# Errors

November 2025

# General Catch All Today

Unit 3 starts today!!

Using this class to hit on the small topics you need to see in order to be well rounded...

- Different Error Rates
- Effect Size
- Multiple Comparisons
- p-hacking

# Error

In statistical testing there is two main types of errors. An **error** in hypohtesis test means your decision is incorrect, but not because you messed up.

- Sometimes our samples are just unlucky
  - Happens more often than you'd think...
- Or the difference is really hard to detect
  - Eg finding relativity when a baseball pitcher throws a ball

These errors come in two types...

# Type 1 Error Rate

A **type 1 error rate** is when you claim to have strong evidence against the null hypothesis $H_0$ when $H_0$ is true

- $H_0$ is actually true
- Your p-value was small by chance
- Incorrectly claimed strong evidence against $H_0$

Less of a problem since we moved towards "strength of evidence". Rejecting/failing to reject $H_0$ is where this is most painful.

- Surprisingly we control this error rate
- "Reject at the 5% level", it's that 5%

# Type 1 Error Rate: More

Fun Fact: The p-value is a random variable that is uniformly distributed 0-1 if the null hypothesis is true

- Moderate and stronger evidence is .05 and smaller
- So 1 in 20 of my samples should have moderate (or better) evidence against $H_0$
- Again, we didn't do anything wrong

Unfortunately this is actually the error rate that is less interesting from a researcher stand point.....

# Type 2 Error Rate

A **type 2 error rate** is when you claim to have little to no evidence against $H_0$ even though $H_0$ is false

- $H_0$ is wrong
- p-value is large by chance
- Leads to conclusions where the researcher's idea is wrong

Why is this one the more interesting on to me?

# Type 2 Error Rate: More

Why is this one the more interesting on to me?

Because if we knew probability we could demostrate $H_0$ is wrong then that would inform us if we should run the experiment/study a priori.

- Eg We can run an expensive physics experiment but we'd only have a 12% probability of getting "moderate" evidence
- Eg if we increase our sample size by 15 the probability we detect a difference with moderate evidence increases to 85%

Why don't we know this one?

# Type 2 Error Rate: More

Why is this one the more interesting on to me?

Because if we knew probability we could demostrate $H_0$ is wrong then that would inform us if we should run the experiment/study a priori.

- Eg We can run an expensive physics experiment but we'd only have a 12% probability of getting "moderate" evidence
- Eg if we increase our sample size by 15 the probability we detect a difference with moderate evidence increases to 85%

Why don't we know this one?

Because we'd have to know where the true population parameter is at which we never truly do.

# Example Graph

If $H_0$ is true (and the assumptions are met) we know what the sampling distribution will look like and where it's centered. We can also find where a 5% p-value would be.

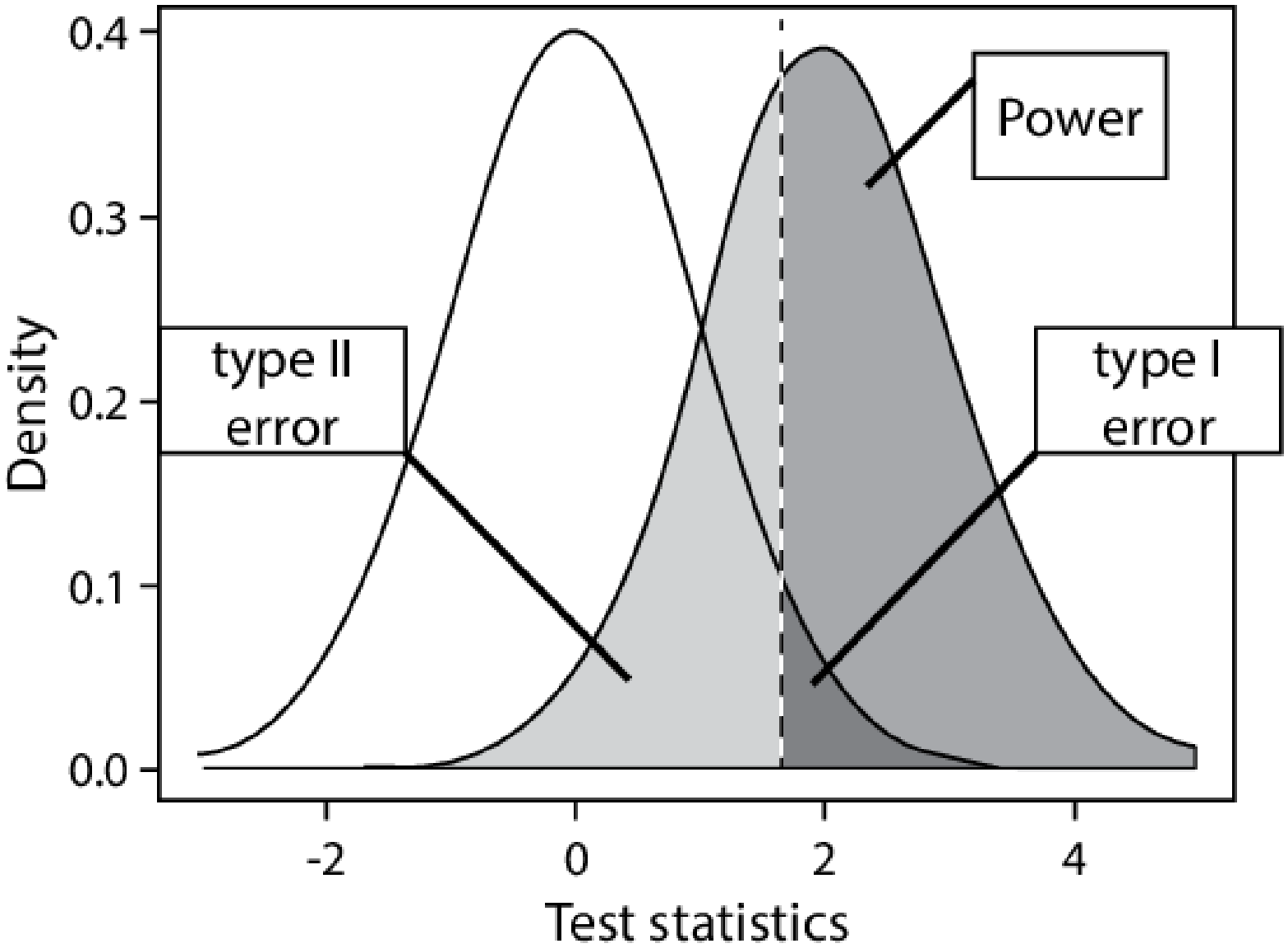
We have complete control

---

If $H_0$ is wrong then we don't now where the real sampling distribution is located...A large chunk could be in the "little to no evidence" range our of sampling distribution.

We have no control

# Example Graph



# Question

What does a really small p-value tell us about our hypothesis?

# Question

What does a really small p-value tell us about our hypothesis?

Not that much honestly, basically just that we can detect a statistical difference (eg home team winning rate is higher than 50%)

# Question

Does a really small p-value tell us the difference is meaningful?

# Question

Does a really small p-value tell us the difference is meaningful?

Note really....

# Large sample size.....

```
> holding <- rnorm(10000000,
+                  mean = .001,
+                  sd = .5)
> t.test(holding)

        One Sample t-test

data:  holding
t = 5.071, df = 1e+07, p-value = 3.957e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.0004919728 0.0011118618
sample estimates:
   mean of x
0.0008019173
```

# Effect Size

Instead of trying to figure out if a change is meaningful based on a p-value alone it's useful to look at the effect size. The **effect size** is the estimated effect/difference that occurs between two groups

- Big Effect Size: A farmer's field doubles his yield from 1000 bushels to 2000 bushels
- Small Effect Size: A farmer's field increases his yield from 1000 bushels to 1002 bushels
- Having a small p-value on the second bullet point won't make a difference in reality

# Effect Size Example

Cardiovascular health is an issue in this country so let's assume we have two drugs, A and B

- Drug A's p-value was .002
- Drug B's p-value was .03

Which would your prefer?

# Effect Size Example

Cardiovascular health is an issue in this country so let's assume we have two drugs, A and B

- Drug A...
  - p-value is .002
  - sample size = 3000
- Drug B...
  - p-value is .03
  - sample size = 200

Which would your prefer?

# Effect Size Example

Cardiovascular health is an issue in this country so let's assume we have two drugs, A and B

- Drug A...
  - p-value is .002
  - sample size = 3000
  - Reduction in heart attacks: .03
- Drug B...
  - p-value is .03
  - sample size = 200
  - Reduction in heart attacks: .07

Which would your prefer?

# Effect Size Example

There is no solution to the last slide.

Drug A

- has a small reduction in heart attack rates we believe
- But we are fairly confident there *is* a reduction
- And we have a large sample size so unexpected things should be happening hopefully

Drug B

- larger reduction in heart attacks we think
- Less confident that the reduction isn't 0
- Smaller sample size means random fluctuations might still be happening

# Effect Size Example

There is no solution to the last slide.

---

Drug A

- has a small reduction in heart attack rates we believe
- But we are fairly confident there *is* a reduction
- And we have a large sample size so unexpected things should be happening hopefully

---

Drug B

- larger reduction in heart attacks we think
- Less confident that the reduction isn't 0
- Smaller sample size means random fluctuations might still be happening

---

The best thing we can do is be honest and report what we saw.

# Multiple Comparisons

Let's work under the assumption that we can only get published if we have a p-value below .05 for the next few slides

There are many, many reasons why this isn't good but how the world is vs how it should be are two different things

# Multiple Comparisons

Earlier I said the p-value was uniformly distributed from 0-1.

What is the probability that a randomly grabbed sample will produce a p-value smaller than 5%, assuming $H_0$ is true?

# Multiple Comparisons

Earlier I said the p-value was uniformly distributed from 0-1.

What is the probability that a randomly grabbed sample will produce a p-value smaller than 5%, assuming $H_0$ is true?

5%? Yes!

# Multiple Comparisons

Earlier I said the p-value was uniformly distributed from 0-1.

What is the probability that either of two randomly grabbed samples will produce a p-value smaller than 5%, assuming $H_0$ is true?

# Multiple Comparisons

Earlier I said the p-value was uniformly distributed from 0-1.

What is the probability that either of two randomly grabbed samples will produce a p-value smaller than 5%, assuming $H_0$ is true?

We publish is either is below 5% or if both are below 5%. So to not publish both must be above 5% so we get...

1- $.95^2 = 0.0975$

The probability I'd get statistical sig. at 5% when $H_0$ is true and there is two tests is a little less than 10%...

# Multiple Comparisons

Earlier I said the p-value was uniformly distributed from 0-1.

What is the probability that any of 50 randomly grabbed samples/tests will produce a p-value smaller than 5%, assuming $H_0$ is true?

We publish is either is below 5% or if both are below 5%. So to not publish both must be above 5% so we get...

$1 - .95^50 = 0.9231$

The probability I'd get statistical sig. at 5% when $H_0$ is true and there is 50 tests is more than 90%

# Multiple Comparisons

Can someone say what the problem is?

# Multiple Comparisons

Can someone say what the problem is?

It's easier to detect a non-existent difference if we do many (multiple) tests (comparisons).

# Multiple Comparisons

Can someone say what the problem is?

It's easier to detect a non-existent difference if we do many (multiple) tests (comparisons).

What can we do to fix this?

# Multiple Comparisons

What can we do to fix this? Lot actually...

- Bonferroni's Correction
  - We scale (mulitple) our cut off (here 5%) by the number of tests we are running
  - The sum of the probability of committing a type 1 error rate across all tests is 5%
  - Can be verrrry conservative
- Sidak's Correction
  - Relability theory solution
  - The probability that no test is sig. is held at 95%
- Tukey's HSD and Fisher's LSD
  - Tweak's to t-tests to better control the p-value
- False Discovery Rate
  - Major breakthrough in statistical testing came out of this recently (Emmanual Candes + student's Knockoff method)
  - Tries to limit how many type 1 errors you do instead of controlling for a single type 1 error

# Multiple Comparisons

We can also do nothing....

# Multiple Comparisons

We can also do nothing....

Nothing?!?! Won't that make it super easy to commit a type 1 error rate??

# p-hacking

We can also do nothing....

Nothing?!?! Won't that make it super easy to commit a type 1 error rate??

Yes, it will and that will let us publish more easily. Driving p-values smaller than they should be (in a reasonable analysis) is called **p-hacking** and p-hacking is wrong m'kay

# p-hacking: Publish or Perish

Turns out basing scientific research on an arbitrary statistic leads to issues...in particular people just want small p-values.

"The p-value is small so that means we passed the assumptions, right?"
-One of the Worst Stat Consulting Clients I had

**The quality/appropriateness of a model and the p-values are not related!!**

# p-hacking: How to Avoid

p-hacking is most often a product of ignorance or temptation

1. Check your assumptions before you look at your p-values
   - If you fail your assumptions with a small p-value you won't be tempted
2. *Kempthorne Principle*: a statistical analysis should reflect the physical realities of the experiment
3. Be steadfast in that non-sig. results doesn't mean non-interesting
4. If you are running dozens of tests...it's probably by accident

# Conclusion

In conclusion....

- Don't let small p-values be your goal
- Any analysis must be judged by
  - it's appropriateness
  - it's meeting the assumptions
  - it's ability to answer the question of interest
- Effect size, sample size, p-values together holistically give insight