STA 2023

Module 9
More About Tests and Intervals

Learning Objectives

Upon completing this module, you should be able to

1. Define and apply Type I Error and Type II Error.
2. Recognize the power of the test is the probability that we reject the null hypothesis when it's false.
3. Recognize a small P-value is an indicator of evidence against the null hypothesis.
4. Utilize the significance level (alpha level) to establish the level of proof.
5. Identify the connection between hypothesis test and confidence interval.

Let's Have a Quick Review

Null and Alternative Hypotheses; Hypothesis Test

Null hypothesis: A hypothesis to be tested. We use the symbol H_0 to represent the null hypothesis.

Alternative hypothesis: A hypothesis to be considered as an alternative to the null hypothesis. We use the symbol H_a to represent the alternative hypothesis.

Hypothesis test: The problem in a hypothesis test is to decide whether the null hypothesis should be rejected in favor of the alternative hypothesis.

What are Hypotheses?

- Hypotheses are working models that we adopt temporarily.
- Our starting hypothesis is called the null hypothesis.
- The null hypothesis, that we denote by H_0, specifies a population model parameter of interest and proposes a value for that parameter.
- We usually write down the null hypothesis in the form H_0: *parameter* = *hypothesized* value.
- The alternative hypothesis, which we denote by H_A, contains the value of the parameter that we consider plausible when we reject the null hypothesis.

Testing Hypotheses

- The first step in defining the null and alternative hypotheses is to determine which parameter is being tested. A parameter describes a population. Examples are the population proportion, the population standard deviation and a population mean.
- The next step is to define the null hypothesis, specifies a population model parameter of interest and proposes a value for that parameter.
  - We might have, for example, H_0: *parameter* = 0.20
  - We want to compare our data to what we would expect, given that H_0 is true.
- Note that the null hypothesis is a statement of equality (with the equal sign.)

What are the Three Possible Alternative Hypotheses?

- There are three possible alternative hypotheses:
  - H_A: *parameter* < *hypothesized value*
  - H_A: *parameter* ≠ *hypothesized value*
  - H_A: *parameter* > *hypothesized value*

A Second Look at Null Hypotheses

- 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.

A Second Look at Null Hypotheses (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.

A Second Look at Null Hypotheses (cont.)

- 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-value?

- 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.

How to think about P-value? (cont.)

- 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.

How to think about P-value? (cont.)

- 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.”

Alpha Level = Significance Level

- 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 Level = Significance Level (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 α.

Alpha Level = Significance Level (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 Level = Significance Level (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."

Alpha Level = Significance Level (cont.)

- The P-value gives us far more information than just stating that we reject or fail to reject the null.
- In fact, by providing a P-value to the reader, we 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.

Statistically Significant

- 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.

Statistically Significant (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 Interval 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.

Confidence Interval and Hypothesis Tests (cont.)

- 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 $100 - C$%.

Confidence Interval 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 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:

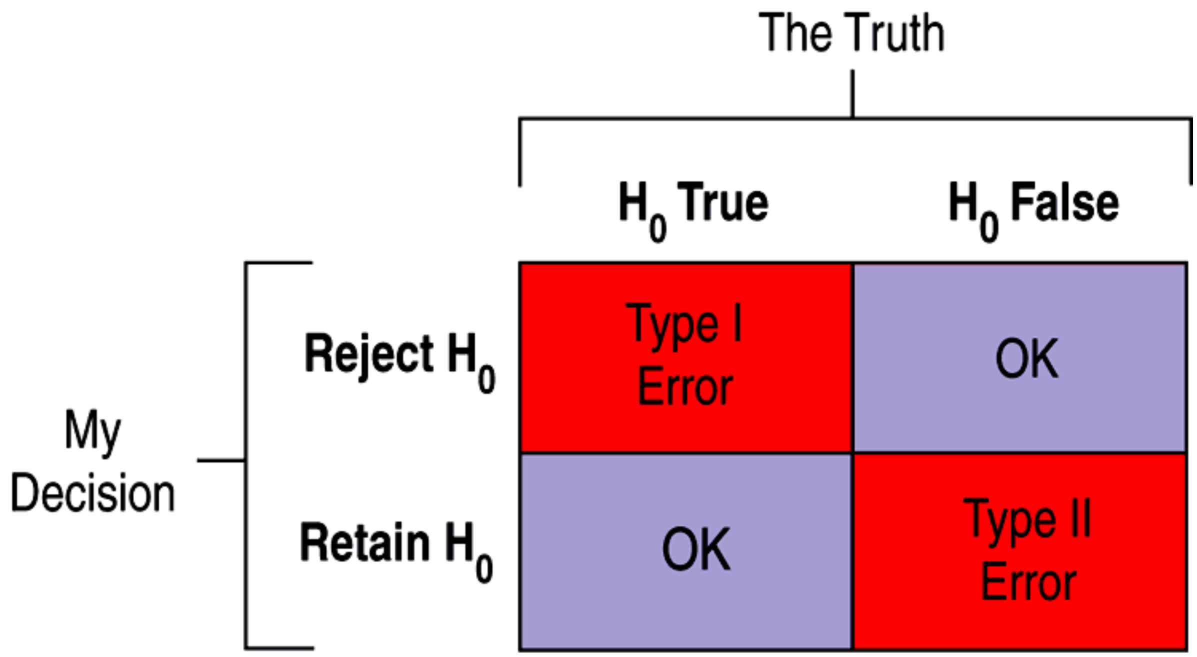
I. The null hypothesis is true, but we mistakenly reject it. (Type I error)

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

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:



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 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.

Making Errors (cont.)

- One way to focus our attention 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 of a Test

- 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 β is the probability that a test *fails* to reject a false null hypothesis and power is the probability that it does reject.

Power of a Test (cont.)

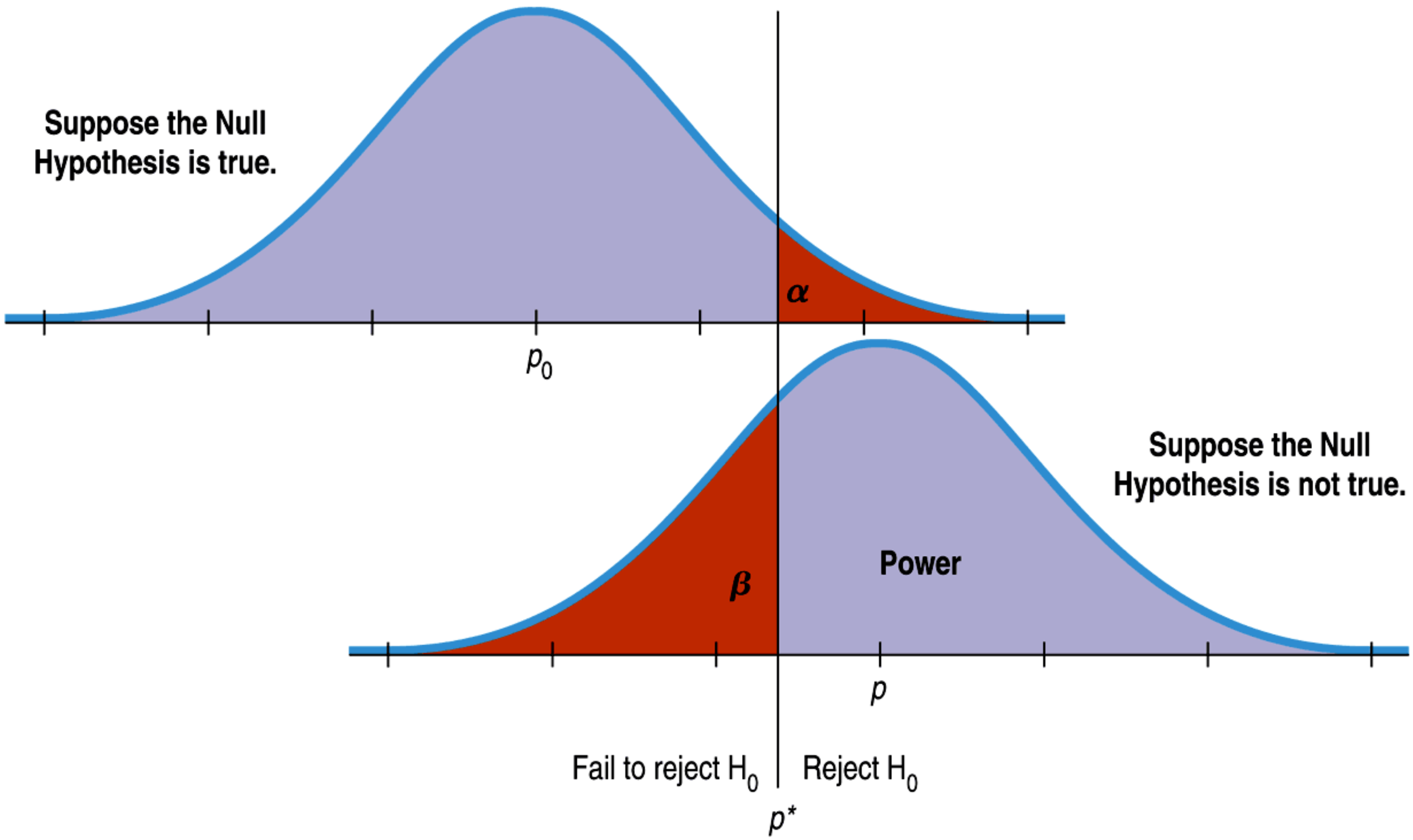
- Whenever a study fails to reject its null hypothesis, the test's power comes into question.
- When we calculate power, we imagine that the null hypothesis is false.
- The value of the power depends on how far the truth lies from the null hypothesis value.
  - The distance between the null hypothesis value, p_0, and the truth, p, is called the effect size.
  - Power depends directly on effect size.

Power of a Test (cont.)

- The larger 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.

Power of a Test (cont.)

This diagram shows the relationship between these concepts:

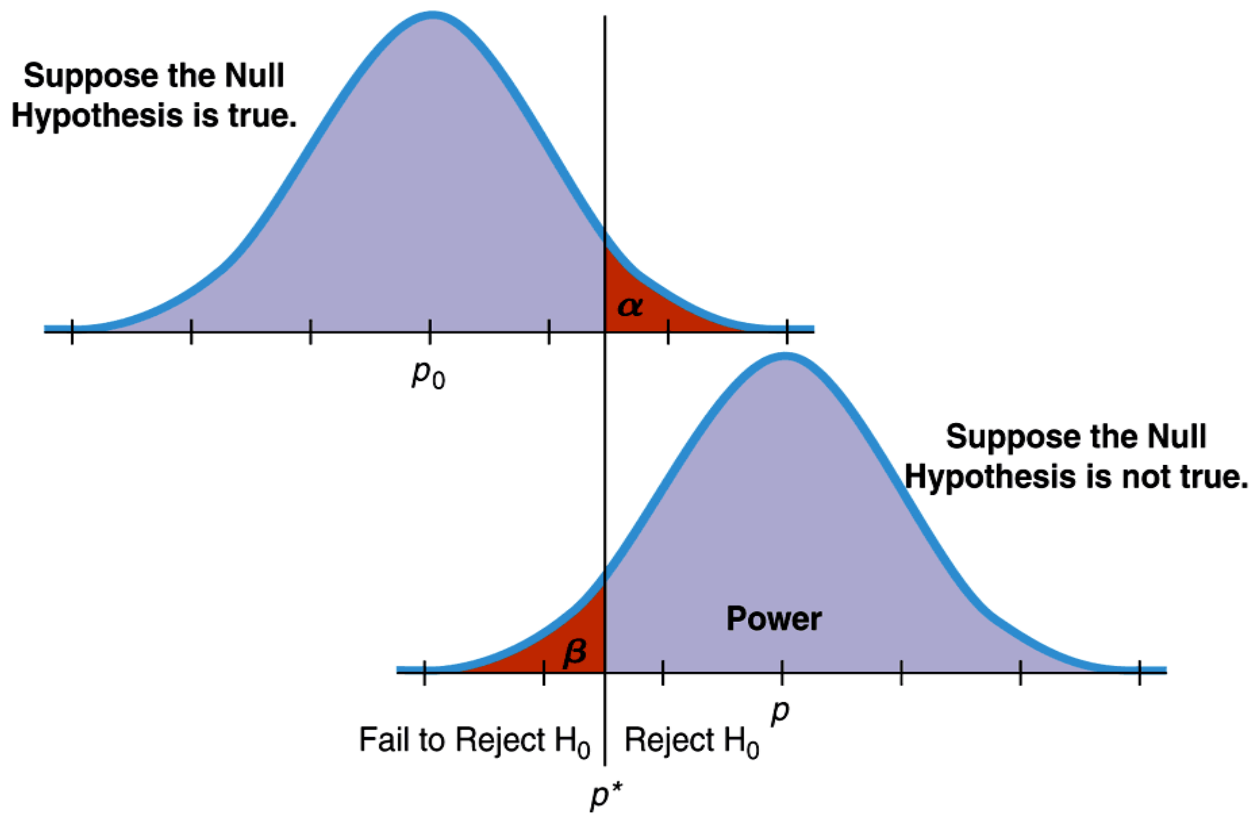


Can We Reduce Both Type I and Type II Errors?

- 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.

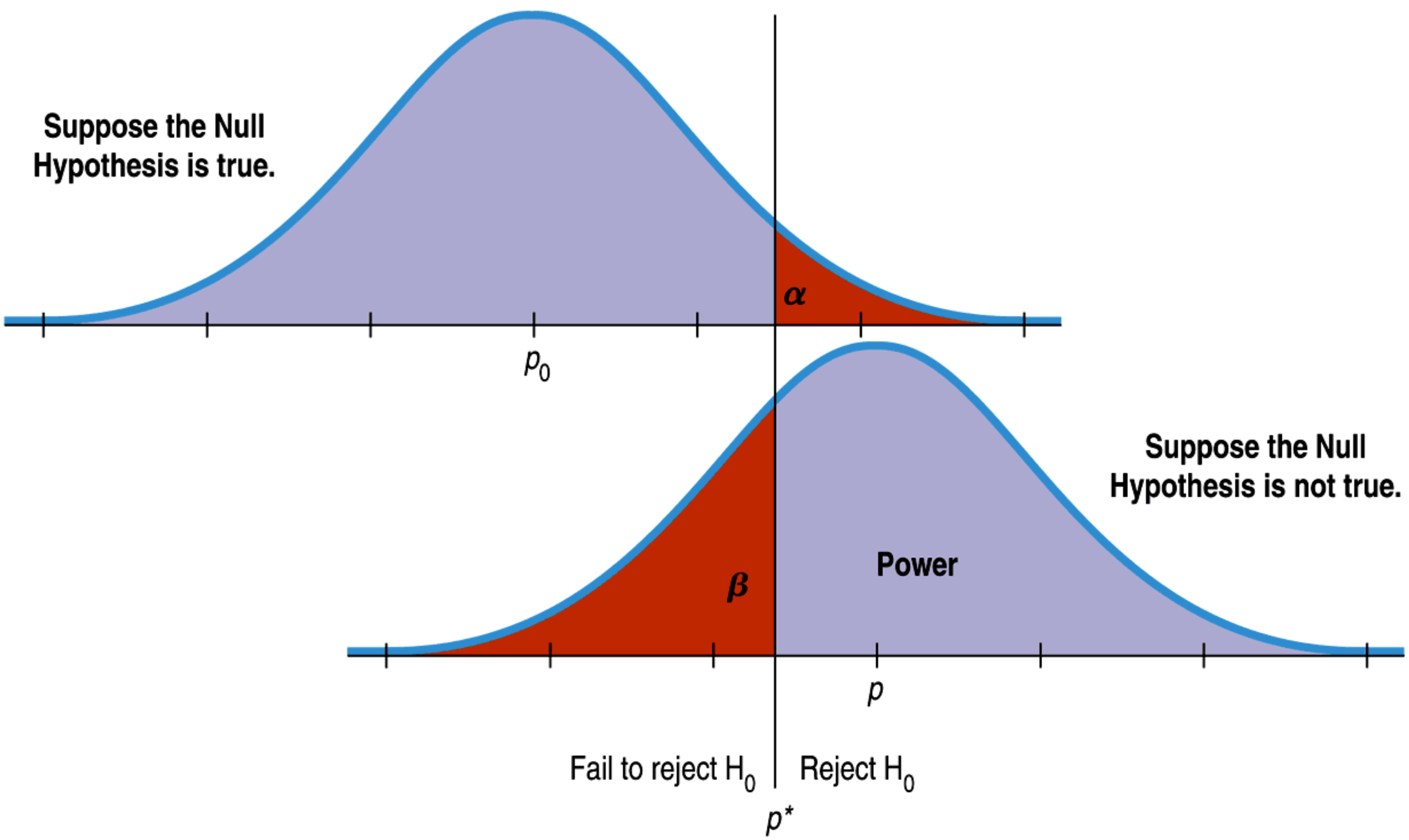
Can We Reduce Both Type I and Type II Errors? (cont.)

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:



Can We Reduce Both Type I and Type II Errors? (cont.)

This diagram shows the relationship between these concepts:



What are Type I Error and Type II Error, again?

Type I and Type II Errors

Type I error: Rejecting the null hypothesis when it is in fact true.

Type II error: Not rejecting the null hypothesis when it is in fact false.

Note: The probability of making type I error, that is, of rejecting a true null hypothesis, is called the significance level of a hypothesis test.

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?

We have learned to:

1. Define and apply Type I Error and Type II Error.
2. Recognize the power of the test is the probability that we reject the null hypothesis when it’s false.
3. Recognize a small P-value is an indicator of evidence against the null hypothesis.
4. Utilize the significance level (alpha level) to establish the level of proof.
5. Identify the connection between hypothesis test and confidence interval.

Credit

Some of these slides have been adapted/modified in part/whole from the slides of the following textbooks.

- Weiss, Neil A., Introductory Statistics, 8th Edition
- Bock, David E., Stats: Data and Models, 3rd Edition