Day 10

Outline

1. “Null Hypothesis Significance Testing”
2. When Null Hypothesis Significance Testing goes horribly wrong

Null Hypothesis Significance Testing

Recall

- identifying a parameter is not "too hard"
- identify its value under H_0 is trivial
- However, identifying its value under H_a is difficult in practice
- Under N-P (Neyman-Pearson): define minimum effect size
- But often, we have no idea
- This is what we have been doing for the past 70 years and it does not require any subject knowledge. "It just works". Will also allow us to get the `P` value.

NHST: just give the inequality in alternative hypothesis H_a

Suppose H_0: $\theta = \theta_0$

- $\theta \rightarrow$ arbitrary parameter
- $\theta_0 \rightarrow$ its value under H_0

N-P:

- H_1: $\theta = \theta_1$
- $\theta_1 \rightarrow$ its value under H_1

NHST: Choose from

- H_a: $\theta > \theta_0$
  - Theory says θ should be bigger
  - One-tailed, one-sided hypothesis testing
    - One-tailed testing: The critical area of a distribution is either $<$ or $>$ a certain value but not both
- H_a: $\theta < \theta_0$
  - Theory says θ should be smaller
  - One-tailed, one-sided hypothesis testing
- H_a: $\theta \neq \theta_0$
  - No idea what to expect or theory suggests arguments for both $>$ **AND** $<$
  - Two-tailed, two-sided hypothesis testing
    - Two-tailed the sample is greater than or less than a certain range of values:

NOTE: method of collecting data tells us what θ is (what θ_0 is) and may suggest H_a

When in doubt: use the H_a version with $\neq$

Next Step : Distribution

Define a test statistic whose value will be computed from sample data

N-P: Find its distribution under both H_0 & H_1

NHST: Find its distribution under H_0 but we don't know its distribution under H_a

Next Step : Critical Region

Define a critical region of the test statistic such that if the observed value is in critical region, accept H_1

NHST: Define a critical region such that if in critical region, reject H_0.

If not in critical region, fail to reject H_0

Example (Theory) : Jury

Start off assuming innocence (H_0) [Null Hypothesis]

- Prosecution presents evidence (test statistic observed value)
- Jury decides if it enough evidence
  - Enough evidence (in critical region) $\rightarrow$ reject the assumption of innocence and declare guilty (reject H_0 and acccept H_a)
  - Not enough evidence (not in critical region) $\rightarrow$ fail to reject presumptions of innocence. He might still be guilty but the evidence is not damming enough to convince us otherwise. We fail to reject H_0 or the Null Hypothesis

In NHST, define significance level (not α but works like α). Here, the significance level is the probability of rejecting the null hypothesis which is generally the same value of α

This can go horribly wrong because power is not taken into account also, there is not a big enough sample size to correctly draw a conclusion.

P-Value

A measure of the "strength" of the evidence against H_0. This is related to power?

ALWAYS COMPUTED AFTER OBSERVATION

Official definition: probability of obtaining our observed value of the test statistic, or a value as or more favorable to H_a, if H_0 is true.

- $P(X \geq x_{observed} \mid H_0 \text{ is true})$ when H_a: $\theta > \theta_0$
- $P(X \leq x_{observed} \mid H_a : \theta < \theta_0)$

Things go weird for two-tailed tests

Usually: P(X is equally or less likely than $x_{observed} \mid H_0$ is true) but sometimes we get one-tailed p-values

"How likely is it that I got this lucky or luckier?"

When P-Value $\leq$ significance level

1. H_0 is true and I got really lucky $\leftarrow$ I will make a Type I Error
   - H_0 may in fact be false but you have circumstantial evidence to promote the notion that H_0 is true
2. H_0 is not true and H_a
   - This is the ideal situation as we do not make any false assumptions about H_0

Either way, I reject H_0 and conclude H_a is true

When P-Value $>$ significance level

1. H_0 is true
2. H_0 is not true, but we don't have "unlikely enough" evidence
   - In this case, we acquit an guilty man. Even though we damn well know he did it but the prosecution did not provide damning evidence to conclude that he is guilty.

Either way, we fail to reject H_0 (make no conclusion so default to assumption H_0 is true)

Example (Book Exercise 8.20 [Application])

Study of children, program intended to increase consumption of whole grains. At end of program, sample of 86 children got a snack.

- 48 children chose whole grain
- 38 chose regular

Suppose that before program, children were equally likely to pick either snack. Do we have enough evidence to claim the program works as intended?

Step 1 (Identify Parameter of Interest)

Use it to write H_0 and H_a.

- `p` = Proportion of all children who choose whole grain (generalized results)
- H_0 : `p = 0.5`
- H_a : `p > 0.5`

Step 2 (Identify Test Statistic and its sampling distribution under H_0)

Let `X` = number of success (number of children in sample choosing whole grain)

$X \sim B(n = 86\,, p = 0.5)$

Step 3 (Observe data and calculate value of test statistic)

$X_{\text{observed}} = 48$

Step 4 (Calculate EITHER the critical region or the P-Value)

P-Value is way easier to compute when you have software

```
binom.test(x = 48, n = 86, p = 0.5, alternative = "g")
```

From software: p-value = 0.166

Step 5 (Determine whether or not to reject H_0)

5% is our cut-off. If the value is LESS than 5% then we **reject the Null Hypothesis.**

Since $0.166 > 0.05$, our results are likely enough under H_0 therefore, we fail to reject H_0

Step 6 (Write what "reject H_0" or "fail to reject H_0" means in context)

We do not have "statistically signifiant" evidence to claim that the program is working. It is reasonable to continue with the assumption that children are still equally likely to pick a healthy snack. We failed to reject the notion that they will be more likely to pick a healthy snack.

When Null Hypothesis Significance Testing goes horribly wrong

1. Very small samples
2. Very large samples
   - Neyman-Pearson: $p = 0.5$ vs $p = 0.50001$
   - NHST: $p = 0.5$ vs $p > 0.5$
3. Significance level is not arbiter of importance (2 and 3 are practically the same)
4. Lots of tests
5. P-hacking