no. 5, pp. 502–511, 2000.
- [3] J. McGowan, “Health Education: Does the Buteyko Institute Method make a difference?,” *Thorax*, vol. 58, 2003.
- [4] Gallagher, Visser, Sepulveda, Pierson, and H. Heymsfield, “How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups?,” *American Journal of Epidemiology*, vol. 143, no. 3, pp. 228–239, 1996.
- [5] R. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.
- [6] T. Allison and D. Cicchetti, “Sleep in mammals: ecological and constitutional correlates,” *Arch. Hydrobiol.*, vol. 75, p. 442, 1975.
- [7] Source: Harvard Business Review, [http://blogs.hbr.org/cs/2009/06/new\\_twitter\\_research\\_men\\_follo.html](http://blogs.hbr.org/cs/2009/06/new_twitter_research_men_follo.html), April 1, 2011.
- [8] Source: Yahoo News, [http://news.yahoo.com/s/ac/20110315/tc\\_ac/8066912\\_happy\\_birthday\\_twitter](http://news.yahoo.com/s/ac/20110315/tc_ac/8066912_happy_birthday_twitter), April 1, 2011.
- [9] Source: CIA Factbook, <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2091rank.html>, October 22, 2010.
- [10] B. Turnbull, B. Brown, and M. Hu, “Survivorship of heart transplant data,” *Journal of the American Statistical Association*, vol. 69, pp. 74–80, 1974.
- [11] R. Rabin, “Risks: Smokers found more prone to dementia,” October 29 2010. <http://www.nytimes.com/2010/11/02/health/research/02risks.html>.
- [12] D. Graham, R. Ouellet-Hellstrom, T. MaCurdy, F. Ali, C. Sholley, C. Worrall, and J. Kelman, “Risk of acute myocardial infarction, stroke, heart failure, and death in elderly medicare patients treated with rosiglitazone or pioglitazone,” *JAMA*, vol. 304, no. 4, p. 411, 2010.

- [13] U.S. Census Bureau, 2005-2009 American Community Survey.
- [14] Majority of Republicans No Longer See Evidence of Global Warming, October 27, 2010, <http://people-press.org/reports/questionnaires/669.pdf>.
- [15] USPSTF, "Screening for breast cancer: U.s. preventive services task force recommendation statement," *Annals of Internal Medicine*, vol. 151, pp. 716–726, 2009.
- [16] J. A. Paulos, "Mammogram math," December 2009. *New York Times*, 13 December 2009.
- [17] S. Johnson and D. Murray, "Empirical Analysis of Truck and Automobile Speeds on Rural Interstates: Impact of Posted Speed Limits," in *Transportation Research Board 89th Annual Meeting*, 2010.
- [18] Source: SAMHSA, Office of Applied Studies, National Survey on Drug Use and Health, 2007 and 2008, <http://www.oas.samhsa.gov/NSDUH/2k8NSDUH/tabs/Sect2peTabs1to42.htm#Tab2.5B>.
- [19] Source: Public Fact Sheet, Chickenpox (Varicella), <http://www.mass.gov/Eeohhs2/docs/dph/cdc/factsheets/chickenpox.pdf>.
- [20] Source: What Frightens America's Youth?, <http://www.gallup.com/poll/15439/What-Frightens-Americas-Youth.aspx>.
- [21] G. Heinz, L. Peterson, R. Johnson, and C. Kerk, "Exploring relationships in body dimensions," *Journal of Statistics Education*, vol. 11, no. 2, 2003.
- [22] Source: [http://www.ets.org/Media/Tests/GRE/pdf/gre\\_0809\\_interpretingscores.pdf](http://www.ets.org/Media/Tests/GRE/pdf/gre_0809_interpretingscores.pdf).
- [23] NAEP Data Explorer, April 16, 2011.
- [24] A. Romero-Corral, V. Somers, J. Sierra-Johnson, R. Thomas, M. CollazoClavell, J. Korinek, T. Allison, J. Batsis, F. Sert-Kuniyoshi, and F. Lopez-Jimenez, "Accuracy of body mass index in diagnosing obesity in the adult general population," *International Journal of Obesity*, vol. 32, no. 6, pp. 959–966, 2008.
- [25] Road Rules: Re-Testing Drivers at Age 65?, <http://maristpoll.marist.edu/34-road-rules-re-testing-drivers-at-age-65>.
- [26] Civil War at 150: Still Relevant, Still Divisive, <http://pewresearch.org/pubs/1958/civil-war-still-relevant-and-divisive-praise-confederate-leaders-flag>.
- [27] Is College Worth It?, <http://pewresearch.org/pubs/1993/survey-is-college-degree-worth-cost-debt-college-presidents-higher-education-system>.
- [28] Public option gains support, October 20, 2009.
- [29] Perceived Insufficient Rest or Sleep Among Adults United States, 2008, October 30, 2009.
- [30] L. Ellis and C. Ficek, "Color preferences according to gender and sexual orientation," *Personality and Individual Differences*, vol. 31, no. 8, pp. 1375–1379, 2001.
- [31] Drilling for oil and natural gas off the coast of California, <http://www.surveyusa.com>.

- [32] Poll: 4 in 5 Support Full-Body Airport Scanners, November 15, 2010, [http://www.cbsnews.com/8301-503544\\_162-20022876-503544.html](http://www.cbsnews.com/8301-503544_162-20022876-503544.html).
- [33] Four in 10 Americans Believe in Strict Creationism, December 17, 2010, <http://www.gallup.com/poll/145286/Four-Americans-Believe-Strict-Creationism.aspx>.
- [34] R. Schmidt, R. Hansen, J. Hartiala, H. Allayee, L. Schmidt, D. Tancredi, F. Tassone, and I. Hertz-Picciotto, "Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism," *Epidemiology*, vol. 22, no. 4, p. 476, 2011.
- [35] R. Rabin, "Patterns: Prenatal vitamins may ward off autism," June 13 2011. [http://www.nytimes.com/2011/06/14/health/research/14patterns.html?\\_r=1&ref=research](http://www.nytimes.com/2011/06/14/health/research/14patterns.html?_r=1&ref=research).
- [36] Facebook privacy, <http://www.surveyusa.com>.
- [37] Source: Fueleconomy.gov, <http://www.fueleconomy.gov/mpg/MPG.do?action=browseList2&make=Toyota&model=Prius>.
- [38] Source: R Dataset, <http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/chickwts.html>.
- [39] Source: U.S. Department of Energy, Fuel Economy Data, <http://www.fueleconomy.gov/feg/download.shtml>.
- [40] R. Oldham-Cooper, C. Hardman, C. Nicoll, P. Rogers, and J. Brunstrom, "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake," *The American journal of clinical nutrition*, vol. 93, no. 2, p. 308, 2011.
- [41] Gallup, "Americans' views of egypt sharply more negative," February 8, 2011. [http://www.gallup.com/poll/File/146006/Egypt\\_Favorability\\_Feb\\_08\\_2011.pdf](http://www.gallup.com/poll/File/146006/Egypt_Favorability_Feb_08_2011.pdf).
- [42] CDC, "2008 assisted reproductive technology report." <http://www.cdc.gov/art/ART2008/index.htm>.
- [43] Source: Mythbusters, Season 3, Episode 28, <http://www.yourdiscovery.com/video/mythbusters-top-10-is-yawning-contagious>.
- [44] D. Hand, *A handbook of small data sets*. Chapman & Hall/CRC, 1994.
- [45] Source: R Dataset, <http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/trees.html>.
- [46] J. Benson, "Season of birth and onset of locomotion: Theoretical and methodological implications," *Infant behavior and development*, vol. 16, no. 1, pp. 69–81, 1993.
- [47] Source: Association of Turkish Travel Agencies, [http://www.tursab.org.tr/en/statistics/foreign-visitors-figure-tourist-spendings-by-years\\_1083.html](http://www.tursab.org.tr/en/statistics/foreign-visitors-figure-tourist-spendings-by-years_1083.html).
- [48] Source: Starbucks.com, data collected on March 10, 2011, <http://www.starbucks.com/menu/nutrition>.
- [49] Source: American Fact Finder, generated on December 27, 2010, <http://www.factfinder.census.gov>.

- [50] J. Malkevitch and L. Lesser, *For All Practical Purposes: Mathematical Literacy in Today's World*. WH Freeman & Co, 2008.
- [51] Source: R Dataset, <http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/ToothGrowth.html>.
- [52] Source: <http://www.ma.hw.ac.uk/~stan/aod/library>.
- [53] Source: R Dataset, <http://stat.ethz.ch/R-manual/R-patched/library/MASS/html/quine.html>.

## Appendix B

# End of chapter exercise solutions

### 1 Introduction to data

**1.1** (a) Control: 56%. Treatment: 70%. (b) There is a 14% difference between the pain reduction rates in the two groups. It appears that patients in the treatment group are more likely to show improvement and, at a first glance, acupuncture appears to be an effective treatment for migraines. (c) It's hard to say. The difference is somewhat large, but the sample is somewhat small.

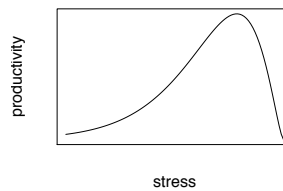
**1.3** (a) 143,196 eligible subjects who were born in Southern California between 1989 and 1993. (b) The variables are measurements of CO, NO<sub>2</sub>, ozone, and particulate matter less than 10 $\mu$ m (PM<sub>10</sub>) collected at air-quality-monitoring stations as well as the birth weights of the babies. All of these variables are continuous numerical variables. (c) Does air pollution exposure have an effect on preterm births?

**1.5** (a) 202 black and 504 white adults who resided in or near New York City, were ages 20-94 years, and had BMIs of 18-35 kg/m<sup>2</sup>. (b) Age (numerical, continuous), sex (categorical), ethnicity (categorical), weight, height, waist and hip circumference, length of tibia, body density and volume, total body water (numerical, continuous). (c) How useful is BMI for predicting body fatness across age, sex and ethnic groups?

**1.7** (a) A participant in the survey. (b) 1,691 participants. (c) gender (gender of the participant), age (age of the participant, in years), marital (marital status of the participant), grossIncome (gross income of the participant, in £), smoke (whether or not the participant smokes), amtWeekends (number of cigarettes smoked on weekend, # of cigarettes / day), amtWeekdays (number of cigarettes smoked on a week day, # of cigarettes / day).

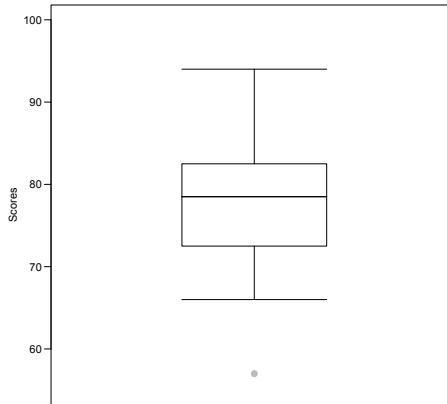
**1.9** gender (categorical), age (originally numerical, continuous, though it was recorded as a discrete numerical variable), marital-Status (categorical), grossIncome (originally numerical, continuous, but recorded as categorical), smoke (categorical), amtWeekends (numerical, discrete), amtWeekdays (numerical, discrete).

**1.11** We would expect productivity to increase as stress increases, but up to a point, after that productivity would decrease as stress continued to increase. The exact shape of your plot may be a little different.



**1.13** (a) Population mean,  $\mu_x = 5.5$ ; sample mean,  $\bar{x} = 6.25$ . (b) Population mean,  $\mu_x = 52$ ; sample mean,  $\bar{x} = 58$ .

**1.15** (a) Decrease. (b) 73.6. (c) The new score,  $x_{25}$ , is more than 1 standard deviation away from the previous mean, and this will tend to increase the standard deviation of the data. While possible, it is mathematically rather tedious to calculate the new standard deviation.



**1.17** The distribution of amount of cigarettes smoked on weekends and on weekdays are both right skewed. The median of both distributions is between 10 and 15 cigarettes, the first quartile is between 5 and 10 cigarettes, and the third quartile is between 15 and 20 cigarettes. Hence the IQR of both distributions is roughly about 10 cigarettes. There are potential outliers above 40 cigarettes per day, giving both distributions a long right tail. We can also see that there are more respondents who smoke only a few cigarettes (0 to 5) on the weekdays, about 80 people, than on weekends, about 60 people. Another feature that is visible from the histograms are peaks at 10 and 20 cigarettes. This may be because most people do not keep track of exactly how many cigarettes they smoke, but round their answers to half a pack (10 cigarettes) or a whole pack (20 cigarettes). Due to these peaks, the distributions could be classified as bimodal.

**1.19**  $s_{amtWeekends} = 0, s_{amtWeekdays} = 4.18$ . Variability of the amount of cigarettes smoked is higher on weekdays than on the weekends for this sample.

**1.21** (a) 6 (b) 6.5

**1.23** Plot below.

**1.25** (a) The distribution is unimodal and symmetric with a mean around 60 and a standard deviation of roughly 3; matches the box plot (2). (b) The distribution is uniform and values range from 0 to 100; matches box plot (3) which shows a symmetric distribution in this range. Also, each 25% chunk of the box plot have about the same width and there are no suspected outliers. (c) The distribution is unimodal and right skewed with a median between 1 and 2. The IQR of the distribution is roughly 1; matches box plot (1).

**1.27** (a) Since median is defined as the 50<sup>th</sup> percentile and about 50% of the data is in the first bar, we would expect median to be between 0 and 20. Q1 is also between 0 and 20 as the 25<sup>th</sup> percentile is in the first bar as well. Q3, defined as the 75<sup>th</sup> percentile, is located between 40 and 60. (b) The distribution is right-skewed, so the long tail will pull the mean above the median.

**1.29** It appears that marathon times decreased greatly between 1970-1975 and remained somewhat steady thereafter. Males consistently had shorter marathon times than females throughout the years. From the box plots of males and females, we could tell that males ran faster “on average”, however, we could not tell that the winning male time for each year was better than the winning female time. We also could not tell from the histogram or the box plot that marathon times have been decreasing for males and females throughout the years.

**1.31** (a) The distribution is right skewed with potential outliers on the positive end, therefore the median and the IQR are appropriate measures of center and spread. (b) The distribution is somewhat symmetric and probably does not have outliers, therefore the mean and the standard deviation are appropriate measures of center and spread. (c) The distribution would be right skewed. There would be some students who do not consume any alcohol but this is the minimum (there cannot be students who consume fewer than 0 drinks). There would be a few students who consume many more drinks than their peers, giving the distribution a long right tail. Due to the skew, the median and the IQR would be appropriate measures of center and spread. (d) The distribution would be right skewed. Most employees would make something on the order of the median salary, but we would expect to have some high level executives making a lot more. The distribution would have a long right tail, and the median and the IQR would be more more appropriate measures of center or spread.

**1.33** (a) As well as the order of the categories, we can also see the relative frequencies in the bar plot. These proportions are not readily available in the pie chart. (b) None. (c) Bar plot, so that we can also see the relative frequencies of the categories in this graph.

**1.35** (a) Proportion of patients who are alive at the end of the study is higher in the treatment group than in the control group. Therefore survival is not independent of whether or not the patient got a transplant. (b) The shape of the distribution of survival times in both groups is right skewed with outliers on the high end. The median survival time for the control group is much lower than the median survival time for the treatment group; patients who got a transplant typically lived longer. The maximum survival time for the treatment group is much higher (about 5 years) than the maximum survival time for the control group. Even though the maximum survival time for

the control group is about 4 years, this observation is an outlier. Overall, very few patients without transplants made it beyond a year while nearly half of the transplant patients survived at least one year. It should also be noted that while the first and third quartiles of the treatment group is higher than those for the control group, the IQR for the treatment group is much bigger, indicating that there is more variability in survival times in the treatment group.

**1.37** (a) The population is all adults 20 and older living in the greater New York City area. The sample is the 202 black and 504 white men and women who resided in or near New York City and had BMIs of 18-35 kg/m<sup>2</sup>. (b) The population is all Californians registered to vote in the 2010 midterm elections. The sample is the 1000 registered California voters who were surveyed for this study.

**1.39** (a) This is an observational study. (b) Wealth is one lurking variable. Countries with individuals who can widely afford internet probably also can afford basic medical care. (Note: Answers may vary.)

**1.41** (a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds his sample may not be representative of the population. (b) Convenience sample. Under coverage bias, his sample may not be representative of the population since it consists only of his friends.

**1.43** (a) Non-responders are most likely parents who have busier schedules and have difficulty spending time with their kids after school. (b) The women who are not reached 3 years later are most likely renters (as opposed to homeowners) who may be in a lower socio-economic status. (c) There is no control group and there may be lurking variables. For example, it may be that these people who go running are generally healthier and/or do other exercises.

**1.45** No, this was an observational study, and we cannot make such a causal statement based on an observational study.

**1.47** Prepare two cups for each participant one containing regular Coke and the other containing Diet Coke. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (regular) and B (diet). (Be sure to randomize A and B for each trial!) Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates how much she liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of the beverage to make this a double-blind experiment. (Answers may vary.)

**1.49** (a) Experiment. (b) Treatment: exercise twice a week, control: no exercise. (c) Yes, the blocking variable is age. (d) No. (e) Since this is an experiment, we can make a causal statement. Since the sample is random, the causal statement can be generalized to the population at large. However, we should be cautious about making a causal statement because of a possible placebo effect. Note that this study could not actually be conducted since people cannot be required to participate in a clinical trial.

**1.51** (a) False. Instead of comparing counts, we should compare percentages of people in each group who suffered a heart attack. (b) True. (c) False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. (We cannot say changing the drug a person is on affects her risk, which is why part (b) is true.) (d) True.

---



---

## 2 Probability

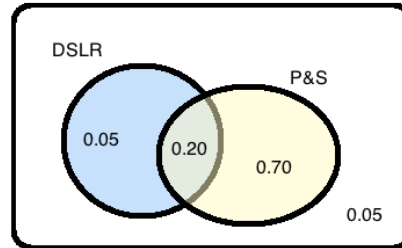
**2.1** False. The tosses are independent trials.

**2.3** (a) 10 tosses. With a low number of flips the variability in the number of heads observed is much larger, so a result further from 50% is more likely. (b) 100 tosses. With more flips, the observed proportion of heads would probably be closer to 50% and therefore above 40%. (c) 100 tosses. The

more flips, the less variability away from 50%. (d) 10 tosses. Fewer flips mean more volatility and a greater chance of getting far from 50% and below 30%.

**2.5** (a)  $1/1024$ . (b)  $1/1024$ . (c)  $1023/1024$ .

**2.7** (a) Figure below. (b) 5% (c) 70% (d) 95% (e) 5% (f) No, there are bloggers who own both types of cameras.



**2.9** (a) Not mutually exclusive. If the class is not graded on a curve, then independent. If graded on a curve, dependent. (b) Not mutually exclusive, most likely dependent. (c) No. See the answer to (a) when the course is not graded on a curve.

**2.11** (a) 0.26. (b) 0.23. (c) Assuming that the education level of the husband and wife are independent, 0.0598. (d) Independence, which may not be a reasonable assumption since people often marry others with a comparable level of education.

**2.13** (a) Sum greater than 1. (b) OK mathematically. (c) Sum less than 1. (d) Negative probabilities make no sense. (e) OK. (f) Probabilities cannot be less than 0 or greater than 1.

**2.15** Approximate answers are OK. Answers are only estimates based on the sample. (a) 0.42. (b) 0.15. (c) 0.37. (d) 0.06.

**2.17** (a) The distribution is right skewed, with a median between \$35,000 and \$49,999. The IQR of the distribution is about \$27,500. There are probably outliers on the high end due to the nature of the data. (b) 62.2%. (c) Assuming gender and income are independent: 25.5%. (d)  $P(\text{less than } \$50,000 \text{ and female}) = 29.4\%$ . The independence assumption does not appear to be valid. If gender and income were independent, we would expect the 25.5% of the sample to be female

and make less than \$50,000, but actually a higher proportion fall into this category.

**2.19** No,  $P(\text{DSLR} \mid \text{point\&shoot}) = 0.22$ , which is not equal to  $P(\text{DSLR})$ .

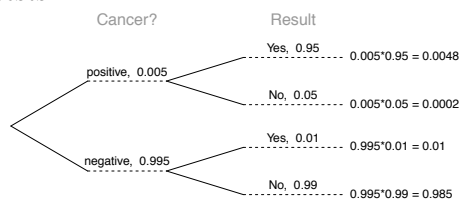
**2.21** (a) 0.2825. (b) 0.1905. (c) 0.4167. (d) No, because  $P(\text{black hair} \mid \text{brown eyes}) \neq P(\text{black hair} \mid \text{blue eyes})$ . (Other explanations are possible.)

**2.23** (a) 0.65. (b) 0.72. (c) Under the assumption of independence of gender and hamburger preference: 0.468. While it is possible there is some mysterious connection between burger choice and finding a partner, independence is probably a reasonable assumption. (e) 0.514.

**2.25** Female, most cats smaller than 2.5kg are female.

**2.27** 0.6049.

**2.29** (a) Tree diagram below. (b) 0.68. (c) 0.32. (d) Your test results have come in. While the test came back positive, this is not conclusive. A positive test result can occur even when a patient has no disease; occasionally a test will be wrong. For this reason, we will need to run some additional tests.



**2.31** (a) 0.3. (b) 0.3. (c) 0.3. (d) 0.09. (e) Yes, each draw is from the same set of marbles.

**2.33** (a) 0.0909. (b) 0.3182. (c) 0.4545. (d) 0. (e) 0.2879.

**2.35** 0.0519.

**2.37** (a) 13. (b) No, this would be unreliable. The students are not a random sample.

**2.39** (a) Table below. Expected winnings: \$3.59. SD: 3.37. (b) EV: -\$1.41, SD: \$3.37. (c) No. The expected net profit is negative, so on average you expect to lose money.

Event	3 hearts	3 blacks	Else
$X$	\$50	\$25	\$0
$P(X)$	0.0129	0.1176	0.8695
$X * P(X)$	0.65	2.94	0
$(X - E(X))^2 P(X)$	0.1115	0.0497	11.2062

**2.41** (a) EV: -\$0.16, SD: \$2.99. (b) EV: -\$0.16, SD: \$1.73. (c) Expected values are the same but the standard deviations are different. The standard deviation from the game where winnings and losses are tripled is higher, making this game riskier.

**2.43** (a) Table to the right. Expected winnings: -\$0.54 (b) No, he is expected to lose money on average.

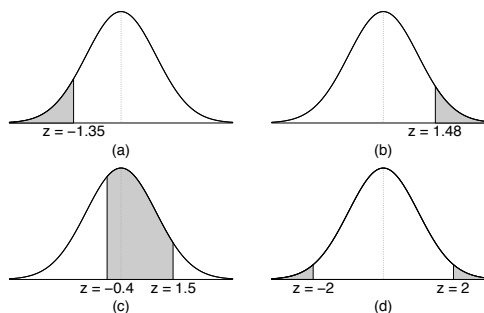
Event	2,...,9	J, Q, K	Ace	A♣
$X$	-2	1	3	23
$P(X)$	0.6923	0.2308	0.0577	0.0192

**2.45** \$4.26.

**2.47** (a) Mean: \$3.90, SD: \$0.34. (b) Mean: \$27.30, SD: \$0.89.

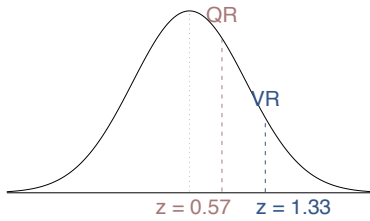
### 3 Distributions of random variables

**3.1** Plots below. (a) 0.0885. (b) 0.0694. (c) 0.5886. (d) 0.0456.



**3.3** (a) Verbal:  $N(\mu = 462, \sigma = 119)$ , Quant:  $N(\mu = 584, \sigma = 151)$ . (b)  $Z_{VR} = 1.33$ ,  $Z_{QR} = 0.57$ . Plots below. (c) She scored 1.33 standard deviations above the mean on the Verbal Reasoning section and 0.57 standard deviations above the mean on the Quantitative Reasoning section. (d)  $\text{Perc}_{VR} = 91\%$ ,  $\text{Perc}_{QR} = 72\%$ . (e) Verbal Reasoning. (f) VR: 9%, QR: 28%. (g) We cannot compare the raw scores since they are on different scales. Her scores will be measured relative to the merits of other students on each exam, so it is helpful to consider the Z score. Comparing her percentiles is more

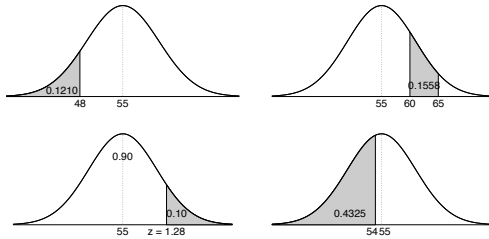
appropriate for determining how well she did compared to others.



**3.5** Answers to (b) and (c) would not change, though we would not draw a Normal curve on which to show these scores. We could not answer parts (d) and (e) since the normal probability table is only valid for the normal model.

**3.7** (a) 711. (b) 400.

**3.9** Figures below. (a) 0.1210. (b) 0.1558. (c) 62.68 inches. (d) 43.25%

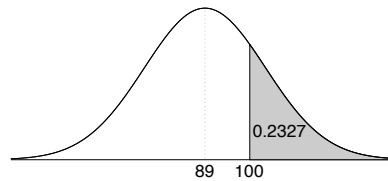


**3.11** (a) 0.1401. (b) 70.6°F or colder.

**3.13** (a) 0.67. Using 0.68 is also okay, but your answers for part (c) will differ a little from the listed solution. (b)  $x = \$1800$ ,  $\mu = \$1650$ . (c)  $\sigma = \$223.88$ .

**3.15** (a) 0.2327. Figure below. (b) If you are bidding on only one auction and set a maximum bid price that is too low, chances are someone will outbid you and you won't win the auction. If your maximum bid price is too high, you may win the auction but you may be paying more than is necessary. If you are bidding on more than one auction and your maximum bid price is too low, chances are you won't win any of the auctions. However, if your maximum bid price is too high, you may win more than one auction and end up with multiple copies of the book. (c) An answer roughly equal to the 10<sup>th</sup> percentile would be reasonable. Regrettably, no percentile cutoff point guarantees

beyond any possible event that you win at least one auction. However, you may pick a higher percentile if you want to be more sure of winning an auction. (d) Using the 10<sup>th</sup> percentile: \$69.80. Answers may vary but should correspond to the answer given in part (c).



**3.17** 70% of the data are within 1 SD, 95% are within 2 SD, and 100% are within 3 SD of the mean. The data approximately follow the 68-95-99.7% Rule.

**3.19** The distribution is unimodal, symmetric, and approximately follows the 68-95-99.7% Rule. The superimposed normal curve seems to approximate the distribution pretty well. The points on the Normal probability plot also seem to follow a straight line. There is one possible outlier on the lower end that is apparent in both graphs, but it is not too extreme. We can say that the distribution is nearly normal.

**3.21** No, in poker cards are dealt without replacement, and they have more than two categories.

**3.23** Approximate answers are OK. (a) 0.13. (b) 0.12. (c)  $\mu = 2.04$ ,  $\sigma = 1.46$ . (d)  $\mu = 3.33$ ,  $\sigma = 2.79$ . (e) When  $p$  was smaller, i.e. the event was rarer, the expected number of trials before a success and the standard deviation increased.

**3.25** (a) 0.096. (b)  $\mu = 8$ ,  $\sigma = 7.48$ .

**3.27** (a) Yes, it meets the four required conditions. (b) 0.203. (c) 0.203. (d) 0.167. (e) 0.997.

**3.29** (a)  $\mu = 34.85$ ,  $\sigma = 3.25$ . (b) Yes, since 45 is more than 3 standard deviations from the mean. (c) 0.0015 (an answer of 0.0009 would be okay if using a normal approximation, since the conditions for the approximation are satisfied). In part (b), we had determined that it would be unusual to ob-

serve 45 or more 18-20 year olds who have consumed alcoholic beverages among a random sample of 50, and the we calculated a very low probability for this event.

**3.31** (a) 0.5160. (b) 0.1234. (c) 0.8483. (d) No, otherwise there is a 15.17% chance that 2 or more will be afraid of spiders in any particular tent.

**3.33** (a) 0.109. (b) 0.219. (c) 0.137. (d) 0.551. (e) 0.084. (f) Since 2 is  $\frac{2-4}{1.06} = -1.89$  standard deviations below the expected number of brown eyed children, strictly speaking this would not be considered unusual. However, it should be noted that the z-score for this value is pretty close to 2, making this observation borderline unusual.

**3.35** The probability model is below.

Y	-3	-1	1	3
P(Y)	0.1458	0.3936	0.3543	0.1063

**3.37** (a)  $(1/5)*(1/4)*(1/3)*(1/2)*(1/1) = 1/(5!) = 1/120$ . (e)  $120 = 5!$ . (c)  $8! = 40,320$ .

**3.39** (a) 0.0804 using the geometric distribution. (b) 0.0322 using the binomial distribution. (c) 0.0193 using the negative binomial distribution.

**3.41** (a) Negative Binomial ( $n = 4$ ,  $p = 0.55$ ): Of the four trials considered here, the last trial must be a success and there were exactly 2 successes. (b) 0.1838. (c)  $\binom{3}{1} = \frac{3!}{2!1!} = 3$ . (d) In the binomial model we have no restrictions on the outcome of the last trial while in the negative binomial model the last trial is fixed. Therefore we are interested in the number of ways of orderings of the other  $k - 1$  successes in the first  $n - 1$  trials.

**3.43** (a) Poisson with  $\lambda = 75$ . (b)  $\mu = \lambda = 75$ ,  $\sigma = \sqrt{\lambda} = 8.66$ . (c) No, since 60 is within 2 standard deviations of the mean.

**3.45**  $P(X = 70) = \frac{75^{70}e^{-75}}{70!} = 0.0402$

**4.3** The point estimates are the corresponding sample values. (a)  $\bar{x} = 13.65$ , median = 14. (b)  $s = 1.91$ ,  $IQR_{estimate} = 2$ . (c) Use the Z score to evaluate ( $Z_{16} = 1.23$ ,  $Z_{18} = 2.28$ ), so 18 credits is unusually high but 16 is not, where we use 2 standard deviations from the mean as a cutoff for deciding what is unusual.

**4.5** No, sample point estimates only approximate the population parameter, and they vary from one sample to another.

**4.7** Standard error,  $SE_{\bar{x}} = \frac{1.91}{\sqrt{100}} = 0.191$ .

**4.9** (a)  $SE_{\bar{x}} = 2.89$  (b) The Z score is 1.73 (absolute value is less than 2), so \$80 is consistent.

**4.11** (a) Independence is met by the random sampling assumption and that the sample is less than 10% of the population. The sample size is also sufficiently large. We cannot check the assumption that the distribution isn't extremely skewed. (b) (19.862, 20.058). (c) We are 90% confident that the true mean amount of coffee in Starbucks venti cups is between 19.862 ounces and 20.058 ounces. (d) 90% of random samples of size 50 will yield confidence intervals that capture the true mean amount of coffee in Starbucks venti cups. (e) Yes, 20 ounces is included in the interval. (f) A 95% confidence interval would be wider. All else kept constant, when confidence level increases so does the margin of error and hence the interval becomes wider. We cast a wider interval.

**4.13** (a) Less. (b) We can infer from the sample statistics that the distribution is skewed, so no we cannot. (c) The only condition that may not be met for normality of the mean relates to skew: it is unclear if the distribution is extremely skewed or not. We'll suppose the skew is strong but not too extreme, something we may like to look into further. Solution: 0.0985. (d) Decreases the standard error by a factor  $\sqrt{2}$ .

**4.15** When the confidence level increases, so does the margin of error and the width of the interval. A wide interval may be undesirable even if the confidence level is higher.

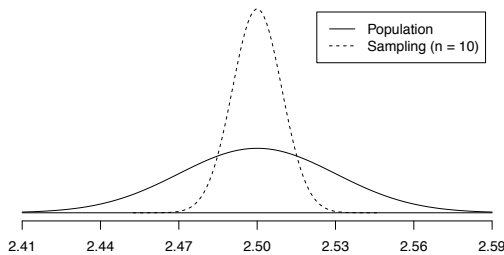
---

## 4 Foundations for inference

**4.1** (a) Mean. (b) Mean. (c) Proportion. (d) Mean. (e) Proportion.

**4.17** (a) False, since we need only check whether the skew is not too extreme. (b) False, we are 100% sure the average for *these* patients is in this interval. (c) True. (d) False, the confidence interval is not about sample means. (e) False, as the confidence level increases, so does the width of the interval. (f) True. (g) False, since in calculation of the standard error we divide the standard deviation by square root of the sample size, we would need to quadruple the sample size.

**4.19** (a) 0.0004. (b) Since the sample is random and the 10% condition is met, we can assume that how much one penny weighs is independent of another. Since the population distribution is normal, and hence not extremely skewed, sampling distribution of means will be nearly normal even though  $n < 50$ .  $N(\mu = 2.5, \sigma_{\bar{x}} = 0.0095)$ . (c) Approximately 0. (d) Plot below. (e) The sample or sampling distributions would not be approximately normal.



**4.21** (a) From the histogram:  $P(X > 5) = \frac{350+100+25+20+5}{3000} = \frac{500}{3000} = 0.17$ . It's okay if your answer differs a little. (b) Two different answers are reasonable. 1) The conditions are reasonably met. We know the population standard deviation, so we can know the standard error (SD of  $\bar{x}$ ) with certainty. The population distribution is also only slightly skewed, so a sample of 15 would probably have a sampling distribution for the mean that is nearly normal. Solution: 0.0956. 2) If you had said the normality condition for  $\bar{x}$  was questionable because the population distribution was not normal, that is also an acceptable answer. (c) Assumptions/conditions are certainly met. Solution: 0.1788.

**4.23** (a)  $H_0: \mu = 8$  (On average New Yorkers sleep 8 hrs a night),  $H_A: \mu < 8$  (On average New Yorkers sleep less than 8 hrs a night). (b)  $H_0: \mu = 15$  (The average amount of company time spent not working is 15 minutes),  $H_A: \mu > 15$  (The average amount of company time spent not working is greater than 15 minutes).

**4.25** The hypotheses should be about the population mean ( $\mu$ ), not the sample mean. If he believes that \$1.3 million is an overestimation, the alternative hypothesis should be *less than* and not *greater than*. The correct way to set up these hypotheses is as follows:  $H_0: \mu = \$1.3 \text{ million}$ ,  $H_A: \mu < \$1.3 \text{ million}$ .

**4.27** (a) 180 minutes is not in the interval, so this is implausible. (b) 2.2 hours (132 minutes) is in the interval, so we conclude the estimated wait time of 2.2 hours is reasonable. (c) A 99% confidence interval will be wider than a 95% confidence interval. Hence even without calculating the interval we can tell that 132 minutes would be in it.

**4.29** (a)  $H_0$ : Anti-depressants do not work for the treatment of Fibromyalgia.  $H_A$ : Anti-depressants work for the treatment of Fibromyalgia. (b) Concluding that anti-depressants work for the treatment of Fibromyalgia when they actually do not. (c) Concluding that anti-depressants do not work for the treatment of Fibromyalgia when they actually do. (d) If she makes a Type I error, she will continue taking medication that does not actually treat her disorder. If she makes a Type II error, she will stop taking medication that could treat her disorder.

**4.31** (a) Yes, if we assume there isn't too much skew, which is certainly reasonable once we realize the possible percentages, which must be between 0% and 100%, are bounded within 3 standard deviations from the mean. (b)  $H_0: \mu = 0.25$ ,  $H_A: \mu \neq 0.25$ .  $Z = -7.71 \rightarrow$  two-sided p-value  $\approx 0$ . Reject  $H_0$ : the evidence indicates that the percentage of time college students spend on the internet for coursework has changed over the last decade. (c) If the percentage of time

college students spend on the Internet for course work has actually remained at 25%, the probability of getting a random sample of 238 college students where the average percentage of time they spend on the Internet for course work is 10% or less or 40% or more is approximately 0. (d) Type I, since we may have incorrectly rejected  $H_0$ .

**4.33**  $H_0: \mu = 7$ ,  $H_A: \mu \neq 7$ .  $Z = -1.04 \rightarrow$ single tail = 0.1492  $\rightarrow$  p-value =  $2 * 0.1492 = 0.2984$ . There isn't sufficient evidence that the average lifespan of all ball bearings produced by this machine is not 7 hours. The manufacturer's claim is not implausible.

**4.35**  $\bar{x} = 36.05$ .

**4.37** (a) The distribution is unimodal and right skewed, median is between 5 and 10 years old, and the IQR is roughly 10. There are potential outliers on the higher end. (b) When the sample size is small, the sampling distribution is right skewed, just like the population distribution. As the sample size increases, the sampling distribution gets more and more unimodal and symmetric, just like the CLT suggests.

**4.39** (a) If the skew is not too strong, the assumptions are met. (b)  $H_0: \mu = 432$ ,  $H_A: \mu < 432$ .  $Z = -3.28 \rightarrow$  p-value (single tail) = 0.0005. Since the p-value  $< \alpha$ , we reject  $H_0$ . There is evidence that the average amount savings of all customers who switch their insurance is less than \$432. (c) Yes, the insurance company's claim may be an overestimate since the hypothesis test result indicated there was strong evidence that the average savings is less than the advertised amount. (d) (\$376.47, \$413.53). (e) Yes, the hypothesis test was statistically significant and \$432 was not in the confidence interval.

**4.41** (a) The only condition we cannot check is for extreme skew. Here, we will assume this is not an issue; in practice, this is something we should verify. (b)  $H_0: \mu = 500$ ,  $H_A: \mu \neq 500$ .  $Z = -3.86 \rightarrow$ single tail  $\approx 0 \rightarrow$  p-value  $\approx 2 * 0 = 0$ . Since the p-value  $< \alpha$  (0.05), we reject  $H_0$ . The data provide strong evidence that the average in-

crease in reading speed is not 500% (it is below 500% based on the data). (c) No, the company's claim of an average of 500% increase in reading speed does not appear to be accurate. (d) 371.88% to 458.12%. (e) Yes. The hypothesis test rejected that the average increase was 500%, and 500% was not in the confidence interval.

**4.43**  $n \geq 693$ .

## 5 Large sample inference

**5.1** (a) Hypothesis test for paired data. (b)  $H_0: \mu_{diff} = 0$  (There is no difference in average daily high temperature between January 1, 1968 and January 1, 2008),  $H_A: \mu_{diff} > 0$  (Average daily high temperature in January 1, 1968 was lower than average daily high temperature in January, 2008.) (c) Independence is satisfied since we have a random sample that is less than 10% of the possible locations we could collect such measurements in the continental U.S. There is also a one-to-one correspondence between the observations in the data set, making it appropriate for a paired analysis. The sample size is sufficiently large ( $n = 51$ ). If we had the data in hand, we would also check for extreme skew.  $Z = 1.60$ , p-value = 0.0548. (d) Fail to reject  $H_0$ . The data do not provide strong evidence of temperature warming in the continental US. However, it should be noted that the p-value is very close to 0.05. (e) Type II. If we made such an error and concluded that there isn't strong evidence for temperature warming in the continental US, but in reality average temperature on January 1, 2008 is higher than average temperature on January 1, 1968. (f) Yes.

**5.3** (a)  $H_0: \mu_B = \mu_A$ , another way to write this is  $\mu_B - \mu_A = 0$  (The population mean of number of cigarettes smoked per day did not change after the Surgeon General's report),  $H_A: \mu_1 > \mu_2$ , another way to write this is  $\mu_1 - \mu_2 > 0$  (The population mean of number of cigarettes smoked per day decreased after the Surgeon General's report) (b) Independence is satisfied since we have

two random samples that are less than 10% of their respective populations. The sample sizes are sufficiently large. If we had the data in hand, we would also check for extreme skew.  $Z = 1.89$ , p-value = 0.0294. (c) Reject  $H_0$ . There is sufficient evidence that the number of cigarettes smoked per day decreased after the Office of the Surgeon General's report. (d) No, we cannot make a causal connection because this is observational data. (e) Type I, since we may have incorrectly rejected  $H_0$ .

**5.5** Independence is satisfied since we have two independent random samples that are each less than 10% of the population. The sample sizes are sufficiently large. If we had the data in hand, we would also check for extreme skew. We are 90% confident that the average score in 2004 was 0.16 to 5.84 points lower than the average score in 2008.

**5.7**  $H_0: \mu_M = \mu_W$ ,  $H_A: \mu_M < \mu_W$ .  $Z = -97.35$ , p-value  $\approx 0$ . Reject  $H_0$ . The data provide strong evidence that average body fat percentage for women is higher.

**5.9** (a) False, for proportions need to check success/failure condition, not  $n \geq 50$ . (b) True. (c) False, only 1.65 standard errors away from the mean, and we use 2 as a cutoff for what is called unusual. (d) True. (e) False, standard error would decrease only by a factor of  $\sqrt{2}$ .

**5.11** (a) True. (b) False, standard error would decrease only by a factor of  $\sqrt{2}$ . (c) True. (d) True. (e) False, success/failure condition is not satisfied.

**5.13** (a) Parameter: proportion of all graduates from this university who found a job within one year of graduating. Point estimate: 0.87. (b) Independence is satisfied since we have a random sample that is less than 10% of the population. Normality is satisfied since the success-failure condition is met. CI: (0.837, 0.903). (c) We are 95% confident that the true proportion of graduates from this university who found a job within one year of completing their undergraduate degree is between 83.7% and 90.3%. (d) 95% of random samples of 400

would produce a confidence interval that includes the true proportion of students at this university who found a job within one year of graduating from college. (e) It would be wider. (f) It would be narrower.

**5.15** (a) She needs a minimum of 3,394 subjects and therefore needs to set aside a minimum of \$67,880. (b) It will be wider.

**5.17** (a)  $ME = 1.96 * \sqrt{0.66 * 0.34 / 1,1018} \approx 0.03$ . (b) No, for two reasons. The point estimate is slightly below 67%, and 67% is contained in the interval.

**5.19** (a)  $H_0: p = 0.5$ ,  $H_A: p > 0.5$ . Assuming the sample is random. The sample is simple random and from <10% of the population, so independence is reasonable. The success/failure condition is also met.  $Z = 4.66$ , p-value  $\approx 0$ . Since the p-value is small, we reject  $H_0$ . The data provide strong evidence that majority of the Americans think the Civil War is still relevant. (b) If in fact only 50% of Americans thought the Civil War is still relevant, the probability of obtaining a random sample of 1,507 Americans where 56% think it is still relevant would be approximately 0. (c) We are 90% confident that 54% to 58% of all Americans think that the Civil War is still relevant. This agrees with the conclusion of the earlier hypothesis test since the interval lies above 50%.

**5.21** (a)  $H_0: p = 0.5$ ,  $H_A: p < 0.5$ . The assumptions and conditions are satisfied.  $Z = -0.73$ , p-value = 0.2327. Since the p-value is large, we fail to reject  $H_0$ . The data do not provide strong evidence that less than half of American adults who decide to not go to college make this decision because they cannot afford college. (b) Yes, since we failed to reject  $H_0$ .

**5.23** (a) The assumptions and conditions are satisfied. We are 80% confident that the 44.5% to 51.5% of all Americans who decide not to go to college do so because they cannot afford it. This agrees with the conclusion of the earlier hypothesis test since the interval includes 50%. (b) 1,818.

**5.25** (a)  $H_0: p = 0.3$ ,  $H_A: p > 0.3$ . The assumptions and conditions are satisfied.  $Z = 1.89$ , p-value = 0.0294. Since the p-value is small, we reject  $H_0$ . The data provide strong evidence that the rate of sleep deprivation for New Yorkers is higher than the rate of sleep deprivation in the population at large. (b) If in fact 30% of New Yorkers were sleep deprived, the probability of getting a random sample of 300 New Yorkers where more than 105 are sleep deprived would be 0.0294.

**5.27** (a)  $H_0: p = 0.18$ ,  $H_A: p \neq 0.18$ . The assumptions and conditions are satisfied.  $Z = 0.74$ , p-value = 0.4592. Since the p-value is large, we fail to reject  $H_0$ . The data do not provide strong evidence that the percentage of students at this university who smoke has changed over the last five years. (b) Type II, since we may have incorrectly failed to reject  $H_0$ .

**5.29** (a)  $H_0: p = 0.65$ ,  $H_A: p > 0.65$ . Assuming that the 250 < 10% of high school graduates at this school district, all conditions and assumptions are satisfied.  $Z = 1.26$ , p-value = 0.1038. Since the p-value is large, we fail to reject  $H_0$ . The data do not provide strong evidence that the percentage of students in this rural school district who go out of state for college has increased. (b) If in fact 65% of students in this school district went out of state for college, the probability of getting a random sample of 250 students where 172 or more of them go out of state for college would be 0.1038.

**5.31** (a) The assumptions and conditions are satisfied. 95% CI: (0.138, 0.270). (b) We are 95% confident that the proportion of students from the rural school district who plan to go out of state for college is 13.8% to 27% higher than the proportion of students from the urban school district who do.

**5.33** (a)  $H_0: p_D = p_I$ ,  $H_A: p_D > p_I$ . The assumptions and conditions are satisfied.  $Z = 11.29$ , p-value  $\approx 0$ . Since the p-value is very small, we reject  $H_0$ . The data provide strong evidence that the proportion of Democrats who support the plan is higher

than the proportion of Independents who support the plan. (b) Type I, since we may have incorrectly rejected  $H_0$ . (c) No, rejecting the null hypothesis of  $p_1 = p_2$  is equivalent to rejecting that  $p_1 - p_2 = 0$ . Therefore we would not expect a confidence interval for the difference between the two proportions to include 0. (d) We are 95% confident that the proportion of Democrats who support the plan is 23% to 33% higher than the proportion of Independents who do. (e) True.

**5.35** The assumptions and conditions are satisfied. We are 95% confident that the proportion of Californians who are sleep deprived is 1.7% less to 0.1% more than the proportion of Oregonians who are sleep deprived. Since the confidence interval includes 0, we would not reject a null hypothesis that the two population proportions equal to each other.

**5.37** (a) True. (b) False, the interval only estimates the difference in population parameters. (c) False, to get the 95% confidence interval for  $(p_{\text{placebo}} - p_{\text{medication}})$ , all we have to do is to swap the bounds of the original confidence interval and take their negatives. (d) True. (e) False, the confidence interval for the difference between the proportions of success includes 0, so we cannot reject the hypothesis of no difference.

**5.39** (a) College grads: 35.2%. Non-grads: 33.9%. (b)  $H_0: p_{CG} = p_{NCG}$ ,  $H_A: p_{CG} \neq p_{NCG}$ . The assumptions and conditions are satisfied.  $Z = 0.37$ , p-value = 0.7114. Since the p-value is large, we fail to reject  $H_0$ . The data do not provide strong evidence of a difference between the proportions of college graduates and non-college graduates who support off-shore drilling in California.

**5.41** (a) We are 90% confident that the proportion of Republicans who support the use of full-body scans at airports is 3% lower to 7% higher than the proportion of Democrats who do. (b) No, this does not prove it; though the data does not provide strong evidence to the contrary.

**5.43** (a)  $H_0$ : The distribution of the format of the book used by the students follows the professor's predictions.  $H_A$ : The distribution of the format of the book used by the students does not follow the professor's predictions. (b)  $E_{hard\ copy} = 75.6$ ,  $E_{print} = 31.5$ ,  $E_{online} = 18.9$ . (c) We are not told explicitly that the sample is random, however, we have no reason to believe that this class is not representative of all introductory statistics students. We can safely assume that  $126 < 10\%$  of all introductory statistics students. We may think it is reasonable to suppose the students are independent. However, the professor probably should have included a question asking whether the student decisions relied on any other students' decisions when they purchased, printed, or read the book online. All expected counts are at least 10. Format of the book used is a categorical variable. (d)  $\chi^2 = 2.32$ ,  $df = 2$ , p-value  $> 0.3$ . (e) Since the p-value is large, we reject  $H_0$ . The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.

**5.45** (a) 47.5. (b) 296.6. (c) 21.0.

**5.47** (a)  $H_0$ : There is no difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy.  $H_A$ : There is some difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy. (b)  $E_{row\ 1,col\ 1} = 95.2$ ,  $E_{row\ 1,col\ 2} = 85.8$ ,  $E_{row\ 2,col\ 1} = 158.8$ ,  $E_{row\ 2,col\ 2} = 143.2$ . The assumptions and conditions are satisfied.  $\chi^2 = 8.85$ ,  $df = 1$ ,  $0.001 < \text{p-value} < 0.005$ . Since the p-value is small, we reject  $H_0$ . There is strong evidence a difference in the rates of autism of children of mothers who did and did not use prenatal vitamins during the first three months before pregnancy. (c) The title of this newspaper article makes it sound like using prenatal vitamins can prevent autism, which is a causal statement. Since this is an observational study, we cannot make causal statements based on

the findings of the study. A more accurate title would be "Mothers who use prenatal vitamins before pregnancy are found to have children with a lower rate of autism".

**5.49**  $H_0$ : The opinion of college grads and non-grads is not different on the topic of drilling for oil and natural gas off the coast of California.  $H_A$ : Opinions regarding the drilling for oil and natural gas off the coast of California has an association with college education.  $E_{row\ 1,col\ 1} = 151.5$ ,  $E_{row\ 1,col\ 2} = 134.5$ ,  $E_{row\ 2,col\ 1} = 162.1$ ,  $E_{row\ 2,col\ 2} = 143.9$ ,  $E_{row\ 3,col\ 1} = 124.5$ ,  $E_{row\ 3,col\ 2} = 110.5$ . The assumptions and conditions are satisfied.  $\chi^2 = 11.46$ ,  $df = 2$ ,  $0.001 < \text{p-value} < 0.005$ . Since the p-value is small, we reject  $H_0$ . There is strong evidence that there is some difference in rate of support for drilling for oil and natural gas off the Coast of California based on whether or not the respondent graduated from college. Support for off-shore drilling and having graduated from college do not appear to be independent.

## 6 Small sample inference

**6.1** (a)  $t_{41}^* = 1.68$  (b)  $t_{20}^* = 2.53$  (c)  $t_{28}^* = 2.05$  (d)  $t_{11}^* = 3.11$

**6.3** With a larger critical value, the confidence interval ends up being wider.

**6.5** (a)  $H_0$ :  $\mu = 8$  (New Yorkers sleep 8 hrs per night on average.),  $H_A$ :  $\mu < 8$  (New Yorkers sleep less than 8 hrs per night on average.) (b) Independence is satisfied since the sample is random and less than 10% of the population. The distribution doesn't appear to be strongly skewed.  $T = -1.75$ ,  $df = 24$ . (c) If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hrs per night or less is between 0.025 and 0.05. (d) Reject  $H_0$ , the data provide strong evidence that New Yorkers sleep less than 8 hours per night on average. (e) No.

- 6.7** (a) We are 90% confident that New Yorkers on average sleep 7.47 to 7.99 hours per night. (b) Yes.
- 6.9** (a)  $H_0: \mu = 1900$ ,  $H_A: \mu \neq 1900$ . Independence assumption is met since the sample is random and less than 10% of the population. We are told to assume normality.  $T = -1.66$ ,  $df = 29$ ,  $0.10 < \text{p-value} < 0.20$ . Since the p-value  $> 0.05$ , we fail to reject  $H_0$ . The data do not provide strong evidence of a change in the average calorie intake of diners at this restaurant. (b) We are 95% confident that diners at this restaurant consume an average of 1690 calories to 1922 calories per meal. (c) Yes.
- 6.11**  $\bar{x} = 56.91$ .
- 6.13** No, distributions are extremely skewed.
- 6.15** (a) p-value  $< 0.005$ , we reject  $H_0$ . (b) p-value is about 0.01, we reject  $H_0$ . (c)  $0.025 < \text{p-value} < 0.05$ , we reject  $H_0$ . (d) p-value  $> 0.20$ , we fail to reject  $H_0$ .
- 6.17** (a) We are 95% confident that those in the group that got the weight loss pill lost 0.92 lbs less to 4.92 lbs more than those in the placebo group. (c) No. (d) No.
- 6.19** (a) Chicken that were fed linseed on average weigh 218.75 grams while those that were given horsebean weigh on average 160.20 grams. Both distributions are relatively symmetric with no apparent outliers. There is more variability in the weights of chicken that were given linseed. (b)  $H_0: \mu_L = \mu_H$ ,  $H_A: \mu_L \neq \mu_H$ . Independence is satisfied since both samples are random and less than 10% of their prospective populations. The distributions do not appear to be extremely skewed and the samples are independent of each other.  $T = 3.02$ ,  $df = 10$ ,  $0.01 < \text{p-value} < 0.02$ . Reject  $H_0$ , the data provide strong evidence of a difference between the average weights of chicken that were fed linseed and horsebean. (c) Type I, since we may have incorrectly rejected  $H_0$ . (d) Yes.
- 6.21** (a)  $H_0: \mu_A = \mu_M$ ,  $H_A: \mu_A \neq \mu_M$ .  $T = 5.46$ ,  $df = 25$ , p-value  $< 0.01$ . Reject  $H_0$ , the data provide strong evidence that there is a difference in the average city mileage between cars with automatic and manual transmissions.
- 6.23** We are 95% confident that on the highway cars with manual transmission get on average 5.53 to 10.33 MPG more than cars with automatic transmission.
- 6.25**  $H_0: \mu_T = \mu_C$ ,  $H_A: \mu_T \neq \mu_C$ .  $T = 2.69$ ,  $df = 21$ ,  $0.01 < \text{p-value} < 0.02$ . Since the p-value  $< 0.05$ , we reject  $H_0$ . The data provide strong evidence that the amount of biscuits consumed by the patients in the treatment and control groups are different.
- 6.27** (a)  $H_0: p = 0.69$ ,  $H_A: p \neq 0.69$ . (b)  $\hat{p} = 0.57$ . (c) The success-failure condition is not satisfied. (d) Each student can be represented with a card. Take 100 cards, 69 black cards representing those who follow the news about Egypt and 31 red cards representing those who do not. Shuffle the cards and draw with replacement (shuffling each time in between draws) 30 cards representing the 30 high school students. Calculate the proportion of black cards in this sample,  $\hat{p}_{sim}$ , i.e. the proportion of those who follow the news. Repeat 10,000 times and plot the resulting sample proportions. The p-value will be two times the proportion of simulations where  $\hat{p}_{sim} \leq 0.57$ . (Note: answers may vary, and in practice we would use a computer to simulate.) (e) p-value  $\approx 0.27$  (Note: answers may vary a little.) Fail to reject  $H_0$ . The data do not provide strong evidence that the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.
- 6.29** (a)  $H_0: p_P = p_C$ ,  $H_A: p_P \neq p_C$ . (b) -0.35. (c) Doubling the one tail, the p-value is about 0.03. Reject  $H_0$ . The data provide strong evidence that people react differently under the two scenarios.
- 

## 7 Introduction to linear regression

- 7.1** (a) The relationship is linear therefore the residuals plot will show randomly distributed residuals around 0 with constant variance. (b) The scatterplot shows a fan

shape, with higher variability in  $y$  for lower  $x$ . Therefore the residuals plot will also show a fan shape, wider around lower  $x$ , narrower around higher  $x$ . There may also be characteristics indicating nonlinearity for points on the left.

**7.3** (2) and (5) show a strong correlation. Even though (1) and (4) show a strong association, the relationship is not linear therefore correlation would not be strong. (3) and (6) show very weak or no relationship. Answers may vary slightly, e.g. one persons *moderate* may be equivalent to another persons *strong*.

**7.5** (a) Exam 2, since the points cluster closer to the line in the second scatterplot. (b) Exam 2 and the final are relatively close to each other chronologically, or Exam 2 may be cumulative so has greater similarities in material to the final exam.

**7.7** (a) 4. (b) 3. (c) 1. (d) 2.

**7.9** (a) The relationship is positive, weak, and possibly linear. There appears to be one outlier, a student who is about 63 inches tall whose fastest speed is 0 mph. This is probably a student who doesn't drive. (b) There is no obvious explanation why simply being tall should lead a person to drive faster. However, one possible outside factor may be gender. Males tend to be taller than females on average, and and personal experiences (anecdotal) may suggest they drive faster (confirmed in sociological studies). (c) It appears that males are taller on average than females and they also drive faster. The gender variable is a lurking variable for the positive association we observe between fastest driving speed and height.

**7.11** (a) There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. (b) Changing the units will not change the form, direction or strength of the relationship between the two variables. (c) Since changing units doesn't affect correlation,  $R = 0.636$ .

**7.13** (a) There is a moderately strong, positive, linear relationship between shoulder girth and height. (b) Changing the units,

even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.

**7.15** (a)  $R = 1$ . (b)  $R = 1$ . (c)  $R = 1$ .

**7.17** (a) There is a positive, very strong, linear association between number of tourists and spending. (b) Explanatory: number of tourists (in thousands), response: spending (in million \$). (c) We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism.

**7.19** Even though the relationship appears linear in the scatterplot, the residuals plot actually shows a non-linear relationship, therefore we should not fit a least squares line to these data.

**7.21** (a)  $\widehat{\text{travel time}} = 51 + 0.726 * \text{distance}$ . (b)  $b_1$ : For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time.  $b_0$ : When the distance traveled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles. Here, the  $y$ -intercept serves only to adjust the height of the line and is meaningless by itself. (c) 126 minutes. (d) 42 minutes, underestimate. (e) No, extrapolation.

**7.23** Approximately 40% of the variability in travel time is accounted for by the model, i.e. explained by distance traveled.

**7.25** No, there is an outlier that appears to have substantial pull on the line. We'll see more on this topic in the next section. The residuals does not show a random scatter around 0, which further suggests that a linear model may not be appropriate.

**7.27** (a) Influential. (b) Leverage. (c) Neither influential nor leverage.

**7.29** Neither influential nor high leverage.

**7.31** (a) The relationship appears to be strong, positive and linear. There is one potential outlier, the student who had 9 cans of beer. (b)  $\widehat{BAC} = -0.0127 + 0.0180 * \text{beers}$ .  $b_1$ : For each additional can of beer con-

sumed, the model predicts an additional 0.0180 grams per deciliter BAC.  $b_0$ : Students who don't have any beer are expected to have a blood alcohol content of -0.0127. It is not possible to have a negative blood alcohol content. Here, the  $y$ -intercept serves only to adjust the height of the line and is meaningless by itself. (c)  $H_0: \beta_1 = 0$ ,  $H_A: \beta_1 > 0$ . p-value  $\approx 0$ . Reject  $H_0$ . Number of cans of beer consumed and blood alcohol content are positively correlated and the true slope parameter is indeed greater than 0. (d) Approximately 79% of the variability in blood alcohol content can be explained by number of cans of beer consumed.

**7.33** (a)  $H_0: \beta_1 = 0$ ,  $H_A: \beta_1 \neq 0$ .  $T = 35.25$ ,  $df = 168$ , p-value  $\approx 0$ . Reject  $H_0$ . Wives' and husbands' ages are correlated and the true slope parameter is indeed greater than 0. (b)  $\widehat{ageWife} = 1.5740 + 0.9112 * ageHusband$ . (c)  $b_1$ : For each additional year in husband's age, the model predicts an additional 0.9112 years in wife's age.  $b_0$ : Men who are 0 years old are expected to have wives who are on average 1.5740 years old. The intercept here is meaningless and serves only to adjust the height of the line.

**7.35** (a)  $R = 0.94$ . The slope is positive, so  $R$  must also be positive. (b) 51.69, since  $R^2$  is high, the prediction based on this regression model is reliable. (c) No, extrapolation.

**7.37** (a)  $R = -0.53$ . The slope is negative, so  $R$  must also be negative. (b)  $H_0: \beta_1 = 0$ ,  $H_A: \beta_1 \neq 0$ .  $T = 4.32$ ,  $df = n - 2 = 49$ , p-value  $\approx 0.0001$ . Reject  $H_0$ . Percent homeownership and percent of the population living in an urban setting are correlated and the true slope parameter is indeed greater than 0. (c) The calculations and plotted line are not shown. The regression line does not adequately fit these data. (d) There is a fan shaped pattern apparent in this plot, which indicates non-constant variability in the residuals (little variability when  $x$  is small, more variability when  $x$  is large). Since the residuals have changing variability as we move across the plot, we should seek more appropriate statistical methods if we want to obtain a reliable estimate of the

best fitting straight line.

---



---

## 8 Multiple regression and ANOVA

**8.1** (a)  $\widehat{weight} = 248.64 + 74.94 * casein$ . (b) The estimated mean weight of chicks who are on casein feed is 74.94 grams higher than those who are given other feeds. Casein: 323.58 grams, No casein: 248.64 grams. (c)  $H_0$ : The true coefficient for **casein** is zero ( $\beta_1 = 0$ ).  $H_A$ : The true coefficient for **casein** is not zero ( $\beta_1 \neq 0$ ).  $T = 3.23$ , and the p-value is approximately 0.0019. With such a low p-value, we reject  $H_0$ . The data provide strong evidence that the true slope parameter is different than 0, and hence there appears to be a statistically significant relationship between feed type (casein or other) and the average weight of chicks.

**8.3** (a)  $\widehat{bwt} = 123.05 - 8.94 * smoke$ . (b) The estimated body weight of babies born to smoking mothers is 8.94 ounces lower than those who are born to non-smoking mothers. Smoker: 114.11 ounces, Non-smoker: 123.05 ounces. (c)  $H_0$ : The true coefficient for **smoke** is zero ( $\beta_1 = 0$ ).  $H_A$ : The true coefficient for **smoke** is not zero ( $\beta_1 \neq 0$ ).  $T = -8.65$ , and the p-value is approximately 0. Since p-value is very small we reject  $H_0$ . The data provide strong evidence that the true slope parameter is different than 0. There is strong evidence that the linear relationship between birth weight and smoking is real.

**8.5** (a)  $\widehat{bwt} = -80.41 + 0.44 * gestation - 3.33 * parity - 0.01 * age + 1.15 * height + 0.05 * weight - 8.40 * smoke$ . (b)  $\beta_1$ : The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day in length of pregnancy, all else held constant.  $\beta_3$ : The model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant. (c) Parity might be correlated with one of the other variables in the model, which introduces collinearity and complicates model estimation. (d) -0.58.

**8.7** (a)  $R^2 = 1 - (249.28/332.57) = 0.2504$ .  $R_{adj}^2 = 1 - (249.28/(1236 - 6 - 1))/(332.57/(1236 - 1)) = 0.2468$ .

**8.9** (a) There does not appear to be a significant relationship between the age of the mother and the birth weight of the baby since the p-value for the `age` variable is relatively high. We might consider removing this variable from the model. (b) No, all variables in the model now appear to have a significant relationship with the outcome therefore we would not need to removed any more variables.

**8.11** Based on the p-value alone, either gestation or smoke should be added to the model first. However, since the adjusted  $R^2$  for the model with gestation is higher, it would be preferable to add gestation in the first step of the forward-selection algorithm. (Other explanations are possible. For instance, it would be reasonable to only use the adjusted  $R^2$ .)

**8.13** (1) Normality of residuals: The normal probability plot shows a nearly straight line of points, providing evidence that the nearly normal assumption is reasonable. (2) Constant variance of residuals: The scatterplot of the absolute values of residuals versus the fitted values suggests that there may be a few outliers, some with lower than average fitted values and some with higher than average fitted values. (3) The residuals should be independent: The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that this as-

sumption is met. (4) Each variable should be linearly related to the outcome (i.e. we don't see any nonlinear trends): No nonlinear trends are evident. However, there are some outliers at the extremes of length of gestation and weight of the mother, so we should carefully examine these particular cases. There is some concern regarding constant variance across the parity groups.

We have two main concerns: outliers and constant variance. None of the outliers are exceptionally extreme, and there are a very large number of observations, so the influence of the outliers is probably mitigated (though we may want to study them more carefully, if possible). Additionally, while the constant variance assumption is violated across the parity groups, this violation is not very extreme. It is probably still reasonable to report the results while noting this model violation.

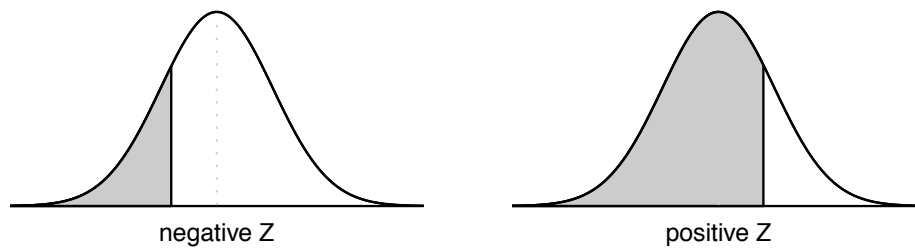
**8.15** Based on the side-by-side boxplots shown in Exercise 6.19, the constant variance assumption appears to be reasonable. Because the chicks were randomly assigned to their groups (and presumably kept separate from one another), independence of observations is also reasonable.  $H_0: \mu_1 = \mu_2 = \dots = \mu_6$ .  $H_A$ : The average weight ( $\mu_i$ ) varies across some (or all) groups.  $F_{5,65} = 15.36$  and the p-value is approximately 0. With such a small p-value, we reject  $H_0$ . The data provide strong evidence that the average weight of chicks varies across some (or all) groups.

# Appendix C

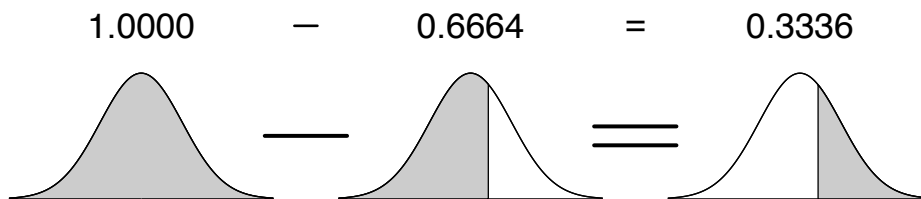
## Distribution tables

### C.1 Normal Probability Table

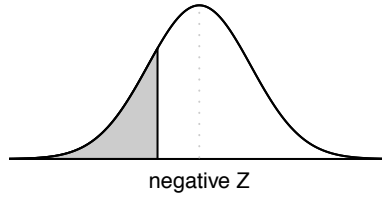
The area to the left of  $Z$  represents the percentile of the observation. The normal probability table always lists percentiles.



To find the area to the right, calculate 1 minus the area to the left.

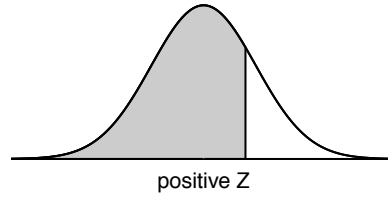


For additional details about working with the normal distribution and the normal probability table, see Section 3.1.



Second decimal place of $Z$										$Z$
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	-3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	-3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	-3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	-3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	-3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	-2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	-2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	-2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	-2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	-2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	-2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	-2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	-2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	-2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	-2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	-1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	-1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	-1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	-1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	-1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	-1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	-1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	-1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	-1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	-0.9
0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119	-0.8
0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420	-0.7
0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743	-0.6
0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085	-0.5
0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446	-0.4
0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821	-0.3
0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207	-0.2
0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602	-0.1
0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000	-0.0

\*For  $Z \leq -3.50$ , the probability is less than or equal to 0.0002.



Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

\*For  $Z \geq 3.50$ , the probability is greater than or equal to 0.9998.

## C.2 t Distribution Table

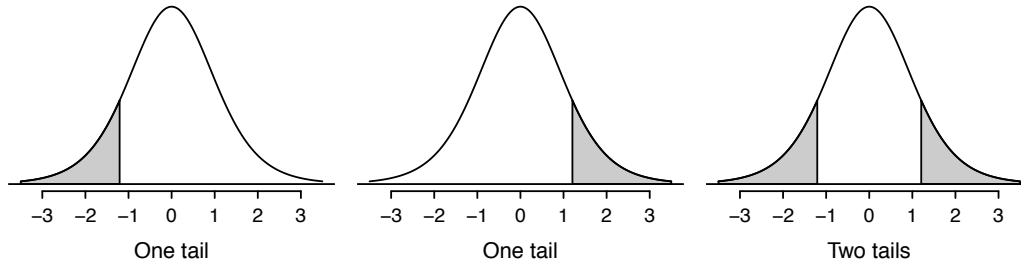


Figure C.1: Three  $t$  distributions.

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75

		0.100	0.050	0.025	0.010	0.005
one tail						
two tails		0.200	0.100	0.050	0.020	0.010
df						
31		1.31	1.70	2.04	2.45	2.74
32		1.31	1.69	2.04	2.45	2.74
33		1.31	1.69	2.03	2.44	2.73
34		1.31	1.69	2.03	2.44	2.73
35		1.31	1.69	2.03	2.44	2.72
36		1.31	1.69	2.03	2.43	2.72
37		1.30	1.69	2.03	2.43	2.72
38		1.30	1.69	2.02	2.43	2.71
39		1.30	1.68	2.02	2.43	2.71
40		1.30	1.68	2.02	2.42	2.70
41		1.30	1.68	2.02	2.42	2.70
42		1.30	1.68	2.02	2.42	2.70
43		1.30	1.68	2.02	2.42	2.70
44		1.30	1.68	2.02	2.41	2.69
45		1.30	1.68	2.01	2.41	2.69
46		1.30	1.68	2.01	2.41	2.69
47		1.30	1.68	2.01	2.41	2.68
48		1.30	1.68	2.01	2.41	2.68
49		1.30	1.68	2.01	2.40	2.68
50		1.30	1.68	2.01	2.40	2.68
60		1.30	1.67	2.00	2.39	2.66
70		1.29	1.67	1.99	2.38	2.65
80		1.29	1.66	1.99	2.37	2.64
90		1.29	1.66	1.99	2.37	2.63
100		1.29	1.66	1.98	2.36	2.63
150		1.29	1.66	1.98	2.35	2.61
200		1.29	1.65	1.97	2.35	2.60
300		1.28	1.65	1.97	2.34	2.59
400		1.28	1.65	1.97	2.34	2.59
500		1.28	1.65	1.96	2.33	2.59
$\infty$		1.28	1.65	1.96	2.33	2.58

### C.3 Chi-Square Probability Table

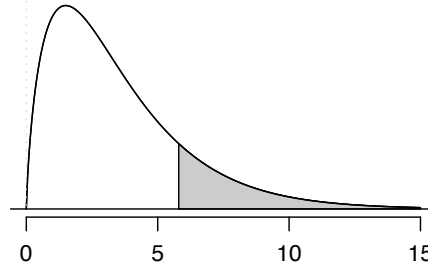


Figure C.2: Areas in the chi-square table always refer to the right tail.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	12	14.01	15.81	18.55	21.03	24.05	26.22	28.30	32.91
	13	15.12	16.98	19.81	22.36	25.47	27.69	29.82	34.53
	14	16.22	18.15	21.06	23.68	26.87	29.14	31.32	36.12
	15	17.32	19.31	22.31	25.00	28.26	30.58	32.80	37.70
	16	18.42	20.47	23.54	26.30	29.63	32.00	34.27	39.25
	17	19.51	21.61	24.77	27.59	31.00	33.41	35.72	40.79
	18	20.60	22.76	25.99	28.87	32.35	34.81	37.16	42.31
	19	21.69	23.90	27.20	30.14	33.69	36.19	38.58	43.82
	20	22.77	25.04	28.41	31.41	35.02	37.57	40.00	45.31
	25	28.17	30.68	34.38	37.65	41.57	44.31	46.93	52.62
	30	33.53	36.25	40.26	43.77	47.96	50.89	53.67	59.70
	40	44.16	47.27	51.81	55.76	60.44	63.69	66.77	73.40
	50	54.72	58.16	63.17	67.50	72.61	76.15	79.49	86.66

# Index

- F** test, **326**
- ., 151
  
- Addition Rule, **57**
- adjusted  $R^2$ , **313**
- alternative hypothesis ( $H_A$ ), **156**
- analysis of variance (ANOVA), **294, 321**
- anecdotal evidence, **27**
- associated, **8**
  
- backward-elimination, **315**
- bar plot, **20**
- Bayesian statistics, **81**
- Bernoulli random variable, **118**
- bias, **28**
- bimodal, **12**
- binomial distribution, **122**
- blind, **36**
- blocking, **34**
- blocks, **34**
- Bonferroni correction, **331**
  
- case, **3**
- categorical, **5**
- chi-square distribution, **214**
- chi-square probability table, **215**
- cluster sample, **32**
- clusters, **32**
- cohort, **30**
- collections, **58**
- collinear, **312**
- column proportion, **20**
- column totals, **20**
- complement, **62**
- condition, **72**
- conditional probability, **71**
- confidence interval, **149**
- confident, **149**
- confounder, **31**
- confounding factor, **31**
- confounding variable, **31**
  
- contingency table, **20**
- continuous, **5**
- control, **34**
- control group, **2, 36**
- convenience sample, **29**
- correlation, **280**
  
- d, 171
- data, **1**
- data density, **11**
- data fishing, **324**
- data matrix, **4**
- data snooping, **324**
- deck of cards, **59**
- degrees of freedom, **313**
- degrees of freedom (**df**), **242**
- degrees of freedom (df), **214**
- density, **68**
- dependent, **8, 29**
- deviation, **13**
- df, *see* degrees of freedom (df)
- discrete, **5**
- disjoint, **57**
- distribution, **10, 68**
- double-blind, **36**
  
- error, **146**
- events, **58**
- expected value, **84**
- experiment, **30**
- experiments, **34**
- explanatory, **29**
- exponentially, **119**
- extrapolation, **287**
  
- face card, **59**
- factorial, **122**
- failure, **118**
- first quartile, **15**
- forward-selection, **316**
- frequency table, **20**

- full model, **314**
- General Addition Rule, **60**
- generalized linear models, **132**
- high leverage, **290**
- histogram, **11**
- hollow histograms, **25**
- hypotheses, **156**
- independent, **8, 29, 64**
- independent and identically distributed (iid), **119**
- influential point, **290**
- interquartile range, **15**
- interquartile range (IQR), **15**
- joint probability, **70**
- Law of Large Numbers, **56**
- least squares criterion, **283**
- least squares line, **283**
- left skewed, **12**
- levels, **5**
- linear combination, **88**
- long tail, **12**
- lurking variable, **31**
- margin of error, **173**
- marginal probabilities, **70**
- mean, **10**
- mean square between groups (*MSG*), **325**
- mean square error (*MSE*), **326**
- median, **15, 16**
- midterm election, **290**
- mode, **12**
- mosaic plot, **22**
- multimodal, **12**
- multiple comparisons, **331**
- multiple regression, **311**
- mutually exclusive, **57**
- $n$  choose  $k$ , **122**
- negative binomial distribution, **128**
- negatively associated, **8**
- nominal, **5**
- non-response, **29**
- non-response bias, **29**
- normal curve, **104**
- normal distribution, **104**
- normal probability plot, **114**
- normal probability table, **107**
- null hypothesis ( $H_0$ ), **156**
- null value, **157**
- numerical, **4**
- observational study, **30**
- observational unit, **3**
- one-sided, **162**
- one-way ANOVA, **332**
- ordinal, **5**
- outcome, **2, 56**
- outlier, **17**
- outliers, **17**
- p-value, **161**
- paired, **192**
- parameters, **105**
- patients, **36**
- percentile, **15, 107**
- permutation test, **261**
- pie chart, **24**
- placebo, **2, 36**
- placebo effect, **2, 36**
- point estimate, **144**
- point-slope, **285**
- Poisson distribution, **131**
- pooled estimate, **210**
- pooled standard deviation, **254**
- population, **9, 26**
- population mean, **144**
- population parameters, **145**
- positive association, **8**
- power, **178**
- practically significant, **179**
- predictor, **274**
- primary, **76**
- probability, **56**
- probability density function, **68**
- probability distribution, **61**
- probability of a success, **118**
- probability sample, *see* sample
- Product Rule for independent processes, **65**
- prosecutor's fallacy, **325**
- prospective study, **31**
- quantile-quantile plot, **114**
- R-squared, **288**
- random process, **56**
- random variable, **83**

- randomization technique, **38**
- randomized experiment, **30, 34**
- rate, **132**
- relative frequency table, **20**
- replicate, **34**
- representative, **29**
- residual plot, **279**
- residuals, **277**
- response, **29**
- retrospective studies, **31**
- right skewed, **12**
- robust estimates, **18**
- row proportions, **20**
- row totals, **20**
- running mean, **145**
  
- sample, **9, 26**
- sample mean, **144**
- sample proportion, **118**
- sample space, **62**
- sampling distribution, **146**
- sampling variation, **144**
- scatterplot, **6, 9**
- secondary, **76**
- segmented bar plot, **22**
- sets, **58**
- side-by-side box plot, **25**
- significance level, **160**
- simple random sample, **28**
- simulation, **38**
- skewed to the high end, **12**
- skewed to the positive end, **12**
- skewed to the right, **12**
- standard deviation, **13, 86**
- standard error (SE), **146**
- statistically significant, **179**
- stepwise, **315**
- strata, **32**
- stratified sampling, **32**
- study participants, **36**
- success, **118**
- success-failure condition, **202**
- suits, **59**
- sum of squared errors (*SSE*), **326**
- sum of squares between groups, **326**
- sum of squares total (*SST*), **326**
- summary statistic, **3**
- symmetric, **12**
  
- t table, **243**
  
- table proportions, **71**
- tail, **12**
- test statistic, **175**
- the outcome of interest, **72**
- third quartile, **15**
- time series, **318**
- transformation, **19**
- treatment group, **2, 36**
- tree diagram, **76**
- trial, **118**
- two-sided, **162**
- Type 1 Error, **160**
- Type 2 Error, **160**
  
- unbiased, **172**
- unimodal, **12**
- unit of observation, **3**
  
- variables, **3**
- variance, **13, 86**
- Venn diagrams, **59**
- volunteers, **36**
  
- whiskers, **16**
- with replacement, **82**
- without replacement, **82**
  
- Z score, **106**