Hypothesis Testing pt. 4

Decision Making

Grinnell College

Strength of Evidence Approach to Testing

Up until this point hypothesis testing has followed this basic process:

1. Begin with a null hypothesis, $H_0 : \mu = \mu_0$
2. Collected data and compute a statistic, i.e, $\overline{x}$
3. Compare our statistic against the null distribution, i.e., $T = \frac{\overline{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$
4. Derive a p-value based on the statistic and the distribution
5. Write a summary talking about 'strength of evidence'

We are going to look at an alternative approach where we change Step 5.

Decision Making

For the remainder of these slides we will talk about this alternative method but I want to point out the following things.

- Many statisticians over the past few years have embraced the 'strength of evidence' approach
- There are many problems that come with the following approach
- However, it is still a relatively common thing you will encounter outside of this class
- I will test you on both methods, and will make clear which one I want you to use for a given problem

Decision Making – Motivation

Based on the evidence we have collected, we must ultimately decide between one of two decisions:

1. There is sufficient evidence to reject H_0 in favor of H_A
   - data seems unlikely if H_0 is true
2. There is *not* sufficient evidence to reject H_0
   - data largely agrees with 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

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

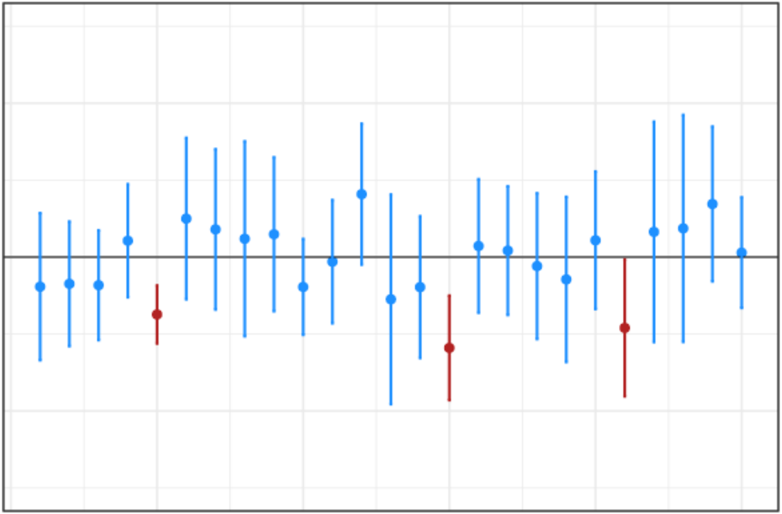
Type I Errors

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

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

A Type I error will occur, for example, when our constructed confidence does not contain μ_0 when $\mu_0 = \mu$ (true mean equals hypothesized mean)

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 $\alpha = 0.05$

We *reject* our null hypothesis when p-value $< \alpha$

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:

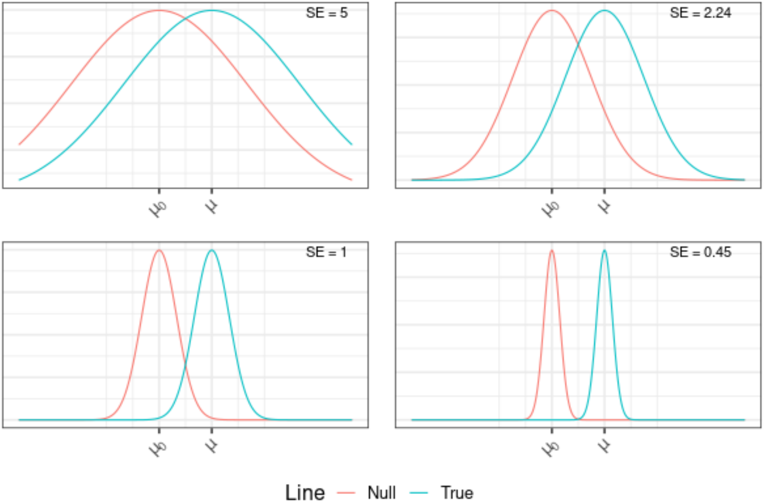
- An intervention has a clinical effect, but it is not detected
- An email is considered spam, but the filter does not detect it

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

Type II Errors

Sample size and population variance affect how easy it is to tell true mean (μ) apart from hypothesized mean (μ_0) $\rightarrow$ affect SE $= \frac{\sigma}{\sqrt{n}}$



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 Professor Nolte, who in turn stole this 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

Issues with Decision Making – Significance Level

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 $p = 0.08$?

- What about .049 and .051?

Issues with Decision Making – Significance Level

There is an unholy obsession with using a Type I error rate of $\alpha = 0.05$ in many disciplines

- the 0.05 value is arbitrary – why 1 in 20?

Back in the 1920s, Ronald A. Fisher (who had a big hand in making many of the methods we are covering in this class) proposed that p-values between 0.01 and 0.05 (or lower) were reasonable and he made many tables that used these cutoffs in his HUGELY famous book *Statistical Methods for Research Workers*

- so many scientists and statisticians followed his results that the 0.05 became incredibly common place
- but Fisher did not intend for everyone to use any one specific cutoff!

Issues with Decision Making – Significance Level

Fisher intended for the signifance level α to be adjusted to the seriousness of getting a wrong conclusion

- does a new medication work better than another
  - may want $\alpha = .01$
- is the percent of red cars that drive through campus more than 10%
  - might be ok with $\alpha = .05$ or even .10

File Drawer Effect

Publishing only results that show a significant finding disturbs the balance of findings in favor of positive results.[2]

File Drawer Effect is when research doesn't get published because the results are not deemed significant (usually p-values > 0.05)

When research is only 'publishable' if p-values are below the $\alpha = .05$ level, it leads to lots of scientific discovery going unreported

- even when we haven't found the effect we wanted to (i.e. there isn't a difference) we may have still learned something valuable!

[2] Song, F.; Parekh, S.; Hooper, L.; Loke, Y. K.; Ryder, J.; Sutton, A. J.; Hing, C.; Kwok, C. S.; Pang, C.; Harvey, I. (2010). "Dissemination and publication of research findings: An updated review of related biases"

P-hacking

P-Hacking: "various techniques that researchers can use to increase the chances of finding statistically significant results in their study, even if the results are not actually meaningful. This is a form of data manipulation that can lead to the publication of false positive results."[3]

- increasing sample size
  - can detect even minute differences with very large sample sizes
  - these small differences may not *really* matter
- throwing out outliers in the data set
- "cherry picking" - doing a whole bunch of tests and choosing the one with the smallest p-value
- post-data hypothesis construction

[3] https://www.physiotutors.com/wiki/p-hacking/

P-hacking

"It's critical to note that P-hacking can happen accidentally and can stem from a researcher's lack of statistical knowledge or the pressure to publish promising results. Yet, it may also be a choice made consciously with a goal in mind. Researchers should pre-register their study design and analytic plan, report all the findings, and apply the proper statistical techniques to account for multiple comparisons in order to prevent p-hacking." [4]

[4] https://www.physiotutors.com/wiki/p-hacking/

Multiple Comparisons

Consider conducting 8 hypothesis tests, each with a Type I error rate 5%

For any given test, the probability of *not* making an error is

$$P(\text{No type I error}) = 0.95$$

What is the probability that I make at least one Type I error?

$$\begin{aligned} P(\text{At least one Type I error}) &= 1 - P(\text{Probability of no Type I errors}) \\ &= 1 - (1 - 0.05)^8 = 1 - (.95)^8 \\ &= 33.6\% \end{aligned}$$

That is, instead of making a Type I error 1 in 20 times, we are now making it 1 in 3 times

Issues in Decision Making

The issues mentioned over the last few slide indicate that we should be a bit skeptical of the 'decision making' approach as I've described it.

The 'strength of evidence' approach is gaining more traction, especially amongst statistics instructors.

- avoids arbitrary significance thresholds
- encourages publication of a wider variety of results
- less pressure for researchers to *manipulate* their data to get published