7.7: The Two Errors in Null Hypothesis Significance Testing

Before going into details about how a statistical test is constructed, it's useful to understand the philosophy behind it. We hinted at it when pointing out the similarity between a null hypothesis test and a criminal trial, but let's be explicit.

Ideally, we would like to construct our test so that we never make any errors. Unfortunately, since the world is messy, this is **never possible**. Sometimes you're just really unlucky: for instance, suppose you flip a coin 10 times in a row and it comes up heads all 10 times. That feels like very strong evidence that the coin is biased, but of course there's a 1 in 1024 chance that this would happen even if the coin was totally fair. In other words, in real life we *always* have to accept that there's a chance that we made the wrong statistical decision. The goal behind statistical hypothesis testing is not to *eliminate* errors, because that's impossible, but to *minimize* them.

At this point, we need to be a bit more precise about what we mean by "errors". Firstly, let's state the obvious: it is either the case that the null hypothesis is true, or it is false. The means are either similar or they ar not. The sample is either from the population, or it is not. Our test will either reject the null hypothesis or retain it. So, as the Table 7.7.1 illustrates, after we run the test and make our choice, one of four things might have happened:

Table 7.7.1- Statistical Decision Versus Reality

Reality Versus Your Sample	Reality: Means are Different (Null Hypothesis is False)	Reality: Means are Similar (Null Hypothesis is True)
Your Sample: Means are Different (Reject Null Hypothesis)	Correct! :)	Error (Type I) :(
Your Sample: Means are Similar (Retain Null Hypothesis)	Error (Type II) :(	Correct! :)

As a consequence there are actually *two* different types of error here. If we reject a null hypothesis that is actually true, then we have made a ***type I error***. On the other hand, if we retain the null hypothesis when it is in fact false, then we have made a ***type II error***. Note that this does not mean that you, as the statistician, made a mistake. It means that, even when all evidence supports a conclusion, just by chance, you might have a wonky sample that shows you something that isn't true.

Errors in Null Hypothesis Significance Testing

Type I Error

- Reject a true null hypothesis.
  - The sample is from the population, but we say that it's not (rejecting the null).
- Saying there is a mean difference when there really isn't one!
- alpha (α, a weird a)
- False positive

Type II Error

- Retain a false null hypothesis.
  - The sample is from a different population, but we say that the means are similar (retaining the null).
- Saying there is not a mean difference when there really is one!
- beta (β, a weird B)
- Missed effect

Why the Two Types of Errors Matter

Null Hypothesis Significance Testing (NHST) is based on the idea that large mean differences would be rare if the sample was from the population. So, if the sample mean is different enough (greater than the critical value) then the effect would be rare enough (< .05) to reject the null hypothesis and conclude that the means are different (the sample is not from the population). However, about ***5%*** of the times when we reject the null hypothesis, saying that the sample is from a different population, because ***we are wrong***. Null Hypothesis Significance Testing is not a "sure thing." Instead, we have a known error rate (5%). Because of

this, replication is emphasized to further support research hypotheses. For research and statistics, “replication” means that we do many experiments to test the same idea. We do this in the hopes that we might get a wonky sample 5% of the time, but if we do enough experiments we will recognize the wonky 5%.

Remember how statistical testing was kind of like a criminal trial? Well,a criminal trial requires that you establish “beyond a reasonable doubt” that the defendant did it. All of the evidentiary rules are (in theory, at least) designed to ensure that there’s (almost) no chance of wrongfully convicting an innocent defendant. The trial is designed to protect the rights of a defendant: as the English jurist William Blackstone famously said, it is “better that ten guilty persons escape than that one innocent suffer.” In other words, a criminal trial doesn’t treat the two types of error in the same way: punishing the innocent is deemed to be much worse than letting the guilty go free. A statistical test is pretty much the same: the single most important design principle of the test is to *control* the probability of a type I error, to keep it below some fixed probability (we use 5%). This probability, which is denoted α, is called the ***significance level*** of the test (or sometimes, the *size* of the test).

Introduction to Power

So, what about the type II error rate? Well, we’d also like to keep those under control too, and we denote this probability by β (beta). However, it’s much more common to refer to the ***power*** of the test, which is the probability with which we reject a null hypothesis when it really is false, which is 1−β. To help keep this straight, here’s the same table again, but with the relevant numbers added:

Table 7.7.2- Statistical Decision Versus Reality with Alpha and Beta

Reality Versus Your Sample	Reality: Means are Different (Null Hypothesis is False)	Reality: Means are Similar (Null Hypothesis is True)
Your Sample: Means are Different (Reject Null Hypothesis)	Correct! :) $1-\beta$ (power of the test)	Error (Type I) :(α (type I error rate)
Your Sample: Means are Similar (Retain Null Hypothesis)	Error (Type II) :(β (type II error rate)	Correct! :) $1-\alpha$ (probability of correct retention)

“powerful” hypothesis test is one that has a small value of β, while still keeping α fixed at some (small) desired level. By convention, scientists usually use 5% (p = .05, α levels of .05) as the marker for Type I errors (although we also use of lower α levels of 01 and .001 when we find something that appears to be really rare). The tests are designed to ensure that the α level is kept small (accidently rejecting a null hypothesis when it is true), but there’s no corresponding guarantee regarding β (accidently retaining the null hypothesis when the null hypothesis is actually false). We’d certainly like the type II error rate to be small, and we try to design tests that keep it small, but this is very much secondary to the overwhelming need to control the type I error rate. As Blackstone might have said if he were a statistician, it is “better to retain 10 false null hypotheses than to reject a single true one”. To add complication, some researchers don't agree with this philosophy, believing that there are situations where it makes sense, and situations where I think it doesn’t. But that’s neither here nor there. It’s how the tests are built.

Can we decrease the chance of Type I Error *and* decrease the chance of Type II Error? Can we make fewer false positives *and* miss fewer real differences?

Unfortunately, no. If we want fewer false positive, then we will miss more real effects. What we can do is increase the power of finding any real differences. We'll talk a little more about Power in terms of statistical analyses next.

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)

This page titled 7.7: The Two Errors in Null Hypothesis Significance Testing is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michelle Oja.