Chapter 21 – More About Intervals and Tests

1. One sided or two?

a) Two sided. Let p be the percentage of students who prefer Diet Coke.

H_0 : 50% of students prefer Diet Coke. (p = 0.50)
H_A : The percentage of students who prefer Diet Coke is not 50%. ($p \neq 0.50$)

b) One sided. Let p be the percentage of teenagers who prefer the new formulation.

H_0 : 50% of students prefer the new formulation. (p = 0.50)
H_A : More than 50%of students prefer the new formulation. ($p > 0.50$)

c) One sided. Let p be the percentage of people who plan to vote for the override.

H_0 : 2/3 of the residents intend to vote for the override. ($p = 2/3$)
H_A : More than 2/3 of the residents intend to vote for the override. ($p > 2/3$)

d) Two sided. Let p be the percentage of days the market goes up.

H_0 : The market goes up on 50% of days. (p = 0.50)
H_A : The percentage of days the market goes up is not 50%. ($p \neq 0.50$)

2. Which alternative?

a) Two sided. Let p be the percentage of students who prefer plastic.

H_0 : 50% of students prefer plastic.. (p = 0.50)
H_A : The percentage of students who prefer plastic is not 50%. ($p \neq 0.50$)

b) Two sided. Let p be the percentage of juniors planning to study abroad.

H_0 : 10% of juniors plan to study abroad. (p = 0.10)
H_A : The percentage of juniors plan to study abroad is not 10%. ($p \neq 0.10$)

c) One sided. Let p be the percentage of people who experience relief.

H_0 : 22% of people experience headache relief with the drug. (p = 0.22)
H_A : More than 22% of people experience headache relief with the drug. ($p > 0.22$)

d) One sided. Let p be the percentage of hard drives that pass all performance tests.

H_0 : 60% of hard drives pass all performance tests. (p = 0.60)
H_A : The percentage of drives that pass all performance tests is greater than 60%. ($p > 0.60$)

3. 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.

4. Another P-value.

If the rate of seat belt usage after the campaign is the same as the rate of seat belt usage before the campaign, there is a 17% chance of observing a rate of seat belt usage after the campaign this large or larger in a sample of the same size by natural sampling variation alone.

5. 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 $\alpha = 0.01$, 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.

6. 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.

7. Significant?

a) If 90% of children have really been vaccinated, there is only a 1.1% chance of observing 89.4% 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 90%, but a 95% confidence interval would show that the true proportion is between 88.9% and 89.9%. Most likely a decrease from 90% to 89.9% would not be considered important. The 90% figure was probably an approximate figure anyway.

8. 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.

9. Success.

a) **Independence assumption:** One man's response is not likely to have any effect on another man's response.
Randomization condition: The men were contacted through a random telephone poll.
10% condition: 1302 men represent less than 10% of all men.
Success/Failure condition: $n\hat{p} = 39$ and $n\hat{q} = 1263$ 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 for whom work is their most important measure of success.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = \left(\frac{39}{1302}\right) \pm 2.326 \sqrt{\frac{\left(\frac{39}{1302}\right)\left(\frac{1263}{1302}\right)}{1302}} = (1.9\%, 4.1\%)$$

We are 98% confident that between 1.9% and 4.1% of all men believe that work is their most important measure of success.

b) Since 5% is not in the interval, there is strong evidence that fewer than 5% of all men use work as their primary measure of success.

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

10. Is the Euro fair?

a) **Independence assumption:** The Euro spins are independent. One spin is not going to effect the others. (With true independence, it doesn't make sense to try to check the randomization condition or 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.

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.

11. Approval 2007.

a) **Independence assumption:** One response is not likely to effect on another response.
Randomization condition: The adults were randomly selected.
10% condition: 1125 adults represent less than 10% of all adults.
Success/Failure condition: $n\hat{p}$ = (1125)(0.30) = 338 and $n\hat{q}$ = (1125)(0.70) = 787 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 George W. Bush's approval rating.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = (0.30) \pm 1.960 \sqrt{\frac{(0.30)(0.70)}{1125}} = (0.274, 0.327)$$

We are 95% confident that George W. Bush's approval rating is between 27.4% and 32.7%.

b) Since 27% is not within the interval, this is not a plausible value for George W. Bush's approval rating. There is evidence against the null hypothesis.

12. Superdads.

a) **Independence assumption:** One man's response is not likely to have any effect on another man's response.
Randomization condition: The men were contacted through a random telephone poll.
10% condition: 712 men represent less than 10% of all men.
Success/Failure condition: $n\hat{p}$ = (712)(0.22) = 157 and $n\hat{q}$ = (712)(0.78) = 555 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 stay-at-home dads.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = (0.22) \pm 1.96\sqrt{\frac{(0.22)(0.78)}{712}} = (19.0\%, 25.1\%)$$

We are 95% confident that between 19.0% and 25.1% of all men identify themselves as stay-at-home dads.

b) Since the confidence interval extends well below 25%we can't be confident that over 25% of men are stay-at-home dads. The company should not buy the ads.

c) Since, 25% is within the confidence interval, Spike could claim that it's possible that the true proportion of stay-at-home dads is 25%, but we can never prove that the null hypothesis is true.

13. 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) **Independence assumption:** If hip dysplasia is hereditary, and puppies brought to the vaccination clinic are from the same litter, independence might be an issue.
Randomization condition: The veterinarian considers the 42 puppies to be a random sample of all puppies.
10% condition: 42 puppies represent less than 10% of all puppies.
Success/Failure condition: As previously mentioned, this condition is not met, since there aren't at least 10 puppies with hip dysplasia in the sample.

Since the other conditions are met, we can construct a one-proportion plus-four z-interval to estimate the percentage of puppies with early hip dysplasia.

$$\tilde{p} = \frac{y+2}{n+4} = \frac{7}{46} = 0.152$$

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}} = (0.152) \pm 1.96\sqrt{\frac{(0.152)(0.848)}{46}} = (4.8\%, 25.6\%)$$

We are 95% confident that between 4.8% and 25.6% of puppies have early hip dysplasia.

14. Fans.

a) We cannot construct a confidence interval for the percentage of home team fans entering the stadium, since only 9 people were not home fans. The Success/Failure condition is not satisfied.

b) **Independence assumption:** We have no reason to believe that one person being a home fan will affect another person.
Randomization condition: The people standing in line were randomly selected.
10% condition: 81 people are likely to be less than 10% of the fans in the stadium.
Success/Failure condition: As previously mentioned, this condition is not met, since there aren't at least 10 non-home fans in the sample.

Since the other conditions are met, we can construct a one-proportion plus-four z-interval to estimate the percentage of home fans at the game.

$$\tilde{p} = \frac{y+2}{n+4} = \frac{75}{85} = 0.882$$

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}} = (0.882) \pm 1.96\sqrt{\frac{(0.882)(0.118)}{85}} = (81.4\%, 95.1\%)$$

We are 95% confident that between 81.4% and 95.1% of all people are home fans.

15. 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.

16. 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.

17. 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.

18. 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.

19. Homeowners 2005.

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.

20. 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.

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 $1 - P(\text{Type II error}) = 1 - 0.08 = 0.92$.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

27. Dropouts.

a) The test is one-tailed. We are testing to see if a decrease in the dropout rate is associated with the software.

b) H_0 : The dropout rate does not change following the use of the software. ($p = 0.13$)
H_A : The dropout rate decreases following the use of the software. ($p < 0.13$)

c) The professor makes a Type I error if he buys the software when the dropout rate has not actually decreased.

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

e) The power of the test is the probability of buying the software when the dropout rate has actually decreased.

28. 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.

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.

29. Dropouts, part II.

a) H_0 : The dropout rate does not change following the use of the software. (p = 0.13)
H_A : The dropout rate decreases following the use of the software. (p < 0.13)

Independence assumption: One student's decision about dropping out should not influence another's decision.
Randomization condition: This year's class of 203 students is probably representative of all stats students.
10% condition: A sample of 203 students is less than 10% of all students.
Success/Failure condition: np = (203)(0.13) = 26.39 and nq = (203)(0.87) = 176.61 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 $\mu_{\hat{p}} = p = 0.13$ and $\sigma(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.13)(0.87)}{203}} \approx 0.0236$.

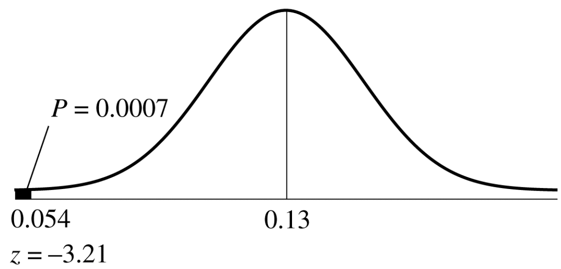
We can perform a one-proportion z-test. The observed proportion of dropouts is $\hat{p} = \frac{11}{203} \approx 0.054$.

Since the P-value = 0.0007 is very low, we reject the null hypothesis. There is strong evidence that the dropout rate has dropped 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.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{pq}{n}}}$$

$$z = \frac{0.054 - 0.13}{\sqrt{\frac{(0.13)(0.87)}{203}}}$$

$$z \approx -3.21$$



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

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = \left(\frac{11}{203}\right) \pm 1.960 \sqrt{\frac{\left(\frac{11}{203}\right)\left(\frac{192}{203}\right)}{203}} = (2.3\%, 8.5\%)$$

We are 95% confident that the dropout rate is between 2.3% and 8.5%. Since 15% is not contained in the interval, this provides evidence that the dropout rate has changed following the implementation of the software program.

b) The chance of observing 11 or fewer dropouts in a class of 203 is only 0.07% if the dropout rate in the population is really 13%.

30. 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 the ad 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 consisted of 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 $\sigma(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.20)(0.80)}{600}} \approx 0.0163$.

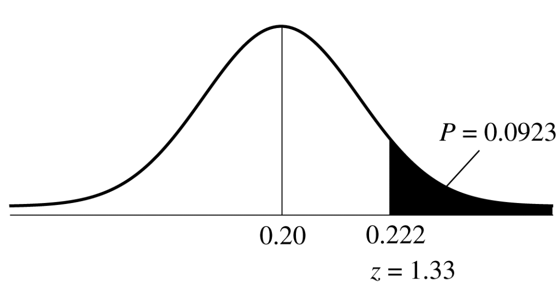
We can perform a one-proportion z-test. The observed proportion of residents who remembered the ad is $\hat{p} = \frac{133}{600} \approx 0.222$.

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.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{pq}{n}}}$$

$$z = \frac{0.222 - 0.20}{\sqrt{\frac{(0.20)(0.80)}{600}}}$$

$$z \approx 1.33$$



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.

31. 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.

32. 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 out test.

33. 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) The player's shots can be considered Bernoulli trials. 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. Use *Binom*(10, 0.60).

Let X = the number of shots made out of n = 10.

$$\begin{aligned}
P(\text{makes at least 9 out of 10}) &= P(X \geq 9) \\
&= P(X=9) + P(X=10) \\
&= {}_{10}C_9(0.60)^9(0.40)^1 + {}_{10}C_{10}(0.60)^{10}(0.40)^0 \\
&\approx 0.0464
\end{aligned}$$

b) The coach made a Type I error.

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. Use *Binom*(10, 0.80).

Let X = the number of shots made out of n = 10.

$$\begin{aligned} P(\text{makes at least 9 out of 10}) &= P(X \geq 9) \\ &= P(X=9) + P(X=10) \\ &= {}_{10}C_9(0.80)^9(0.20)^1 + {}_{10}C_{10}(0.80)^{10}(0.20)^0 \approx 0.376 \end{aligned}$$

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.

34. 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 is better than the old, and breaks less than 40% of the time. (p < 0.40)

a) The fired pieces can be considered Bernoulli trials. 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. Use *Binom*(10, 0.40).

Let X = the number of broken pieces out of n = 10.

$$\begin{aligned} P(\text{at most one breaks}) &= P(X \leq 1) \\ &= P(X=0) + P(X=1) \\ &= {}_{10}C_0(0.40)^0(0.60)^{10} + {}_{10}C_1(0.40)^1(0.60)^9 \approx 0.0464 \end{aligned}$$

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 is 2 or more pieces break. Use *Binom*(10, 0.80).

Let X = the number of broken pieces out of n = 10.

$$\begin{aligned} P(\text{at least 2 break}) &= P(X \geq 2) \\ &= P(X=2) + \cdots + P(X=10) \\ &= {}_{10}C_2(0.20)^2(0.80)^8 + \cdots + {}_{10}C_{10}(0.20)^{10}(0.80)^0 \\ &\approx 0.6242 \end{aligned}$$

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.