Decision Error

Grinnell College

April 3, 2024

Review

Decision Making

For now, let's not worry about p-values (*we will revist), instead, let's go back to binary thinking since, in actuality, we must ultimately decide between one of two decisions:

1. There is sufficient evidence to reject H_0
2. There is *not* sufficient evidence to reject H_0

Decision Making

Just as our confidence intervals were correct or incorrect, so to may be our decision regarding H_0. In this case, however, there are two distinct ways in which our decision can be incorrect:

1. H_0 is *TRUE* (i.e., there is no effect), yet we reject anyway
2. H_0 is *FALSE* (i.e., there is an effect), yet we fail to reject it

Decision Making

These two types of errors are known as Type I and Type II errors, respectively:

1. H_0 is *TRUE* (i.e., there is no effect), yet we reject anyway
   - Type I error
   - False positive
   - Evidence leads to wrong conclusion
2. H_0 is *FALSE* (i.e., there is an effect), yet we fail to reject it
   - Type II error
   - False negative
   - Not enough evidence to conclude

Decision Making

Test Result	True State of Nature	
	H_0 True	H_0 False
Fail to reject H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

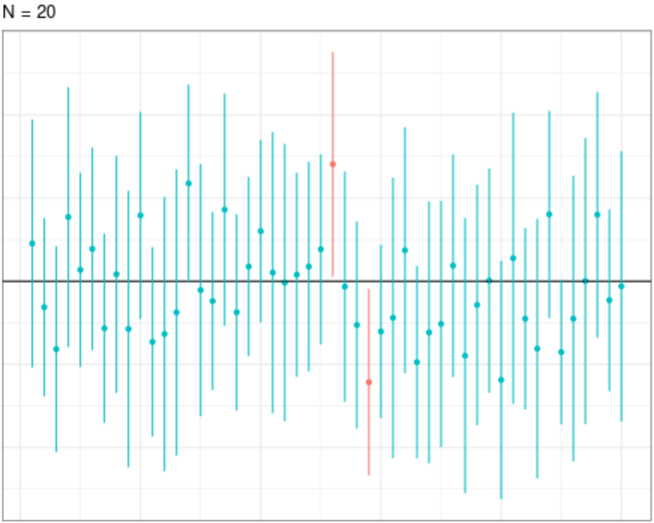
Type I Errors

A Type I error describes a situation in which we incorrectly identify a null effect:

- Conclude that an intervention works when it does not
- Conclude that there is a relationship between two variables when there are not

A Type I error will occur, for example, when our constructed confidence does not contain μ_0 when in actuality it should

Type I Errors



Type I Error Rate

We can control the rate at which we commit Type I errors with adjusting the *level of significance*, denoted α.

This is also called the *Type I error rate*

The Type I error rate has a *one-to-one* correspondence with our confidence intervals: a 95% confidence interval will permit a Type I error 5% of the time, corresponding to $\alpha = 0.05$

Type I Error Rate

Before we begin a study, we specify a threshold of evidence required to reject H_0

For example, we may specify at onset that we want confidence of $1 - \alpha = 0.95$, or, equivalently, a Type I error of $\alpha = 0.05$

So long as our p-value is such that $p < \alpha$, we can be certain in the long run that our Type I error rate is bounded by α

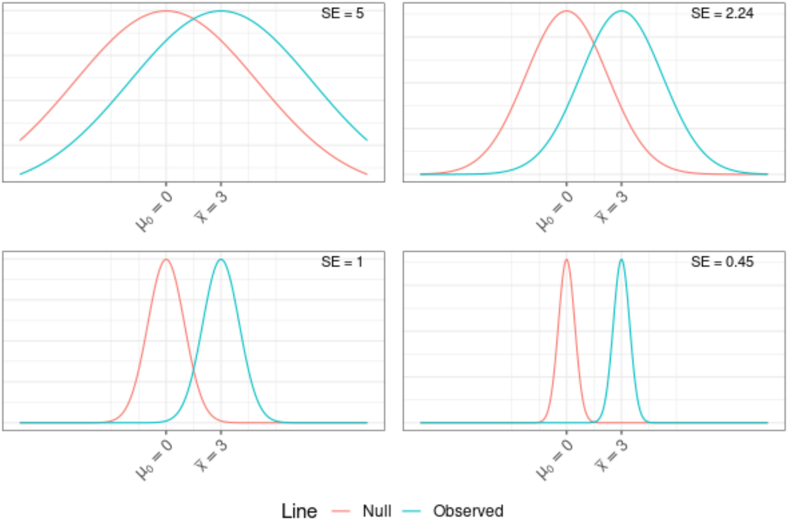
Type II Errors

A Type II error describes a situation in which the null hypothesis is false, yet based on the evidence gathered we fail to reject it:

- An intervention has a clinical effect, but it is not detected
- An email is considered spam, but the filter does not detect it

Typically, a Type II error is the result of one or more factors:

- Too few observations in our sample
- The population has large variability
- The effect size is small


SE = 5
SE = 2.24
SE = 1
SE = 0.45
$\mu_0 = 0$
$\bar{x} = 3$
Line
Null
Observed

Type II Error Rate

The Type II error rate is typically denoted β

More frequently, we consider the rate at which Type II errors do not occur $(1 - \beta)$, a term we refer to as *power*

A study that is unable to detect a true effect is said to be *underpowered*

Power

Consider the following analogy[1]: you send a child into the basement to find an object

- What is the probability that she actually finds it?
- This will depend on three things:
  - How long does she spend looking?
  - How big is the object she is looking for?
  - How messy is the basement?

[1]Stolen from Patrick Breheny who credits the text *Intuitive Biostatistics*, which in turn credits John Hartung for this example

Power

If the child spends a long time looking for a large object in a clean, organized basement, she will most likely find what she's looking for

If a child spend a short amount of time looking for a small object in a messy, chaotic basement, it's probably that she won't find it

Each of these has a statistical analog:

- How long she spends looking? = How big is the sample size?
- How big is the object? = How large is the effect size?
- How messy is the basement? = How noisy/variable is the data?

Drawing Conclusions

As we never truly know whether H_0 is correct or not, we must simultaneously be prepared to combat both types of error

Test Result	True State of Nature	
	H_0 True	H_0 False
Fail to reject H_0	Correct $(1-\alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Correct $(1-\beta)$

- Type I error $= P(\text{Reject } H_0 | H_0 \text{ true}) =$ false alarm
- Type II error $= P(\text{Fail to reject } H_0 | H_A \text{ true}) =$ missed opportunity

Example

Suppose that an investigator sets out to test 200 null hypotheses where exactly half of them are true and half of them are not. Additionally, suppose the tests have a Type I error rate of 5% and a Type II error rate of 20%

1. Out of the 200 hypothesis tests carried out, how many should be expect to be Type I errors?
2. How many would be Type II errors?
3. Of the 200 tests, how many times would the investigator correctly *fail to reject* the null hypothesis?
4. Out of all of the tests in which the null hypothesis was rejected, for what percentage was the null hypothesis actually true?

Base Rate Fallacy – Medical Testing

The previous example hints at the existence of a common error in interpretation known as the *base rate fallacy*.

Imagine that we have a dianostic test for an infectious disease which has a Type I error rate of 5% and a Type II error rate of 1% (99% power). Then consider two scenarios:

- Scenario 1: We use it to test for the disease on population A of 1,000 people where 40% are infected
- Scenario 2: We use it to test for the disease on population B of 1,000 people where 2% are infected

p-values

Although the $\alpha = 0.05$ is customary for Type I error rate and a cut-off for "statistical significance", this is no substitute for correctly evaluating context

For example, a highly publicized study in 2009 involving a vaccine protecting against HIV found that, analyzed one way, the data suggested a p-value of 0.08. Computed a different way, it resulted in a p-value of 0.04

Debate and controversy ensued, primarily because the consequence of using a particular method was the difference between a result being on other side of the $p < \alpha$ threshold

But is there really that much a difference between $p = 0.04$ and $p = 0.08$?

Review

Based on the evidence observed, we will ultimately make one of two decisions:

1. Reject H_0
2. Fail to reject H_0

Depending on the true state of H_0, we can be incorrect in two ways:

1. Type I Error (α): H_0 is true, yet we reject anyway
2. Type II Error (β): H_0 is false, yet we fail to reject it

References

- Patrick Breheny 2022 BIOS 4120 course notes (thank you)