Section 9.1: Steps for Performing a Significance Test

9.1 H_0 or H_a?

a) null hypothesis
b) alternative hypothesis
c) (a) H_0: $p = 0.50$; H_a: $p \neq 0.50$
(b) H_0: $p = 0.24$; H_a: $p < 0.24$

9.2 H_0 or H_a?

a) Null hypothesis
b) Alternative hypothesis
c) Alternative hypothesis

9.3 Burden of proof

H_0: The mean toxicity level equals the threshold. (“no effect”); H_0 specifies a single value, the threshold, for the parameter.

H_a: The mean toxicity level exceeds the threshold.

9.4 Electricity prices

In this study, the population parameter is the mean dollar amount of the monthly electricity bill for a residential household in the United States this year, denoted by μ. The null hypothesis is H_0: $\mu = 114$, and the alternative hypothesis is H_a: $\mu \neq 114$.

9.5 Low-carbohydrate diet

a) This is an alternative hypothesis because it has a range of parameter values.
b) The relevant parameter is the mean weight change, μ. H_0: $\mu = 0$; this is a null hypothesis.

9.6 Examples of hypotheses

The examples given by each student will differ.

9.7 Proper hypotheses?

a) The null and alternative hypotheses are always about population parameters (e.g., $\hat{p}$ or μ) and never about sample statistics such as $\hat{p}$ or $\bar{x}$. The correct hypotheses are: H_0: $p = 0.5$, H_a: $p > 0.5$. For example: more than half of students in the university are females.

b) The alternative hypothesis needs to specify a range of parameters. The correct hypotheses are: H_0: $\mu = 10$, H_a: $\mu \neq 10$. For example: the number of occupied hotel rooms per day are different than 10.

c) By convention, the null hypothesis (H_0) statement is written as an equality, the range of values in the alternative hypothesis should be appropriately associated with the null value. The correct hypotheses are: H_0: $p = 0.10$, H_a: $p < 0.10$. For example: less than 10% of people who take a loan from a financial institution will fail to pay their loans when due.

9.8 *z* test statistic

The data give strong evidence against the null hypothesis. Most scores fall within three standard errors of the mean, and this sample proportion falls over three standard errors from the null hypothesis value.

9.9 P-value

a) This P-value does not give strong evidence against the null hypothesis.
b) This extreme P-value does give strong evidence against the null hypothesis.

Section 9.2: Significance Tests About Proportions

9.10 Customer satisfaction

The parameter of interest is the proportion, p, of the car workshop’s customers who are not satisfied with the service provided. The null hypothesis is that the proportion, p, of the car workshop’s customers who are not satisfied with the service provided is 0.5, and the alternative hypothesis is that the proportion of the car workshop’s customers who are not satisfied with the service provided is more than 0.5.
H_0: $p = 0.5$ and H_a: $p > 0.5$.

9.11 Believe in astrology?

The parameter of interest is, p, the true probability of a correct prediction. The null hypothesis is that the astrologer will successfully predict the personality profile 1/4 of the time and the alternative hypothesis is that the astrologer will successfully predict the personality profile more than 1/4 of the time. H_0: $p = 1/4$ and H_a: $p > 1/4$

9.12 Get P-value from *z*

a) 0.15
b) 0.30
c) 0.85
d) None of these P-values gives strong evidence against H_0. All of them indicate that the null hypothesis is plausible.

9.13 Get more P-values from *z*

a) (i) 0.006
(ii) 0.012
(iii) 0.994
b) Yes, the P-values in (i) and (ii) indicate that the test statistic is very extreme, strong evidence against H_0.

9.14 Find test statistic and P-value

a) The standard error is $\sqrt{p_0(1-p_0)/n} = \sqrt{0.5(1-0.5)/100} = 0.05$, so $z = \dfrac{0.35-0.50}{0.05} = -3.0$.
b) The P-value is 0.001.
c) If the null hypothesis were true, the probability would be 0.001 of getting a test statistic at least as extreme as the value observed. This does provide strong evidence against H_0. This is a very small P-value; it does not seem plausible that $p = 0.50$.

9.15 Dogs and cancer

a) H_0: $p = 1/5$
b) H_a: $p \neq 1/5$
c) H_a: $p > 1/5$
d) P-value ≈ 0.000. The probability of obtaining a sample proportion of 81 or more successes in 83 trials is essentially 0; thus, there is strong evidence that the probability of a correct selection is greater than with random guessing.

9.16 Religion important in your life?

1. The response is categorical with outcomes "yes" or "no" to the statement that young adults pray daily; p represents the probability of a yes response. The poll was a random sample of 1679 18-29-year-olds and $np_0 = n(1-p_0) = 1679(0.5) = 839.5 \geq 15$.
2. $H_0: p = 0.5$; $H_a: p \neq 0.5$
3. $\hat{p} = 0.45$, so $z = \dfrac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{0.45-0.5}{\sqrt{0.5(0.5)/1679}} = -4.10$. The sample proportion is 4.1 standard errors below the null hypothesis value.
4. The P-value ≈ 0 is the probability of obtaining a sample proportion at least as extreme as the one observed, if the null hypothesis is true.
5. Since the P-value is approximately 0, the sample data supports the alternative hypothesis. There is very strong evidence that the percentage of 18-29-year-olds who pray daily is not 50%.

9.17 Another test of astrology

a) Let p be the proportion of adults who guess correctly.
$H_0: p = 1/3$; $H_a: p > 1/3$
b) $\hat{p} = 28/83 = 0.337$; The standard error is $\sqrt{p_0(1-p_0)/n} = \sqrt{0.333(1-0.333)/83} = 0.052$, so

$$z = \frac{0.337-0.333}{0.052} = 0.08.$$

9.17 (continued)

c) The P-value is 0.47. If the null hypothesis were true, the probability would be 0.47 of getting a test statistic at least as extreme as the value observed. It is plausible that the null hypothesis is correct. I would not conclude that people are more likely to select their "correct" horoscope than if they were randomly guessing.

9.18 Opinion on fracking a year earlier

a) Let p be the proportion of the population that opposes increased use of fracking.
$\mathrm{H}_0: p = 0.5$; $\mathrm{H}_a: p < 0.5$

b) $\hat{p} = 740/1506 = 0.491$; The standard error is $\sqrt{p_0(1-p_0)/n} = \sqrt{0.50(1-0.50)/1506} = 0.0129$, so $z = \frac{0.491-0.50}{0.0129} = -0.670$; the sample proportion of 0.491 is less than one standard error less than the null hypothesis value of 0.50.

c) The P-value is 0.251. Because the P-value is larger than the significance level of 0.05, we do not reject the null hypothesis. There is no evidence from the survey that in 2013 those opposing fracking are in the minority. The probability would be 0.251 of getting a test statistic at least as extreme as the value observed if the null hypothesis were true, and the population proportion were 0.50.

d) $np_0 = n(1-p_0) = 1506(0.5) = 735 \geq 15$; The sample of 1506 respondents must be a random sample representative of the U.S. population in 2013 (which it was).

e) The P-value for the two-sided alternative is 2(0.251) = 0.502.

9.19 Testing a headache remedy

a) $\hat{p} = 22/30 = 0.733$; The standard error is $\sqrt{p_0(1-p_0)/n} = \sqrt{0.50(1-0.50)/30} = 0.091$, so $z = \frac{0.733-0.50}{0.091} = 2.56$.

b) When we look up the P-value, we find that the proportion beyond that z-score is 0.005. Doubled to include both tails, the P-value is 0.01. If the null hypothesis were true, the probability would be 0.01 of getting a test statistic at least as extreme as the value observed. We have strong evidence that the population proportion of children who had more pain relief with the drug differs from 0.50.

c) We have to meet three assumptions to use this test. The data must be categorical, the data must be obtained using randomization, and the sample size must be large enough that the sampling distribution of the sample proportion is approximately normal (i.e., expected successes and failures both at least fifteen under H_0). $np_0 = n(1-p_0) = 30(0.5) = 15 \geq 15$, But in this case, the data were not obtained using randomization, but rather from a convenience sample which might not be representative of the population.

9.20 Gender bias in selecting managers

a) Let p be the probability that the company selects a female $(\mathrm{H}_0: p = 0.4)$. No effect here means a probability of selecting females that is in accordance with the proportion of eligible females. (Note: It is also correct to let p be the probability of selecting a male $(\mathrm{H}_0: p = 0.6)$.

b) There is a gender bias if p is different (i.e., either smaller or larger) than 0.4, so we test $\mathrm{H}_0: p = 0.4$ and $\mathrm{H}_a: p \neq 0.4$. (Or, if p was selected as the probability of selecting a male in (a), then we have $\mathrm{H}_0: p = 0.6$ and $\mathrm{H}_a: p \neq 0.6$.)

c) The large-sample analysis is justified because the expected successes and failures are both at least fifteen under H_0, $np_0 = 40(0.4) = 16 \geq 15$ and $n(1-p_0) = 40(0.6) = 24 \geq 15$. $\hat{p} = 12/40 = 0.30$; The standard error is $\sqrt{p_0(1-p_0)/n} = \sqrt{0.40(1-0.40)/40} = 0.077$, so $z = \frac{0.3-0.4}{0.077} = -1.29$.

d) The P-value in the table refers to the alternative hypothesis $(\mathrm{H}_a: p \neq 0.4)$. (Note: If you chose p as the probability of selecting a male in (a) and tested the alternative $(\mathrm{H}_a: p \neq 0.6)$, it will have the same P-value.) The P-value of 0.1967 is large. The sample proportion of 0.3 falls not too far (1.29 standard errors) from the hypothesized value of 0.4, indicating no unusual data in light of the null hypothesis.

9.20 (continued)

If the null-hypothesis were true, observing a test statistic (or sample proportion) this extreme or even more extreme is not that unlikely (19.7%).

e) Because the P-value of 0.1967 is larger than the significance level, we would not reject the null hypothesis. There is insufficient evidence that the proportion of females selected for management training is different from the proportion of 40% eligible for that training.

9.21 Gender discrimination

Plausible values for probability of a female to be selected range from 0.16 to 0.44. This is in accordance with the decision reached in (e) of the previous exercise not to reject the null hypothesis $(\text{H}_0: p = 0.4)$ in favor of the alternative hypothesis $(\text{H}_a: p \neq 0.4)$ because 0.4 is a plausible value for that probability.

9.22 Complementary and integrative health (CIH) among nurses in Iran

a) The relevant variable is whether nurses in hospitals in Iran apply or do not apply CIH methods. The parameter is the population proportion, p = proportion of nurses who have never applied CIH methods.

b) $H_0: p = 0.5$ and $H_a: p \neq 0.5$; the sample size is adequate because there are at least 15 successes (nurses who have never applied for CIH methods) and failures (nurses who applied for CIH methods).

c) $\hat{p} = 57.3\%$; The standard error is $\sqrt{p_0(1-p_0)/n} = \sqrt{0.5(1-0.5)/157} = 0.04.$ So,

$$z = \frac{0.573 - 0.5}{0.04} = 1.825.$$

d) The P-value is 0.068. This P-value is not that extreme. If the null hypotheses were true, the probability would be 0.068 of getting a test statistic at least as extreme as the value observed. It is plausible that the null hypothesis is correct. Because the probability is 0.068 that we would observe our test statistic or one more extreme due to random variation, there is not strong evidence that the population proportion who would have never applied CIH methods differs from 0.50.

9.23 Performance of Egypt's president

The variable is whether someone "highly approve" the Egyptian president's performance. The parameter of interest is the proportion among all citizens aged 18 years old and above who "highly approve" the Egyptian president's performance.

1) Assumptions: The data are categorical (highly approve or others) and obtained randomly; the expected successes and failures are both at least 15 under
H_0: $np = (0.5)(709) \geq 15,$ and $n(1-p) = (0.5)(709) \geq 15.$

2) Parameter: p = population proportion of citizens who highly approve the president's performance.
Hypotheses: $H_0: p = 0.5;$ $H_a: p \neq 0.5.$

3) Test statistic: $z = \dfrac{0.51 - 0.5}{\sqrt{0.5(1-0.5)/709}} = 0.53.$

4) P-value: 0.596.

5) Conclusion: We cannot reject the null hypothesis at a significance level of 0.05; we have strong evidence that the population proportion of Egyptians who highly approve the president's performance is 0.5.

9.24 Which cola?

a) The test statistic ("Z-Value") is calculated by taking the difference between the sample proportion and the null proportion and dividing it by the standard error.

b) We get the "P-value" by looking up the "Z-value" in Table A or using technology. We have to determine the two-tail probability from the standard normal distribution below –1.286 and above 1.286. The P-value of 0.20 tells us that if the null hypothesis were true, a proportion of 0.20 of samples would fall at least this far from the null hypothesis proportion of 0.50. This is not very extreme; it is plausible that the null hypothesis is correct, and that Coke is not preferred to Pepsi.

c) It does not make sense to accept the null hypothesis. It is possible that there is a real difference in the population that we are not detecting in our test (perhaps because the sample size is not very large), and we can never accept a null hypothesis. A confidence interval shows that 0.50 is one of many plausible values.

9.24 (continued)

d) The 95% confidence interval tells us the range of plausible values, whereas the test merely tells us that 0.50 is plausible.

9.25 How to sell a burger

1) Assumptions: The data are categorical (higher sale with coupons versus higher sales with posters); we'll assume the data are obtained randomly; the expected successes and failures are both at least fifteen under H_0; $np_0 = n(1-p_0) = 50(0.5) = 25 \geq 15$.
2) Hypotheses: $H_0: p = 0.5$; $H_a: p \neq 0.5$
3) Test Statistic: $z = \dfrac{0.56 - 0.50}{\sqrt{0.5(1-0.5)/50}} = 0.85$
4) P-value: 0.40
5) Conclusion: If the null hypothesis were true, the probability would be 0.40 of getting a test statistic at least as extreme as the value observed. It is plausible that the null hypothesis is correct, and that coupons do not lead to higher sales than do posters.

♦♦9.26 A binomial headache

This P-value gives strong evidence against the null hypothesis. It would be very unlikely to have a sample proportion of 1.00 if the actual population proportion were 0.50.

♦♦9.27 P-value for small samples

a) This has the binomial distribution because there are two possible outcomes, each trial holds the same probability of success, and the *n* trials are independent.

b) The P-value represents the probability of observing the test statistic $x = 22$, or a value even more extreme if the population proportion is 1/7.
Because the random variable *X* is binomial, the P-value is $P(x = 22) + P(x = 23) + \ldots + P(x = 54)$.

Section 9.3: Significance Tests About Means

9.28 Which *t* has P-value = 0.05?

a) $t = -2.145$ or $t = 2.145$

b) $t = 1.762$

c) $t = -1.762$

9.29 Practice mechanics of a test

a) 0.0026 or 0.03

b) 0.013

c) 0.987

9.30 Effect of *n*

The P-value would be larger when $t = 1.20$ than when $t = 2.40$ because the *t*-value of 1.20 is less extreme.

9.31 Photovoltaic solar energy in Europe

a) The sample mean in this exercise ≈ 1.34.

b) *P*-value ≈ 0.15.

c) The relatively high level of 15% of the *P*-value suggests the non-rejection of H_0. At approximately 0, the *P*-value is below any reasonable significance level for the test.

💻9.32 Female work week

a) The relevant variable is the number of hours worked by females in the previous week; the parameter of interest is the mean number of hours, μ, worked by females in the United States in the previous week.

b) $H_0: \mu = 40$; $H_a: \mu \neq 40$

9.32 (continued)

c) $t = \dfrac{37.0 - 40}{15.1/\sqrt{583}} = -4.8;$ P-value ≈ 0. The P-value is the probability of observing a test statistic this extreme or even more extreme (equivalently, a sample mean this far away from the null value of 40 or even further away, in both directions) when the null hypothesis is true, the P-value is very small (less than 0.1%).

d) Because the P-value is less than the significance level of 0.01, we have sufficient evidence to reject the null hypothesis and conclude that the mean working week for females in the United States is different from 40 hours.

9.33 StatCrunch for statistics

a) The relevant variable in this study is the number of hours of study per week on StatCrunch. The parameter: mean number μ of hours of study per week on StatCrunch in the entire class.

b) The null hypothesis is H_0: $\mu = 7$, and the alternative hypothesis is H_a: $\mu > 7$.

c) $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{5.9 - 100}{4.99/\sqrt{15}} = -0.854$. The observed sample mean of 5.9 falls 0.854 standard errors below the null hypothesis value of 7.

d) *P*-value = probability in upper tail = 0.40. The probability of observing a test statistic this large or larger (equivalently a sample mean that is this number of standard errors high or even higher) when the null hypothesis is true is not that small (about 40%). This does not provide evidence to reject the null hypothesis. There is insufficient evidence to conclude that the mean number of hours of study per week on StatCrunch is larger than 7.

9.34 Lake pollution

a) $\bar{x} = \dfrac{2000 + 1000 + 3000 + 2000}{4} = \dfrac{8000}{4} = 2000$

$$s = \sqrt{\frac{(2000-2000)^2 + (1000-2000)^2 + (3000-2000)^2 + (2000-2000)^2}{4-1}} = \sqrt{\frac{2,000,000}{3}} = 816.5$$

$se = s/\sqrt{n} = 816.5/\sqrt{4} = 408.25$

b) $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{2000 - 1000}{816.5/\sqrt{4}} = 2.45$

c) The P-value is 0.046 for a one-sided test. This is smaller than 0.05, so we have enough evidence to reject the null hypothesis at a significance level of 0.05. We have relatively strong evidence that the wastewater limit is being exceeded.

d) The one-sided analysis in (b) implicitly tests the broader null hypothesis that $\mu \le 1000$. We know this because if it would be unusual to get a sample mean of 2000 if the population mean were 1000, we know that it would be even more unusual to get this sample mean if the population mean were less than 1000.

9.35 Weight change for controls

1) Assumptions: Random sample on a quantitative variable having a normal population distribution. Here, the data are not likely produced using randomization. Population distribution may be skewed, but the test is two-sided so it is robust to a violation of the normal population assumption.

2) Hypotheses: H_0: $\mu = 0$; H_a: $\mu \ne 0$

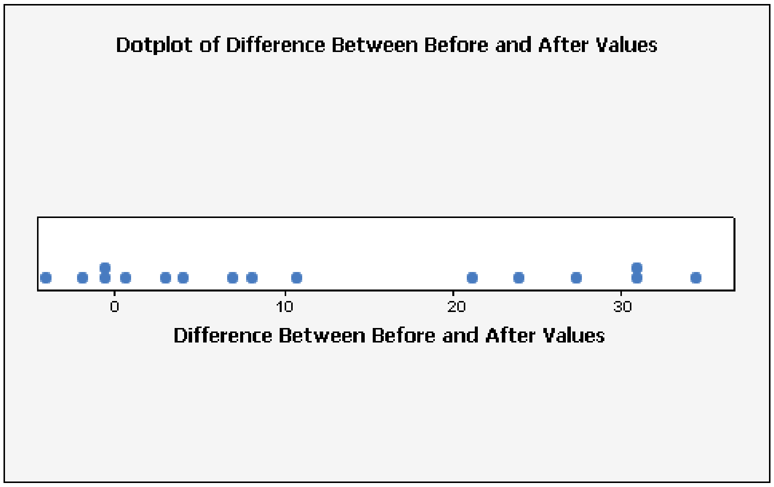
3) Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{-0.5 - 0}{8.0/\sqrt{26}} = -0.32$

4) P-value: 0.75

5) Conclusion: If the null hypothesis were true, the probability would be 0.75 of getting a test statistic at least as extreme as the value observed. Based on this large P-value, it is plausible that the null hypothesis is correct, and that there was no change in mean weight in the control group.

9.36 Water fluoridation

a) The difference scores are 31, 8.1, 10.8, 27.4, 0.6, –4.2, –0.6, –0.4, –2, 4, 3, 7, 21.2, 34.5, 24, 31. The sample size is small so it we cannot obtain from the plot, but there is no evidence of severe non-normality.



b) 1) Assumptions: *The data (NCF change scores) are randomly obtained from a normal population distribution*. Here, the data are not likely produced using randomization, but are likely a convenience sample. The two-sided test is robust if the population distribution is not normal.

2) Hypotheses: H_0: $\mu = 0$; H_a: $\mu \neq 0$

3) Test statistic (mean and standard deviation of sample calculated using software):

$$t = \frac{12.21 - 0}{13.62/\sqrt{16}} = 3.586.$$

4) *P*-value: 0.0027

5) Conclusion: If the null hypotheses were true, the probability would be 0.0027 of getting a test statistic at least as extreme as the value observed. There is strong evidence that NCF levels were lower before than after.

c) The assumption of random production of data does not seem valid for this example. A convenience sample limits reliability in applying this inference to the population at large.

9.37 Too little or too much wine?

1) Assumptions: The data are produced using randomization, from a normal population distribution.
2) Hypotheses: H_0: $\mu = 5.1$; H_a: $\mu \neq 5.1$
3) Test statistic: $t = \dfrac{5.065 - 5.1}{0.0870/\sqrt{4}} = -0.80$
4) P-value: 0.48
5) Conclusion: If the null hypothesis were true, the probability would be 0.48 of getting a test statistic at least as extreme as the value observed. There is not enough evidence to support that the true mean differs from 5.1 ounces.

9.38 Selling a burger

1) Assumptions: The data are produced using randomization, from a normal population distribution. The two-sided test is robust for the normality assumption.
2) Hypotheses: H_0: $\mu = 0$; H_a: $\mu \neq 0$
3) Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{3000 - 0}{4000/\sqrt{10}} = 2.37$
4) P-value: 0.04
5) Conclusion: If the null hypothesis were true, the probability would be 0.04 of getting a test statistic at least as extreme as the value observed. Because the P-value of 0.04 < 0.05, there is sufficient evidence that the coupons led to higher sales than did the outside posters.

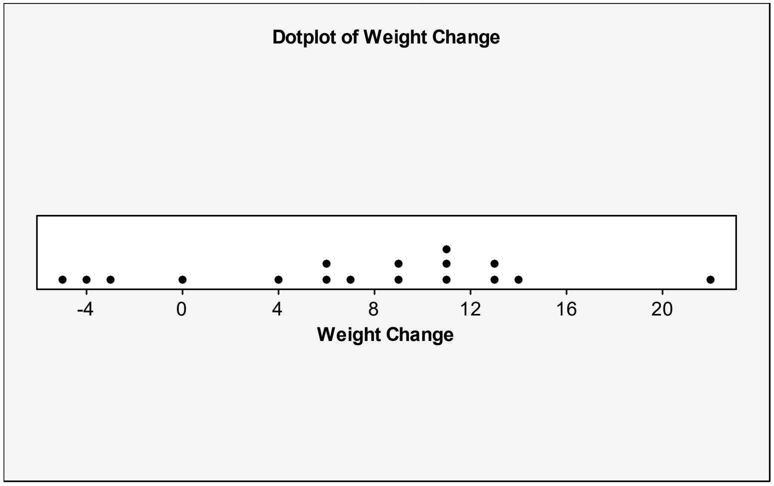
9.39 Assumptions important?

a) The confidence interval does not include 0 and also indicates that coupons led to higher sales than did the outside posters.

b) A one-sided test might be problematic. If the population distribution is highly non-normal (such as very skewed) the method is not robust for a one-sided test.

9.40 Anorexia in teenage girls

a) Most of the data fall between 4 and 14. The sample size is small so we cannot tell too much from the plot, but there is no evidence of severe non-normality.



b) Technology verifies these statistics.

c) 1) Assumptions: The data are quantitative and are produced randomly and the population distribution should be approximately normal.

2) Hypotheses: $H_0: \mu = 0$; $H_a: \mu \neq 0$

3) Test statistic: $t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{7.29 - 0}{7.18/\sqrt{17}} = 4.19$

4) P-value: 0.001

5) This extreme P-value suggests that we have strong evidence against the null hypothesis that family therapy has no effect. If the null hypothesis were true, the probability would be only 0.001 of getting a test statistic at least as extreme as the value observed.

9.41 Sensitivity study

From technology, after changing 20.9 to 2.9, the test statistic changes from 2.21 to 1.98, and the P-value changes from 0.04 to 0.06. The test statistic is less extreme, and the P-value is no longer smaller than 0.05. We can no longer reject the null hypothesis. The conclusion does depend on the single observation of 20.9.

9.42 Test and CI

Results of 99% confidence intervals are consistent with results of two-sided tests with significance levels of 0.01. A confidence interval includes the most plausible values for the population mean to a 99% degree of confidence. If the test rejects the null hypothesis with significance level 0.01, then the 99% confidence interval does not contain the value in the null hypothesis.

Section 9.4: Decisions and Types of Errors in Significance Tests

9.43 Dr. Dog

a) For the significance level of 0.05, we would reject the null hypothesis. We have strong evidence that dogs can detect urine from bladder cancer patients at a rate higher than would be expected by chance.

b) If we made an error, it was a Type I error. A Type I error would indicate that we concluded that dogs could detect urine from bladder cancer patients, but they really were not able to do so any better than chance.

9.44 Error probability

a) The probability of Type I error would be 0.05.

b) If *P*-value was 0.3, H_0 is not rejected. If this test resulted in a decision error, it was a Type II error.

9.45 Fracking errors

a) A Type I error would occur if we concluded that those opposing fracking are in the minority when in fact they are not.

b) A Type II error would occur if we failed to reject the null hypothesis when it was false. Thus, we determined that it was plausible that the null hypothesis was correct and said that there is no evidence of a minority of those who oppose fracking when, in fact, they are in the minority.

9.46 Anorexia errors

a) A Type I error would occur if we rejected the null hypothesis when it was true. Thus, we concluded that the therapy had an effect when in fact it did not.

b) A Type II error would occur if we failed to reject the null hypothesis when it was false. Thus, we determined that it was plausible that the null hypothesis was correct, that the therapy might have no effect, when in fact it does.

9.47 Anorexia decision

a) We would decide to reject the null. We would have strong evidence that the population mean weight change post-therapy is greater than 0.

b) If this decision were in error, it would be a Type I error.

c) If the significance level were instead 0.01, we would decide not to reject the null hypothesis. If this decision were in error, it would be a Type II error.

9.48 Errors in the courtroom

a) If H_0 is rejected, we conclude that the defendant is guilty.

b) A Type I error would result in finding the defendant guilty when he/she is actually innocent.

c) If we fail to reject H_0, the defendant is found not guilty.

d) A Type II error would result in failing to convict a defendant who is actually guilty.

9.49 Errors in medicine

a) If H_0 is rejected, we conclude that the new drug is not safe.

b) A Type I error would result in finding the new drug is not safe when it actually is safe.

c) If we fail to reject H_0, we conclude that the drug is safe.

d) A Type II error would result in failing to find that the new drug is not safe when it actually is not safe.

9.50 Decision errors in prostate cancer detection

a) A Type I error is a false positive because we have rejected the null hypothesis that there is no disease, but we were wrong. The man does not have prostate cancer. The consequence would be that the man would have treatment, or at least further testing, when he did not need any.

b) A Type II error is a false negative because we have failed to reject the null hypothesis that there is no disease, but we were wrong. The man does have prostate cancer. The consequence would be failing to detect cancer and treat the cancer when it actually exists.

c) The disadvantage of this tactic is that more men who do have prostate cancer will have false negative tests and not receive necessary treatment.

9.51 Detecting pregnancy

a) A Type I error would occur if we diagnose pregnancy when it is not present. This would be that a woman change her life style take some vitamins when she did not need any, and likely to incur some financial losses for additional medical tests or for buying products for the expected newborn.

b) A Type II error would occur if we fail to diagnose pregnancy when a woman is actually pregnant. This would mean that a woman who was actually pregnant would not receive necessary assistance on time for the delivery stage and for the care of the expected newborn.

c) The probability of 99% refers to the probability of a making a correct diagnostic, whether it is a correct positive or a correct negative.

9.51 (continued)

d) The 1% refers to the probability that a woman received a positive test result given that she is not pregnant or that a woman received a negative test result given that she is actually pregnant. This is the probability of a Type I error or a Type II error.

9.52 Which error is worse?

a) When rejecting the null results in the death penalty, a Type I error is worse than a Type II error. With a Type II error, a guilty man or woman goes free, whereas with a Type I error, an innocent man or woman is put to death.

b) When rejecting the null hypothesis results in treatment for breast cancer, a Type II error is worse than a Type I error. With a Type I error, someone might receive additional tests (e.g., biopsy) before ruling out breast cancer, but with a Type II error, someone might not receive life-saving treatment when they need it.

Section 9.5: Limitations of Significance Tests

9.53 Misleading summaries?

a) Researcher A: P-value = 2P(Z > 2.0) = 0.046

b) Researcher B: P-value = 2P(Z > 1.90) = 0.057

c) Researcher A's result has a P-value less than 0.05; thus, it is "statistically significant." Researcher B's P-value is not less than 0.05, and is not, therefore, "statistically significant." Results that are not different from one another in practical terms might lead to different conclusions if based on statistical significance alone.

d) If we do not see these two P-values, but merely know that one is statistically significant and one is not, we are not able to see that the P-values are so similar.

e) For A, the 95% confidence interval is $\hat{p} \pm z_{.025}\sqrt{\hat{p}(1-\hat{p})/n} = 0.550 \pm 1.96\sqrt{0.550(1-0.550)/400}$, or (0.501, 0.599).

For B: the 95% confidence interval is $\hat{p} \pm z_{.025}\sqrt{\hat{p}(1-\hat{p})/n} = 0.5475 \pm 1.96\sqrt{0.5475(1-0.5475)/400}$, or (0.499, 0.596).

This method shows the enormous amount of overlap between the two confidence intervals. The plausible values for the population proportions are very similar in the two cases, which we would not realize by merely reporting whether the null was rejected in a test.

9.54 Practical significance

a) Test statistic: $t = \dfrac{\bar{x}-\mu_0}{s/\sqrt{n}} = \dfrac{498-500}{100/\sqrt{25,000}} = -3.16$

b) P-value = 2P(Z < –3.16) = 0.0016

c) This result is statistically significant because the P-value is very small, but it is not practically significant because the sample mean of 498 is very close to the null hypothesis mean of 500.

9.55 Effect of *n*

a) Test statistic: $t = \dfrac{\bar{x}-\mu_0}{s/\sqrt{n}} = \dfrac{4.09-4.0}{1.43/\sqrt{25}} = 0.31$

b) This test statistic is associated with a P-value of 0.76. We cannot reject the null hypothesis; it is plausible that the null hypothesis is correct and that the population mean score is 4.0.

c) The 95% confidence interval is $\bar{x} \pm t_{.025}\left(s/\sqrt{n}\right) = 4.09 \pm 2.0639\left(1.43/\sqrt{25}\right)$, or (3.5, 4.7).

d) (i) This illustrates that a decrease in sample size increases the P-value greatly; a finding that might be statistically significant with a large sample size might not be with a small sample size.

(ii) In addition, confidence intervals become wider as sample sizes become smaller.

9.56 Fishing for significance

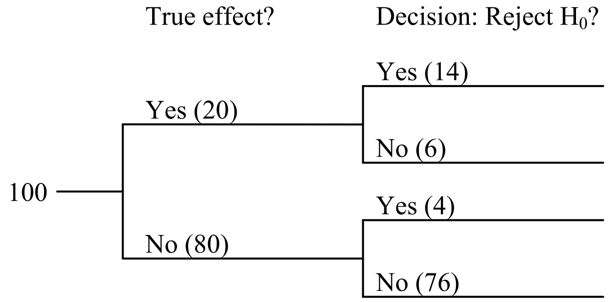
This is misleading because, with a significance level of 0.05, we would expect 5% of tests to be significant just by chance if the null hypothesis is true, and for 60 tests this is 0.05(60) = 3 tests.

9.57 Selective reporting

If we report only results that are “statistically significant,” these are the only ones of which the public becomes aware. There might be many studies on the same subject that did not reject the null and did not get published. The public is not able to identify situations in which only one of twenty or so studies on the same phenomenon had a significant finding. In such a case, the finding that did get published might be an example of a Type I error.

9.58 How many medical discoveries are Type I errors?

The following tree diagram is based on 100 studies.



The proportion of actual Type I errors (of cases where the null is rejected) would be about 4/(4 + 14) = 0.22.

9.59 Interpret medical research studies

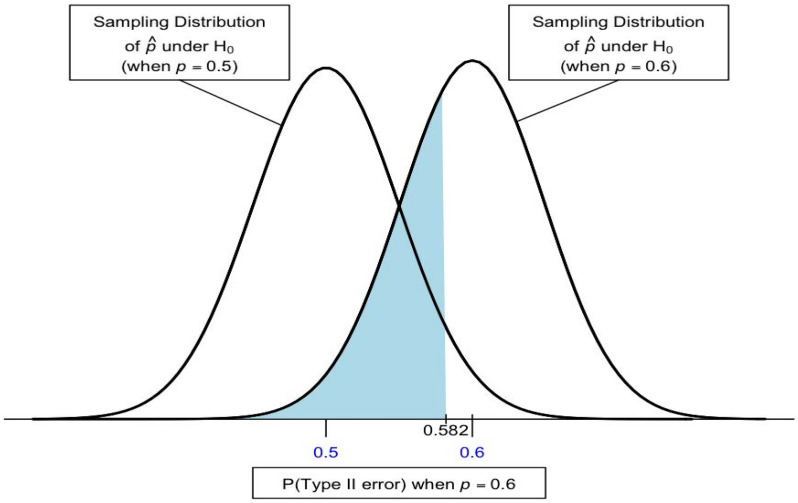
At a significance level of 5%, if the probability that each trial gives a false positive result is 1 in 20, then by testing 20 different colors it is now likely that at least one jelly bean test will give a false positive. To be precise, the probability of having no false positive in 20 tests is $(0.95)^{20}$ = 35.85%. Probability of having no false positive in 21 tests (counting the test without color discrimination) is $(0.95)^{21}$ = 34.06%. Because only the positive results get reported, this limits the value any single study has – especially if the mechanism linking the two things is not known. Research that suggests an impact of a therapy or drug tends to get the most media coverage if it is a very large difference. It is quite possible that studies that found a smaller difference did not get media coverage, or studies that failed to find a difference did not get published at all.

Section 9.6: The Likelihood of a Type II Error and the Power of a Test

9.60 Find P(Type II error)

a) A one-tailed test would have a *z*-score of 1.645 at the cutoff. Here, the standard error would be $\sqrt{\hat{p}(1-\hat{p})/n} = \sqrt{0.5(1-0.5)/100} = 0.050$. The value 1.645 standard errors above 0.50 is 0.50 + 1.645(0.050) = 0.582.

b)



9.60 (continued)

c) $z = \dfrac{(0.582 - 0.60)}{\sqrt{.60(1-.60)/100}} = -0.37$; Table A tells us that 0.36 falls beyond this *z*-score; thus, P(Type II error) = 0.36.

9.61 Gender bias in selecting managers

a) The cutoff *z*-score for a 0.05 significance level and a one-tailed test (in a negative direction) is –1.645. Here, the standard error would be $\sqrt{p_0(1-p_0)/n} = \sqrt{0.4(1-0.4)/50} = 0.069$. The value 1.645 standard errors below 0.069 is 0.4 – 1.645(0.069) = 0.286.

b) The standard error would now be $\sqrt{p_0(1-p_0)/n} = \sqrt{0.2(1-0.2)/50} = 0.0566$. If $p = 0.20$, the *z*-score for 0.286 in reference to 0.20 is $z = \dfrac{0.286 - 0.20}{0.0566} = 1.52$. If we look this *z*-score up on a table, we find that the proportion of this curve that is not in the rejection area is 0.06. Thus, the Type II error has probability of 0.06.

9.62 Balancing Type I and Type II errors

a) The cutoff for a 0.01 significance level and a one-tailed test is 2.33. Here, the standard error would be $\sqrt{p_0(1-p_0)/n} = \sqrt{(1/3)(1-1/3)/116} = 0.0438$. The value 2.33 standard errors above 0.333 is 0.333 + 2.33(0.0438) = 0.435.

b) If $p = 0.50$, the *z*-score for 0.435 in reference to 0.50 is $z = \dfrac{0.435 - 0.5}{\sqrt{0.5(1-0.5)/116}} = -1.40$. If we look this *z*-score up on a table, we find that the proportion of this curve that is not in the rejection area is 0.08.

9.63 P (Type II error) increase

a) We reject H_0 when we get a sample proportion that is $0.3 + 1.645\sqrt{\dfrac{0.3 \times 0.7}{200}}$(= 0.3533) or larger.

i. When $p = 0.4$, $z = \dfrac{0.3544 - 0.4}{\sqrt{0.4(1-0.4)/200}} = -1.32$, P(Type II error) = $P(z < -1.32) = 0.094$.

ii. When $p = 0.35$, $z = \dfrac{0.3544 - 0.35}{\sqrt{0.35(1-0.35)/200}} = 0.13$, P(Type II error) = $P(z < 0.13) = 0.552$.

b) We reject H_0 when we get a sample proportion that is $0.3 + 1.645\sqrt{\dfrac{0.3 \times 0.7}{100}}$(= 0.3754) or larger.

i. When $p = 0.4$, $z = \dfrac{0.3754 - 0.4}{\sqrt{0.4(1-0.4)/100}} = -0.5$, P(Type II error) = $P(z < -0.5) = 0.308$.

ii. When $p = 0.35$, $z = \dfrac{0.3754 - 0.35}{\sqrt{0.35(1-0.35)/100}} = 0.53$, P(Type II error) = $P(z < 0.53) = 0.703$.

c) When the parameter value is close to the value in H_0, there is not much difference between these two values. Thus, the chances of being able to reject the null hypotheses are smaller, and the chances of failing to reject are larger. If we fail to reject the null hypothesis, but the null hypothesis is not true, this is a type II error. As the parameter value moves away from H_0, the chances of rejecting the null hypothesis go up, and the chances of failing to reject (and therefore the chances of a type II error) go down. Furthermore, when the sample size decreases both the threshold of rejecting H_0 and the standard deviation of the proportion distribution for a given value of p increases. These lead to a lower value of the P(type II error).

9.64 Type II error with two-sided H$_a$

a) The standard error is $\sqrt{p_0(1-p_0)/n} = \sqrt{(1/3)(1-1/3)/116} = 0.0438$. For H$_a$, a test statistic of $z =$ 1.96 has a P-value (two-tail probability) of 0.05. We reject H$_0$ when $|\hat{p}-1/3| \geq 1.96(se) \Rightarrow$ $|\hat{p}-1/3| \geq 1.96(0.0438) \Rightarrow |\hat{p}-1/3| \geq 0.086,$ hence we need $\hat{p} \geq 0.086+1/3$ or $\hat{p} \leq 1/3-0.086$, or $\hat{p} \geq 0.419$ or $\hat{p} \leq 0.248$. When H$_0$ is false, a Type II error occurs if $0.248 < \hat{p} < 0.419$.

b) We can calculate z-scores for each of these proportions. $z = \dfrac{0.248-0.50}{0.0464} = -5.43;$ the probability that $\hat{p}$ is less than this z-score is 0. $z = \dfrac{0.419-0.50}{0.0464} = -1.75;$ the probability that $\hat{p}$ is greater than this z-score is 0.96.

c) The probability of a Type II error is the portion of the curve (for the parameter 0.50) that is not over the rejection area, so P(Type II error) = 1 – 0.96 = 0.04.

9.65 Power for knee osteoarthritis treatment

a) The null hypothesis is *H*$_0$: The Agilium Freestep AFO has no effect on the lever arm of the ground reaction force (GRF), and the alternative hypothesis is *H*$_a$: The Agilium Freestep AFO reduces the lever arm of the GRF.

b) With 80% probability, the proposed biomechanical study will lead to a rejection of the null hypothesis *H*$_0$ in favor of *H*$_a$.

c) Not rejecting *H*$_0$, when in fact the Agilium Freestep AFO helps to reduce the lever arm of the GRF.

9.66 Exploring Type II errors

a) As *n* increases, the probability of a Type II error decreases.

b) The sample size is around *n* = 85.

c) The probability will decrease.

d) The power increases.

Chapter Problems: Practicing the Basics

9.67 H$_0$ or H$_a$?

a) null hypothesis

b) alternative hypothesis

c) alternative hypothesis

d) null hypothesis

9.68 Write H$_0$ and H$_a$

a) p is the proportion of fathers taking parental leave. $\text{H}_0: p = 0.15; \text{H}_a: p > 0.15$

b) μ is the mean CO$_2$ emission from cars purchased last year. $\text{H}_0: \mu = 160; \text{H}_a: \mu < 160$

c) μ is the mean retirement age for female workers. $\text{H}_0: \mu = 60; \text{H}_a: \mu \neq 60$

9.69 ESP

1) Assumptions: The data are categorical (correct versus incorrect guesses) and are obtained randomly. The expected successes and failures are both fifteen under H$_0$; $np_0 = n(1-p_0) = 30(0.5) = 15,$ so the sample size condition is met.

2) Hypotheses: $\text{H}_0: p = 0.5; \text{H}_a: p > 0.5,$ p is the probability of a correct guess.

3) Test statistic: $z = \dfrac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{18/30-0.50}{\sqrt{0.5(1-0.5)/30}} = 1.1$

4) P-value: 0.137.

5) Conclusion: If the null hypothesis were true, the probability would be 0.137 of getting a test statistic at least as extreme as the value observed. Fail to reject H$_0$, there is not strong evidence with this P-value that the probability of a correct guesses is higher than 0.50.

9.70 Free throw accuracy

a) Assumptions: The data are categorical (make first only versus make second only) and are obtained randomly; the expected successes and failures are both at least fifteen under H_0; $np_0 = n(1 - p_0) = 82(0.5) = 41 \geq 15$.

Hypotheses: $H_0: p = 0.5$; $H_a: p \neq 0.5$, p is the proportion of pairs of shots in which only one shot was made in which the first shot went in.

b) Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \dfrac{0.415 - 0.50}{\sqrt{0.5(1 - 0.5)/82}} = -1.55$

c) The P-value is 0.12; If the null hypothesis were true, the probability would be 0.12 of getting a test statistic at least as extreme as the value observed. Fail to reject H_0, it is plausible that the null hypothesis is correct and that the population proportion of first free shots made (out of all pairs in which only one shot was made) is 0.50.

9.71 2016 Irish Exit Poll

a) 1) Assumptions: The data are categorical (Fine Gael, Fianna Fáil or others) and obtained randomly; the expected successes and failures are both at least fifteen under H_0; $np = (0.25)(5260) \geq 15$, and $n(1 - p) = (0.25)(5260) \geq 15$.

2) p = population proportion of voters for Fianna Fáil. Hypotheses: $H_0: p = 0.25$; $H_a: p \neq 0.25$.

3) Test statistic: $z = \dfrac{0.229 - 0.25}{\sqrt{0.25(1 - 0.25)/5260}} = -3.52$.

4) P-value: 0.0005

5) Conclusion: We can reject the null hypothesis at a significance level of 0.05; we have strong evidence that the population proportion of voters who chose Fianna Fáil is different from 0.25.

b) If the sample size had been 500, the test statistic would have been $z = \dfrac{0.229 - 0.250}{\sqrt{0.25(1 - 0.25)/500}} = -1.08$, and the P-value would have been 0.30. We could not have rejected the null hypothesis under these circumstances.

c) The result of a significance test can depend on the sample size. As the sample size increases, the standard error decreases (because the sample size is the denominator of the standard error equation; dividing by a larger number leads to a smaller result). A smaller standard error leads to a larger z-score and a smaller P-value.

9.72 Protecting the environment?

a) The assumptions are that the data are categorical, that they are obtained using randomization, and that the sample size is large enough that the sampling distribution of the sample proportion is approximately normal (i.e., expected successes and failures both at least fifteen under H_0; $np_0 = n(1 - p_0) = 1085(0.5) = 542.5 \geq 15$. The data are categorical (yes, no), the text tells us that we can assume that the GSS data are randomly obtained, and the sample size is large enough.

b) Hypotheses: $H_0: p = 0.5$; $H_a: p \neq 0.5$; the point estimate of p is $\hat{p} = 0.423$. The value of the test statistic is –5.07.

c) The P-value $\approx$ 0.000, indicating that $\hat{p}$ is quite extreme. If the null hypothesis were true, the probability would be close to 0 of getting a test statistic at least as extreme as the value observed. There is strong evidence that the population proportion who would answer yes is different from 0.50.

d) It is not plausible that $p = 0.50$. The test statistic is too extreme for 0.50 to be plausible.

e) The advantage of the confidence interval is that it provides a range of plausible values.

9.73 Majority supports gay marriage

1) Assumptions: The data are categorical (yes versus no); the sample is a random sample of 1690 adults; the expected number of yes and no responses are both at least 15 under H$_0$; $np_0 = n(1-p_0) = 1690(0.5) = 854 \geq 15$.
2) Hypotheses: $\mathrm{H}_0: p = 0.5$; $\mathrm{H}_a: p \neq 0.5$, p is the proportion of Americans agreeing that homosexuals should be able to marry.
3) Test statistic: $z = \dfrac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{955/1690-0.5}{\sqrt{0.5(1-0.5)/1690}} = 5.35$
4) P-value: 0.000
5) Conclusion: At the 0.05 significance level, we have strong evidence (P-value < 0.001) that the proportion of Americans agreeing with the statement that homosexuals should have the right to marry is different from 0.5. With a sample proportion of 0.565, a clear majority now supports gay marriage.

9.74 Plant inheritance

1) Assumptions: The data are categorical (green versus yellow) and are obtained randomly; the expected successes and failures are both at least fifteen under H $_a$; $np_0 = 1103(0.75) = 827.25 \geq 15$ and $n(1-p_0) = 1103(0.25) = 275.75 \geq 15$.
2) Hypotheses: $\mathrm{H}_0: p = 0.75$; $\mathrm{H}_a: p \neq 0.75$, p is the proportion of green seedlings.
3) Test statistic: $z = \dfrac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{854/1103-0.75}{\sqrt{0.75(1-0.75)/1103}} = 1.86$
4) P-value: 0.06
5) Conclusion: If the null hypothesis were true, the probability would be 0.06 of getting a test statistic at least as extreme as the value observed. Fail to reject H$_0$, it is plausible that the null hypothesis is correct.

9.75 Zika Virus

a) The population is the set of people in the United States during January 3–March 5, 2016, who traveled to or moved from areas with active Zika virus transmission. p = proportion of people who would report no signs or symptoms, Hypotheses: H_0: $p = 0.50$; H_a: $p > 0.50$.
b) Test statistic: $z = \dfrac{0.564-0.5}{\sqrt{0.5(1-0.5)/4534}} = 8.62$; *P*-value < 0.0001; there is very strong evidence that the population proportion of those report no signs or symptoms is higher than 0.50.

9.76 Start a hockey team

1) Assumptions: The data are categorical (male versus female) and are obtained randomly; the expected successes and failures are both at least fifteen under H$_0$; $np_0 = 100(0.55) = 55 \geq 15$ and $n(1-p_0) = 100(0.55) = 45 \geq 15$.
2) Hypotheses: $\mathrm{H}_0: p = 0.55$; $\mathrm{H}_a: p \neq 0.55$, p is proportion of university students that are male.
3) Test statistic: $z = \dfrac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{80/100-0.55}{\sqrt{0.55(1-0.55)/100}} = 5.0$
4) P-value: 0.0000
5) Conclusion: If the null hypothesis were true, the probability would be close to 0 of getting a test statistic at least as extreme as the value observed. There is extremely strong evidence that the probability of selecting a male was higher than 0.55 and thus the sample was not random.

9.77 Interest charges on credit card

The output first tells us that we are testing "$p = 0.50$ versus not $p = 0.50$." That tells us that the null hypothesis is that the two cards are preferred equally and the alternative hypothesis is that one is preferred more than the other. The printout then tells us that X is 40. This is the number of people in the sample who preferred the card with the annual cost. 100 is the "N," the size of the whole sample. The "Sample p" of 0.40000 (rounds to 0.40) is the proportion of the sample that preferred the card with the annual cost. The "95.0% CI" is the 95% confidence interval, the range of plausible values for the population proportion. The "Z-Value" is the test statistic. The sample proportion is 2.00 standard errors below the proportion as per the null hypothesis, 0.50. Finally, the "P-Value" tells us that if the null hypothesis were true, the proportion 0.0455 of samples would fall at least this far from the null hypothesis proportion of 0.50. This is barely extreme enough to reject the null hypothesis with a significance level of 0.05. We have evidence that the population proportion of people who prefer the card with the annual cost is below 0.50. The majority of the customers seem to prefer the card without the annual cost, but with the higher interest rate.

9.78 Jurors and gender

a) Hypotheses: $H_0: p = 0.53$; $H_a: p \neq 0.53$, p is proportion of jurors who are women.

b) Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{0.125 - 0.53}{\sqrt{0.53(1-0.53)/40}} = -5.1$

c) The P-value is 0.000. If the null hypothesis were true, the probability would be almost 0 of getting a test statistic at least as extreme as the value observed.

d) This P-value is more extreme than the significance level of 0.01. We can reject the null hypothesis; we have strong evidence that women are not being selected in numbers proportionate to their representation in the jury pool.

9.79 Type I and Type II errors

a) In the previous exercise, a Type I error would have occurred if we had rejected the null hypothesis, concluding that women were being passed over for jury duty, when they really were not. A Type II error would occur if we had failed to reject the null, but women really were being picked disproportionate to their representation in the jury pool.

b) If we made an error, it was a Type I error.

9.80 Levine = author?

a) 1) Assumptions: The data are categorical (whereas versus not whereas) and are obtained randomly; the expected successes and failures are both at least fifteen under H_0; $np_0 = 300(0.10) = 30 \geq 15$ and $n(1-p_0) = 300(0.90) = 270 \geq 15$.

2) Hypotheses: $H_0: p = 0.10$; $H_a: p \neq 0.10$, p is proportion of sentences beginning with *whereas*.

3) Test statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{0.00 - 0.10}{\sqrt{0.10(1-0.10)/300}} = -5.8$

4) P-value: 0.000

5) Conclusion: If the population proportion is 0.10, we would expect a sample proportion this extreme almost none of the time. Reject H_0, it seems unlikely that Levine wrote this document.

b) The assumptions for this conclusion to be valid are in part (a)-(1).

💻9.81 Practice steps of test for mean

a) (i) $\bar{x} = \dfrac{3+7+3+3+0+8+1+12+5+8}{10} = \dfrac{50}{10} = 5.0$

(ii) $s = \sqrt{\dfrac{(3-5)^2 + (7-5)^2 + \cdots + (5-5)^2 + (8-5)^2}{10-1}} = \sqrt{\dfrac{124}{9}} = 3.71$

(iii) $se = s/\sqrt{n} = 3.71/\sqrt{10} = 1.17$

(iv) $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{5.0 - 0}{3.71/\sqrt{10}} = 4.26$

(v) $df = n - 1 = 10 - 1 = 9$

9.81 (continued)

b) The P-value of 0.002 is less than the significance level of 0.05. We can reject the null hypothesis. We have very strong evidence that the population mean is not 0.

c) If we had used the one-tailed test, $H_a: \mu > 0,$ the P-value would be 0.002/2 = 0.001, also less than the significance level of 0.05. Again, we have very strong evidence that the population mean is positive.

d) If we had used the one-tailed test, $H_a: \mu < 0,$ the P-value would be 1 – 0.001 = 0.999, far from the significance level of 0.05. It would be plausible that the null hypothesis is correct; we cannot conclude that the population mean is negative.

9.82 Two ideal children?

a) The test statistic value is $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{2.518 - 2}{0.875/\sqrt{1417}} = 22.29.$

b) The P-value is the probability of observing a *t* test statistic as extreme as the one observed (22.29) given the null hypothesis is true. The P-value is extremely small, so observing such a test statistic (and sample mean) is extremely unlikely if H_0 is true. We reject H_0 and conclude that the ideal number of children is different from two.

9.83 Hours at work

a) Hypotheses: $H_0: \mu = 40$; $H_a: \mu \neq 40$

b) (i) SE Mean, 0.445, is the standard error of the sampling distribution of the mean.
(ii) T = 0.60, is the test statistic, the distance (measured in standard errors) of the sample mean of 40.27 from the null value of 40 hours.
(iii) The P-value, 0.548, is the probability of observing a sample mean of 40.27 or more extreme (on either side) when the null hypothesis is true. This is rather large, so the sample mean is not unusually extreme if H_0 is true. There is no evidence to reject the null hypothesis.

c) The confidence interval shows that 40 is a plausible value for the hours in a workweek in the population. This is consistent with result of the hypothesis test, which does not reject $H_0: \mu = 40.$

9.84 Females liberal or conservative?

1) Assumptions: The data are quantitative, have been produced randomly, and have an approximate normal population distribution.

2) Hypotheses: $H_0: \mu = 4.0$; $H_a: \mu \neq 4.0$

3) Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{4.06 - 4.00}{1.37/\sqrt{1345}} = 1.61,$ (Note: when using technology with the original data, a more precise value is $t = 1.72$.)

4) P-value: 0.11; (0.08 using technology.)

5) Conclusion: With a significance level of 0.05, we would not reject H_0 (P-value = 0.11 > 0.05). We have insufficient evidence to support the claim that the mean rating for females has changed from 4.

9.85 Blood pressure

a) 1) Assumptions: The data are quantitative, have been produced randomly, and have an approximate normal population distribution.

2) Hypotheses: $H_0: \mu = 130$; $H_a: \mu \neq 130$

3) Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{150 - 130}{8.37/\sqrt{6}} = 5.85$

4) P-value: 0.002

5) Conclusion: If the null hypothesis were true, the probability would be 0.002 of getting a test statistic at least as extreme as the value observed. There is very strong evidence that the population mean is different from 130; reject H_0, we can conclude that Vincenzo Baronello’s blood pressure is not in control.

b) The assumptions are outlined in Step 1 in (a). Blood pressure readings are quantitative data. These data are the last six times he monitored his blood pressure. This might be considered a random sample of possible readings for that point in time. We do not know whether the population distribution is normal, but the two-sided test is robust for violations of this assumption.

9.86 Increasing blood pressure

H_a: $\mu > 130$; the P-value is 0.002/2 = 0.001, which is more extreme. We can still conclude that the blood pressure is not in control.

💻**9.87 Tennis balls in control?**

a) Technology indicates a test statistic of $t = –5.5$ and a P-value of 0.001.
b) For a significance level of 0.05, we would conclude that the process is not in control. The machine is producing tennis balls that weigh less than they are supposed to.
c) If we rejected the null hypothesis when it is in fact true, we have made a Type I error and concluded that the process is not in control when it actually is.

9.88 Catalog sales

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{10-15}{10/\sqrt{100}} = -5.0;\ \text{P-value} \approx 0.000$$

If the null hypothesis were true, the probability would be close to 0 of getting a test statistic at least as extreme as the value observed. There is strong evidence that mean sales for the catalog differed from the mean of $15 from past catalogs.

💻**9.89 Wage claim false?**

1) Assumptions: The data are quantitative. The data seem to have been produced using randomization. We also assume an approximately normal population distribution.
2) Hypotheses: H_0: $\mu = 500$; H_a: $\mu \neq 500$
3) Test statistic: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{441.11 - 500}{12.69/\sqrt{9}} = -13.9$
4) P-value: 0.000
5) Conclusion: If the null hypothesis were true, the probability would be almost 0 of getting a test statistic at least as extreme as the value observed. Reject H_0, there is extremely strong evidence that the population mean is different than 500; with a sample mean of 441.11, we can conclude that the mean income is less than $500 per week.

9.90 CI and test

a) We can reject the null hypothesis for any of these significance levels (i.e., 0.10, 0.05, and 0.01).
b) None of the intervals based on the three confidence levels (i.e., 0.90, 0.95, and 0.99) would contain 500.
c) If a finding is statistically significant, then the confidence interval associated with that significance level will not include the value in H_0.
d) (i) A Type I error would occur if we conclude that the mean income for all senior-level assembly-line workers is different from $500 when it actually is not.
   (ii) A Type II error would occur if we fail to reject the null hypothesis that the mean income for all senior-level assembly-line workers is $500 when it actually is different from $500.

9.91 CI and test connection

a) We can reject the null hypothesis.
b) It would be a Type I error.
c) A 95% confidence interval would not contain 100. When a value is rejected by a test at the 0.05 significance level, it does not fall in the 95% confidence interval.

9.92 Net migration of EU citizens

a) We could explain that there is a less than 5% chance that we would find a net migration rate this extreme if there had been no change in net migration for the EU citizens. It is unlikely that there has been no change.
b) It would have been informative to have the actual *P*-value because not only would we know that there is a smaller than 0.05 chance of finding a rate this extreme if there had been no change, but we would also know the actual probability of finding a rate at least this extreme.
c) We cannot conclude that a practically important change in EU citizens' net migration has occurred because significance levels only tell us that a change has occurred beyond chance; it does not tell us the size of that change in practical terms. However, in this case, it is given that the size of that change is about 10,000 additional migrants.

9.93 How to reduce chance of error?

a) The researcher can control the probability of a Type I error by choosing a smaller significance level. This will decrease the probability of a Type I error.

b) If a researcher sets the probability equal to 0.00001, the probability of a Type I error is low, but it will be extremely difficult to reject the null hypothesis, even if the null hypothesis is not true.

9.94 Legal trial errors

a) A Type I error in a trial setting would occur if we convicted a defendant who was not guilty. A Type II error would occur if we failed to convict a guilty defendant.

b) To decrease the chance of a Type I error, we could decrease the significance level. In doing this, it is more difficult to reject the null hypothesis (i.e., find someone guilty). Thus, there will be more guilty people who are not found guilty, a Type II error.

9.95 P(Type II error) with smaller *n*

a) The standard error is $\sqrt{p_0(1-p_0)/n} = \sqrt{(1/3)(1-1/3)/60} = 0.061$.

b) When P(Type I error) = significance level = 0.05, $z = 1.645$, and the value 1.645 standard errors above 1/3 is 1/3 + (1.645)(0.061) = 0.433.

c) The standard error is now $\sqrt{p_0(1-p_0)/n} = \sqrt{0.5(1-0.5)/60} = 0.0645$ and $z = \dfrac{0.433-0.5}{0.0645} = -1.03$.

The probability that $\hat{p}$ falls below this *z*-score is 0.15. The Type II error is larger when *n* is smaller, because a smaller *n* results in a larger standard error and makes it more difficult to have a sample proportion fall in the rejection region. If we're less likely to reject the null with a given set of proportions, we're more likely to fail to reject the null when we should reject it.

Chapter Problems: Concepts and Investigations

💻9.96 Student data

a) The P-value for this significance test is 0.000. If the null hypothesis were true, the probability would be almost 0 of getting a test statistic at least as extreme as the value observed; we have strong evidence that the population mean political ideology is not 4.0.

b) The P-value for this significance test is 0.001. If the null hypothesis were true, the probability would be almost 0 of getting a test statistic at least as extreme as the value observed; we have strong evidence that the population proportion favoring affirmative action differs from 0.50.

The one page report will be different for each student.

💻9.97 Class data

The report will be different for each student.

💻9.98 Gender of best friend

The short reports will vary, but will report the following information:

$z = \dfrac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \dfrac{0.1064-0.5}{\sqrt{0.5(1-0.5)/1381}} = -29.25$, P-value = 0.000, 95% confidence interval: (0.08, 0.13) for the proportion having an opposite gender best friend. The confidence interval is more informative than the significance test. Not only does it tell us that the value at the null hypothesis (0.50) is implausible; it actually gives us a range of plausible values.

9.99 NBA home court advantage

$z = \dfrac{0.537-0.685}{\sqrt{0.685(1-0.685)/1230}} = -11.17$. *P*-value = 0.0095. Since the *P*-value is too small, we reject the null hypothesis and conclude that the home team's win percentage has significantly changed from 1976–77 to 2014–15. We can add and subtract the margin of error, (1.96) (0.0143) = 0.0279, from the sample proportion to obtain a 95% confidence for the population proportion. The interval is (0.5091, 0.5649). The interval also supports the alternative hypothesis since the hypothesized value of 0.685 (the 1976–77's value) is not in the confidence interval. The test merely indicates whether $p = 0.685$ is plausible whereas the confidence interval displays the range of plausible values.

9.100 Statistics and scientific objectivity

We can examine experimentally the claims of quack scientists, such as astrologers. If we fail to find a statistically significant effect (and this is replicated over and over), we start to get information that a given effect does not exist.

9.101 Two-sided or one-sided?

a) Once a researcher sees the data, he or she knows in which direction the results lie. At this point, it is "cheating" to decide to a do a one-tailed test. In this scenario, one has actually done a two-tailed test, then cut the P-value in half upon seeing the results, making it easier to reject the null hypothesis. The decision of what type of test to use must be made before seeing the data.

b) A result that is statistically significant with a P-value of 0.049 is not greatly different from one that is not statistically significant with a P-value of 0.051. The decision to use such a cutoff is arbitrary, is dependent on sample size, and is dependent on the random nature of the sample. Such a policy leads to the inflation of significant findings in the public view. If there really is no effect, but many studies are conducted, eventually someone will achieve significance, and then the journal will publish a Type I error.

9.102 No significant change and P-value

The mean change in glucose level that this particular study found is not unusual. If, in fact, there is no effect of using a cell phone on the glucose levels (i.e., if the true mean change is zero), observing such a mean change or an even more dramatic one happens 63% of the time.

9.103 Subgroup lack of significance

The sample size (n) has an impact on the P-value. The subgroups have smaller sample size, so for a particular size of effect will have a smaller test statistic and a larger P-value.

9.104 Curcumin and muscle soreness

Given a significance level of 0.05, we are going to get a test statistic in the rejection region 5% of the time when the null hypothesis is true. Thus, in every twenty studies, we are likely to have one that is a Type I error. Type I errors, particularly those with larger effects, tend to get more exposure both in peer-reviewed journals and by the media. Thus, the early published research in a given area is likely to be that found a significant effect. Only after the public becomes interested in it, might we then hear about other studies with null results.

9.105 Overestimated effect

The studies with the most extreme results will give the smallest P-values and be most likely to be statistically significant. If we could look at how results from all studies vary around a true effect, the most extreme results would be out in a tail, suggesting an effect much larger than it actually is.

9.106 Choosing α

a) We might prefer a smaller significance level because we want to diminish the chances of a Type I error. We don't want to run the risk of taking people off a drug that works until we are sure that our drug works at least as well.

b) The disadvantage of a smaller significance level is that it is more difficult to reject the null hypothesis. There's a higher chance of making a Type II error, and failing to find support for a drug that actually is better.

9.107 Why not accept H_0?

When we do not reject H_0, we should not say that we accept H_0. Just because the sample statistic was not extreme enough to conclude that the value in H_0 is unlikely doesn't mean that the value in H_0 is the actual value. As a confidence interval would demonstrate, there is a whole range of plausible values for the population parameter, not just the null value.

9.108 Report P-value

Knowing the exact P-value is more informative and, often, less misleading. For example, a very small P-value, such as 0.0001, tells us that this significance test provided stronger evidence than if the P-value were just beyond the cutoff, such as 0.049. On the other hand, if one P-value is just beyond the cutoff and one just barely fails to fall beyond the cutoff, we know that in practical terms they are giving about the same amount of evidence, even though one allows us to reject the null hypothesis and one does not.

9.109 Significance

Statistical significance means that we have strong evidence that the true parameter value is either above or below the value in H_0; this need not indicate practical significance. Practical significance means that the true parameter is sufficiently different from the value in H_0 to be important in practical terms. Examples will vary.

9.110 More doctors recommend

a) The company could conduct a significance test of the null hypothesis $H_0: p = 0.75$ vs. $H_a: p > 0.75$, and support their claim by quoting the P-value for this test.

b) (i) If this claim is based on a random sample of 40 doctors, it is more impressive than if it is based on 4 doctors. A higher *n* means a smaller standard error, and therefore, a larger test statistic. A larger test statistic is more likely to fall in the rejection region.

(ii) If this claim is based on a random sample of 40 doctors nationwide, it would be more impressive than if based on a sample of all 40 doctors who work in a particular hospital. In the former case, we are more able to generalize. If all of the doctors are in one hospital, we do not know if this finding would be true for the whole population. A more representative sample provides stronger evidence.

9.111 False-positive biopsy

With the probability of a 10-year cumulative risk for a false-positive biopsy being about 7%, it would not be unusual for a woman to receive a false positive over the course of having had many mammography screenings. Likewise, when conducting many significance tests with a Type I error of 0.05, it would not be unusual to have some show statistical significance (i.e., support the alternative hypothesis) even though the null hypothesis is in fact true.

9.112 Bad P-value interpretations

Proper interpretation of the P-value: If the null hypothesis is correct (and the population mean is 100) there is a 0.057 chance that we would obtain a sample mean at least this far from the population mean (a sample mean at least as far from the population mean as 104 is). This is pretty unlikely, but not beyond the typical cutoff value of 0.05.

a) 0.057 is not the probability that the null hypothesis is correct; it's the probability that we would obtain a sample mean at least this extreme, <u>if</u> the null hypothesis is correct. We calculate probabilities for test statistic values – not for hypotheses about parameters. We actually never know whether the null hypothesis is or isn't true.

b) $\bar{x}$ is the sample mean. The probability that the sample mean equals 104, regardless of whether the null hypothesis is true, is 100%. The researchers would have calculated the <u>actual</u> sample mean from the actual sample data. This is not an inference. The probability that we would obtain a sample mean of at least as extreme as 104 if the null hypothesis is true is 0.057.

c) This probability of 0.057 refers to the likelihood of getting a sample at least this extreme if the null hypothesis = 100, not if the null hypothesis does <u>not</u> equal 100.

d) The probability of a Type I error is the probability to which we set α. If we were to set α to a level of 0.05, the most common level, the probability of a Type I error would be 0.05.

e) It is never a good idea to accept H_a because, even though our P-value of 0.057 is not smaller than 0.05, it is very possible that the population mean is not 100. Remember – the null hypothesis contains only a single value! If we were to set up a 95% confidence interval around the sample mean of 104, we would see a lot of plausible values that the population mean could be.

f) In order to reject H_a at the $\alpha = 0.05$ level, the P-value would have to be less than 0.05, but this P-value is greater than 0.05.

9.113 Interpret P-value

The P-value tells us the probability of getting a test statistic this large if the value at the null hypothesis represents the true parameter. In this case, it is 0.057. We can reject the null hypothesis if the P-value is at or beyond the significance level α. If it were any number below this, there would not be enough evidence.

9.114 Incorrectly posed hypotheses

This notation is for sample statistics, not population parameters. Hypotheses are about populations, not samples, and therefore use parameters, not statistics.

9.115 Multiple choice: Small P-value

The best answer is (b).

9.116 Multiple choice: Probability of P-value

The best answer is (a).

9.117 Multiple choice: Pollution

The best answer is (a).

9.118 Multiple choice: Interpret P(Type II error)

The best answer is (c).

9.119 True or false

False

9.120 True or false

True

9.121 True or false

False

9.122 True or false

True

9.123 True or false

False

9.124 True or false

True

9.125 True or false

False

9.126 True or false

True

♦♦9.127 Standard error formulas

If the sample probability is 0, then the standard error is 0, and the test statistic is infinity, which does not make sense. A significance test is conducted by supposing the null is true, so in finding the test statistic we should substitute the null hypothesis value, giving a more appropriate standard error.

♦♦9.128 Rejecting true H_0?

a) The distribution is binomial, with $n = 100$, $p = 0.05$. We would expect the researcher to reject the null hypothesis about $np = 100(0.05) = 5$ times.

b) If she rejects the null hypothesis in five out of 100 tests, it is plausible that the null hypothesis is correct in every case. We'd expect about 5 rejections merely by chance when the null is true each time.

Chapter Problems: Student Activities

9.129

The results will be different for each class.

9.130

The results will be different for each class.