Chapter 19 – More About Tests and Intervals

Section 19.1

1. **True or false**

   a) False. It provides evidence against it but does not show it is false.

   b) False. The P-value is not the probability that the null hypothesis is true.

   c) True.

   d) False. Whether a P-value provides enough evidence to reject the null hypothesis depends on the risk of a Type I error that one is willing to assume (the α level).

2. **False or true**

   a) False. It's likely that you'll get a reasonably high P-value, but not sure.

   b) True.

   c) False. It says nothing about the probability that the null hypothesis is true.

   d) False. It simply provides no evidence that it's false.

3. **P-values.**

   a) False. A high P-value shows that the data are consistent with the null hypothesis, but provides no evidence for rejecting the null hypothesis.

   b) False. It results in rejecting the null hypothesis, but does not prove that it is false.

   c) False. A high P-value shows that the data are consistent with the null hypothesis but does not prove that the null hypothesis is true.

   d) False. Whether a P-value provides enough evidence to reject the null hypothesis depends on the risk of a type I error that one is willing to assume (the α level).

4. **More P-values.**

   a) True.

   b) False. A high P-value shows that the data are consistent with the null hypothesis but does not prove that the null hypothesis is true.

   c) False. No P-value ever shows that the null hypothesis is true (or false).

   d) False. If the null hypothesis is true, you will get a P-value below 0.01 about once in a hundred hypothesis tests.

Section 19.2

5. **Alpha true and false.**

   a) True.

   b) False. The alpha level is set independently and does not depend on the sample size.

   c) False. The P-value would have to be less than 0.01 to reject the null hypothesis.

   d) False. It simply means we do not have enough evidence at that alpha level to reject the null hypothesis.

6. **Alpha false and true.**

   a) False. A lower alpha level lowers the probability of a Type I error, but increases the probability of a Type II error.

   b) True. **c)** True.

   d) True.

Copyright © 2020 Pearson Education, Inc.

7. Critical values.

a) $z = \pm 1.96$

b) $z = 1.645$

c) $t = \pm 2.03$

d) $z = 2.33$ (n is not relevant for values of z^*)

e) $z = -2.33$

8. More critical values.

a) $t = \pm 2.00$

b) $z = 1.645$

c) $z = \pm 2.58$

d) $z = -2.33$ (n is not relevant for values of z^*)

e) $z = -2.33$

Section 19.3

9. Significant?

a) If 98% of children have really been vaccinated, there is practically no chance of observing 97.4% of children (in a sample of 13,000) vaccinated by natural sampling variation alone.

b) We conclude that the proportion of children who have been vaccinated is below 98%, but a 95% confidence interval would show that the true proportion is between 97.1% and 97.7%. Most likely a decrease from 98% to 97.7% would not be considered important. The 98% figure was probably an approximate figure anyway. However, if the 98% figure was not as estimate, and with 1,000,000 kids per year vaccinated, even 0.1% represents 1,000 kids, so this may be important.

10. Significant again?

a) If 15.9% is the true percentage of children who did not attain the grade level standard, there is only a 2.3% chance of observing 15.1% of children (in a sample of 8500) not attaining grade level by natural sampling variation alone.

b) Under old methods, 1352 students would not be expected to read at grade level. With the new program, 1284 would not be expected to read at grade level. This is only a decrease of 68 students. The costs of switching to the new program might outweigh the potential benefit. It is also important to realize that this is only a *potential* benefit.

Section 19.4

11. Errors.

a) Type I error. The actual value is not greater than 0.3 but they rejected the null hypothesis.

b) No error. The actual value is 0.50, which was not rejected.

c) Type II error. The actual value was 55.3 points, which is greater than 52.5.

d) Type II error. The null hypothesis was not rejected, but it was false. The true relief rate was greater than 0.25.

12. More errors.

a) Type I error. The actual mean was not greater than 25.

b) No error. The actual proportion is greater than 0.80 so they were correct in not rejecting the null hypothesis.

c) No error. The actual proportion is not equal to 0.5.

d) Type II error. They should have rejected the null hypothesis since 0.60 is less than 0.70.

Chapter Exercises

13. P-value.

If the effectiveness of the new poison ivy treatment is the same as the effectiveness of the old treatment, the chance of observing an effectiveness this large or larger in a sample of the same size is 4.7% by natural sampling variation alone.

Copyright © 2020 Pearson Education, Inc.

14. Another P-value.

If harsher penalties and ad campaigns have made no difference in seat-belt use, there is a 17% chance of seeing an observed difference this large or larger by natural sampling variation.

15. Alpha.

Since the null hypothesis was rejected at $\alpha = 0.05$, the P-value for the researcher's test must have been less than 0.05. He would have made the same decision at $\alpha = 0.10$, since the P-value must also be less than 0.10. We can't be certain whether or not he would have made the same decision at $\alpha = 0.10$, since we only know that the P-value was less than 0.05. It may have been less than 0.01, but we can't be sure.

16. Alpha again.

Since the environmentalists failed to reject the null hypothesis at $\alpha = 0.05$, the P-value for the environmentalists' test must have been greater than 0.05. We can't be certain whether or not they would have made the same decision at $\alpha = 0.10$, since we only know that the P-value was greater than 0.05. It may have been greater than 0.10 as well, but we can't be sure. They would have made them same decision at $\alpha = 0.01$, since the P-value must also be greater than 0.01.

17. Groceries.

a) **Randomization condition:** We will assume that the Yahoo survey was conducted randomly.
10% condition: 2400 is less than 10% of all men.
Success/Failure condition: $n\hat{p} = 1224$ and $n\hat{q} = 1176$ are both greater than 10, so the sample is large enough.

Since the conditions are met, we can use a one-proportion *z*-interval to estimate the percentage of men who identify themselves as the primary grocery shopper in their household.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = \left(\frac{1224}{2400}\right) \pm 2.326\sqrt{\frac{\left(\frac{1224}{2400}\right)\left(\frac{1176}{2400}\right)}{2400}} = (48.6\%, 53.4\%)$$

We are 98% confident that between 48.6% and 53.4% of all men identify themselves as the primary grocery shopper in their household.

b) Since 45% is not in the interval, there is strong evidence that more than 45% of all men identify themselves as the primary grocery shopper in their household.

c) The significance level of this test is $\alpha = 0.01$. It's an upper tail test based on a 98% confidence interval.

18. Is the Euro fair?

a) **Independence assumption:** The Euro spins are independent. One spin is not going to affect the others. (With true independence, it doesn't make sense to try to check the randomization condition and the 10% condition. These verify our assumption of independence, and we don't need to do that!)

Success/Failure condition: $n\hat{p} = 140$ and $n\hat{q} = 110$ are both greater than 10, so the sample is large enough.

Since the conditions are met, we can use a one-proportion *z*-interval to estimate the proportion of heads in Euro spins.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = \left(\frac{140}{250}\right) \pm 1.960\sqrt{\frac{\left(\frac{140}{250}\right)\left(\frac{110}{250}\right)}{250}} = (0.498, 0.622)$$

We are 95% confident that the true proportion of heads when a Euro is spun is between 0.498 and 0.622.

Copyright © 2020 Pearson Education, Inc.

18. (continued)

b) Since 0.50 is within the interval, there is no evidence that the coin in unfair. 50% is a plausible value for the true proportion of heads. (That having been said, I'd want to spin this coin a few hundred more times. It's close!)

c) The significance level is $\alpha = 0.05$. It's a two-tail test based on a 95% confidence interval.

19. Approval 2016.

a) **Randomization condition:** The adults were randomly selected.
10% condition: 1500 adults represent less than 10% of all adults.
Success/Failure condition: $n\hat{p} = (1500)(0.57) = 855$ and $n\hat{q} = (1500)(0.43) = 645$ are both greater than 10, so the sample is large enough.

Since the conditions are met, we can use a one-proportion *z*-interval to estimate Barack Obama's approval rating.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = (0.57) \pm 1.960\sqrt{\frac{(0.57)(0.43)}{1500}} = (0.545, 0.595)$$

We are 95% confident that Barack Obama's approval rating is between 54.5% and 59.5%.

b) Since 52% is not within the interval, this is not a plausible value for Barack Obama's approval rating. There is evidence against the null hypothesis.

20. Hard times.

a) **Randomization condition:** The men were contacted through a random poll.
10% condition: 800 men represent less than 10% of all men.

Success/Failure condition: $n\hat{p} = (800)(0.09) = 72$ and $n\hat{q} = (800)(0.91) = 728$ are both greater than 10, so the sample is large enough.

Since the conditions are met, we can use a one-proportion *z*-interval to estimate the percentage of men who have taken a second job to help pay the bills.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = (0.09) \pm 1.96\sqrt{\frac{(0.09)(0.91)}{800}} = (7.0\%, 11.0\%)$$

We are 95% confident that between 7.0% and 11.0% of all men have taken a second job to help pay the bills.

b) Since 6% is not in the interval, the pundit's claim is not plausible.

21. Dogs.

a) We cannot construct a confidence interval for the rate of occurrence of early hip dysplasia among 6-month old puppies because only 5 of 42 puppies were found with early hip dysplasia. The Success/Failure condition is not satisfied.

b) We could not use a bootstrap confidence interval. The sample size is too small.

22. Fans.

a) We cannot construct a confidence interval for the percentage of home team fans entering the stadium, since only 8 people were not home fans. The Success/Failure condition is not satisfied.

b) We could not use a bootstrap confidence interval. The sample size is too small.

Copyright © 2020 Pearson Education, Inc.

23. Loans.

a) The bank has made a Type II error. The person was not a good credit risk, and the bank failed to notice this.

b) The bank has made a Type I error. The person was a good credit risk, and the bank was convinced that he/she was not.

c) By making it easier to get a loan, the bank has reduced the alpha level. It takes less evidence to grant the person the loan.

d) The risk of Type I error is decreased and the risk of Type II error has increased.

24. Spam.

a) Type II. The filter decided that the message was safe, when in fact it was spam.

b) Type I. The filter decided that the message was spam, when in fact it was not.

c) This is analogous to lowering alpha. It takes more evidence to classify a message as spam.

d) The risk of Type I error is decreased and the risk of Type II error has increased.

25. Second loan.

a) Power is the probability that the bank denies a loan that could not have been repaid.

b) To increase power, the bank could raise the cutoff score.

c) If the bank raised the cutoff score, a larger number of trustworthy people would be denied credit, and the bank would lose the opportunity to collect the interest on these loans.

26. More spam.

a) The power of the test is the ability of the filter to detect spam.

b) To increase the filter's power, lower the cutoff score.

c) If the cutoff score is lowered, a larger number of real messages would end up in the junk mailbox.

27. Homeowners 2015.

a) The null hypothesis is that the level of home ownership does not rise. The alternative hypothesis is that it rises.

b) In this context, a Type I error is when the city concludes that home ownership is on the rise, but in fact, the tax breaks don't help.

c) In this context, a Type II error is when the city abandons the tax breaks, thinking they don't help, when in fact they were helping.

d) A Type I error causes the city to forego tax revenue, while a Type II error withdraws help from those who might have otherwise been able to buy a house.

e) The power of the test is the city's ability to detect an actual increase in home ownership.

28. Alzheimer's.

a) The null hypothesis is that a person is healthy. The alternative is that they have Alzheimer's disease. There is no parameter of interest here.

b) A Type I error is a false positive. It has been decided that the person has Alzheimer's disease when they don't.

c) A Type II error is a false negative. It has been decided that the person is healthy, when they actually have Alzheimer's disease.

d) A Type I error would require more testing, resulting in time and money lost. A Type II error would mean that the person did not receive the treatment they needed. A Type II error is much worse.

28. (continued)

e) The power of this test is the ability of the test to detect patients with Alzheimer's disease. In this case, the power can be computed as $1 - P(\text{Type II error}) = 1 - 0.08 = 0.92$.

29. Testing cars.

H_0: The shop is meeting the emissions standards.
H_A: The shop is not meeting the emissions standards.

a) Type I error is when the regulators decide that the shop is not meeting standards when they actually are meeting the standards.

b) Type II error is when the regulators certify the shop when they are not meeting the standards.

c) Type I would be more serious to the shop owners. They would lose their certification, even though they are meeting the standards.

d) Type II would be more serious to environmentalists. Shops are allowed to operate, even though they are allowing polluting cars to operate.

30. Quality control.

H_0: The assembly process is working fine.
H_A: The assembly process is producing defective items.

a) Type I error is when the production managers decide that there has been an increase in the number of defective items and stop the assembly line, when the assembly process is working fine.

b) Type II error is when the production managers decide that the assembly process is working fine, but defective items are being produced.

c) The factory owner would probably consider Type II error to be more serious, depending of the costs of shutting the line down. Generally, because of warranty costs and lost customer loyalty, defects that are caught in the factory are much cheaper to fix than defects found after items are sold.

d) Customers would consider Type II error to be more serious, since customers don't want to buy defective items.

31. Cars again.

a) The power of the test is the probability of detecting that the shop is not meeting standards when they are not.

b) The power of the test will be greater when 40 cars are tested. A larger sample size increases the power of the test.

c) The power of the test will be greater when the level of significance is 10%. There is a greater chance that the null hypothesis will be rejected.

d) The power of the test will be greater when the shop is out of compliance "a lot". Larger problems are easier to detect.

32. Production.

a) The power of the test is the probability that the assembly process is stopped when defective items are being produced.

b) An advantage of testing more items is an increase in the power of the test to detect a problem. The disadvantages of testing more items are the additional cost and time spent testing.

c) An advantage of lowering the alpha level is that the probability of stopping the assembly process when everything is working fine (committing a Type I error) is decreased. A disadvantage is that the power of the test to detect defective items is also decreased.

d) The power of the test will increase as a day passes. Bigger problems are easier to detect.

Copyright © 2020 Pearson Education, Inc.

33. Equal opportunity?

H_0: The company is not discriminating against minorities.
H_A: The company is discriminating against minorities.

a) This is a one-tailed test. They wouldn't sue if "too many" minorities were hired.

b) Type I error would be deciding that the company is discriminating against minorities when they are not discriminating.

c) Type II error would be deciding that the company is not discriminating against minorities when they actually are discriminating.

d) The power of the test is the probability that discrimination is detected when it is actually occurring.

e) The power of the test will increase when the level of significance is increased from 0.01 to 0.05.

f) The power of the test is lower when the lawsuit is based on 37 employees instead of 87. Lower sample size leads to less power.

34. Stop signs.

H_0: The new signs provide the same visibility than the old signs.
H_A: The new signs provide greater visibility than the old signs.

a) The test is one-tailed, because we are only interested in whether or not the signs are more visible. If the new design is less visible, we don't care how much less visible it is.

b) Type I error happens when the engineers decide that the new signs are more visible when they are not more visible.

c) Type II error happens when the engineers decide that the new signs are not more visible when they actually are more visible.

d) The power of the test is the probability that the engineers detect a sign that is truly more visible.

e) When the level of significance is dropped from 5% to 1%, power decreases. The null hypothesis is harder to reject, since more evidence is required.

f) If a sample of size 20 is used instead of 50, power will decrease. A smaller sample size has more variability, lowering the ability of the test to detect falsehoods.

35. Software for learning.

a) The test is one-tailed. We are testing to see if an increase in average score is associated with the software.

b) H_0: The average score does not change following the use of software. ($\mu = 105$)
H_A: The average score increases following the use of the software. ($\mu > 105$)

c) The professor makes a Type I error if he buys the software when the average score has not actually increased.

d) The professor makes a Type II error if he doesn't buy the software when the average has actually increased.

e) The power of the test is the probability of buying the software when the average score has actually increased.

36. Ads.

a) H_0: The percentage of residents that have heard the ad and recognize the product is 20%. ($p = 0.20$)
H_A: The percentage of residents that have heard the ad and recognize the product is greater than 20%. ($p > 0.20$)

b) The company wants more evidence that the ad is effective before deciding it really is. By lowering the level of significance from 10% to 5%, the probability of Type I error is decreased. The company is less likely to think that the ad is effective when it actually is not effective.

Copyright © 2020 Pearson Education, Inc.

36. (continued)

c) The power of the test is the probability of correctly deciding more than 20% have heard the ad and recognize the product when it's true.

d) The power of the test will be higher for a level of significance of 10%. There is a greater chance of rejecting the null hypothesis.

e) Increasing the sample size to 600 will lower the risk of Type II error. A larger sample size decreases variability, which helps us notice what is really going on. The company will be more likely to notice when the ad really works.

37. Software, part II.

a) H_0: The average score does not change following the use of software. (μ = 105)
H_A: The average score increases following the use of the software. (μ > 105)

Randomization condition: This year's class of 203 students is probably representative of all stats students. **Nearly Normal condition:** We don't have the scores from the 203 individuals, so we can't check a plot of the data. However, with a sample this large, the Central Limit Theorem allows us to model the sampling distribution of the means with a *t*-distribution.

The mean score was 108 points, with a standard deviation of 8.7 points. Since the conditions for inference are satisfied, we can model the sampling distribution of the mean score with a Student's *t*-model, with 203 – 1 = 202 degrees of freedom, $t_{202}\left(108, \frac{8.7}{\sqrt{203}}\right)$. We will perform a one-sample *t*-test.

$$t = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{108 - 105}{\frac{8.7}{\sqrt{203}}} = 4.91$$

The value of *t* is approximately 4.91, which results in a P-value of less than 0.0001, so we reject the null hypothesis. There is strong evidence that the mean score has increased since use of the software program was implemented. As long as the professor feels confident that this class of stats students is representative of all potential students, then he should buy the program.

If you used a 95% confidence interval to assess the effectiveness of the program:

$$\bar{y} \pm t_{n-1}^{*}\left(\frac{s}{\sqrt{n}}\right) = 108 \pm t_{202}^{*}\left(\frac{8.7}{\sqrt{203}}\right) \approx (106.8, 109.2)$$

We are 95% confident that the mean score is between 106.8 and 109.2. Since 105 is above the interval, this provides evidence that the mean score has increased following the implementation of the software program.

b) The mean score on the exam only increased by 1 to 4 points. This small difference might not be enough to be worth the cost of the program.

Copyright © 2020 Pearson Education, Inc.

38. Testing the ads.

a) H_0: The percentage of residents that remember the ad is 20%. ($p = 0.20$)
H_A: The percentage of residents that remember is greater than 20%. ($p > 0.20$)

Independence assumption: It is reasonable to think that randomly selected residents would remember the ad independently of one another.
Randomization condition: The sample was 600 randomly selected residents.
10% condition: The sample of 600 is less than 10% of the population of the city.
Success/Failure condition: $np = (600)(0.20) = 120$ and $nq = (600)(0.80) = 480$ are both greater than 10, so the sample is large enough.

The conditions have been satisfied, so a Normal model can be used to model the sampling distribution of the proportion, with $\mu_{\hat{p}} = p = 0.20$ and $SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.20)(0.80)}{600}} \approx 0.0163.$

We can perform a one-proportion z-test. The observed proportion of residents who remembered the ad is $\hat{p} = \frac{133}{600} \approx 0.222$.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{pq}{n}}} = \frac{0.222 - 0.20}{\sqrt{\frac{(0.20)(0.80)}{600}}} \approx 1.33$$

Since the P-value = 0.0923 is somewhat high, we fail to reject the null hypothesis. There is little evidence that more than 20% of people remember the ad. The company should not renew the contract.

b) There is a 9.23% chance of having 133 or fewer of 600 people in a random sample remember the ad, if in fact only 20% of people in the population do.

39. TV safety.

a) This is an upper-tail test. We hope to show that the TV stand will hold 500 or more pounds easily.

b) The inspectors will commit a Type I error if they decide the TV stands are safe when they are not.

c) The inspectors will commit a Type II error if they decide the TVs are unsafe when they are actually safe.

40. Catheters.

a) This is a two-sided test. If catheters are too big, they won't fit through the vein. If the catheters are too small, they might not have enough structural integrity to work well.

b) The quality control people will commit a Type I error if they decide that the catheters are not the correct size when they actually have the correct diameter. The manufacturing process is stopped needlessly.

c) The quality control people will commit a Type II error if they decide that the catheters have the correct diameter when they actually are too small or too large. Catheters that do not meet specifications are produced and sold, possibly injuring patients, or simply not working properly.

41. TV safety, revisited.

a) To decrease the likelihood of producing an unsafe TV stand, they should decrease α. This lower the chance of making a Type I error.

b) The power of the test is the ability of the inspectors to determine that the TV stand is safe when it is actually capable of holding more than 500 pounds.

c) The company could increase the power of the test by lowering the standard deviation by testing more stands. This could prove costly and would require more time to test. They could also increase α, but then they will commit Type I errors with greater frequency, approving stands that cannot hold 500 pounds of more. Finally, they could require that TV stands have a higher weight capacity than 500 pounds as the standard. Again, that might prove costly, since they would be rejecting many more stands that were safe.

Copyright © 2020 Pearson Education, Inc.

42. Catheters, again.

a) If the significance level were changed to $\alpha = 0.01,$ this would increase the probability of Type II error. Requiring a higher standard of proof would mean that more catheters would be rejected, even when they met the diameter specification.

b) The power of the test is the probability of detecting a catheter than does not meet the 2.00 mm specification.

c) As the diameters got farther and farther away from 2.00 mm, the power would increase. It would become easier to detect that the catheters did not meet the diameter specification when the were much too big or small.

d) To increase the power of the test, they could sample more catheters or increase the significance level.

43. Two coins.

a) The alternative hypothesis is that your coin produces 30% heads.

b) Reject the null hypothesis if the coin comes up tails. Otherwise, fail to reject.

c) There is a 10% chance that the coin comes up tails if the null hypothesis is true, so alpha is 10%.

d) Power is our ability to detect the 30% coin. That coin will come up tails 70% of the time. That's the power of our test.

e) To increase the power and lower the probability of Type I error at the same time, simply flip the coin more times.

44. Faulty or not?

a) The null hypothesis is that the drive is good. The alternative hypothesis is that the drive is bad.

b) Reject the null hypothesis if the computer fails the test. Otherwise, fail to reject.

c) There is a 4% chance that the computer fails the test, even if the drive is good, so alpha is 4%.

d) Power is the ability to detect faulty drives. Faulty drives fail the test 65% of the time. That's the power of our test.

45. Hoops.

H_0: The player's foul-shot percentage is only 60%. ($p = 0.60$)
H_A: The player's foul-shot percentage is better than 60%. ($p > 0.60$)

a) There are only two possible outcomes, make the shot and miss the shot. The probability of making any shot is constant at $p = 0.60$. Assume that the shots are independent of each other.

There are 10 different ways to make 9 out of 10. To look at it another way, there are 10 ways to choose the shot to miss.

$$\begin{aligned} P(\text{makes at least 9 out of 10}) &= P(\text{makes 9}) + P(\text{makes 10}) \\ &= 10(0.60)^9(0.40)^1 + (0.60)^{10} \\ &\approx 0.0464 \end{aligned}$$

b) The coach made a Type I error.

Copyright © 2020 Pearson Education, Inc.

45. (continued)

c) The power of the test can be calculated for specific values of the new probability of success. Each true value of p has a power calculation associated with it. In this case, we are finding the power of the test to detect an 80% foul-shooter.

There are 10 different ways to make 9 out of 10. To look at it another way, there are 10 ways to choose the shot to miss.

$$\begin{aligned} P(\text{makes at least 9 out of 10}) &= P(\text{makes 9}) + P(\text{makes 10}) \\ &= 10(0.80)^9(0.20)^1 + (0.80)^{10} \approx 0.376 \end{aligned}$$

The power of the test to detect an increase in foul-shot percentage from 60% to 80% is about 37.6%.

d) The power of the test to detect improvement in foul-shooting can be increased by increasing the number of shots, or by keeping the number of shots at 10 but increasing the level of significance by declaring that 8, 9, or 10 shots made will convince the coach that the player has improved. In other words, the coach can increase the power of the test by lowering the standard of proof.

46. Pottery.

H_0: The new clay is no better than the old, and breaks 40% of the time. ($p = 0.40$)
H_A: The new clay breaks less than 40% of the time. ($p < 0.40$)

a) There are only two possible outcomes, broken and unbroken. The probability of breaking is constant at $p = 0.40$. It is reasonable to think that the pieces break independently of each other.

There are 10 ways to have exactly one broken piece.

$$\begin{aligned} P(\text{at most one breaks}) &= P(\text{no broken pieces}) + P(\text{1 broken piece}) \\ &= (0.60)^{10} + 10(0.40)^1(0.60)^9 \approx 0.0464 \end{aligned}$$

b) The artist made a Type I error.

c) The probability Type II error can be calculated for specific values of the new probability of success. Each true value of p has a Type II error calculation associated with it. In this case, we are finding the probability of Type II error if the pieces break only 20% of the time instead of 40% of the time. She won't notice that the clay is better if 2 or more pieces break.

This gets a little tricky. The best way to calculate this is to use the complement. Having at least 2 break is every outcome except having no breaks or 1 break.

$$\begin{aligned} P(\text{at least 2 break}) &= 1 - P(\text{0 or 1 break}) \\ &= 1 - (P(0) + P(1)) \\ &= 1 - \left((0.80)^{10} + 10(0.20)^1(0.80)^9\right) \\ &\approx 0.6242 \end{aligned}$$

The probability that she makes a Type II error (not noticing that the clay is better) is approximately 0.6242.

d) The power of the test to detect improvement in the clay can be increased by increasing the number of pieces fired, or by keeping the number of pieces at 10 but increasing the level of significance by declaring that 0, 1, or 2 broken pieces will convince the artist that the player has improved. In other words, the artist can improve the power by lowering her standard of proof.

Copyright © 2020 Pearson Education, Inc.

47. Chips Ahoy! bootstrapped.

The sample mean of the original data set was 1238.19 chips. If the null hypothesis were true, and the true mean were 1000 chips, we didn't see means that extreme or more extreme in any of our 10,000 bootstrap samples. That's less than 1/10,000 times, or a P-value < 0.00001.

48. Farmed salmon bootstrapped.

a) The mean mirex level in the sample was 0.09134 ppm. The histogram of shifted bootstrapped mean mirex levels shows simulated means as extreme or more extreme than 0.09134 ppm in 0.38% + 0.18% = 0.56% of bootstrapped samples. That corresponds to a P-value of 0.0056.

b) If the EPA is really only concerned about mirex contamination above 0.08 ppm, the percentage of shifted bootstrapped means above the sample mean of 0.09134 is 0.38%, a P-value of 0.0038.

Copyright © 2020 Pearson Education, Inc.