Chapter 21

More About Tests and Intervals
Zero In on the Null

Null hypotheses have special requirements.

To perform a hypothesis test, the null must be a statement about the value of a parameter for a model.

We then use this value to compute the probability that the observed sample statistic—or something even farther from the null value—might occur.
Zero In on the Null (cont.)

- How do we choose the null hypothesis? The appropriate null arises directly from the context of the problem—it is not dictated by the data, but instead by the situation.

- One good way to identify both the null and alternative hypotheses is to think about the *Why* of the situation.

- To write a null hypothesis, you can’t just choose any parameter value you like.
  - The null must relate to the question at hand—it is context dependent.
Zero In on the Null (cont.)

- There is a temptation to state your new *claim* as the null hypothesis.
  - However, you cannot prove a null hypothesis true.
- So, it makes more sense to use what you want to show as the *alternative*.
  - This way, when you reject the null, you are left with what you want to show.
How to Think About P-Values

- A P-value is a conditional probability—the probability of the observed statistic *given* that the null hypothesis is true.
  - The P-value is NOT the probability that the null hypothesis is true.
  - It’s not even the conditional probability that null hypothesis is true given the data.
- Be careful to interpret the P-value correctly.
What to Do with a High P-Value

- When we see a small P-value, we could continue to believe the null hypothesis and conclude that we just witnessed a rare event. But instead, we trust the data and use it as evidence to reject the null hypothesis.

- However big P-values just mean what we observed isn’t surprising. That is, the results are now in line with our assumption that the null hypothesis models the world, so we have no reason to reject it.

- A big P-value doesn’t prove that the null hypothesis is true, but it certainly offers no evidence that it is not true.

- Thus, when we see a large P-value, all we can say is that we “don’t reject the null hypothesis.”
Example p. 499 #1

- In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypothesis?
  - a) A business student conducts a taste test to see whether students prefer Diet Coke or Diet Pepsi.
Example p. 499 #1

- In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypothesis?
  
  a) A business student conducts a taste test to see whether students prefer Diet Coke or Diet Pepsi.

  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 $50\%$. ($p \neq 0.50$)
Example p. 499 #1

In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypothesis?

c) A budget override in a small town requires a two-thirds majority to pass. A local newspaper conducts a poll to see if there’s evidence it will pass.
Example p. 499 #1

- In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypothesis?

- c) A budget override in a small town requires a two-thirds majority to pass. A local newspaper conducts a poll to see if there’s evidence it will pass.

  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. ($p > 2/3$)
Example p. 499 #3

A medical researcher tested a new treatment for poison ivy against the traditional ointment. He concluded that the new treatment is more effective. Explain what the P-value of 0.047 means in this context.
Example p. 499 #3

A medical researcher tested a new treatment for poison ivy against the traditional ointment. He concluded that the new treatment is more effective. Explain what the P-value of 0.047 means in this context.

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.
Alpha Levels

- Sometimes we need to make a firm decision about whether or not to reject the null hypothesis.
- When the P-value is small, it tells us that our data are rare *given the null hypothesis*.
- How rare is “rare”? 
Alpha Levels (cont.)

- We can define “rare event” arbitrarily by setting a threshold for our P-value.
  - If our P-value falls below that point, we’ll reject $H_0$. We call such results statistically significant.
  - The threshold is called an alpha level, denoted by $\alpha$. 
Alpha Levels (cont.)

- Common alpha levels are 0.10, 0.05, and 0.01.
  - You have the option—almost the *obligation*—to consider your alpha level carefully and choose an appropriate one for the situation.

- The alpha level is also called the *significance level*.
  - When we reject the null hypothesis, we say that the test is “significant at that level.”
Alpha Levels (cont.)

- What can you say if the P-value does not fall below $\alpha$?
  - You should say that “The data have failed to provide sufficient evidence to reject the null hypothesis.”
  - Don’t say that you “accept the null hypothesis.”
Recall that, in a jury trial, if we do not find the defendant guilty, we say the defendant is “not guilty”—we don’t say that the defendant is “innocent.”
Alpha Levels (cont.)

- The P-value gives the reader far more information than just stating that you reject or fail to reject the null.
- In fact, by providing a P-value to the reader, you allow that person to make his or her own decisions about the test.
  - What you consider to be statistically significant might not be the same as what someone else considers statistically significant.
  - There is more than one alpha level that can be used, but each test will give only one P-value.
Significant vs. Important

- What do we mean when we say that a test is statistically significant?
  - All we mean is that the test statistic had a P-value lower than our alpha level.
- Don’t be lulled into thinking that statistical significance carries with it any sense of practical importance or impact.
Significant vs. Important (cont.)

- For large samples, even small, unimportant ("insignificant") deviations from the null hypothesis can be statistically significant.
- On the other hand, if the sample is not large enough, even large, financially or scientifically "significant" differences may not be statistically significant.
- It’s good practice to report the magnitude of the difference between the observed statistic value and the null hypothesis value (in the data units) along with the P-value on which we base statistical significance.
Confidence Intervals and Hypothesis Tests

- Confidence intervals and hypothesis tests are built from the same calculations.
  - They have the same assumptions and conditions.
- You can approximate a hypothesis test by examining a confidence interval.
  - Just ask whether the null hypothesis value is consistent with a confidence interval for the parameter at the corresponding confidence level.
Because confidence intervals are two-sided, they correspond to two-sided tests.

In general, a confidence interval with a confidence level of \( C \)% corresponds to a two-sided hypothesis test with an \( \alpha \)-level of 100 – \( C \)%.
Confidence Intervals and Hypothesis Tests (cont.)

- The relationship between confidence intervals and one-sided hypothesis tests is a little more complicated.
  - A confidence interval with a confidence level of $C\%$ corresponds to a one-sided hypothesis test with an $\alpha$-level of $\frac{1}{2}(100 - C)\%$. 
Example p. 499 #5

- A researcher developing scanners to search for hidden weapons at airports has concluded that a new device is significantly better than the current scanner. He made this decision based on a test using a $\alpha = 0.05$. Would he have made the same decision at $\alpha = 0.10$? How about $\alpha = 0.01$? Explain.
Example p. 499 #5

- A researcher developing scanners to search for hidden weapons at airports has concluded that a new device is significantly better than the current scanner. He made this decision based on a test using a $\alpha = 0.05$. Would he have made the same decision at $\alpha = 0.10$? How about $\alpha = 0.01$? Explain.

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.
Example p. 500 #9

- In August 2004, *Time* magazine reported the results of a random telephone poll commissioned by the Spike network. Of the 1302 men who responded, only 39 said that their most important measure of success was their work.

- a) Estimate the percentage of all American males who measure success primarily from their work. Use a 98% confidence interval. Check the conditions first.
Example p. 500 #9

- In August 2004, *Time* magazine reported the results of a random telephone poll commissioned by the Spike network. Of the 1302 men who responded, only 39 said that their most important measure of success was their work.

- b) Some believe that few contemporary men judge their success primarily by their work. Suppose we wished to conduct a hypothesis test to see if the fraction has fallen below the 5% mark. What does your confidence interval indicate? Explain.
In August 2004, *Time* magazine reported the results of a random telephone poll commissioned by the Spike network. Of the 1302 men who responded, only 39 said that their most important measure of success was their work.

c) What is the level of significance of this test? Explain.
A 95% Confidence Interval for Small Samples

- When the Success/Failure Condition fails, all is not lost.
- A simple adjustment to the calculation lets us make a confidence interval anyway.
- All we do is add four *phony* observations, two successes and two failures.
- So instead of \( \hat{p} = \frac{y}{n} \), we use the adjusted proportion

\[
\tilde{p} = \frac{y + 2}{n + 4}
\]
A Better Confidence Interval for Proportions (cont.)

- Now the adjusted interval is

\[
\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}
\]

- The adjusted form gives better performance overall and works much better for proportions of 0 or 1.

- It has the additional advantage that we no longer need to check the Success/Failure Condition.
*A Better Confidence Interval for Proportions (cont.)

- Ex: pg. 490-491

\[ \tilde{p} = \frac{y + 2}{n + 4} \]

\[ \tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \]

- Only 3 failures -> Success/Failure Condition not met
- Add 2 successes and 2 failures
Making Errors

Here’s some shocking news for you: nobody’s perfect. Even with lots of evidence we can still make the wrong decision.

When we perform a hypothesis test, we can make mistakes in two ways:

I. The null hypothesis is true, but we mistakenly reject it. (Type I error)
   False Positive (patient is healthy but test positive for disease)

II. The null hypothesis is false, but we fail to reject it. (Type II error)
   False negative (infected person is diagnosed as disease free)
Making Errors (cont.)

- Which type of error is more serious depends on the situation at hand. In other words, the gravity of the error is context dependent.

- Here’s an illustration of the four situations in a hypothesis test:

```
My Decision

Reject H₀  Retain H₀

H₀ True  Type I Error  OK
H₀ False  OK  Type II Error
```

The Truth
Making Errors (cont.)

- How often will a Type I error occur?
  - Since a Type I error is rejecting a true null hypothesis, the probability of a Type I error is our $\alpha$ level.

- When $H_0$ is false and we reject it, we have done the right thing.
  - A test’s ability to detect a false hypothesis is called the power of the test.
Making Errors (cont.)

- When $H_0$ is false and we fail to reject it, we have made a Type II error.
  - We assign the letter $\beta$ to the probability of this mistake.
  - It’s harder to assess the value of $\beta$ because we don’t know what the value of the parameter really is. (We base everything on $H_0$ being true, not false)
  - There is no single value for $\beta$--we can think of a whole collection of $\beta$’s, one for each incorrect parameter value.
Making Errors (cont.)

- One way to focus our attention on a particular $\beta$ is to think about the **effect size**.
  - Ask “*How big a difference would matter?*”

- We could reduce $\beta$ for all alternative parameter values by increasing $\alpha$.
  - This would reduce $\beta$ but increase the chance of a Type I error.

- This tension between Type I and Type II errors is inevitable.

- The only way to reduce both types of errors is to collect more data. Otherwise, we just wind up trading off one kind of error against the other.
Power

- The **power** of a test is the probability that it correctly rejects a false null hypothesis.
- When the power is high, we can be confident that we’ve looked hard enough at the situation.
- The power of a test is $1 - \beta$; because $\beta$ is the probability that a test *fails* to reject a false null hypothesis and power is the probability that it does reject.
Power (cont.)

- Whenever a study fails to reject its null hypothesis, the test’s power comes into question.
- When we calculate power, we imagine that the null hypothesis is false.
- The value of the power depends on how far the truth lies from the null hypothesis value.
  - The distance between the null hypothesis value, \( p_0 \), and the truth, \( p \), is called the effect size.
  - Power depends directly on effect size.
A Picture Worth \( \frac{1}{P(z > 3.09)} \) Words

- Obtaining a larger sample size decreases the probability of a Type II error, so it increases the power (a test’s ability to detect a false hypothesis).
- It also makes sense that the more we’re willing to accept a Type I error, the less likely we will be to make a Type II error.
A Picture Worth Words (cont.)

\[ P(z > 3.09) = \frac{1}{P(z > 3.09)} \]

- This diagram shows the relationship between these concepts:

Suppose the Null Hypothesis is true.

Suppose the Null Hypothesis is not true.

\( P \)

\( p \)

\( p^* \)
Reducing Both Type I and Type II Error

- The previous figure seems to show that if we reduce Type I error, we must automatically increase Type II error.
- But, we can reduce both types of error by making both curves narrower.
- How do we make the curves narrower? Increase the sample size.
Reducing Both Type I and Type II Error (cont.)

- This figure has means that are just as far apart as in the previous figure, but the sample sizes are larger, the standard deviations are smaller, and the error rates are reduced:
Reducing Both Type I and Type II Error (cont.)

- Original comparison of errors:
- Comparison of errors with a larger sample size:
What Can Go Wrong?

- Don’t interpret the P-value as the probability that $H_0$ is true.
  - The P-value is about the data, not the hypothesis.
  - It’s the probability of observing data this unusual, given that $H_0$ is true, not the other way around.
- Don’t believe too strongly in arbitrary alpha levels.
  - It’s better to report your P-value and a confidence interval so that the reader can make her/his own decision.
What Can Go Wrong? (cont.)

- Don’t confuse practical and statistical significance.
  - Just because a test is statistically significant doesn’t mean that it is significant in practice.
  - And, sample size can impact your decision about a null hypothesis, making you miss an important difference or find an “insignificant” difference.

- Don’t forget that in spite of all your care, you might make a wrong decision.
What have we learned?

- There’s a lot more to hypothesis testing than a simple yes/no decision.
  - Small P-value, indicates evidence against the null hypothesis, not that it is true.
  - Alpha level establishes level of proof, determines the critical value $z$ that leads us to reject null hypothesis.
- Hypothesis test gives an answer to decision about parameter; confidence interval tells us plausible values of that parameter.
What have we learned?

- And, we’ve learned about the two kinds of errors we might make and seen why in the end we’re never sure we’ve made the right decision.
- Type I error, reject a true null hypothesis, probability is $\alpha$.
- Type II error, fail to reject a false null hypothesis, probability is $\beta$.
- Power is the probability we reject a null hypothesis when it is false, $1 - \beta$.
- Larger sample size increase power and reduces the chances of both kinds of errors.
Example p. 500 #15

Before lending someone money, banks must decide whether they believe the applicant will repay the loan. One strategy used is a point system. Loan officers assess information about the applicant, totaling points they award for the person’s income level, credit history, current debt burden, and so on. The higher the point total, the more convinced the bank is that it’s safe to make the loan. Any applicant with a lower point total than a certain cutoff score is denied a loan.

We can think of this decision as a hypothesis test. Since the bank makes its profit from the interest collected on repaid loans, their null hypothesis is that the applicant will repay the loan and therefore should get the money. Only if the person’s score falls below the minimum cutoff will the bank reject the null and deny the loan. This system is reasonably reliable, but, of course, sometimes there are mistakes.
Example p. 500 #15

a) When a person defaults on a loan, which type of error did the bank make?

b) Which kind of error is it when the bank misses an opportunity to make a loan to someone who would have repaid it?

c) Suppose the bank decides to lower the cutoff score from 250 points to 200. Is that analogous to choosing a higher or lower value of $\alpha$ for a hypothesis test? Explain.

d) What impact does this change in the cutoff value have on the chance of each type of error?
Exercise 15 describes the loan score method a bank uses to decide which applicants it will lend money. Only if the total points awarded for various aspects of an applicant’s financial condition fail to add up to a minimum cutoff score set by the bank will the loan be denied.

a) In this context, what is meant by the power of the test?

b) What could the bank do to increase the power?

c) What’s the disadvantage of doing that?
A clean air standard requires that vehicle exhaust emissions not exceed specified limits for various pollutants. Many states require that cars be tested annually to be sure they meet these standards. Suppose state regulators double-check a random sample of cars that a suspect repair shop has certified as okay. They will revoke the shop’s license if they find significant evidence that the shop is certifying vehicles that do not meet standards.

- a) In this context, what is a Type I error?

- b) In this context, what is a Type II error?
Example p. 501 #21

- A clean air standard requires that vehicle exhaust emissions not exceed specified limits for various pollutants. Many states require that cars be tested annually to be sure they meet these standards. Suppose state regulators double-check a random sample of cars that a suspect repair shop has certified as okay. They will revoke the shop’s license if they find significant evidence that the shop is certifying vehicles that do not meet standards.

  - c) Which type of error would the shop’s owner consider more serious?

  - d) Which type of error might environmentalists consider more serious?
A company is sued for job discrimination because only 19% of the newly hired candidates were minorities when 27% of all applicants were minorities. Is this strong evidence that the company’s hiring practices are discriminatory?

a) Is this a one-tailed or a two-tailed test? Why?

b) In this context, what would a Type I error be?

c) In this context, what would a Type II error be?

d) In this context, what is meant by the power of the test?
Example p. 502 #25

- A company is sued for job discrimination because only 19% of the newly hired candidates were minorities when 27% of all applicants were minorities. Is this strong evidence that the company’s hiring practices are discriminatory?

- e) If the hypothesis is tested at the 5% level of significance instead of 1%, how will this affect the power of the test?

- f) The lawsuit is based on the hiring of 37 employees. Is the power of the test higher than, lower than, or the same as it would be if it were based on 87 hires?
Homework

- Chapter 21 Homework: p. 499 # 2(1), 4(3), 6(5), 10(9), 16(15), 18(17), 22(21), 26(25)