Chapter 1
Randomization Tests: Schistosomiasis

Activity Solutions

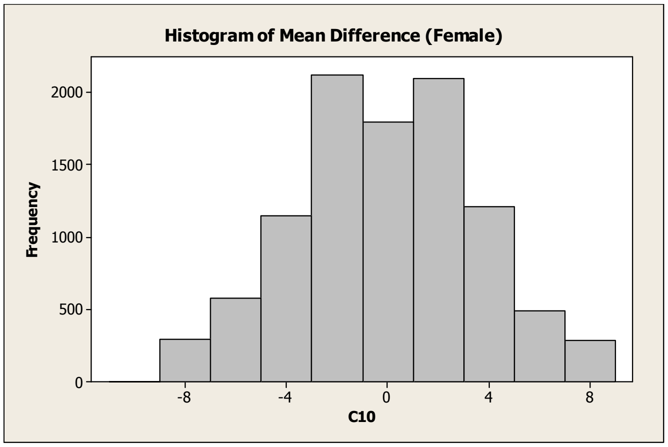
1. The male control group tends to have more worms than any other group. This group also has a much larger variance than the other groups. For both males and females, the control group has more worms than the treatment group does. There are no clear outliers (possibly the worm count of 50; however, the spread seems somewhat reasonable for that group).

2.

```
Variable      Mean  StDev  Minimum     Q1  Median     Q3  Maximum
Female-Trt    4.40   3.91     1.00   1.50    2.00   8.50    10.00
Female-Ctl   12.00   4.30     7.00   8.50   10.00  16.50    17.00
Male-Trt      6.60   2.88     3.00   4.00    6.00   9.50    10.00
Male-Ctl     29.00  12.19    13.00  19.50   28.00  39.00    47.00
```

3-8. Answers will vary.

9-10. Answers may vary, but after running the macro for 10,000 times, we obtained 284 mean differences greater than or equal to 7.6, which translates to an empirical *p*-value of 284/10,000 = 0.0284. Simulated results should be close to the exact *p*-value, which based on all 252 permutations is 7/252 = 0.02778.



This diagram looks fairly symmetric with respect to the vertical line passing through zero. Normal distribution may fit, except that the diagram should have theoretically seen its peak around zero.

11. As our empirical *p*-value from Question 10 (0.0284) indicates, the probability of obtaining a mean difference greater than or equal to 7.6 among the female mice by random allocation alone is unlikely.

12. The answer from Question 11 leads us to believe that K11777 has a positive, inhibitory effect on the schistosome worm in female mice. K11777 reduces the mean number of worms in female mice. This is because random allocation alone would produce a mean group difference larger than or equal to 7.6 only less than 3% of the time, which suggests that something other than chance accounts for the difference in group means. Since the only other distinction is whether or not the mice were treated, it is logical to say that K11777 is effective in reducing the number of worms in female mice.

13. Answers will vary, but the simulated *p*-value should be close to the exact *p*-value of 14/252= 0.5556.

14. No, the two values will not be identical since the distribution of the mean differences will not be perfectly symmetric about 0.

15. The distribution is roughly symmetric.

16. Small sample sizes of size 5 female mice each.

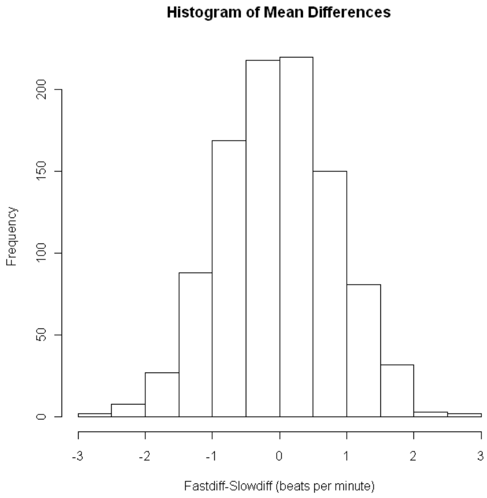
Extended Activity Solutions

17. Answers may vary, but using 10,000 repetitions, we obtained the one-sided *p*-value= 0.0486 through a simulation. Note that the exact one-sided p-value is 6/120 = 0.05. Thus, it is not very likely that the mean difference would be at least as great as the one observed by random chance alone.

18. Answers may vary, but using 10,000 repetitions, we obtained the one-sided *p*-value=.193 through simulation. Note that the exact one-sided p-value is 20/120 = 0.1667. Thus, it is fairly likely that random chance could explain the difference at least as great the one we observed. Based on the results of the permutation tests in Questions 17-18, the mean age of those laid off appears to be greater than the mean age of those who were not laid off; however, the median ages of those who were laid off do not appear to be greater than the median age of those who were not laid off.

19. They should not allow the data to “suggest” a particular direction for the effect of fast music on pulse rates.

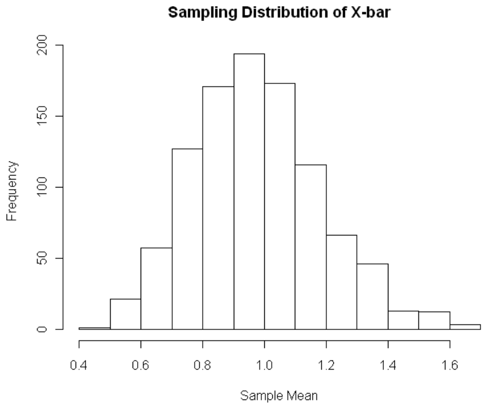
20. Refer to the R or Minitab instructions for program or macro. Note that answers may vary slightly from ours.



a) You can shade the area under the histogram to the right of 1.86. This represents the *p*-value for the test. We obtained a *p*-value of 0.016.

b) Based on the *p*-value of 0.016, we can conclude that listening to fast music increased the average pulse rate more than listening to slow music.

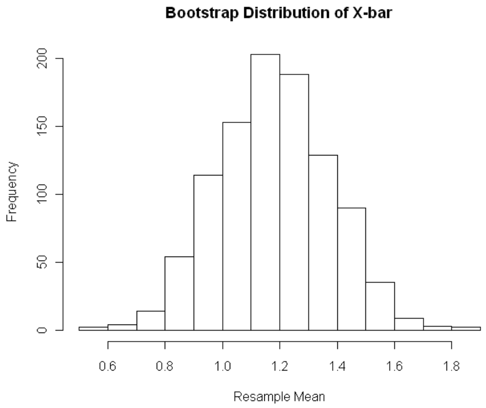
21. a) Refer to the R or Minitab instructions for the code to produce the histogram of the sampling distribution of the sample mean. Note that answers may vary slightly from ours. We get:



b) The central limit theorem tells us that the mean of the sampling distribution is 0.974, the standard deviation is $1.3153/\sqrt{40} = .208$, and the shape should be approximately normal.

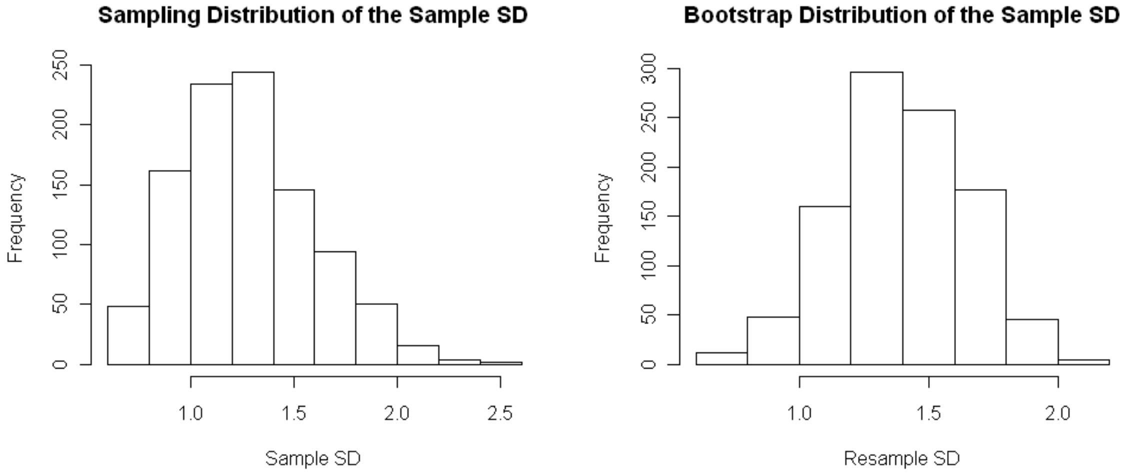
c) Based on our sampling distribution, the mean is .977 and the standard deviation is .2058. Both values are close to what we would expect by the CLT.

22. b) Refer to the R or Minitab instructions for the code to produce the histogram of the bootstrap distribution. Note that answers may vary slightly from ours. We get:



c) The shape of the bootstrap distribution is a bit more symmetric than the sampling distribution, but the mean and the standard deviation of the bootstrap distribution are 1.18 and .196, respectively.

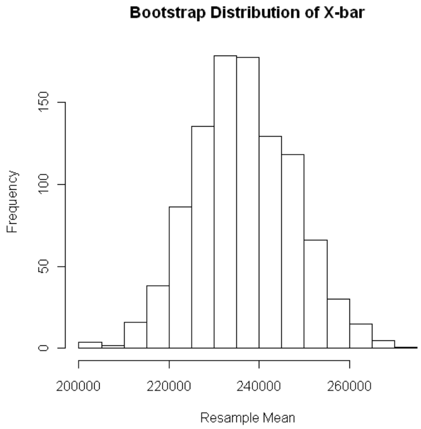
23. Refer to the R or Minitab instructions for the code to produce histograms of the sampling distribution and bootstrap distribution of the sample standard deviation. Note that answers may vary slightly from ours.



The bootstrap distribution is more symmetric than the sampling distribution, which is slightly right-skewed. The mean and standard deviation of the sampling distribution are 1.276 and 0.327, respectively. The mean and standard deviation of the bootstrap distribution are 1.394 and 0.257, respectively.

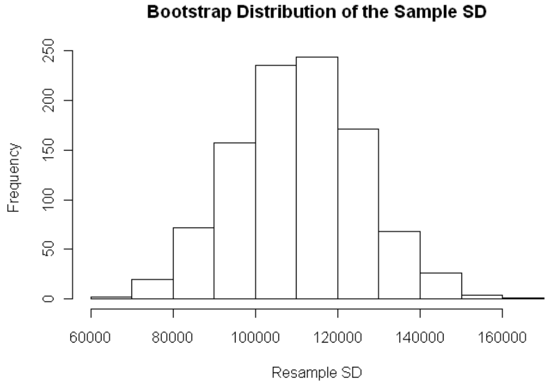
24. Refer to the R or Minitab instructions for constructing the bootstrap distributions and confidence intervals. Note that answers may vary slightly from ours.

a) Bootstrap distribution for the sample mean salary:



95% Bootstrap t Confidence Interval for the population mean salary: (214,486.0, 258,666.2)
95% Bootstrap Percentile Confidence Interval for the population mean salary: (215,272.4, 258,801.6)

b) Bootstrap distribution for the sample standard deviation salary:



95% Bootstrap t Confidence Interval for the population standard deviation salary: (80,084.9, 140,825.9)
95% Bootstrap Percentile Confidence Interval for the population standard deviation salary:
(81,339.5, 141,038.3)

c) 95% confidence interval for the population mean salary using the t-procedure (Equation 1.2): (213,357.6, 258,028.8). This interval is similar to that found in Part (a). This is expected, since the sample size is large.

d) Equation 1.2 is derived from the fact that the sample average salary is approximately normally distributed. The same cannot be said about the sampling distribution of the sample standard deviation, so Equation 1.2 cannot be used.

e) Both bootstrap distributions are symmetric; however, the centers and spread of the distributions are different. The mean and standard deviation for the bootstrap distribution of the sample mean are 236,576.1 and 11,132.9, respectively. The mean and standard deviation for the bootstrap distribution of the sample standard deviation are 110,455.4 and 15,306.0, respectively.

25. Refer to the R or Minitab instructions for performing the Wilcoxon rank sum test. Different software will provide slightly different solutions. The *p*-value is .06, so there is marginal evidence that the distributions of salaries for pitchers and first basemen are different.

26. Different software will provide slightly different solutions. $2 \times P(W \leq 18) = 0.046$.

27. Refer to the R or Minitab instructions for performing t-tests. The two-sample t-test (not assuming equal variances) yields a *p*-value of 0.02266. The conclusion is similar to that of Question 25, although the evidence is now stronger in favor of a significant difference in the average salaries. Because of the small sample sizes and possible lack of normality of the salaries, the nonparametric procedure is more appropriate.

28. Refer to the R or Minitab instructions for performing the Kruskal-Wallis test. The *p*-value for the Kruskal-Wallis test is 0.0185, providing evidence that the distribution of the salaries differs by position.

29. Test 1: a permutation tests results in a p-value = 0.643 (probability of observing a difference of group means at least as extreme as 442.98)

Test 2: a permutation tests results in a p-value =0.001 (probability of observing a difference of group means at least as extreme as 13326.9)

Test 3: a permutation tests results in a p-value < 0.001. (probability of observing a difference of group means at least as extreme as 12883.9)

We can conclude that the prices of Cadillacs are significantly different than Pontiacs and Buicks

30.

Case	Test 1	Test 2	Test 3	Probability
1	F	F	F	0.8574
2	F	F	R	0.0451
3	F	R	F	0.0451
4	F	R	R	0.002375
5	R	F	F	0.0451
6	R	F	R	0.002375
7	R	R	F	0.002375
8	R	R	R	0.000125

31. 1-.0.9^3 = 0.271.

32. 1-.0.95^6 = 0.2649.

33. We can conclude that the prices of Cadillacs are significantly different than Pontiacs and Buicks; we did not find a difference between Pontiacs and Buicks.

34. Three of the tests still reject; Cadillacs are significantly different than other makes.

35. 1-.0.95^21 = 0.6594

Exercise Solutions

E.1 Yes, it is important that all 20 mice come from the same population of mice, if we want to say that there is some causal relationship between the treatment and the lower worm counts. If all 20 mice did not come from the same population, then the different populations could create differences in worm counts. In that case, the observed difference between groups could be due to different populations, not different treatments.

E.2 No, the results should not be trusted. If the mice were not randomly assigned to each group, then there may be some other variable that could affect the results of the study. For instance, perhaps the first five mice were less healthy, making them slower and easier to catch.

E.3 Cause and effect cannot be inferred from this study because children who watch more television may be affected by other variables that are causing their obesity. For instance, they may watch more television because they aren't involved in sports and therefore have more free time to watch television, and the lack of physical activity makes them more likely to become obese. If this were just an observational study, the subjects were not randomly allocated to groups so there is a chance of other variables influencing the results.

E.4 A random sample is a sample of data where each subject is randomly chosen from some population. A randomized experiment is where the experimental units are randomly allocated to treatment groups.

E.5 In a population model, units are selected at random from a population or populations. In a randomization model, a fixed number of experimental units are randomly allocated to treatments.

E.6 In a randomization model, there are a fixed number of samples. A fixed number (likely half) of experimental units are randomly allocated to each treatment. If there is one unusual unit in the sample and it is assigned to the first group, it will not be assigned to the second group. Thus, the two groups are not completely independent, as they are formed from a fixed collection of experimental units.

E.7 No, the data from a sample will tend to have a shape that reflects the population from which it was collected. As the sample size increases, the sample data will more closely follow the shape of the original population. There is no guarantee that the original population was normally distributed.

E.8 The distribution of the sample mean will be normally distributed. According to the central limit theorem (assuming a finite standard deviation), as the size of the sample data increases, the sample mean tends to follow a normal distribution where the sample mean is the same as the population mean, and the sample standard deviation is the population standard deviation divided by the square root of the sample size.

E.9 Graphical techniques should be used to visualize the data before a parametric test is conducted to make sure that the data follow the assumptions of the parametric test. For example, if the data are highly skewed or has outliers, the sample mean may not be normally distributed. If these assumptions are not met, then the *p*-value given by the test may not be accurate.

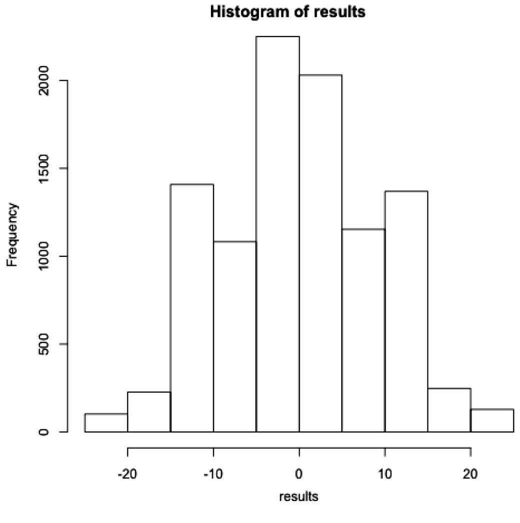
E.10 If the *p*-value in the schistosomiasis study was 0.85, then the difference in the group means would not be significant at the 0.05 significance level, and the data would be consistent with the null hypothesis. We cannot conclude that there is no difference between the treatment and control means, so we would fail to reject the null hypothesis. For instance, there could still be a difference between the treatment and control means that we failed to detect due to the sample size.

E.11 **a)** Answers will vary. In our simulation, the *p*-value obtained was 0.0825; based on this *p*-value, we would fail to reject the null hypothesis that the difference in the group medians was due to the random chance alone at the 0.05 significance level. This is different from the results obtained in Question 9, where the null hypothesis was rejected.

b) Answers will vary. In our simulation, the *p*-value obtained was 0.4285; based on this *p*-value, we would fail to reject the null hypothesis (of equivalent standard deviations) at the 0.05 significance level. It would seem that the differences in the variances of each group could be due to chance alone.

E.12 **Testing Male Mice**

a)

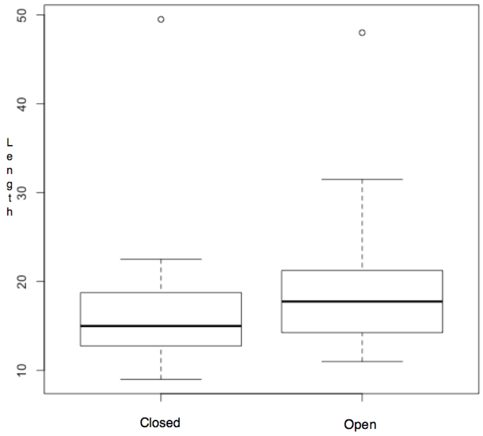


Based on a simulated the *p*-value of 0.0042 (will vary with each simulation), we would reject the null hypothesis that the difference in group means is due to random chance alone. We conclude that there is a difference in the true treatment and control means. From this, it would seem that K11777 does inhibit schistosome viability, and so would be an effective treatment for reducing worm counts in this population of mice. We can make this causal conclusion because we know that this is a randomized experiment with a group of mice that come from the same population. However, we cannot say that these results would hold for mice of a different population.

b) The *p*-value obtained in our simulation was 0.0282 (will vary slightly with each simulation). Based on this *p*-value, we would still reject the null hypothesis, but the results are not as conclusive as in Part (a).

c) The simulated *p*-value obtained was 0.4176 (will vary slightly with each simulation). Based on this *p*-value, we would fail to reject the null hypothesis. We do not have evidence to conclude that there is a difference between the true variances of the two groups.

E.13 a)

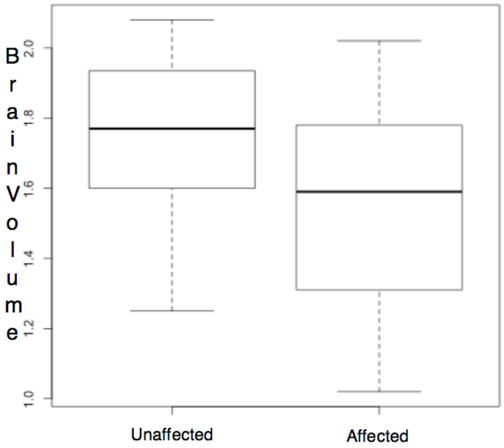


The standard deviation = 7.158 and the mean = 16.419 for the closed-nest birds. The standard deviation = 6.56, and the mean = 18.852 for birds with open nests. So, it would appear that open nests are built by larger birds, contrary to what Moore believes.

b) Excluding those birds for which there wasn't length data, the two sided *p*-value obtained was 0.0992 (answers will vary slightly). Based on this *p*-value, we would fail to reject the null hypothesis that the length of the bird varies based on the type of nest built. However, this was an observation based on available evolutionary data; due to this, even if the *p*-value was significant, we could not infer a causal relationship between length and nest type. Other variables could be affecting what type of nest the birds create, as there was no random allocation of units to groups or random sampling of the population of birds.

E.14 a) This data should be analyzed as a matched pairs design. Each pair of twins is genetically identical because they are monozygotic, thus the samples would not be independent of each other.

b)



On the left is the unaffected group and on the right is the affected group.

The mean of the brain volume for the affected group and for the unaffected group is 1.56 and 1.758667, respectively, with standard deviations of 0.3012593 and 0.2424243 respectively. Based on this, it would seem that the true mean of the unaffected group's brain volume is larger than the affected group.

c) The resulting *p*-value for the one-sided randomization test (based on 10,000 simulations) is .0023, indicating that we should reject the null hypothesis and conclude that the true mean brain volume for unaffected individuals is larger than the true mean brain volume of those individuals affected by schizophrenia.

E.15 Comparing Parametric and Nonparametric Tests

a) R-code results provide t = -1.5951, df = 55.035, *p*-value = 0.1164. Minitab results in a two-sided *p*-value = 0.116 using 55 df.

95 percent confidence interval: (-5.4879680 0.6235526)

These results are consistent with the randomization test. That is, we fail to reject the null hypothesis that there is no difference in nest type as bird length varies.

b) t = 2.28 with 27 df, *p*-value = 0.015

Using a one-sided t-test, we reject the null hypothesis at the 0.05 significance level; this result agrees with the results of the randomization test.

E.16 Means versus Medians in Rank-based Tests

a)

rank	group	number	rank	group	number	rank	group	number
1	1	1	10	2	10	19	3	19
2	1	2	11	2	11	20	3	20
3	1	3	12	2	12	21	3	21
4	1	4	13	2	13	22	3	22
5	1	5	14	2	14	23	3	23
6	1	6	15	2	15	24	3	24
7	1	7	16	2	16	25	3	25
8	1	8	17	2	17	26	3	26
9	1	9	18	2	18	27	3	27
30	1	46	29	2	37	28	3	28
31	1	47	38	2	58	45	3	65
32	1	48	39	2	59	46	3	66
33	1	49	40	2	60	47	3	67
34	1	50	41	2	61	48	3	68
35	1	51	42	2	62	49	3	69
36	1	52	43	2	63	50	3	70
37	1	53	44	2	64	51	3	71
54	1	342	53	2	193	52	3	72

The sum of the ranks of each group is 367, 495, 623 respectively for groups one, two, and three. In each group, the mean = 43.5 and the median = 27.5. From the Kruskal-Wallis test, we obtain a test statistic of 7.3553.

Using R: Kruskal-Wallis chi-squared = 7.3553, df = 2, *p*-value = 0.02528

Although we are testing a null hypothesis that the true group medians are equal and all the sample group medians are the same, the Kruskal-Wallis test provides a *p*-value that rejects the null hypothesis at the 0.05 significance level. Based on this, it is clear that the Kruskal-Wallis test is not appropriate for this data, as the standard deviations for the three groups vary greatly (77.78, 43.69, and 23.17 for groups 1, 2, and 3, respectively).

E.17 Rank Based Bird Nest Tests

a) W = 718, *p*-value = 0.02969.

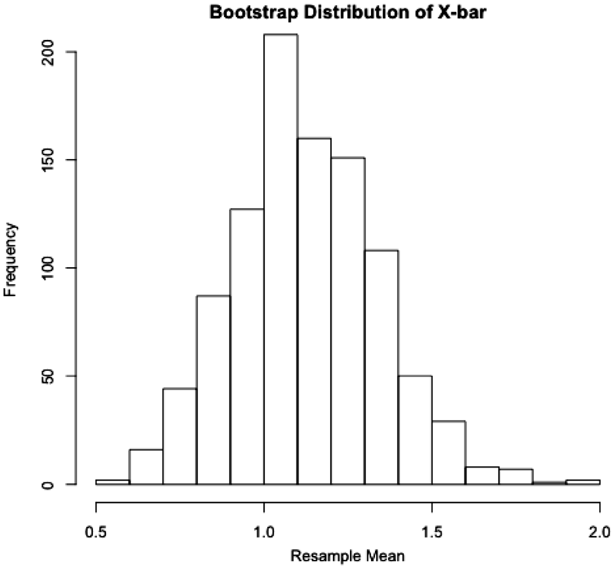
Thus, we conclude that the true location shift is not equal to 0.

b) Kruskal-Wallis chi-squared = 4.6339, df = 3, *p*-value = 0.2007 (results will vary slightly depending on how groups are formed).

Data were placed into 4 groups: cavity, cup, spherical, and everything else was put into a last group. Data for which there was no length data were excluded. Based on this *p*-value, we would fail to reject the null hypothesis that the distribution of bird size is the same for each nest type. From this, we can say that the data are consistent with the null hypothesis that bird length doesn't vary with nest type.

E.18 Bootstrap Confidence Intervals

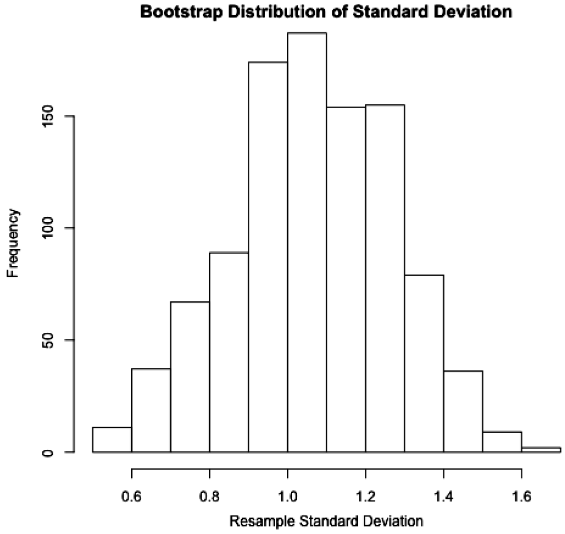
a)



The 95% bootstrap t confidence interval for the mean is (0.68765, 1.56168). Answers will vary slightly based on the simulation.

b) The 95% bootstrap percentile confidence interval for the mean: (0.720425, 1.556814). Simulated results will vary. The bootstrap distribution is still slightly right-skewed as the re-sampling was done from a strongly right-skewed data set, and so not centered on the observed statistic (0.974425). The two confidence intervals are close together, but there is still reason to question their reliability.

c)

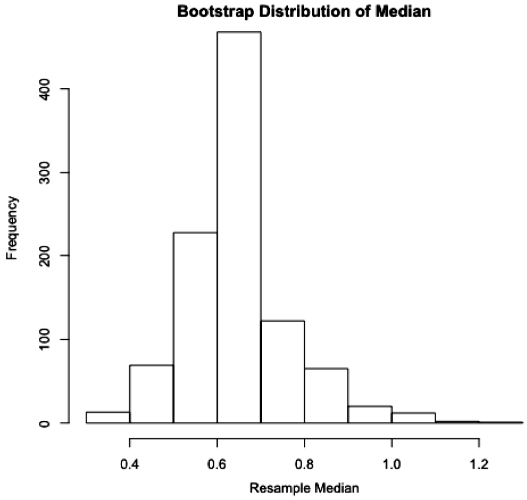


The 95% bootstrap t confidence interval for the standard deviation is (0.64888, 1.47469). Answers will vary with each simulation.

d) The 95% bootstrap percentile confidence interval for the standard deviation is (0.6547394, 1.435706). Simulated results will vary. The bootstrap distribution of the standard deviation is slightly right-skewed and not centered on the observed statistic (the standard deviation of the data set was 1.315), so these results may not be reliable.

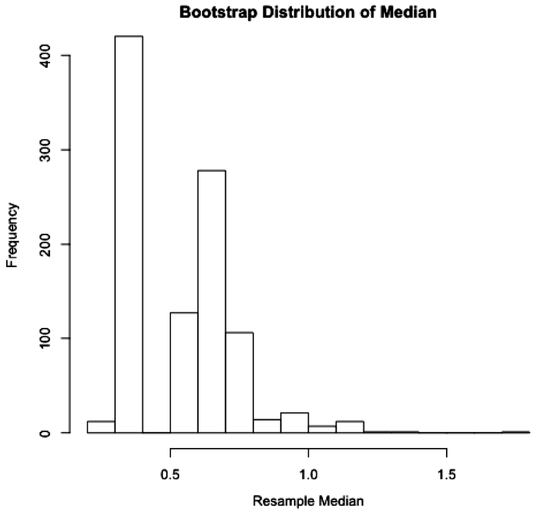
E.19 Medians and Trimmed Means in Bootstrap Confidence Intervals

a)



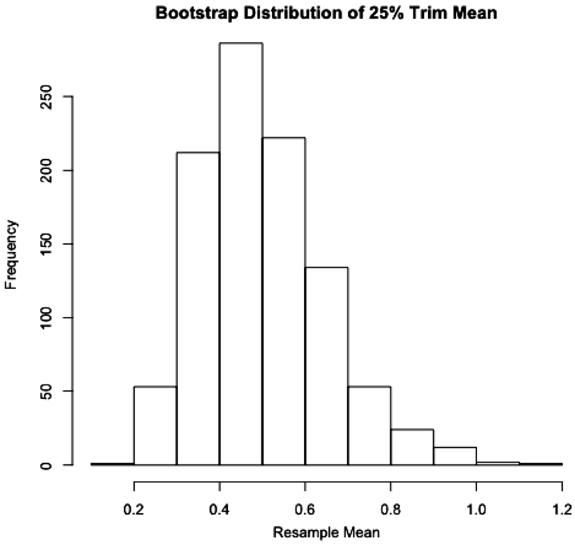
The bootstrap distribution is strongly right-skewed and biased, so bootstrap confidence intervals are unlikely to be reliable. There are very few values for the median, the data are not normal, and there are outliers.

b)



This second bootstrap distribution of the median appears to have two centers, is strongly right-skewed, is not symmetric, is not normal, and may be biased. Bootstrap distributions of medians are unlikely to be normally distributed because, for any random sample of a data set, there are only so many possibilities for the median when you take a thousand samples of that sample with re-sampling, so some values may show up many more times than others creating two peaks, but a few other values will show up creating skewness and outliers, which will cause the distribution to be non-normal.

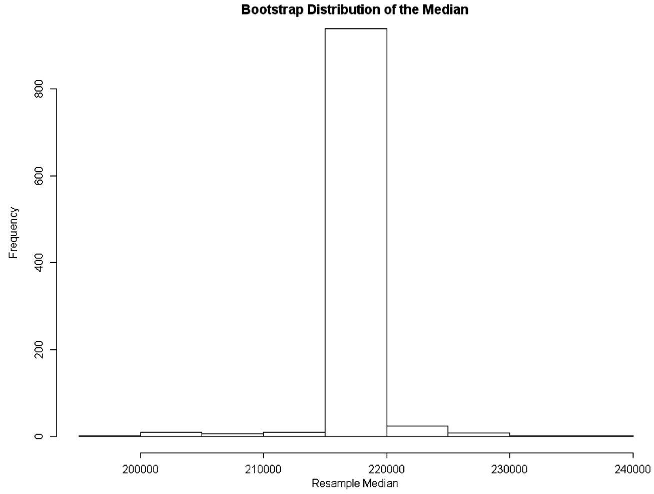
c)



The bootstrap distribution of the 25% trim mean is somewhat right-skewed, and seems slightly biased with respect to the 25% trim mean of the data set (25% trim mean is 0.5812164). The distribution is not symmetrical and does not appear normal, which suggests that the confidence intervals may not be reliable. The bootstrap t confidence interval was (0.194769, 0.8074788), and the bootstrap percentile confidence interval was (0.2720696, 0.8465589).

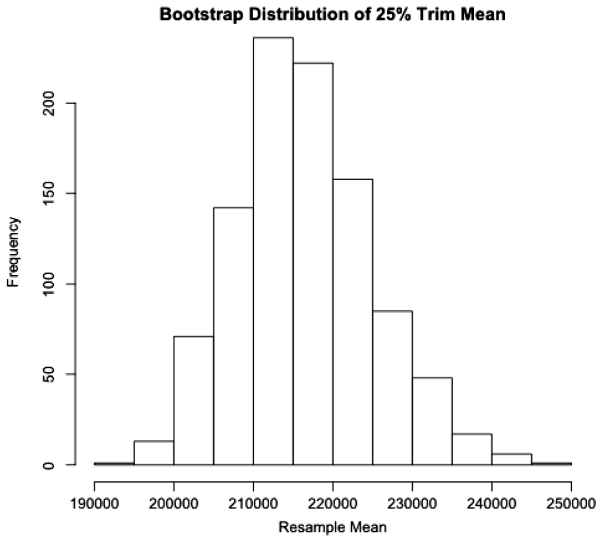
E.20 Medians and Trimmed Means in Bootstrap Confidence Intervals

a)



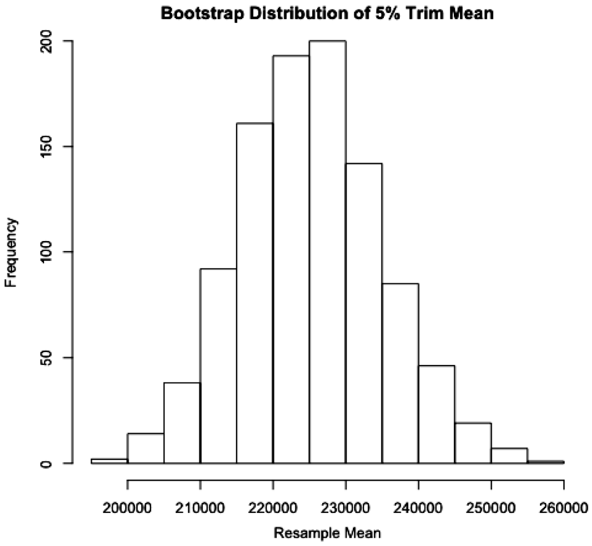
This bootstrap distribution does not have the shape of a normal distribution. Thus, confidence intervals would not be appropriate to use.

b)



The bootstrap distribution of the 25% trimmed mean appears to be slightly right-skewed, but not very biased with respect to the 25% trimmed mean of the sample. It seems that it would be appropriate to make the confidence interval.

c)



The distribution appears to be approximately normal, symmetrical, and only slightly right-skewed, but the distribution is slightly biased (the 5% trimmed mean is 224390.2 based on the random sample). However, it seems that it would be appropriate to create bootstrap confidence intervals.

d) A. Bootstrap t confidence interval (215410.6, 224091.9)
Bootstrap percentile confidence interval (213999.9, 220049.5)

B. Bootstrap t confidence interval (199391.5, 233508.6)
Bootstrap percentile confidence interval (201336.5, 234423.8)

C. Bootstrap t confidence interval (205512.3, 244870.2)
Bootstrap percentile confidence interval (206335.7, 245497.1)

E.21 Multiple Comparisons

a) The *p*-value = 0.023 (answers will vary slightly; exact solution is 0.0317). Based on this *p*-value and a significance level of 0.05, we would reject the null hypothesis that there is no difference between the salaries of pitchers and first baseman. Rather, the data are consistent with the alternative hypothesis that there is a difference between the salaries for these two groups.

b) The *p*-value obtained was 0.347 (answers will vary-exact solution is 0.3254). Based on this *p*-value, we would fail to reject the null hypothesis at the 0.05 significance level. From this, the data are consistent with the null hypothesis that there is no difference between the salaries of pitchers and catchers.

c) The *p*-value = 0.007(answers will vary-exact solution is 0.0079). From this, we would reject the null hypothesis at the 0.05 significance level. We can conclude that the data are consistent with the alternative hypothesis, and that there is a difference in the salary of these two groups.

d) If each of the aforementioned three tests uses an alpha level of 0.05, the probability that at least one of t he tests will inappropriately reject the null hypothesis is 0.1426.

e) If we use the Bonferroni method with an overall alpha level of 0.10, the critical value would be 0.0167.

Using Bonferroni methods, we would now reject the hypothesis in Part (a); conclusions from Parts (b) and (c) would not change.

Chapter 1
Nonparametric Methods: Schistosomiasis

Investigation

This chapter asks students to use nonparametric tests to determine if a new drug is helpful in reducing schistosomiasis (shis-tuh-soh-mahy-uh-sis), a disease occurring in humans caused by parasitic flatworms. Schistosomiasis affects millions of people, especially children in developing countries. The disease can cause death, but more commonly results in chronic and debilitating symptoms, caused primarily by the body's immune reaction to parasite eggs lodged in the liver, spleen, and intestines.

Goals

This chapter introduces randomization tests, permutation tests, and bootstrap methods. We demonstrate that these techniques typically require fewer assumptions and provide results that are often more accurate than those from traditional techniques (especially when the sample data are skewed, the sample size is small or when we want to conduct inference for something other than the population mean). Rank based tests, such as the Wilcoxon Rank Sum Test and the Kruskal-Wallis Test, are also discussed in the extended activities.

This chapter also reviews the key concepts of statistical inference, thus students may feel the material in this chapter moves somewhat slower than other chapters if they are already comfortable with hypothesis tests.

Suggested Schedule

Below is an aggressive outline that has been used for Chapter 1. This class met three times a week for 50 minute time periods. While it is helpful to teach this chapter in a computer lab, these materials have also been taught in rooms where only the instructor has a computer.

I require students to submit answers to all the questions from the initial activities, but grade only a few problems. Class time is used to discuss (or give solutions to) questions where several students had questions, review key concepts, and provide additional examples (often selected from the homework problems).

Day 1-2: Before class the students (often in groups of 2-3) are expected to read the introduction and complete the first two problems. We start the class by discussing the case study, then review the goals of hypothesis testing and discuss why the traditional two-sample t-test is not appropriate for this study. Students (typically in groups) work through Questions 3-6 in class and answers are recorded on the board (or on the instructor's computer) but not graded. *Bring 3x5 note cards to class for students to use.*

While Questions 3-6 are rather simplistic and time consuming, I always take the time to work through these in detail in class. Many students leave the introductory statistics course not clearly understanding the key concepts behind statistical inference. These four questions provide a concrete example that can be referred to throughout the course to reiterate the true meaning of these concepts.

If time allows we will also work through Questions 7-12 in class. Even though step by step instructions are given, *this is one place where students may become easily frustrated*, since many of my students are not familiar with writing even a small computer program. Many students may not complete all the problems in class, so this is one time that I am careful to give time outside of class where I will be available to answer questions.

Note: If you are starting this chapter the first day of class, it may be best to briefly introduce the research question and have the students work through problems 3-6 in class. Then have the students catch up on the reading before the second day.

Day 2: The second day is used to review the simulation in Questions 7-12. I use the instructor's computer and projector to walk through the meaning of each step in the short program and ask how changing certain aspects of the program will impact the results. Often I will allow 15 minutes at the end of class to help any student groups who could not get their programs to work. If class is conducted in a computer lab, students who did get their programs to work can use class time to continue to work through the rest of the chapter.

Day 3-4: Student groups submit Questions 13-18 at the beginning of class. We spend 10-15 minutes comparing and interpreting p-values from their simulations and the two-sample t-tests. I also spend time reviewing the concepts of random allocation and random sampling. If time allows I lecture with new examples (often using one of the end-of-chapter exercises), discuss the extended activities, or allow students to work through a few of the extended activities. At this point I have students read through the gender discrimination research project and have them start collecting data on their university of choice.

Day 4-5: I typically allow 1-2 days to discuss a few topics from the extended activities, particularly Section 1.13 (Multiple Comparisons) and Section 1.9 (Using Bootstrap Methods to Create Confidence Intervals). While not required, the multiple comparison activities are helpful for students to work through before they complete the project. I often briefly discuss other nonparametric techniques without requiring students to work through all the questions. While students work on these activities, they are also expected to be collecting data, conducting an analysis and preparing their report.

Day 6: Student groups submit their 3-page summary report for their discrimination research project. In addition, at the beginning of class students are expected to submit the first few questions of the next chapter (Chapter 2 or other chapter). Class discussion focuses on Chapter 2.

Extended Activities

Some of the optional extended activities are more complex than the questions asked in the schistosomiasis study.

- Sections 1.6 and 1.7 can be completed as out of class homework problems.

- Often I lecture through Sections 1.8 -1.10 instead of having students work through these activities on their own. Students tend to struggle with these sections.
- Traditional non-parametric techniques (such as *Wilcoxon* and *Kruskal-Wallis*) are very useful to be aware of because they are often used in research. However, these techniques are somewhat ancillary to the goals of this chapter and after students have completed this course, they usually have little difficulty understanding these techniques on their own.
- Section 1.13 (Multiple Comparisons) is useful to work through before conducting the project, but not required

Research Project

While a data set labeled `faculty` is available for this project, students are much more interested in this project when they find their own data based on a local university. The `faculty` data set contains messy data. Expect students to ask why particular outliers exist. The one female English instructor earning $178,798 was working as an associate dean while still considered part of her original department. While information on other faculty was not completely available, it seems reasonable that some are emeritus or have part time teaching positions.

Students will find it difficult to limit their report three pages (including graphs); they tend to struggle with succinctly presenting their results. This is a very useful exercise in encouraging students to clearly and concisely communicate key results of a study. I point out that while academics sometimes value very thorough lengthy documents, in my experience working as an industry consultant, all the top executives that I know prefer to be given short summary documents that help them understand the key points on the issue.

Since this is their first project, I tend to make it shorter and worth fewer points than later projects. While it would be useful to give students an opportunity to rewrite this document, instead I often lecture on common errors in the assignment and provide an example of a well written summary paper. The following can be used as a model solution in class.

SAMPLE PROJECT SOLUTION

Chapter 1 Project: Faculty Salaries

Conclusion:

We found no evidence of gender discrimination in the salaries paid to the regular, full-time faculty in the English and statistics departments.

Method:

We examined only regular, full-time faculty; the following persons were excluded from the analysis:

- The three lecturers and adjuncts in the English department. Without knowing details about each of these instructors' teaching loads and contracts, it is difficult to determine if there is gender bias in their pay.

- The English professor earning $178,798, although a member of the English department, is currently serving as an associate provost, and presumably her salary reflects her administrative position rather than her role as an English professor.

- Three faculty members (one in statistics and two in English) for whom salary information was not provided.

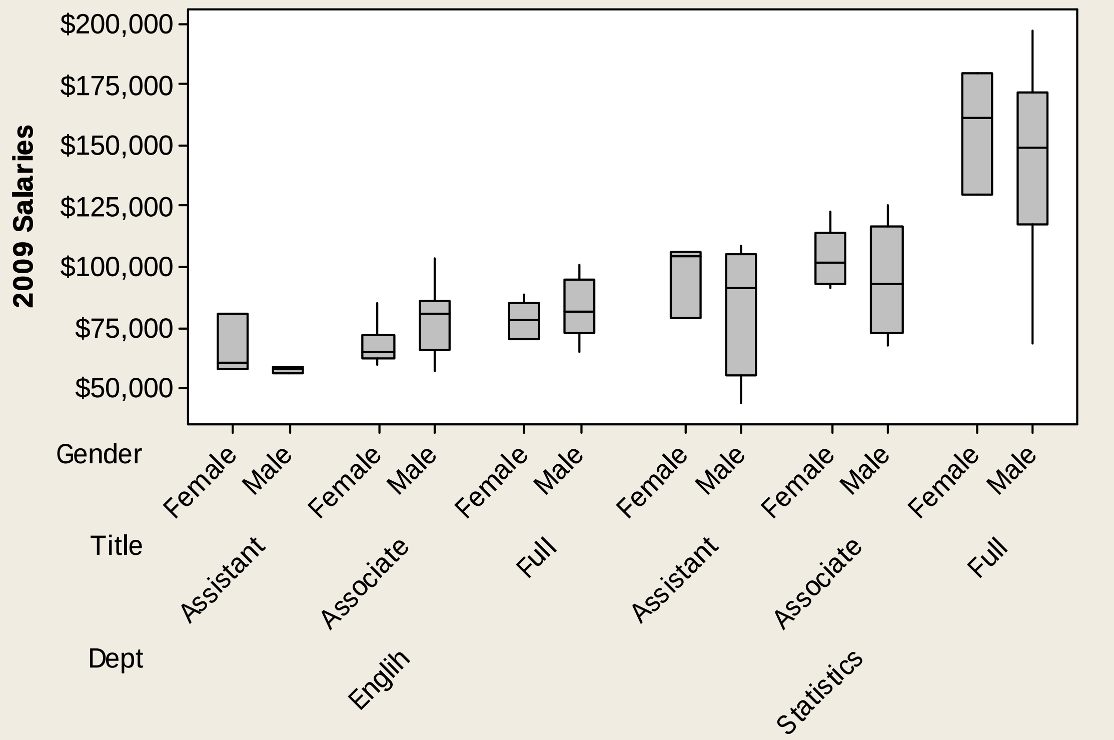
- The male full professor, whose salary was $21,268. This salary is so low that he was presumably not acting as a full-time member of the faculty.

"Distinguished professors" and "university professors" were grouped with full professors for the purposes of this analysis.

Figure 1 graphs the remaining salary data by gender, faculty rank (title), and department. These categories were chosen because faculty rank and department are likely to explain much of the variations in salaries. It is not valid to compare the salaries of male full professors in the statistics department with the salaries of female assistant professors in the English department, and then conclude that the males' higher salaries are due to gender.

Figure 1 shows that the members of the male-dominated statistics department are paid more than the members of the female-dominated English department, leading to an overall higher average salary for the male professors ($106,040 for male professors versus $90,899 for female professors). However, within the six department and faculty rank combinations there is no clear pattern showing that male professors out-earn the female professors .

Figure 1: Boxplots of Faculty Salaries by Gender, Title and Department



Two-sided randomization tests were used to determine if gender differences in average salaries could be explained by random chance. Two-sided tests were chosen to account for the possibility that there may have been gender bias in *favor* of female professors; a one-sided test would only have shown statistical significance in the cases where the male professors' salaries were significantly higher than the female professors' salaries, whereas a two-sided test would also show statistical significance if the female professors' salaries were significantly higher than the males. Randomization tests were chosen because the small number of faculty in each group ruled out the use of a t-test.

Table 1 shows the results of the randomization tests conducted for each of the six department and faculty rank combinations. In each test, the salaries were randomly assigned to either a male or a female (in the same proportion as the actual members of the department), and then the difference between the average male and female salaries were calculated. After this was done 10,000 times, the p-value represents the percentage of times where the absolute value of the difference exceeded the absolute value of the actual difference. This p-value estimates how often the actual difference would have occurred by random chance alone.

Results:

Although the average female and average male salaries do differ in each of the six department and faculty rank combinations, the difference was not statistically significant for any of the six. Thus in each test, random chance could have explained the observed differences between salaries for males and females.

Table 1: Comparison of Faculty Salaries by Department and Rank

Department and rank	**Average female salary**	**Female sample size**	**Average male salary**	**Male sample size**	**Difference**	***p-value***
English assistant professors	$66,327	4	$57,545	3	$8,782	.6074
English associate professors	$67,714	6	$77,901	7	-$10,187	.1885
English full professors	$78,036	5	$82,748	7	-$4,712	.4929
Statistics assistant professors	$96,464	3	$83,899	5	$12,565	.6068
Statistics associate professors	$103,184	5	$94,171	7	$9,013	.4048
Statistics full professors	$157,155	3	$143,875	16	$13,280	.5276

Conclusions should not be drawn about the entire university based on an analysis of two departments that were not randomly selected. Also, this analysis is only able to consider the people who are currently members of these departments at these ranks. This analysis cannot speak to possible biases (including gender biases) in which faculty are hired, which faculty receive tenure, how quickly faculty are promoted, or which faculty choose to leave for higher-paying jobs elsewhere. Any biases such as these, if they exist, might influence the results of this analysis. (For example, if lower-paid female faculty have left the university [voluntarily or through tenure denial], then their lower salaries cannot be included in this analysis, and the average female salary will appear to be higher.)

However, for these two departments, and with the above caveats, there is no evidence of gender discrimination in salaries.