Types of Errors in Hypothesis Testing

statisticsbyjim.com/hypothesis-testing/types-errors-hypothesis-testing

Jim Frost July 9, 2018

Hypothesis tests use sample data to make inferences about the properties of a population. You gain tremendous benefits by working with random samples because it is usually impossible to measure the entire population.

However, there are tradeoffs when you use samples. The samples we use are typically a miniscule percentage of the entire population. Consequently, they occasionally misrepresent the population severely enough to cause hypothesis tests to make errors.

In this blog post, you will learn about the two types of errors in hypothesis testing, their causes, and how to manage them.

Potential Outcomes in Hypothesis Testing

Hypothesis testing is a procedure in inferential statistics that assesses two mutually exclusive theories about the properties of a population. For a generic hypothesis test, the two hypotheses are as follows:

- **Null hypothesis**: There is no effect
- **Alternative hypothesis**: There is an effect.

The sample data must provide sufficient evidence to reject the null hypothesis and conclude that the effect exists in the population. Ideally, a hypothesis test fails to reject the null hypothesis when the effect is not present in the population, and it rejects the null hypothesis when the effect exists.

Statisticians define two types of errors in hypothesis testing. Creatively, they call these errors Type I and Type II errors. Both types of error relate to incorrect conclusions about the null hypothesis.

The table summarizes the four possible outcomes for a hypothesis test.

	Test Rejects Null	Test Fails to Reject Null
Null is True	**Type I Error False Positive**	**Correct decision No effect**
Null is False	**Correct decision Effect exists**	**Type II error False negative**

Related post: How Hypothesis Tests Work: P-values and the Significance Level

Fire alarm analogy for the types of errors

A fire alarm provides a good analogy for the types of hypothesis testing errors. Preferably, the alarm rings when there is a fire and does not ring in the absence of a fire. However, if the alarm rings when there is no fire, it is a false positive, or a Type I error in statistical terms. Conversely, if the fire alarm fails to ring when there is a fire, it is a false negative, or a Type II error.

Using hypothesis tests correctly improves your chances of drawing trustworthy conclusions. However, errors are bound to occur.

Unlike the fire alarm analogy, there is no sure way to determine whether an error occurred after you perform a hypothesis test. Typically, a clearer picture develops over time as other researchers conduct similar studies and an overall pattern of results appears. Seeing how your results fit in with similar studies is a crucial step in assessing your study's findings.

Now, let's take a look at each type of error in more depth.

Type I Errors: False Positives

When you see a p-value that is less than your significance level, you get excited because your results are statistically significant. However, it could be a type I error. The supposed effect might not exist in the population. Again, there is usually no warning when this occurs.

Why do these errors occur? It comes down to sample error. Your random sample has overestimated the effect by chance. It was the luck of the draw. This type of error doesn't indicate that the researchers did anything wrong. The experimental design, data collection, data validity, and statistical analysis can all be correct, and yet this type of error still occurs.

Even though we don't know for sure which studies have false positive results, we *do* know their rate of occurrence. The rate of occurrence for Type I errors equals the significance level of the hypothesis test, which is also known as alpha (α).

The significance level is an evidentiary standard that you set to determine whether your sample data are strong enough to reject the null hypothesis. Hypothesis tests define that standard using the probability of rejecting a null hypothesis that is actually true. You set this value based on your willingness to risk a false positive.

Related post: How to Interpret P-values Correctly

Using the significance level to set the Type I error rate

When the significance level is 0.05 and the null hypothesis is true, there is a 5% chance that the test will reject the null hypothesis incorrectly. If you set alpha to 0.01, there is a 1% of a false positive. If 5% is good, then 1% seems even better, right? As you'll see, there is a

tradeoff between Type I and Type II errors. If you hold everything else constant, as you reduce the chance for a false positive, you increase the opportunity for a false negative.

Type I errors are relatively straightforward. The math is beyond the scope of this article, but statisticians designed hypothesis tests to incorporate everything that affects this error rate so that you can specify it for your studies. As long as your experimental design is sound, you collect valid data, and the data satisfy the assumptions of the hypothesis test, the Type I error rate equals the significance level that you specify. However, if there is a problem in one of those areas, it can affect the false positive rate.

Warning about a potential misinterpretation of Type I errors and the Significance Level

When the null hypothesis is correct for the population, the probability that a test produces a false positive equals the significance level. However, when you look at a statistically significant test result, you cannot state that there is a 5% chance that it represents a false positive.

Why is that the case? Imagine that we perform 100 studies on a population where the null hypothesis is true. If we use a significance level of 0.05, we'd expect that five of the studies will produce statistically significant results—false positives. Afterward, when we go to look at those significant studies, what is the probability that each one is a false positive? Not 5 percent but 100%!

That scenario also illustrates a point that I made earlier. The true picture becomes more evident after repeated experimentation. Given the pattern of results that are predominantly not significant, it is unlikely that an effect exists in the population.

Type II Errors: False Negatives

When you perform a hypothesis test and your p-value is greater than your significance level, your results are not statistically significant. That's disappointing because your sample provides insufficient evidence for concluding that the effect you're studying exists in the population. However, there is a chance that the effect is present in the population even though the test results don't support it. If that's the case, you've just experienced a Type II error, which is also known as beta (β).

What causes Type II errors? Whereas Type I errors are caused by one thing, sample error, there are a host of possible reasons for Type II errors—small effect sizes, small sample sizes, and high data variability. Furthermore, unlike Type I errors, you can't set the Type II error rate for your analysis. Instead, the best that you can do is estimate it before you begin your study by approximating properties of the alternative hypothesis that you're studying. When you do this type of estimation, it's called power analysis.

To estimate the Type II error rate, you create a hypothetical probability distribution that represents the properties of a true alternative hypothesis. However, when you're performing a hypothesis test, you typically don't know which hypothesis is true, much less the specific properties of the distribution for the alternative hypothesis. Consequently, the true Type II error rate is usually unknown!

Type II errors and the power of the analysis

The Type II error rate (beta) is the probability of a false negative. Therefore, the inverse of Type II errors is the probability of correctly detecting an effect. Statisticians refer to this concept as the power of a hypothesis test. Consequently, 1 – β = the statistical power. Analysts typically estimate power rather than beta directly.

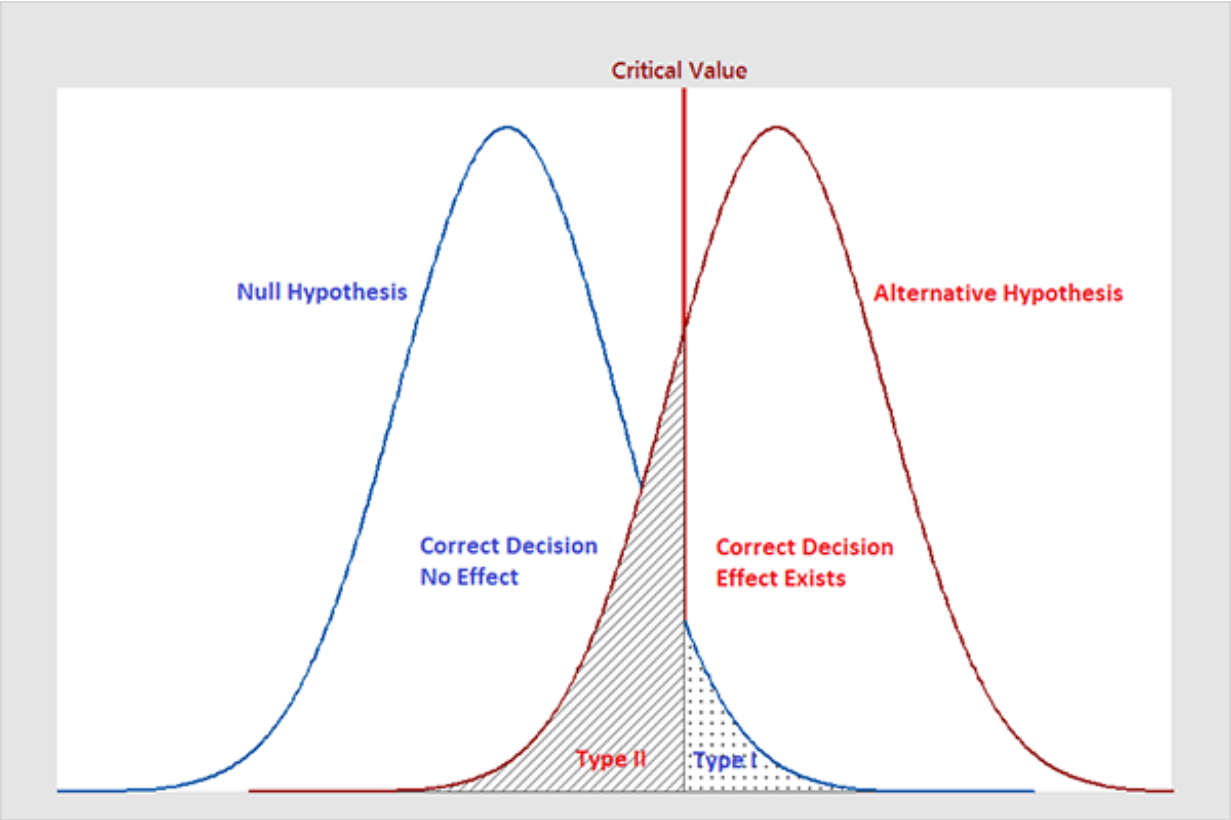
If you read my post about power and sample size analysis, you know that the three factors that affect power are sample size, variability in the population, and the effect size. As you design your experiment, you can enter estimates of these three factors into statistical software and it calculates the estimated power for your test.

Suppose you perform a power analysis for an upcoming study and calculate an estimated power of 90%. For this study, the estimated Type II error rate is 10% (1 – 0.9). Keep in mind that variability and effect size are based on estimates and guesses. Consequently, power and the Type II error rate are just estimates rather than something you set directly. These estimates are only as good as the inputs into your power analysis.

Low variability and larger effect sizes decrease the Type II error rate, which increases the statistical power. However, researchers usually have less control over those aspects of a hypothesis test. Typically, researchers have the greatest control over sample size, which makes it the critical way to manage your Type II error rate. Holding everything else constant, increasing the sample size reduces the Type II error rate and increases power.

Graphing Type I and Type II Errors

The graph below illustrates the two types of errors using two sampling distributions. The critical region line represents the point at which you reject or fail to reject the null hypothesis. Of course, when you perform the hypothesis test, you don't know which hypothesis is correct. And, the properties of the distribution for the alternative hypothesis are usually unknown. However, use this graph to understand the general nature of these errors and how they are related.



The distribution on the left represents the null hypothesis. If the null hypothesis is true, you only need to worry about Type I errors, which is the shaded portion of the null hypothesis distribution. The rest of the null distribution represents the correct decision of failing to reject the null.

On the other hand, if the alternative hypothesis is true, you need to worry about Type II errors. The shaded region on the alternative hypothesis distribution represents the Type II error rate. The rest of the alternative distribution represents the probability of correctly detecting an effect—power.

Moving the critical value line is equivalent to changing the significance level. If you move the line to the left, you're increasing the significance level (e.g., α 0.05 to 0.10). Holding everything else constant, this adjustment increases the Type I error rate while reducing the Type II error rate. Moving the line to the right reduces the significance level (e.g., α 0.05 to 0.01), which decreases the Type I error rate but increases the type II error rate.

Is One Error Worse Than the Other?

As you've seen, the nature of the two types of error, their causes, and the certainty of their rates of occurrence are all very different.

A common question is whether one type of error is worse than the other? Hypothesis tests are designed to be able to control Type I errors while Type II errors are much less defined. Consequently, many statisticians state that it is better to fail to detect an effect when it exists than it is to conclude an effect exists when it doesn't. That is to say, there is tendency to assume that Type I errors are worse.

However, reality is more complex than that. You should carefully consider the consequences of each type of error for your specific test.

Suppose you are assessing the strength of a new jet engine part that is under consideration. Peoples lives are riding on the part's strength. A false negative in this scenario merely means that the part is strong enough but the test fails to detect it. This situation does not put anyone's life at risk. On the other hand, Type I errors are worse in this situation because they indicate the part is strong enough when it is not.

Now suppose that the jet engine part is already in use but there are concerns about it failing. In this case, you want the test to be more sensitive to detecting problems even at the risk of false positives. Type II errors are worse in this scenario because the tests fail to detect the problem and leave these problematic parts in use for longer.

Using hypothesis tests effectively requires that you understand their error rates. By setting the significance level and estimating the power of your test, you can manage both error rates so they meet your requirements.