Unit 6: Hypothesis and Confidence for Proportions

Introduction: Now that we have had some practice with sampling distributions and demonstrated the importance of replication in sampling, we can apply this to making inferences about data. Two ways to infer statistically are called Confidence Intervals and Hypothesis Testing. While both are carfully constructed methods of making inferences, they are based on statistics we already know: means, proportions, standard deviations, and z-scores. We need to be careful not to call them conclusions because the data we collect is not to be taken as proof of an argument one way or the other and because such wording will matter on the AP exam. We will not be proving anything but rather stating that we did or did not find 'significant evidence'. What determines significance will also be discussed here.

6-1 Confidence Intervals for Population Proportions

Goals: 1. Review biased and unbiased statistics
2. Construct confidence intervals for population proportions
3. Correctly interpret confidence intervals and confidence levels
4. Relate margin of error and sample size

First, a quick review of ***point estimates*** and ***biased/unbiased statistics***:

- A **point estimate** of a population parameter is a *single number* that is based on sample data (we've called these **statistics** all year).

Example 1 Determine the point estimates (in symbols) for the following parameters:

(a) The point estimate for the population mean, μ

(b) The point estimate of the population standard deviation, σ

(c) The point estimate of the population variance, σ^2

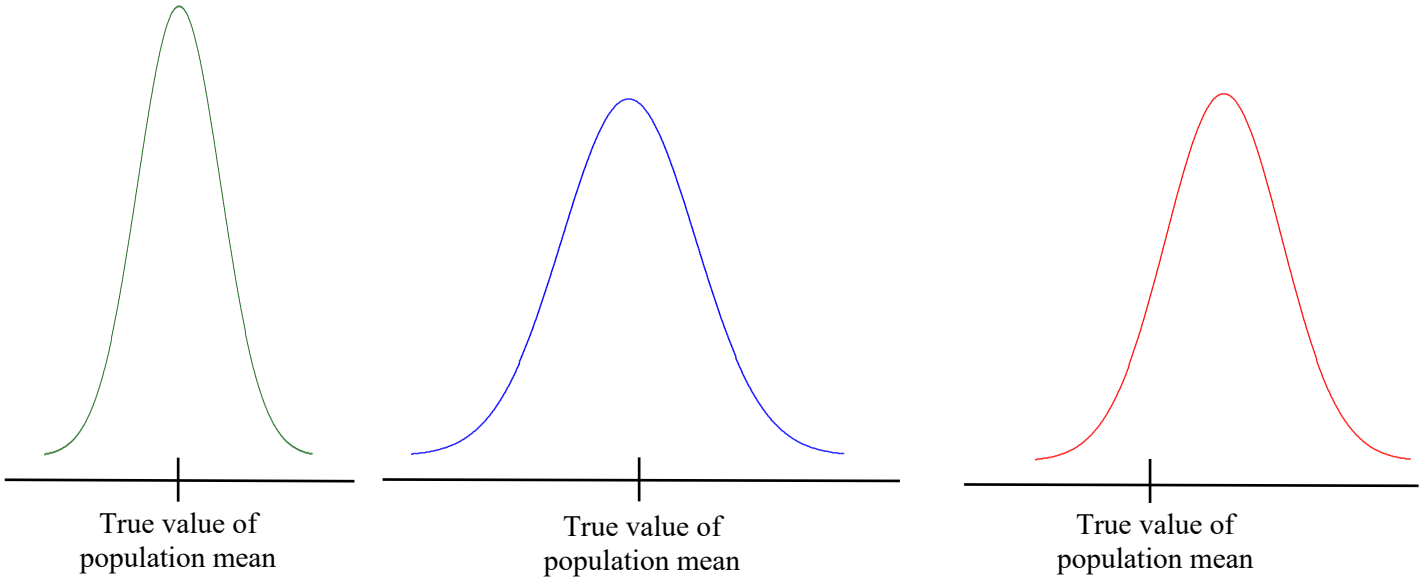
(d) The point estimate of the population proportion, p

Answers: (a) $\bar{x}$ (b) s (c) s^2 (d) $\hat{p}$

- A statistic whose mean value is equal to the value of the population characteristic being estimated is said to be an **unbiased statistic**. A statistic that is not unbiased is said to be **biased**.

- Given a choice between several unbiased statistics that could be used for estimating a population characteristic, the best statistic to use is the one with the smallest standard deviation.

Which of the following are unbiased statistics? Which of the unbiased statistics is the best fit?

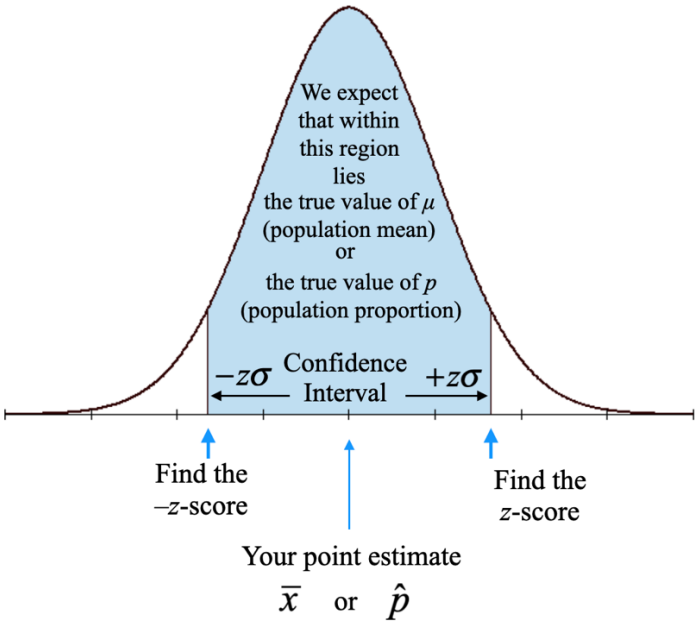


Example 2: We are interested in the proportion of red M & M's in a giant candy bowl. We don't have time to count every single bead. We estimate this proportion by taking repeated samples of appropriate size to get a point estimate of the true proportion.

What is a reasonable range of values for p?

This estimated range of values is what we call a **Confidence Interval (CI)**

Remember that unlike Unit 5, in real life we usually do not have the population parameter such as p or μ so we need to sample in order to get our best estimate.

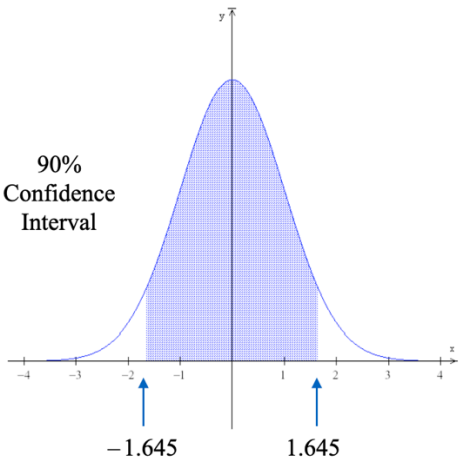


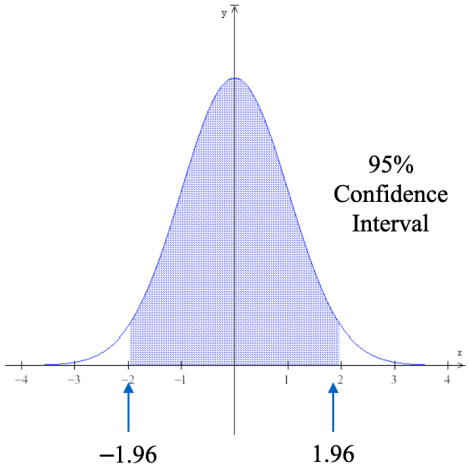
In the example shown to the left, we have plotted a normal distribution with our point estimate as the mean. This is a value that we will have obtained from our sample.

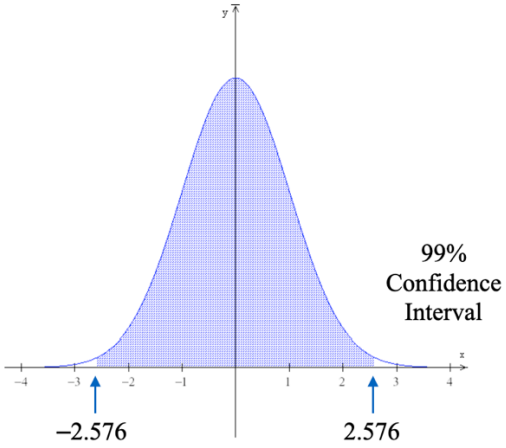
Using z-scores we get from our calculator or a table, we can choose the appropriate size of our confidence interval as a percentage.

If we want an x% confidence interval, we can then say that we are x% confident that the true mean or proportion lies within that shaded interval

Shown below are graphs of three common confidence interval percentages. Note that the size of each interval is determined by the z-score for that percentage, an easy value to find.







Here we also need to note the difference between a confidence interval and a confidence level:

- A **confidence interval (CI)** for a population characteristic is an *interval estimate* of plausible values for the characteristic.

- The **confidence level** associated with a confidence interval estimate is the success rate of the *method* used to construct the interval.

Since all of this is based on finding sample proportions and means as we did in the previous chapter, we still have to check our assumptions. Let's review them here and discuss how to apply them to confidence intervals.

Large – Sample Confidence Level for p:

Assumptions:

1. p is the sample proportion from a **random sample (or the sample represents the population)**

2. the sample size ***n* is large** ($n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$)
3. **the sample size is small relative to the population size** if the sample is selected without

replacement (*i.e.*, n is at most 10% of the population size). This allows us to sample without replacement.

Now let's look at **constructing a confidence interval:**

One way to look at this and other confidence interval formulas is to go back to the z-score equation:

$$z-\text{score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

Here, instead of means, we will rewrite it in terms of proportions:

$$z = \frac{\hat{p} - p}{\sigma_p}$$

in which p is our true (but unknown) population and $\hat{p}$ is our sample proportion that we are hoping will give us our best approximation of p. A little algebra gives us $p = \hat{p} - z\sigma_p$ but since we do not know p then we also don't know σ_p and since we are trying to find an interval in which we are confident the true value of p is, we have to consider that p could be greater than or less than our sample $\hat{p}$. This means that our z value could be positive or negative and we will have this:

Confidence Interval for a Population Proportion

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The desired confidence level determines which z critical value is used. The three most commonly used confidence levels, 90%, 95%, and 99%, use z critical values of 1.645, 1.96, and 2.58, respectively. By the way, you can find these numbers on your calculator using *invNorm* since they are just z values from a standard normal distribution.

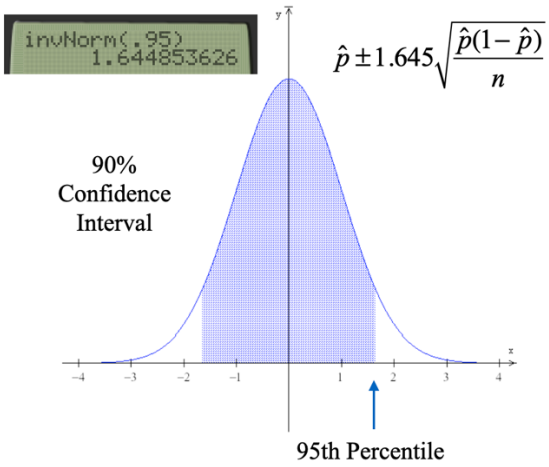
Confidence Level (CI)	z Critical Value (C.V.)
90%	1.645
95%	1.96
99%	2.58

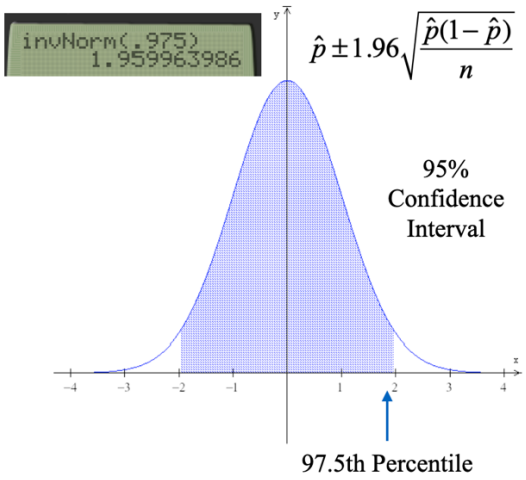
You will use this general formula for ALL the confidence intervals that you construct (there are many more than proportion CIs).

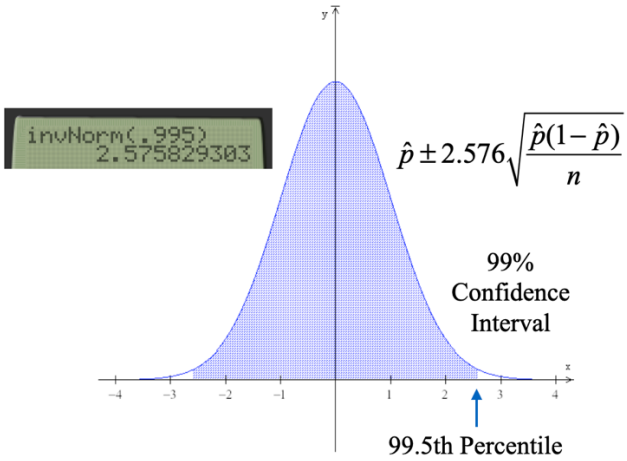
statistic ± (critical value) * (standard deviation)

Since we will be dealing with proportions in this unit our formula will be based on this interval:

$$p = \hat{p} \pm z\sigma$$







Notice that we replace p with $\hat{p}$ in these formulas. In most cases we don't know the true mean/proportion much less the standard deviation. This is why sampling distributions and thus confidence intervals are so important. We are making our best, educated effort to find the true mean/proportion and most of the time the only real data we have are our sample data

Example 3 Affirmative action in university admission is a controversial topic. To assess public opinion on this issue, investigators conducted a survey of 1013 randomly selected U.S. adults. It was reported that 537 of the 1013 people surveyed believed that affirmative action programs should be continued. Calculate the point estimate of p, the true proportion of US adults who believe that affirmative action programs should be continued. Construct a 95% confidence interval around our point estimate.

$$\hat{p} = \frac{537}{1013} = .530 \rightarrow \sigma_{\hat{p}} = \sqrt{\frac{0.530(1-0.530)}{1013}} = 0.0157$$

CI: $0.530 \pm 1.96\sqrt{\frac{0.530(1-0.530)}{1013}} = (0.499, 0.561)$

Interpretation for Large Sample Proportion Confidence Intervals
"We are __% confident that *p*, the true proportion of _____, is between ___% and ___%."

Interpretation for the Confidence *Level* of a Large Sample Proportion Confidence Intervals
"We used a method to construct this estimate that in the long run will successfully capture the true value of *p* ___% of the time."

General Work Flow → Assumptions → Construction of Interval → Interpretation(s)

Calculator: Stat → Tests → A. 1-PropZInt → Enter Success → Enter *n* →

Enter Confidence Level → Highlight Calculate → Enter

Example 4 An AP article on potential violent behavior reported the results of a survey of 750 workers who were employed full time. Of those surveyed, 125 indicated that they were so angered by a co-worker during the past year that he or she felt like hitting the person (but didn't). Assuming that it is reasonable to regard this sample of 750 as a random sample from the population of full-time workers, we can use this information to construct an estimate of *p*, the true proportion of full-time workers so angered in the last year that they wanted to hit a colleague. Construct a 90% C.I. for *p*. Interpret this interval and level.

$$\hat{p} = \frac{125}{750} = 0.167 = \frac{1}{6} \rightarrow \sigma_{\hat{p}} = \sqrt{\frac{0.167(1-0.167)}{750}} = 0.0136$$

CI: $0.167 \pm 1.645\sqrt{\frac{0.167(1-0.167)}{750}} = (0.144, 0.189)$

We are 90% confident that *p*, the true proportion of colleagues angry enough to hit, is between 14.4% and 18.9%

We used a method to construct this estimate that in the long run will successfully capture the true value of *p* 90% of the time

You can check your interval on your calculator, but keep in mind, to get **full credit** on your AP exam FRQ, you **MUST** show your work!

Example 5 There are 533 successes in a random sample of size 1000. How would you calculate a 90% confidence interval for this sample?

(a) $0.533 \pm 1.960\sqrt{\frac{(0.533)(0.467)}{1000}}$

(b) $0.533 \pm 1.645\sqrt{\frac{(0.533)(0.467)}{1000}}$

(c) $0.533 \pm 1.960\sqrt{\frac{(0.533)(0.467)}{533}}$

(d) $0.533 \pm 1.960\sqrt{\frac{0.533}{1000}}$

(e) Not enough information to answer

Example 6 A company wants to find out what kinds of transportation its employees use to get to work. It conducts a survey of 537 employees, and 243 say they ride the bus. Construct a 95% confidence interval for the proportion of employees who ride the bus to and from work.

(a) (0.410, 0.495)
(b) (0.4521, 0.4539)
(c) (2.168, 2.252)
(d) (0.3962, 0.5098)
(e) (0.4316, 0.4744)

Margin of Error

You will hear a term in statistics - **margin of error**. When you hear this term, think of the end piece of your confidence interval, the part you "plus or minus".

Margin of Error = (critical value)*(standard deviation)

$$z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Ex7 A company wants to find out what kinds of transportation its employees use to get to work. It conducts a survey of 537 employees, and 243 say they ride the bus. To plan parking space distribution, management wants to estimate the proportion of employees who take the bus to within a 3% margin of error at a 90% confidence level. What's the sample size necessary to construct this interval?

(a) $n = 751$
(b) $n = 752$
(c) $n = 744$
(d) $n = 745$
(e) $n = 1058$

Example 8 A consumer group is interested in estimating the proportion of packages of ground beef sold at a particular store that have an actual fat content exceeding the fat content stated on the label. How many packages of ground beef should be tested to estimate this proportion to within 0.05 with 95% confidence?

MOE = $z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < 1.96\sqrt{\frac{0.5(1-0.5)}{n}} < 0.05$ $\hat{p} = 0.5$ recall that in the absense of a sample proportion we use 0.5 because it gives the maximum value for $\hat{p}(1-\hat{p})$

A little algebra give us $\frac{0.25}{n} < 0.0006508$ → $n > 384.16$ → $n = 385$

- The **standard error** of a statistic is the estimated standard deviation of the statistic.

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Checkpoint

1. You want to construct a 99% confidence interval for a sample of size 498. What's your critical z - value?

(a) 2.326
(b) 1.960
(c) 2.576
(d) Depends on the standard error
(e) Depends on the point estimate

2. The sample mean, $\bar{x}$, is called a _____ of the population mean μ.
(a) point estimate
(b) margin of error
(c) critical z - value
(d) confidence level
(e) interval estimate

3. Management at a seaside resort is publishing a brochure and wants to include a statement about the proportion of clear days during their peak season. Out of a random sample of 150 days from over the last two peak seasons, 117 days were recorded as clear. They want to estimate the proportion of clear days to within a 5% margin of error with a 95% confidence interval. What's the sample size necessary to construct this interval?

(a) 263
(b) 264
(c) 383

(d) 384
(e) 385

4. In a survey funded by Burroughs-Welcome, 750 of 1000 adult Americans said they didn't believe they could come down with a sexually transmitted disease. Construct a 95% confidence interval estimate of the proportion of adult Americans who don't believe they can contract and STD.

(a) (0.728, 0.772)
(b) (0.723, 0.777)
(c) (0.718, 0.782)
(d) (0.713, 0.787)
(e) We cannot construct an interval because the necessary assumptions have not been met.

5. A 90% confidence interval for a population proportion means that

(a) we are 90% confident that the interval will contain all possible sample proportions with the sample sample size taken from the given population.
(b) we are 90% confident that the population proportion will be the same as the sample proportion used in constructing the interval.
(c) we are 90% confident that the population proportion is included in the interval.
(d) None of the above

6. Suppose that a random sample of 100 high school classrooms in the state of California is selected and a 95% confidence interval for the proportion that has Internet access is (0.62, 0.78). Which of the following is a correct interpretation of the 95% confidence level?

(a) The method used to construct the interval will produce an interval that includes the value of the population proportion about 95% of the time in repeated sampling.
(b) We are 95% confident that the sample proportion is between 0.62 and 0.78.
(c) There is a 95% chance that the proportion of all high school classrooms in California that have Internet access is between 0.62 and 0.78.
(d) We are 95% confident that the proportion of all high school classrooms in California that have Internet access is between 0.62 and 0.78.
(e) None of the above is the correct interpretation of the confidence level.

6-1 Homework

1. What is the appropriate z critical value for each of the following confidence levels?
   a. 95%. b. 90%. c. 99%. D. 80%. E 85%

2. According to an Associated Press poll of 1002 randomly selected adults, a total of 82% said that reality TV shows are either 'totally made up' or 'mostly distorted'. Compute and interpret a 90% bound on the error of estimation for the reported percentage.

3. A sleep health study in 2016 that surveyed a representative sample of 734 college students found that 125 reported that they sleep with their cell phones near their bed and check their phones for something other than the time at least twice during the night. Compute and interpret a 95% confidence level for the proportion of students who check their phone at least twice a night for something other than the time.

4. The article "Kids Digital Day: Almost 8 Hours" (USA Today, January 20, 2010) summarized results from a national survey of 2002 Americans age 8 to 18. The sample was selected in a way that was expected to result in a sample representative of Americans in this age group.
   (a) Of those surveyed, 1321 reported owning a cell phone. Use this information to construct and interpret a 90% confidence interval estimate of the proportion of all Americans age 8 to 18 who own a cell phone.
   (b) Of those surveyed, 1522 reported owning an MP3 music player. Use this information to construct and interpret a 90% confidence interval estimate of the proportion of all Americans age 8 to 18 who own an MP3 music player.
   (c) Explain why the confidence interval from Part (b) is narrower than the confidence interval from Part (a) even though the confidence level and the sample size used to compute the two intervals was the same.

5. A question posed in the article "Academic Cheating, Aided by Cell Phones or Web, Shown to be Common" (Los Angeles Times, June 17, 2009) was: What proportion of college students have used cell phones to cheat on an exam? Suppose you have been asked to estimate this proportion for students enrolled at a large university. How many students should you include in your sample if you want to estimate this proportion to within .02 with 95% confidence?

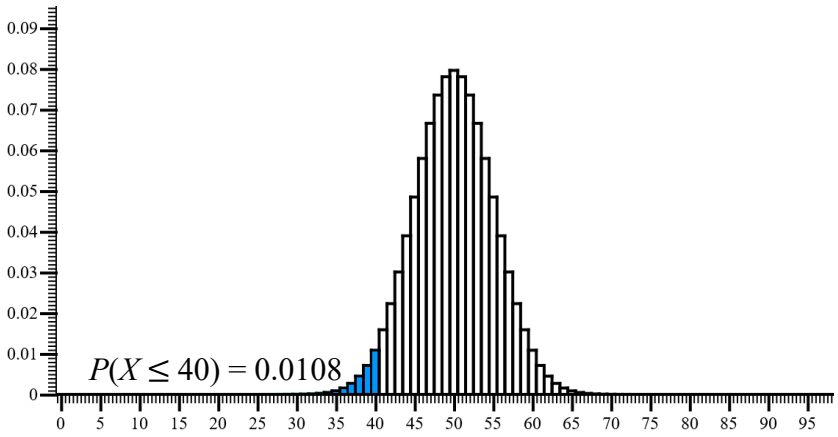
6-2 Hypothesis Testing for Population Proportions

Goal: 1. Conduct a test of significance for a population proportion.
2. Conduct a hypothesis test for a population proportion

This section introduces the specifics of **Hypothesis Testing**. Now that we've seen how to narrow our estimations of confidence intervals, we would like to get one step closer to being able to either draw an inference (not a conclusion) or at least eliminate some range of point estimates as possible parameters when sampling. In order to do this, we should first start with some kind of hypothetical parameter and subject it to random sample testing. What follows is sound statistical methodology for testing such hypotheses.

Example 1: Mr. Murphy gives you a two-sided coin and asks you to toss it 100 times recording your results. You conduct your 100 tosses and get 40 heads. Mr. Murphy claims that the coin is a fair one and that randomness alone accounts for the 40-60 outcome. Can you draw any conclusions or make any inferences about his claim?

First of all, what is the probability that you will get 40 heads in 100 tosses? Let's take a look at a binomial distribution graph. Below is a histogram of probabilities of every possible outcome when tossing a fair coin 100 times.



Using the Binomial Theorem and our calculators we get *binompdf*(100, 0.5, 40) = 0.0108 but this is just one specific result. The probability that you would get 40 or fewer heads is *binomcdf*(100, 0.5, 40) = 0.02844 which is still very unlikely. You are left with two options:

(a) The coin is not a truly fair coin *so we reject Mr. Murphy's claim*
(b) The coin is fair and this is just a fluke result *so we do not reject Mr. Murphy's claim*

Note that we did not use the term *accept*, we *reject* or we do not *reject*. We will discuss this important distinction soon.

Mr. Murphy's claim is what we call the ***null hypothesis***. We can only have one other hypothesis which is that Mr. Murphy's claim is not true. This is called the ***alternate hypothesis.*** We use probability to determine the likelihood of our sample outcome and decide whether to reject or not reject the null hypothesis.

To be sure, there is no way to be absolutely certain based solely on these results whether or not the coin is fair but we can make a reasonable inference here that the coin is not a fair one.

For the sake of the above example we used a binomial probability. For the rest of this course, we will be dealing only with probability distributions that can be assumed to be *normal* and make our inferences using *z*-scores.

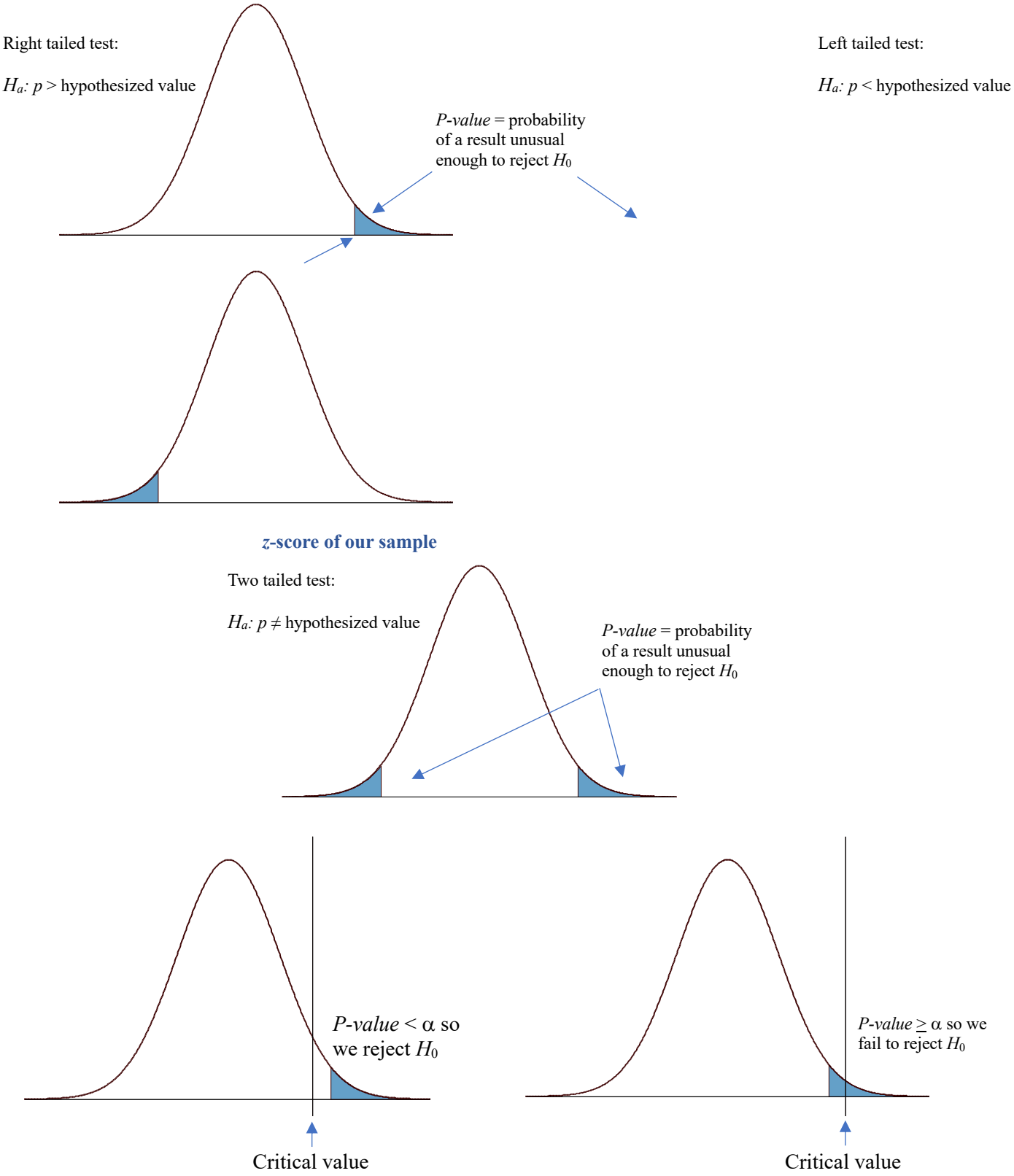
Example 2 The article "Credit Cards and College Students: Who Pays, Who Benefits?" described a study of credit card payment practices of college students. According to the authors of the article, the credit card industry asserts that at most 50% of college students carry a balance from month to month. However, the authors of the article report that, in a random sample of 310 college students, 217 carried a balance each month. Does this sample provide sufficient evidence to reject the industry claim?

Single Sample *z* Test for *p*

Null Hypothesis: H_0: p = hypothesized value or what we claim in the null hypothesis

Test Statistic: $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ Note that this is just the formula for a *z*-score $z_i = \dfrac{x_i - \mu}{\sigma}$

Alternate Hypothesis:	***P* – value:**
H_a: $p >$ hypothesized value	Area under *z* curve to right of calculated *z*
H_a: $p <$ hypothesized value	Area under *z* curve to left of calculated *z*
H_a: $p \neq$ hypothesized value	(1) 2(area to right of *z*) if *z* is positive, or (2) 2(area to left of *z*) if *z* is negative



Assumptions:

1. The sample proportion is from a **random sample** or **sample represents population**.

2. $np \geq 10$ and $n(1-p) \geq 10$ (for Normality)

3. The sample size is no more than 10% of the population size (SSSRTP)

STEPS IN HYPOTHESIS TESTING

1. Define the population characteristic (i.e. parameter) about which hypotheses are to be tested.

2. State the null hypothesis H_0.

3. State the alternative hypothesis H_a.

4. State the significance level α for the test.

5. Check all assumptions.

6. State the name of the test.

7. *State degrees of freedom if applicable* **(not applicable with proportions)**.

8. Display the test statistic to be used without any computation at this point. $z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$

9. Compute the value of the test statistic, showing specific numbers used.

10. Calculate the P – value.

11. Sketch a picture of the situation.

12. State the conclusion in two sentences -
    A. Summarize in theory discussing H_0. Always start by stating the P – value compared to the significance level, α, of the test
        - If the P – value is **less than** α, then we **reject the null hypothesis (H_0)** at the significance level we tested.
        - If the P – value is **greater than** α, then we **fail to reject the null hypothesis (H_0)** at the significance level we tested.

    B. Summarize in context discussing H_a .

        - **If we reject H_0** state that "we have evidence that the proportion of _____ is …, therefore, the *(initial claim)* is incorrect."
        - **If we fail to reject H_0** state that "we have insufficient evidence that the proportion of _____ is …, therefore, we cannot reject the *(initial claim)*."

These 12 steps will be given to you throughout Unit 6 but you will need to have them memorized by the end of the unit because you will need them from now on.

Important Note: We *NEVER* accept the null hypothesis, we either reject or fail to reject it.

Example 1 The article "Credit Cards and College Students: Who Pays, Who Benefits?" described a study of credit card payment practices of college students. According to the authors of the article, the credit card industry asserts that at most 50% of college students carry a balance from month to month. However, the authors of the article report that, in a random sample of 310 college students, 217 carried a balance each month. Does this sample provide sufficient evidence to reject the industry claim? We will answer this question by carrying out a hypothesis test using a 0.05 significance level. Does this sample provide sufficient evidence to reject the industry claim with a 0.05 significance level?

Example 2 Owners of a very large lake recently stocked the lake with bass and proudly proclaimed that 80% of the bass caught in the lake meet the required 15-inch minimum length (smaller fish must be thrown back). At the lake, 10 fishermen caught 51 bass, of which they were allowed to keep 27.

(a) Find a 95% confidence interval for the proportion of bass that meet the required 15-inch minimum length.
(b) Do the fishermen have evidence to show that the lake's proportion of bass that meet the required 15-inch minimum length is different from the owners' claim?
(c) Use your confidence interval to justify your decision in part (a).

Checkpoint
Multiple Choice

Questions #1-3: A major videocassette rental chain is considering opening a new store in an area that currently does not have any such stores. The chain will open if there is evidence that more than 5,000 of the 20,000 households in the area are equipped with videocassette recorders (VCRs). It conducts a telephone poll of 300 randomly selected households in the area and finds that 96 have VCRs.

1. State the test of interest to the rental chain:

(a) $H_0: p = 0.32$, $H_a: p > 0.32$ (b) $H_0: \mu = 5000$, $H_a: \mu > 5000$ (c) $H_0: p = 5000$, $H_a: p > 5000$

(d) $H_0: p = 0.25$, $H_a: p > 0.25$ (e) $H_0: \mu = 0.25$, $H_a: \mu > 0.25$

2. The P-value associated with the test statistic in this problem is approximately equal to

(a) 0.0026
(b) 0.0013
(c) 0.0051
(d) 0.1000
(e) 0.0125

3. The decision on the hypothesis test using a 3% level of significance is

(a) no decision should be made
(b) to reject H_0
(c) to fail to reject H_0
(d) to accept H_0
(e) to accept H_0

4. A *P*-value tells you:

(a) The probability that your results are statistically significant.
(b) The probability that your sample mean is equal to the population mean.
(c) The probability that you'd get results as extreme as you did, from random variation alone.
(d) The significance level.
(e) Whether to use a binomial, normal, or geometric distribution.

Free Response

5. The Public Policy Institute of California reported that 71% of people nationwide prefer to live in a single-family home. To determine whether the preferences of Californians are consistent with this nationwide figure, a random sample of 2002 Californians were interviewed. Of those, 1682 said they consider a single-family home the ideal. Can we reasonably conclude that the proportion of Californians who prefer a single-family home is different from the national figure? We will answer this question by carrying out a hypothesis test with $\alpha = 0.01$

6-2 Homework

1. Pairs of *P*-values and significance levels, α, are given. For each pair, state whether the observed *P*-value leads to rejection of H_0 at the given significance level.

a. *P*-value = 0.084, $\alpha = 0.05$
b. *P*-value = 0.003, $\alpha = 0.001$
c. *P*-value = 0.498, $\alpha = 0.05$
d. *P*-value = 0.084, $\alpha = 0.10$
e. *P*-value = 0.039, $\alpha = 0.01$
f. *P*-value = 0.218, $\alpha = 0.10$

2. Assuming a random sample from a large population, for which of the following null hypotheses and
sample sizes *n* is the large-sample proportion *z* test appropriate:
a. H_0: $p = 0.2$, $n = 25$
b. H_0: $p = 0.6$, $n = 210$
c. H_0: $p = 0.9$, $n = 100$
d. H_0: $p = 0.05$, $n = 75$

3. A press release about a paper that appeared in *The Journal of Youth and Adolescence* (August 16, 2013) was titled "Video Games Do Not Make Vulnerable Teens More Violent." The press release includes the following statement about the study described in the paper: "Study finds no evidence that violent video games increase antisocial behavior in youths with pre-existing psychological conditions." In the context of a hypothesis test with null hypothesis being that video games do not increase antisocial behavior, explain why the title of the press release is misleading.

4. The article "Theaters Losing Out to Living Rooms" (San Luis Obispo Tribune, June 17, 2005) reported that 470 of 1000 randomly selected adult Americans thought that the quality of movies being produced was getting worse.

a. Is there convincing evidence that fewer than half of adult Americans believe that movie quality is getting worse? Use a significance level of 0.05.
b. Suppose that the sample size had been 100 instead of 1000, and that 47 thought that the movie quality was getting worse (so that the sample proportion is still .47). Based on this sample of 100, is there convincing evidence that fewer than half of adult Americans believe that movie quality is getting worse? Use a significance level of .05.
c. Write a few sentences explaining why different conclusions were reached in the hypothesis tests of Parts (a) and (b).

5. The report "California's Education Skills Gap: Modest Improvements Could Yield Big Gains" (Public Policy Institute of California, April 16, 2008, www.ppic.org) states that nationwide, 61% of high school graduates go on to attend a two-year or four-year college the year after graduation. At that time the college-going rate for high school graduates in California was estimated to be 55%. Suppose that the estimate of 55% was based on a random sample of 1500 California high school graduates in 2009. Can we reasonably conclude that the proportion of California high school graduates in 2009 who attended college the year after graduation is different from the national figure? Use a significance level of $\alpha = 0.01$ to answer this question.

6. According to a report from the Institute for Higher Education titled, "Average Won't Do: Performance Trends in California Higher Education as a Foundation for Action" (January 2014), 53% of students graduating from California high schools go on to attend a 2 or 4-year college the year after graduation. A representative (random) sample of 1500 Bay Area students estimated the college going rate to be 50.6%. Use a hypothesis test with significance level $\alpha = 0.05$ to determine if there is sufficient evidence to suggest that the Bay Area proportion is different from that of the state.

7. If the significance level were $\alpha = 0.1$, would this change your inference based on the evidence? Explain.

6-3 Errors in Hypothesis Testing

Goals: 1. Determine Type I and Type II errors in context.
2. Recommendation of appropriate levels of significance.

A Quick Review of Hypothesis Testing

- A **test hypotheses or test procedure** is a method for using sample data to decide between two competing claims (hypothesis) about a population characteristic.

- The **null hypothesis**, denoted by H_0, is a claim about a population characteristic that is initially *assumed to be true*.

- The **alternate hypothesis**, denoted by H_a, is the competing claim.

- The two possible conclusions are then,
  - *reject* (infer that H_a is true) or
  - fail to reject (infer that H_0 is true).

- The form of a null hypothesis is

$$H_0: \quad \text{population characteristic} = \text{hypothesized value}$$

The alternate hypothesis has one of the following three forms:

$$H_a: \quad \text{population characteristic} > \text{hypothesized value}$$
$$H_a: \quad \text{population characteristic} < \text{hypothesized value}$$
$$H_a: \quad \text{population characteristic} \neq \text{hypothesized value}$$

- Hypotheses are **always** based on **parameters**, NEVER statistics.

Example 1 Testing to see whether a particular brand of lightbulb lasts longer than 1000 hours, we hypothesize that p is the proportion of light bulbs that last longer than 1000 hours. Suppose you are testing to see if more than half of the light bulbs last longer than 1000 hours. How would we write our null and alternative hypotheses?

Answer: $H_0: p = 0.5$
$H_a: p > 0.5$

Type I vs. Type II Errors

Errors can be made when testing hypotheses. In any hypothesis test we have 4 possibilities: two are correct and two are errors.

- **Type I Error** : the error of rejecting H_0 when H_0 is true (also called a false positive)
- **Type II Error**: the error of failing to reject H_0 when H_0 is false (also called a false negative)

	Fail to reject H_0	**Reject H_0**
H_0 true	*We rightly failed to reject H_0*	Type I error
H_a true	Type II error	*We rejected H_0 and we were right*

Example 2 A certain university has decided to introduce the use of plus and minus with letter grades, as long as there is evidence that more than 60% of the faculty favor the change. A random sample of faculty will be selected, and the resulting data will be used to test the relevant hypotheses.

(a) If p represents the true proportion of all faculty that favor a change to plus-minus grading, which of the following pair of hypotheses should the administration test. Explain your choice.

$$H_0 : p = 0.6 \quad H_a : p < 0.6 \qquad \text{or} \qquad H_0 : p = 0.6 \quad H_a : p > 0.6$$

(b) Explain the consequences of both a Type I and Type II error in this test
AP Note: Always write errors in terms of H_a

Answer: (a) The correct choice is II because we begin with the presumption that the percentage is 60 but we are seeking evidence that it is more.

Answer: (b) If we made a Type I error then we found evidence that more than 60% of the faculty were in favor of change to plus-minus grades when not more than 60% favor it. A Type II error would mean that we did not find significant evidence that more than 60% of the faculty favored the change when in fact the percentage is more than 60.

Example 3 The U.S. Department of Transportation reported that during a recent period, 77% of all domestic passenger flights arrived on time (meaning within 15 minutes of the scheduled arrival). Suppose that an airline with a poor on-time record decides to offer its employees a bonus if, in an upcoming month, the airline's proportion of on-time flights exceeds the overall industry rate of 0.77. Let p be the true proportion of the airline's flights that are on time during the month of interest.

(a) What are the null and alternate hypotheses?

(b) What are the Type I and Type II errors in this context?

(c) What are the consequences of these errors?

Answers: (a) $H_0 : p = 0.77$
$H_a : p > 0.77$

(b) A Type I error would be that the airline finds convincing evidence that more than 77% of domestic flights arrived on time when the percentage was not that high. A Type II error would be that the airline did not find significant evidence that the on time arrival percentage was higher than 77% when it actually was.

(c) The consequence of a Type I error would be that employees would receive an undeserved bonus. The consequence of a Type II error would be that employees would not receive a bonus that they had in fact earned.

- The probability of a Type I error is denoted by α (alpha) and is called the **level of significance** of the test. Thus, a test with $\alpha = 0.01$ is said to have a level of significance of 0.01 or to be a level 0.01 test.

- The probability of a Type II error is denoted by β (beta).

- An ideal test procedure would result in both $\alpha = 0$ and $\beta = 0$ -- but in order for this to happen we would need to conduct a census.

- After assessing the consequences of Type I and Type II errors, identify which error to control - *using a smaller α increases β, and vice versa.*

- The probability of *correctly* rejecting a null hypothesis H_0 is the complement of a Type II error. This is called the **Power** of a test and its probability is $1 - \beta$

There are 4 ways for power to increase:

1. Increase α because β will go down
2. Increase the sample size n
3. Decrease the standard error size (which would happen with #2 anyway)
4. The larger the discrepancy between the hypothesized parameter value and the true parameter value, the larger the power.

Example 4 An AP Stats student claims that 75% of SI students say that Dr. Quattrin's website is their go-to site for math enlightenment. Unconvinced by this claim, Mr. Murphy bets the student a commons cookie that it is much less and the student accepts provided that Mr. Murphy can find convincing statistical evidence that this percentage is less. They agree that if Mr. Murphy wins the bet but the student later finds that the evidence is wrong then Mr. Murphy will owe the student three commons cookies.

(a) What should Mr. Murphy's null and alternate hypotheses be for his test?

Let p be the true proportion of SI math students that use Dr. Quattrin's website as their first choice for math enlightenment

$H_0 : p = 0.75$
$H_a : p < 0.75$

(b) At least how many students should Mr. Murphy survey in order to validate his test? Explain

$np \geq 10 \rightarrow n(0.75) \geq 10$
$n(1-p) \geq 10 \rightarrow n(0.25) \geq 10$

$n = 40$ which is also less than 10% of the SI student body

(c) Mr. Murphy will choose between a significance level of 0.05 or 0.01 for his test. Which carries more risk? Explain.

Since the significance level α is the probability of a Type I Error which would mean that Mr. Murphy would owe the student three commons cookies, he will want that probability to be as low as possible. Mr. Murphy should choose $\alpha = 0.01$

(d) The student secretly calculates a 0.43 probability that even if the true percentage is 68, Mr. Murphy will fail to find evidence of it. What is the power of this test?

The power of a test is the complement of the probability of a Type II error. The student calculated that to be $\beta = 0.43$ so the power of the test is 0.57.

Checkpoint
1. What type of error occurs if you reject H_0, when, in fact, it is true?

(a) Type 1 error
(b) Type 2 error
(c) Type 3 error
(d) either a Type 1 or Type 2 error, depending on the level of significance.
(e) either a Type 1 or Type 2 error, depending on whether the test is one tail or two tail.

2. A psychologist claims that more than 6.1 percent of the population suffers from professional problems due to extreme shyness. Determine the null and alternate hypotheses.

(a) $H_0: \; p < 6.1\%$
$H_a: \; p \geq 6.1\%$

(b) $H_0: \; p = 6.1\%$
$H_a: \; p < 6.1\%$

(c) $H_0: \; p > 6.1\%$
$H_a: \; p \leq 6.1\%$

(d) $H_0: \; p = 6.1\%$
$H_a: \; p > 6.1\%$

(e) $H_0: \; p = 6.1\%$
$H_a: \; p \neq 6.1\%$

3. In hypothesis testing,

(a) the less the likelihood of a Type I error, the less the likelihood of Type II error
(b) the less the likelihood of a Type I error, the more the likelihood of Type II error
(c) the likelihood Type II errors will not be affected by the likelihood Type I errors
(d) the sum of the probabilities of Type I and Type II errors must equal 1
(e) the probability of committing a Type I error is β

4. In the past, the mean running time for a certain type of flashlight battery has been 9.6 hours. The manufacturer has introduced a change in the production method and wants to perform a significance test to determine whether the mean running time has increased as a result. The hypotheses are:

$$H_0: \; \mu = 9.6 \text{ hours}$$
$$H_a: \; \mu > 9.6 \text{ hours}$$

If the hypothesis test concludes the production change increases battery life, the manufacturer will change production methods and pass the increased cost onto the consumer. From the standpoint of the consumer, what α and β levels should be chosen?

(a) $\alpha = 0.05, \beta = 0.05$
(b) $\alpha = 0.07, \beta = 0.05$
(c) $\alpha = 0.10, \beta = 0.01$
(d) $\alpha = 0.05, \beta = 0.07$
(e) $\alpha = 0.01, \beta = 0.10$

5. Rejecting a true alternate hypothesis

(a) is a Type I error.
(b) has the probability of $1 - \beta$ of occurring.
(c) has the probability of α of occurring.
(d) is a Type II error.
(e) is a correct decision.

6-3 Homework

1) Working in a science lab on the 3rd floor of SI for hours on end, Mr. Maychrowitz devises a quick home test that will actually detect a common cold with what he thinks will be an accuracy of at least 98%. His AP Chem students past and present bring in friends from other high schools to see this potential marvel of modern testing and to be guinea pigs. They are able to try out his test on 600 students to assess his claim of 98% accuracy. Several have promised that if they find significant evidence that the accuracy is less than 98%, they will personally donate hours of their own time to help him in the lab to improve the accuracy.

   (a) What should their null and alternate hypotheses be for his test?

   (b) Why do they need so many students? What would be the minimum number in order to validate the test?

   (c) If they choose a level of significance of 0.05, what is the minimum z-value that would force them to reject his claim?

   (d) What is the probability that their test will lead to them *mistakenly* rejecting his claim of 98%?

   (e) One of his star students calculates that if Mr. M's 98% claim is 1% too high, the probability of their test not finding sufficient evidence of it is 0.57. What is the power of this test?

2) Referring to #1, define Type I and Type II errors. State and explain which error would be more costly and how that would affect how the level of significance is chosen.

3) A manufacturer of PC motherboards receives large shipments of CPU's from a supplier. In order to efficiently inspect for defects, when each shipment arrives, a sample is selected for inspection. If the proportion p of defective circuits exceeds 0.01, then the shipment is sent back as being of inferior quality.

   (a) Treating this as a hypothesis test of large sample proportions, state the null and alternate hypotheses.

   (b) State the minimum sample size to validate this sample.

   (c) In this context, define Type I and Type II errors.

   (d) From the calculator manufacturer's point of view, which type of error is considered more serious?

   (e) From the printed circuit supplier's point of view, which type of error is considered more serious?

6-4 Two Proportion Confidence Intervals & Hypothesis Tests

Goals: 1. Run a hypothesis test for the difference in means between two proportions.
2. Construct a confidence interval for the difference in means between two proportions.

Constructing a confidence interval for the difference between two proportions is not much different than confidence intervals from Unit 6-1. We are still using the product of the corresponding z-score and the standard deviation but in this case we have to adjust for two proportions by using the standard deviation formula we learned in Unit 5, applying the Pythagorean relationship between the standard deviations of two independent variables. Furthermore, while the three steps of checking our assumptions remains the same, we now have to check each distribution separately.

Constructing a Large Sample Confidence Interval for the Difference of Two Proportions

Given two independently selected random samples of large enough size, we can construct a confidence interval by adding the product of the z-score (which we're now calling the *critical value*) that corresponds to the percentile of the chosen interval and the standard deviation of the difference between the proportions:

Large Sample Confidence Interval: $\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Example 1 Some people seem to believe that you can fix anything with duct tape. Even so, many were skeptical when researchers announced that duct tape may be a more effective and less painful alternative to liquid nitrogen, which doctors routinely use to freeze warts. A study was conducted at the Madigan Army Medical Center where patients with warts were randomly assigned to either the duct tape treatment or the more traditional freezing treatment. Those in the duct tape group wore duct tape over the wart for 6 days, then removed the tape, soaked the area in water, and used an emery board to scrape the area. This process was repeated for a maximum of 2 months or until the wart was gone. Data consistent with values in the study are summarized in the following table:

Treatment	n	Number with Wart Successfully Removed
Liquid nitrogen freezing	100	60
Duct Tape	104	88

Construct a 99% confidence interval for the difference in proportions of successfully removed warts between duct tape and liquid nitrogen

$\hat{p}_1$ = proportion of warts removed with duct tape = $\frac{88}{104} \approx 0.846$

$\hat{p}_2$ = proportion of warts removed by LN freezing $\frac{60}{100} \approx 0.6$

$$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$0.846 - 0.6 \pm 2.576\sqrt{\frac{0.846(1-0.846)}{104} + \frac{0.6(1-0.6)}{100}} = (0.0905, 0.4018)$$

We are 99% confident that the difference between the proportion of warts removed by duct tape and warts removed by freezing will be between 9% and 40%

Example 2 Researchers at the National Cancer Institute released the results of a study examined the effect of weed-killing herbicides on house pets. Dogs, some of whom were from homes where the herbicide was used on a regular basis, were examined for the presence of malignant lymphoma. The following data was reported:

Group	**Sample Size**	**Number with Lymphoma**
Exposed	827	473
Unexposed	130	19

Use the given data to construct a 90% confidence interval for the difference between the proportion of exposed dogs that develop lymphoma and the proportion of unexposed dogs that develop lymphoma.

$\hat{p}_1 =$ proportion of exposed dogs $= \frac{473}{827} \approx 0.572$

$\hat{p}_2 =$ proportion of unexposed dogs $= \frac{19}{130} \approx 0.146$

$$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$0.572 - 0.146 \pm 1.645\sqrt{\frac{0.572(1-0.572)}{827} + \frac{0.146(1-0.146)}{130}} = (0.3675, 0.4841)$$

Large Sample Hypothesis Test for the Difference of Two Proportions

Conducting a hypothesis test with two independent random variables is largely the same process. Up until now, we've only discussed a one proportion z-test and it's corresponding test statistic. Now we will see how this changes with a two proportion z-test:

Steps for a Two Sample Hypothesis z-test

1. Define the population characteristics for each proportion to be tested
2. State the null hypothesis H_0
3. State the alternate hypothesis H_a
4. State the significance level for the test α
5. Check all assumptions (for each proportion)
6. State the name of the test to be used
7. State degrees of freedom if applicable
8. Write the test statistic (the formula you will use to find the z-value)
9. Calculate the test statistic showing your work
10. Calculate the P-value
11. Sketch a picture of the situation (Let the reader know which tail test you are using)
12. State the conclusion in two sentences:
    I. Reject or fail to reject
    II. State evidence in favor of or against

Before conducting our hypothesis tests, we need to see the differences in the test statistic from one sample proportions.

$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \dfrac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}$ but what is $\hat{p}_c$? This is the combined estimate of the common population proportion and the formula for it is $\hat{p}_c = \dfrac{x_1 + x_2}{n_1 + n_2}$ in which the x values are the number of successes in each sample. We can also write the formula this way: $\hat{p}_c = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$. This will be much easier to see once we revisit some examples

Now let's look at the previous examples as hypothesis tests

Example 4: Referring back to example 1, conduct a hypothesis test with significance level 0.05

Treatment	n	Number with Wart Successfully Removed
Liquid nitrogen freezing	100	60
Duct Tape	104	88

Do the data suggest that freezing is less successful than duct tape in removing warts? Let the level of significance be 0.01.

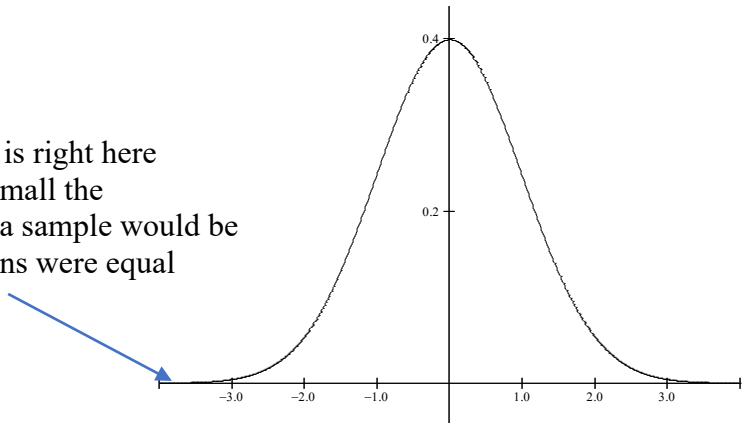
1. p_1 = proportion of warts removed by duct tape
   p_2 = proportion of warts removed by freezing
2. State the null hypothesis H_0: $p_1 = p_2$
3. State the alternate hypothesis H_a: $p_1 > p_2$
4. State the significance level for the test $\alpha = 0.05$

5. Check all assumptions (for each proportion)
6. 2 proportion z test
7. State degrees of freedom if applicable
8. $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \dfrac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}$ $\qquad \hat{p}_c = \dfrac{60+88}{100+104} = \dfrac{148}{204} \approx 0.7255$
9. $z = \dfrac{0.846 - 0.6}{\sqrt{\dfrac{0.7255(1-0.7255)}{100} + \dfrac{0.7255(1-0.7255)}{104}}} \approx 3.938$

10. *normalcdf*(3.938, 1E99) = 0.000041 < 0.05

11. Sketch a picture of the situation

The critical z value is right here which shows how small the probability of such a sample would be if the two proportions were equal



12. State the conclusion in two sentences:
    I. Because the P-value is less than alpha we reject the null hypothesis with significance level 0.05
    II. We have significant evidence that a larger proportion of warts can be removed with duct tape than with freezing

Example 2 Researchers at the National Cancer Institute released the results of a study examined the effect of weed-killing herbicides on house pets. Dogs, some of whom were from homes where the herbicide was used on a regular basis, were examined for the presence of malignant lymphoma. The following data was reported:

Group	Sample Size	Number with Lymphoma
Exposed	827	473
Unexposed	130	19

Do the data suggest with a significance level of 0.05 a difference between the proportion of exposed dogs that develop lymphoma and the proportion of unexposed dogs that develop lymphoma.

Checkpoint
Multiple Choice
1. In a random sample of 200 University of Manitoba graduate students, it was found that 66% of them had previously attended some other college or university. In a random sample of 100 University of Waterloo graduate students, it was found that 35% of them had previously attended some other college or university. A 95% confidence interval for estimating the difference in proportions of graduate students who had previously attended some other college or university between the University of Manitoba and the University of Waterloo is:

(a) $(0.66-0.35)\pm 1.96\sqrt{(0.3366)(0.6633)\left(\frac{1}{200}+\frac{1}{100}\right)}$

(b) $(0.66-0.35)\pm 1.96\sqrt{\frac{(0.66)(0.34)}{200}+\frac{(0.35)(0.65)}{100}}$

(c) $(0.66-0.35)\pm 1.96\sqrt{(0.5566)(0.4433)\left(\frac{1}{200}+\frac{1}{100}\right)}$

(d) $(0.33-0.35)\pm 1.96\sqrt{(0.5566)(0.4433)\left(\frac{1}{200}+\frac{1}{100}\right)}$

(e) $(0.33-0.35)\pm 1.645\sqrt{(0.5566)(0.4433)\left(\frac{1}{200}+\frac{1}{100}\right)}$

The next two questions refer to the following situation:
One criticism of reforestation efforts after timber harvesting is that too few of the seedling survive. An experiment was conducted to assess if mulching the slash (limbs, roots, small branches, etc.) and

leaving the mulch on the ground improves the survival rate compared to just leaving the slash on the ground. It is believed that mulching will cause the material to break down sooner and release the nutrients to the seedlings. A total of 500 seedlings were randomly assigned to the two treatments and the two year survival rate was measured. Of the 250 seedling receiving the "mulching" treatment, 75 survived; of the 250 seedlings receiving the "control" treatment, 55 survived.

2. The null and alternate hypotheses are: (m = mulch, c = control)

(a) $H_0 : p_m = 0.22, H_a : p_m > 0.22$
(b) $H_0 : \mu_m = 0.22, H_a : \mu_m > 0.22$
(c) $H_0 : p_m - p_c = 0, H_a : p_m - p_c > 0$
(d) $H_0 : \mu_m - \mu_c = 0, H_a : \mu_m - \mu_c > 0$
(e) $H_0 : p_m - p_c = 0, H_a : p_m - p_c \neq 0$

3. The value of the test statistic and the p-value are:

(a) 2.76, 0.003
(b) 2.05, 0.042
(c) 2.76, 0.006
(d) 2.05, 0.021
(e) 2.05, 0.011

Free Response

1. Even though landlords participating in a telephone survey indicated that they would generally be willing to rent to persons with AIDS, it was wondered whether this was true in actual practice. To investigate, researchers independently selected two random samples of 80 advertisements for rooms for rent from newspaper advertisements in three large cities. An adult male caller responded to each ad in the first sample of 80 and inquired about the availability of the room and was told that the room was still available in 61 of these calls. The same caller also responded to each ad in the second sample. In these calls, the caller responded to each ad in the second sample. In these calls, the caller indicated that he was currently receiving some treatment for AIDS and was about to be released from the hospital and would require a place to live. The caller was told that a room was available in 32 of these calls. Based on this information, the study concluded that "reference to AIDS substantially decreased the likelihood of a room being described as available." Do the data support this conclusion? Carry out a hypothesis test with $\alpha = 0.01$

6-4 Homework

1. The report "Audience Insights: Communicating to Teens (Aged 12–17)" (www.cdc.gov, 2009) described teens' attitudes about traditional media, such as TV, movies, and newspapers. In a representative sample of American teenage girls, 41% said newspapers were boring. In a representative sample of American teenage boys, 44% said newspapers were boring. Sample sizes were not given in the report.

   a. Suppose that the percentages reported had been based on a sample of 58 girls and 41 boys. Is there convincing evidence that the proportion of those who think that newspapers are boring is different for teenage girls and boys? Carry out a hypothesis test using $\alpha = 0.05$.

   b. Suppose that the percentages reported had been based on a sample of 2000 girls and 2500 boys. Is there convincing evidence that the proportion of those who think that newspapers are boring is different for teenage girls and boys? Carry out a hypothesis test using $\alpha = 0.05$.

c. Explain why the hypothesis tests in Parts (a) and (b) resulted in different conclusions.

2. Some commercial airplanes recirculate approximately 50% of the cabin air in order to increase fuel ef- ficiency. The authors of the paper "Aircraft Cabin Air Recirculation and Symptoms of the Common Cold" (Journal of the American Medical Association [2002]: 483–486) studied 1100 airline passengers who flew from San Francisco to Denver between January and April 1999. Some passengers traveled on airplanes that recirculated air and others traveled on planes that did not recirculate air. Of the 517 passengers who flew on planes that did not recirculate air, 108 reported post-flight respiratory symptoms, while 111 of the 583 passengers on planes that did recirculate air reported such symptoms. Is there sufficient evidence to conclude that the proportion of passengers with post-flight respiratory symptoms differs for planes that do and do not recirculate air? Test the appropriate hypotheses using a $\alpha = 0.05$. You may assume that it is reasonable to regard these two samples as being independently selected and as representative of the two populations of interest.

3. Public Agenda conducted a survey of 1379 parents and 1342 students in grades 6–12 regarding the im- portance of science and mathematics in the school curriculum (Associated Press, February 15, 2006). It was reported that 50% of students thought that understanding science and having strong math skills are essential for them to succeed in life after school, whereas 62% of the parents thought it was crucial for today's students to learn science and higher-level math. The two samples—parents and students—were selected independently of one another. Is there sufficient evidence to conclude that the proportion of parents who regard science and mathematics as crucial is different than the corresponding proportion for students in grades 6–12? Test the relevant hypotheses using a significance level of 0.05.