4.5 CONFIDENCE INTERVALS AND HYPOTHESIS TESTS

In Chapter 3 we examine methods to construct confidence intervals for population parameters. We sample (with replacement) from the original sample to create a *bootstrap distribution* of possible values for a sample statistic. Based on this distribution, we produce a range of plausible values for the parameter so that we have some degree of certainty that the interval will capture the actual parameter value for the population.

In this chapter we develop methods to test claims about populations. After specifying null and alternative hypotheses, we assess the evidence in a sample by constructing a *randomization distribution* of possible sample statistics that we might see by random chance, if the null hypothesis were true. If the original sample falls in an unlikely location of the randomization distribution, we have evidence to reject the null hypothesis in favor of the alternative.

You have probably noticed similarities in these two approaches. Both use some sort of random process to simulate many samples and then collect values of a sample statistic for each of those samples to form a distribution. In both cases we are generally concerned with distinguishing between “typical” values in the middle of a distribution and “unusual” values in one or both tails. Assuming that the values in a bootstrap or randomization distribution reflect what we might expect to see if we could generate many sets of sample data, we use the information based on our original sample to make some inference about what actually might be true about a population, parameter, or relationship.

Randomization and Bootstrap Distributions

In Data 4.8 we consider measurements of body temperature for a sample of $n=50$ individuals to test $H_0{:}\mu=98.6$ vs $H_a{:}\mu\neq 98.6$, where μ is the average body temperature. The mean in the sample is $\bar{x}=98.26$, so we construct a randomization distribution by adding the difference, 0.34, to each of the sample values, creating a “population” that matches the null mean of 98.6, and then sampling with replacement from that new sample. The original sample mean (98.26) is well out in the tail of this randomization distribution (estimated $\text{p-value}=0.0016$). This shows significant evidence in the sample to reject H_0 and conclude that the average body temperature probably differs from 98.6°F.

Now suppose that we use the original data to find a 95% confidence interval for the average body temperature, μ, by constructing a bootstrap distribution. This involves sampling (with replacement) from the original sample and computing the mean for each sample. How does this differ from the randomization distribution we use in the test? The procedures are exactly the same, except that one set of values has been shifted by 0.34°F. The two distributions are displayed in Figure 4.32. Note that any of the bootstrap samples might have been selected as a sample in the randomization distribution, with the only difference being that each of the values would be 0.34° larger in the randomization sample to account for the adjustment to a null mean of 98.6°F.

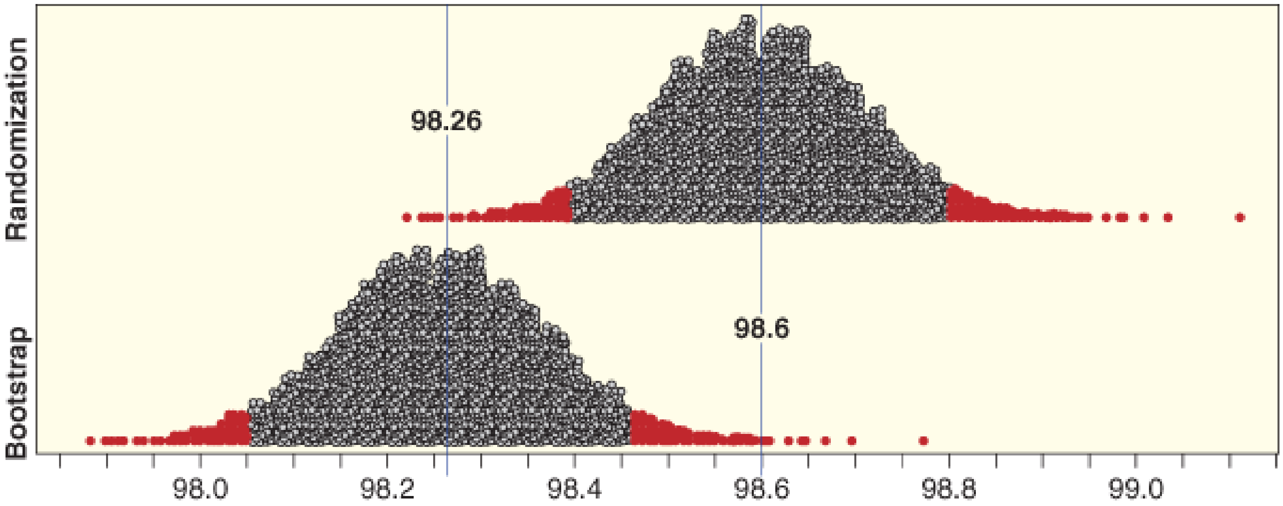

Figure 4.32 ***Bootstrap and randomization distributions for body temperatures with H_0:μ=98.6***

To find a 95% confidence interval from the bootstrap distribution of Figure 4.32 we need to find values with just 2.5% of the samples beyond them in each tail. This interval goes from 98.05 to 98.47. Thus, based on this sample, we are relatively sure that mean body temperature for the population is somewhere between 98.05°F and 98.47°F.

Note that, looking at the bootstrap confidence interval, the hypothesized value, μ=98.6, is *not* within the 95% confidence interval and, looking at the randomization distribution for the test, the mean of the sample, $\bar{x} = 98.26$, falls in the extreme tail of the distribution. This is not a coincidence! If 98.6°F is not a plausible value for the population mean, we should see this with both the confidence interval and the hypothesis test. The values in the lower and upper 2.5% tails of the randomization distribution (including the original sample mean of $\bar{x} = 98.26$) are values of sample means that would be extreme if H_0 were true and thus would lead to rejecting H_0:μ=98.6 at a 5% level. The values in the lower and upper 2.5% tails of the bootstrap distribution (including the null mean of μ=98.6) are values of means that would be outside of the 95% confidence bounds and thus are considered unlikely candidates to be the actual mean for the population.

Example 4.35

Suppose we observe the same data (so $\bar{x} = 98.26$) but are instead testing H_0:μ=98.4 versus H_a:μ≠98.4. How would Figure 4.32 change? Would the confidence interval contain the null value of μ=98.4? Would we reject the null hypothesis?

Solution

Since the bootstrap distribution and corresponding confidence interval don't depend on the hypotheses, they would remain unchanged. When testing H_0:μ=98.4 the randomization samples would only be shifted to the right by 0.14 to be centered at 98.4, as shown in Figure 4.33. Now we see that the hypothesized value, μ=98.4 is contained within the 95% confidence interval and the sample mean, $\bar{x} = 98.26$, falls in the “typical” region of the randomization distribution, so the null hypothesis would not be rejected at a 5% level.

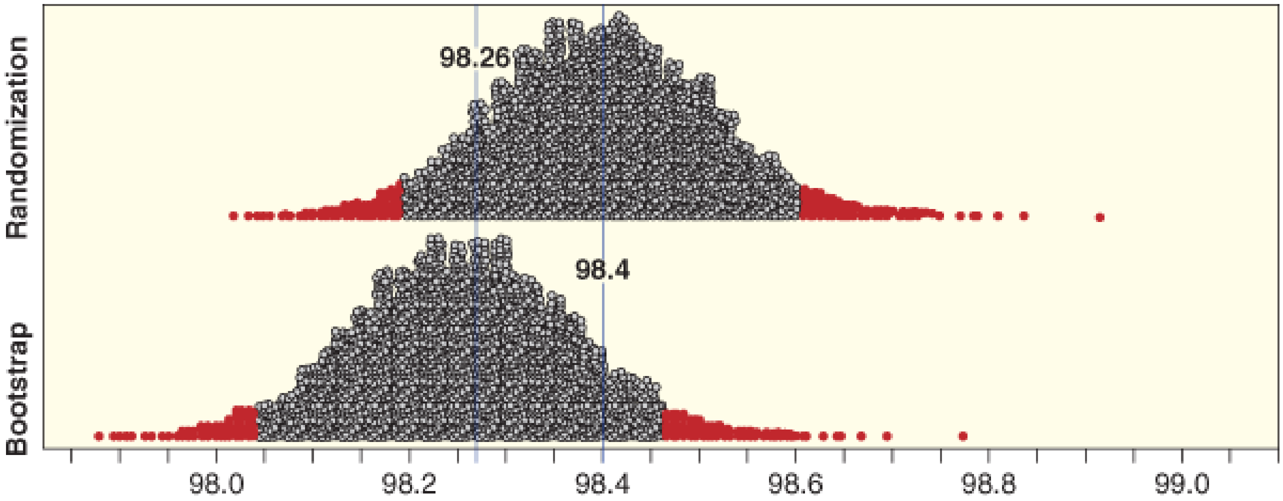

Figure 4.33 ***Bootstrap and randomization distributions for body temperatures with $H_0: \mu = 98.4$***

Connecting Confidence Intervals and Hypothesis Tests

In general, we see that a sample statistic lies in the tail of the randomization distribution when the null hypothesized parameter lies in the tail of the bootstrap distribution, and that the sample statistic lies in the typical part of the randomization distribution when the null hypothesized parameter lies in the typical part of the bootstrap distribution (i.e., in the confidence interval). While this relationship is precise for a mean, the idea extends (somewhat more loosely) to any parameter. We summarize this relationship between two-tailed tests and confidence intervals in the following box.

Connection between Confidence Intervals and Hypothesis Tests

The formal decision to a two-tailed hypothesis test is related to whether or not the hypothesized parameter value falls within a confidence interval:

- When the parameter value given in H_0 falls *outside* of a 95% confidence interval, we should reject H_0 at a 5% level in a two-tailed test based on the same sample.
- When the parameter value specified by H_0 falls *inside* of a 95% confidence interval, we should not reject H_0 at a 5% level in a two-tailed test based on the same sample.

One way to interpret this relationship between confidence intervals and tests is to view the values in a confidence interval as the plausible values for a parameter—those that would not be rejected if formally tested against a two-tailed alternative. This relationship is very flexible: It can be applied to different parameters and we can use different significance levels by adjusting the confidence level accordingly. For example, a 1% test would correspond to seeing if the hypothesized value is within a 99% confidence interval and a significance level of 10% would use a 90% confidence interval. Note that, especially when doing confidence intervals and tests using simulation methods, the correspondence is not exact. For example, the precise boundaries for the 2.5% points in the tails of either a randomization or a bootstrap distribution will vary slightly depending on the particular batch of simulated samples.

Example 4.36

The Comprehensive Assessment of Outcomes in Statistics[38]https://app.gen.umn.edu/artist/caos.html. (CAOS) exam is a standardized test for assessing students' knowledge of statistical concepts. The questions on this exam have been tested extensively to establish benchmarks for how well students do when answering them. One of the tougher questions, dealing with misinterpretations of a confidence interval, is answered correctly by about 42% of all statistics students. A statistics instructor gets the results for 30 students in a class and finds that 17 of the students ($\hat{p} = 17/30 = 0.567$) answered the confidence interval question correctly. Based on these sample results a 95% confidence interval for the proportion of students with this instructor who get the question correct goes from 0.39 to 0.75. We assume that the 30 students who answered the question are a representative sample of this instructor's students.

(a) Based on this confidence interval, is the instructor justified in saying the proportion of his students who get the question correct is different from the baseline national proportion of $p=0.42$?

(b) This question is in a multiple-choice format with four possible answers, only one of which is correct. Can the instructor conclude that his students are not just guessing on this question?

Solution

(a) If the hypotheses are $H_0: p=0.42$ and $H_a: p\neq 0.42$, we see that the null proportion is within the 95% confidence interval, (0.39, 0.75), so using a 5% significance level we do not reject H_0. The instructor would not have sufficient evidence to conclude that the proportion correct for his students is different than 42%.

(b) If students are just guessing, the proportion correct for a question with four choices is $p=0.25$. Since 0.25 is not within the 95% confidence interval, we reject H_0 and the instructor can conclude (using a 5% significance level) that the proportion of correct answers for this question is different from 0.25. The students are doing better than merely guessing at random.

About 59% of Americans favor a ban on smoking in restaurants

Example 4.37

In a Gallup poll of American adults in August 2010, 59% of the respondents favored a total ban on smoking in restaurants.[39]http://www.gallup.com/poll/141809/Americans-Smoking-Off-Menu-Restaurants.aspx. In a similar survey a decade earlier the proportion who favored such a ban was only 40%. We use these two samples to construct a 95% confidence interval for the difference in proportion of support for a smoking ban in restaurants between these two years, $p_2 - p_1$, where p_2 is the proportion in 2010 and p_1 is the proportion in 2000. The confidence interval for the difference in proportions is 0.147 to 0.233.

(a) Does this confidence interval provide sufficient evidence at a 5% level that the proportion of Americans supporting a ban on smoking in restaurants was different in 2010 than it was in 2000?

(b) What conclusions (if any) could we draw if the significance level was 10% or 1%?

Solution

(a) When testing $H_0 : p_2 = p_1$, the null difference in proportions is $p_2 - p_1 = 0$. Since the 95% confidence interval for $p_2 - p_1$ does not include zero, we have sufficient evidence (at a 5% level) to reject H_0 and conclude that the proportion of Americans favoring the smoking ban changed over the decade.

Since the confidence interval includes only positive differences, we can go even further and conclude that the proportion supporting such a ban was *higher* in 2010 than it was in 2000. This conclusion may seem more appropriate for a one-tailed test, but note that a sample statistic which is far enough in the tail to reject H_0 for a two-tailed test will also reject H_0 for a one-tailed test in that direction.

(b) Since part (a) indicates that we should reject H_0 at a 5% significance level, we know we would also reject H_0 at the larger 10% level and draw the same conclusion. However, we cannot reach a firm decision for a 1% test based only on the results of the 95% confidence interval for the difference in proportions. Since that is a stricter significance level, we would need to either construct a 99% confidence interval for the difference or carry out the actual details of the hypothesis test to make a decision at the 1% level.

Since we can use a confidence interval to make a conclusion in a hypothesis test, you might be wondering why we bother with significance tests at all. Couldn't we just always compute a confidence interval and then check whether or not it includes some hypothesized value for a parameter? If we adopted this approach, we could make a reject-not reject decision, but we lose information about the strength of evidence. For example, when actually doing a hypothesis test for the situation in Example 4.37, the p-value is less than 0.0001, indicating very strong evidence that the proportion of Americans who support a total ban on smoking in restaurants has increased over the decade from 2000 to 2010. On the other hand, the question of interest is often “how big is the difference?” not just does a difference exist at all. In that case the confidence interval for the difference in proportions, (0.147, 0.233), is more useful than just knowing that the p-value is very small. Confidence intervals and hypothesis tests are both important inference procedures, and which is most relevant in a particular situation depends on the question of interest.

Practice Problems 4.5Q

Practical vs Statistical Significance

Suppose that a company offers an online tutorial course to help high school students improve their scores when retaking a standardized exam such as the Scholastic Aptitude Test (SAT). Does the online course improve scores? We might use a hypothesis test to determine if there is an improvement in scores and a confidence interval to determine the size of the improvement. Suppose we set up an experiment to measure the change in SAT score by randomly assigning students to either take the course or just study on their own before retaking the SAT. We let μ_c be the mean change in SAT scores for those taking the online course and μ_{nc} be the mean change for those who just prepare on their own with no course. This gives the hypotheses

$$\begin{aligned} H_0&: \mu_c = \mu_{nc} \\ H_a&: \mu_c > \mu_{nc} \end{aligned}$$

Suppose that we randomly assign 2000 students to take the online course and another 2000 students to a “no course” group. Figure 4.34 shows histograms of the score changes for both groups. Although some students in both groups do worse (i.e., have a negative change) when they retake the exam, in general students tend to do better the second time. The mean change for the sample of students taking the online course is $\bar{x}_c = 42.7$ points improvement and for the other group without the course the sample mean change is $\bar{x}_{nc} = 38.5$ points. The difference is $D = 42.7 - 38.5 = 4.2$ points and a randomization distribution shows the upper tail p-value is about 0.0038. For any reasonable significance level this is a small p-value so we have very strong evidence to reject H_0 and conclude that the mean improvement in SAT scores is higher for students who use the online course.

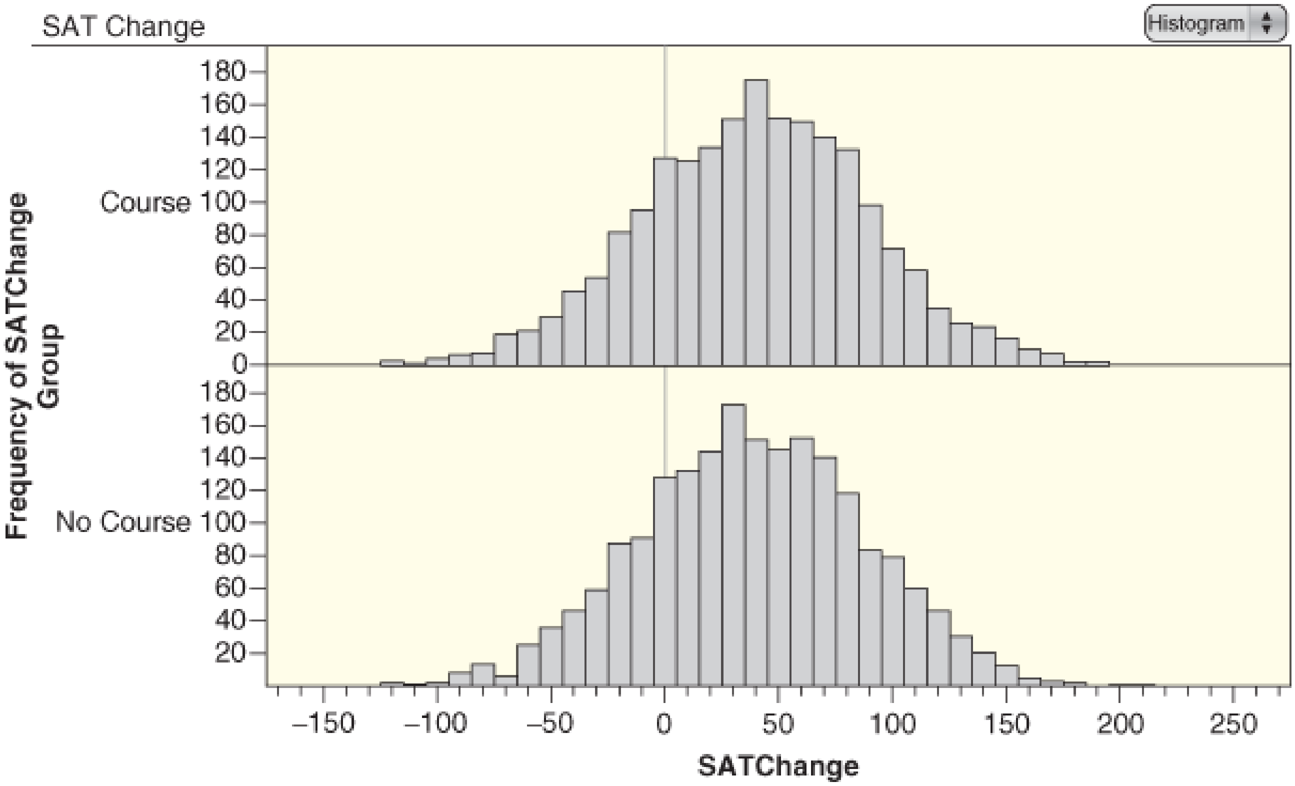

Figure 4.34 ***Hypothetical SAT score changes for groups of 2000 students with/without an online course***

We not only care about significance but also want to know how much higher the average improvement is for students who use the online course. For this, we compute an interval estimate. A 95% confidence interval for difference in mean improvement in SAT scores for students who use the online course minus students who don't is $(1.04, 7.36)$ points. Is an improvement between 1 and 7 points worth it?

Now suppose that the online prep course costs $3000 and takes more than 50 hours to complete. Would you be willing to spend that time and money to earn (on average) roughly 4 more points than you might get by preparing on your own (on an exam that is scored out of 800 points)? Would that magnitude of a score change really make much difference in how your SAT score is viewed by a college admissions officer?

Example 4.38

In testing whether an online prep course for the SAT test improves scores, we saw that the average increase is 4.2 points and the p-value for the test is 0.0038. Are the results statistically significant? Are the results practically significant?

Solution ▶

Since the p-value is very low, at 0.0038, the results are definitely statistically significant. Since the average improvement is only 4.2 points, however, the results are probably not practically significant. It is probably not worth taking the online course for such a small change.

This hypothetical example demonstrates that a difference that is *statistically* significant might not have much *practical* significance. Especially when the sample sizes are large, a rather small difference (such as 4 points on an 800-point SAT exam) might turn out to be statistically significant. That does not necessarily mean that the difference is going to be particularly important to individuals making a decision (such as whether or not to take the online course). While some small differences may be important and large samples can help reveal the true effects, we should not make the mistake of automatically assuming that anything that is statistically significant is practically significant. Conversely, for smaller samples, a difference that appears large may be the result of random chance and not statistically significant.

The Problem of Multiple Testing

In Section 3.2 we see that a 95% confidence interval will capture the true parameter 95% of the time, which also means that 5% of these confidence intervals will miss the true parameter. Similarly, in Section 4.3, we see that if the null hypothesis is true, then 5% of hypothesis tests using $\alpha=0.05$ will incorrectly reject the null hypothesis. (Recall that α is the probability of a Type I error, which is rejecting a true null hypothesis.) It is important to remember that intervals will not always capture the truth and results can be deemed statistically significant even when the null hypothesis is in fact true.

These issues become even more important when doing multiple hypothesis tests. Of all hypothesis tests conducted for a true null hypothesis, using $\alpha=0.05$, 5% of the tests will lead to rejecting the null hypothesis! In other words, if you do 100 hypothesis tests, all testing for an effect that doesn't exist (the null is true), about 5% of them will incorrectly reject the null.

⚠

If we use a significance level of $\alpha=0.05$, about 5% of tests that are testing true null hypotheses will incorrectly reject the null hypothesis.

Example 4.39

Opening an Umbrella Indoors

Is it really bad luck to open an umbrella indoors? Suppose researchers all over the world set out to actually test this idea, each randomizing people to either open an umbrella indoors or open an umbrella outdoors, and somehow measure "luck" afterward. If there are 100 people all testing this phenomenon at $\alpha=0.05$, and if opening an umbrella indoors does *not* bring bad luck, then about how many people do you expect to get statistically significant results?

Solution ▶

If the null hypothesis is true (opening an umbrella indoors has no effect on luck), then about 5% of the hypothesis tests will get p-values less than 0.05 just by random chance, so about 5 of the 100 people testing this phenomenon will get statistically significant results.

If multiple hypothesis tests are conducted for an effect that doesn't exist, some of them may get significant results just by chance. The more hypothesis tests being conducted, the higher the chance that at least one of those tests will make a Type I error. This problem is known as the problem of *multiple testing*.

The Problem of Multiple Testing

When multiple hypothesis tests are conducted, the chance that at least one test incorrectly rejects a true null hypothesis increases with the number of tests.

If the null hypotheses are all true, α of the tests will yield statistically significant results just by random chance.

This issue is made even worse by the fact that usually only significant results are published. This problem is known as *publication bias*: Usually only significant results are published, while no one knows of all the studies producing insignificant results. Consider the umbrella example. If the five statistically significant studies are all published, and we do not know about the 95 insignificant studies, we might take this as convincing evidence that opening an umbrella indoors really does cause bad luck. Unfortunately this is a very real problem with scientific research.

⚠

Often, only significant results are published. If many tests are conducted, some of them will be significant just by chance, and it may be only these studies that we hear about.

The problem of multiple testing can also occur when one researcher is testing multiple hypotheses.

DATA 4.9 **Genes and Leukemia**

Genome association studies, tests for whether genes are associated with certain diseases or other traits, are currently widely used in medical research, particularly in cancer research. Typically, DNA is collected from a group of people, some of whom have the disease in question, and some of whom don't. These DNA data are made up of values for thousands of different genes, and each gene is tested to see if there is a difference between the diseased patients and the healthy patients. Results can then be useful in risk assessment, diagnosis, and the search for a cure. One of the most famous genome association studies tested for genetic differences between patients with two different types of leukemia (acute myeloid leukemia and acute lymphoblastic leukemia).[40]Golub, T.R., et al., "*Molecular Classification of Cancer:*

Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 1999; 286: 531-537. In this study, scientists collected data on 7129 different genes for 38 patients with leukemia.

Example 4.40

Genes and Leukemia

Data 4.9 refers to a study in which data included information on 7129 genes, and each gene was tested for a difference between the two types of leukemia.

(a) If all tests used a significance level of $\alpha=0.01$, and if there are no genetic differences between the two types of leukemia, about how many of the genes would be found to be significantly different between the two groups?

(b) Do we have reason to believe that all of the genes found to be statistically significant are actually associated with the type of leukemia?

(c) In the actual study, 11% of tests for each gene yielded p-values less than 0.01. Do we have reason to believe that there is some association between genes and the type of leukemia?

Solution

(a) If there are no genetic differences between the two types of leukemia, then we would expect about 0.01 of the tests to yield statistically significant results just by random chance. We expect about $0.01\times7129\approx71$ of the genes to be found to be significantly different between the two groups, even if no differences actually exist.

(b) Because we expect 71 genes to be found significant just by random chance even if no associations exist, we should not believe that all genes found to be statistically significant are actually associated with the type of leukemia.

(c) If there were no association between genes and leukemia, we would only expect about 1% of the tests to yield p-values less than 0.01. Because 11% of the genes yielded p-values below 0.01, some of them are probably truly associated with the type of leukemia.

There are many ways of dealing with the problem of multiple testing,[41]One common way, known as Bonferroni's correction, is to divide the significance level by the number of tests. For $\alpha=0.05$ and 100 tests, a p-value would have to be less than $0.05/100=0.0005$ to be deemed statistically significant. but those methods are outside the scope of this text. The most important thing is to be aware of the problem, and to realize that when doing multiple hypothesis tests, some are likely to be significant just by random chance.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Interpret a confidence interval as the plausible values of a parameter that would not be rejected in a two-tailed hypothesis test
- Determine the decision for a two-tailed hypothesis test from an appropriately constructed confidence interval
- Recognize that statistical significance is not always the same as practical significance
- Explain the potential problem with significant results when doing multiple tests

Exercises for Section 4.5

SKILL BUILDER 1

In Exercises 4.146 to 4.149, hypotheses for a statistical test are given, followed by several possible confidence intervals for different samples. In each case, use the confidence interval to state a conclusion of the test for that sample and give the significance level used.

4.146 Hypotheses: $H_0:\mu=15$ vs $H_a:\mu\neq15$

(a) 95% confidence interval for μ: 13.9 to 16.2

(b) 95% confidence interval for μ: 12.7 to 14.8

(c) 90% confidence interval for μ: 13.5 to 16.5

4.147 Hypotheses: $H_0:p=0.5$ vs $H_a:p\neq0.5$

(a) 95% confidence interval for p: 0.53 to 0.57

ANSWER ⊕

WORKED SOLUTION ⊕

(b) 95% confidence interval for p: 0.41 to 0.52

ANSWER ⊕

WORKED SOLUTION ⊕

(c) 99% confidence interval for p: 0.35 to 0.55

ANSWER ⊕

WORKED SOLUTION ⊕

4.148 Hypotheses: $H_0:\rho=0$ vs $H_a:\rho\neq0$. In addition, in each case for which the results are significant, give the sign of the correlation.

(a) 95% confidence interval for ρ: 0.07 to 0.15

(b) 90% confidence interval for ρ: -0.39 to -0.78

(c) 99% confidence interval for ρ: -0.06 to 0.03

4.149 Hypotheses: $H_0:\mu_1=\mu_2$ vs $H_a:\mu_1\neq\mu_2$. In addition, in each case for which the results are significant, state which group (1 or 2) has the larger mean.

(a) 95% confidence interval for $\mu_1 - \mu_2$: 0.12 to 0.54

ANSWER ⊕

WORKED SOLUTION ⊕

(b) 99% confidence interval for $\mu_1 - \mu_2$: -2.1 to 5.4

ANSWER ⊕

WORKED SOLUTION ⊕

(c) 90% confidence interval for $\mu_1 - \mu_2$: -10.8 to -3.7

ANSWER ⊕

WORKED SOLUTION ⊕

SKILL BUILDER 2

In Exercises 4.150 to 4.152, a confidence interval for a sample is given, followed by several hypotheses to test using that sample. In each case, use the confidence interval to give a conclusion of the test (if possible) and also state the significance level you are using.

4.150 A 95% confidence interval for p: 0.48 to 0.57

(a) $H_0: p = 0.5$ vs $H_a: p \neq 0.5$

(b) $H_0: p = 0.75$ vs $H_a: p \neq 0.75$

(c) $H_0: p = 0.4$ vs $H_a: p \neq 0.4$

4.151 A 99% confidence interval for μ: 134 to 161

(a) $H_0: \mu = 100$ vs $H_a: \mu \neq 100$

ANSWER ⊕

WORKED SOLUTION ⊕

(b) $H_0: \mu = 150$ vs $H_a: \mu \neq 150$

ANSWER ⊕

WORKED SOLUTION ⊕

(c) $H_0: \mu = 200$ vs $H_a: \mu \neq 200$

ANSWER ⊕

WORKED SOLUTION ⊕

4.152 A 90% confidence interval for $p_1 - p_2$: 0.07 to 0.18

(a) $H_0: p_1 = p_2$ vs $H_a: p_1 \neq p_2$

(b) $H_0: p_1 = p_2$ vs $H_a: p_1 > p_2$

(c) $H_0: p_1 = p_2$ vs $H_a: p_1 < p_2$

4.153 **Approval Rating for Congress**

In a Gallup poll[42]http://www.gallup.com/poll/141827/Low-Approval-Congress-Not-Budging.aspx. conducted in August 2010, a random sample of $n=1013$ American adults were asked “Do you approve or disapprove of the way Congress is handling its job?” The proportion who said they approve is $\hat{p}=0.19$, and a 95% confidence interval for Congressional job approval is 0.166 to 0.214. If we use a 5% significance level, what is the conclusion if we are:

(a) Testing to see if there is evidence that the job approval rating is different than 20%. (This happens to be the average sample approval rating from the six months prior to this poll.)

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Testing to see if there is evidence that the job approval rating is different than 14%. (This happens to be the lowest sample Congressional approval rating Gallup ever recorded through the time of the poll.)

ANSWER ⊕

WORKED SOLUTION ⊕

4.154 Car Window Skin Cancer?

A new study suggests that exposure to UV rays through the car window may increase the risk of skin cancer.[43]“*Surprising Skin Cancer Risk: Too Much Driving*,” *LiveScience.com*, May 7, 2010, reporting on Butler, S. and Fosko, S., “*Increased Prevalence of Left-Sided Skin Cancers*,” *Journal of the American Academy of Dermatology*, published online, March 12, 2010. The study reviewed the records of all 1050 skin cancer patients referred to the St. Louis University Cancer Center in 2004. Of the 42 patients with melanoma, the cancer occurred on the left side of the body in 31 patients and on the right side in the other 11.

(a) Is this an experiment or an observational study?

(b) Of the patients with melanoma, what proportion had the cancer on the left side?

(c) A bootstrap 95% confidence interval for the proportion of melanomas occurring on the left is 0.579 to 0.861. Clearly interpret the confidence interval in the context of the problem.

(d) Suppose the question of interest is whether melanomas are more likely to occur on the left side than on the right. State the null and alternative hypotheses.

(e) Is this a one-tailed or two-tailed test?

(f) Use the confidence interval given in part (c) to predict the results of the hypothesis test in part (d). Explain your reasoning.

(g) A randomization distribution gives the p-value as 0.003 for testing the hypotheses given in part (d). What is the conclusion of the test in the context of this study?

(h) The authors hypothesize that skin cancers are more prevalent on the left because of the sunlight coming in through car windows. (Windows protect against UVB rays but not UVA rays.) Do the data in this study support a conclusion that more melanomas occur on the left side because of increased exposure to sunlight on that side for drivers?

4.155 Print vs E-books

Suppose you want to find out if reading speed is any different between a print book and an e-book.

(a) Clearly describe how you might set up an experiment to test this. Give details.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Why is a hypothesis test valuable here? What additional information does a hypothesis test give us beyond the descriptive statistics we discussed in Chapter 2?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Why is a confidence interval valuable here? What additional information does a confidence interval give us beyond the descriptive statistics of Chapter 2 and the results of a hypothesis test described in part (b)?

ANSWER ⊕

WORKED SOLUTION ⊕

(d) A similar study[44]Neilsen, J., “*iPad and Kindle Reading Speeds*,” www.useit .com/alertbox/ipad-kindle-reading.html, accessed July 2010. has been conducted and reports that “the difference between Kindle and the book was significant at the $p<.01$ level, and the difference between the iPad and the book was marginally significant at $p=.06$.” The report also stated that “the iPad measured at 6.2% slower reading speed than the printed book, whereas the Kindle measured at 10.7% slower than print. However, the difference between the two devices [iPad and Kindle] was not statistically significant because of the data's fairly high variability.” Can you tell from the first quotation which method of reading (print or e-book) was faster in the sample or do you need the second quotation for that? Explain the results in your own words.

ANSWER ⊕

WORKED SOLUTION ⊕

4.156 Are You “In a Relationship”?

A new study[45]Roan, S., “*The True Meaning of Facebook's ‘In a Relationship’*,” *Los Angeles Times*, February 23, 2012, reporting on a study in *Cyberpsychology, Behavior, and Social Networking*. shows that relationship status on Facebook matters to couples. The study included 58 college-age heterosexual couples who had been in a relationship for an average of 19 months. In 45 of the 58 couples, both partners reported being in a relationship on Facebook. In 31 of the 58 couples, both partners showed their dating partner in their Facebook profile picture. Men were somewhat more likely to include their partner in the picture than vice versa. However, the study states: “Females' indication that they are in a relationship was not as important to their male partners compared with how females felt about male partners indicating they are in a relationship.” Using a population of college-age heterosexual couples who have been in a relationship for an average of 19 months:

(a) A 95% confidence interval for the proportion with both partners reporting being in a relationship on Facebook is about 0.66 to 0.88. What is the conclusion in a hypothesis test to see if the proportion is different from 0.5? What significance level is being used?

(b) A 95% confidence interval for the proportion with both partners showing their dating partner in their Facebook profile picture is about 0.40 to 0.66. What is the conclusion in a hypothesis test to see if the proportion is different from 0.5? What significance level is being used?

4.157 Testing for a Gender Difference in Compassionate Rats

In Exercise 3.80, we found a 95% confidence interval for the difference in proportion of rats showing compassion, using the proportion of female rats minus the proportion of male rats, to be 0.104 to 0.480. In testing whether there is a difference in these two proportions:

(a) What are the null and alternative hypotheses?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Using the confidence interval, what is the conclusion of the test? Include an indication of the significance level.

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Based on this study would you say that female rats or male rats are more likely to show compassion (or are the results inconclusive)?

ANSWER ⊕

WORKED SOLUTION ⊕

4.158 Testing for a Home Field Advantage in Soccer

In Exercise 3.108, we see that the home team was victorious in 70 games out of a sample of 120 games in the FA premier league, a football (soccer) league in Great Britain. We wish to investigate the proportion p of all games won by the home team in this league.

(a) Use *StatKey* or other technology to find and interpret a 90% confidence interval for the proportion of games won by the home team.

(b) State the null and alternative hypotheses for a test to see if there is evidence that the proportion is different from 0.5.

(c) Use the confidence interval from part (a) to make a conclusion in the test from part (b). State the confidence level used.

(d) Use *StatKey* or other technology to create a randomization distribution and find the p-value for the test in part (b).

(e) Clearly interpret the result of the test using the p-value and using a 10% significance level. Does your answer match your answer from part (c)?

(f) What information does the confidence interval give that the p-value doesn't? What information does the p-value give that the confidence interval doesn't?

(g) What's the main difference between the bootstrap distribution of part (a) and the randomization distribution of part (d)?

4.159 Change in Stock Prices

Standard & Poor's maintains one of the most widely followed indices of large-cap American stocks: the S&P 500. The index includes stocks of 500 companies in industries in the US economy. A random sample of 50 of these companies was selected, and the change in the price of the stock (in dollars) over the 5-day period from August 2 to 6, 2010 was recorded for each company in the sample. The data are available in **StockChanges**.

(a) Is this an experiment or an observational study? How was randomization used in the study, if at all? Do you believe the method of data collection introduced any bias?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Describe one way to select a random sample of size 50 from a population of 500 stocks.

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Figure 4.35 shows a boxplot of the data. Describe what this plot shows about the distribution of stock price changes in this sample.

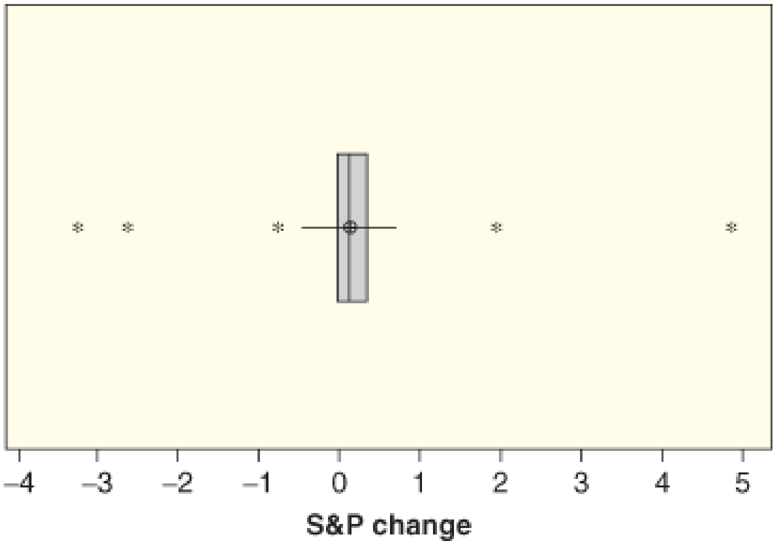

Figure 4.35 *Changes in stock prices on the S&P 500 over a 5-day period*

ANSWER ⊕

WORKED SOLUTION ⊕

(d) Give relevant summary statistics to describe the distribution of stock price changes numerically.

ANSWER ⊕

WORKED SOLUTION ⊕

(e) Use *StatKey* or other technology to calculate a 95% confidence interval for the mean change in all S&P stock prices. Clearly interpret the result in context.

ANSWER ⊕

WORKED SOLUTION ⊕

(f) Use the confidence interval from part (e) to predict the results of a hypothesis test to see if the mean change for all S&P 500 stocks over this period is different from zero. State the hypotheses and significance level you use and state the conclusion.

ANSWER ⊕

WORKED SOLUTION ⊕

(g) Now give the null and alternative hypotheses in a test to see if the average 5-day change is positive. Use *StatKey* or other technology to find a p-value of the test and clearly state the conclusion.

ANSWER ⊕

WORKED SOLUTION ⊕

(h) If you made an error in your decision in part (g), would it be a Type I error or a Type II error? Can you think of a way to actually find out if this error occurred?

ANSWER ⊕

WORKED SOLUTION ⊕

4.160 How Long Do Mammals Live?

Data 2.2 includes information on longevity (typical lifespan), in years, for 40 species of mammals.

(a) Use the data, available in **MammalLongevity**, and *StatKey* or other technology to test to see if the average lifespan of mammal species is different from 10 years. Include all details of the test: the hypotheses, the p-value, and the conclusion in context.

(b) Use the result of the test to determine whether $\mu=10$ would be included as a plausible value in a 95% confidence interval of average mammal lifespan. Explain.

4.161 How Long Are Mammals Pregnant?

Data 2.2 includes information on length of gestation (length of pregnancy in days) for 40 species of mammals.

(a) Use the data, available in **MammalLongevity**, and *StatKey* or other technology to test to see if the average gestation of mammals is different from 200 days. Include all details of the test: the hypotheses, the p-value, and the conclusion in context.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Use the result of the test to indicate whether $\mu=200$ would be included as a plausible value in a 95%

confidence interval of average mammal gestation time. Explain.

ANSWER ⊕

WORKED SOLUTION ⊕

4.162 Weight Loss Program

Suppose that a weight loss company advertises that people using its program lose an average of 8 pounds the first month and that the Federal Trade Commission (the main government agency responsible for truth in advertising) is gathering evidence to see if this advertising claim is accurate. If the FTC finds evidence that the average is less than 8 pounds, the agency will file a lawsuit against the company for false advertising.

(a) What are the null and alternative hypotheses the FTC should use?

(b) Suppose that the FTC gathers information from a very large random sample of patrons and finds that the average weight loss during the first month in the program is $\bar{x} = 7.9$ pounds with a p-value for this result of 0.006. What is the conclusion of the test? Are the results statistically significant?

(c) Do you think the results of the test are practically significant? In other words, do you think patrons of the weight loss program will care that the average is 7.9 pounds lost rather than 8.0 pounds lost? Discuss the difference between practical significance and statistical significance in this context.

4.163 Do iPads Help Kindergartners Learn: A Subtest

The Auburn, Maine, school district conducted an early literacy experiment in the fall of 2011. In September, half of the kindergarten classes were randomly assigned iPads (the intervention group) while the other half of the classes got them in December (the control group.) Kids were tested in September and December and the study measures the average difference in score gains between the control and intervention group.[46]Reich, J., "*Are iPads Making a Significant Difference? Findings from Auburn Maine*," *Ed Tech Researcher*, February 17, 2012. The experimenters tested whether the mean score for the intervention group was higher on the HRSIW subtest (Hearing and Recording Sounds in Words) than the mean score for the control group.

(a) State the null and alternative hypotheses of the test and define any relevant parameters.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) The p-value for the test is 0.02. State the conclusion of the test in context. Are the results statistically significant at the 5% level?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) The effect size was about two points, which means the mean score for the intervention group was approximately two points higher than the mean score for the control group on this subtest. A school board member argues, "While these results might be statistically significant, they may not be practically significant." What does she mean by this in this context?

ANSWER ⊕

WORKED SOLUTION ⊕

4.164 Do iPads Help Kindergartners Learn: A Series of Tests

Exercise 4.163 introduces a study in which half of the kindergarten classes in a school district are randomly assigned to receive iPads. We learn that the results are significant at the 5% level (the mean for the iPad group is significantly higher than for the control group) for the results on the HRSIW subtest. In fact, the HRSIW subtest was one of 10 subtests and the results were not significant for the other 9 tests. Explain, using the problem of multiple tests, why we might want to hesitate before we run out to buy iPads for all kindergartners based on the results of this study.

4.165 Eating Breakfast Cereal and Conceiving Boys

Newscientist.com ran the headline "Breakfast Cereals Boost Chances of Conceiving Boys," based on an article which found that women who eat breakfast cereal before becoming pregnant are significantly more likely to conceive boys.[47]Mathews, F., Johnson, P.J., and Neil, A., "*You Are What Your Mother Eats: Evidence for Maternal Preconception Diet Influencing Foetal Sex in Humans*," *Proceedings of the Royal Society B: Biological Sciences*, 2008; 275: 1643,1661-1668. The study used a significance level of $\alpha=0.01$. The researchers kept track of 133 foods and, for each food, tested whether there was a difference in the proportion conceiving boys between women who ate the food and women who didn't. Of all the foods, only breakfast cereal showed a significant difference.

(a) If none of the 133 foods actually have an effect on the gender of a conceived child, how many (if any) of the individual tests would you expect to show a significant result just by random chance? Explain.

(*Hint:* Pay attention to the significance level.)

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Do you think the researchers made a Type I error? Why or why not?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Even if you could somehow ascertain that the researchers did not make a Type I error, that is, women who eat breakfast cereals are actually more likely to give birth to boys, should you believe the headline "Breakfast Cereals Boost Chances of Conceiving Boys"? Why or why not?

ANSWER ⊕

WORKED SOLUTION ⊕

4.166 Approval from the FDA for Antidepressants

The FDA (US Food and Drug Administration) is responsible for approving all new drugs sold in the US. In order to approve a new drug for use as an antidepressant, the FDA requires two results from randomized double-blind experiments showing the drug is more effective than a placebo at a 5% level. The FDA does not put a limit on the number of times a drug company can try such experiments.

Explain, using the problem of multiple tests, why the FDA might want to rethink its guidelines.

4.167 Does Massage Really Help Reduce Inflammation in Muscles?

In Exercise 4.132, we learn that massage helps reduce levels of the inflammatory cytokine interleukin-6 in muscles when muscle tissue is tested 2.5 hours after massage. The results were significant at the 5% level. However, the authors of the study actually performed 42 different tests: They tested for significance with 21 different compounds in muscles and at two different times (right after the massage and 2.5 hours after).

(a) Given this new information, should we have less confidence in the one result described in the earlier exercise? Why?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Sixteen of the tests done by the authors involved measuring the effects of massage on muscle metabolites. None of these tests were significant. Do you think massage affects muscle metabolites?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Eight of the tests done by the authors (including the one described in the earlier exercise) involved measuring the effects of massage on inflammation in the muscle. Four of these tests were significant. Do you think it is safe to conclude that massage really does reduce inflammation?

ANSWER ⊕

WORKED SOLUTION ⊕