Principles of hypothesis testing

- This is one of two big ideas in inferential statistics, the other one being Estimating parameters of a population.

- There is a lot of discussion / debate on how to interpret p-value. The technique was initially created by Fisher and Neyman who had different views on how the null hypothesis should work - so while there is an orthodox way to view it, the finer details of its interpretation may vary from text-to-text.

- Research hypothesis: a substantive and testable scientific claim. You should be able to measure 1 number (or calculate 1 number based on a combination of measurements) and use that number as a proxy to say whether or not the hypothesis describes the state of the world.

- Most of the fundamental questions in research are ontological questions (e.g. what is X? what is dark matter? what is intelligence?). However, when it comes to doing experiments, it is far easier to measure the relationships between two variables (i.e. is there a relationship between X and Y?)

- **research hypotheses**: the broad messy question that you're trying to answer (e.g. ESP exists)

- **Statistical hypothesis**: a mathematical claim about the characteristics of a data generating mechanism (e.g. $\theta \neq 0.5$ where θ is the probability that a participant will guess the correct card / color / side of coin / whatever)

The Null and alternative hypothesis

- When evaluating a statistical hypothesis, there are actually 2 hypotheses at work:

Null hypothesis

- this corresponds to the exact opposite of what you want to believe (ex: H_0 = ESP does NOT exist)

  - This null hypothesis makes an assumption about what you would measure in the dataset (e.g. $\theta = 0.5$)
  - Start the experiment by assuming that this is the case

Alternative hypothesis

- This is the thing that you want to believe about the world (ex: H_A = ESP DOES exist)
  - this also makes an assumption about what you would measure from the dataset (e.g. $\theta \neq 0.5$)

THE GOAL OF A HYPOTHESIS TEST IS TO SHOW THAT THE NULL HYPOTHESIS IS PROBABLY FALSE

The null hypothesis is deemed to be true unless we can prove beyond a reasonable doubt that it's false.

Errors in hypothesis testing

	retain H_0	reject H_0
H_0 is true	correct decision	type I error
H_0 is false	type II error	correct decision

- type 1: reject a null hypothesis that is true
- type 2: fail to reject a null hypothesis that is false

How to remember type I vs type II error:

If he's going to jail but shouldn't be
you've committed an error
of the 1st degree (Type I)

If you set him free
when he's actually guilty
your evidence was too few
and you've committed a type II

source: https://www.dataday.life/blog/statistics/mnemonic-for-remembering-type-i-type-ii-errors/

- α = probability of type I error (aka **significance level**)
- β = probability of type II error.
  - **power** = $1 - \beta$ = probability that we'll that we reject the null when it's really false

	retain H_0	reject H_0
H_0 is true	$1 - \alpha$ (probability of correct retention)	α (type I error)
H_0 is false	β (type II error)	$1 - \beta$ (correct decision)

- tests are designed to ensure that α is kept small, but there's no guarentee for β

Test statistics and sampling distributions

- In principle- we have a quantity X that we calculate by looking at the data. Given some value of X, we make a decision about whether we believe that the null hypothesis is correct, or to reject the null hypothesis. This thing that we calculate is called the **Test statistic**
- now that we have the test statistic, we need to determine the **sampling distribution of the test statistic** - sometimes this is very nontrivial
  - This distribution will tell us what values of X our null hypothesis would lead us to expect

Critical regions and critical values

- **critical region** - values of X that would lead us to reject the null
  - X should be far from the expected value of X for the null hypothesis
  - if the null hypothesis is True, then X came from that sampling distrubution
  - if the null hypothesis is False, then X did NOT come from that sampling distrubution
  - if $\alpha = 0.05$, then the critical region must cover 5% of the sampling distrubution
- **critical values** - X-values that represent the boundaries of the critical region

What about type II error?

- A type II error (β) occurs when we incorrectly accept the null hypothesis
- When measuring type II error, we actually measure $1 - \beta$, which we call the **power** of a test
- β depends on the true value of the test statistic
- Therefore, people will calculate the power for every possible value of the test statistic and generate a **power function**

Effect size

- A way to measure how similar the true state of the world is to the null hypothesis.
- take the ESP experiment as an example:

    - $H_0 : \theta = 0.5$
    - if we reject the null with $\hat{\theta} = 0.8$, that is way more impressive than rejecting the null with $\hat{\theta} = 0.55$.
- If we want to increase the power of a test, finding ways to increase effect size is a good start
- normally, people don't worry about calculating this (very difficult to know or even define)

steps for doing a hypothesis test

- choose an α level
- come up with a test statistic that compares H_0 to H_a
- figure out the sampling distribution of the test statistic assuming the null hypothesis is true
- calculate the critical region that produces an approprate α level

p-values

Neyman's interpretation

-

References

- Learning statistics with R chapter 11
-