9.1 Intro to Significance Tests

In this chapter we are _____________________________________________________

Goal of a significance test: to assess evidence provided by data about a claim concerning a parameter.

**Scenario 1** Consider a factory making ingots. Ingots are huge pieces of metal often weighing more than 20,000 pounds! They must be cast in one huge piece, and cracking is a huge problem. Imagine your company specializes in creating these casts, and you have a failure rate of 20%. In other words, you need to recycle and start over 20% of your projects. You hire engineers and chemists to change the process to decrease the cracking problem. After implementing their changes, the next 400 ingots you cast have only a 17% failure rate. Should you declare victory? Has the cracking rate really decreased, or was this merely due to luck?

The 400 ingots are a ___________________. Is the 17% failure rate simply a result of _______________________________ or instead is my new process enough to convince me that the true cracking rate (the true \( p \)), is really below 20%.

**Scenario 2** A recent study on “The relative age effect and career success: Evidence from corporate CEOs” (Economics Letters 117 (2012)) suggests that people born in June and July are under-represented in the population of corporate CEOs. This “is consistent with the ‘relative-age effect’ due to school admissions grouping together children with age differences up to one year, with children born in June and July disadvantaged throughout life by being younger than their classmates born in other months.” In their sample of 375 corporate CEOs, only 45 (12%) were born in June and July. Is this convincing evidence that the true proportion \( p \) of all corporate CEOs born in June and July is smaller than 2/12?

What is the evidence that people born in June and July are underrepresented?

Give two explanations for why the sample proportion was below 2/12.

How can we decide which of the two explanations is more plausible?
In this chapter we speak in the language of __________________________.

Specifically, we will talk about a __________________ and an ______________ hypothesis.

What notation is used for each? What are some common mistakes when stating hypotheses?

For each scenario, define the parameter of interest and state appropriate hypotheses:

(a) The CEO study from the previous above.

(b) Mike is an avid golfer who would like to improve his play. A friend suggests getting new clubs and lets Mike try out his 7-iron. Based on years of experience, Mike has established that the mean distance that balls travel when hit with his old 7-iron is $\mu = 175$ yards with a standard deviation of $\sigma = 15$ yards. He is hoping that this new club will make his shots with a 7-iron more consistent (less variable), and so he goes to the driving range and hits 50 shots with the new 7-iron.

What is the difference between and one-sided and a two-sided alternative hypothesis? How can you decide which to use?

So, after we create our hypotheses, we calculate a ___________________. It is the probability of getting a statistic (the evidence) at least as extreme as the one observed in a study, assuming the null hypothesis is true.

In the CEO example, the $P$-value = 0.008. Interpret this value.

Example: When Mike was testing a new 7-iron, the hypotheses were $H_0: \sigma = 15$ versus $H_a: \sigma < 15$ where $\sigma$ = the true standard deviation of the distances Mike hits golf balls using the new 7-iron. Based on a sample of shots with the new 7-iron, the standard deviation was $s_x = 13.9$ yards.

(a) What is the evidence that the new club is better? What are the two explanations for the evidence?

(b) A significance test using the sample data produced a $P$-value of 0.28. Interpret the $P$-value in this context.
What are the two possible conclusions for a significance test?

What are some common errors that students make in their conclusions?
- Don’t forget context.
- Don’t accept the null hypotheses. We are saying that $H_0$ value is one of the plausible values for the parameter. Think about confidence intervals.
- Remember to link the conclusion to the p-value.

What is a significance level? When are the results of a study statistically significant?

Example: For his second semester project in AP Statistics, Zenon decided to investigate whether students at his school prefer name-brand potato chips to generic potato chips. After collecting data, Zenon performed a significance test using the hypotheses $H_0: p = 0.5$ versus $H_a: p > 0.5$ where $p$ = the true proportion of students at his school who prefer name-brand chips. The resulting $P$-value was 0.074. What conclusion would you make at each of the following significance levels?
(a) $\alpha = 0.10$

(b) $\alpha = 0.05$

What should be considered when choosing a significance level? (bottom line ... it’s pretty subjective!)

Consider a jury trial. What are the two possible errors? Which is worse?

In a significance test, what two errors can we make? Which error is worse?
Type I Error:
Type II Error:

Hint: Type II is when we fail “II” reject $H_0$.

Not anyone’s fault—no mistake was made in the calculations, etc.
Describe a Type I and a Type II error in the context of the CEO example. Which error could the researchers have made? Explain.

What is the probability of a Type I error? What can we do to reduce the probability of a Type I error? Are there any drawbacks to this?

- To reduce P(Type I error), demand that the evidence be very convincing before rejecting $H_0$. For example, demand that a murder conviction requires 10,000 eye witnesses. But, many guilty people would go free because there aren’t 10,000 eye witnesses. Making it harder to reject $H_0$ means that we are more likely to make a Type II error.

9.2 Significance Tests for Proportions

This section is about the details of testing a claim about a population proportion.

What are the three conditions for conducting a significance test for a population proportion? How are these different than the conditions for constructing a confidence interval for a population proportion?

Note: we always do calculations assuming the null hypothesis is true (so we use $p$ not $\hat{p}$)

What is a test statistic? What does it measure?

What are the four steps for conducting a significance test? What is required in each step?
What test statistic is used when testing for a population proportion?

What happens when the data (evidence) don’t support \( H_a \) at all?

Example: According to an article in the San Gabriel Valley Tribune (February 13, 2003), “Most people are kissing the ‘right way.’” That is, according to a study, the majority of couples prefer to tilt their heads to the right when kissing. In the study, a researcher observed a random sample of 124 kissing couples and found that 83/124 of the couples tilted to the right. Is this convincing evidence that couples really do prefer to kiss the right way?

Two-sided tests for a proportion

Example: Benford’s law and fraud

When the accounting firm AJL and Associates audits a company’s financial records for fraud, they often use a test based on Benford’s law. Benford’s law states that the distribution of first digits in many real-life sources of data is not uniform. In fact, when there is no fraud, about 30.1% of the numbers in financial records begin with the digit 1. However, if the proportion of first digits that are 1 is significantly different from 0.301 in a random sample of records, AJL and Associates does a much more thorough investigation of the company. Suppose that a random sample of 300 expenses from a company’s financial records results in only 68 expenses that begin with the digit 1. Should AJL and Associates do a more thorough investigation of this company?
Describe a Type I and Type II error in this context.

- **I:** Finding convincing evidence that the true proportion of expenses that begin with 1 is different than 0.301, when it really isn’t.
- **II:** Not finding convincing evidence that the true proportion of expenses that begin with 1 is different than 0.301, when it really is.

Can you use confidence intervals to decide between two hypotheses? What is an advantage to using confidence intervals for this purpose? Why don’t we always use confidence intervals?

*CI’s give more info*

*CI’s only match two-sided tests and SEs are slightly different*

Example: **Benford’s law and fraud**

A 95% confidence interval for the true proportion of expenses that begin with the digit 1 for the company in the previous Alternate Example is (0.180, 0.274). Does the interval provide convincing evidence that the company should be investigated for fraud?

Can you use your calculator to solve these?

One-Proportion zTest on Calculator... Sure, go ahead. Just make sure you name the procedure and report the test statistic and p-value.

In a recent year, 73% of first-year college students responding to a national survey identified “being very well-off financially” as an important personal goal. A state university finds that 132 of an SRS of 200 of its first-year students say that this goal is important. Is there convincing evidence at the \( \alpha = 0.05 \) significance level that the proportion of all first-year students at this university who think being very well-off is important differs from the national value, 73%?
Type II Errors and the Power of a Test

What is the power of a test? How is power related to the probability of a Type II error? Will you be expected to calculate the power of a test on the AP exam?

\[
\text{Power} = \text{probability of avoiding a type II error} = \text{probability of finding convincing evidence that } H_a \text{ is true when } H_0 \text{ is really true} = P(\text{reject } H_0 \text{ when parameter } = \text{some alternative value}) = 1 - P(\text{Type II error})
\]

In the potato example, suppose that the true proportion of blemished potatoes is \( p = 0.10 \). This means that we should reject \( H_0 \) because \( p = 0.10 > 0.08 \). *Write out Type I and II errors on the board.*

(a) Will the inspector be more likely to find convincing evidence that \( p > 0.08 \) if he looks at a small sample of potatoes or a large sample of potatoes? How does sample size affect power?

Large sample—more data gives a better chance of making a correct decision.

Bigger sample size = more power

(b) Will the inspector be more likely to find convincing evidence that \( p > 0.08 \) if he uses \( \alpha = 0.10 \) or \( \alpha = 0.01 \)? How does the significance level affect power?

If alpha goes up, beta goes down, and power \((1 - \beta)\) goes up.

Or: if alpha = 0.10, it will be easier to reject Ho \((p\text{-values only need to be less than 0.10, not 0.01})\). Because we will reject Ho more often, we are at less risk of a Type II error so more power.

(c) Suppose that a second shipment of potatoes arrives and the proportion of blemished potatoes is \( p = 0.50 \). Will the inspector be more likely to find convincing evidence that \( p > 0.08 \) for the first shipment \((p = 0.10)\) or the second shipment \((p = 0.50)\)? How does “effect size” affect power?

“Effect size” is the difference between the hypothesized value and the truth. Bigger effect size = more power. Easier to see the shipment is bad when the actual proportion of bad potatoes is far from the hypothesized value \((0.50 \text{ is farther from 0.08 than 0.10})\).

(d) Is there anything else that affects power?
Other sources of variability (good use of control/blocking/stratifying) helps to account for sources of variability, making power go up. Ex: River problem, cholesterol activity

(e) Suppose that the true proportion of blemished potatoes is \( p = 0.11 \). If \( \alpha = 0.05 \), the power of the test is 0.76. Interpret this value.

\[ \text{Given that the true proportion of blemished potatoes is } p = 0.11, \text{ there is a 0.76 probability of finding convincing evidence that } p > 0.08. \]

(f) What is the probability of a Type II error for this test? Interpret this value.

\[ P(\text{Type II}) = 1 - 0.76 = 0.24. \text{ Interpret}. \]

In the Benford’s Law and Fraud example from above, suppose that \( p = 0.25 \). That is, 25% of all financial records at this company begin with the digit 1. When \( \alpha = 0.05 \), the power of the test is 0.58.

(a) Interpret this value.

\[ \text{Given that the true proportion of expenses that begin with 1 is } p = 0.25, \text{ there is a 0.58 probability that there will be convincing evidence that the true proportion is different than 0.301.} \]

(b) How can AJL and Associates increase the power of their test?

\[ \text{Increase sample size, use alpha } = 0.10, \text{ use a reasonable stratified sample} \]

(c) For what values of \( p \) would the power of the test be greater than 0.58, assuming everything else stayed the same?

\[ \text{Any value of } p \text{ that is farther away from 0.301 than 0.25. So, } p < 0.25 \text{ and } p > 0.352. \]

9.3 Significance Tests for a Population Mean

What are the three conditions for conducting a significance test for a population mean? How are these different than the conditions for calculating a confidence interval for a population mean?
What test statistic do we use when testing a population mean?

What is a t distribution, anyway? Describe the shape, center, and spread of the t distributions (page 512)

- **Shape:** symmetric, unimodal, but not quite Normal. Heavier tails. Approaches standard Normal distribution as df increase.
- **Center:** 0, since t is a standardized score
- **Spread:** greater than a standard Normal distribution, but gets closer to Normal as the df increase. This means we need to go farther than 1.96 SD to have 95% confidence. More spread since it is calculated from two variables, not 1, and more variables gives more variability!

How do you calculate P-values using the t distributions?

Example: Abby and Raquel like to eat sub sandwiches. However, they noticed that the lengths of the “6-inch sub” sandwiches they get at their favorite restaurant seemed shorter than the advertised length. To investigate, they randomly selected 24 different times during the next month and ordered a “6-inch” sub. Here are the actual lengths of each of the 24 sandwiches (in inches):

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4.50</td>
<td>4.75</td>
<td>4.75</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>5.50</td>
</tr>
<tr>
<td>5.50</td>
<td>5.50</td>
<td>5.50</td>
<td>5.50</td>
<td>5.75</td>
<td>5.75</td>
<td>5.75</td>
</tr>
<tr>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.50</td>
<td>6.75</td>
<td>6.75</td>
</tr>
</tbody>
</table>

(a) Do these data provide convincing evidence at the \( \alpha = 0.10 \) level that the sandwiches at this restaurant are shorter than advertised, on average?

(b) Given your conclusion in part (a), which kind of mistake—a Type I or a Type II error—could you have made? Explain what this mistake would mean in context.
Alternate Example: *Don’t break the ice*

In the children’s game Don’t Break the Ice, small plastic ice cubes are squeezed into a square frame. Each child takes turns tapping out a cube of “ice” with a plastic hammer, hoping that the remaining cubes don’t collapse. For the game to work correctly, the cubes must be big enough so that they hold each other in place in the plastic frame but not so big that they are too difficult to tap out. The machine that produces the plastic cubes is designed to make cubes that are 29.5 millimeters (mm) wide, but the actual width varies a little. To ensure that the machine is working well, a supervisor inspects a random sample of 50 cubes every hour and measures their width. The Fathom output summarizes the data from a sample taken during one hour.

(a) Interpret the standard deviation and the standard error provided by the computer output.
(b) Do these data give convincing evidence that the mean width of cubes produced this hour is not 29.5 mm? Use a significance test with $\alpha = 0.05$ to find out.
(c) Calculate a 95% confidence interval for $\mu$. Does your interval support your decision from (b)?
Paired Data and Using Tests Wisely

Example: Is the express lane faster?

For their second semester project in AP Statistics, Libby and Kathryn decided to investigate which line was faster in the supermarket: the express lane or the regular lane. To collect their data, they randomly selected 15 times during a week, went to the same store, and bought the same item. However, one of them used the express lane and the other used a regular lane. To decide which lane each of them would use, they flipped a coin. If it was heads, Libby used the express lane and Kathryn used the regular lane. If it was tails, Libby used the regular lane and Kathryn used the express lane. They entered their randomly assigned lanes at the same time, and each recorded the time in seconds it took them to complete the transaction. Carry out a test to see if there is convincing evidence that the express lane is faster.

Since these data are paired, we will consider the differences in time (regular − express). Here are the 15 differences. In this case, a positive difference means that the express lane was faster.

<table>
<thead>
<tr>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
</tr>
<tr>
<td>246</td>
</tr>
<tr>
<td>−46</td>
</tr>
<tr>
<td>121</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>55</td>
</tr>
<tr>
<td>79</td>
</tr>
<tr>
<td>−94</td>
</tr>
<tr>
<td>−17</td>
</tr>
<tr>
<td>95</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>129</td>
</tr>
<tr>
<td>−39</td>
</tr>
<tr>
<td>42</td>
</tr>
</tbody>
</table>

State: We want to test the following hypotheses at the $\alpha = 0.05$ significance level: $H_0: \mu_d = 0$ versus $H_a: \mu_d > 0$ where $\mu_d$ = the true mean difference (regular − express) in time required to purchase an item at the supermarket. EVIDENCE FOR $H_a$: $\bar{x} = 42.7 > 0$.

Plan: If conditions are met, we will perform a paired t test for $\mu_d$.

- Random A random sample of times to make the purchases was selected, and the students were assigned to lanes at random.
- 10% Since we randomly selected the times to conduct the study from an infinite number of possible times, the differences should be independent.
- Normal/Large Sample The graphs do not show much skewness and there are no outliers, so it is reasonable to use t procedures for these data.

Do:

- Test statistic $t = \frac{42.7 - 0}{84.0} / \sqrt{\frac{1}{15}} = 1.97$
- P-value $P(t > 1.97)$ using the t distribution with $15 - 1 = 14$ degrees of freedom. Using technology, P-value = $t_{cdf}(1.97, 100, 14) = 0.034$.

Conclude: Since the P-value is less than $\alpha$ (0.034 < 0.05), we reject the null hypothesis. There is convincing evidence that the express lane is faster than the regular lane.
What is the difference between statistical and practical significance?

What is the problem of multiple tests?


Show 60 minutes video: http://www.cbsnews.com/video/watch/?id=7399362n

Follow up regarding FDA: http://marginalrevolution.com/marginalrevolution/2015/08/is-the-fda-too-conservative-or-too-aggressive.html