# S520 and Similar Courses Past Exams Questions and Answers[1]

## Questions

1. (F2018) In the National Basketball Association (NBA), the champions of the Eastern Conference (E) play against the champions of the Western Conference (W) in a series of games called the NBA finals. The first team to win four games wins the series and are crowned champions for the season. Thus, the series must last at least four games and can last no more than seven games. Let's define an outcome by identifying which Conference's pennant winner won each game in order from first to last. For example, last May the Golden State Warriors (West) beat the the Cleveland Cavaliers (East) in four games so the outcome was WWWW, a year ago they did it in five games: WWWEW, and the previous year the Cavaliers beat the Warriors in seven games: WWEWEEE.

   (a) (1 point) Write down the outcomes in event $A$, where

   $$A = \{\text{at most five games are played in the series}\}$$

   (b) (3 points) How many outcomes are possible such that the series last exactly 6 games?

   (c) (3 points) Assume that the Western champions have 0.7 probability of winning any given game, what is the probability that the series last exactly 6 games?

   (d) (3 points) Assume that the Western champions have 0.7 probability of winning any given game, what is the probability that they win the series in a 7th and decisive game?

2. (F2016.) As part of an early study of early detection of breast cancer in the 1960s, 31,000 women in New York were offered screening mammograms. Of these women, 20,200 accepted the screening, while 10,800 refused. Of those that chose to have screening, 23 died from breast cancer in the five years following screening. Of those that refused screening, 16 died of breast cancer in the five years following screening.

   Does the above data prove that screening reduces women's deaths from breast cancer? Explain why or why not.

3. (F2017) Of 40,000 students who applied to Indiana University (IU), 33,800 were born in the United States (US), while the rest were from other countries. After the review process, a U.S. applicant has a 70% chance of getting admitted, but a non-U.S. applicant has only 53% chance. Let A be the event that an applicant is from the U.S., and let B be the event that an applicant is admitted.

   (a) (3 points) An applicant is selected at random. Given that she is from a country other than the US, what is the probability that she was admitted to IU?

   (b) (4 points) What is the probability that a randomly chosen applicant will be admitted?

   (c) (3 points) An applicant is selected at random. Given that she was admitted to IU, what is the probability that she is from the US?

[1] The questions presented here have been developed by faculty members of the Department of Statistics at Indiana University. You are only authorize to use them for your personal use. DO NOT reproduce and/or distribute this material.

4. (F2017) Six students take seven dance lessons during the semester. There are three female students: Dora(D), Eliza(E), and Fiona(F); and three male students: Alvin(A), Boris(B), and Charlie(C). For each lesson, couples (one male and one female) are randomly selected before the lesson starts. Eliza is always happy when she is matched with Alvin, she is also happy when matched with Charlie as long as Alvin-Fiona are not matched together, and she is never happy when her partner is Boris. The other five students are indifferent about the chosen partner.

   (a) (2 points.) At any given dance lesson, what is the probability that Eliza is happy?

   (b) (6 points.) Let $H$ : {Eliza is happy in exactly 4 lessons}, and
   $J$ : {Eliza is happy in most dance lessons}
   Find $P(H)$, $P(J)$, and $P(H^c \cup J)$.

   (c) (2 points.) Now assume that the female student chooses her male partner, but the order of choice is random (i.e., Eliza is equally likely to choose first, second, or third). Is the probability of Eliza being happy lower, equal, or higher than before? Explain why (you do not need to calculate this probability).

5. (Su2017) I toss a fair coin six times

   (a) (2 points.) What is the probability exactly four of the tosses show heads?

   (b) (4 points.) What is the probability that there are more heads than tails?

   (c) (4 points.) Given that there are more tails than heads, what is the conditional probability exactly four of the tosses show tails?

6. (S2017) I have four fair dice, each with six faces numbered 1, 2, 3, 4, 5, 6. I toss all four dice independently.

   (a) (2 points) What is the probability that all four dice show 1?

   (b) (4 points) What is the probability that at least two of the dice show 1?

   (c) (4 points) What is the probability that the sum of the four dice is 6 or less?

7. (F2016) There are five Power Rangers. Of the five, three are male (Red, Black, Blue) and two are female (Pink, Yellow.)

   I randomly select two of the Power Rangers, without replacement.

   (a) What is the probability that the Pink Ranger is one of the two selected Power Rangers?

   (b) What is the probability at least one of the two selected Power Rangers is female?

   (c) Given that at least one of the two selected Power Rangers is female, what is the conditional probability the Pink Ranger is selected?

   (d) Are the events "the Pink Ranger is one of the two selected" and "at least one of the two selected Power Rangers is female" independent? Explain why or why not.

8. (Su2016) According to the Breast Cancer Surveillance Consortium (`breastscreening.cancer.gov`), out of the population "women aged 50–54 who have screening mammograms,"

- 0.428% have breast cancer;
- Of those with breast cancer, 82.6% correctly test positive on the mammogram;
- Of those without breast cancer, 90.4% correctly test negative on the mammogram.

(a) What is the probability that a randomly selected woman from the population both has breast cancer and test positive?

(b) What is the probability that a randomly selected woman from the population tests positive?

(c) Given that a randomly selected woman tests positive, what is the probability she has breast cancer?

(d) Suppose that out of a large sample from the population, ten women test positive. What is the probability that at least one of these ten women has breast cancer?

9. (S2016) According to a 2015 Pew survey, 84% of U.S. adults use the Internet. The rate of Internet use ranges from 96% of young adults (18 to 29) to 58% of seniors (65 or older.) For the rest of this question, assume Pew's figures are correct.

   (a) I select three U.S. adults at random. What is the probability that at least one of them does NOT use the Internet?

   (b) I select at random a young adult and a senior. What is the probability that the senior uses the Internet and the young adult does not?

   (c) (I select at random a young adult and a senior. Given that exactly one of the two uses the Internet, what is the *conditional* probability that it is the senior?

10. (F2015)

   (a) I toss six fair coins. What is the probability exactly four of the coins show heads?

   (b) I toss six fair coins. What is the probability that there are more heads than tails?

   (c) I toss six fair coins. Given that there are more heads than tails, what is the conditional probability exactly four of the coins show heads?

11. (Su2015) I have four fair dice, each with six faces numbered 1, 2, 3, 4, 5, 6. I toss all four dice independently.

   (a) What is the probability that all four dice show 1?

   (b) What is the probability that at least two of the dice show 1?

   (c) What is the probability that the sum of the four dice is 6 or less?

12. (Su2015.) About 1% of men in their fifties have or will develop prostate cancer (within a certain timeframe.) Under one method for screening for prostate cancer:

   - Out of men in their fifties that have or will develop prostate cancer, 96% test positive (the rest test negative.)
   - Out of men in their fifties that do not have and will not develop prostate cancer, 94% test negative (the rest test positive.)

I randomly select a man in his fifties who has both been screened using this method and tested positive. Based on the information above, what is the probability he has or will develop prostate cancer?

13. (S2015.) In the last minute of Super Bowl XLIX, the Seattle Seahawks, who are behind by four points, must decide whether to run or pass. To simplify, suppose that in each case, there are three mutually exclusive outcomes: a touchdown, no gain, and turnover.

    If the Seahawks run, the probabilities of each outcome are:

    - Touchdown: 60%
    - No gain: 39%
    - Turnover: 1%

    If the Seahawks pass, the probabilities are:

    - Touchdown: 65%
    - No gain: 31%
    - Turnover: 4%

    No matter whether they decide to run or pass, the probability of winning depends only on whether there is a touchdown, no gain, or a turnover. The probabilities are:

    - After a touchdown, the Seahawks have a 95% chance of winning.
    - After no gain, the Seahawks have an 80% chance of winning.
    - After a turnover, the Seahawks have no chance of winning.

    (a) Suppose the Seahawks decide to run. What is the probability they win?

    (b) Suppose the Seahawks decide to pass. What is the probability they win?

    (c) Based on your answers to parts (a) and (b), should the Seahawks run or pass?

**Ch 4**

14. (F2018) Theresa and John are getting married. For the wedding party, they reserve a ballroom with a maximum capacity of 200 people, and this number should include the bride, the groom, and their four parents, who are certain to attend. Assume that friends invited to the party will accept the invitation with probability 0.7. In addition, if a friend of Theresa's accepts the invitation she/he will actually attend with probability 0.9, and if a friend of John's accepts he/she will attend with probability 0.8.

    (a) (2 points) Assume that only Theresa's friends are invited, and she invites 300 friends. Let $X$ and $Y$ be random variables that represent the number of friends accepting the invitation and attending the wedding party, respectively. Write down the distribution of $X$ and $Y$ using the notation learned in class.

    (b) (2 points) What is the expected number of friends who will accept the invitation? What is the expected number of friends who will attend?

    (c) (3 points) If only Theresa's friends are invited, and she invites 300 friends, what is the probability that she will end up with more guests than she can accommodate?

    (d) (3 points) If only John's friends are invited, and he invites 320 friends, what is the probability that the ballroom won't be at full capacity during the wedding party?

15. (Su2018) Let $X$ be a discrete random variable with probability mass function

$$P(X = x) = \begin{cases} k/12 & x = 1 \\ k/6 & x = 2 \\ k/4 & x = 3 \\ k/3 & x = 5k \\ 0 & \text{otherwise.} \end{cases}$$

where $k$ is a constant. Let $X_1, \ldots, X_{100}$ be an iid sequence of random variables with the same distribution as $X$. Let $\bar{X}$ be the sample mean of $X_1, \ldots, X_{100}$.

    (a) (2 points.) What is the value of $k$?

    Note: If you are unsure how to obtain $k$, for the next parts let $X$ be the number of pips you observe when rolling a fair die. (20% penalty)

    (b) (2 points.) Find $E(X)$ and $Var(X)$

    (c) (3 points.) Find $E(\bar{X})$ and $Var(\bar{X})$

    (d) (3 points.) Find the approximated probability that $\bar{X}$ is greater than 4.

16. (S2018) Based on the United States Preventive Services Task Force, 10% of men ages 55 to 69 years have prostate cancer. Elevated levels of prostate-specific antigen (PSA) in the man's blood are used to test of prostate cancer. Unfortunately, PSA level analysis is not a definitive test for prostate cancer; for example, a man with no prostate cancer has a 12% chance of testing positive while a man with prostate cancer has a 15% chance of testing negative. Let A be the event that a man has prostate cancer and let B be the event that a man test positive.

(a) (3 points) Build the probability tree using events A and B. State explicitly the location of unconditional and conditional probabilities and their corresponding values.

(b) (3 points) A man is selected at random. Given that he does have prostate cancer, what is the probability of a false-negative when using PSA? (i.e. the probability that he has prostate cancer and test negative).

(c) (4 points) What is the probability that a randomly chosen man test positive?

(d) (Optional, it replaces 3 lost points elsewhere) A man is selected at random. Given that he tests positive, what is the probability that he has prostate cancer?

17. (F2017) Let $X$ be a discrete random variable WITH probability mass function

$$f(x) = P(X = x) = \begin{cases} 3mx & x = 1 \\ 2mx & x = 2 \\ mx & x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

(a) (2 points) Find $m$.

(b) (4 points)Write down an graph $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

(c) (4 points) Find the expected value and variance of $X$.

18. (Su2017) Let $X$ be a discrete random variable with probability mass function

$$f(x) = P(X = x) = \begin{cases} kx & x \in \{3, 4, 6, 7\} \\ 0 & \text{otherwise.} \end{cases}$$

where $k$ is a constant.

(a) (4 points) Find $k$.

(b) (4 points) Find the expected value of $k^2X + k$.

(c) (2 points) Show/explain that the probability of the underlying sample space is one, P(S) = 1.
(Hint: Start by showing that S can be decomposed in disjoint events and their sum of probabilities provides the desired result).

19. (S2017.) In the 1890s, the biologist and statistician Raphael Weldon rolled twelve six-sided dice 26,306 times, for a total of 315,672 rolls. (Video games had not been invented at the time.) If the dice were perfectly fair, the probability of getting either a five or a six on any one roll should be 1/3, so we would expect 105,224 rolls to be five or six. In fact, 106,602 of the rolls turned out to be five or six — that's a difference of 1,378.

(a) Suppose the dice were perfectly fair and each die roll was independent of every other die roll. What's the probability (to one significant figure) of getting 106,602 or more fives or sixes (combined) in 315,672 rolls?

(b) Suppose the dice were perfectly fair and each die roll was independent of every other die roll. Let $X$ be a random variable representing the number of rolls that are either five or six out of 315,672 rolls. What's the probability that $X$ differs from its expected value by at least 1,378? (That is, what's $P(X \leq 103846 \text{ or } X \geq 106602)$?)

(c) Were the dice perfectly fair?

20. (S2017) I toss a coin seven times, independently. The coin is fair, so the probability of getting a head is 0.5 for each toss. Let $X$ be a binomial random variable representing the number of heads in the seven tosses.

(a) What is the probability I get exactly five heads?

(b) Give R code to find $F(2)$, the cumulative distribution function of $X$ at 2. (If you can't remember the R code, you may calculate the numerical answer.)

(c) Find the expected value and standard deviation of $X$.

(d) Are the events "I get all heads on the first four tosses" and "I get all tails on the last four tosses" independent? Explain why or why not.

21. (S2017) Let $X$ be a discrete random variable with probability mass function

$$f(x) = P(X = x) = \begin{cases} 4k & x = 0 \\ 3k & x = 1 \\ 2k & x = 2 \\ k & x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find $k$.

Note: If you can't solve part (a), you may write your answers for the rest of the question in terms of $k$ for a small penalty.

(b) Complete $F(y)$, the cumulative distribution function (CDF) of $X$, by filling in the blanks below:

$$F(y) = \begin{cases} \qquad\qquad & y < 0 \\ & 0 \leq y < 1 \\ & 1 \leq y < 2 \\ & 2 \leq y < 3 \\ & y \geq 3 \end{cases}$$

(c) Find the expected value of $X$.

(d) Find the variance of $X$.

22. (F2016) 48% percent of U.S. adults have visited a library in the past year. I select two U.S. adults at random. Let $X$ be a discrete random variable representing the number of the adults I selected who have visited a library in the last year.

(a) Find $P(X = 0)$, $P(X = 1)$, and $P(X = 2)$.

(b) Write down an expression for $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

(c) Find the expected value and variance of $X$.

23. (F2016.) One semester I gave a 20-question test for psychic powers to a class of 80 students. Each student (independently) guessed "left" or "right," then (using the random number generator in R) a star appeared on either the left or the right side on the screen. This process was repeated 20 times. At the end, each student counted up the number of times they correctly guessed the side of the screen the star would appear on. The highest score was 15.

    (a) Suppose this test is taken by a student who does not have psychic powers. What is the probability she gets 15 or more correct?

    (b) Suppose this test is taken by eighty students who do **NOT** have psychic powers. What is the probability *at least one student* gets 15 or more correct?

24. (Su2016) Let $X$ be a discrete random variable representing the "grade points" earned by students in a certain class. Suppose that $X$ has probability mass function

$$f(x) = P(X = x) = \begin{cases} 0.15 & x = 4 \\ 0.3 & x = 3.7 \\ 0.35 & x = 3.3 \\ 0.2 & x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

    (a) Write down an expression for $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

    (b) Find the expected value of $X$.

    (c) Find the variance of $X$.

25. (Su2016.) Suppose that among likely voters in an upcoming election, 46% support Billary, 39% support Ronald, while the remaining 15% support Other. (You may assume that the population of likely voters is very large.)

    (a) I randomly select two likely voters (in no particular order.) What is the probability that both support Ronald?

    (b) I randomly select two likely voters (in no particular order.) What is the probability that one supports Billary and the other supports Ronald?

    (c) I randomly select 1000 likely voters (in no particular order.) What is the probability that at least 480 of the sample support Billary?

26. (S2016) Let $X$ be a discrete random variable with probability mass function

$$f(x) = P(X = x) = \begin{cases} 0.1 & x = 3 \\ 0.4 & x = 4 \\ 0.3 & x = 5 \\ 0.2 & x = 6 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Write down a mathematical expression for $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$. (If you can't find the mathematical expression, draw a graph of the piecewise function for partial credit.)

(b) Find the expected value of $X$.

(c) Find the variance of $X$.

27. (S2016.) Hillary and Donald make a bet about an upcoming election. If Hillary wins the election, Donald pays her \$1. If Donald wins the election, Hillary pays him \$3. If anyone else wins, they do not exchange any money.

A website gives the following probabilities for the election:

| Winner | Probability |
|---|---|
| Hillary | 54% |
| Donald | 16% |
| Somebody else | 30% |

If the model is right, then according to expected value, whom does the bet favor?

28. (F2015.) A controversial issue in evolutionary psychology (and statistics) is whether and to what extent characteristics of parents, such as physical beauty, affect the probability of giving birth to a girl. Gelman and Weakliem (2009) counted all children born to the members of People Magazine's 50 Most Beautiful People list from 1995–2000 (up to August 2007,) so, for example, Julia Roberts contributes two boys and a girl to the data set. They found that of the 329 children born to these celebrities, 157 were girls. In the general population, 48.5% of births are girls.

(a) Suppose the sexes of all children are independent, and each child has a 48.5% probability of being a girl. Use the binomial distribution to find the probability that at least 157 out of 329 children are girls.

(b) Explain why the sexes of the children in the study might not all be independent.

29. (F2015) Let $X$ be a discrete random variable with probability mass function

$$f(x) = P(X = x) = \begin{cases} kx & x \in \{1, 2, 3, 4\} \\ 0 & \text{otherwise.} \end{cases}$$

where $k$ is a constant.

(a) Find $k$.

(b) Find $F(y)$, the cumulative distribution function of $X$, for all $y \in (-\infty, \infty)$.

(c) Find the expected value and the variance of $X$.

30. (F2015.) I give a ten question true/false statistics test to a class of ten chimpanzees. The chimpanzees, who do not know any statistics, randomly guess true or false, independently for each question and independently of each other. A score of at least 8 out of 10 is required to pass the test. What is the probability that at least one of the chimpanzees passes the test?

31. (Su2015.) A roulette player repeatedly bets on '00'. The probability on the wheel landing on '00' is 1/38. If this happens, the player wins \$35. Otherwise, the player loses \$1.

    (a) The player bets on '00' for two hundred spins. Let $X$ be a random variable representing the number of times the player wins. What are the expected value and standard deviation of $X$?

    (b) For the player to make a profit, the wheel must land on '00' on at least six of the 200 spins. What is the probability the player makes a profit?

    (c) Suppose the player goes to the casino and bets on "00" 50,000 times. What is the probability the player makes a profit?

32. (S2015) Let $X$ be a discrete random variable with probability mass function

$$f(x) = P(X = x) = \begin{cases} 0.1 & x = -3 \\ 0.2 & x = -2 \\ 0.3 & x = 0 \\ 0.3 & x = 1 \\ 0.1 & x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Complete $F(y)$, the cumulative distribution function (CDF) of $X$, by filling in the blanks below:

$$F(y) = \begin{cases} 0 & y < -3 \\ & -3 \le y < -2 \\ & -2 \le y < 0 \\ & 0 \le y < 1 \\ & 1 \le y < 3 \\ 1 & y \ge 3 \end{cases}$$

(b) Find the expected value of $X$.

(c) Find the variance of $X$.

33. (S2014) Let $X$ be a discrete random variable with probability mass function

$$P(X=x)=\begin{cases} 0.1 & x=1 \\ 0.2 & x=2 \\ 0.2 & x=3 \\ 0.3 & x=4 \\ 0.2 & x=5 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find $F(y)$, the cumulative distribution function of $X$, for all $y \in (-\infty, \infty)$.

(b) Find the expected value and the variance of $X$.

(c) Let $X_1$ and $X_2$ be independent random variables with the same distribution as $X$. What is the probability that $X_1 + X_2 = 4$?

34. (S2013) Let $X$ be a discrete random variable with probability mass function

$$P(X=x)=\begin{cases} kx & x \in \{1,2,3,4,5\} \\ 0 & \text{otherwise.} \end{cases}$$

where $k$ is a constant.

(a) Find $k$.

(b) Find $F(y)$, the cumulative distribution function of $X$, for all $y \in (-\infty, \infty)$.

(c) Find the expected value and variance of $X$.

**Ch 5**

35. (F2017) The distribution of ACT scores nationwide is approximately normal. Let $W$ represent the ACT scores with $W \sim N(21.5, 29.2)$. Let's assume that a student is likely to succeed in college is her ACT score is among the top 20%.

    (a) (2 points.) What is the minimum score needed for a student to be likely to succeed in college?

    (b) (4 points.) If a random sample of 100 students is taken, what is the probability that at least 30 are likely to succeed in college?

    (c) (4 points.) The SAT scores are also approximately normally distributed. Let $Y$ represent the SAT scores and $Y = 240 + 58.5W$. Find the expected value and variance of $Y$ and the probability of obtaining an SAT score higher than 1900.

36. (S2017.) The tempo (speed) of a piece of music is usually measured in beats per minute (BPM.) A study[2] found that the BPM in disco songs was approximately normal with mean 120 and standard deviation 20. (Treat BPM as a continuous random variable for the purpose of this question.) The best tempo for dancing is considered to be 115 to 135 BPM, while anything above 160 BPM is exhausting.

    Suppose a DJ gets lazy and puts her MP3 player containing a large, representative collection of disco songs on shuffle.

    (a) Suppose the MP3 player randomly selects a disco song. What is the probability that the BPM is between 115 and 135?

    (b) Suppose the MP3 player randomly selects five disco songs. What is the probability that all five have a BPM between 115 and 135?

    (c) Suppose the MP3 player randomly selects ten disco songs. What is the probability at least one of the ten songs has a BPM over 160?

37. (F2016) Let $X$ be a continuous random variable with the probability density function (PDF)

$$f(x) = \begin{cases} 0.1 & 0 \leq x < 8 \\ 0.05 & 8 \leq x < 12 \\ 0 & \text{otherwise.} \end{cases}$$

    To answer the following questions, it may help to draw a graph of this PDF.

    (a) Find $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

    (b) Find the median of $X$.

    (c) Find the expected value of $X$.

[2] https://github.com/nikhilunni/BPMFinder/

38. (F2016) Let $X$ be a random variable with the cumulative distribution function (CDF)

$$F(y) = \begin{cases} 0 & y < 0 \\ 0.25y^2 & 0 \leq y < 1 \\ 0.25y & 1 \leq y < 4 \\ 1 & y \geq 4. \end{cases}$$

Calculus shows that $E(X) = 49/24$ and $E(X^2) = 43/8$. You may use these results below.

(a) Is $X$ discrete, continuous, or neither? Explain.

(b) Find the standard deviation of $X$.

(c) Let $A$ and $B$ be independent random variables with the same distribution as $X$. Let $Y = A - B$. What are the expected value and standard deviation of $Y$?

39. (F2016.) The *exponential distribution* is sometimes used to model times between events. For example, the times between the emissions of alpha particles from a smoke alarm follow an exponential distribution.

An exponential random variable with rate parameter $\lambda$ is a continuous random variable with CDF

$$F(y) = \begin{cases} 0 & y < 0 \\ 1 - e^{-\lambda y} & y \geq 0 \end{cases}.$$

Note that an exponential random variable can only take non-negative values.

The CDF can be found in R using the function `pexp()`. For example, `pexp(3, rate=1)` gives $F(3)$ for an exponential random variable with rate $\lambda = 1$.

Let $X$ be an exponential random variable with rate 1.

(a) Find $P(X > 3)$.

(b) Find $P(1 < X < 3)$.

(c) Let $Y = X^2$. Find $P(Y > 4)$.

(d) Let $X_1, X_2, X_3, X_4$, and $X_5$ all be independent exponential random variables, each with rate parameter $\lambda = 1$. Find the probability that no more than three out of $(X_1, X_2, X_3, X_4, X_5)$, are less than 3.

40. (Su2016) Let $X_1$ be a standard normal random variable. Let $X_2$ be a normal random variable with expected value $-5$ and variance 25. Suppose that $X_1$ and $X_2$ are independent. The following R output may be useful:

```
> pnorm(1)
[1] 0.8413447
```

(a) Find $P(X_1 > 1)$.

(b) Find $P(X_2 > 0)$.

(c) Find the probability that both $X_1$ and $X_2$ are greater than zero.

(d) Let $Y = X_1 + X_2$. Find the expected value, variance, and standard deviation of $Y$.

41. (Su2016.) Let $X$ be a standard normal random variable. Let $Y = X^3$.

    (a) Find $P(-1.5 < X < 2.5)$.

    (b) Find $P(-1 < Y < 1)$.

    (c) Find $P(Y > 2)$.

42. (S2016.) Let $X$ be a standard normal random variable. Let $Y = X^3$. Let $Z = |Y|$.

    (a) Find $P(1 < X < 2)$.

    (b) Find $P(1 < Y < 2)$.

    (c) Find $P(1 < Z < 2)$.

43. (S2016.) Students' scores on the SAT Math approximately follow a normal distribution with mean 511 and standard deviation 117.

    (a) I randomly select a student taking the SAT Math. What is the probability they score more than 665?

    (b) I randomly and independently select two students taking the SAT Math. What is the probability the sum of their scores is more than 1330?

    (c) I randomly and independently select ten students taking the SAT Math. What is the probability the highest of the ten scores is more than 665?

44. (S2016) Let $X$ be a continuous random variable with the following probability density function (PDF):

$$f(x) = \begin{cases} 0.15 & -5 \leq x < 0 \\ 0.05 & 0 \leq x \leq 5 \\ 0 & \text{otherwise.} \end{cases}$$

To answer the following questions, it may help to draw a graph of this PDF.

    (a) Find $F(y)$, the cumulative distribution function of $X$. Write it in the form given below:

$$F(y) = \begin{cases} ? & y < -5 \\ ? & -5 \leq y < 0 \\ ? & 0 \leq y < 5 \\ ? & y \geq 5 \end{cases}$$

    (b) Find $P(|X| < 2.5)$.

    (c) Let $EX$ be the expected value of $X$. Is $EX$ less than zero, equal to zero, or greater than zero? Either give an intuitive explanation or calculate $EX$.

45. (Su2015) Let $X$ be a continuous random variable with the following probability density function (PDF):

$$f(x) = \begin{cases} (x+6)/12 & -4 \leq x \leq -2 \\ x/12 & 2 \leq x \leq 4 \\ 0 & \text{otherwise.} \end{cases}$$

To answer the following questions, it may help to draw a graph of this PDF.

(a) Find $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

(b) Find $P(|X| < 3)$.

(c) Is the expected value of $X$ less than zero, equal to zero, or greater than zero? Explain.

**Note**: This problem is harder than what you should expect in the exam!

46. (Su2015.) Let $X$ be a standard normal random variable. Let $Y = X^2$.

(a) Find $P(1 < X < 2)$.

(b) Find $P(1 < Y < 2)$.

47. (S2015) Let $X$ be a continuous random variable with the following probability density function (PDF):

$$f(x) = \begin{cases} |x| & -1 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

To answer the following questions, it may help to draw a graph of this PDF.

(a) Find $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

(b) Find the expected value of $X$.

(c) Let $U$ be a uniform random variable on $[-1, 1]$:

$$f(u) = \begin{cases} 0.5 & -1 \le u \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Which has higher standard deviation, $X$ or $U$? Why? (You do not have to do calculate the SDs exactly.)

48. (S2015.)

(a) Let $X$ be a normal random variable with mean $-5$ and standard deviation 10. Find $P(X > 0)$.

(b) Let $Y$ be a standard normal random variable. Find $P(|Y| > 1.5)$.

(c) Let $Z_1, Z_2, \ldots, Z_{10}$ be ten independent standard normal random variables. Find the probability that at least six of them are positive.

49. (F2014) Let $X$ be a continuous random variable with pdf

$$f(x) = \begin{cases} 0.1 & 0 \le x < 2 \\ 0.2 & 2 \le x < 6 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find $F$, the cumulative distribution function of $X$.

(b) The *median* of $X$ is the value $m$ such that $F(m) = 0.5$. Find the median of $X$.

(c) Find the expected value of $X$.

**Note**: Question is phrased this way because the median is not formally introduced until chapter 6.

50. (S2013) Let $X$ be a continuous random variable with probability density function (PDF)

$$f(x) = \begin{cases} \frac{1}{2}x & 0 \leq x < 1 \\ \frac{1}{2} & 1 \leq x < 2 \\ \frac{1}{2}(3 - x) & 2 \leq x < 3 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find $F$, the cumulative distribution function of $X$.

(b) Find the expected value of $X$.

(c) Find the standard deviation of $X$. (Note: Harder than I would ask these days.)

## Ch 6

51. (Su2018) Use the data frame `congress_age` from the library `fivethirtyeight`.[3] The variable of interest is `age`

    (a) (6 points.) Using the variable `age`, graph and/or compute the following:

    - Histogram and kernel density estimate
    - Plug-in estimates of the population mean, variance, median,
    - Boxplot
    - Interquartile range (iqr)
    - The ratio of the iqr to the square root of the plug-in estimate of the variance
    - Normal probability plot (QQ-Plot).

    (b) (2 points.) Based on your results in part (a), do you think that this sample was drawn from a normal distribution? Explain?

    (c) (2 points.) If you select a random congressman from these data, what is the probability that her/his age is at least 35 but no more than 45 years? (note that you should include 35 and 45 in the group of interest to obtain this probability).

52. (S2017) On weekday mornings, I leave the house at 11:40 am to catch the bus. The bus goes past my house at an approximately uniform time between 11:40 am and 11:55 am. Let $X$ be a uniform continuous random variable representing the time (in minutes) that I have to wait for the bus. The expected value of $X$ is 7.5 minutes.

    (a) Find $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

    (b) Find the first quartile, the median, and the third quartile of $X$.

    (c) (Hard.) Suppose that I am sick of always being early and decide to get to the bus stop at 11:45 am. If I miss the bus, the next one passes my house at an approximately uniform time between 12:10 pm and 12:25 pm. Let $Y$ be the length of time (in minutes) that I wait for a bus under this new strategy. Is the expected value of $Y$ more, less, or the same as 7.5? What is the expected value of $Y$? (You may assume that bus arrival times are independent.)

53. (Su2016) Suppose that in a certain year, an alien attack will happen at a random time on July 4th. Let $X$ be a continuous random variable representing the time of the attack in hours after midnight. Recall that there are 24 hours in a day.

    (a) Write down $f(x)$, the probability density function (PDF) of $X$, for all $x$-values from $-\infty$ to $\infty$.

    (b) Find $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

    (c) President Bill Pullman wants to know a time such that there's a 60% chance the attack has happened by that time. Find the 0.6-quantile of $X$.

[3] The raw data behind the story "Both Republicans And Democrats Have an Age Problem"

54. (Su2016) Let $X$ be a continuous random variable with the following probability density function (PDF):

$$f(x) = \begin{cases} 0.02x & 0 \leq x \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

To answer the following questions, it may help to draw a graph of this PDF.

(a) Find $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

(b) Find $P(X > 5)$.

(c) Is the expected value of $X$ bigger, smaller, or the same as the median? Explain.

55. (Su2016) Let $X$ be a continuous random variable with the following probability density function (PDF):

$$f(x) = \begin{cases} 0.01x + 0.05 & 0 \leq x \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

To answer the following questions, it may help to draw a graph of this PDF.

(a) Find $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

(b) Find $P(X > 5)$.

(c) Is the expected value of $X$ bigger, smaller, or the same as the median? Explain.

56. (S2016) Let $X$ be a uniform random variable with probability density function

$$f(x) = \begin{cases} c & 5 \leq x \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

where $c$ is a constant.

(a) Find $c$.

(b) Find $F(y)$, the cumulative distribution function (CDF) of $X$, for all $y$-values from $-\infty$ to $\infty$.

(c) What is the interquartile range of $X$?

57. (F2015) Let $X$ be a continuous random variable with cumulative distribution function

$$F(y) = \begin{cases} 0 & y < 1 \\ 1 - \frac{1}{y^2} & y \geq 1. \end{cases}$$

and probability density function

$$f(x) = \begin{cases} \frac{2}{x^3} & x \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the median of $X$.

(b) Find the interquartile range of $X$.

(c) Is the expected value of $X$ less than, equal to, or greater than its median? Explain.

58. (F2015.) Let $X$ be a standard normal random variable. Let $Y = |X|$.

    (a) What is the median of $Y$?

    (b) What is $P(1 < Y < 2)$?

    (c) What is the 0.95-quantile of $Y$?

59. (S2015) Let $X$ be a continuous random variable with probability density function

$$f(x) = \begin{cases} 2(x-1) & 1 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases}$$

    (a) Find the median of $X$.

    (b) Find $P(X \leq 1.5)$.

    (c) Suppose we observe independent random variables $X_1, \ldots, X_{10}$, all with the same distribution as $X$. What is the probability that the *maximum* of $\{X_1, \ldots, X_{10}\}$ is less than or equal to 1.5?

60. (F2014.) Let $X$ be a standard normal random variable. Let $Y = X^2$.

    (a) Find $P(Y > 1)$.

    (b) Find the 0.9-quantile of $Y$.

61. (S2014.) Let $A, B, C, D$, and $E$ be independent standard normal random variables. What is the probability that at least two of the five variables are greater than 1?

**Ch9**

62. (S2017) I downloaded data on the number of citations for a random sample of 1000 journal articles published in 1981. (The data is from the ISI Citation Indexes.) I ran some analysis on the data in R, and produced the following output:

```
> citations = scan("citations.txt")
Read 1000 items
> summary(citations)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    1.00    9.06    7.25  300.00
> var(citations)
[1] 565.2476
> # Number of articles with no citations
> sum(citations == 0)
[1] 460
```

(a) Is the distribution of the number of citations (i) exactly normal, (ii) approximately normal, or (iii) not close to normal? How do you know?

(b) Find an approximate 95% confidence interval for the mean number of citations.

(c) Find an approximate 95% confidence interval for the *proportion* of journal articles with no citations.

63. (S2017) The Cooperative Congressional Election Study is a large survey carried out in the U.S. after major elections. In the survey for 2016, 45,242 respondents said they voted in the Presidential election. Of them, 18,755 said they voted for Donald Trump.

Recall that the standard error of a sample proportion is

$$\sqrt{\frac{\text{sample proportion} \times (1 - \text{sample proportion})}{n}}.$$

(a) What percentage of the sample voted for Donald Trump? (To get credit for this question, you must give your answer as a percentage and you must round appropriately.)

(b) Treat the data as a simple random sample of all voters. Using the Central Limit Theorem or otherwise, find a 95% confidence interval based on this data for the percentage of Presidential election voters who voted for Trump. If you can't calculate the interval, give R code.

(c) Suppose we wish to test the null hypothesis that 46.1% of Presidential election voters voted for Trump. Write down null and alternative hypotheses in mathematical notation for this test.

(d) The $P$-value for this test is two times the binomial probability (under the null hypothesis) that out of 45,242 random Presidential election voters, 18,755 or fewer voted for Trump. Write down R code to find the $P$-value.

(e) The $P$-value for your test in part (c) is basically zero, implying that the null hypothesis should be rejected. However, we know that in fact, 46.1% of Presidential election voters did vote for Trump. What explains the discrepancy? (Hint: “Voter fraud” is not the correct answer.)

64. (F2016) A survey in Denmark studied the first three children in a random sample of 154,443 families with three or more children. One variable the observed was the number of the children in each family that were girls (out of 3.) The distribution of the number of girls is given in the table below.

| Number of girls in the first three children | Number of families with that number of girls |
|:---:|---:|
| 0 | 23236 |
| 1 | 58529 |
| 2 | 53908 |
| 3 | 18770 |
| Total | 154443 |

Let $X$ be the number of girls in the first three children of a random Denmark family with at least three children. Let $\mu$ be the expected value of $X$.

(a) Find the mean and variance of the sample. You may use either the plug-in or the sample version of the standard deviation; it doesn't matter.

(b) Find a 95% confidence interval for $\mu$. (If you could not solve part (a) correctly, use a sample mean of 1.5 and a sample standard deviation of 0.75.)

(c) Suppose we wish to perform a similar survey in the United States to find the average number of girls in the first three children in families with at least three children. If we want a 95% confidence interval of width 0.02, how many families should we sample?

65. (F2016) In 2014 in the US, there were 4,010,532 recorded births in total. Of these, 327,680 were in June. June has 30 days and 2014 had 365 days. For this question, treat the births in 2014 as an IID sample from a larger population of births (this is not literally true but is a sufficient approximation.)

(a) Suppose we wish to test the hypothesis that the probability of being born in June is proportional to the number of days in June. Write down mathematical null and alternative hypotheses for this test.

(b) Recall that `qnorm(0.975)` is about 1.96. Using the Central Limit Theorem, calculate an approximate 95% confidence interval for the probability a child is born in June.

(c) The $P$-value for the test of your hypotheses in (a) is 0.0004. Using both this and your confidence interval, explain what you can conclude about the probability of being born in June.

66. (Su2016) The basketball player Steph Curry sometimes shoots free throws with his mouthguard in his mouth, and sometimes shoots free throws with his mouthguard outside of his mouth. His free throw statistics for one season were:

- Free throws with mouthguard in: 110 completed, 13 missed (89.4%)
- Free throws with mouthguard out: 198 completed, 16 missed (92.5%)

His observed free throw completion rate was slightly higher when his mouthguard was outside his mouth. However, we should check whether the difference could be plausibly explained as luck.

(a) Find an approximate 95% confidence interval for the probability that Curry completes a free throw with his mouthguard *in*. Give a numerical answer. Hint: Recall that the estimated standard error of a proportion is

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

(b) Find an approximate 95% confidence interval for the probability that Curry completes a free throw with his mouthguard *out*.

(c) Suppose we wish to test the null hypothesis that Curry's probability of completing a free throw is the same with his mouthguard in as it is with his mouthguard out. The $P$-value for such a test is 0.33. What does this $P$-value tells you? Explain.

67. (Su2016) Three of the best-known types of cricket matches are Test cricket, ODI (One Day International) cricket, and Twenty20 International cricket. A coin toss takes place at the start of all of these games. Does the team that wins the coin toss have an advantage?

There have been 1,442 Test cricket matches in which there has been a winner and in which the result of the coin toss in the beginning of the game has been recorded. (Draws and other games with no winner are excluded.) In these matches, the team that won the coin toss won the match 760 times, while the team that lost the coin toss won the match 682 times.

There have been 3,591 ODI cricket matches in which there has been a winner and in which the result of the coin toss in the beginning of the game has been recorded. In these matches, the team that won the coin toss won the match 1804 times, while the team that lost the coin toss won the match 1787 times.

Suppose we wish to use the binomial to test whether the team that wins the coin toss has a higher chance of winning than the team that loses the toss. We will do this separately for both Test matches and ODI matches.

(a) Is there evidence that in Test matches, the team that wins the coin toss has a higher chance of winning than the team that loses the toss? Use the following R output to find a $P$-value to three decimal places, and give a conclusion:

```
> pbinom(681, 1442, 0.5)
[1] 0.01872661
> pbinom(682, 1442, 0.5)
[1] 0.02127576
> pbinom(759, 1442, 0.5)
[1] 0.9787242
> pbinom(760, 1442, 0.5)
[1] 0.9812734
```

(b) Is there evidence that in ODI matches, the team that wins the coin toss has a higher chance of winning than the team that loses the toss? Use the following R output to find a $P$-value to three decimal places, and give a conclusion:

```
> pbinom(c(1786, 1787, 1803, 1804), 3591, 0.5)
[1] 0.3819477 0.3947366 0.6052634 0.6180523
```

(c) There have not yet been enough Twenty20 International matches to draw accurate conclusions about the effect of the coin toss. Let $p_{20}$ be the probability that the team that wins the coin toss in a Twenty20 International wins the game. Suppose we wished to find a 95% confidence interval for $p_{20}$ that had a length (lower bound to upper bound) of 0.05. Approximately how many games would we need to observe? (Ignore games with no winner.)

68. (S2016) In one year in the United States, 4.247 million babies were born. Of these, 2.173 million were male and 2.074 million were female. With very few exceptions (e.g. identical twins), the sexes of the babies are independent, so we can use the binomial distribution to model the number of babies that are female. Let $p$ be the probability that a random (future) newborn is female. Recall that the estimated standard error of a proportion is

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

(a) What percentage of the babies were female? (To get credit for this question, you must give your answer as a percentage and you must round appropriately.)

(b) Suppose we wish to test the null hypothesis that the probability a baby is female is 50%. Write down null and alternative hypotheses in mathematical notation for this test.

(c) The $P$-value for this test is two times the binomial probability (under the null hypothesis) that out of 4.247 million babies, 2.074 million or fewer are female. Write down R code to find the $P$-value.

(d) Using the Central Limit Theorem or otherwise, find a 95% confidence interval for the probability that a birth is female. If you can't calculate the interval, give R code.

(e) The $P$-value for your test in part (c) is basically zero. From this and your confidence interval, write in a sentence your conclusion about the probability that a random newborn is female.

69. (F2015) As you may know, the range of household incomes in Bloomington is wide — from a large number of student households with very low incomes to a small number of households of very well-paid senior university employees. Suppose the City of Bloomington surveys a random sample of 100 of its households regarding annual household income.

(a) Is the distribution of annual household income in Bloomington (i) exactly normal, (ii) approximately normal, or (iii) not close to normal? Explain.

(b) When taking a sample of 100 Bloomington households, will the sampling distribution of the sample mean annual household income be (i) exactly normal, (ii) approximately normal, or (iii) not close to normal? Explain.

(c) Suppose the county finds that a 95% confidence interval for mean annual household income is \$65,000 to \$95,000. Does this mean that 95% of households have annual household income between \$65,000 to \$95,000? Explain.

70. (Su2015) In a controversial 2011 paper in the Journal of Personality and Social Psychology, a researcher claimed to have found evidence of precognition. In one experiment, Cornell students sat in front of computer screens with two windows. They were asked to click the window that they thought had a picture behind it; after clicking, a picture would then randomly appear behind one of the two windows. In all, the students correctly clicked the window with the picture 1844 out of 3600 times. However, when the pictures were erotic, the students correctly clicked the window with the picture 828 out of 1560 times.

Note: Recall that the standard error of a proportion is

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

(a) Using the Central Limit Theorem, calculate a 95% confidence interval for the proportion of the time students click correctly for all pictures.

(b) Using the Central Limit Theorem, calculate a 95% confidence interval for the proportion of the time students click correctly for erotic pictures.

(c) Another team of researchers wanted to repeat the experiment for erotic pictures, but with a larger sample size, so that the width of a 95% confidence interval for the proportion would be 0.02. How large a sample would they need?

71. (S2015) A controversial 2007 paper in the Journal of Theoretical Biology was titled "Beautiful parents have more daughters," and claimed that children of beautiful parents were much more likely to be female than children of the rest of the population. To investigate the matter further, Gelman and Weakliem (2009) performed the following study:

- Consider the celebrities on People Magazine's 50 Most Beautiful People list from 1995–2000.
- Count the number of boys and girls these beautiful people had (up to August 2007.)
- In the general population, the percentage of births that are girls is 48.5 percent. Test the null hypothesis that children of the Most Beautiful People have a 48.5% chance of being girls.

Collecting the data, they found the Most Beautiful People had 157 girls out of 329 children.

(a) Write R code to find a $P$-value for the test specified above.

(b) Using the Central Limit Theorem, find a 95% confidence interval for the probability that a child of the Most Beautiful People is a girl. Give a numerical answer.

(c) Are beautiful parents more likely to have more daughters than other parents? What can you conclude? (Do not merely write "reject" or "do not reject.")

## Ch 10

72. (S2017) Twelve pairs of identical twins take tests to measure their aggression on a scale from 0 to 100, where 100 is the most aggressive. The researchers' question is whether first born and second born twins differ in mean aggressiveness. They are willing to assume that first born scores, the second born scores, and their differences are all approximately normal. Plots show no strong skewness or outliers, so we can do an appropriate $t$-test.

| First born | Second born |
|---|---|
| 86 | 88 |
| 71 | 77 |
| 77 | 76 |
| 68 | 64 |
| 91 | 96 |
| 72 | 72 |
| 77 | 65 |
| 91 | 90 |
| 70 | 65 |
| 71 | 80 |
| 88 | 81 |
| 87 | 72 |

Table 1: Aggressiveness of twelve pairs of identical twins.

I entered the data into R as vector `first` and `second`. Some summary statistics appear below:

```
> summary(first)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  68.00   71.00   77.00   79.08   87.25   91.00
> sd(first)
[1] 8.887768
> summary(second)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  64.00   70.25   76.50   77.17   82.75   96.00
> sd(second)
[1] 10.37333
> summary(first - second)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -9.000  -2.750   1.000   1.917   5.500  15.000
> sd(first - second)
[1] 7.153617
```

(a) Is this a problem with one independent sample or two independent samples? Explain.

(b) Find an approximate 95% confidence interval for the mean difference in first-born aggression minus second-born aggression. The following R output will be useful:

```
> qt(.975, df = 11)
[1] 2.200985
```

(c) The $P$-value for a two-tailed $t$-test of the null hypothesis that there's no mean difference is 0.37. From this, your confidence interval, and the sample size, carefully state what you can and cannot conclude. (You must give a full and complete answer for full credit.)

73. (F2016) In a randomized experiment in Georgia, a treatment group of 592 convicts received cash payments upon being released from prison, while a control group of 154 convicts received no money upon release. In the first year after release, the members of the treatment group averaged 16.8 weeks of paid work, with a standard deviation of 15.9 weeks. The members of the control group averaged 24.3 weeks of paid work, with a standard deviation of 17.3 weeks. The samples were large and right-skewed.

We wish to test the hypothesis that the treatment and control will, on average, result in the same number of weeks worked.

(a) To allow interpretation, the researcher would prefer not to transform the data. Explain why we may do a Welch's $t$-test even though the samples are right-skewed.

(b) Calculate the test statistic, and give R code to find the $P$-value. Notes: The correct number of degrees of freedom is about 225. We do not have the full data set so we cannot use `t.test()`.

(c) Calculate a 95% confidence interval for the average difference in weeks worked between the treatment and control. (Hint: Degrees of freedom will be very large, so you can use a normal distribution in place of a $t$-distribution.)

74. (F2016) An experiment was performed to examine the effect of caffeine on blood flow. Eight participants had their blood flow measured both before and after consuming caffeine. The results (expressed as percentage change, where a positive value indicates an increase) were:

$$-1.85, -0.25, -0.88, -1.46, -1.05, -1.67, -1.74, -0.33$$

This data is consistent with an approximately normal distribution.

(a) Suppose we had the "before" and "after" measurements. Explain why we should use the percentage changes in a one-sample test rather than using the "before" and "after" measurements in a two-sample Welch's $t$-test.

(b) The sample standard deviation ($s$) of the percentage changes above is 0.63. Using the fact that the R code `qt(0.975, df=7)` gives a value of 2.36, find a 95% confidence interval for the mean percentage change, and explain what this confidence interval means without using the word "confident."

(c) (Requires sign test.) Suppose we wish to perform a two-tailed sign test of the null hypothesis that the median percentage change is zero. Find a numerical value for the $P$-value of this test, and explain what this $P$-value means.

75. (F2016) A statistician studying the 2016 Presidential election uses R to build a probability model that's too complicated to use for exact calculations. He wants to know the model's expected difference between Hillary Clinton's percentage of the two-party vote in Pennsylvania and her percentage of the two-party vote in Michigan. He runs 4000 simulations from the model. Let $X_i$ be the percentage of the Pennsylvania two-party vote obtained by Hillary Clinton in the $i$th simulation. Let $Y_i$ be the percentage of the Michigan two-party vote obtained by Clinton in the $i$th simulation. Let $D = E(X_i - Y_i)$.

    The simulation results are given in vectors called `Penn` and `Mich` (numbers are in percentage points.) Summaries appear below:

```
> summary(Penn)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  47.57   51.42   52.33   52.32   53.21   57.37
> summary(Mich)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  47.43   51.05   52.01   52.01   52.98   56.27
> summary(Penn - Mich)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.4860 -0.4839  0.3183  0.3075  1.0940  4.3290
> var(Penn)
[1] 1.781306
> var(Mich)
[1] 1.998398
> var(Penn - Mich)
[1] 1.362384
```

    (a) Give the test statistic for a two-tailed test of the hypothesis that $D = 0$. Will the $P$-value be (i) close to zero, (ii) close to 1, or (iii) close to 0.5?

    (b) Find an approximate 95% confidence interval for $D$.

    (c) Suppose the statistician wishes to estimate $D$ with a 95% confidence interval of width 0.02 percentage points. To the nearest thousand, how many simulations will he need?

76. (Su2016) It has long been asserted that the average body temperature was 98.6 degrees Fahrenheit. A 1992 study aimed to test this hypothesis. (The data presented here is fictionalized but similar to the study data.) The body temperatures of a sample of 130 adults were taken to one decimal place. The mean temperature of the sample was 98.5 degrees, the median was 98.3 degrees, and the standard deviation was 0.73 degrees. On the previous page, Figure 76 shows a histogram of the data, while Figure 76 shows a normal quantile plot of the data.

    (a) From the information provided, does it seem like the distribution of body temperatures is (i) exactly normal, (ii) approximately normal, or (iii) not close to normal? Explain your choice.

    (b) Let $\mu$ be the population mean body temperature (not the median!) We wish to test $H_0 : \mu = 98.6$ against $H_1 : \mu \neq 98.6$. Assuming this is a random sample, calculate a test statistic and give R code for the $P$-value of this test.

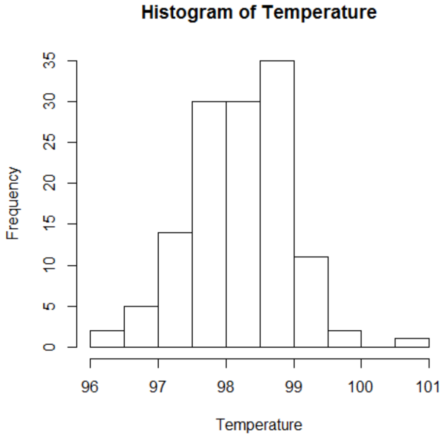


Figure 1: Histogram of body temperatures of 130 adults.

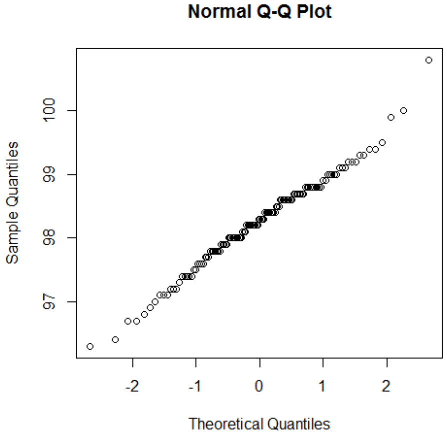


Figure 2: Normal quantile plot of body temperatures of 130 adults.

(c) Construct an approximate 95% confidence interval for the population mean body temperature. (Give a numerical answer — apply the Central Limit Theorem if necessary.) Summarize the evidence for or against the null hypothesis.

77. (Su2016) In a study of a wave power generator, experiments were carried out on scale models in a wave tank to establish how two different choices of mooring method affected the bending stress produced in part of the device. The wave tank could simulate a wide range of sea states and the model system was subjected to the same sample of sea states with each of two mooring methods, one of which was considerably cheaper than the other. The question of interest is whether bending stress differs for the two mooring methods.

The data frame waves contains the following:

```
> waves
   method1 method2
1     2.23    1.82
2     2.55    2.42
3     7.99    8.26
4     4.09    3.46
5     9.62    9.77
6     1.59    1.40
7     8.98    8.88
8     0.82    0.87
9    10.83   11.20
10    1.54    1.33
11   10.75   10.32
12    5.79    5.87
13    5.91    6.44
14    5.79    5.87
15    5.50    5.30
16    9.96    9.82
17    1.92    1.69
18    7.38    7.41
```

Here method1 and method2 are the stresses (in Newton meters) for the two different mooring methods. Some summary statistics follow:

```
> summary(waves$method1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.820   2.310   5.790   5.736   8.732  10.830
> summary(waves$method2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.870   1.970   5.870   5.674   8.725  11.200
> summary(waves$method2 - waves$method1)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.63000 -0.20750 -0.11500 -0.06167  0.08000  0.53000
```

```
> var(waves$method1)
[1] 11.82893
> var(waves$method2)
[1] 12.54725
> var(waves$method2 - waves$method1)
[1] 0.08414412
```

The data appears close to normal.

(a) We want to test for an average difference in stress between the methods. Explain why we should use a one-sample $t$-test rather than a two-sample $t$-test here.

(b) Write down the null and alternative hypotheses for a one-sample $t$-test, and calculate an appropriate $t$-statistic.

(c) The $P$-value for the test is 0.38 and a 95% confidence interval for the average difference (Method 2 minus Method 1) goes from $-0.21$ to 0.08. What do you conclude about the two methods?

78. (Su2016)

(a) A study wished to examine whether average systolic blood pressure differed between Americans with diabetes and Americans without diabetes. In the following R output, `diabetes.BPS` is a vector of systolic blood pressure for 744 Americans with diabetes, while `nodiabetes.BPS` is a vector of systolic blood pressure for 7802 Americans without diabetes. What can you conclude from this R output?

```
> t.test(diabetes.BPS, nodiabetes.BPS)

Welch Two Sample t-test

data:  diabetes.BPS and nodiabetes.BPS
t = 14.56, df = 851.56, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.279825 12.171466
sample estimates:
mean of x mean of y
 127.9516  117.2260
```

(b) The same study also measured diastolic blood pressure for the same samples. In the following R output, `diabetes.BPD` is the vector of diastolic blood pressures for Americans with diabetes, while `nodiabetes.BPD` is the vector of diastolic blood pressures for Americans without diabetes. What can you conclude from this R output?

```
> t.test(diabetes.BPD, nodiabetes.BPD)

Welch Two Sample t-test
```

```
data:  diabetes.BPD and nodiabetes.BPD
t = 1.1101, df = 895, p-value = 0.2673
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4641581  1.6728508
sample estimates:
mean of x mean of y
 68.03898  67.43463
```

(c) (Requires chapter 10.2.) The pulses of a sample of 747 Americans with diabetes were measured by taking a count over 30 seconds and multiplying by 2. Relevant data is in Table 1. Find a 95% confidence interval for the median pulse rate for all Americans with diabetes, carefully stating whether endpoints are included or excluded. The following R output will help:

```
> pbinom(346,747,0.5)
[1] 0.02405488
> pbinom(347,747,0.5)
[1] 0.02851132
```

```
> pbinom(399,747,0.5)
[1] 0.9714887
> pbinom(400,747,0.5)
[1] 0.9759451
```

| Pulse | Number of people |
|---|---:|
| 68 or less | 313 |
| 70 | 47 |
| 72 | 33 |
| 74 | 30 |
| 76 or more | 324 |

Table 2: Distribution of pulse rate in survey of 747 Americans with diabetes. Data adapted from NHANES survey. Only even numbers were observed.

79. (S2016) Rosene (1950) studied how quickly hairs on radish roots absorbed water when they were immersed. For each of eleven radishes, she measured the rate of influx of water for a young root hair and an old root hair on that radish. The data is given below.

| Radish | Old | Young |
|---|---:|---:|
| A | 0.89 | 2.13 |
| B | 0.49 | 1.16 |
| C | 0.91 | 2.60 |
| D | 0.80 | 1.58 |
| E | 0.56 | 1.53 |
| F | 0.79 | 1.70 |
| G | 0.47 | 2.67 |
| H | 0.50 | 2.64 |
| I | 1.08 | 2.19 |
| J | 1.65 | 2.54 |
| K | 1.94 | 4.46 |

Table 3: Radish root hair absorption data. Rates are in cubic microns per square micron per minute.

For each pair, the "Young" number is bigger than the "Old" number, so even without a test it's clear that young roots take in water more quickly. But how much more quickly?

(a) Explain why we should study this data using techniques for one sample rather than techniques for two independent samples.

(b) A normal probability (qqnorm) plot of the differences (old minus young) is on the next page. Explain why:

i. We should be hesitant do a $t$-test on these differences (old minus young);

ii. We should not take the logs of these differences (old minus young.)

(c) Instead of using the differences, we can use the *ratios*: young divided by old. The ratios come from a distribution that's much closer to normal, so we can use one-sample $t$ inference. Write R code that enters the data and finds a 95% confidence interval for the average value of this ratio. The first line is given for you:

```
old = c(89,  49,  91,  80,  56 , 79,  47,  50, 108, 165, 194)
```

80. (F2015) Wilder and Rypstra (2004) tested the effect of praying mantis excrement on the behavior of wolf spiders. They put 12 wolf spiders in individual containers. Each container had two semicircles of filter paper: one semicircle that had been smeared with praying mantis excrement, and one without excrement. They observed each spider for one hour, and measured its walking speed while it was on each half of the container. They used a $t$-test at level $\alpha = 0.05$ to see if, on average, there was a difference between walking speed on the paper with excrement and on the paper without excrement.

    (a) What is the experimental unit? What measurements are taken on the experimental units? Is this a problem with one or two independent samples?

    (b) Give null and alternative hypotheses for an appropriate two-tailed $t$-test. If the null hypothesis is true, what is the distribution of the $t$-statistic?

    (c) The $P$-value (significance probability) was calculated to be 0.053, so the null hypothesis was not rejected. From this and the other information given, is it correct to conclude that we are sure that wolf spiders' walking speed is not affected by praying mantis excrement? Explain.

81. (F2015) Boxes of cereal are advertised as having a net weight of 8 ounces. The weights of boxes are assumed to be normally distributed. A new cereal box-filling machine is purchased, and we wish to be sure that on average, it puts at least the correct amount of cereal in the boxes. We randomly select 16 boxes, and find they have weights with sample mean 8.10 ounces and sample standard deviation 0.20 ounces.

    (a) Suppose your data included all the box weights. How would you check the normal distribution assumption? Explain what would indicate a violation of this assumption. (Note: You do not have to perform the check on the given data.)

    (b) Perform a $t$-test of the null hypothesis that the average net weight of the boxes of cereal produced by the new machine is less than or equal to 8 ounces, against the alternative that it is greater than 8 ounces. Test at level $\alpha = 0.05$. You may wish to use the fact that the R command `pt(2, df=15)` gives the output 0.968.

    (c) Suppose five of the sixteen boxes weighed less than 8 ounces (and none were exactly equal to 8 ounces.) We wish to perform a sign test of the null hypothesis that the median net weight of the boxes of cereal produced by the new machine is less than or equal to 8 ounces, against the alternative that it is greater than 8 ounces. Using the table of `pbinom` values below, find the numerical value for the sign test $P$-value.

82. (Su2015 Requires Trosset chapter 10.2.) The lecturer of a large statistics class decides to ask a question on the final very similar to one on the He hypothesizes that on average, the

| | |
|---|---|
| `pbinom(0, 16, 0.5)` = 0.000 | `pbinom(9, 16, 0.5)` = 0.773 |
| `pbinom(1, 16, 0.5)` = 0.000 | `pbinom(10, 16, 0.5)` = 0.895 |
| `pbinom(2, 16, 0.5)` = 0.002 | `pbinom(11, 16, 0.5)` = 0.962 |
| `pbinom(3, 16, 0.5)` = 0.011 | `pbinom(12, 16, 0.5)` = 0.989 |
| `pbinom(4, 16, 0.5)` = 0.038 | `pbinom(13, 16, 0.5)` = 0.998 |
| `pbinom(5, 16, 0.5)` = 0.105 | `pbinom(14, 16, 0.5)` = 1.000 |
| `pbinom(6, 16, 0.5)` = 0.227 | `pbinom(15, 16, 0.5)` = 1.000 |
| `pbinom(7, 16, 0.5)` = 0.402 | `pbinom(16, 16, 0.5)` = 1.000 |
| `pbinom(8, 16, 0.5)` = 0.598 | |

| Difference (final minus midterm) | Number of students |
|---|---|
| $-4$ | 1 |
| $-3$ | 1 |
| $-2$ | 1 |
| $-1.5$ | 1 |
| $-1$ | 8 |
| 0 | 19 |
| 1 | 16 |
| 2 | 7 |
| 3 | 9 |
| 4 | 6 |
| 6 | 1 |

Table 4: Final question score minus midterm question score for 70 statistics students.

students will do as well on the final question as they did on the midterm question. After the final, he collects the data for 70 students, given in Table 1 below.

He calculate the following summary statistics:

- Midterm question: mean 8.4, SD 1.57
- Final question: mean 9.38, SD 0.99
- Difference (final minus midterm): mean 0.98, SD 1.82

(a) Treating the data as a sample from a larger population, find a 98% confidence interval for the *mean* difference in scores between the final and the You may use the fact that the R code `qt(0.99, df=69)` gives 2.38.

(b) Treating the data as a sample from a larger population, find a 98% confidence interval for the *median* difference in scores between the final and the You may use the fact that the R code `pbinom(25, 70, 0.5)` gives 0.01.

(c) From the data and from your analysis, what can you conclude about the class' performance on the final question compared to the midterm question? (You must give a full answer for full credit.)

83. (Su2015) The Breakfast Club believes the ages of its members are approximately normally distributed. It sends out a survey to a random sample of 10 of its members. The ages of the sample are:

$$18, 13, 18, 16, 12, 16, 17, 17, 20, 18.$$

The following R output will be useful for this question:

```
> ages = c(18, 13, 18, 16, 12, 16, 17, 17, 20, 18)
> sd(ages)
[1] 2.415229
> pt(0.655, df=9)
[1] 0.7355744
> qt(0.975, df=9)
[1] 2.262157
```

(a) Test the null hypothesis that the mean age of all members of the Breakfast Club is no more than 16, finding a $P$-value and giving a conclusion.

(b) Find a 95% confidence interval for the mean age of all members of the Breakfast Club.

(c) The Deadbeat Club sends out a survey to a random sample of 10 of their members. The ages of the sample are:

$$13, 20, 26, 41, 67, 13, 79, 15, 20, 15.$$

Explain why a $t$-test would not be the best choice to test the hypothesis that the mean age of all members of the Deadbeat Club is no greater than 30.

84. (S2015) To find the average number of cars owned per household in the U.S., a random survey of $n = 100$ American households is taken. The results are given in the table below.

| Number of cars | Frequency |
|---|---|
| 0 | 9 |
| 1 | 33 |
| 2 | 38 |
| 3 | 14 |
| 4 | 4 |
| 5 | 2 |
| Total | 100 |

(a) Find the sample mean.

(b) The standard deviation of the number of cars owned by households in the sample is 1.06. Using a $z$-interval or otherwise, find a 95% confidence interval for the mean number of cars owned per household in the U.S.

(c) If calculated correctly, your confidence interval from part (b) will not contain any whole numbers. Your friend says: "This can't be right, because you can only own a whole number of cars." Briefly explain to your friend why they are mistaken.

85. (S2015) We wish to test if the center of a population is zero. We observe a random sample of size 15:

```
x = c(4.5, 7.2, 8.7, 7.0, 7.4, 11.1, 9.9, 5.2, -1.2, 9.8, 3.3, 10.3, 10.1, 8.4, 6.1)
```

(a) Give R code to draw a graph to check the assumption that the data come from an approximately normal distribution, and explain what to look for in that graph.

(b) Suppose we are willing to assume the population is approximately normally distributed. Give R code to find a two-tailed $P$-value.

(c) (Requires Trosset chapter 10.2.) Using the binomial distribution or otherwise, find the numerical value of the two-tailed $P$-value for a sign test of the hypothesis that the median is zero. (If you can't find the numerical value, you may give R code for reduced credit.)

86. (S2015 Requires Trosset chapter 10.2.) The following are monthly rents (in dollars) paid by a sample of students:

$$333, 678, 414, 364, 1997, 764, 874, 1063, 725, 715$$

We wish to know if the center of the rent distribution is \$1,000. Figure 1 shows a normal quantile plot of the data.

(a) Explain why a $t$-test is inappropriate for this data.

(b) Perform a sign test of the hypothesis that the median monthly rent is \$1,000. You may use the fact that the output of the R command `pbinom(2, 10, 0.5)` is 0.055.

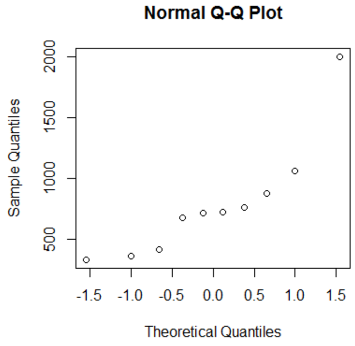


Figure 3: Normal quantile plot of data in question 3.

(c) Find an approximate 98% confidence interval for the median monthly rent. You may use the fact that the output of the R command `pbinom(1, 10, 0.5)` is approximately 0.01.

(d) Explain what you can conclude from your analysis above.

87. (F2014) At a major chain of cafes, the average amount of caffeine per serving in regular coffee is known to be 188 milligrams. In a study published in the Journal of Analytical Toxicology, twelve servings of decaf coffee from the chain were measured for their caffeine content. The results (in milligrams), sorted from lowest to highest, were:

$$3.0, 3.2, 3.3, 4.1, 12.0, 12.5, 12.7, 13.0, 13.0, 13.4, 13.4, 15.8$$

(a) Draw a boxplot or other graph of the data. Does the sample appear to come from a normal population?

(b) Find the $P$-value (significance probability) for a sign test to try to prove that the median caffeine in this chain's decaf coffee is less than 188 milligrams.

(c) Your friend says that the study doesn't prove that the chain's decaf coffee averages less caffeine than their regular coffee, as the sample size is too small. Is your friend right? Explain.

88. (S2014 Requires Trosset chapter 10.2.) The following are the daily precipitation (in hundredths of an inch) in Snoqualmie Falls, WA, for a random sample of 16 days:

$$1, 3, 160, 116, 32, 30, 48, 45, 0, 0, 0, 9, 9, 4, 27, 39$$

(a) Show that a $t$-test/confidence interval is inappropriate for this data. (It is not necessary to draw a graph, but it may help you.)

(b) Find the $P$-value for a sign test of the null hypothesis that the median daily precipitation in Snoqualmie Falls is no more than 0.5 hundredths of an inch.

(c) Find an approximate 98% confidence interval for the median daily precipitation.

The following table of output of the R function `pbinom` may be useful:

| | |
|---|---|
| `pbinom(0,16,0.5)` | 0.00002 |
| `pbinom(1,16,0.5)` | 0.0003 |
| `pbinom(2,16,0.5)` | 0.002 |
| `pbinom(3,16,0.5)` | 0.011 |
| `pbinom(4,16,0.5)` | 0.038 |
| `pbinom(5,16,0.5)` | 0.105 |
| `pbinom(11,16,0.5)` | 0.962 |
| `pbinom(12,16,0.5)` | 0.989 |
| `pbinom(13,16,0.5)` | 0.998 |
| `pbinom(14,16,0.5)` | 0.9997 |
| `pbinom(15,16,0.5)` | 0.99998 |
| `pbinom(16,16,0.5)` | 1 |

**Ch 11**

89. (S2017) The Republican National Committee ran a randomized experiment to test two different colors of button, red and green, for visitors to donaldjtrump.com to select donation amounts. Their findings were:

    - Red button: 6855 visitors, mean revenue: \$4.76 per visitor (sample standard deviation \$29.)
    - Green button: 6639 visitors, mean revenue: \$6.60 per visitor (sample standard deviation \$49.)

    (a) Are the red button revenue distribution and the green button revenue distribution approximately Normal, or not close to Normal? How do you know?

    (b) Find a 95% confidence interval for the *difference* in mean revenue obtained by using a green button instead of a red button. (Hint: Because the sample sizes are large, a confidence interval will extend about ±1.96 standard errors from the point estimate.)

    (c) Two options for a significance test for this data include Welch's two-sample $t$-test and Student's two-sample $t$-test. Which of the two would you use, and why?

    (d) Will the $P$-value for a two-sample $t$-test of the null hypothesis that the green button and the red button give the same mean revenue be close to 0, close to 0.5, or close to 1? Explain.

90. (S2017) A high-profile psychology paper claimed that being in a "high-power pose" like having your hands on your hips (as opposed to a "low-power pose" like sitting with your arms crossed) meant that you felt more powerful, sought more risk, and had higher testosterone levels. However, these results were based on small sample sizes. Here, we will only consider the effect of pose on testosterone.

    Ranehill et al. performed a randomized experiment to see if these findings could be replicated. 104 subjects were randomly assigned to the high-power poses, while 94 were randomly assigned to the low-power pose. The testosterone of the subjects was measured both before and after the pose.

The distributions aren't normal, but the sample sizes are large enough that $t$-tests are justifiable. (A nonparametric test might be slightly better.)

(a) Explain why it's better to do a test on the *differences* between the measurements of testosterone before and after the pose, rather than just using the testosterone after.

(b) I loaded the two samples of testosterone differences (after minus before) into R as vectors `highpower` and `lowpower`. Some R output follows. What can you conclude about the effect of pose on testosterone?

```
> summary(lowpower)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-47.940  -2.013   2.002   6.767  15.200 101.100
> summary(highpower)
     Min.   1st Qu.    Median      Mean   3rd Qu.
-105.2000   -3.5570    0.2215    2.6900   10.1600
     Max.
  62.3100
> t.test(highpower, lowpower, alt = "greater")

Welch Two Sample t-test

data:  highpower and lowpower
t = -1.4098, df = 195.84, p-value = 0.9199
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -8.856176       Inf
sample estimates:
mean of x mean of y
 2.689615  6.766606

> t.test(highpower, lowpower)

Welch Two Sample t-test

data:  highpower and lowpower
t = -1.4098, df = 195.84, p-value = 0.1602
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.780046  1.626064
sample estimates:
mean of x mean of y
 2.689615  6.766606
```

(c) The sample mean for the low-power treatment is actually quite high. A one-sample $t$-test of the alternative hypothesis that a low power pose has a positive effect on testosterone gives the following $P$-value:

```
> t = mean(lowpower) / (sd(lowpower) / sqrt(94))
> 1 - pt(t, df = 93)
[1] 0.0005914953
```

Should we conclude that a low-power pose increases testosterone, or is this test inappropriate? Explain.

91. (S2017) In a genetic inheritance study discussed by Margolin (1988), blood was collected from samples of individuals from several ethnic groups, and mean sister chromatid exchange (MSCE) was measured for each individual. We wish to use a $t$-test to compare the average MSCE for the groups labeled "Native American" and "Caucasian" :

   - Native American: 8.50, 9.48, 8.65, 8.16, 8.83, 7.76, 8.63
   - Caucasian: 8.27, 8.20, 8.25, 8.14, 9.00, 8.10, 7.20, 8.32, 7.70

   The Native American individuals had mean MSCE 8.57 with standard deviation 0.54. The Caucasian individuals had mean MSCE 8.13 with standard deviation 0.49.

   (a) What is the experimental unit (i.e. the individuals in the study)? What measurements are taken on the experimental units? Is this a problem with one or two independent samples?

   (b) Carefully define a parameter (or parameters) of interest for the study. Give null and alternative hypotheses for an appropriate significance test, and calculate the test statistic.

   (c) The $P$-value for the significance test turns out to be 0.11. Does this prove that there is no difference in mean MSCE between Native Americans and Caucasians? Explain why or why not.

92. (F2016) In part of a study reported by Perotta and Finch (1972), the blood films of 16 patients with severe renal anemia and 10 patients with functional heart disease were measured for red blood cell counts. The percentage changes in red blood cell counts are entered into R. Because of non-normality, logs are taken:

```
> log(Renal)
 [1]  0.7884574  0.4187103  0.4317824 -0.2613648 -1.0788097
 [6] -0.7985077 -0.9416085 -1.2378744 -1.7147984 -1.8325815
[11] -1.4696760 -1.4271164 -1.7719568 -2.5257286 -3.9120230
[16] -3.9120230
> log(Heart)
 [1]  0.6097656 -0.8209806 -1.2039728 -2.8134107 -1.6094379
 [6] -1.9661129 -2.3025851 -2.4079456 -2.8134107 -3.2188758
```

   We calculate summary statistics:

```
> mean(log(Renal))
[1] -1.32782
```

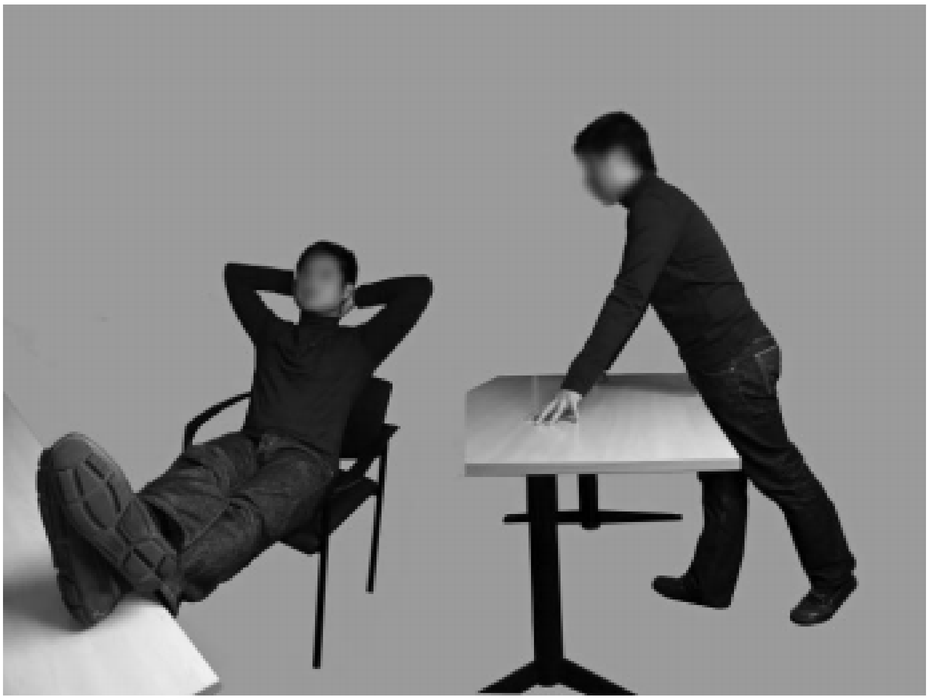


**Fig. 1.** The two high-power poses used in the study. Participants in the high-power-pose condition were posed in expansive positions with open limbs.

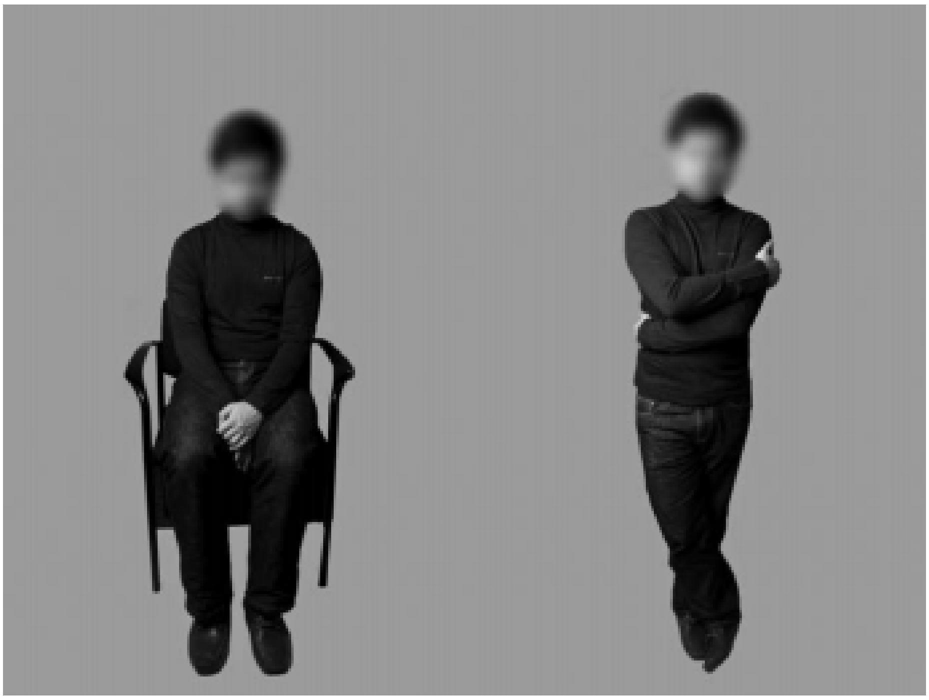


**Fig. 2.** The two low-power poses used in the study. Participants in the low-power-pose condition were posed in contractive positions with closed limbs.

```
> sd(log(Renal))
[1] 1.358551
> mean(log(Heart))
[1] -1.854697
> sd(log(Heart))
[1] 1.147175
```

Suppose we wish to show that the mean log percentage change is *greater* for renal patients than for heart patients.

(a) Write down mathematical null and alternative hypotheses for such a test. Carefully define the parameters you use.

(b) Calculate the observed test statistic and the degrees of freedom for a Welch's $t$-test.

(c) Give R code to produce the correct $P$-value for this Welch test.

(d) The $P$-value for the test turns out to be 0.15, while a 95% confidence interval for the difference in the means of the log percentage changes is $-0.5$ to 1.6. What do you conclude?

93. (Su2016) A randomized experiment was carried out to study a cognitive behavioral treatment for depression called Beat the Blues. 52 patients were assigned to the treatment, while the remaining 48 patients were in the control group. The effectiveness of the treatment was measured using the BDI-II test, which gives a score from 0 to 63 (where 0–13 is "minimal depression" and 29–63 is severe depression.) The data is in R as the vectors `depress.treatment` for the treatment group and `depress.control` for the control group. We can't take a log transformation, as the data contains zeroes, but a square root transformation is a possibility. Summary statistics appear below:

```
> summary(depress.treatment)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   6.188  12.000  14.170  20.000  44.500
> summary(depress.control)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
   1.75    9.00   17.50   17.72   23.50   44.50       3
> summary(sqrt(depress.treatment))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.487   3.463   3.492   4.472   6.671
> summary(sqrt(depress.control))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
  1.323   3.000   4.183   3.987   4.848   6.671       3
```

(Note that there are three missing values in the control group due to participants dropping out of the program; you may ignore these.)

Normal QQ plots of the raw data and transformed data appear on the next page.

(a) What would be the advantage of performing a test on the raw data without transformation? What would be the advantage of performing a test on the transformed data instead?

(b) Write down R code to perform an appropriate *one-tailed* test on either the raw or the transformed data (your choice.) Write down hypotheses and state the assumptions of your test. (Note: If your code does not give the correct one-tailed $P$-value, describe how to get the correct $P$-value.)

(c) The $P$-value of the test that I would carry out is 0.042. What does this tell you about the effectiveness of Beat the Blues?

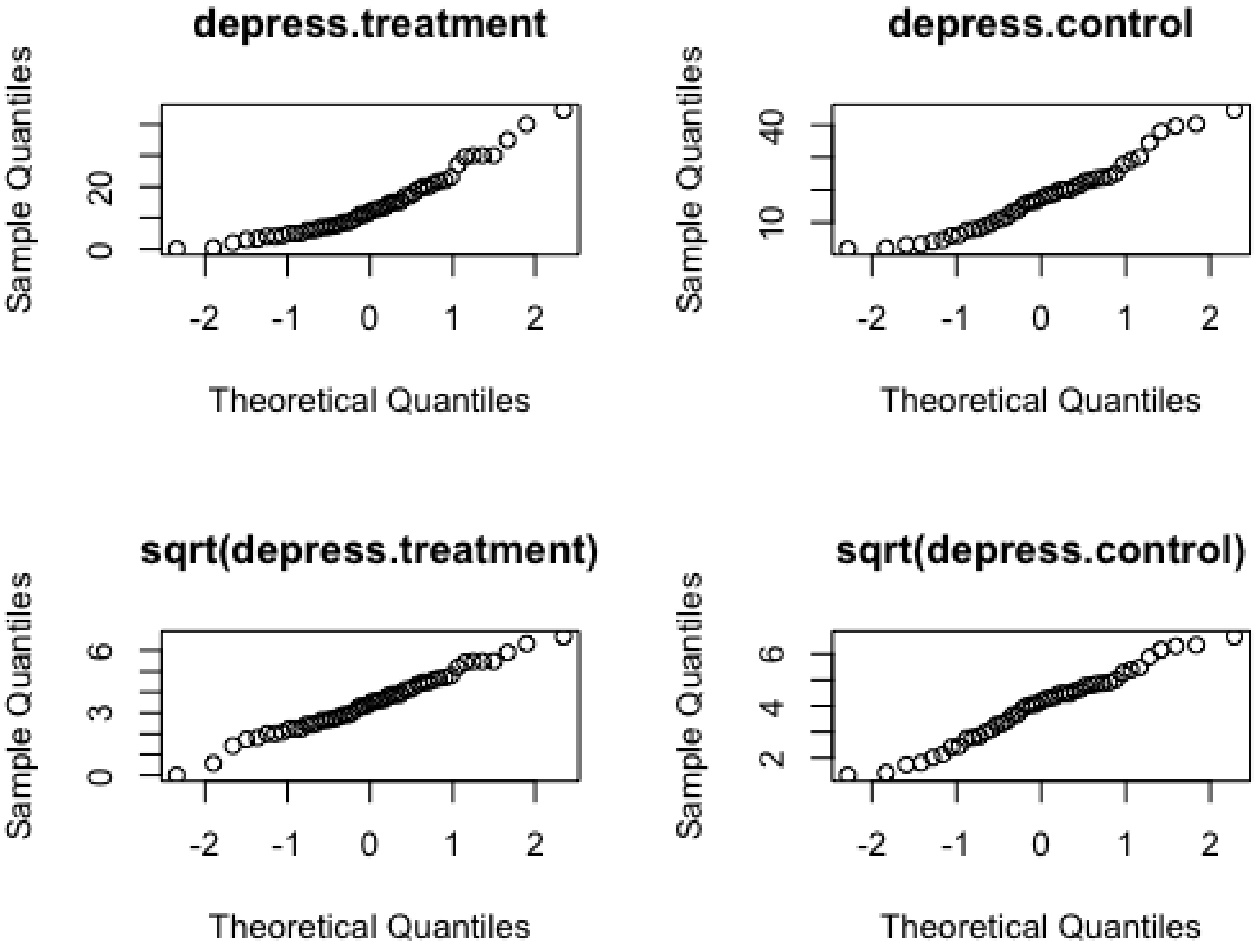


Figure 4: Normal QQ plots of the raw data and transformed data for treatment and control groups in the Beat the Blues study.

94. (S2016) Kolb (1965) reported a study designed to investigate the research hypothesis that children who are underachievers are also low in motivation and/or concern for achievement. For the purpose of the experiment, underachievers were defined as high school boys with an IQ above 120 but a grade average below C. The subjects were randomly divided into two groups for a summer program. Both groups were given academic training, but one group was also given achievement motivation training (AMT.) The academic performance of each subject was evaluated at the end of the summer program by administering tests. Artificial overall scores on these tests follow. We wish to test the theory that achievement motivation has a *positive* effect on academic performance.

- AMT: 66 72 69 80 78 86 70 53 48 76 59 61 65 70 75 63 68 45
- No AMT: 81 85 92 71 68 77 55 90 93 84 73 65

(a) In this problem, is there one independent sample or two independent samples? Explain.

(b) Suppose we decide to perform Welch's $t$-test. State the assumptions of the Welch test, and explain how we would check whether they are satisfied.

(c) Write down null and alternative hypotheses for a one-tailed Welch's $t$-test of this data.

(d) A student analyzes the data and produces the R output below. However, the student has made a big mistake, so the $P$-value is wrong. What is the correct $P$-value? What can you conclude about the effect of AMT?

```
> t.test(amt, noamt, alt="less")

Welch Two Sample t-test

data:  amt and noamt
t = -2.5684, df = 22.328, p-value =
0.008706
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.631954
sample estimates:
mean of x mean of y
 66.88889  77.83333
```

95. (S2016) In January 2016, American National Election Studies (ANES) performed a survey of 570 U.S. men and 629 U.S women. Participants were asked to rate their feelings toward various individuals and groups on a "Feeling Thermometer" scale from 0 to 100, where 100 was most favorable and 0 was most unfavorable. One group they were asked to rate was "scientists."

Figure 5 shows density plots for the two samples.

(a) As far as you can tell from the graphs, do either mens' feelings or womens' feelings toward scientists have a normal distribution when expressed on the Feeling Thermometer scale? Explain why or why not.

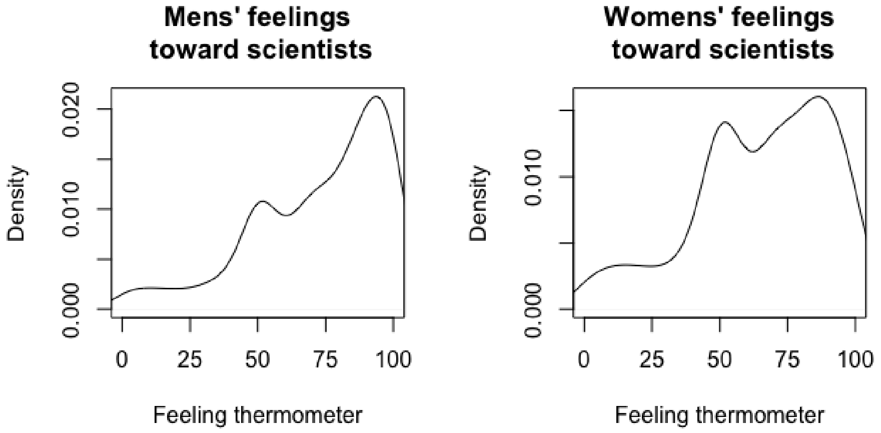


Figure 5: Density plots for feelings of samples of men and women toward scientists. The "Feeling Thermometer" is a 0 to 100 scale, where a higher number expresses more positive feelings.

(b) Suppose we wish to test whether men and women have the same average feeling toward scientists. Is Welch's *t*-test appropriate? If you think so, explain why; if not, describe the test you would do instead.

(c) Suppose we decide to perform Welch's *t*-test. Some R output follows, where `male.sci` contains the Feeling Thermometer scores for the sample of men and `female.sci` contains the Feeling Thermometer scores for the sample of women. In two or three sentences, explain this output to a political scientist who has never used R.

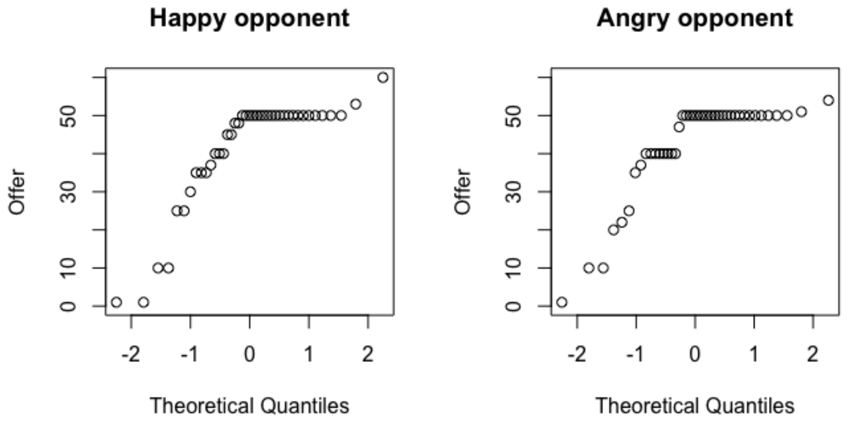
```
> summary(male.sci)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   53.25   80.00   72.79   93.00  100.00
> summary(female.sci)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   56.00   79.00   72.83   92.00  100.00
> t.test(male.sci, female.sci)

Welch Two Sample t-test

data:  male.sci and female.sci
t = -0.026876, df = 1169.7, p-value = 0.9786
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.767701  2.692901
sample estimates:
mean of x mean of y
 72.78772  72.82512
```

96. (S2016) In an experiment to study bargaining, Slowik and Voracek (following a study by van Dijk et al.) let a sample of 83 students play a game. Each subject (Player One) was asked to bargain with another participant (Player Two) over 100 chips by offering the other participant some of the chips. No one offered less than 1 or more than 60. However, before Player One made the offer, 41 of the students (selected at random) were told that Player Two was happy with them, while the other 42 were told that Player Two was angry with them. (In actual fact, Player Two did not exist.)

Did Player Two's emotion affect Player One's offer? We study this with the help of R. The variable `happy` contains the number of chips offered by the 41 players who were told Player Two was happy. The variable `angry` contains the number of chips offered by the 42 players who were told Player Two was angry. Numerical summaries appear below:

```
> summary(happy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   37.00   50.00   41.78   50.00   60.00
> summary(angry)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   40.00   50.00   42.43   50.00   54.00
> sd(happy)
[1] 14.31173
> sd(angry)
[1] 12.80462
```

(a) Normal QQ plots for both the happy and angry groups appear below. Are the samples from normal populations? Explain.



(b) You talk to a Statistics professor, who tells you it's okay to analyze the data using Welch's $t$-test. Let $\Delta$ be the expected value for the happy group minus the expected value for the angry group. Calculate the numerical value of the $t$-statistic for a test of the null hypothesis that $\Delta$ is zero.

(c) The $P$-value for Welch's $t$-test is about 0.83. (If we had instead done a test that did not assume normality, the $P$-value would be similar to this.) What can you conclude? Note: The following R output (with some values removed) may be useful.

```
> t.test(happy, angry)

Welch Two Sample t-test

data:  happy and angry
t = ???, df = 79.549, p-value = 0.8286
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.585368  5.289201
sample estimates:
mean of x mean of y
 41.78049  42.42857
```

(d) Write R code to find the Welch's $t$-test $P$-value *without* using the function `t.test()`. You may assume the variables `happy` and `angry` are already in your workspace.

97. (F2015) We wish to know if the average number of hours studied per week by IU freshmen is the same as the average number of hours studied per week by IU seniors. Suppose we sample 16 IU freshmen and 25 IU seniors. The freshmen have a sample mean of 9.0 hours with sample standard deviation 4 hours, while the seniors have a sample mean of 12.0 hours with sample standard deviation 5 hours. The samples appear to come from distributions that are close to normal.

    Let $\Delta$ be the population average number of hours studied per week by IU seniors minus the population average number of hours studied per week by IU freshmen.

    (a) We wish to test $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$. Should we use Welch's two-sample $t$-test or Student's two-sample $t$-test? Explain your choice.

    (b) Calculate the $t$-statistic for the test you chose in part (a).

    (c) The correct number of degrees of freedom for the $t$-test is about 37. The R command `qt(.975, df=37)` gives a value of 2.026. Use this information to find an approximate 95% confidence interval for $\Delta$. Is there evidence that $\Delta \neq 0$?

98. (F2015) A randomized experiment is carried out to compare two treatments for ulcers. Treatment A is carried out on a sample of 25 patients. Their healing times are approximately normally distributed with sample mean 70 days and standard deviation 32 days. Treatment B is carried out on a sample of 36 patients. Their healing times are approximately normally distributed with sample mean 55 days and standard deviation 21 days. We wish to test for a difference between the two treatments.

    (a) Which kind of test would you use? Explain your choice.

    (b) Calculate the test statistic for the test you chose in part (a), and give R code to produce the $P$-value for a two-tailed test.

(c) Calculate an approximate 95% confidence interval for the difference in average healing time between Treatment A and Treatment B. You may use the fact that the R command `qt(0.975, df)`, where df is the correct number of degrees of freedom, gives the output 2.02.

99. (F2015) A survey of math and statistics majors after graduation is carried out. A sample of 400 math majors has average salary $47,300, with standard deviation $8,000. A sample of 400 statistics majors has average salary $48,600, with standard deviation $9,000.

    (a) The distribution of salaries is not normal; however, we can use techniques based on the normal distribution to calculate confidence intervals and perform hypothesis tests. Why is this?

    (b) Find an approximate 95% confidence interval for the average difference in starting salaries between math and statistics majors.

    (c) A two-tailed test of the null hypothesis that math and statistics majors have the same average starting salary gives a $P$-value of 0.03. A career advisor then claims that almost all statistics majors have higher starting salaries than almost all math majors. Is this claim justified? Explain clearly.

100. (Su2015) Data was collected on the salaries of random samples from the following two populations:

    - Employees of IU Athletics (sample size 50): sample mean $81,090, SD $73,339.
    - IU Bloomington faculty (sample size 100): sample mean $87,040, SD $50,051.

    Boxplots of the data, along with the raw data, are shown on the following page.

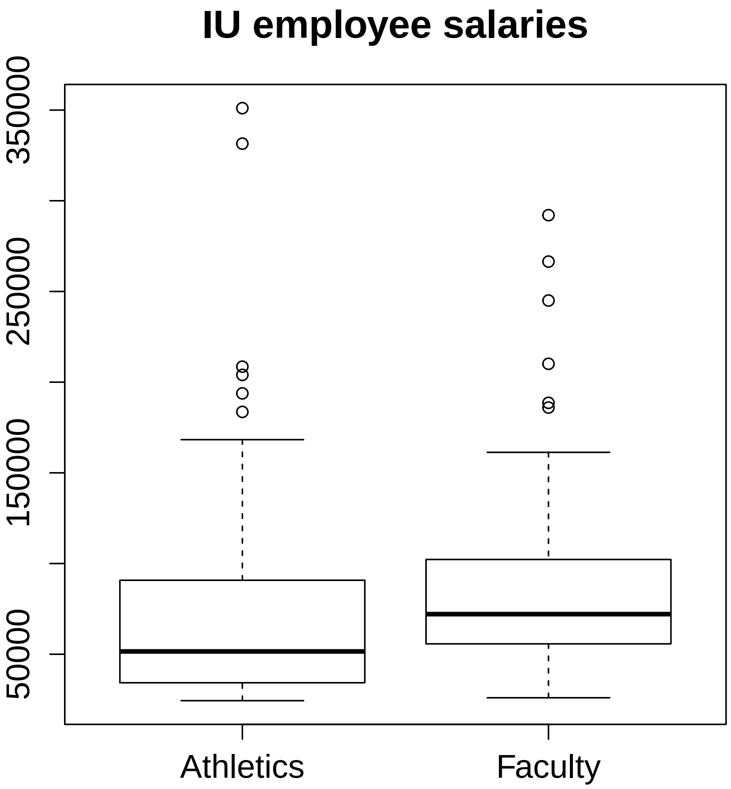
    (a) Which of these approximately follows a normal distribution: athletics salaries, faculty salaries, neither, or both? Explain your answer.

    (b) Find an approximate 95% confidence interval for the mean salary of employees of IU Athletics, and an approximate 95% confidence interval for the mean salary of IU Bloomington faculty. You may either assume the sample sizes are large enough to use techniques based on the Central Limit Theorem, or else use the following R output:

    ```
    > qt(.975, df=49)
    [1] 2.009575
    > qt(.975, df=99)
    [1] 1.984217
    ```

    (c) The $P$-value of a two-tailed test of the hypothesis that the two populations have the same mean is 0.61. An administrator draws the following conclusion: “Employees of IU Athletics and IU Bloomington faculty have similar salary distributions.” As far as you can tell from the data, is this conclusion true or false? Justify your answer to the administrator.

    Raw data (sorted):


IU employee salaries
350000
250000
150000
50000
Athletics
Faculty

```
> ATH
 [1]  24398  24398  25438  27997  28434  29474  31272  32136  32656  33349
[11]  33446  34272  34272  36414  37170  39000  40307  41000  41850  42656
[21]  45008  47586  47861  48450  51038  52020  52530  54060  56100  57222
[31]  58157  65280  71088  76626  83030  86626  88434  90780 102000 109033
[41] 122400 135000 143432 168300 183600 193800 204000 208539 331500 351054
> FAC
  [1]  26000  26010  27411  29304  30000  31378  32640  32640  35000  38610
 [11]  45726  45920  46000  47940  50989  51000  51000  51000  51510  52005
 [21]  52020  52988  54915  54959  55410  56000  56100  56100  56794  56999
 [31]  57818  58106  59000  59260  59405  59558  60573  61200  63189  63240
 [41]  65149  66079  66576  68328  68597  69693  69855  70030  70441  71680
 [51]  72520  73765  76146  78650  79110  80000  81451  81600  82820  85000
 [61]  86018  86071  86190  86553  88022  89543  90535  92887  92948  93806
 [71]  93823  97893  98072  99000 101595 102894 104567 112524 112779 113491
 [81] 114250 116548 118025 122873 124864 135334 138859 139098 152893 153147
 [91] 154462 160650 160899 161300 186000 188600 210108 245000 266472 292060
```

101. (Su2015) In the late 1990s and early 2000s, steroid use in Major League Baseball was widespread. Many observers expected home run rates to decline after stricter drug penalties were introduced.

    The files `hr2000.txt` and `hr2010.txt` contain data on simple random samples of 50 Major League Baseball batters (out of those that played a full season) from 2000 and 2010 respectively. Each file contains two variables:

    - `Homeruns`: the number of home runs hit by the player that year
    - `Appearances`: the number of plate appearances made by the player that year

    Test the (alternative) hypothesis that for players that played a full season, the average *rate* of home runs per plate appearance was lower in 2010 than in 2000. State any assumptions you make and give a $P$-value and a conclusion.

102. (Su2015) A large sample of statistics students is randomly split into two groups, each containing 100 students. Group A is shown *The Joy of Stats*,[4] a documentary about the history and philosophy of statistics. Group B is shown the 1992 Steven Seagal movie *Under Siege*. After the screenings, all students are given the same statistics test. Group A has an average score of 61 with a sample standard deviation of 10. Group B has an average score of 59 with a sample standard deviation of 13. Both sets of scores are approximately normal.

    (a) Is there a significant difference between the averages of the two groups? Calculate a test statistic and give R code for the $P$-value of an appropriate two-tailed test.

    (b) Give R code for a 90% confidence interval for the difference in averages between the two groups.

[4] `http://www.gapminder.org/videos/the-joy-of-stats/`

(c) Suppose that your $P$-value in part (a) comes out to be 0.22 and your confidence interval in part (b) comes out to be $-0.7$ to 4.7 points higher for the inspirational movie group. What do you conclude based on this analysis?

103. (Su2015) Gray-Donald et al. (1985) performed a controlled experiment to examine the effect of withholding supplementary bottle-feedings from newborn babies in a Montreal hospital. Newborns were assigned pseudo-randomly to one of two groups — a "traditional" ward where babies received supplementary bottle-feedings, and an "experimental" ward where babies that did not (in order to encourage breast-feeding.) 393 babies were assigned to the traditional ward, while 388 babies were assigned to the experimental ward. Newborn babies tend to lose some weight in the first few days after birth; losing too much weight is bad. The experimenters measured the percentage of birthweight lost by each baby in the study. Their results were:

- Traditional group: Mean percentage weight loss: 5.1%, SD of percentage weight loss: 2.0%
- Experimental group: Mean percentage weight loss: 6.0%, SD of percentage weight loss: 2.0%

(a) The experimenters wished to test whether the two populations the groups came from had the same or different mean percentage weight loss. State mathematical null and alternative hypotheses and state what kind of test you would do.

(b) Calculate the numeric value of the test statistic of your test in part (a), and give R code for the $P$-value.

(c) The $P$-value for this test turns out to be nearly zero. What do you conclude about the effect of withholding bottle-feedings on newborn babies' weights? (Do not merely write "reject" or "do not reject.")

104. (S2015) A group of 24 stroke patients is randomly split into a treatment group and a control group. The treatment group receives aerobic exercise, while the control group does not. The VO2 (a measure of fitness, in mL/kg/min) of each patient in both groups is measured before and after the treatment. The changes in VO2 (after minus before, positive is good) for the treatment group are

$$-2.3, -0.7, -0.2, 0.1, 0.5, 0.8, 0.9, 1.6, 2.0, 3.9, 4.5, 6.0$$

while the changes for the control group are

$$-2.9, -1.5, -0.9, -0.8, -0.7, -0.5, -0.2, 0.2, 0.6, 1.2, 1.9, 2.8.$$

(a) What test would you use to test this hypothesis? What are the assumptions of this test?

(b) Perform a one-tailed test of the kind you named in part (a), at level 0.05.

(c) A doctor, who does not know statistics, wishes to know what can be concluded from the study. Write a brief non-technical explanation for her.

105. (F2014) A randomized experiment is carried out to test two diets for patients with diverticulosis. Diet A is given to 15 patients, while Diet B is given to 12 patients. The outcome studied is the transit time through the alimentary canal. The patients on Diet A had a mean transit time of 68.4 hours, with standard deviation 16.47 hours. The patients on Diet B had a mean transit time of 83.42 hours, with standard deviation 17.63 hours. Normal probability plots (qqnorm) of each sample give approximately straight lines.

    (a) Which type of test would you use to test for a difference between the two diets? Explain your choice.

    (b) Calculate the test statistic for the test you chose in part (a), and give R code to produce the $P$-value (significance probability) for a two-tailed test.

    (c) Calculate an approximate 95% confidence interval for the difference in mean transit time between Treatment A and Treatment B. You may use the fact that the R command `qt(0.975, df)`, where df is the correct number of degrees of freedom, gives the output 2.07.

106. (F2014) Economists at the Federal Reserve Bank of Atlanta performed a survey on the incomes of nonsmokers and smokers. They looked at the hourly wages of 96,994 nonsmokers and 24,340 smokers. The nonsmokers had an average hourly wage of \$16.26, with a standard deviation of \$15.26. The smokers had an average hourly wage of \$13.10, with a standard deviation of \$20.75.

    (a) The distribution of hourly wages is not normal. However, we can use techniques based on the normal distribution to calculate confidence intervals and perform hypothesis tests. Why is this?

    (b) Find an approximate 95% confidence interval for the difference in average hourly wages between nonsmokers and smokers. (If you cannot compute the interval by hand, give R code.)

    (c) A news article on the research had the headline “If you want to earn more money, quit smoking.” Does the data above show that smoking causes lower wages? If not, what else could explain the difference?

107. (S2014) The heights of men and women are approximately normally distributed. Suppose we wish to estimate the average difference in heights between men and women attending a certain university. A random sample of seven men has average height 68.5 inches with standard deviation 3.0 inches, while a random sample of seven women has average height 65.5 inches with standard deviation 2.5 inches.

    (a) Explain why we would use the $t$-distribution and not the standard normal distribution to calculate $P$-values and confidence intervals in this case.

    (b) According to Welch’s approximation, the number of degrees of freedom is 11.6. Using the fact that the R code `qt(.975,df=11.62)` gives the value 2.187, find an approximate 95% confidence interval for the average difference in heights between men and women at the university.

(c) The confidence interval you calculated in part (b) contains the value 0. Should you conclude that there is no difference between the average height of men and the average height of women at the university? Explain why or why not.

## Ch 13

108. (S2016) For adult women in the U.S., height (in centimeters) and $\log_e$ of weight (in kilograms) follow close to a linear relationship. The regression line to predict log weight from height is

$$\log \text{Weight} = 2.5546 + 0.0107 \times \text{Height}$$

The data approximately satisfies the assumptions of the simple linear regression model.

(a) What is the regression prediction for the log weight of a woman who is 160 cm tall?

(b) What is the regression prediction for the weight (in kilograms) of a woman who is 170 cm tall?

(c) A student interprets the slope of the regression line as follows: "If you grow one centimeter, then your log weight will go up by 0.0107." Is this a correct interpretation? If not, give a correct interpretation.

109. (F2015) The International Rice Research Institute is breeding new lines of rice that are resistant to insects. One project involved breeding 374 lines of rice. The null hypothesis was that 25% would be resistant to insects, 50% would be mixed, and 25% would be susceptible to insects.

(a) If the null hypothesis is correct, how many of the 374 lines would we expect to be resistant? How many would be expect to be mixed? How many would we expect to be susceptible?

When the experiment was performed, they found 97 plants were resistant, 184 were mixed, and 93 plants were susceptible.

(b) Calculate the chi-squared statistic (either $G^2$ or $X^2$) and give R code to find the $P$-value.

(c) The $P$-value turns out to be about 0.92. What do you conclude?

110. (Su2015) In a population of numbers that obeys the second-digit Benford's law, the probability $f(x)$ that the second digit is $x$ is given in Table 8 for $x$ from 0 to 9.

Populations of vote counts often, but do not always, follow the second-order Benford's law. There is controversy among some political scientists as to whether inconsistency with the second-order Benford's law is a reliable indicator of vote fraud. Table 8 also gives frequencies for the number of times the second digit $x$ was observed in counts of the number of votes in the 2012 U.S. Presidential Election in 4586 county-level regions.

(a) Suppose the 2012 vote count data comes from a distribution that follows the second-digit Benford's law. For each second digit from 0 to 9, how many of the 4586 regions would we expect to have that second digit?

(b) Calculate the chi-square statistic for the test of the hypothesis that the vote counts follow the second digit Benford's law.

(c) The $P$-value for the chi-square test turns out to be 0.947. What do you conclude?

| $x$ (Second digit) | $f(x)$ | Number of regions |
|---|---|---|
| 0 | 0.120 | 573 |
| 1 | 0.114 | 505 |
| 2 | 0.109 | 492 |
| 3 | 0.104 | 478 |
| 4 | 0.100 | 459 |
| 5 | 0.097 | 455 |
| 6 | 0.093 | 440 |
| 7 | 0.090 | 393 |
| 8 | 0.088 | 407 |
| 9 | 0.085 | 384 |

Table 5: Second digit Benford's law probabilities, along with frequencies for the number of times the second digit $x$ was observed in counts of the number of votes in the 2012 U.S. Presidential Election in 4586 county-level regions.

111. (Su2015) A random sample of 1,232 California men aged 35 to 44 was surveyed. Two categorical variables that were measured were:

- Marital status: married, formerly married (widowed/divorced/separated), never married.
- Employment status: employed, unemployed, not in labor force.

The following table gives the results:

| | Married | Formerly married | Never married |
|---|---|---|---|
| Employed | 790 | 98 | 209 |
| Unemployed | 56 | 11 | 27 |
| Not in labor force | 21 | 7 | 13 |

Test the independence of marital status and employment status, giving a test statistic, a $P$-value, and a conclusion.

112. (S2015) Radlet (1981) studied effects of racial characteristics on whether individuals convicted of homicide received the death penalty. When the victim was white, 30 defendants received the death penalty, while 184 did not. When the victim was black, 6 defendants received the death penalty, while 106 did not. One way of studying this data would be to do a chi-square test of the null hypothesis that the variables "race of victim" (white or black) and "death penalty status" (yes or no) are independent.

(a) Suppose that race of victim and death penalty status are independent. Given the above data, how many out of 326 individuals convicted of homicide would you expect to fall in the following categories:

- White victim, death penalty

- White victim, no death penalty
- Black victim, death penalty
- Black victim, no death penalty

(b) Calculate the chi-squared statistic (either Pearson's or the likelihood ratio) for a chi-squared test of independence, and give R code for the $P$-value.

(c) The $P$-value turns out to be 0.018. What do you conclude about the relationship between the death penalty and the race of the victim? (Do not merely write "reject" or "do not reject.")

113. (S2015) In Alameda County, California, 42% of residents (for this question, counting only people over 21) are aged 21 to 40, 23% are aged 41 to 50, 16% are aged 51 to 60, and 19% are aged over 60.

(a) Suppose we select 66 Alameda County adults at random. What is the probability that at least half of the 66 are over 50? What is the probability that at least half of the 66 are over 60?

A panel of 66 jurors is chosen from among Alameda County residents.

(b) If the jurors were chosen at random from among the population of Alameda County residents, how many of the 66 jurors would we expect to fall in each of the four age groups?

In actuality, 5 jurors were aged 21 to 40, 9 were aged 41-50, 19 were aged 51-60, and 33 were aged over 60.

(c) Was the panel of jurors selected at random from the population of Alameda County residents? Test this hypothesis, giving a test statistic, a $P$-value, and a conclusion.

114. (F2014) In a population of numbers that obeys Benford's law, the probability that the first non-zero digit is $x$ is given by

$$f(x) = P(X = x) = \log_{10}(1 + 1/x)$$

for $x \in \{1, \dots, 9\}$. Table 6 gives the numerical values of $f(x)$ to 4 decimal places.

(a) Suppose the first digits of the number of followers of Twitter accounts obey the law (excluding accounts with no followers). If there are 38,663,000 Twitter accounts, how many would we expect to have each first digit from 1 to 9?

(b) Table 6 shows the first digit of the number of Twitter followers for 38,663,000 Twitter accounts. Calculate the chi-square statistic for the test of the hypothesis that the first digits of the number of Twitter followers follow Benford's law.

115. (F2014) 415 members of the House of Representatives voted on the Civil Rights Act in 1964: 289 for, 126 against. (For the purposes of this question, treat these votes as a sample from a larger hypothetical population.) We can break down the vote by party:

- Democrats: 153 for, 91 against

| $x$ (First digit) | $f(x)$ | Number of accounts (thousands) |
|---|---|---|
| 1 | 0.3010 | 12614 |
| 2 | 0.1761 | 6443 |
| 3 | 0.1249 | 4563 |
| 4 | 0.0969 | 3581 |
| 5 | 0.0792 | 2951 |
| 6 | 0.0669 | 2533 |
| 7 | 0.0580 | 2227 |
| 8 | 0.0512 | 1988 |
| 9 | 0.0458 | 1763 |

Table 6: Distribution of the first digit of the number of Twitter followers for 38,663,000 Twitter accounts.

- Republicans: 136 for, 35 against

We can also break down the vote by geography (north/south):

- Northern: 281 for, 32 against
- Southern: 8 for, 94 against

(a) Perform a chi-squared test of independence of vote and party. State which kind of chi-squared test you are using, and give the test statistic, the $P$-value, and a conclusion.

(b) Perform a chi-squared test of independence of vote and geography (north/south). State which kind of chi-squared test you are using, and give the test statistic, the $P$-value, and a conclusion.

(c) Based on this data, a political commentator claims that out of northern Democrats, northern Republicans, southern Democrats, and southern Republicans, the group most likely to vote against the Civil Rights Act was southern Democrats. Does this follow from the data? Explain why or why not.

116. (S2013) 1749 graduate students are admitted to the six largest departments of a university (Table 7).

(a) Suppose department and gender are independent. If there are 1749 admissions, how many men and how many women would you expected to be admitted to each of the six departments? Is the sample size large enough to perform a chi-square test of independence?

(b) Perform a chi-squared test of independence of department and gender. State which kind of chi-squared test you are using, and give the test statistic and $P$-value.

(c) Clearly explain the meaning of your result in part (b) to a university employee who does not know any statistics.

| Department | Men admitted | Women admitted |
|---|---|---|
| A | 512 | 89 |
| B | 353 | 17 |
| C | 120 | 202 |
| D | 138 | 131 |
| E | 53 | 94 |
| F | 16 | 24 |

Table 7: Number of men and women admitted to the six largest departments at a large university.

**Ch 15**

117. (S2017) A random sample of 60 students at a large elementary school is selected to take a standardized math test. Data on height (in inches) and math scores (on a 0 to 100 scale) is collected, and appears approximately bivariate normal. The results are summarized below:

- Heights of sample: $\bar{x} = 48, s_X = 3.75$
- Math scores of sample: $\bar{y} = 50, s_Y = 15$
- Sample correlation: $r = 0.6$

(a) Find the slope and intercept of the regression line for predicting math scores from height.

(b) Find an approximate 95% confidence interval for $\beta_1$, the slope of the regression line. Use the fact that the R code `qt(.975, df=58)` gives the number 2.00.

(c) The confidence interval you calculated in (b) does not contain zero, so there's strong evidence that math scores typically increase with height. Does this mean a quick period of growth in height will cause a student to get better at math? Explain.

118. (F2016) A 1903 study found that the heights of husbands had mean 68 inches with SD 2.7 inches. The heights of their wives had mean 63 inches and SD 2.5 inches. The correlation of the heights of husbands and wives was 0.25. The data was approximately bivariate normal.

(a) Find the slope and intercept of the regression line to predict the height of a husband in 1903 from the height of his wife.

(b) Find the slope and intercept of the regression line to predict the height of a wife in 1903 from the height of her husband.

(c) Suppose a randomly selected husband is 66 inches tall. Write R code to find the probability that his wife is taller than him.

119. (Su2016) Two variables are measured on a random sample of about 3000 U.S. adults:

- Hours of TV watched per day ("TV") — mean 2.88 hours, SD 1.69 hours
- Body mass index ("BMI"): weight (in kilograms) divided by height (in meters) squared — mean 28.67, SD 6.70

We use BMI instead of weight because this better fits the assumptions of the simple linear regression model (although it is not bivariate normal.) The relationship between the two variables is approximately linear. The correlation between them is 0.1381.

(a) Find the regression line to predict BMI from hours of TV watched. Write your answer as an equation.

(b) According to your model, what is the prediction for the BMI of a person who watches two hours of TV a day?

(c) According to your model, what is the prediction for the *weight* of a person of height 1.6 meters who watches four hours of TV a day?

120. (F2015) In a National Football League (NFL) regular season, each team plays 16 games. Let "team wins" be the number of regular season wins by a team in a particular season (taking a tie as half a win.) Since on average teams win half their games, the distribution of team wins has mean 8. Assume the distribution of team wins stays about the same from year to year.

There is a positive correlation between a team's wins one year and their wins the next ($r = 0.327$.) Because of this, we can use regression to predict a team's win one year by using their wins the previous year.

(a) Find the regression line to predict a team's wins one year from their wins the previous year. (Hint: You do not need to know the standard deviations, but if you cannot work out how to do the problem without standard deviations, make a reasonable guess.)

(b) In 2013, Houston had 2 wins, while in 2014 they had 9 wins. In 2013, Dallas had 8 wins, while in 2014 they had 12 wins. Use regression to predict 2014 wins for Houston and Dallas based on their 2013 wins. Which team exceeded their prediction by a larger margin?

(c) A cable sports analyst who does not know statistics suggests a different prediction system — simply predict a team will win as many games one year as they did the previous year. Explain convincingly to the analyst why in the long run, this prediction system will not be as accurate as a regression line.

121. (Su2015) In a study of 2.5 million students who took the SAT in the 1990s, researchers from the University of Minnesota (Sackett et al., 2009) created a numeric index of socioeconomic status (SES) with mean zero and SD 1. For the students in the study, the average high school GPA was 3.21, with a standard deviation of 0.66.[5] Together, SES and high school GPA had an approximately bivariate normal distribution, with correlation 0.4.

(a) Suppose that a rich student has an SES of +1. Use regression to predict their high school GPA.

(b) Suppose that a poor student has an SES of −1. Use regression to predict their high school GPA.

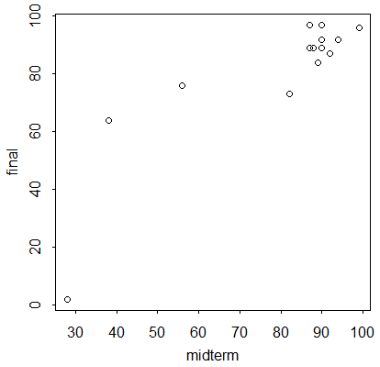
[5] Remember, high school GPA may be above 4.0.

(c) Suppose a student had a high school GPA of 4.0. Write R code to estimate the probability the student is of above-average SES.

122. (S2014) The following table gives midterm and final exam scores for a sample of fifteen statistics students.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Midterm | 87 | 90 | 28 | 82 | 90 | 38 | 56 | 92 | 90 | 88 | 87 | 90 | 99 | 89 | 94 |
| Final | 89 | 92 | 2 | 73 | 97 | 64 | 76 | 87 | 92 | 89 | 97 | 89 | 96 | 84 | 92 |

(a) A scatter plot of the data is below. Does the data seem to come from a bivariate normal distribution? Explain why or why not.



(b) (Requires sign test.) Find a 99.26% sign test confidence interval for the median *improvement* in score in the final relative to the Use the fact that that if $X$ is a binomial random variable with $n = 15$ and $p = 0.5$, then $P(X \leq 2) = 0.0037$.

(c) The confidence interval you found in (b) contains zero. A professor concludes that none of the students did any better or worse on the final compared to the Is this a reasonable conclusion to draw? Explain.

123. (F2014) A Major League Baseball team plays 162 games each season. There are 30 teams. Each season, the number of wins by Major League Baseball teams has an approximately normal distribution with mean 81 and standard deviation 11.7. The correlation between a team's wins one season and their wins the next season is 0.54.

(a) Suppose a baseball executive believed the best prediction of a team's wins in 2015 should be equal to their wins in 2014. For example, he predicts that the Los Angeles Angels, who had the most wins in 2014 with 98, would have 98 wins in 2015. Using the data given, explain to the baseball executive (who knows very little statistics) why this particular prediction is likely too high.

(b) Use regression to predict the Los Angeles Angels' 2015 wins using only the above data.

(c) The executive looks at the regression predictions for all MLB players and sees that no team is predicted to win more than 91 games. The executive suspects the predictions are too low, because in every full season since 1961, at least one team has won at least 96 games. Explain to the executive, who knows very little statistics, why his suspicions are misplaced.

124. (S2013) A class of 50 engineering students took two midterm tests. Gabby missed the first test and Kohei missed the second test. The 48 students who took both tests scored an average of 65 points on Test One, with a standard deviation of 15 points, and an average of 68 points on Test Two, with a standard deviation of 15 points. The scatter diagram of their scores is roughly ellipsoidal, with a correlation coefficient of $r = 0.6$.

Because Kohei and Gabby each missed one of the tests, the professor needs to guess how each would have performed on the missing test in order to compute their semester grades.

(a) Kohei scored 85 points on Test One. He suggests that his missing score on Test Two be replaced with his score on Test One, 85 points, arguing that, if anything, this is unfair to him, since on average students did better on Test Two than Test One. Explain to Kohei, who knows very little statistics, why we might predict his Test Two score to be less than 85.

(b) Show that the regression prediction for Kohei's Test Two score is 80 points.

(c) Gabby scored 80 points on Test Two. She argues that since Kohei scored 85 on Test One and was predicted to score 80 on Test Two, she would have scored 85 points on Test One. Explain to Gabby, who knows very little statistics, why this is not a good prediction, and give a better prediction.

## Answers

1. To be completed

2. No, the data does not prove this. The study is not randomized, so there may be important differences between the women who accepted screening and the women who refused besides the screening itself. For example, it could be that the women who refused screening were younger, and younger women are less likely to get breast cancer. Thus the results can be explained even if the screening has no effect.

3. (a) $P(B|A^c) = 0.53$

   (b) $P(B) = P(A \cap B) + P(A^c \cap B) = P(A)P(B|A) + P(A^c)P(B|A^c) = \frac{33800}{40000}(0.7) + \frac{6200}{40000}(0.53) = 0.674$

   (c) $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.592}{0.674} = 0.878$

4. (a) Using the letters given, let's use $EA$ to represent the couple {Eliza,Alvin}, for example. $3 \times 2 \times 1 = 6$ possible arrangements of 3 couples. Out of those, three make Eliza happy (two containing $EA$, and the other one being $\{EC, FB, DA\}$). Hence $P(\{\text{Eliza happy}\}) = 3/6 = 0.5$

   (b) Let $X = \{\text{Number of lessons Eliza is happy}\}$, so $X \sim \mathit{Binomial}(7, 0.5)$.
   $P(H) = P(X = 4) =$ `dbinom(4, 7, .5) = 0.273`.
   $P(J) = P(X \geq 4) = 1 -$ `pbinom(3, 7, .5)` $= 0.5$.
   $P(H^c \cup J) = P(X \neq 4 \text{ or } X \geq 4) = P(X \in X(S)) = P(S) = 1$

   (c) The probability is higher than before. Observe that if Eliza picks last it's equivalent to the previous game (other students are indifferent, making the choice random). If Eliza picks first or second, she can improve her probability of being happy (note that if she picks first she is guaranteed of being happy by always picking Alvin).

5. (a) If $X = \{\text{Number of heads}\}$, then $X \sim \mathit{Binomial}(6, .5)$.
   $P(X = 4) =$ `dbinom(4,6,.5)` $= 0.234$

   (b) $P(X > 3) = 1 - P(X \leq 3) = 1 -$ `pbinom(3,6,.5)` $= 0.344$

   (c) $P(X = 4|X \geq 4) = \frac{P(X=4 \text{ and } X\geq4)}{P(X\geq4)} = \frac{P(X=4)}{P(X\geq4)} = \frac{0.234}{0.344} = 0.682$

6. (a) $X = \{\text{Number of dice showing 1}\}$, so $X \sim \mathit{Binomial}(4, 1/6)$.
   $P(X = 4) = (1/6)^4 = 0.00077$

   (b) $P(X \geq 2) = 1 - P(X \leq 1) = 1 -$ `pbinom(1, 4, 1/6)` $= 0.868$

   (c) With exactly two dice showing ones, there is only one case, e.g., {1,1,2,2} (and 6 possible arrangements). With exactly three dices showing ones, there are two cases, {1,1,1,2} and {1,1,1,3} (with 4 possible arrangements for each case), and finally one case with all four dice showing ones. Hence, $6 + 4 + 4 + 1 = 15$, and the probability is $15/6^4 = 0.0112$

7. (a) Two of the five Rangers are selected, so the probability is 2/5.

   (b) The probability both selected Rangers are male is $3/5 \times 2/4 = 3/10$. So the probability at least one Ranger is female is 7/10.

(c)

$$\begin{aligned}
P(\text{Pink selected}|\text{at least one female}) &= \frac{P(\text{Pink selected} \cap \text{at least one female})}{P(\text{at least one female})} \\
&= \frac{P(\text{Pink selected}}{P(\text{at least one female})} \\
&= \frac{2/5}{7/10} \\
&= \frac{4}{7}.
\end{aligned}$$

(d)

$$\begin{aligned}
P(\text{Pink selected} \cap \text{at least one female}) &= 2/5 \\
P(\text{Pink selected}) \times P(\text{at least one female}) &= 2/5 \times 7/10 = 7/25.
\end{aligned}$$

These two numbers are not equal, so the events are not independent. (Intuitively, of course knowing that the Pink Ranger was selected changed your probability that at least one female is selected, since that latter probability is now 1.)

8. (a) $0.00428 \times 0.826 = 0.003535$ or about 0.35%.
   (b) In addition to the 0.003535 who have cancer and test positive, $0.99572 \times 0.096 = 0.09559$ don't have cancer and test positive. That's a total of 0.09912, or about 9.9%.
   (c) This is $0.003535/0.09912 = 0.03567$, about 3.6%.
   (d) The probability that none have cancer is $(1 - 0.03567)^{10} = 0.6955$. So the probability that at least one has breast cancer is one minus this, or about 30%.

9. (a) $1 - 0.84^3 = 0.4073$.
   (b) Once you've chosen a young adult and a senior, it doesn't matter what their original selection probabilities were: it's like selecting one young adult at random from all young adults, and then selecting one senior at random from all seniors. So this is just $0.04 \times 0.58 = 0.0232$.
   (c) $P(\text{young adult does but not senior}) = 0.96 \times 0.42 = 0.4032$. P(senior does | one does) $= 0.0232/(0.0232 + 0.4032) = 0.0232/0.4264 = 0.0544$.

10. (a) By counting or the binomial, this is $\binom{6}{4}/2^6 = 15/64$.
   (b) Let $X$ be the number of heads. We need to find $P(X = 4) + P(X = 5) + P(X = 6)$. This is
   $$\frac{\binom{6}{4} + \binom{6}{5} + \binom{6}{6}}{2^6} = \frac{22}{64}.$$
   (c) From the definition of conditional probability, this is the answer to (a) divided by the answer to (b), which is 15/22.

11. (a) $(1/6)^4 = 1/1296$.

(b) $1 - P(\text{none do}) - P(\text{one does}) = 1 - (5/6)^4 - 4 \times (1/6) \times (5/6)^3 = 171/1296$ or 13.2%.

(c) There's one way to get 4 (1111), four ways to get 5 (four 1112s in some order), and ten ways to get 6 (four 1113s, six 1122s), giving a total probability of 15/1296.

12.

$$\begin{aligned} P(\text{cancer and positive}) &= .01 \times .96 = 0.0096 \\ P(\text{no cancer and positive}) &= .99 \times .06 = 0.0594 \\ P(\text{cancer}|\text{positive}) &= \frac{.0096}{.0096 + .0594} = 0.139. \end{aligned}$$

13. (a) $(0.6 \times 0.95) + (0.39 \times 0.8) = 0.882$

(b) $(0.65 \times 0.95) + (0.31 \times 0.8) = 0.818$

(c) Run, because it has a higher win probability.

**Ch4**

14. To be completed

15. To be completed

16. To be completed

17. (a) Using the corresponding $x$ values, observe that we need $3m(1) + 2m(2) + m(3) = 1$, so $10m = 1$ and $m = .1$. Hence the pmf is:

$$f(x) = P(X = x) = \begin{cases} 0.3 & x = 1 \\ 0.4 & x = 2 \\ 0.3 & x = 3 \\ 0 & \text{otherwise.} \end{cases}$$

The graph for the CDF is:

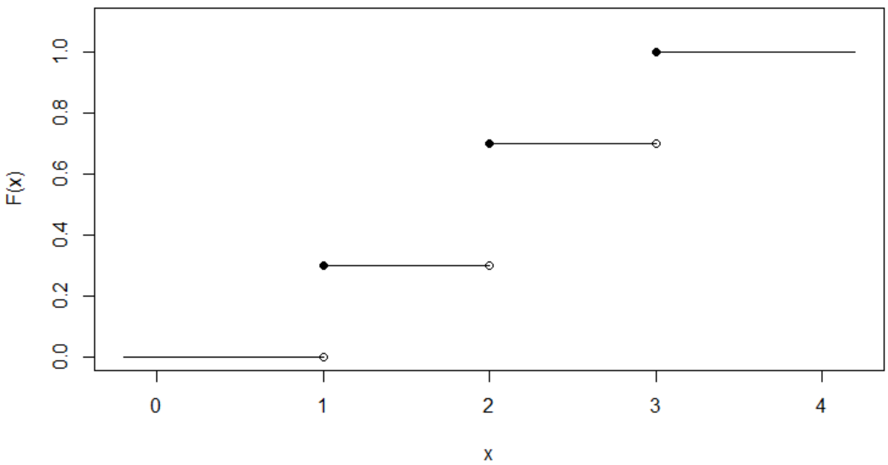
$$F(x) = P(X \le x) = \begin{cases} 0 & x < 1 \\ .3 & 1 \le x < 2 \\ .7 & 2 \le x < 3 \\ 1 & x \ge 3 \end{cases}$$

(c) $EX = 1(.3) + 2(.4) + 3(.3) = 2$ and
$Var(X) = (1-2)^2(.3) + (2-2)^2(.4) + (3-2)^2(.3) = .6$

18. (a) $3k + 4k + 6k + 7k = 20k = 1$ hence $k = 1/20$

(b) $EX = 3(3k) + 4(4k) + 6(6k) + 7(7k) = (3^2 + 4^2 + 6^2 + 7^2)k = 110k = 5.5$ and
$E(k^2X + k) = k^2EX + k = (1/20)^2(5.5) + (1/20) = 0.064$

(c) (You should not expect questions like this in the exam, too theoretical) Let $A_j = \{s \in S : X(s) = j\}$ for $j = 3, 4, 6, 7$, e.g., $A_4$ is the event of all the outcomes such that the random variable assigns the number 4. By construction, for any $s \in S$ with $X(s) = j$, $s \in A_j$, and if $j \ne k$ then $s \notin A_k$. It follows that $A_j$ and $A_k$ are disjoint when $j \ne k$ and $S = \cup_{j \in X(S)} A_j$, and the result follows.



(b)

19. (a) Let $X$ be the number of fives or sixes in 315,672 rolls. If the dice were perfectly fair and each die roll was independent of every other die roll, then $X$ is a Binomial$(315672, 1/3)$ random variable. The probability that $X$ is at least 106,602 is

```
> 1 - pbinom(106601, 315672, 1/3)
[1] 1.020838e-07
```

or about 1 in 10 million.

(b) Add together $P(X \leq 103846) + P(X \geq 106602)$:

```
> pbinom(103846, 315672, 1/3) + (1 - pbinom(106601, 315672, 1/3))
[1] 1.983372e-07
```

or about 2 in 10 million.

(c) Either the dice weren't perfectly fair, or the die rolls weren't independent, or a 1 in 5 million event occurred. Out of these, it's easiest to believe that the dice were collectively very slightly biased toward 5 and/or 6.

20. (a) $C(7,5) \cdot 0.5^5 \cdot 0.5^2 = 21/128$ or 16.4%.

(b) `pbinom(2, 7, 0.5)` or $[C(7,0) + C(7,1) + C(7,2)]/2^7$, which is 29/128 or 22.7%.

(c) $EX = np = 3.5$, $\text{Var}(X) = np(1-p) = 1.75$, $\text{SD}(X) = \sqrt{1.75} = 1.32$.

(d) They're not independent. If $A$ is all heads on the first four tosses and $B$ is all heads on the last four tosses, then $P(A) > 0$ and $P(B) > 0$ but $P(A \cap B) = 0$ since they can't both happen on the same set of seven tosses. So $P(A) \cdot P(B) \neq P(A \cap B)$, implying dependence.

21. (a) $4k + 3k + 2k + k = 1$, so $10k = 1$, so $k = 0.1$.

(b)

$$F(y) = \begin{cases} 0 & y < 0 \\ 0.4 & 0 \leq y < 1 \\ 0.7 & 1 \leq y < 2 \\ 0.9 & 2 \leq y < 3 \\ 1 & y \geq 3 \end{cases}$$

(c) $E(X) = 0.4 \times 0 + 0.3 \times 1 + 0.2 \times 2 + 0.1 \times 3 = 1$.

(d) $E(X^2) = 0.4 \times 0^2 + 0.3 \times 1^2 + 0.2 \times 2^2 + 0.1 \times 3^2 = 2$.
$\text{Var}(X) = E(X^2) - [EX]^2 = 2 - 1^2 = 1$.

22. (a) By a tree or the binomial,

$$\begin{aligned} P(X = 0) &= 0.52 \times 0.52 = 0.2704 \\ P(X = 1) &= 2 \times 0.48 \times 0.52 = 0.4992 \\ P(X = 2) &= 0.48 \times 0.48 = 0.2304. \end{aligned}$$

(b)

$$F(y) = \begin{cases} 0 & y < 0 \\ 0.2704 & 0 \leq y < 1 \\ 0.7696 & 1 \leq y < 2 \\ 1 & y \geq 2. \end{cases}$$

(c) This is easiest to find using the properties of the binomial: $EX = np = 0.96$ and $\text{Var}(X) = np(1 - p) = 0.4992$.

23. (a) `1 - pbinom(14, 20, 0.5)` = 0.0201 or about 2%.

(b) `1 - pbinom(0, 80, p)`, where `p` is the answer to part (a). This is 0.812 or about 81%, i.e., quite likely.

24. (a)

$$F(y) = P(X \leq y) = \begin{cases} 0 & y < 3 \\ 0.2 & 3 \leq y < 3.3 \\ 0.55 & 3.3 \leq t < 3.7 \\ 0.85 & 3.7 \leq y < 4 \\ 1 & y \geq 4 \end{cases}$$

(b) $EX = 0.15 \times 4 + 0.3 \times 3.7 + 0.35 \times 3.3 + 0.2 \times 3 = 3.465$.

(c) $EX^2 = 0.15 \times 4^2 + 0.3 \times 3.7^2 + 0.35 \times 3.3^2 + 0.2 \times 3^2 = 12.1185$.
$\text{Var}(X) = 12.1185 - 3.465^2 \approx 0.112$.

25. (a) By the multiplication rule, $0.39 \times 0.39 = 0.1521$, or about 15%.

(b) Either the first supports Billary and the second supports Ronald, or the first supports Ronald and the second supports Billary. The total probability is
$(0.46 \times 0.39) + (0.39 \times 0.46) = 0.3588$ or about 36%.

(c) The number in the sample who support Billary is (approximately) binomial with $n = 1000$ and $p = 0.46$. The probability this random variable is at least 480 can be found in R as `1- pbinom(479, 1000, 0.46)`, which gives 0.1081 or about 11%.

26. (a)

$$F(y) = P(X \leq y) = \begin{cases} 0 & y < 3 \\ 0.1 & 3 \leq y < 4 \\ 0.5 & 4 \leq y < 5 \\ 0.8 & 5 \leq y < 6 \\ 1 & y \geq 6 \end{cases}$$

(b) $0.1 \times 3 + 0.4 \times 4 + 0.3 \times 5 + 0.2 \times 6 = 0.3 + 1.6 + 1.5 + 1.2 = 4.6$.

(c) $EX^2 = 0.1 \times 3^2 + 0.4 \times 4^2 + 0.3 \times 5^2 + 0.2 \times 6^2 = 0.9 + 6.4 + 7.5 + 7.2 = 22$.
$\text{Var}(X) = 22 - 4.6^2 = 0.84$.

27. Let $X$ be the gain (in dollars) by Hillary. Then
$EX = (0.54 \times 1) + (0.16 \times -3) + (0.3 \times 0) = 0.06$. Since this is positive, the bet favors Hillary.

Note: LOL.

28. (a) `1 - pbinom(156, 329, 0.485)` = 63.2%.

(b) There are many possible reasons for this. For example, it may be that children of the same parent are not independent in sex. If Julia Roberts' first child is a boy, perhaps that changes our probability that her second child will also be a boy. Or there could be identical twins, etc., in the data set. (Such departures from independence are likely to be small, however.)

29. (a) $1k + 2k + 3k + 4k = 1$, so $k = 10$.

(b)

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{10} & 1 \leq x < 2 \\ \frac{3}{10} & 2 \leq x < 3 \\ \frac{6}{10} & 3 \leq x < 4 \\ 1 & x \geq 4 \end{cases}$$

(c)

$$\begin{aligned} EX &= (1 \times 1/10) + (2 \times 2/10) + (3 \times 3/10) + (4 \times 4/10) = 3 \\ EX^2 &= (1^2 \times 1/10) + (2^2 \times 2/10) + (3^2 \times 3/10) + (4^2 \times 4/10) = 10 \\ \text{Var}\, X &= 10 - 3^2 = 1 \end{aligned}$$

30.
```
> prob.pass = 1 - pbinom(7, 10, 0.5)
> 1 - pbinom(0, 10, prob.pass)
[1] 0.4301586
```

There's a 43% chance at least one of the chimps passes.

31. (a) $X$ is Binomial$(200, 1/38)$, so $EX = np = 200/38 = 5.26$, $\text{Var}(X) = 1850/361$, SD $= 2.26$.

(b) 1 - pbinom(5, 200, 1/38) = 43.1%.

(c) 1 - pbinom(1388, 50000, 1/38) = 2.18%.

32. (a)

$$F(y) = \begin{cases} 0 & y < -3 \\ 0.1 & -3 \leq y < -2 \\ 0.3 & -2 \leq y < 0 \\ 0.6 & 0 \leq y < 1 \\ 0.9 & 1 \leq y < 3 \\ 1 & y \geq 3 \end{cases}$$

(b) $(0.1 \times -3) + (0.2 \times -2) + (0.3 \times 0) + (0.3 \times 1) + (0.1 \times 3) = -0.1$

(c) $EX^2 = (0.1 \times (-3)^2) + (0.2 \times (-2)^2) + (0.3 \times 0^2) + (0.3 \times 1^2) + (0.1 \times 3^2) = 2.89$
$\text{Var } X = 2.9 - 0.1^2 = 2.89$

33. (a) Since $X$ is discrete, the CDF is a step function, with the height found by adding the probabilities of the $x$-values up to and including the $y$-value in question. This gives:

$$F(y) = \begin{cases} 0 & y < 1 \\ 0.1 & 1 \leq y < 2 \\ 0.3 & 2 \leq y < 3 \\ 0.5 & 3 \leq y < 4 \\ 0.8 & 4 \leq y < 5 \\ 1 & y \geq 5 \end{cases}$$

(b) $EX = .1 + .4 + .6 + 1.2 + 1 = 3.3$
$EX^2 = .1 + .8 + 1.8 + 4.8 + 5 = 12.5$
$\text{Var } X = 12.5 - 3.3^2 = 1.61$

(c)

$$\begin{aligned} P(X_1 + X_2 = 4) &= P(X_1 = 1, X_2 = 3) + P(X_1 = 2, X_2 = 2) + P(X_1 = 3, X_3 = 1) \\ &= (.1 \times .2) + (.2 \times .2) + (.2 \times .1) \\ &= .08 \end{aligned}$$

34. (a) $k = 1/(1 + 2 + 3 + 4 + 5) = 1/15$.

(b)

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{15} & 1 \leq x < 2 \\ \frac{3}{15} & 2 \leq x < 3 \\ \frac{6}{15} & 3 \leq x < 4 \\ \frac{10}{15} & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

(c) $EX = 1/15 + 4/15 + 9/15 + 16/15 + 25/15 = 11/3$.
$EX^2 = 1/15 + 8/15 + 27/15 + 81/15 + 125/15 = 15$. $\mathrm{Var}\, X = 15 - (11/3)^2 = 14/9$.

**Ch 5**

35. (a) (This question correspond to chapter 6) `qnorm(.8, 21.5, sqrt(19.2))`=25.188

(b) `1-pbinom(29, 100, 0.2)`=0.0112

(c) $EY = 240 + 58.5EW = 240 + 58.5(21.5) = 1497.75$.
$VarY = 58.5^2 VarW = 58.5^2(29.2) = 99929.7$. $P(Y > 1900)$=
`1-pnorm(1900,1497.75,sqrt(99929.7))` = 0.102

36. (a) This is $F(135) - F(115)$:

```
> pnorm(135, 120, 20) - pnorm(115, 120, 20)
[1] 0.372079
```

or about 37%.

(b) Let $p_a$ be the answer to (a). The probability all five songs are in the best tempo range for dancing is $p_a^5$, or about 0.71%. It's pretty unlikely.

(c) Let $p_c$ be the probability that one randomly selected disco song has a BPM of less than or equal to 160. This is $F(160)$:

```
> pc = pnorm(160, 120, 20)
> pc
[1] 0.9772499
```

The probability that at least one of the ten songs has a BPM over 160 is one minus the probability that *zero* of the ten songs has a BPM over 160 (i.e. the probability that all ten songs have a BPM of less than or equal to 160):

```
> 1 - pc^10
[1] 0.205569
```

which is about 21.6%. There's a decent chance that at least one song will have a BPM over 160, which might lead to lawsuits.

37. (a)

$$F(y) = \begin{cases} 0 & y < 0 \\ 0.1y & 0 \le y < 8 \\ 0.05y + 0.4 & 8 \le y < 12 \\ 1 & y \ge 12. \end{cases}$$

(b) $F(8) = 0.8$, so the median $q_2$ is clearly in $0 < q_2 < 8$. Setting $0.1q_2 = 0.5$, we get a median of 5.

(c) Taking a weighted average, $EX = 4 \times 0.8 + 10 \times 0.2 = 5.2$.

38. (a) $F(y)$ is a continuous function, so $X$ is continuous. (Alternatively, $F(y)$ is differentiable at all points except $y = 1$ and $y = 4$, so a PDF exists.)

(b) $\mathrm{Var}(X) = 43/8 - 49/24^2 \approx 1.207$; $\sigma_X \approx 1.1$.

(c) $EY = EA - EB = EX - EX = 0$. $\mathrm{Var}(Y) = \mathrm{Var}(A) + \mathrm{Var}(B) \approx 2.41$; $\sigma_Y \approx 1.55$.

39. (a) $e^{-3}$ or `1 - pexp(3, rate=1)` gives 0.0498 or about 5%.

    (b) $(1 - e^{-3}) - (1 - e^{-1})$ or `pexp(3, rate=1) - pexp(1, rate=1)` gives 0.318 or about 32%.

    (c) $P(Y > 4) = P(X^2 > 4) = P(X > 2)$, since $X$ is non-negative. This is `1 - pexp(2, rate=1)` $= 0.135$ or 13.5%.

    (d) Let $Y$ be the number out of $(X_1, X_2, X_3, X_4, X_5)$ that are less than 3. Then $Y$ is a binomial random variable with $n = 5$ and $p =$ `pexp(3)`. The probability "no more than three" are less than 3 is $P(Y \leq 3) =$ `pbinom(3, 5, pexp(3))` $= 0.0224$ or 2.2%.

40. (a) This is $1 - F(1) = 1 - 0.8413 = 0.1587$, or about 16%.

    (b) This is the probability that $X_2$ is more than one standard deviation above average, so it's just $0.1587 \approx 16\%$ again.

    (c)

$$\begin{aligned} P(X_1 > 0, X_2 > 0) &= P(X_1 > 0) \times P(X_2 > 0) \\ &= 0.5 \times 0.1587 \\ &= 0.0793 \end{aligned}$$

    So it's about 8%.

    (d) $EY = 0 + (-5) = -5, \mathrm{Var}(X) = 1 + 25 = 26, \sigma = \sqrt{26} \approx 5.1$.

41. (a) We can calculate this as the difference between the CDF of the standard normal at $y = 2.5$ and the CDF at $y = -1.5$:

```
> pnorm(2.5) - pnorm(-1.5)
[1] 0.9269831
```

    This gives about 93%.

    (b) $|Y| < 1$ if and only if $|X| < 1$. This is easily calculated using `pnorm`:

```
> pnorm(1) - pnorm(-1)
[1] 0.6826895
```

    So the probability is about 68%.

    (c) $Y > 2$ if and only if $X > 2^{1/3}$. We find this again using `pnorm()`:

```
> 1 - pnorm(2^(1/3))
[1] 0.1038489
```

    The probability is about 10%.

42. (a) `pnorm(2) - pnorm(1)` $= 0.136$.

    (b) $P(1 < X^3 < 2) = P(1 < X < 2^{\frac{1}{3}}) =$ `pnorm(1.26) - pnorm(1)` $= 0.0548$.

    (c) $P(1 < |Y| < 2) = P(1 < Y < 2) + P(-2 < Y < -1) = 2 \times P(1 < Y < 2)$ (by symmetry) $= 0.110$.

43. (a) `1 - pnorm(665, 511, 117)` $= 0.0940$.

(b) The sum is a normal random variable with mean $511 + 511 = 1022$, variance $117^2 + 117^2 = 27378$, and SD $\sqrt{27378}$. The probability is `1 - pnorm(1330, 1022, sqrt(27378))` $= 0.0313$.

(c) "The highest of the ten scores is more than 665" means the same thing as "at least one of the ten scores is more than 665." The number of students who scored more than 665 is binomial with $n = 10$ and $p$ given by the answer in part (a). The probability can thus be found by

```
p = 1 - pnorm(665, 511, 117)
1 - pbinom(0, 10, p)
```

which gives a probability of 0.628.

44. (a) By geometry or calculus,

$$F(y) = \begin{cases} 0 & y < -5 \\ 0.15(y+5) = 0.15y + 0.75 & -5 \le y < 0 \\ 0.75 + 0.05y & 0 \le y < 5 \\ 1 & y \ge 5 \end{cases}$$

(b) From a picture, $P(|X| < 2.5) = P(|X| > 2.5)$, so both are 1/2.

(c) From a picture and thinking about center of balance, less than zero. If you want to calculate it, there are a couple of options. By calculus:

$$\begin{aligned} EX &= \int_{-5}^{5} x \cdot f(x)\,dx \\ &= \int_{-5}^{0} 0.15x\,dx + \int_{0}^{5} 0.05x\,dx \\ &= [0.075x^2]_{-5}^{0} + [0.025x^2]_0^5 \\ &= -1.875 + 0.625 \\ &= -1.25. \end{aligned}$$

Or by argument: It's a mixture of two uniforms. The negative rectangle has area 0.75 and the positive rectangle has area 0.25. The centers of the two rectangles are $-2.5$ and $+2.5$. The expected value is thus

$$0.75 \times -2.5 + 0.25 \times 2.5 = -1.875 + 0.625 = -1.25.$$

45. (a) By geometry or calculus,

$$F(y) = \begin{cases} 0 & y < -4 \\ \frac{(y+4)(y+8)}{24} & -4 \le y < -2 \\ 1/2 & -2 \le y < 2 \\ \frac{y^2+8}{24} & 2 \le y < 4 \\ 1 & y \ge 4. \end{cases}$$

(b) From a picture, $P(|X| < 3) = P(|X| > 3)$, so both are 1/2.

(c) From a picture and thinking about center of balance, greater than zero.

46. (a)
```
> pnorm(2) - pnorm(1)
[1] 0.1359051
```

(b)
```
> F2 = pnorm(sqrt(2)) - pnorm(-sqrt(2))
> F1 = pnorm(sqrt(1)) - pnorm(-sqrt(1))
> F2 - F1
[1] 0.1600113
```

47. (a)

$$F(y) = \begin{cases} 0 & y < -1 \\ \frac{1}{2} - \frac{1}{2}x^2 & -1 \leq y < 0 \\ \frac{1}{2} + \frac{1}{2}x^2 & 0 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$$

(b) By symmetry: 0.

(c) $X$ does, because more of its mass is far away from its center.

48. (a) `1 - pnorm(0, -5, 10)` $= 0.309$

(b) `pnorm(-1.5) + (1 - pnorm(1.5))` $= 0.134$

(c) `1 - pbinom(5, 10, 0.5)` $= 0.377$

49. (a)

$$F(y) = \begin{cases} 0 & y < 0 \\ 0.1y & 0 \leq y < 2 \\ 0.2y - 0.2 & 2 \leq y < 6 \\ 1 & y \geq 6 \end{cases}$$

(b) Setting $F(q_2) = 0.5$, we get a median of $q_2 = 3.5$.

(c) Taking the weighted average of the centers of the two blocks, this is $0.2 \times 1 + 0.8 \times 4 = 3.4$.

Note: Question is phrased this way because the median is not formally introduced until chapter 6.

50. (a) For $0 \leq x \leq 1$,

$$F(x) = \int_0^x \frac{1}{2}u\,du = \frac{1}{4}x^2.$$

For $1 \leq x \leq 2$,

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= P(X \leq 1) + P(1 < X \leq x) \\ &= F(1) + \int_1^x \frac{1}{2}\,du \\ &= \frac{1}{4} + \left(\frac{1}{2}x - \frac{1}{2}\right) \\ &= \frac{1}{2}x - \frac{1}{4}. \end{aligned}$$

For $2 \leq x \leq 3$,

$$F(x) = F(2) + \int_2^x \frac{1}{2}(3-x)\,du = -\frac{5}{4} + \frac{3x}{2} - \frac{1}{4}x^2.$$

This gives:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4}x^2 & 0 \leq x < 1 \\ \frac{1}{2}x - \frac{1}{4} & 1 \leq x < 2 \\ -\frac{5}{4} + \frac{3x}{2} - \frac{1}{4}x^2 & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

(b)

$$\begin{aligned} EX &= \int_0^1 \frac{1}{2}x^2\,dx + \int_1^2 \frac{1}{2}x\,dx + \int_2^3 \frac{1}{2}x(3-x)\,dx \\ &= \frac{1}{6}x^3\Big|_0^1 + \frac{1}{4}x^2\Big|_1^2 + \left(\frac{3x^2}{4} - \frac{x^3}{6}\right)\Big|_2^3 \\ &= \frac{1}{6} + 1 - \frac{1}{4} + \frac{9}{4} - \frac{5}{3} \\ &= \frac{3}{2}. \end{aligned}$$

(c)

$$\begin{aligned} EX^2 &= \int_0^1 \frac{1}{2}x^3\,dx + \int_1^2 \frac{1}{2}x^2\,dx + \int_2^3 \frac{1}{2}x^2(3-x)\,dx \\ &= \frac{1}{8}x^4\Big|_0^1 + \frac{1}{6}x^3\Big|_1^2 + \left(\frac{x^3}{2} - \frac{x^4}{8}\right)\Big|_2^3 \\ &= \frac{1}{8} + \frac{4}{3} - \frac{1}{6} + \frac{27}{8} - 2 \\ &= \frac{8}{3} \\ \operatorname{Var} X &= \frac{8}{3} - \left(\frac{3}{2}\right)^2 \\ &= \frac{5}{12} \\ \operatorname{sd}(X) &= \sqrt{\frac{5}{12}} \\ &\approx 0.645. \end{aligned}$$

## Ch 6

51. To be completed

52. (a)

$$F(y) = \begin{cases} 0 & y < 0 \\ \frac{1}{15}y & 0 \le y < 15 \\ 1 & y \ge 15 \end{cases}$$

(b) These are just a quarter, half, and three-quarters of 15: 3.75, 7.5, and 11.25 respectively.

(c) Two-thirds of the time a bus arrives between 11:45 and 11:55, in which case the expected waiting time is 5 minutes. The other third of the time, I have to wait until some time between 12:10 and 12:25, or between 25 and 40 minutes, for an expected value of 32.5 minutes. The overall expected value is thus $2/3 \times 5 + 1/3 \times 32.5 = 14$ minutes, 10 seconds.

53. (a) Lacking any other information, use a Uniform$(0, 24)$ distribution.

$$f(x) = \begin{cases} \frac{1}{24} & 0 \le x < 24 \\ 0 & \text{otherwise.} \end{cases}$$

(b)

$$F(y) = P(X \le y) = \begin{cases} 0 & y < 0 \\ \frac{y}{24} & 0 \le y < 24 \\ 1 & y \ge 24 \end{cases}$$

(c) Setting $F(q) = 0.6$, we get $q = 24 \times 0.6 = 14.4$, or 2:24 pm.

54. (a) By calculus or from the area of a triangle,

$$F(y) = P(X \le y) = \begin{cases} 0 & y < 0 \\ 0.01x^2 & 0 \le y < 10 \\ 1 & y \ge 10 \end{cases}$$

(b) $F(5) = 0.25$, so $P(X > 5) = 1 - 0.25 = 0.75$.

(c) The left skew means the expected value is below the median. (Calculation gives a median of $\sqrt{50} \approx 7.1$ and an expected value of $20/3 \approx 6.7$.)

55. (a) By calculus or from the area of a triangle,

$$F(y) = P(X \le y) = \begin{cases} 0 & y < 0 \\ 0.005x^2 + 0.05x & 0 \le y < 10 \\ 1 & y \ge 10 \end{cases}$$

(b) $F(5) = 0.375$, so $P(X > 5) = 1 - 0.375 = 0.625$.

(c) The left skew means the expected value is below the median. (Extremely tedious calculation gives a median of $\sqrt{125} - 5 \approx 6.2$ and an expected value of $35/6 \approx 5.8$.)

56. (a) $c = 1/5$.

(b)

$$F(y) = \begin{cases} 0 & y < 5 \\ (y-5)/5 & 5 \le y < 10 \\ 1 & y \ge 10. \end{cases}$$

(c) $8.75 - 6.25 = 2.5$.

57. (a) To find the median $m$, set $F(m) = 0.5$. This gives

$$\begin{aligned} 1 - \frac{1}{m^2} &= 0.5 \\ 0.5 &= \frac{1}{m^2} \\ m &= \sqrt{2}. \end{aligned}$$

(b) Find a general expression for the $\alpha$-quantile:

$$\begin{aligned} 1 - \frac{1}{q^2} &= \alpha \\ 1 - \alpha &= \frac{1}{q^2} \\ q &= \sqrt{\frac{1}{1-\alpha}}. \end{aligned}$$

Plugging in 0.25 and 0.75 for $\alpha$ gives $q_1 = \sqrt{4/3}$, $q_3 = 2$. The IQR is $2 - \sqrt{4/3} \approx 0.85$.

(c) The distribution is right-skewed, so $EX$ will be greater than the median. We can check this with calculus:

$$\begin{aligned} EX &= \int_1^\infty \frac{2}{x^2}\,dx \\ &= [-2/x]_1^\infty \\ &= 2. \end{aligned}$$

58. (a) `qnorm(.75)` $= 0.674$.

(b) The probability $X$ is between 1 and 2 is `pnorm(2) - pnorm(1)` $= 0.136$. The probability $Y$ is between 1 and 2 is double this, which is 0.272.

(c) `qnorm(0.975)` $= 1.96$. (We'll be seeing a lot of this number.)

59. (a) Call the median $m$. We need $P(X \le m) = 0.5$. From geometry, $P(X \le x) = (x-1)^2$, so $m = 1 + 1/\sqrt{2} \approx 1.71$.

(b) $(1.5 - 1)^2 = 0.25$.

(c) $0.25^{10} = 1/2^{20}$ or about one in a million.

60. (a) `pnorm(-1) + (1 - pnorm(1))` $= 31.7\%$.

(b) `qnorm(.95)`$^2 = 2.71$.

61. `1 - pbinom(1, 5, 1-pnorm(1))` gives 0.181.

**Ch9**

62. (a) Not close to normal. Most articles have 0 or 1 citations, yet the mean is 9, so the distribution is very right-skewed and thus non-normal.

    (b) $9.06 \pm 1.96 \times \sqrt{565/1000} = 7.6$ to $10.5$.

    (c) $\hat{p}$ is 0.46. Interval is $0.46 \pm 1.96\sqrt{.46 \times .54/1000} = 0.42$ to $0.49$, or 42% to 49%.

63. (a) 18755/45242 is 41.5%.

    (b) The interval is $0.415 \pm 1.96\sqrt{.415 \times .585/45242}$, or 41.0% to 41.9%.

    (c) Let $p$ be the proportion of Presidential election voters that voted for Trump.

$$\begin{aligned} H_0 &: p = 0.461 \\ H_1 &: p \neq 0.461 \end{aligned}$$

    (d) `2 * pbinom(18755, 45242, 0.461)`

    (e) The most plausible reason is that the survey was biased in some way. It could be that Trump voters were less likely to respond to the survey than others, or it could be the study was not a simple random sample, or it could be that some respondents did not reply accurately when asked whether they voted and who they voted for.

64. (a) The sample mean is $(0 \times 232326 + 1 \times 58529 + 2 \times 53908 + 3 \times 18770)/154443 = 1.4417$. To find the plug-in variance, first find the sample average of $x^2$: $(0^2 \times 232326 + 1^2 \times 58529 + 2^2 \times 53908 + 3^2 \times 18770)/154443 = 2.869$. Then the plug-in variance is $2.869 - 1.4417^2 = 0.7906$ and the plug-in SD is 0.8891. (The sample variance and SD are the same to four significant figures.)

    (b) The interval is $1.4417 \pm 1.96 \times 0.8891/\sqrt{154443}$, or 1.437 to 1.446.

    (c) Set $0.02 = 2 \times 1.96 \times 0.8891/\sqrt{n}$. Rearranging and solving for $n$ gives a required sample size of around 30,000 families with three or more children.

65. (a)

$$\begin{aligned} H_0 &: p = 30/365 \approx 0.0822 \\ H_1 &: p \neq 30/365 \end{aligned}$$

    (b) $\hat{p} = 327680/4010532 = 0.08170$. The CI is $\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which is $0.08170 \pm 1.96 \times 0.00014$, which is 0.0814 to 0.0820.

    (c) It's a little less than 30/365, but only a little.

66. (a)

$$110/123 \pm 1.96 \times \sqrt{\frac{110/123 \times 13/123}{123}}$$

gives a 95% confidence interval from 84.0% to 94.9%.

(b)

$$198/214 \pm 1.96 \times \sqrt{\frac{198/214 \times 16/214}{214}}$$

gives a 95% confidence interval from 89.0% to 96.0%.

(c) The $P$-value is not small, so the data is consistent with the null hypothesis that free throw success doesn't depend on mouthguard position. So the difference could just be luck. (It's impossible to prove it's just luck from the data, but it probably is.)

67. (a) We want the probability 760 or more wins out of 1442 if the games are 50-50. This is `1 - pbinom(759, 1442, 0.5)`, or $0.021 = 2.1\%$. Yes, there's evidence that the team that wins the coin toss has a higher chance of winning than the team that loses the toss.

(b) We want the probability of 1804 or more wins out of 3591 if the games are 50-50. This is `1 - pbinom(1803, 3591, 0.5)`, or $0.395 = 39.5\%$. This isn't small, so there's insufficient evidence that the team that wins the coin toss has a higher chance of winning than the team that loses the toss.

(c) Use $p = 0.5$ for the purpose of estimation. We set

$$2 \times 1.96 \times \sqrt{\frac{0.5 \times 0.5}{n}} = 0.05$$

Solving for $n$, we get

$$\begin{aligned} n &= \left(\frac{2 \times 1.96 \times 0.5}{0.05}\right)^2 \\ &\approx 1537. \end{aligned}$$

68. (a) $\hat{p} = 2.074/4.247 = 0.488 = 48.8\%$.

(b)

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p \neq 0.5 \end{aligned}$$

(c) `2 * pbinom(2074000, 4247000, 0.5)`

(d) $\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. This is 0.4879 to 0.4889, or 48.8% to 48.9%.

(e) The probability that a random newborn is female isn't 50% — it's actually a bit less than 49%.

69. (a) Not close to normal. If many household incomes are low (with a lower limit at zero) and a few are very high (with no upper limit), then the distribution will be right-skewed, and thus non-normal. (Note: since no data was provided, it was possible to get full credit for saying the distribution was approximately normal if you had a *very* good argument.)

(b) Approximately normal. 100 is a large enough sample for the Central Limit Theorem to kick in, even for a skewed distribution. The Central Limit Theorem says the distribution of the sample mean approaches the normal, though for any finite sample size the distribution won't be exactly normal (unless the population is normal, which we have already established it isn't).

(c) No. The confidence interval is for the mean: it says we are 95% confident that the interval encloses the population mean[6]. It does not directly say anything about other aspects of the population distribution. The first sentence of the question as well as common sense imply a large proportion of households have an annual income well below $65,000, so the proportion of households that have annual income in the narrow range $65,000 to $95,000 will be much less than 95%.

70. (a) The sample proportion is $\hat{p} = 1844/3600 \approx 0.5122$. We estimate the standard error of this proportion as

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.0083$$

Using 1.96 standard errors, the 95% confidence interval is 49.6% to 52.9%.

(b) The sample proportion is $\hat{p} = 828/1560 \approx 0.5308$. We estimate the standard error of this proportion as

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.013$$

Using 1.96 standard errors, the 95% confidence interval is 50.6% to 55.6%.

(c) The width of the interval is $2 \times 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.02$. Solving for $n$ gives

$$n = \frac{1.96^2\hat{p}(1-\hat{p})}{0.01^2}$$

We don't know what $\hat{p}$ will be in the new study, but if we guess that it will be the same as the previous study, this gives $n = 9568$. If we think that quantum erotophysics does not exist, we might instead guess $\hat{p} = 0.5$, which would give $n = 9604$. In either case, a sample of size 10,000 should suffice. (This sounds like a lot, but we can get e.g. 50 students to do 200 pictures each.)

71. (a)

```
1 - pbinom(157, 329, 0.485) # One-tailed
2 * pbinom(157, 329, 0.485) # Two-tailed
```

(b) The sample proportion is 0.4772. Its standard error is

$$se = \sqrt{\frac{.4772 \times .5228}{329}} = 0.0275.$$

A 95% normal confidence interval is $.4772 \pm 1.96 \times .0275$, giving 42.3% to 53.1%.

[6] In frequentist statistics, this does not mean there is a 95% chance that the population mean is in the interval $65,000 to $95,000. The population mean is not random, so it's either in the interval or it isn't. The randomness is in the interval: you wouldn't get the same interval if you repeated the sampling. Approximately 95% of intervals created this way will contain the population mean.

(c) 47.7% of the sample were girls. This is less than 48.5%, so there's definitely no indication in this data that beautiful parents are more likely to have daughters than the general population. (The interval is wide enough that it's not impossible that they do have more daughters, but even if this is so, the effect isn't huge.)

72. (a) There's one independent sample. The twins are sampled in pairs, not independently of each other.

(b) An approximate 95% confidence interval for the mean difference is $1.917 \pm 2.201 \times 7.153/\sqrt{12}$, or $-2.63$ to 6.46.

(c) The $P$-value is not small, so the data is consistent with the null hypothesis of no average difference in aggression between first born and second born twins. However, the sample size is small, meaning the power of the test is low — regardless of whether there was or wasn't a difference, we were unlikely get a small $P$-value, so we certainly can't be sure that there's absolutely no difference between the aggression of first born and second born twins. The confidence interval suggests it's plausible that first born twins could be six points more aggressive on average, or that second born twins could be two points more aggressive on average, a wide range of possibilities on a 100-point scale.

73. (a) By the Central Limit Theorem, the distribution of the mean of a large sample is approximately normal. Here, we're looking at the difference between the means of two (basically independent) large samples, which will be approximately normal as well, so a test based on the normal will work. Welch's $t$-test on this data will have a large number of degrees of freedom and thus will be nearly indistinguishable from the CLT test, so it will work as well.

(b) The test statistic is

$$t_W = \frac{16.8 - 24.3}{\sqrt{\frac{15.9^2}{592} + \frac{17.3^2}{154}}} = \frac{-7.5}{1.54} = -4.87.$$

(It could also be $+4.87$ if you do the subtraction the other way around.) The two-tailed $P$-value is `2 * (1 - pt(4.87, df = 225))`, which is of the order of two in a million.

(c) An approximate 95% confidence for $\Delta$ (treatment mean minus control mean) is $-7.5 \pm 1.96 \times 1.54$, which is about $-10.5$ to $-4.5$ weeks.

74. (a) The "before" and "after" measurements are taken on the same experimental units. So they will be dependent: If a person has relatively high blood flow before consuming caffeine, they will likely have relatively high blood flow after. Hence there is one independent sample from one population.

(b) The sample mean is $-1.154$. A 95% confidence interval is $-1.15 \pm 2.36 \times 0.63/sqrt(8)$, which is $-1.68$ to $-0.62$. If we repeated the experiment on eight randomly selected participants many times, then 95% of the time, we would get an interval that contained the true mean percentage change in blood flow (assuming a normal distribution.)

(c) All eight observations are negative; none are positive. The two-tailed $P$-value is double the chance that no observations are above the true median. This is $2 \times 0.5^8 = 1/128$. This is small (less than 1%), so there is strong evidence that the median percentage change is not zero (it's pretty clear that it's negative.)

75. (a)

$$t = \frac{0.3075}{\sqrt{\frac{1.362384}{4000}}} = 16.7$$

A $t$-distribution with 3999 degrees of freedom will be approximately standard normal, and 16.7 is a extreme value for a standard normal. Hence the $P$-value will be close to zero.

(b) $0.3075 \pm 1.96 \times \sqrt{1.362384/4000}$ gives 0.27 to 0.34 percentage points.

(c)

$$n = \left( \frac{2 \times 1.96 \times \sqrt{1.362384}}{0.02} \right)^2 \approx 52000$$

76. (a) Approximately normal. The histogram is similar in shape to the normal histogram, while the normal quantile plot is close to a straight line. However, very few things are exactly normal, and there is no reason to believe this is one of them. (You could argue subjectively that it not close to normal, depending on what you mean by "close.")

(b) The test statistic is $(98.5 - 98.6)/(0.73/\sqrt{130}) = -1.56$. We compare to a $t$ distribution with 129 degrees of freedom. One possible R command for the two-sided $P$-value is `2 * pt(-1.56, df=129)`.

(c) The sample size is large enough to use the standard normal confidence interval. This is

$$(98.5 - 1.96 \times 0.73/\sqrt{130}, 98.5 + 1.96 \times 0.73/\sqrt{130})$$

or 98.37 degrees to 98.63 degrees. The data is compatible with the null hypothesis of an average of 98.6 degrees. (Footnote: In actuality, average body temperature depends strongly on time of day and on method of measurement.)

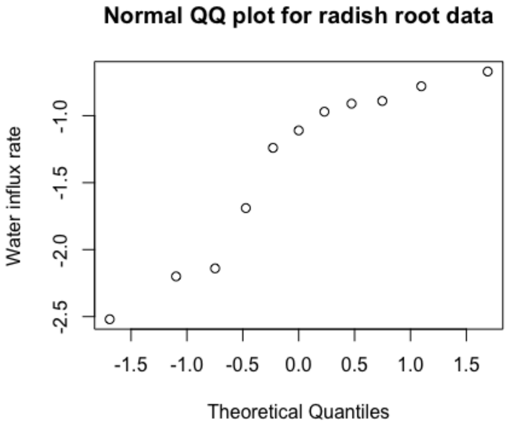
77. (a) Each pair of measurements is taken under the same conditions, so they are strongly dependent. Thus we should reduce to a one-sample problem by taking differences.

(b) Let $\mu$ be the expected value of the difference (method 2 minus method 1.) Since there's no direction given in the question,

$$\begin{aligned} H_0 &: \mu = 0 \\ H_1 &: \mu \neq 0 \\ t &= \frac{5.674 - 5.736}{\sqrt{\frac{0.08414412}{18}}} \\ &= -0.907 \end{aligned}$$

(c) There's no evidence of a difference, so unless very small differences matter, they might as well choose the considerably cheaper method.

78. (a) The $P$-value is tiny (essentially zero.) We can be sure that Americans with diabetes have higher systolic blood pressure on average. We can be confident the difference in averages is of the order of 9 to 12 units (which is quite a bit.)



79.

(b) The $P$-value is 0.27, not small. There's insufficient evidence for a difference in average diastolic blood pressure between Americans with and without diabetes. The confidence intervals for the difference in averages goes from $-0.5$ to 1.7 units, so if there is a difference, it's small.

(c) The confidence interval should go from the 347th lowest value to the 347th highest (401st lowest.) From the table, this is $[70, 74]$, endpoints included.

(a) We're taking two measurements on each radish root, and we shouldn't just assume the measurements are independent.

(b) i. The normal QQ plot is not close to a straight line, and the sample size is small.

ii. You can't log negatives!

(c)
```
old = c(89,  49,  91,  80,  56 , 79,  47,  50, 108, 165, 194)
young = c(213, 116, 260, 158, 153 ,170, 267, 264, 219, 254, 446)
ratios = young / old
# Lower bound
mean(ratios) - qt(0.975, df=10) * sd(ratios) / sqrt(10)
# Upper bound
mean(ratios) + qt(0.975, df=10) * sd(ratios) / sqrt(10)
```

80. (a) The experimental unit is a wolf spider. Two measurements are taken: Walking speed on filter paper with excrement, and walking speed on filter paper without excrement. There is only one independent sample.

(b) Let $\mu$ be the average difference in the walking speed of a wolf spider on filter paper with excrement compared to filter paper without excrement.

$$\begin{aligned} H_0 &: \mu = 0 \\ H_1 &: \mu \neq 0. \end{aligned}$$

If the population is normal, then under the null, the $t$-statistic has a $t$-distribution with 11 degrees of freedom. (If it's not normal then who knows.)

(c) This is too strong a conclusion. We are right on the border of statistical significance, and the sample size is only 12.

81. (a) We would draw a normal quantile plot of the weights. If the points were not close to a straight line, this would be evidence against the normal assumption.

(b) The $t$-statistic is 2. The $P$-value is `1 - pt(2, df=15)`, which would be 0.032. We would reject the null hypothesis, meaning we have evidence the mean is greater than 8 ounces.

(c) Assuming no ties, this would be `pbinom(5, 16, 0.5)` or `1 - pbinom(10, 16, 0.5)`, which is 0.105.

82. (a) This is a problem with one independent sample. The sample size is fairly large, so the $t$-confidence interval will be reasonable. The observed difference in means was 0.98. The estimated standard error of the difference is

$$\frac{1.82}{\sqrt{70}} = 0.218.$$

A 98% confidence interval for the mean difference is $0.98 \pm 2.38 \times 0.218 = 0.98 \pm 0.52$ or 0.46 to 1.5.

(b) By the sign test, a 98% confidence interval for the median difference runs from the 26th lowest to the 26th highest observation (inclusive.) This is 0 to 1.

(c) On the mean, the class scored significantly higher on the final question than on the corresponding midterm question. More specifically, 39 students did better while only 12 did worse. (The median is not too interesting here because of the discreteness of the distribution.)

83. (a) A quick glance at the numbers show no obvious outliers or skewness, so the normal assumption seems OK. The sample has mean 16.5 and SD 2.42. The $t$-statistic is $(16.5 - 16)/(2.42/\sqrt{10}) = 0.6547$. The $P$-value is `1 - pt(0.6547, df=9)`, or about 0.26. There's no evidence against the null hypothesis.

(b) The confidence interval is $16.5 \pm$ `qt(0.975, df=9)` $\times\ 2.42/\sqrt{10}$, or 14.8 to 18.2 years.

(c) This sample is skewed and so clearly non-normal, so we should avoid the standard normal and $t$ methods, especially with such a small sample.

84. (a) $(0 \times 9 + 1 \times 33 + 2 \times 38 + 3 \times 14 + 4 \times 4 + 5 \times 2)/100 = 1.77$.

(b) The standard error of the average is 0.106. A 95% confidence interval goes up and down about 1.96 standard errors from the sample average, which is 1.56 to 1.98 (the $t$-interval is the same to 3 sig. figs.)

(c) Averages can be decimals!

85. (a) After loading the data and running `qqnorm(x)`, we check to see if the plot is an approximately straight line (except perhaps at the corners.) If so, the data may be consistent with a normal distribution. On the other hand, if there is systematic bend in the QQ plot, the data isn't consistent with a normal distribution.

(b) 
```
T = mean(x) / (sd(x) / sqrt(15))
P.value = 2 * (1 - pt(abs(T), df = 14))
```

(c) There are 14 positive data points, so if $Y$ has a Binomial$(15, 0.5)$ distribution, the two-tailed $P$-value is double the smaller of $P(Y \leq 14)$ and $P(Y \geq 14)$. Clearly $P(Y \geq 14)$ is the smaller of these two. We then use the binomial to calculate:

$$\begin{aligned} P(Y = 14) &= C(15, 14)(1/2)^{14}(1/2)^1 = 15/(2^15) \\ P(Y = 15) &= C(15, 15)(1/2)^{15}(1/2)^0 = 1/(2^15) \end{aligned}$$

The $P$-value is thus $2 \times 16/(2^{15}) = 1/1024 \approx 0.001$, strong evidence that the median is not zero.

86. (a) The QQ plot is bendy.

(b) There are two observations above \$1000. A two-tailed $P$-value is $2 \times 0.055 = 0.11$, which is not usually considered significant.

(c) Sort the observations. The 98% CI will run from the 2nd lowest to the 2nd highest value, which gives \$364 to \$1063.

(d) The median may or may not be \$1000 (the sample median is \$720, which isn't that close.) The wide interval suggests we shouldn't read too much into such a small sample.

87. (a) The sample doesn't come from a normal population. It looks bimodal. (A histogram is better than a boxplot for showing this, but takes longer to draw.)

(b) None of the samples had caffeine over 188 mg ($Y = 0$). The probability of this happening by chance is $(1/2)^{12} = 1/4096$.

(c) Assuming your friend believes that science can ever prove anything, he or she is wrong. The $P$-value above takes the sample size into account — if you can reject so definitively with a small sample, you'll only reject more definitively with a larger one. Caveats: (1) The test was for a median, not a mean, but it's difficult to imagine the mean is an order of magnitude larger than the median. (2) We don't know how representative the sample was — for example, if all samples were taken from one cafe, the result may not generalize to other cafes.

88. (a) The distribution is strongly right-skewed (this is evident from a boxplot or other graph of the distribution). This violates the normality assumption of the $t$-test, and since the sample is small this is actually important.

(b) The $P$-value is the probability of 13 or more heads on 16 coin flips: `1 - pbinom(12,16,0.5)` = `pbinom(3,16,0.5)` = 0.011.

(c) This will run from the 4th smallest to 4th largest observation, i.e. 1 to 45 hundredths of an inch. (This is really a 97.8% confidence interval; if you are paranoid, a 99.6% confidence interval runs from the 3rd smallest to 3rd largest observation, or 0 to 48.)

**Ch 11**

89. (a) The distribution is not close to Normal. With a Normal distribution, 68% of the data should be within one standard deviation of the mean, with 16% more than an SD below and 16% more than an SD above. Well, one standard deviation below the mean for (for example) the red button distribution is about $-\$24$ dollars. But you can't have negative donations, despite what Clinton supporters would have liked. So the distribution can't be close to Normal.

(b) This is a 2-sample problem. The point estimate is $\hat{\Delta} = 6.60 - 4.76 = \$1.84$. Its standard error is

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{29^2}{6855} + \frac{49^2}{6639}} = \$0.696.$$

So a 95% confidence interval for $\Delta$, the difference in expected revenue per visitor, is $1.84 \pm 1.96 \times 0.696$, or \$0.48 to \$3.20.

(c) Do Welch's test, because you should never do Student's test. (Although to be honest, the sample sizes are large enough that nothing horrible will happen if you use Student's test despite the variances being unequal.)

(d) Welch's $t$-statistic is $1.84/0.696 \approx 2.64$. Under the null, this statistic should come from an approximate standard normal distribution, since the sample sizes are huge. It is somewhat uncommon for a standard normal random variable $Z$ to be as big in magnitude as $\pm 2.64$, so the $P$-value, which is $P(|Z| > 2.64)$, will be close to 0. (With R, we find it's 0.008.)

90. (a) Individuals will vary greatly in their baseline testosterone levels (e.g. most men will have higher levels than most women.) If what we care about is the *change* in testosterone induced by power poses, then using the testosterone after just adds noise to the data that's irrelevant to the treatments, reducing the power of the test.

(b) Let $\Delta$ be the expected change in testosterone after the high power pose minus the expected change in testosterone after the low power pose. The main pair of hypotheses we're testing is $H_0 : \Delta \leq 0$ vs. $H_1 : \Delta > 0$, i.e. we're trying to show that the high power pose has a positive effect on testosterone relative to the low power pose. The $P$-value for this test is 0.92, which is not small. So the data is consistent with the null that $\Delta \leq 0$. In fact, it was the low power sample that saw the greater mean increases in testosterone. The sample sizes are fairly large by the standards of psychology experiments. The two-sided confidence interval shows that it's hard to believe that even if the high power pose did increase mean testosterone by more, it would do so by more than a couple of points on a scale such that 100-point changes were observed. We conclude there's nothing in this data set in favor of the original researchers' conclusion regarding the effects of power poses on testosterone.

(c) No, this is a bad take. Firstly, it's not kosher to choose your hypotheses after looking at the data, and there's nothing that suggests that this was a hypothesis the researchers wanted to test before doing the experiment. Secondly, if we ignore the high power group, we no longer have a randomized controlled experiment and we no longer have a comparison group. This means we can't draw cause-and-effect conclusions — instead of or in addition to the low power pose, any other number of things to do with how the study was run could be affecting testosterone. If we really cared, we could run a new experiment comparing the low-power pose to a control, but since there does not seem to be any scientific reason why the low-power pose should affect testosterone positively while the high-power pose doesn't, this would not seem to be worth the effort.

91. (a) Experimental unit: Individual (Native American or Caucasian). Measurement: MSCE. Two independent samples.

    (b) Let $\Delta$ be the population mean MSCE in Native Americans minus the population mean MSCE in Caucasians. Then $H_0 : \Delta = 0$ and $H_1 : \Delta \neq 0$. (You could also define the parameters in terms of $\mu_{NA}$ and $\mu_C$, the mean MSCE for Native Americans and Caucasians respectively. Then $H_0 : \mu_{NA} = \mu_C$ and $H_1 : \mu_{NA} \neq \mu_C$.) In either case,

$$t_w = \frac{8.57 - 8.13}{\sqrt{\frac{0.54^2}{7} + \frac{0.49^2}{9}}} = 1.68.$$

    (If you use the raw data instead, $t_W$ is 1.70.)

    (c) No, we can only say the data is consistent with no difference in means. The sample sizes are far too small to prove the difference in means is zero or negligible.

92. (a) Let $\mu_r$ be the population mean log percentage change in red blood cell count for renal anemia patients and $\mu_h$ be the population mean log percentage change in red blood cell count for heart disease patients. Let $\Delta = \mu_r - \mu_h$. The hypotheses are:

$$\begin{aligned} H_0 &: \Delta \leq 0 \\ H_1 &: \Delta > 0 \end{aligned}$$

    (b) The Welch statistic is

$$t_W = \frac{-1.32782 - (-1.854697)}{\sqrt{\frac{1.358551^2}{16} + \frac{1.147175^2}{10}}} = 1.06$$

    The number of degrees of freedom is

$$\hat{\nu} = \frac{\left(\frac{1.359^2}{16} + \frac{1.147^2}{10}\right)^2}{(1.359^2/16)^2/15 + (1.147^2/10)^2/9} = 21.7.$$

    (c) `1 - pt(1.06, df=21.7)`

    (d) The data is consistent with a null hypothesis that $\Delta = 0$, so we have not proved there's a difference between renal and heart patients. However, the sample sizes are both very small, so it could just be that we lacked the power required to get a significant result. The back-transformed confidence interval shows the median renal anemia percentage change could be anywhere from 0.6 to 5 times the median heart disease percentage change.

93. (a) Inference on the raw data is easier to interpret (effects on a square root scale do not have an easy interpretation), at least for people used to working with this depression scale. However, the square-root data seems to give straighter normal QQ plots.

(b) Let $\mu_T$ be the expected value for the square root of the treatment score and $\mu_C$ be the expected value for the square root of the control score. Let $\Delta = \mu_T - \mu_C$. We test the hypotheses

$$\begin{aligned} H_0 &: \Delta \geq 0 \\ H_1 &: \Delta < 0 \end{aligned}$$

We could perform a Welch test, which assumes we have two independent samples from normal populations. The code `t.test(sqrt(depress.treatment), sqrt(depress.control), alt="less")` gives the correct $P$-value. (A nonparametric test on the untransformed data would arguably be better than this.)

(c) There's some evidence that the treatment works, though as usual larger sample sizes would be useful.

94. (a) Two. We treat the treatment and control groups as coming from two populations.

(b) The major one is normality, which we check by (for example) drawing normal QQ plots and checking they have no systematic bends. (In addition, we require independence, but this is something we usually assume rather than check.)

(c) Let $\Delta$ be the AMT population mean (expected value) minus the No AMT population mean (expected value.)

$$\begin{aligned} H_0 &: \Delta \leq 0 \\ H_1 &: \Delta > 0. \end{aligned}$$

(d) The test takes the wrong tail. The $P$-value should be $1 - 0.008706 = 0.991$. It appears that AMT *lowers* scores on this test (though the sample sizes here are small.)

95. (a) Neither appears normal. For one thing, normal distributions are symmetric, and the density plots do not look symmetric.

(b) Welch's $t$-test would be fine — there are two independent samples, and with very large sample sizes, the normality assumption is unimportant. A Central Limit Theorem ($z$) test would give much the same result.

(c) For samples of 570 men and 629 women, the average Feeling Thermometer towards scientists is 72.79 for men and 7283 for women. A Welch $t$-test gives a $P$-value of 0.98, meaning the data is compatible with the hypothesis that men and women have the same population mean for this variable. A confidence interval goes from $-2.8$ to 2.7, meaning there *could* be a difference in averages of a few points in one direction or the other, but such a difference would seem fairly small.

96. (a) Not normal — they're not close to a straight line. (The main issue is that lots of people offer exactly 50 chips.)

(b) The observed difference in means is $\hat{\Delta} = 41.78 - 42.43 = -0.65$. We estimate the standard error of $\hat{\Delta}$ as

$$\begin{aligned} se(\hat{\Delta}) &= \sqrt{\frac{14.31173^2}{41} + \frac{12.80462^2}{42}} \\ &= 2.983. \end{aligned}$$

The $t$-statistic is

$$\begin{aligned} t_W &= \frac{-0.65}{2.983} \\ &= -0.218. \end{aligned}$$

(c) The $P$-value is not small. This means the data is compatible with the hypothesis that on average, there's no difference in the number of chips given to happy players vs. angry players.

(d)
```
Delta.hat = mean(happy) - mean(angry)
std.error = sqrt(var(happy)/41 + var(angry)/42)
t.W = Delta.hat / std.error
df = (var(happy)/41 + var(angry)/42)^2 /
  ((var(happy)/41)^2/40 + (var(angry)/42)^2/41)
2 * (1 - pt(abs(t.W), df=df))
```

97. (a) The sample standard deviations are a little different, and there's no a priori reason to think that freshmen and seniors necessarily have the same population standard deviation of study time per week. So it's safer to do Welch's test.

(b) The $t$-statistic is

$$T = \frac{12 - 9}{\sqrt{4^2/16 + 5^2/25}} = 2.12.$$

(c) The standard error is

$$\sqrt{4^2/16 + 5^2/25} = \sqrt{2}.$$

So an approximate 95% confidence interval is $3 \pm 2.026\sqrt{2}$, or 0.13 to 5.86 hours. The interval doesn't include zero (equivalently, the $t$–statistic is less than 2.026), so we would reject the hypothesis that $\Delta = 0$ at the 0.05 level. We have some (but not overwhelming) evidence that $\Delta \neq 0$.

98. (a) This is a randomized experiment in which both groups are approximately normal and we have sample standard deviations instead of population standard deviations, so we should do some kind of two sample $t$-test. The standard deviations are somewhat different, so the Welch test is safer.

(b) The $t$-statistic is

$$T = \frac{70 - 55}{\sqrt{\frac{32^2}{25} + \frac{21^2}{36}}} = 2.056.$$

The degrees of freedom is

$$\hat{\nu} = \frac{\left(\frac{32^2}{25} + \frac{21^2}{36}\right)^2}{(32^2/25)^2/24 + (21^2/36)^2/35} = 38.16.$$

R code for the $P$-value is `2 * (1 - pt(2.056, df=38.16))`.

(c) Such a confidence interval is

$$\left(70 - 55 - 2.02\sqrt{\frac{32^2}{25} + \frac{21^2}{36}}, 70 - 55 + 2.02\sqrt{\frac{32^2}{25} + \frac{21^2}{36}}\right)$$

which is 0.26 to 29.7 days.

99. (a) By the Central Limit Theorem, for a large sample, the distribution of the sample average is approximately normal. The difference between the two sample averages here will thus be approximately normal, as the difference of two normally distributed random variables is itself normal.

(b) Such a confidence interval is

$$\left(47300 - 48600 - 1.96 \times \sqrt{\frac{8000^2}{400} + \frac{9000^2}{400}}, 47300 - 48600 + 1.96 \times \sqrt{\frac{8000^2}{400} + \frac{9000^2}{400}}\right)$$

or −\$2480 to −\$120.

(c) No. The test draws a conclusion about averages, not the whole population.

100. (a) Neither. Normal distributions are symmetric, and neither boxplot is close to symmetric — both are strongly skewed to the right.

(b) For employees of IU Athletics:

$$\begin{aligned} 95\% \text{ CI for mean} &= 81090 \pm 2.009575 \times \frac{73339}{\sqrt{50}} \\ &= 81090 \pm 20843 \end{aligned}$$

For IU Bloomington faculty:

$$\begin{aligned} 95\% \text{ CI for mean} &= 87040 \pm 1.984217 \times \frac{50051}{\sqrt{100}} \\ &= 87040 \pm 9931 \end{aligned}$$

The confidence intervals are approximately \$60,000 to \$102,000 for the IU Athletics mean and \$77,000 to \$97,000 for the IU Bloomington faculty mean.

(c) False. The test only tells you that the hypothesis that the population means might be the same can't be rejected. The summary statistics and the boxplots both give strong evidence that the populations differ because of their spreads: The athletics sample is much more spread out than the faculty sample, so it is reasonable to assume the same holds for the respective populations.

101.
```
hr2000 = read.table("Teaching/hr2000.txt", header=TRUE)
hr2010 = read.table("Teaching/hr2010.txt", header=TRUE)
rate2000 = hr2000$Homeruns / hr2000$Appearances
rate2010 = hr2010$Homeruns / hr2010$Appearances
qqnorm(rate2000)
qqnorm(rate2010)
```

```
t.test(rate2010, rate2000, alt="less")
```

Welch's $t$-test assumes independent samples from normal populations. Independence is probably okay (there aren't THAT any Major League Baseball players, but this is unlikely to be a big problem) and the QQ plots are consistent with normality. The test gives $T_w = -2.382$ and a $P$-value of 0.01, so there is evidence against the null hypothesis and for the alternative that the average rate of home runs per plate appearance has decreased.

102. (a) There won't be much difference between the various two-sample tests in this case, but the Welch test is safest. We'll perform this test at level $\alpha = 0.05$. The observed difference in means is 2 points. The standard error is

$$\sqrt{\frac{10^2}{100} + \frac{13^2}{100}} = 1.64$$

giving a $t$-statistic of $2/1.64 = 1.219$. The Welch test degrees of freedom is

$$\hat{\nu} = \frac{\left(\frac{10^2}{100} + \frac{13^2}{100}\right)^2}{\frac{(10^2/100)^2}{99} + \frac{(13^2/100)^2}{99}} = 185.78$$

The two-tailed $P$-value is `2*(1 - pt(1.219, df=185.78))`

(b)
```
2 - qt(0.95, df=185.78) * 1.64
2 + qt(0.95, df=185.78) * 1.64
```

(c) There isn't any evidence of a difference between the group means. That doesn't necessarily mean there's absolutely no difference at all. The confidence interval gives us an idea of how big the difference in means could reasonably be: from 0.7 points in favor of *Under Siege* to 4.7 points in favor of *The Joy of Stats*.

103. (a) Let $\mu_T$ be the (hypothetical) population mean percentage weight lost by newborns when assigned to the traditional ward. Let $\mu_E$ be the (hypothetical) population mean percentage weight lost by newborns when assigned to the traditional ward. Our hypotheses are:

$$\begin{aligned} H_0 &: \mu_T = \mu_E \\ H_1 &: \mu_T \neq \mu_E \end{aligned}$$

With reasonably large samples, we could justify either a two-sample $z$-test or a two-sample $t$-test (either Student's or Welch's — it makes no difference when the sample SDs are equal.)

(b) The $z$- (or $t$-) statistic is

$$z = \frac{5.1 - 6.0}{\sqrt{2^2/393 + 2^2/388}} = 6.29.$$

```
2 * (1 - pnorm(6.29))
# or
2 * (1 - pt(6.29, df = 779))
```

(c) There is strong evidence of a difference in mean percentage weight loss between the two treatments. Withholding bottle feeding seems to result in more weight loss, so maybe that's not a good idea.

104. (a) Since the sample SDs are somewhat different, I'd do a Welch two-sample $t$-test. This assumes independent normal data (a quick look at QQ plots does not reveal any strong non-normality.)

(b)

```
treatment = c(-2.3, -0.7, -0.2, 0.1, 0.5, 0.8, 0.9, 1.6, 2.0, 3.9, 4.5, 6.0)
control = c(-2.9, -1.5, -0.9, -0.8, -0.7, -0.5, -0.2, 0.2, 0.6, 1.2, 1.9, 2.8)
Delta = mean(treatment) - mean(control)
se = sqrt(var(treatment)/12 + var(control)/12)
t.stat = Delta / se
nu = (var(treatment)/12 + var(control)/12)^2 /
  ((var(treatment)/12)^2/11 + (var(control)/12)^2/11)
1 - pt(t.stat, nu)
```

The $P$-value is 0.04.

(c) There's some (but not insuperable) evidence that the treatment improves VO2 relative to the control (the difference in group mean improvements is 1.5 mL/kg/min.) Note that the study is rather small.

105. (a) Welch's two-sample $t$-test is a safe choice. The samples are plausibly from normal populations with unknown variances, and there's no a priori reason to be sure that the variances are equal. (Student's two-sample $t$-test is also justifiable here.)

(b) The observed average difference (Diet B minus Diet A) is 15.02 hours, with a standard error of 6.63 hours. The $t$-statistic is 2.265. The degrees of freedom calculation comes out to be 22.94. The $P$-value is `2*(1-pt(2.265, df=22.94))` (which is about 0.03).

(c) $15.02 \pm 2.07 \times 6.63$, or about 1.3 hours to 28.7 hours longer for Diet B.

106. (a) By the Central Limit Theorem, the averages of large samples have approximately normal distributions. Furthermore, the *difference* between two independent normally distributed random variables is itself normal.

(b) Using the standard normal interval: $3.16 \pm 1.96 \times 0.14$ or \$2.88 to \$3.44.

(c) No. There could be confounding factors — certain personality types might be more likely to smoke and have low wages. Or it could be that are more likely to smoke because they have lower wages.

107. (a) We have small samples, so the CLT doesn't help us. Since the samples are small, we need to account for the extra variation that occurs because we estimate the standard deviation using $s$. This is what the $t$-distribution is for.

(b) We go up and down 2.187 standard errors from the observed difference in means. $(68.5 - 65.5) \pm 2.187 \times \sqrt{\frac{3^2}{7} + \frac{2.5^2}{7}}$, or $-0.2$ inches to 6.2 inches.

(c) No. Based on the data alone, it's plausible that men and women at the university are the same height on average, but it's also plausible that men are on average six inches taller. Based on everything you have ever seen in real life, it's not plausible that men and women at the university average the same height.

**Ch 13**

108. (a) Plugging in gives 4.27.

 (b) $e^{4.37} = 79.3$ kilograms.

 (c) This is wrong — there is nothing in the data that measures *change.* (To put it another way, the student's statement is causal, and the data doesn't come from a controlled experiment.) The number 0.0107 can be interpreted as the difference in averages or predictions of log weight for heights that are 1 cm apart.

109. (a) We would expect 93.5 resistant, 187 mixed, and 93.5 susceptible.

 (b) The likelihood ratio chi-squared statistic is 0.180, while Pearson's chi-squared statistic is 0.182; either is fine. We compare this to a chi-squared distribution with 2 degrees of freedom (number of categories minus 1). The R code `1 - pchisq(0.180, df=2)` gives the $P$-value.

 (c) The data are consistent with the null hypothesis that 25% of lines would be resistant, 50% would be mixed, and 25% would be susceptible — discrepancies between these proportions and the sample proportions can be easily explained as chance variation. (Note: this is not a test of independence!)

110. (a) See the "Expected number of regions" column in the table below.

 (b) Pearson's version of the chi-square statistic is 3.39. (The likelihood ratio chi-square is basically the same.)

 (c) The data is consistent with the second-order Benford's law. There's no evidence of large-scale voter fraud (though that of course does not prove there's no voter fraud. . . )

| $x$ (Second digit) | $f(x)$ | Number of regions | Expected number of regions |
|---|---|---|---|
| 0 | 0.120 | 573 | 550.3 |
| 1 | 0.114 | 505 | 522.8 |
| 2 | 0.109 | 492 | 499.9 |
| 3 | 0.104 | 478 | 476.9 |
| 4 | 0.100 | 459 | 458.6 |
| 5 | 0.097 | 455 | 444.8 |
| 6 | 0.093 | 440 | 426.5 |
| 7 | 0.090 | 393 | 412.7 |
| 8 | 0.088 | 407 | 403.6 |
| 9 | 0.085 | 384 | 389.8 |

Table 8: Second digit Benford's law probabilities, along with frequencies for the number of times the second digit $x$ was observed in counts of the number of votes in the 2012 U.S. Presidential Election in 4586 county-level regions.

111. We create a table of expected values:

Note: This is a little bit dodgy as the expected number for "Formerly married" and "Not in labor force" is small. You would be justified in refusing to perform the chi-squared test at

|  | Married | Formerly married | Never married |
|---|---|---|---|
| Employed | 772 | 103 | 222 |
| Unemployed | 66 | 8.85 | 19 |
| Not in labor force | 28.9 | 3.86 | 8.29 |

this point. (In practice, nothing horrible is going to happen with one slightly small count out of 9.) Pearson's chi-square is 14.2, which compared to a chi-squared distribution with 4 degrees of freedom gives a $P$-value of 0.007. (Note: A nonparametric way of finding the $P$-value that accounts for the small sample size gives basically the same result.) The likelihood ratio chi-square statistic is 13.2, giving a $P$-value of 0.010. Either way, we have evidence against the null hypothesis. That is, we have evidence that marital status and employment status are statistically dependent.

112. (a) In the data provided, 65.6% of victims were white while 34.4% of victims were black, while 11.0% of the convicted received the death penalty and 89.0% did not. If independence held, the expected number for each of the four combinations would be:
   - White victim, death penalty: 7.2% = 23.6
   - White victim, no death penalty: 58.4% = 190.4
   - Black victim, death penalty: 3.8% = 12.4
   - Black victim, no death penalty: 30.6% = 99.6

   (b) The likelihood ratio chi-square statistic is

$$G^2 = 2[30\log(30/23.6) + 184\log(184/190.4) + 6\log(6/12.4) + 106\log(106/99.6)] = 6.25.$$

   If you prefer Pearson's chi-squared test, the test statistic is 5.61. The $P$-value is `1 - pchisq(6.249715, df=1)` = 0.012 (0.018 for Pearson's.)

   (c) There's evidence against the null hypothesis, i.e. the data suggests that race and death penalty are not independent. (It doesn't address the causes of the dependence, and we should be hesitant to come to any such conclusions until we've thought carefully about confounding.)

113. (a)

```
> 1 - pbinom(32, 66, 0.35) # or pbinom(33, 66, 0.65)
[1] 0.008644789
> 1 - pbinom(32, 66, 0.19) # or pbinom(33, 66, 0.89)
[1] 1.404454e-08
```

(b)

```
> expected = 66 * c(0.42, 0.23, 0.16, 0.19)
> expected
[1] 27.72 15.18 10.56 12.54
```

(c)

```
> observed = c(5, 9, 19, 33)
> X2 = sum((observed - expected)^2 / expected)
> X2
[1] 61.26556
> 1 - pchisq(X2, 3)
```

```
[1] 3.154144e-13
> # or
> G2 = 2 * sum(observed * log(observed / expected))
> G2
[1] 59.6437
> 1 - pchisq(G2, 3)
[1] 7.004397e-13
```

No, the jurors are not a random sample.

114. (a)

| First digit | Observed number of accounts (thousands) | Expected number of accounts (thousands) |
|---|---|---|
| 1 | 12614 | 11639 |
| 2 | 6443 | 6808 |
| 3 | 4563 | 4831 |
| 4 | 3581 | 3747 |
| 5 | 2951 | 3061 |
| 6 | 2533 | 2588 |
| 7 | 2227 | 2242 |
| 8 | 1988 | 1978 |
| 9 | 1763 | 1769 |

Table 9: Observed and expected counts of the first digit of the number of Twitter followers for 38,663,000 Twitter accounts.

(b) If we use the likelihood ratio chi-squared test, $G^2 = 127429$. If we use Pearson's chi-squared test, $X^2 = 128811$. (Either way, the $P$-value is basically zero. Contrary to some media reports, the number of Twitter followers doesn't follow Benford's law: there are too many first digit 1's.)

115. (a) Let's do Pearson's chi-square test. There are 244 Democrats and 171 Republicans.

```
ob = c(153, 91, 136, 35)
ex = c(244*289/415, 244*126/415, 171*289/415, 171*126/415)
X2 = sum((ob-ex)^2/ex)
1 - pchisq(X2, df=1)
```

Pearson's chi-square statistic is 13.46. Comparing this to a chi-squared distribution with 1 degree of freedom, we get a $P$-value of 0.0002. Vote and party were not independent. (The likelihood ratio chi-square statistic is 13.87, giving a similar $P$-value.)

(b) There are 313 northerners and 102 southerners.

```
ob = c(281, 32, 8, 94)
ex = c(313*289/415, 313*126/415, 102*289/415, 102*126/415)
X2 = sum((ob-ex)^2/ex)
1 - pchisq(X2, df=1)
```

Pearson's chi-square statistic is 244. Comparing this to a chi-squared distribution with 1 degree of freedom, we get a $P$-value that's essentially zero. Vote and party were not independent. (The likelihood ratio chi-square statistic is 247, also giving a zero $P$-value.)

(c) No. Democrats were more likely to vote against the act than Republicans, and southerners were more likely to vote against the act than northerners, but that doesn't imply that southern Democrats were the most likely to vote against the act. (In fact, while 83 out of 91 southern Democrats voted against, all 11 out of 11 southern Republicans voted against.)

116. (a)
```
men = c(512, 353, 120, 138, 53, 16)
women = c(89,17, 202, 131, 94, 24)
n.men = sum(men)
n.women = sum(women)
n = n.men + n.women
department = (men + women)/n
men.expected = department * n.men
women.expected = department * n.women
```

See Table 10. The smallest expected count is 13, which is sufficient to do a chi-squared test of independence.

| Department | Men admitted | Women admitted |
|---|---|---|
| A | 410 | 191 |
| B | 252 | 118 |
| C | 219 | 103 |
| D | 183 | 86 |
| E | 100 | 47 |
| F | 27 | 13 |

Table 10: Table of expected admissions to the six largest departments at a large university (rounded to whole numbers).

## Ch 15

117. (a) Predicted math score $= 2.4 * height - 65.2$

(b) Standard error $= s_Y/s_X\sqrt{(1-r^2)/(n-2)} \approx 0.42$. CI: $2.4 \pm 2 \times 0.42$, or 1.6 to 3.2 points per inch.

(c) No, this is dumb. Unless there's negligible confounding, regression doesn't give an accurate estimate of cause-and-effect, and here age (for instance) is clearly a confounding variable.

118. (a)

$$\begin{aligned}\text{Predicted husband's height} &= \left(68 - \frac{0.25 \cdot 2.7}{2.5} \times 63\right) + \frac{0.25 \cdot 2.7}{2.5} \times \text{wife's height} \\ &= 50.99 + 0.27 \times \text{wife's height}\end{aligned}$$

(b)

$$\begin{aligned}\text{Predicted wife's height} &= \left(63 - \frac{0.25 \cdot 2.5}{2.7} \times 68\right) + \frac{0.25 \cdot 2.5}{2.7} \times \text{husband's height} \\ &= 47.26 + 0.231 \times \text{husband's height}\end{aligned}$$

(c) The regression prediction for the height of the wife of a 66-inch husband is $47.26 + 0.231 \times 66 = 62.54$ inches. The heights of such wives are approximately normally distributed with mean 62.54 and standard deviation $2.5 \times \sqrt{1 - 0.25^2} = 2.42$ inches. So the required probability is

$$P(Y > 66) \approx 1 - \text{pnorm}(66, 62.4, 2.42) \approx 8\%.$$

119. (a) Slope $= 0.1381 \times 6.70/1.69 = 0.547$. Intercept $= 28.67 - \text{slope} \times 2.88 = 27.09$.

$$\text{Predicted BMI} = 27.09 + 0.547 \times \text{hours of TV watched per day}$$

(b) Substituting 2 into the above equation, we get 28.2.

(c) Substituting 4 into the above equation, we get 29.28. Multiplying by height squared gives a weight of about 75 kilograms.

120. (a)

$$\text{Predicted wins} = 0.327 \times \text{Previous year's wins} + 5.384$$

(b) Houston would have been predicted to win 6.038 games, so their residual was around +3 wins. Dallas would have been predicted to win 8 games, so their residual was +4 wins. Dallas exceeded their prediction by a larger margin.

(c) Regression toward the mean exists: Bad teams have more room to improve and can't get much worse, while good teams have more room to decline and can't get much better.

121. (a) The regression line goes through the point of averages and has slope $0.4 \times 0.66/1 = 0.264$. So a student one unit above average on SES would be predicted to be 0.264 units above average on GPA, which would be 3.474.

(b) Using the same regression line as above, a student one unit below average on SES would be predicted to be 0.264 units below average on GPA, which would be 2.946.

(c) We could do this using the bivariate normal or empirically. To do it empirically, get the data set, pick out all students with GPA close to 4.0 (e.g. between 3.9 and 4.1), and find out what proportion of them have above average SES.

e.g. If the data has two variables called `GPA` and `SES`,

```
fourpointoh = which(GPA > 3.9 & GPA < 4.1)
mean(SES[fourpointoh] > 0)
```

122. (a) It's not bivariate normal. The data doesn't form an ellipse, and both the midterm and final distributions are left-skewed.

(b) Firstly, find the set of differences. Then take the 3rd smallest to 3rd largest. This gives $-5$ to 10.

(c) Nope. I mean that one student who got 2 on the final definitely did worse.

123. (a) Regression to the mean means that we predict the best teams will do worse than they did the previous year.

(b) Predicted 2015 wins $= 0.54 \times (\text{2014 wins}) + 37.26$

(c) 90.6 wins.

(d) While it's likely that some team will win 96 games or more, there's an element of luck involved. Some teams will be better than their projection. We can't tell in advance which team will be the luckiest.

124. (a) "Regression to the mean" means that students that do exceptionally well on one test will, on average, do less exceptionally (but still well) on a related test. Some students that did well on the first test did so by chance.

(b) The regression line has slope 0.6 and intercept $68 - 0.6(65) = 29$. The prediction for Kohei's Test Two is $29 + 0.6(85) = 80$.

(c) Regression is not symmetric: the line to predict Test One from Test Two is not the same as the line to predict Test Two from Test One. In contrast, while students who do well on Test One are predicted to do relatively less well on Test Two, students who do well on Test Two are predicted to have done relatively less well on Test One. The regression line to predict Test One from Test Two has slope 0.6 and intercept $65 - 0.6(68) = 24.2$. The prediction for Gabby's Test One is $24.2 + 0.6(80) = 72.2$.