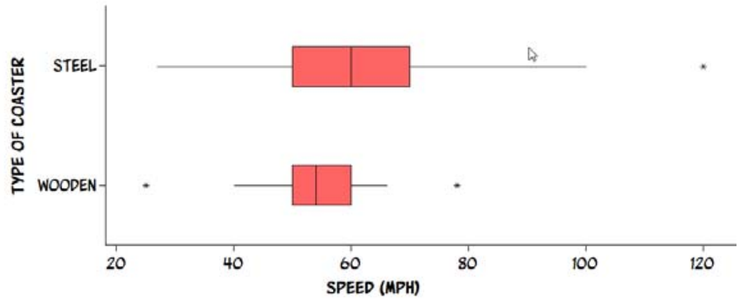
ISCAM 3: CHAPTER 2 EXERCISES

1. Roller Coaster Speeds

The Roller Coaster Database maintains a web site (www.rcdb.com) with data on roller coasters around the world. Some of the data recorded include whether the coaster is made of wood or steel and the maximum speed achieved by the coaster, in miles per hour. The boxplots display the distributions of speed by type of coaster for 145 coasters in the United States, as downloaded from the site in November of 2003.



(a) Do these boxplots allow you to determine whether there are more wooden or steel roller coasters?
(b) Do these boxplots allow you to say which type has a higher percentage of coasters that go faster than 60mph? Explain and, if so, answer the question.
(c) Do these boxplots allow you to say which type has a higher percentage of coasters that go faster than 50mph? Explain and, if so, answer the question.
(d) Do these boxplots allow you to say which type has a higher percentage of coasters that go faster than 45mph? Explain and, if so, answer the question. *Hint*: Think twice on this one.
(e) Which type of coaster has more "outliers"? Explain how you are deciding.
(f) Conjecture as to how the mean, median, interquartile range, and standard deviation will change (if at all) if the faster steel coaster (Top Thrill Dragster in Cedar Point Amusement Park, Sandusky, Ohio) is removed from the data set. Explain your reasoning.

2. Roller Coaster Speeds (cont.)

Reconsider the data in the previous exercise on 139 coasters in the United States, as downloaded from the www.rcdb.com site in November of 2003 (coasters.txt).
(a) Identify the observational units in this study. Then identify the variable of interest here. Also whether it is a quantitative or a categorical variable.
(b) Write a paragraph comparing and contrasting these distributions. Describe the shaper, center, and spread (as best you can) for each distribution, and then also comment on the issue of whether one type of coaster tends to have higher speeds than the other. Remember to state your description in the context of the study.

3. Old Faithful Geyser

Millions of people from around the world flock to Yellowstone Park in order to watch eruptions of Old Faithful geyser. How long does a person usually have to wait between eruptions, and has the timing changed over the years? In particular, scientists have investigated whether a 1998 earthquake lengthened the time between eruptions at Old Faithful. The data in OldFaithful.txt are the inter-eruption times (in minutes) for all 108 eruptions occurring between 6am and midnight on August 1–8 in 1978 (from Weisberg, 1985) and for 95 eruptions for the same week in 2003 (http://www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFAITHFUL).

(a) Use technology to determine the five-number summary for each distribution and produce boxplots on the same scale. What does this analysis reveal about the typical waiting times and the variability in waiting times?
(b) What feature of the distributions is not very well revealed by this analysis?
(c) Do modified boxplot identify any outliers in these distributions?
(d) Suppose the two lowest inter-eruption times in 2003 were removed from the data set, explain how the mean and standard deviations of the inter-eruption times for 2003 would change (larger, smaller, not much change). Explain your reasoning.

4. US Births (cont.)
Return to the USbirthsJan2013.txt data from Investigation 2.1. (Recall more detailed descriptions of the variables can be found here.)
(a) Produce numerical and graphical summaries of the *apgar scores* for the full term babies. Describe what you learn (in context).
(b) Are the *apgar scores* of the premature babies noticeably lower?
(c) Repeat (a) and (b) for the *mother's weight_gain* (in pounds) variable. ["A reported loss of weight is recorded as zero gain."]
(d) Does the mother's weight gain appear to be a predictor of the health of the baby at birth? Justify your reasoning.

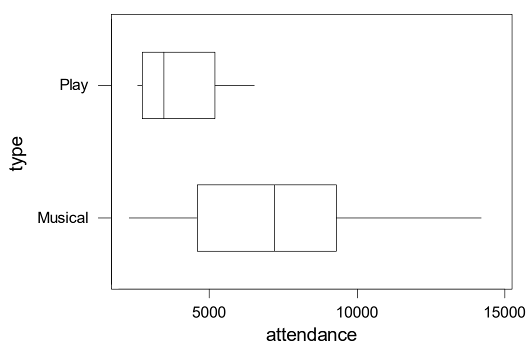
5. Guess the Instructor's Age
The file AgeGuesses.txt contains guesses of an instructor's age by her current students. Let μ represent the average guess of her age by all current students at the university and suppose the sample constitutes a representative sample of all students at this school on this issue.
(a) Produce numerical and graphical summaries of the distribution and describe what you learn (in context).
(b) Use a normal probability plot to decide whether the data has strong deviations from the pattern of a normal distribution.
(c) Use technology to determine a 90% one-sample *t*-interval for these data. Include your output and comment on the validity of this procedure. Provide a one-sentence interpretation of this interval.
(d) Count how many of the class guesses are inside the 90% confidence interval. Compute the percentage of the class guesses that are inside the interval. Is this close to 90%? Should it be?
(e) Calculate and interpret a 90% prediction interval. Include the details of your calculation and comment on the validity of this procedure. How does the prediction interval compare (midpoint, length) to the confidence interval?

6. July Temperatures
The July 8, 2012 edition of the *San Luis Obispo Tribune* listed predicted high temperatures (in degrees Fahrenheit) for that date. One section reported predictions for locations in San Luis Obispo County, another section for locations throughout the state of California, and another section for cities across the United States. The data can be found in the file JulyTemps.txt.
(a) Produce (and submit) dotplots of the predicted high temperatures for the three regions, using the same scale and on the same axis for each dotplot.
(b) Calculate (and report) the mean and median, SD, and IQR of the temperatures for each region.
(c) Based on the graphs and statistics, write a paragraph comparing and contrasting the distributions of predicted high temperatures in the three regions. [*Hint*: As always when describing distributions of quantitative data, be sure to comment on center, variability, shape, and outliers.]

(d) Produce (and submit) histograms of the predicted high temperatures for the three regions, using the same scale for each histogram.
(e) The San Luis Obispo county and California region display some *bi-modality* in their distributions. Describe what this means, and provide an explanation for why it makes sense that these distributions reveal some bi-modality.
(f) Calculate (and report) the five-number summary of the temperatures for each region.
(g) Produce (and submit) boxplots of the predicted high temperatures for the three regions, using the same scale and on the same axis for each boxplot.
(h) Identify the location/city for any outliers revealed in the boxplots. Also use the 1.5×IQR criterion to verify (by hand) that the location/city really is an outlier.
(i) Now change the measurement units to be degrees Celsius rather than degrees Fahrenheit. [*Hint*: Create a new variable by first subtracting 32 from the temperature and then multiplying by 5/9.] Produce (and submit) dotplots of the predicted high temperatures (in degrees Celsius) for the three regions, using the same scale and on the same axis. Comment on how the shapes in these dotplots compare to the original dotplots (when the measurement units were degrees Fahrenheit).
(j) Calculate (and report) the mean and median, SD, and IQR of the temperatures (in degrees Celsius) for each region.
(k) Determine (and describe) how the values of these statistics have changed based on the transformation from degrees Fahrenheit to degrees Celsius. [*Hint*: Be as specific as you can be. For example, do not just say that the SD got smaller.]

7. Broadway Attendance

The boxplots shown reveal the distributions of weekly attendance for Broadway shows in the first week of September in 1999, where the shows have been categorized as "play" or "musical."
(a) Did one type of show (play or musical) tend to have more attendees than the other? Justify your conclusion.



(b) Did one type of show tend to have more variability in their attendance figures than the other? Justify your conclusion.
(c) Which distribution appears to be more skewed? Explain how you are deciding.
(d) For the musicals, the mean was equal to 7121 and the standard deviation was equal to 3126. What are the "measurement units" of these numbers?
(e) For the musicals, between what two values do you expect to find the middle 68% of the attendance figures? Explain.

8. The Empirical Rule

The "Empirical Rule" is actually a famous result for normal distributions, claiming not only that approximately 95% of the observations fall within two standard deviations of the mean, but also that roughly 68% fall within one standard deviation and 99.7% fall within three standard deviations.

(a) In empirical sciences, the "three-sigma rule" claims "nearly all" values are taken to like within three standard deviations of the mean. Is this consistent with the empirical rule?
(b) "Six Sigma" became famous in the 1980s and 1990s for improving manufacturing processes. To allow for changes over time, this asserts a process is out of control if the process mean falls more than 4.5 standard deviations from the nearest specification limit. Use a standard normal probability model to determine how many "defective parts per million opportunities" (DPMO) this allows (one-sided)?
(c) Wikipedia claims that in particle physics, a "five sigma effect" is needed before a result qualifies as a discovery. According to the normal distribution, how often will a five sigma effect occur?

Note: In the *Black Swan*, the author claims that conventional risk models implied the Black Monday crash in 1987 would correspond to a 36-sigma event, instantly suggesting the models were flawed.

9. Sleeping Students (cont.)

Reconsider the students' sleeping times from the Chapter 0 Exercises (SleepStudents.txt).
(a) Determine the five-number summary of sleeping times for each student.
(b) For each student, determine which (if any) of their sleeping times qualify as outliers by the 1.5IQR rule.
(c) Create boxplots of these students' sleeping times on the same scale. Comment on what these boxplots reveal.
(d) What does the dotplot reveal about Amber's sleeping times that the boxplot does not?

10. Sleeping Students (cont.)

Reconsider the students' sleeping times from Exercise 9 (SleepStudents.txt).
(a) Calculate the mean and standard deviation of sleeping times for each student.
(b) For each student, determine the proportion of the 63 sleeping times that fall within one standard deviation of the mean.
(c) For which student does the empirical rule (see Exercise 8) appear to hold most closely? For that student, determine the proportion of sleeping times that fall within two standard deviations of the mean.
(d) Suppose that Katherine gets 10 hours of sleep in a particular night. How many hours more than her mean is this? Also calculate the z-score for this value.
(e) Suppose that Amber gets 13 hours of sleep in a particular night. How many hours more than her mean is this? Also calculate the z-score for this value.
(f) Which of these (10 hours for Katherine or 13 for Amber) is higher above that student's mean? Which has the higher z-score? Explain why your answers are not the same.

11. Sleeping Students (cont.)

Reconsider the students' sleeping times from the previous exercises (SleepStudents.txt). The worksheet also includes a day-of-the-week variable and a variable called *school night*? indicating whether school was in session the next day. For each student, analyze her sleeping times on school nights vs. non-school nights. Write a paragraph summarizing your findings. Also identify which student appears to have the biggest difference in sleeping times between these two kinds of days, and identify which has the least difference.

12. Hypothetical Quiz Scores

Reconsider the hypothetical quiz scores for classes A–D in the Chapter 0 Exercises.
(a) For each class (A–D), calculate the range of the quiz scores.
(b) Is the range a helpful measure here is comparing the variability of these distributions? Explain.

13. Create an Example

(a) Create a hypothetical example of 10 exam scores (say, between 0 and 100 with repeats allowed) such that 90% of the scores are above the mean.
(b) Repeat (a) for the condition that the mean is roughly 40 points less than the median.
(c) Repeat (a) for the condition that the IQR equals 0 and the mean is more than twice the median.

14. Measures of Center and Spread

The *mid-range* of a dataset is defined to be the sum of the minimum and maximum values divided by 2. The *mid-hinge* of a dataset is defined to be the sum of the first and third quartiles divided by 2.
(a) Is mid-range a measure of center or a measure of spread? Explain.
(b) Is mid-hinge a measure of center or a measure of spread? Explain.
(c) Is the mid-range resistant to outliers? Explain.
(d) Is the mid-hinge resistant to outliers? Explain.

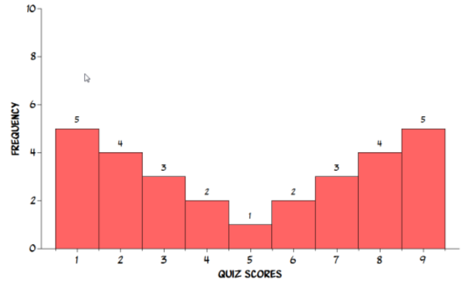
15. Identifying Outliers

Perhaps you are wondering about the motivation behind the "1.5IQR criterion" for identifying outliers.
(a) Determine the 25^{th} and 75^{th} percentiles of the standard normal model. Then calculate the inter-quartile range. Also draw a well-labeled sketch of the standard normal curve and indicate how to find the value of the IQR on the graph.
(b) Using the "1.5IQR" rule for identifying outliers, determine what proportion of the values from a standard normal distribution would be classified as outliers. [*Hint*: Again draw a sketch first, and then identify the "cut-off" points for identifying outliers using your answers from (a).]
(c) Use a simulation as a check on your calculations: First simulate 1000 random values from a standard normal distribution. Then determine the IQR for your 1000 simulated values. Finally, set up an indicator variable to count how many of the values are not outliers. Also draw a boxplot to reveal the outliers. What proportion of the 1000 random values are identified as outliers? Is this close to your answer to (b)?
(d) Now consider a more general normal model with mean μ and standard deviation σ. Determine how your answers to (a) and (b) will change, if it all. Follow up with a technology simulation using a few different values of (μ, σ) as a check on your work. Summarize your results.
(e) Based on your simulation in (c), what proportion of the 1000 random values are more than 1IQR from the respective quartiles? What proportion of the 1000 random values are more than 2IQR from the respective quartiles? Explain why someone might consider 1.5IQR a more reasonable way to identify outliers than 1IQR or 2IQR.
(f) The rule of "3IQR" has also been recommended as a way to identify "extreme" outliers. What proportion of your simulated values are more than 3IQR are from the quartiles?

16. Identifying Outliers (cont.)

Reconsider the previous question. An alternative procedure for identifying outliers is to classify any value more than three standard deviations away from the mean as an outlier.
(a) By this criterion, what proportion of values from a normal distribution will be identified as outliers? Is this more or less than with the 1.5IQR criterion? Much more so?
(b) Repeat (a) if the criterion is to classify any observation more than *two* standard deviations away from the mean as an outlier.
(c) Explain how the 1.5IQR rule is a more "general" criterion than using 2 or 3 standard deviations? [*Hint*: When would the latter condition not be reasonable to apply?]

17. Properties of Center and Spread
The following histogram displays the (hypothetical) quiz scores for a class of n = 29 students.



Suppose we were to give every student 5 bonus points.
(a) How would the mean change? The median?
(b) How would the standard deviation change? The inter-quartile range?
Note: You should explain your answers to (a) and (b) without carrying out the calculations to find these new values.

18. Linear Transformations
Suppose that a *linear* transformation is applied to a set of data, so all of the x_i's are converted into y_i's by the expression $y_i = a + b\,x_i$ for some constants a and b. It can be shown that the mean of the transformed data is $\bar{y} = a + b\bar{x}$ and the standard deviation is SD(y) = |b|SD(x).
(a) Prove these results (using summation notation).
(b) Determine the effect of this linear transformation on the *median* of the data. Justify your answer. Prove that your answer is correct, making sure you thoroughly explain your proof.
(c) Determine the effect of this linear transformation on the *IQR* of the data. Justify your answer. Prove that your answer is correct, making sure you thoroughly explain your proof.

19. Seeding Clouds
The values in CloudSeeding.txt report the volume (acre-feet = "height" of rain across one acre) of rainfall from selected clouds in a 24-hour period. (In Chapter 3 you will compare the treatment groups, but for now just examine the rainfall amounts.)
(a) Produce a graph and describe the distribution of the rainfall amounts.
(b) Apply a log transformation to the rainfall amounts. Comment on the normality of the resulting variable's distribution.
(c) Apply a square root transformation to the rainfall amounts. Which transformation procedures more normally distributed data? Justify your answer.

20. Seeding Clouds (cont.)
Reconsider the previous exercise.
(a) Use technology to take the (natural) log transformation of the rainfall amounts. Calculate and report the mean and median of these transformed values.
(b) Does the mean of the ln(rainfall) amounts equal the ln of the mean of the rainfall amounts? Report calculations to support your answer.
(c) Does the median of the ln(rainfall) amounts equal the ln of the median of the rainfall amounts? Report calculations to support your answer.
(d) Will the relationship that you found in (c) always hold? If so, explain. If not, provide a counterexample.

21. Log Transformations

Suppose that a *logarithmic* transformation is applied to a set of data, so all of the x_i's are converted into y_i's by the expression $y_i = \log(x_i)$.

(a) Explain why you cannot say what effect this would have on the mean of the data.
(b) Describe what effect this would have on the median of the data, and justify your answer.
(c) Between the IQR and standard deviation, for which measure can you say what the effect would be? Describe that effect, and justify your answer.

22. Transformations

Consider a general power transformation, represented by the function $f(x) = x^p$, for some power p.

(a) Explain why using the power $p = 0$ does not make sense.

The log transformation actually "takes the place" of zero on the power transformation scale. You can see this by examining derivatives.

(b) Take the derivative (with respect to x, for a fixed value of p) of $f_p(x) = x^p$.
(c) Take the derivative of $f(x) = \log(x)$.
(d) Explain how these derivatives reveal that $\log(x)$ is comparable to a power of zero on the power transformation scale. [*Hint*: $f'(x)$ has the same exponent on x as $f_p'(x)$ for what value of p?]

23. Body Mass Index

The data in BodyMassIndex.txt are ages (in years), weights (in kg), and heights (in cm) for a sample of adults (Heinz et al., 2003). Body mass index (BMI) is defined to be a person's weight (in kg) divided by the square of their height (in meters). (Divide height in cm by 100 to convert to meters.)

(a) Use technology to calculate the BMI values for this sample of adults by computing

$$BMI = (weight)/(height)^2 \times 10000$$

(b) Produce boxplots and descriptive statistics comparing BMI values between men and women. Write a paragraph summarizing your findings. [Remember to comment on center, spread, and shape.]
(c) Try several transformations (log, square root, reciprocal) of the BMI values for the two sexes combined. Identify which transformation produces an approximately symmetric distribution for the BMI values. Provide graphical displays to support your answer.
(d) Examine histograms of the BMI values for men and women separately. Then repeat this transformation analysis for men and for women separately. For each sex, identify which transformation produces an approximately symmetric distribution for the BMI values. Provide graphical displays to support your answer.

24. Mean IQs

Is it possible for an individual to move from one city to another and have the mean IQ decrease in both cities? If not, explain why not. If so, explain what conditions would be needed to make this happen.

25. Average Children

Suppose that you record the number of children in each of ten families (labeled as A–J) to be:

Family	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Number of children	1	2	1	0	2	2	3	7	4	2

(a) Determine the average (mean) number of children per family.

Now consider the 24 children in these families as the observational units, and consider the variable "number of siblings." Thus, the one child in family A has 0 siblings, each of the two children in family B has 1 sibling, and so on.

(b) Determine the average number of siblings per child.
(c) Some might expect that there would be a clear relationship between these two averages. For example, some might suspect that the average number of siblings would be one less than the average number of children. Give a mathematical explanation for why this is not the case.

26. Average Children (cont.)
Reconsider the previous question. A similar phenomenon can reveal itself with class sizes. The average number of students per class can be very different from the average class size per student. Demonstrate this with a hypothetical example of five classes. Specify the number of students in each class, and then calculate the average number of students per class. Then consider the students as the observational units, with "number of students in that student's class" as the variable, and calculate the average class size per student. Construct your example so that these two averages are quite different, and explain why that happens.

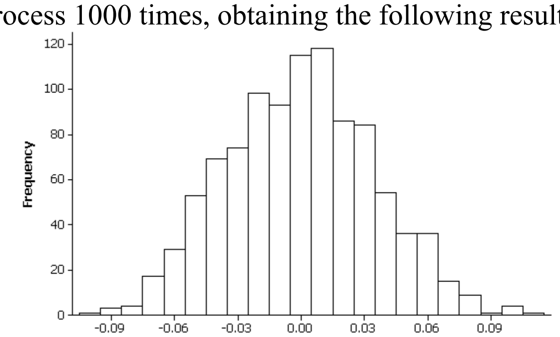
27. Body Mass Index (cont.)
Suppose that the body mass index (BMI) of healthy American males follows a symmetric, mound-shaped distribution with mean 24.5 and standard deviation 3.0 and that the BMI of healthy American females follows a symmetric, mound-shaped distribution with mean 22.5 and standard deviation 3.0.
(a) Between what two values would approximately 95% of males' BMI values fall?
(b) About what percentage of male BMI values fall below 21.5?
(c) About what percentage of male BMI values fall above 30.5?
(d) About what percentage of female BMI values fall between 19.5 and 25.5?
(e) About what percentage of female BMI values fall between 16.5 and 25.5?
(f) Below what value do about 2.5% of female BMI values fall?

28. SATs
Suppose the distribution of SAT scores is mound-shaped and symmetric with a mean of 1500 and a standard deviation of 240, and that the distribution of ACT scores is mound-shaped and symmetric with a mean of 21 and a standard deviation of 5. Suppose Tory scores a 1800 on the SATs and Jeff scores a 28 on the ACT.
(a) Provide a rough sketch, labeling the horizontal axis, of each distribution and indicate where the observed test score falls on the distribution.
(b) Which test taker had a higher score relative to the distribution of scores on that test? Explain. [*Hint*: Compare their *z*-scores.]

29. SATs (cont.)
Recall the previous Exercise, in which you considered SAT scores and ACT scores to have symmetric, mound-shaped distributions. Continue to assume that SAT scores have mean 1500 and standard deviation 240, whereas ACT scores have mean 21 and standard deviation 5.
(a) An ACT score of 21 is equivalent to what SAT score, in terms of z-scores?
(b) An ACT score of 26 is equivalent to what SAT score, in terms of z-scores?
(c) An ACT score of 28 is equivalent to what SAT score, in terms of z-scores?
(d) Let x represent a generic ACT score, and let y represent the SAT score to which x is equivalent, in terms of z-scores. Determine y as a function of x.
(e) Graph the function in (d), and confirm that it satisfies your answers to (a), (b), and (c).

30. Equating *z*-scores
Reconsider the previous exercise. Suppose that two variables both have symmetric, mound-shaped distributions, and you want to find the value of one variable (call it y) that has the same z-score as a given value of the other variable (call it x). Denote the means of the variables by μ_x and μ_y, and denote their standard deviations by σ_x and σ_y.
(a) Derive a function that expresses y as a function of x, μ_x, μ_y, σ_x, and σ_y.
(b) If all else remains unchanged, is y an increasing or a decreasing function of x? Explain both algebraically and intuitively.
(c) Repeat (b), answering whether y is an increasing or a decreasing function of μ_x.
(d) Repeat (b), answering whether y is an increasing or a decreasing function of μ_y.
(e) Repeat (b), answering whether y is an increasing or a decreasing function of σ_x.
(f) Repeat (b), answering whether y is an increasing or a decreasing function of σ_y.

31. Normal Groceries
Suppose you take a random sample of 30 grocery products from two local stores and find that average price difference in these products is \$0.10, with standard deviation \$0.20. To decide if this is a statistically significant average price difference, suppose you simulate selecting random samples of 30 products from a normal distribution with mean 0 and standard deviation of 0.20, compute the sample mean, and then repeat this process 1000 times, obtaining the following results.



(a) Specify the observational units in this graph and provide an appropriate label for the horizontal axis.
(b) Use the Central Limit Theorem to determine the theoretical standard deviation of this distribution. Does your result seem consistent with the above graph? Explain.
(c) Using these simulation results, would you consider \$0.10 a surprising average price difference if the population mean price difference was zero? Explain.
(d) What conclusion would you come to about the average price difference of all the products in these two stores? Explain.
(e) What part, if any, of the above analysis depends on the population following a normal distribution? Explain.

32. Exponential Models
Consider the exponential model, with probability density function $f(x) = (1/\beta)e^{-(x/\beta)}$ for $x > 0$. First consider the model with $\beta = 1$.
(a) Write out and sketch the pdf for this model with $\beta = 1$.
Another function that can be used to describe a probability model is a cumulative distribution function (cdf). The cdf is denoted by $F(x)$ and is defined to be the function that reports the probability that the random variable is less than or equal to the input of the function: $F(x) = P(X \leq x)$.
(b) Determine and sketch a well-labeled graph of the cdf of the exponential model with $\beta = 1$. [*Hint*: What is the functional form of $P(X \leq x)$ for all values of x?]
The median of a continuous probability model is defined to be a value m such that $P(X \leq m) = 0.5$ and $P(X \geq m) = 0.5$.

(c) Use the cdf to determine the median of the exponential model with λ=1. [*Hint*: Set F(m) = 0.5 and solve for m.]

The mean, or **expected value**, of a continuous probability model, denoted as either E(X) or μ, is defined by $\mu = -\int_{-\infty}^{\infty} xf(x)dx,$ where $f(x)$ is the probability density function.

(d) Verify that the mean of the exponential model with β =1 is 1. [*Hint*: Use integration by parts.]

(e) How does the median compare to the mean for this exponential model? Explain why this makes sense, based on the shape of the density function.

(f) Use a Minitab or R simulation to verify your results. First simulate 1000 values from this exponential model.

Minitab	**R**

`MTB> rand 1000 c1;`
`SUBC> expo 1.`
or select **Calc > Random Data > Exponential** and ask for 1000 rows in C1 with a scale parameter of 1 and a threshold parameter of 0.
Note: scale = mean
	`> mydata = rexp(n=1000, rate = 1)`
	Note: rate = 1/mean

Then examine a histogram of the generated values and calculate descriptive statistics. Does the histogram follow the same shape as the density function? Do the median and mean values come close to your theoretical analysis?

33. Exponential Probability Models (cont.)

Reconsider the previous question about the exponential probability model with parameter β =1. Now consider the general exponential model with parameter β.

(a) Determine and sketch a well-labeled graph of the cumulative distribution function.

(b) Determine the median.

(c) Verify that the mean equals the parameter β.

(d) How do the mean and median compare?

(e) Show that the ratio of mean to median is constant regardless of β.

(f) Choose two different values of β (other than 1), and use a simulation to verify your findings. (Include a histogram and descriptive statistics of your generated distributions.)

34. Probability Density Functions

Consider the probability density function (model) for a random variable X given by

$$f(x) = (1+\theta x)/2 \text{ for } -1< x <1 \text{ and } f(x) = 0 \text{ otherwise,}$$

where θ is a parameter restricted to satisfy $-1 \le \theta \le 1$.

(a) Sketch well-labeled graphs of this function when $\theta = 1$, when $\theta = 0$, and when $\theta = -1/2$.

(b) Verify that for any value of θ satisfying $-1 \le \theta \le 1$, the total area under the density curve does equal one.

(c) Explain why this function does not produce a legitimate probability model for values of θ not satisfying $-1 \le \theta \le 1$. [*Hint*: Drawing some sketches of the function for values of θ outside of that interval might be helpful.]

(d) Evaluate $f(0)$. Does this represent the probability of X equaling zero? Explain.

(e) Determine the expected value μ of this model in terms of θ. [*Hint*: Refer to Exercise 32 for the definition of expected value of a continuous probability model.]

35. Uniform Probability Models

A uniform probability model is one whose probability density function is constant (flat) between two endpoints. Let's call the endpoints a and b, where $a < b$. So the pdf has the form $f(x) = k$ when $a \le x \le b$, 0 otherwise, where k is the appropriate constant. For example, the times at which calls are made to a computer help line in a particular hour period could follow a uniform distribution (0, 60) if they are equally likely to occur at any time in that hour period.

(a) Sketch and label a general uniform(a, b) distribution pdf and determine the constant k, as a function of a and b, so that the total area under the density equals one.

(b) Use integration to determine the expected value μ of the uniform distribution. [*Hint*: Refer to Exercise 32 for the definition of expected value of a continuous probability model.]

(c) Interpret this value geometrically (in other words, where in the interval from a to b does the mean value fall). Explain why this makes sense.

(d) It can be shown that the standard deviation of this uniform distribution is the square root of $(b–a)^2/12$. Determine the standard deviation of a uniform distribution on the interval (0, 2), on the interval (0, 10), and on the interval (8, 10).

(e) Explain why the relative values of these three standard deviations make sense.

36. House Prices

Cal Poly students Peter Cerussi and Patrick Ziegler were interested in studying factors that are related to the price of a house. They gathered data from realestate.com on the listed prices of houses for sale in San Luis Obispo, California on November 20, 2003. The prices of eight houses are shown below, and are in the houseprices.xls Excel file.

Price (in $K): 255, 349, 399, 460, 545, 649, 799, 1195

You will now consider other criteria based on the absolute deviations between the data values and your guess. Even if you keep absolute deviations as your basis for a minimization criterion, you can consider functions other than the sum. For example, if you want to be sure that you are never too far off, you might want to minimize the *maximum* of those absolute deviations:

$$MAXAD(m) = \max\{|255-m|, |349-m|, |399-m|, |469-m|, |545-m|, |649-m|, |799-m|, |1195-m|\}.$$

(a) Use the Excel file to investigate the behavior of this MAXAD function. Return the data values (house prices) in column A to their original values, and click on cell E2. Notice that this cell contains a formula for evaluating the MAXAD function. Use the "fill down" feature to evaluate this function for the rest of the m values. Then use Excel to draw a graph of the MAXAD function. Describe its behavior, and comment on whether it has a unique minimum value. Identify where the minimum occurs and what that minimum value is.

(b) Change the maximum house price from 1195 to 895 thousand dollars. Comment on the impact of this change on the MAXAD function and especially on where the function is minimized.

(c) Change the fourth house's price from 469 to 529 thousand dollars, and reevaluate the MAXAD function. Now what has changed, and what has not?

(d) Now change the cheapest house's price from 255 to 305 thousand dollars, and reevaluate the MAXAD function. Now what has changed and what has not?

(e) Based on this analysis, make a conjecture for determining the value that will minimize the maximum of absolute deviations from the mean of the data values.

37. House Prices (cont.)

Reconsider the previous Exercise and the houseprices.xls Excel file. Consider a measure of spread based on absolute deviations: minimizing the *median* of them. Let the function *MEDAD* be defined as:

$$MEDAD(m) = \text{median}\{|255-m|, |349-m|, |399-m|, |469-m|, |545-m|, |649-m|, |799-m|, |1195-m|\}.$$

Use Excel to investigate the behavior of this *MEDAD* function. In particular, describe its shape, identify where the function is minimized for the house prices data, and comment on the effects of changing the maximum, middle, and minimum values on the function.

38. House Prices (cont.)
Reconsider the previous Exercise and the houseprices.xls Excel file. You have already investigated finding a prediction that minimizes the sum of absolute deviations and the sum of squared deviations. With the benefit of technology, we need not limit ourselves to exponents of 1 and 2, however. Use technology to examine the function *SkD*(*m*), defined as:

$$SkD(m) = \sum_{i=1}^{n} |x_i - m|^k$$

(a) First analyze this function where $k = 1.5$. Look at a sketch of the function and describe its shape. What value of m minimizes this function? Is this minimum value between those for when $k = 1$ and when $k = 2$ (the median and mean, respectively, as you found above)?
(b) Choose another value of k, repeat this analysis, and report on your results.

39. Modeling Australian Births
On December 21, 1997, a record number of births were recorded in one 24-hour period in the Mater Mothers' Hospital in Brisbane, Australia hospital. The aussiebirths.txt file includes data on time of birth, sex, and birth weight for each of the 44 babies born that day (from *The Sunday Mail*, as reported at the JSE Datasets website http://www.amstat.org/publications/jse/datasets/babyboom.txt). We want to explore the distribution of *time between births*. Note that the fourth column contains the times of the births (in minutes after midnight).
(a) Use technology to calculate the time between births. Then produce a histogram of the *time between births* variable. Describe the characteristics of this distribution.
(b) We might expect a variable such as birthweight, a biological characteristic, to follow a symmetric, mound-shaped distribution, even a normal distribution. Use technology to overlay a normal probability model on this histogram. Does the normal model do a reasonable job of describing these data?
(c) Now overlay an Exponential curve. Does this probability curve appear to be a better model for these data? Explain. [*Hint*: You may want to change the binning so the first bin starts at zero.]
(d) Suppose we think the *square root of the times between births* follow a normal distribution with mean 5.25 sqrt min and standard deviation 2.5 sqrt min. Use this model to predict how often this hospital would wait more than 80 minutes between births.

40. Modeling Australian Births (cont.)
Reconsider the previous exercise. Now we will focus on the *birth weights* of the babies.
(a) Create a timeplot of the *birthweights*. Do you see any trends in the birthweights over the 24-hour period? Is this what you would expect?
(b) Create separate dotplots or histograms and normal probability plots of the birthweights for the females and for the males. Do either of these look like a normal distribution? Is this what you would expect?
(c) Suppose we think birth weights of males at this hospital generally follow a normal distribution with mean 3375 grams and standard deviation 428 grams. How unusual would it be for a baby to be of low birth weight, 2500 grams?

41. Modeling Australian Births (cont.)
Reconsider the previous exercise.
(a) Produce a normal probability plot of the times between births. Describe how the distribution deviates from normality.
(b) Produce and describe an exponential probability plot of the times between births.
(c) Take a ln transformation of the times between births. Produce a normal probability plot of these transformed data. Does this plot suggest that a normal model might be appropriate for describing the distribution of the ln of the times between births? Explain.
(d) Take a square root transformation of the times between births. Produce a normal probability plot of these transformed data. Does this plot suggest that a normal model might be appropriate for describing the distribution of the square root of the times between births?

42. Normal Distributions?
Consider the (hypothetical) data in the first three columns of the data file GotNormal.txt.
(a) Produce a histogram for each variable, and describe the shape of each distribution.
(b) For each variable, comment on whether a normal model would seem to be appropriate, based on the histogram.
(c) For each variable, produce a normal probability plot. Comment on what these plots reveal about the appropriateness of a normal model for each variable. In particular, use these plots to describe *how* the non-normal distributions deviation from the expected behavior of a normal distribution.

43. Modeling Australian births (cont.)
Reconsider the previous exercise. Suppose we think that the times between births in a Australian hospital are well modeled by an exponential distribution with parameter $\beta = 33$ minutes and you want to determine the probability of more than 1 hour (60 minutes) transpiring between births.
(a) Write the function for the density curve with this value of β.
(b) Integrate this function to determine $P(X \geq 60)$.
(c) Use technology to confirm your calculation (scale = 33, threshold = 0).
(d) Use technology to determine how many of the 43 observed times between births were longer than 60 minutes. How does this relative frequency compare to the probability predicted by the exponential model?

44. Body Mass Index (cont.)
The data in BodyMassIndex.txt are ages (in years), weights (in kg), and heights (in cm) for a sample of adults (Heinz et al., 2003). Body mass index (BMI) is defined to be a person's weight (in kg) divided by the square of their height (in meters).
(a) Examine separate normal probability plots for the BMI values of men and women. Does the normal model appear to be appropriate for either sex? For which sex does it come closer to providing a reasonable model? [*Hint*: You may first need to recalculate the BMI values from the weights and heights.]
(b) Try several transformations (log, square root, reciprocal) of the BMI values for the two sexes separately. With each transformation, examine separate normal probability plots for men and women. For each sex, identify which transformation produces an approximately normal distribution.

45. Hypothetical Waiting Times
Suppose the data in HypoWaitTimes.txt represent the amount of time patients waited in an emergency room prior to seeing a doctor (in minutes).

(a) Produce numerical and graphical summaries of this distribution and describe what they reveal.
(a) Two different models that are often used to describe waiting times and other skewed right distributions are the "Weibull" density function and the "Lognormal" density function.
(b) Add a "Distribution Fit" to your histogram using the Weibull distribution and then the Lognormal distribution. Comment on the behavior of these models.
(c) Use probability plots to determine whether these data are better modeled by a Weibull density function or a Lognormal density function. Justify your conclusion.
(d) For the distribution you choose in (b), use the parameter estimates reported by technology and estimate the probability that a randomly selected person would have to wait more than 240 minutes at this hospital for this fitted distribution.
(e) Use this same distribution to estimate the 90th percentile of waiting times at this hospital.

46. Stock Prices

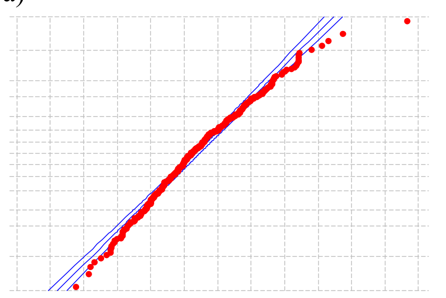
The file StockchangesOct31.txt contains the opening prices and net changes on October 31, 2001 for 3561 stocks listed on the New York Stock Exchange (nyse.com).
(a) Examine a histogram and boxplot of the opening prices. What unusual feature of this distribution is immediately apparent?
(b) Identify (by its stock exchange symbol) the stock with the largest opening price.
(c) Remove this outlier from the analysis, and then produce a histogram and boxplot of the remaining prices. Is there still an outlier that dominates these graphs? If so, identify its stock market symbol.
(d) Remove this second outlier from the analysis, and then produce a histogram and boxplot of the opening prices. Describe the distribution of opening prices now that two outliers have been removed.
(e) Examine visual displays and describe the distribution of net changes, leaving those two outlying stocks out of the analysis.
(f) Examine normal probability plots of the opening prices and net changes. Does the normal model seem to be appropriate for either variable? If not, describe how the distribution(s) deviates from normality.
(g) What percentage of the net changes fall within 1 standard deviation of the mean? Does this provide further evidence about the suitability of the normal model? Explain.
(h) Create a new variable: *percentage change*. [*Hint*: Divide the net change by the opening price and multiply by 100.] Examine visual displays, including a normal probability plot. Comment on its distribution, including whether the normal model would be appropriate for describing these percentage changes.
(i) For BRK A, the opening price was $69,800 and the net change was $1200. Calculate the percentage change for this stock. Is this percentage more than 2 standard deviations from the mean percentage change for the data set in (e)? If not, explain how this stock could be such an extreme outlier in terms of net change, but not percentage change.
(j) Repeat (i) for BRK B.

Note: Typically, when a stock price rises enough, a company will "split" the stock (each new share is worth half the value of the old shares), believing these lower-priced shares will be more attractive to investors. BRK A is the Berkshire Hathaway stock (class A) and BRK B is the Berkshire Hathaway stock (class B). Berkshire Hathaway is run by Warren Buffet, the "oracle of Omaha," who does not believe in stock splits, so the price of shares of these stocks has increased over time while other stocks increasing in value have generally split.

47. Matching Probability Plots to Boxplots

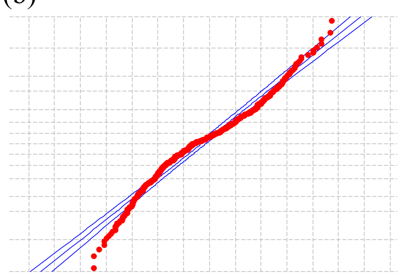
Graphs for three different variables are given below, one boxplot and one normal probability plot for each. Which boxplot corresponds to which normal probability plot? Write a few sentences providing your justification.

a)

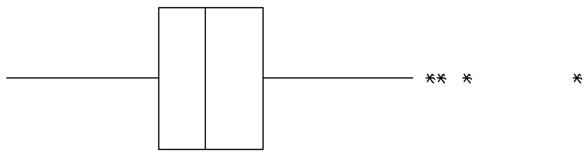


I

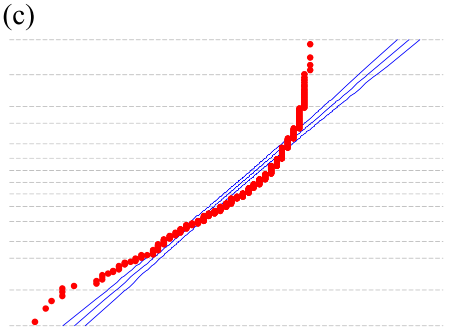
(b)



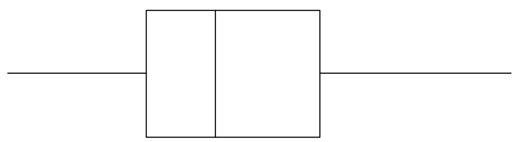
II



(c)



III



48. Backpack Weights (cont.)

Reconsider the backpack data from the Chapter 1 Exercises (backpack.txt).

(a) Construct graphical and numerical summaries to describe the distribution of the weight ratios. Comment on what this preliminary analysis reveals.

(b) Conduct a significance test of whether the sample data suggest that the mean weight ratio among all Cal Poly students is actually less than 0.10. Report hypotheses, comment on technical conditions, and calculate the test statistic and p-value. Include a well-labeled sketch of the sampling distribution for the test statistic and indicate the area represented by the p-value. Also summarize your conclusion and explain how it follows from your test.

(c) Construct and interpret a 90% confidence interval for the population mean of the weight ratios.

(d) Do you have any concerns about sampling bias or non-sampling errors with this study? Explain.

49. College Sleeping Habits

During a Monday class meeting, a statistics professor asked her students to report how much sleep (to the nearest quarter hour) they got the night before. The data are in SleepTimes.txt.

(a) Produce numerical and graphical summaries of the reported sleep times. Write a paragraph summarizing (in context) the most important features of the distribution. Use appropriate symbols to refer to your sample mean and standard deviation (and remember to include measurement units).

(b) Suppose we wanted to test whether these data provide convincing evidence that the average amount of sleep Sunday night by all students at this university is less than 8 hours. Define the parameter of interest and state appropriate null and alternative hypotheses for this research question.
(c) Suppose we are willing to consider this sample to be a representative sample of the population in (b). Roughly outline how we can carry out a simulation analysis to estimate a p-value for this research question.
(d) Open the One Variable with Sampling applet. Population 1 is sleep times for a population of 18,000 students. Describe the behavior of this population, using appropriate symbols to refer to the population mean and standard deviation (and remember to include measurement units).
(e) Check Show Sampling Options and specify a large number of samples. Also specify the Sample Size to match our study. Include a screen capture of the sampling distribution of the sample mean and summarize its behavior.
(f) Verify that this sampling distribution behaves as predicted by the Central Limit Theorem for the sample mean (do the mean and standard deviation of the distribution of the sample mean match the theoretical results)?
(g) Use this sampling distribution to estimate a p-value for our research question. Include a screen capture of your p-value output and state your conclusion in context.
(h) Now select the radio button for population 2. Repeat (d), (e), (f), and (g) for this population.
(i) Have any of the results changed much? Is this surprising? Explain.

50. College Sleeping Habits
Return to the SleepTimes.txt data from the previous exercise.
(a) Carry out a one-sample *t*-test for our null and alternative hypotheses. Include output and be sure to report the test statistic and p-value.
(b) Write a one sentence interpretation of the test statistic (in context).
(c) Write a one sentence interpretation of the p-value (in context).
(d) Does this analysis give a similar conclusion as the previous exercise?
(e) Produce **and interpret** a 95% confidence interval for the parameter. (Be sure to clarify what the parameter is in context.)
(f) How do the results differ if the low outlier (0 hours) is removed from the data set?

51. College Sleeping Habits
Return to the SleepTimes.txt data from the previous exercise.
(a) Do you believe it is valid to calculate a 95% prediction interval from these data? Justify your answer.
(b) Regardless of your answer to a) Calculate and interpret a 95% prediction interval from these data (assuming it's valid). Be sure to show your work.
(c) How does this interval compare to the 95% confidence interval from the previous exercise? Explain why the similarities and differences make sense.
(d) Suppose my sample size is really, really large. How would that impact the *width* of the 95% confidence interval? In other words, write an expression for the half-width of the interval and what that value will approach as *n* increases.
(e) Repeat (d) for the half-width of the prediction interval.
(f) Explain why the differences in (d) and (e) make sense.

52. Smoking Habits
One of the questions in the National Health and Nutrition Examination Surveys (NHANES) study asked subjects about their smoking habits. One of the questions asked whether the person has smoked at least 100 cigarettes in his/her life. The 2328 people who answered "yes" were asked to report the age at which

they started smoking. The responses are tallied in the table below and in the file SmokingStart.txt:

Age	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Count	10	6	10	23	24	99	115	155	255	195	239	377	152	192	120
Age	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Count	72	40	29	64	20	8	17	10	36	2	1	3	4	15	2
Age	**37**	**38**	**40**	**41**	**43**	**45**	**46**	**47**	**49**	**50**	**54**	**55**	**65**	**72**	
Count	4	2	9	2	4	3	1	2	1	1	1	1	1	1	

For now, consider these 2328 smokers to constitute the entire population of interest.

(a) Examine visual displays (histogram, boxplot) of the distribution of ages, and write a paragraph summarizing its features.

(b) Report the mean and median, standard deviation and IQR of these ages. Are these parameters or statistics? What symbols would you use for the mean and standard deviation?

(c) Suppose that we were to take a simple random sample of 40 people from this population of 2328 smokers. Would you expect the sample mean age to equal the population mean exactly? Explain.

(d) Does the Central Limit Theorem for a sample mean apply in this case? In other words, can the CLT tell us about the sampling distribution of the sample mean age if we were to repeatedly take random samples of size 40 from this population? If not, explain. If so, describe what it says in this case, and draw a well-labeled sketch of the sampling distribution.

(e) According to the CLT, what is the probability that the sample mean age of 40 randomly selected people from this population would exceed 20 years? (Show the details of your calculation and/or relevant output from technology.) Shade the region of interest on your sketch, and write a one-sentence summary of the probability.

(f) According to the CLT, what is the probability that the sample mean age would be less than 17.5 years? (Show the details of your calculation and/or relevant output from technology.)

(g) According to the CLT, what is the probability that the sample mean age would fall between 18 and 19 years? (Show the details of your calculation and/or relevant output from technology.)

53. Smoking Habits (cont.)

Reconsider the previous question about ages at which people started to smoke. Continue to regard those 2328 smokers as the entire population of interest, and consider taking a random sample of 40 smokers.

(a) Write and execute a simulation for taking 1000 random samples of size 40 from this population, recording the sample mean age for each sample. Construct a histogram and calculate descriptive statistics for the 1000 sample mean ages.

(b) Are your findings in (a) close to what the CLT would predict? Explain.

(c) Use your simulation results to approximate the probabilities asked for in (e)–(g) of the previous question. Comment on how closely the simulation results match those from the CLT.

54. Smoking Habits (cont.)

Reconsider the previous questions on ages at which people start smoking, but now consider the 2328 smokers to be a random sample from the population of all smokers in the U.S.

(a) Use the sample data to conduct a significance test of whether the mean age at which smokers begin to smoke differs from 18 years. Report the hypotheses in symbols and in words, comment on the technical conditions, and calculate the test statistic and p-value. Include a well-labeled sketch of the sampling distribution for the test statistic and indicate the area represented by the p-value. Also indicate whether the sample mean differs significantly from 18 at the 0.10 level, the 0.05 level, and the 0.01 level. Summarize your conclusions.

(b) Construct and interpret a 95% confidence interval for the population mean age at which smokers begin to smoke.

(c) Do you expect that about 95% of the ages in this sample fall within this interval? Would you expect that about 95% of the ages in the population of American smokers fall within this interval? Explain.
(d) Calculate a 95% prediction interval (show your methods). Provide a one-sentence interpretation of this interval in context.
(e) Do you believe the interval procedure in (e) is valid? Explain why or why not.

55. Margin-of-Error Properties

Consider the margin-of-error for a *t*-interval estimating a population mean μ: $t^* \frac{s}{\sqrt{n}}$.

(a) Explain what each of these symbols (t^*, s, n) represents.
(b) Is the margin-of-error an increasing or decreasing function of t^*, or is it neither? Is it an increasing or decreasing function of the confidence level? Explain both mathematically and intuitively.
(c) Is the margin-of-error an increasing or decreasing function of s, or is it neither? Explain both mathematically and intuitively.
(d) Is the margin-of-error an increasing or decreasing function of n, or is it neither? Explain both mathematically and intuitively.
(e) Does doubling the sample size cut the margin-of-error in half, if everything else remains the same? Explain.

56. Margin-of-Error Properties (cont.)

Reconsider the margin-of-error for a *t*-interval estimating a population mean μ: $t^* \frac{s}{\sqrt{n}}$. Suppose that you want to determine the sample size n needed for the margin-of-error not to exceed some pre-specified bound, M, at a certain confidence level.
(a) Solve for an inequality expressing the necessary sample size n as a function of t^*, s, and the error bound, M.
(b) Is this an increasing or decreasing function of t^*? Of the confidence level? Of s? Of M? Explain why your answers make intuitive sense.

57. Honda Civics
The following data pertain to a sample of 22 used Honda Civics advertised for sale on the web (Kelly Blue Book kbb.com) within 50 miles of the authors' home on August 17, 2015 (also found in the file UsedHondaCivics.txt):

ID#	age (years)	year	Type	mileage	Price	ID#	age (years)	mileage	year	Type	price
1	10	2006	Si	120451	11900	12	4	27513	2012	LX	14995
2	4	2012	LX	21136	16000	13	3	19667	2013	LX	16495
3	4	2012	EX	38422	17300	14	4	24804	2012	EX	16495
4	1	2015	LX	120	20200	15	2	13377	2014	LX	16995
5	4	2012	LX	38353	15100	16	2	14217	2014	EX	18495
6	4	2012	Hybrid	56201	16900	17	4	57309	2012	LX	13495
7	5	2011	LX	41283	13000	18	3	15970	2013	LX	16995
8	3	2013	EX	35370	16200	19	4	52115	2012	EX	14999
9	4	2012	Hybrid	39097	17300	20	2	39494	2014	LX	16995
10	8	2008	EX	76042	11000	21	1	7318	2015	EX	20495
11	7	2009	LX	72204	11500	22	2	12400	2014	LX	16995

(a) Identify the observational units with these data.

(b) Identify the five variables represented here (the *model* is not a variable here). Identify each as categorical or quantitative.
(c) Examine graphical displays and numerical summaries for the age, mileage, and price variables. Comment on the distribution of each variable in this sample.
(d) Treat these as a random sample from the population of all used Honda Civics for sale on the web that day. Would you feel comfortable applying a *t*-interval to estimate the population mean for any of these variables? For all of them? Explain.
(e) Construct and interpret a 95% confidence interval for the population mean price of used Honda Civics for sale on the web.
(f) Would you consider it appropriate to use these data to construct a prediction interval for the price of an individual Honda Civic for sale on the web that day? If not, explain. If so, construct and interpret a 95% prediction interval.
(g) How large a sample would be needed to estimate the population mean price to within ± 500 dollars with 90% confidence? (Use the standard deviation of prices in this sample as your best estimate of the population standard deviation.)
(h) Is there any sample size for which the half-width of a 90% prediction interval for price would be 500 dollars or less? Explain.

58. Breaking Ice

Nenana is a small, interior Alaskan town that holds a famous competition to predict the exact moment that "spring arrives" every year. The arrival of spring is defined to be the moment when the ice on Tanana River breaks, which is measured by a tripod erected on the ice with a trigger to an official clock. A contest is now held everyone to see who can predict the timing of this break. The minute at which the ice breaks has been recorded in every year since 1917. For example, the dates and times for the years 2000-2004 were:

2000	2001	2002	2003	2004
May 1, 10:47am	May 8, 1:00pm	May 7, 9:27pm	April 29, 6:22pm	April 24, 2:16pm

The data file `NenanaIceBreak.txt` contains all of the data from 1917 to 2004, Nenana Ice Classic. 2011. *Nenana Ice Classic: Tanana River Ice Annual Breakup Dates.* Boulder, Colorado USA: National Snow and Ice Data Center.http://dx.doi.org/10.5067/DURFYP131STS. The "date" variable is recorded in days with April 1 being coded as 1.

(a) Treat these data as a random sample from the process by which nature produces the ice-breaking dates each year. Produce a 95% confidence interval for the population mean date. Then translate the endpoints from the coded scale to the actual calendar, and interpret the interval.
(b) Produce a 95% prediction interval for the ice break-up date in an individual year. Again translate the endpoints from the coded scale to the actual calendar, and interpret the interval.
(c) Repeat (a)–(b) for the *time of day* (minutes after midnight) variable with midnight = 0.
(d) In 2015, the Tanana River officially broke up on April 24th at 2:25pm. Did either of your intervals contain this outcome?
(e) Describe a strategy for using the previous data to predict the date and time in 2015.

59. *z* vs. *t*-intervals

Some textbooks recommend that when the sample size is 30 or more, it's ok to use a *z*-interval instead of a *t*-interval, even when you have to estimate the population standard deviation σ with the sample standard deviation *s*, because the intervals do not differ too much. Investigate this recommendation in the $n = 30$ case as follows.
(a) Calculate the widths of a 95% *z*-interval and a 95% *t*-interval (in terms of *s* and *n*). Then calculate the difference in widths and divide by the width of the *t*-interval (the correct one) to determine the

percentage error in the width of the z-interval.
(b) Use simulation with the Simulating Confidence Intervals applet to compare the coverage rates of the two procedures, assuming that the population follows a normal distribution. (Use at least 1000, preferably 10,000 or more, samples to approximate the coverage rate. Choose at least two different values of the sample size to compare.)
(c) Repeat (b), but with a uniformly distributed population.
(d) Repeat (b), now with an exponentially distributed population.
(e) Summarize your findings.

60. Stock Prices

Reconsider the exercise about stock prices (StockChangesOct31.txt). Consider the 3559 stocks' opening prices (after removing the two extreme outliers as you did in the previous exercise) to be the entire population of interest.

(a) Is the population distribution symmetric or skewed?
(b) Determine the mean and standard deviation of this population. Record them with the appropriate symbols.
(c) Suppose that you take many random samples of size $n = 5$ stocks from this population and calculate the sample mean for each sample. Would you expect the sampling distribution to be as skewed as the population, less skewed than the population, or nearly symmetric? Explain.
(d) Write a simulation to take 1000 random samples of size $n = 5$ stocks from this population and to calculate the sample mean for each sample. Produce a histogram, boxplot, and normal probability plot of the sample means. Describe this distribution.
(e) Calculate the mean and standard deviation of these 1000 sample means. Are they close to what you would have expected? Explain.
(f) Repeat (b)–(e) with samples of size $n = 40$ stocks. Also comment on how this empirical sampling distribution compares to that when $n = 5$.
(g) Use the Central Limit Theorem to calculate the theoretical probability that a sample mean opening price would exceed 25, with a random sample of size $n = 40$ from this population.
(h) What proportion of your 1000 simulated sample means exceed 25? Is this close to the probability in (g)?

61. Stock Prices (cont.)

Reconsider the previous exercise, but turn your attention to the "net change" variable rather than opening price. Repeat (a)–(f) for this variable.

62. Sleeping Students (cont.)

Reconsider the data from Exercise 9, concerning the nightly sleeping times of college students over a nine-week period (SleepStudents.txt). Before analyzing the data, Amber suspected that she tended to sleep longer than either Sarah or Katherine.

(a) For each of the 63 nights, determine who got more sleep between Amber and Sarah (or if they got the same amount of sleep). Construct a bar graph to display the results.
(b) Conduct a sign test of whether the data provide strong evidence that Amber tends to get more sleep than Sarah. Report the hypotheses and p-value, and summarize your conclusion. [*Hint*: First eliminate "ties," nights for which they got the same amount of sleep, from your analysis.]
(c) Repeat (a) and (b) for comparing Amber to Katherine.
(d) If you include ties in the analysis, would it change your findings substantially? Address this question by re-running the sign test, first putting the tie on Amber's "side" and then putting it on Katherine's side. Summarize your findings.

63. Golden Rectangles
The ancient Greeks made extensive use of the "golden rectangle" in art and literature. They believed that a width-to-length ratio of 0.618 was aesthetically pleasing. Some have conjectured that American Indians used the same standard. The following data from Hand et. al. (1994) (also in shoshoni.txt) are width-to-length ratios for a sample of 20 beaded rectangles used by the Shoshoni Indians to decorate their leather goods:

0.693	0.662	0.690	0.606	0.570	0.749	0.672	0.628	0.609	0.844
0.654	0.615	0.668	0.601	0.576	0.670	0.606	0.611	0.553	0.933

(a) Produce a histogram and comment on the distribution of these ratios.
(b) Calculate the sample median of these ratios. (Note that the data are not listed in order.)
(c) Conduct a two-sided sign test of whether the sample data suggest that the population median is not 0.618. Report the hypotheses, and show how the p-value is calculated. Also summarize your conclusion.

64. Honda Civics (cont.)
Recall the data on used Honda Civics from Exercise 57 (UsedHondaCivics.txt).
(a) Examine the sample data on the "age" variable. Would a *t*-procedure be appropriate for these data? Explain.
(b) Use the bootstrap procedure to produce a 95% confidence interval for the median age in the population of all used Honda Civics for sale on the web that day.

65. Water Quality (cont.)
Return to the WaterQuality.txt data from Investigation 2.7
(a) Describe how to calculate the 10th percentile of a dataset.
(b) Create and interpret an informal ($\pm$ 2SE) bootstrap confidence interval for the 10th percentile of the population.
(c) Does the interval in (b) provide convincing evidence that the corresponding population parameter is less than 5.0 mg/l? Explain.

66. Inference Subtleties
The following questions address some finer distinctions about the inference procedures you learned in this chapter.
(a) Does the Central Limit Theorem indicate that all samples follow a normal distribution if the sample size is large enough? Explain.
(b) Suppose that the observational units in a study are people in your home state, and the variable of interest in a study is number of siblings. If the sample size is chosen to be in the thousands, would a histogram of the sample data follow a normal distribution? Explain.

67. Modeling Pregnancy Durations
According to the National Vital Statistics Reports, there were 4,130,665 live births in the United States in 2009. The report lists 30,567 pre-term deliveries, meaning the pregnancy lasted for less than 37 weeks, whereas 228,839 lasted for more than 42 weeks ("post-term deliveries"). If we want to model pregnancy durations with a normal distribution, we can use this information to determine the values of the parameters μ and σ.
(a) Of the pregnancies with known gestation periods, what proportion resulted in pre-term deliveries? What proportion resulted in post-term deliveries?

(b) Draw a well-labeled sketch of a normal curve to model these pregnancy durations, with parameters μ and σ still to be determined, but with areas corresponding roughly to the proportions calculated in (a).
(c) Determine the *z*-scores corresponding to the values 37 weeks and 42 weeks, in order for the proportions calculated in (a) to hold.
(d) Set $(37-\mu)/\sigma$ and $(42-\mu)/\sigma$ equal to these *z*-scores. Then solve this system of two equations in two unknowns for μ and σ.

68. Candy Bar Weights
Suppose that a candy bar wrapper reports the weight of the candy bar to be 1.55 ounces. Suppose that the actual weights of the candy bars vary according to a normal distribution with mean $\mu = 1.60$ ounces and standard deviation $\sigma = 0.02$ ounces.
(a) Draw a well-labeled sketch of this model for the distribution of candy bar weights.
(b) According to the model, what proportion of candy bars will weigh less than the wrapper advertises?
Now suppose that the manufacturer wants only 0.1% of the candy bars to weigh less than what the wrapper advertises. At least one of three things must change: the weight listed on the wrapper, the mean weight of the bars in the production process, or the standard deviation of the weights of the bars in the production process.
(c) To accomplish the manufacturer's goal, what weight should be listed on the wrapper, assuming that the mean and standard deviation of the weights in the production process do not change?
(d) What should the mean weight in the production process be changed to, if the weight listed on the wrapper is to remain 1.55 ounces and the standard deviation of the bar weights is not to change?
(e) What should the standard deviation of the candy bar weights in the production process be changed to, if the weight listed on the wrapper is to remain 1.55 ounces and the mean of the bar weights is not to change?
(f) Which of these three options (changing the label value, the mean, or the standard deviation) do you suspect is/are under the manufacturer's control? Explain.
(g) If the manufacturer wants only 0.01% to weigh less than advertised, in what direction would the mean and/or standard deviation need to change? Give an intuitive explanation.

69. Candy Bar Weights (cont.)
Reconsider the previous question, with the original specifications that the wrapper lists the weight as 1.55 ounces and the actual weights of the candy bars vary according to a normal distribution with mean $\mu =$ 1.60 ounces and standard deviation $\sigma = 0.02$ ounces.
(a) In a random sample of 10 candy bars, what is the probability that at least one weighs less than the advertised weight? [*Hint*: Consider the random variable Y = number of the ten bars that weigh less than advertised. What probability distribution does Y have?]
(b) If a random sample of 10 candy bars revealed that 3 weighed less than advertised, would you have reason to doubt that the production process is operating according to its specifications? Explain. [*Hint*: What is the probability of a result at least this extreme occurring by chance alone? Would you consider this result surprising?]

70. Paint Drying Time
Suppose that the drying time for a certain type of paint under specified test conditions is known to be normally distributed with mean 75 minutes and standard deviation 5 minutes. Suppose that chemists have devised a new additive that they hope will reduce the mean drying time (without changing the standard deviation). Suppose that a test is conducted to measure the drying time for a test specimen, and suppose that company executives decide that they will be convinced that the additive is effective only if the drying time on this specimen is less than 70 minutes.

(a) If the additive actually has no effect at all on the drying time, what is the probability that the company executives will mistakenly conclude that it is effective? Include a well-labeled sketch with your calculation.

Now suppose that the additive really is effective and that it reduces the mean drying time to 65 minutes, without changing the standard deviation of 5 minutes.

(b) Draw a well-labeled sketch of the two normal curves on the same scale. (You can sketch these by hand, or you can copy from technology. To get both curves to appear in the applet, check the box for the second mean and sd row and enter the second set of values.)
(c) What is the probability that this test will fail to convince the executives that the additive is effective, even though it actually is?
(d) If you want alter the cut-off value from 70 in order to reduce the error probability in (a) to 0.05, what cut-off value should you choose?
(e) Using this new cut-off value found in (d), what is the probability that the test will fail to convince the executives that the additive is effective, even though it actually is?
(f) How does the probability in (e) compare to that in (c)? Explain why this makes sense.
(g) Suppose that the additive not only reduced the mean drying time to 65 minutes but also reduced the standard deviation to 2 minutes. Re-calculate the error probability in (e). Comment on how it has changed, and explain why this makes sense.

71. Modeling the Body Mass Index

Suppose that the body mass index (BMI) of healthy American males follows a normal distribution with mean 24.5 and standard deviation 3.0 and that the BMI of healthy American females follows a normal distribution with mean 22.5 and standard deviation 3.0.

(a) Sketch (and label) these normal curves on the same scale.
(b) What proportion of healthy American males have a BMI above 25? How about females?
(c) What proportion of healthy American males have a BMI below 20? How about females?
(d) If you learn that an individual has a BMI of 19.6, would you suspect that the person is male or female? Explain.

72. Filling Cereal Boxes

Suppose that a cereal manufacturer advertises that its cereal boxes contain 16 ounces of cereal. The actual weight of the cereal put into boxes by machines follows a normal distribution with mean 16.10 ounces and standard deviation 0.08 ounces.

(a) Produce a well-labeled sketch of this distribution. (Feel free to use technology.)
(b) How many standard deviations below the mean is the advertised weight? (Show how you calculate this.)
(c) What proportion of cereal boxes will be filled with less than the advertised weight? (Also indicate the area corresponding to this proportion on your sketch.)
(d) Determine the weight such that only 1% of cereal boxes weigh less than that weight.

Now suppose that the company executives determine that your answer to (c) is an unacceptably large proportion of boxes that weigh less than advertised. They want to adjust the process of putting cereal into boxes so that only 1% of cereal boxes weigh less than the advertised weight.

(e) Determine the z-score for which only 1% of the values in a normal distribution fall below that z-score.

Suppose for now that the standard deviation of the box-filling weights is not to be changed from 0.08 ounces.

(f) What value of the mean weight should be used in order to achieve their goal that only 1% of cereal boxes weigh less than the advertised weight? (Show/explain your work in this calculation. Make use of your answer to part (e).)
(g) How does this adjusted mean weight compare to the original mean? Explain why the company

executives might be displeased about adjusting the mean weight of the box-filling process in this way.
Now suppose that the company executives are unwilling to increase the mean weight of the box-filling process from 16.10 ounces.
(h) What value of the standard deviation (SD) should be used in order to achieve their goal that only 1% of cereal boxes weigh less than the advertised weight? (Show/explain your work in this calculation. Make use of your answer to part (e).)
(i) How does this adjusted SD compare to the original SD? By what percentage does the SD need to be decreased in order to achieve the goal?
Now suppose that the company executives want to have only 0.5% of cereal boxes weigh less than advertised, and they also want only 0.5% of cereal boxes to weigh more than 16.2 ounces.
(j) Determine the values of the mean and SD that achieve this pair of goals. Show/explain your work in this calculation.
(k) Will this pair of goals require an even more precise (less variable) box-filling process than the previous goal? Explain.

73. Water Quality (cont.)
Reconsider the water quality data from Investigation 2.7
(a) Calculate and interpret a 95% confidence interval based on the "compliant" and "non-compliant" results.
(b) Calculate and interpret a 95% confidence interval based on the dissolved oxygen levels.
(c) Explain to someone who has not taken statistics the difference between these two interval procedures and the interpretations.

74. Water Quality (cont.)
Recall the dissolved oxygen study from Investigation 2.7. Suppose the investigators decided they wanted to be able to detect a non-attainment rate of 25% and that they wanted the Type I and Type II error rates, considering this alternative value of 25%, to be reasonably similar.
(a) Identify what decisions would be represented by a Type I error and by a Type II error in this context. Also, describe possible consequences from each type of error.
(b) Suppose the quality assessment manager states that the probability of a Type I error can be at most 0.15. What is the "cutoff" value for the rejection region? That is, find the smallest x so that $P(X \geq x) \leq 0.15$ when $\pi = 0.10$.
(c) Was the observed sample result (19 out of 34) in this rejection region?
(d) What is the probability of a Type II error for the cutoff value in (b) and the alternative of $\pi_a = 0.25$? Does this false negative rate appear reasonable, considering the investigator's view that the two types of errors are equally serious?
(e) How would the error rates in (b) and (d) change if instead we made the cutoff 5 or more? Explain intuitively, and then confirm your answer with appropriate calculations.
(f) Would the cutoff suggested in (e) be appropriate if the investigators considered a Type I error to be more serious than a Type II error, or vice versa? Explain.
(g) Suppose you wanted to do a similar power analysis focusing on the long-run dissolved oxygen level in the river. What additional information would you need to know?

75. More bootstrap exploration
For most statistics, the bootstrap distribution also provides information about the shape of the sampling distribution. When the bootstrap distribution is slightly skewed, this often indicates skewness in the sampling distribution as well. When the sampling distribution is skewed, we might prefer a confidence interval that reflects that skewness rather than use a symmetric confidence interval.

Consider the following notation:

- θ, the population parameter
- $\hat{\theta}$, the sample estimate of θ
- $\hat{\theta}^*_i$, the i^{th} bootstrap sample estimate of θ.

To create a 95% confidence interval for θ based on the sample estimate $\hat{\theta}$, we need to determine the distance that we plausibly expect $\hat{\theta}$ to fall from θ (at the 5% level). If we knew the critical values c_1 and c_2 such that $P(c_1 \leq (\hat{\theta} - \theta) \leq c_2) = 0.95$, then we could rearrange the inequalities as follows: $P(\hat{\theta} - c_2 \leq \theta \leq \hat{\theta} - c_1) = .95$, to produce a 95% confidence interval for θ.

However, we don't know the theoretical sampling distribution of $\hat{\theta}$ (or of $\hat{\theta} - \theta$), but the key result for bootstrapping is the "plug-in principle": The distribution of $\hat{\theta} - \theta$ is closely approximated by the distribution of $\hat{\theta}^* - \hat{\theta}$. In particular, the percentiles of these two distributions match. So the 97.5th percentile of the distribution of $\hat{\theta} - \theta$ is the same as the 97.5th percentile of the distribution of $\hat{\theta}^* - \hat{\theta}$. Fortunately, we can find the 97.5th percentile of $\hat{\theta}^*$ from the bootstrap sampling distribution and then subtract $\hat{\theta}$ to find the 97.5th percentile of $\hat{\theta}^* - \hat{\theta}$.

(a) Notice that $P(\hat{\theta}^*_{.025} \leq \hat{\theta}^* \leq \hat{\theta}^*_{.975}) = .95$. Use algebra and the key result to obtain a formula for a 95% confidence interval for θ that we can calculate from the sample data (i.e., uses $\hat{\theta}$, $\hat{\theta}^*_{.025}$, and $\hat{\theta}^*_{.975}$).

(b) Apply this formula to determine a 95% confidence interval for the population median treatment time for the heroin data in Investigation 2.9. Compare your interval to that obtained in Investigation 2.9.

76. NBA Salaries

The file NBASalaries2014.xls contains season salaries (in millions of dollars) for all NBA basketball players at the start of the 2014-15 season (posted Oct 16, 2014; downloaded from BallnRoll, July 2015). Players without a team affiliation or without a listed salary were not included in the data set.

(a) Paste the data into the Sampling from a Finite Population applet and copy and paste the histogram of this *population* distribution. Summarize the shape, center, and spread. What *symbols* would you use to refer to the mean and to the standard deviation?

(b) Which do you expect is larger, the mean or the median salary? Why?

(c) The inter-quartile range is 4.562. What are the units of this calculation? Provide a one-sentence interpretation of this number in this context.

(d) Use the applet to draw one sample of 5 players from this population. Report the mean and standard deviation. What *symbols* could you use to refer to these values?

(e) Use the applet to draw 1000 samples of 5 players from this population. Include a screen capture of the histogram of sample means. Explain what this is a distribution of (what is the variable?). Report the mean and standard deviation. Does the distribution appear to be approximately normal?

(f) Use the applet to draw 1000 samples of 10 players from this population. Include a screen capture of the histogram of sample means. How has the standard deviation changed (larger, smaller, or about the same) from (e)? How has the shape changed? Overlay the Normal Distribution – does this appear to be a reasonable prediction of the behavior of this distribution?)

(g) Repeat (f) using samples of 20 players.

(h) Use your distribution in (g) to count how many samples are at least 6 (million dollars).

(i) Use the Central Limit Theorem for sample means to assume a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$. Use this normal distribution (and technology) to approximate the probability that the sample mean is at least 6 (million dollars). Include your work and write a one-sentence interpretation of your result.

(j) Using the appropriate normal approximation and the sample mean you found in (d), approximate a 95% confidence interval for μ using the 2SD short-cut. Show your work. Does your interval include μ?

(k) Assuming the normal model is appropriate, roughly what percentage of the class do you expect to obtain an interval that does contain μ?