SECTION 6.3 Summary

P-values are more informative than the reject-or-not result of a fixed level α test. Beware of placing too much weight on traditional values of α, such as $\alpha = 0.05$.

Very small effects can be highly significant (small P), especially when a test is based on a large sample. A statistically significant effect need not be practically important. Plot the data to display the effect you are seeking, and use confidence intervals to estimate the actual values of parameters.

On the other hand, lack of significance does not imply that H_0 is true, especially when the test has low power.

Significance tests are not always valid. Faulty data collection, outliers in the data, and testing a hypothesis on the same data that suggested the hypothesis can invalidate a test. Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true.

SECTION 6.3 Exercises

For Exercise 6.84, see page 383; and for Exercise 6.85, see page 386.

6.86 A role as a statistical consultant. You are the statistical expert for a graduate student planning her PhD research. After you carefully present the mechanics of significance testing, she suggests using $\alpha = 0.20$ for the study because she would be more likely to obtain statistically significant results and she *really* needs significant results to graduate. Explain in simple terms why this would not be a good use of statistical methods.

6.87 What do you know? A research report described two results that both achieved statistical significance at the 5% level. The P-value for the first is 0.048; for the second it is 0.0002. Do the P-values add any useful information beyond that conveyed by the statement that both results are statistically significant? Write a short paragraph explaining your views on this question.

6.88 Selective publication based on results. In addition to statistical significance, selective publication can also be due to the observed outcome. A recent review of 74 FDA-registered studies of antidepressant agents found 38 studies with positive results and 36 studies with negative or questionable results. All but 1 of the 38 positive studies were published. Of the remaining 36, 22 were not published with another 11 published in such a way as to convey a positive outcome.[29] Describe how this selective reporting can have adverse consequences on health care.

6.89 What a test of significance can answer. Explain whether a test of significance can answer each of the following questions.

(a) Is the sample or experiment properly designed?

(b) Is the observed effect compatible with the null hypothesis?

(c) Is the observed effect important?

6.90 Vitamin C and colds. In a study to investigate whether vitamin C will prevent colds, 400 subjects are assigned at random to one of two groups. The experimental group takes a vitamin C tablet daily, while the control group takes a placebo. At the end of the experiment, the researchers calculate the difference between the percents of subjects in the two groups who were free of colds. This difference is statistically significant ($P = 0.03$) in favor of the vitamin C group. Can we conclude that vitamin C has a strong effect in preventing colds? Explain your answer.

6.91 How far do rich parents take us? How much education children get is strongly associated with the wealth and social status of their parents, termed "socioeconomic status," or SES. The SES of parents, however, has little influence on whether children who have graduated from college continue their education. One study looked at whether college graduates took the graduate admissions tests for business, law, and other graduate programs. The effects of the parents' SES on taking the LSAT test for law school were "both statistically insignificant and small."

(a) What does "statistically insignificant" mean?

(b) Why is it important that the effects were small in size as well as insignificant?

6.92 Do you agree? State whether or not you agree with each of the following statements and provide a short summary of the reasons for your answers.

(a) If the P-value is larger than 0.05, the null hypothesis is true.

(b) Practical significance is not the same as statistical significance.

(c) We can perform a statistical analysis using any set of data.

(d) If you find an interesting pattern in a set of data, it is appropriate to then use a significance test to determine its significance.

6.93 Practical significance and sample size. Every user of statistics should understand the distinction between statistical significance and practical importance. A sufficiently large sample will declare very small effects statistically significant. Consider the study of elite female Canadian athletes in Exercise 6.70 (page 380). Female athletes were consuming an average of 2403.7 kcal/day with a standard deviation of 880 kcal/day. Suppose a nutritionist is brought in to implement a new health program for these athletes. This program should increase mean caloric intake but not change the standard deviation. Given the standard deviation and how caloric deficient these athletes are, a change in the mean of 50 kcal/day to 2453.7 is of little importance. However, with a large enough sample, this change can be significant. To see this, calculate the P-value for the test of

$$H_0: \mu = 2403.7$$
$$H_a: \mu > 2403.7$$

in each of the following situations:

(a) A sample of 100 athletes; their average caloric intake is $\bar{x} = 2453.7$.

(b) A sample of 500 athletes; their average caloric intake is $\bar{x} = 2453.7$.

(c) A sample of 2500 athletes; their average caloric intake is $\bar{x} = 2453.7$.

6.94 Statistical versus practical significance. A study with 7500 subjects reported a result that was statistically significant at the 5% level. Explain why this result might not be particularly important.

6.95 More on statistical versus practical significance. A study with 14 subjects reported a result that failed to achieve statistical significance at the 5% level. The P-value was 0.051. Write a short summary of how you would interpret these findings.

6.96 Find journal articles. Find two journal articles that report results with statistical analyses. For each article, summarize how the results are reported and write a critique of the presentation. Be sure to include details regarding use of significance testing at a particular level of significance, P-values, and confidence intervals.

6.97 Create example of your own. For each case, provide an example and an explanation as to why it is appropriate.

(a) A set of data or experiment for which statistical inference is not valid.

(b) A set of data or experiment for which statistical inference is valid.

6.98 Predicting success of trainees. What distinguishes managerial trainees who eventually become executives from those who, after expensive training, don't succeed and leave the company? We have abundant data on past trainees—data on their personalities and goals, their college preparation and performance, even their family backgrounds and their hobbies. Statistical software makes it easy to perform dozens of significance tests on these dozens of variables to see which ones best predict later success. We find that future executives are significantly more likely than washouts to have an urban or suburban upbringing and an undergraduate degree in a technical field.

Explain clearly why using these "significant" variables to select future trainees is not wise. Then suggest a follow-up study using this year's trainees as subjects that should clarify the importance of the variables identified by the first study.

6.99 Searching for significance. Give an example of a situation where searching for significance would lead to misleading conclusions.

6.100 More on searching for significance. You perform 1000 significance tests using $\alpha = 0.05$. Assuming that all null hypotheses are true, about how many of the test results would you expect to be statistically significant? Explain how you obtained your answer.

6.101 Interpreting a very small *P*-value. Assume that you are performing a large number of significance tests. Let n be the number of these tests. How large would n need to be for you to expect about one P-value to be 0.00001 or smaller? Use this information to write an explanation of how to interpret a result that has $P = 0.00001$ in this setting.

6.102 An adjustment for multiple tests. One way to deal with the problem of misleading P-values when performing more than one significance test is to adjust the criterion you use for statistical significance. The **Bonferroni procedure** does this in a simple way. If you perform two tests and want to use the $\alpha = 5\%$ significance level, you would require a P-value of $0.05/2 = 0.025$ to declare either one of the tests significant. In general, if you perform k tests and want protection at level α, use α/k as your cutoff for statistical significance. You perform six tests and obtain individual P-values 0.083, 0.032, 0.246, 0.003, 0.010, and < 0.001. Which of these are statistically significant using the Bonferroni procedure with $\alpha = 0.05$?

6.103 Significance using the Bonferroni procedure. Refer to the previous problem. A researcher has performed 12 tests of significance and wants to apply the Bonferroni procedure with $\alpha = 0.05$. The calculated P-values are 0.041, 0.569, 0.050, 0.416, 0.002, 0.006, 0.286, 0.021, 0.888, 0.010, < 0.002, and 0.533. Which of these tests reject their null hypotheses with this procedure?

Solutions

6.86. Finding something to be "statistically significant" is not really useful unless the significance level is sufficiently small. While there is some freedom to decide what "sufficiently small" means, $\alpha = 0.20$ would lead the student to incorrectly reject H_0 one-fifth of the time, so it is clearly a bad choice.

6.87. The first test was barely significant at $\alpha = 0.05$, while the second was significant at any reasonable α.

6.88. One can learn something from negative results; for example, a study that finds no benefit from a particular treatment is at least useful in terms of what will *not* work. Furthermore, reviewing such results might point researchers to possible future areas of study.

6.89. A significance test answers only Question b. The P-value states how likely the observed effect (or a stronger one) is if H_0 is true, and chance alone accounts for deviations from what we expect. The observed effect may be significant (very unlikely to be due to chance) and yet not be of practical importance. And the calculation leading to significance *assumes* a properly designed study.

6.90. Based on the description, this seems to have been an experiment (not just an observational study), so a statistically significant outcome suggests that vitamin C is effective in preventing colds.

6.91. **(a)** If SES had no effect on LSAT results, there would still be some difference in scores due to chance variation. "Statistically insignificant" means that the observed difference was no more than we might expect from that chance variation. **(b)** If the results are based on a small sample, then even if the null hypothesis were not true, the test might not be sensitive enough to detect the effect. Knowing the effects were small tells us that the statistically insignificant test result did not occur merely because of a small sample size.

6.92. These questions are addressed in the summary for Section 6.3. **(a)** Failing to reject H_0 does not mean that H_0 is true. **(b)** This is correct; a difference that is statistically significant might not be practically important. (This does not mean that these are opposites; a difference *could* be both statistically and practically significant.) **(c)** This might be technically true, but in order for the analysis to be meaningful, the data must satisfy the assumptions of the analysis. **(d)** Searching for patterns and then testing their significance can lead to false positives (that is, we might reject the null hypothesis incorrectly). If a pattern is observed, we should collect new data to test if it is present.

6.93. In each case, we find the test statistic z by dividing the observed difference $(2453.7 - 2403.7 = 50 \text{ kcal/day})$ by $880/\sqrt{n}$. **(a)** For $n = 100$, $z \doteq 0.57$, so $P = P(Z > 0.57) = 0.2843$. **(b)** For $n = 500$, $z \doteq 1.27$, so $P = P(Z > 1.27) = 0.1020$. **(c)** For $n = 2500$, $z \doteq 2.84$, so $P = P(Z > 2.84) = 0.0023$.

6.94. The study may have rejected $\mu = \mu_0$ (or some other null hypothesis), but with such a large sample size, such a rejection might occur even if the actual mean (or other parameter) differs only slightly from μ_0. For example, there might be no practical importance to the difference between $\mu = 10$ and $\mu = 10.5$.

6.95. We expect more variation with small sample sizes, so even a large difference between $\bar{x}$ and μ_0 (or whatever measures are appropriate in our hypothesis test) might not turn out to be significant. If we were to repeat the test with a larger sample, the decrease in the standard error might give us a small enough P-value to reject H_0.

6.98. When many variables are examined, "significant" results will show up by chance, so we should not take it for granted that the variables identified are really indicative of future success. In order to decide if they are appropriate, we should track this year's trainees and compare the success of those from urban/suburban backgrounds with the rest, and likewise compare those with a degree in a technical field with the rest.

6.100. We expect 50 tests to be statistically significant: Each of the 1000 tests has a 5% chance of being significant, so the number of significant tests has a binomial distribution with $n = 1000$ and $p = 0.05$, for which the mean is $np = 50$.

6.101. $P = 0.00001 = \frac{1}{100,000}$, so we would need $n = 100{,}000$ tests in order to expect one P-value of this size (assuming that all null hypotheses are true). That is why we reject H_0 when we see P-values such as this: It indicates that our results would rarely happen if H_0 were true.

6.102. Using $\alpha/6 \doteq 0.008333$ as the cutoff, the fourth ($P = 0.003$) and sixth ($P < 0.001$) tests are significant.

6.103. Using $\alpha/12 \doteq 0.004167$ as the cutoff, we reject the fifth ($P = 0.002$) and eleventh ($P < 0.002$) tests.