Chapter 21 Summary
More about Tests

What have we learned?

- There's a lot more to hypothesis testing than a simple yes/no decision.
- 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.

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—would occur.
- 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.
- A 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.
- There is a temptation to state your *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.

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"?
- 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 α.
- 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 α?
  - 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."
- 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.

What Not to Say About Significance

- 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.
- 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.

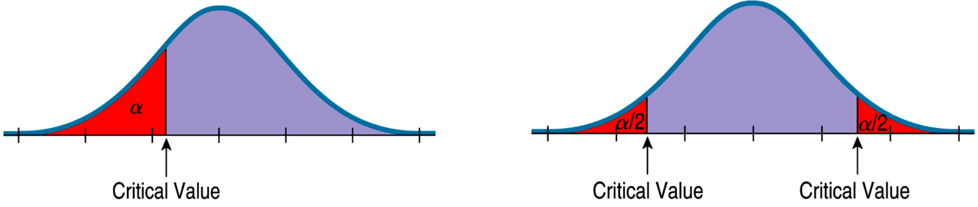
Critical Values Again

- When making a confidence interval, we've found a critical value to correspond to our selected confidence level.
- Prior to the use of technology, P-values were difficult to find, and it was easier to select a few common alpha values and learn the corresponding critical values for the Normal model.
- Rather than looking up your *z*-score in the table, you could just check it directly against these critical values.
  - Any *z*-score larger in magnitude than a particular critical value leads us to reject H_0.
  - Any *z*-score smaller in magnitude than a particular critical value leads us to fail to reject H_0.
- Here are the traditional critical values from the Normal model:

α	1-sided	2-sided
0.05	1.645	1.96
0.01	2.28	2.575
0.001	3.09	3.29

Critical Values Again (cont.)

- When the alternative is one-sided, the critical value puts all of α on one side (see below)



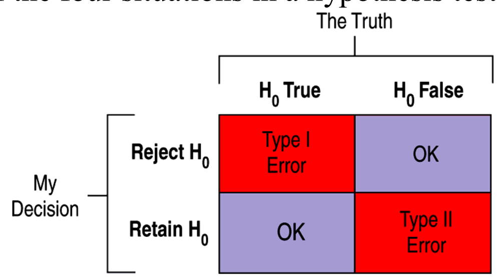
- When the alternative is two-sided, the critical value splits α equally into two tails (see above)

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 α-level of 100 – *C*%.
- The relationship between confidence intervals an 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 α-level of ½(100 – *C*)%.

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)
  II. The null hypothesis is false, but we fail to reject it. (Type II error)
- 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:



		The Truth	
		H_0 True	H_0 False
My Decision	Reject H_0	Type I Error	OK
	Retain H_0	OK	Type II Error

- 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 α 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.)

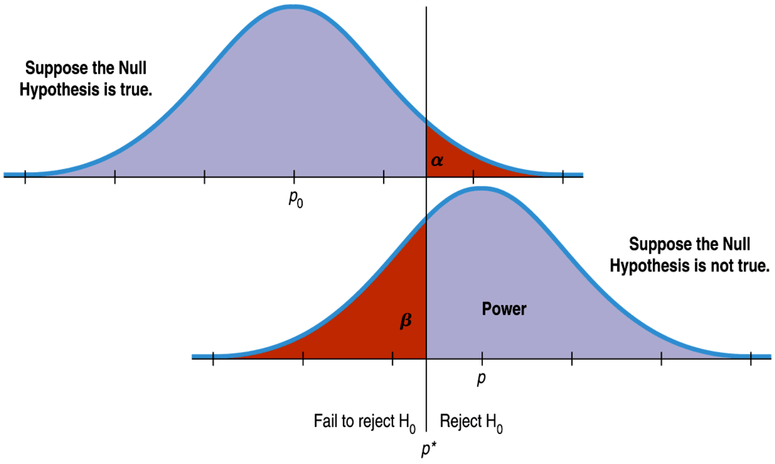
- When H_0 is false and we fail to reject it, we have made a Type II error.
    - We assign the letter β to the probability of this mistake.
    - It's harder to assess the value of β because we don't know what the value of the parameter really is.
    - There is no single value for β--we can think of a whole collection of β's, one for each incorrect parameter value.
- One way to focus our attention on a particular β is to think about the effect size.
    - Ask "*How big a difference would matter?*"
- We could reduce β for *all* alternative parameter values by increasing α.
    - This would reduce β 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 – β.
- 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.

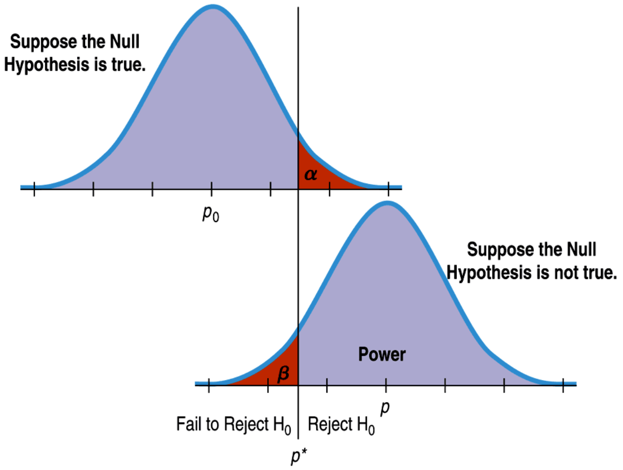
A Picture Worth a Thousand Words

- The large the effect size, the easier it should be to see it.
- Obtaining a larger sample size decreases the probability of a Type II error, so it increases the power.
- 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.
- This diagram shows the relationship between these concepts:

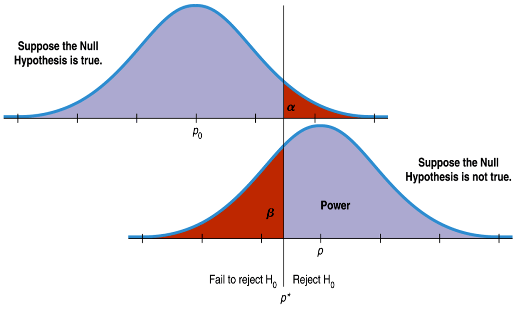

Suppose the Null Hypothesis is true.
α
p_0
Suppose the Null Hypothesis is not true.
Power
β
p
Fail to reject H_0
Reject H_0
p^*

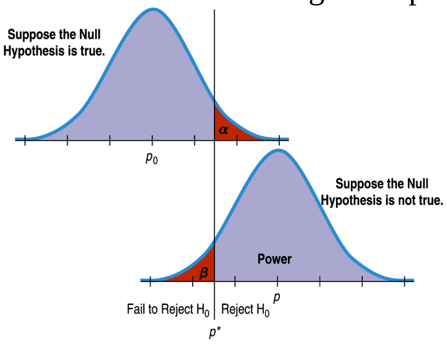
Reducing Both Type I and Type II Error

- 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.
- 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:


Suppose the Null Hypothesis is true.
α
p_0
Suppose the Null Hypothesis is not true.
Power
β
p
Fail to Reject H_0
Reject H_0
p^*

- Original comparison of errors: Comparison of errors with a larger sample size:


Suppose the Null Hypothesis is true.
α
p_0
Suppose the Null Hypothesis is not true.
Power
β
p
Fail to reject H_0
Reject H_0
p^*


Suppose the Null Hypothesis is true.
α
p_0
Suppose the Null Hypothesis is not true.
Power
β
p
Fail to Reject H_0
Reject H_0
p^*

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 the data *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.
- 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.