Week 5: Introduction to Hypothesis Testing

phase 2

Introduction

- We have previously seen how to estimate parameters from our data.

- The next step is to evaluate how likely such estimations are given our assumed world (our model).

- This is the basis of null **hypothesis testing**.

Null hypotheses and alternative hypotheses

- The statement being tested is called the **Null Hypothesis,** H_0. Usually H_0 is a statement of *"There is no effect"*.

Example: *Drug X did not cause, on average, a physiological response, i.e.*

$<\textit{Response}>_{\textit{before}} = <\textit{Response}>_{\textit{after}}$

- The **Alternative Hypothesis**, H_a, is every statement that is true instead of the null hypothesis. H_a usually *"There is some effect"* statement.

Example: *Drug X did cause, on average, a physiological response, i.e.*

$<\textit{Response}>_{\textit{before}} \neq <\textit{Response}>_{\textit{after}}$

Null hypotheses and alternative hypotheses

➢ The statement being tested is called the **Null Hypothesis,** H_0. Usually H_0 is a statement of *"There is no effect"*.

Example: Drug X did not increase, on average, the physiological response, i.e.

$<\mathit{Response}>_{before} <= <\mathit{Response}>_{after}$

No effect depends on what you are testing!

➢ The **Alternative Hypothesis**, H_a, is every statement that is true instead of the null hypothesis. H_a usually *"There is some effect"* statement.

Example: Drug X did increase, on average, the physiological response, i.e.

$<\mathit{Response}>_{before} > <\mathit{Response}>_{after}$

Hypothesis testing is like a jury trial

H_0 : Defendant is innocent.

H_a : Defendant is guilty.

Steps in the trial (Hypothesis Testing):

1. Collect the data.
2. Present the evidence.
3. Make a judgment.

Evidence: "How plausibly could we see these data by chance?"
(Null Hypothesis is true)

Judgment: How unlikely is unlikely enough?

Test statistics

- A test is based on a statistic that estimates the parameter that appears in the hypotheses.
- When H_0 is true, we expect the estimate to take a value near the parameter value specified by H_0. This is the **hypothesized value**.
- Values of the estimate far from the hypothesized value give evidence **against** H_0 . The alternative hypothesis determines which **directions** count against H_0.

Test statistics

- A test is based on a statistic that estimates the parameter that appears in the hypotheses.
- When H_0 is true, we expect the estimate to take a value near the parameter value specified by H_0. This is the **hypothesized value**.
- Values of the estimate far from the hypothesized value give evidence **against** H_0 . The alternative hypothesis determines which **directions** count against H_0.

Caution!!!!

This only talks about the probability of the observed data, not the probability of the null hypothesis…

Hypothesis Testing procedure

1. We start with a **null hypothesis** that represents the status quo.
2. We also have an **alternative hypothesis** H_a that represents our research question, i.e. what we're testing for.
3. We conduct a hypothesis test under the **assumption** that the **null** hypothesis is **true**.
4. If the test results suggest that the data do **not** provide **convincing evidence** for the alternative hypothesis, we **stick** with the **null** hypothesis. If they do, then we **reject** the **null** hypothesis **in favor** of the **alternative**.
5. The burden of proof is on the alternative hypothesis.

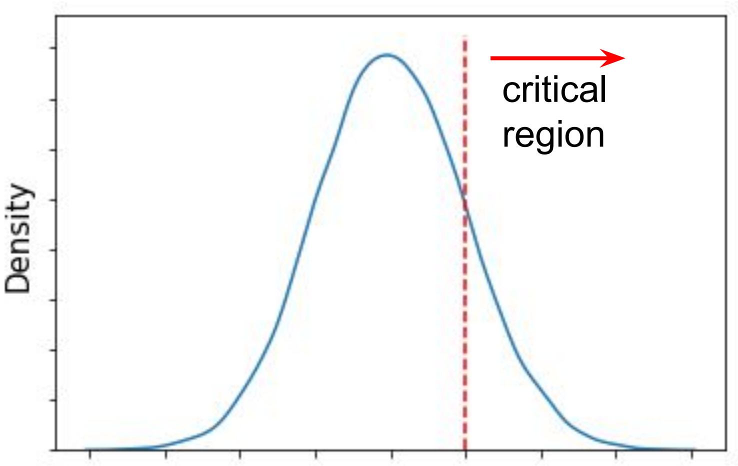
Hypothesis Testing procedure

How much do we need, and how do we quantify this??

4. If the test results suggest that the data do **not** provide **convincing evidence** for the alternative hypothesis, we **stick** with the **null** hypothesis. If they do, then we **reject** the **null** hypothesis **in favor** of the **alternative**.

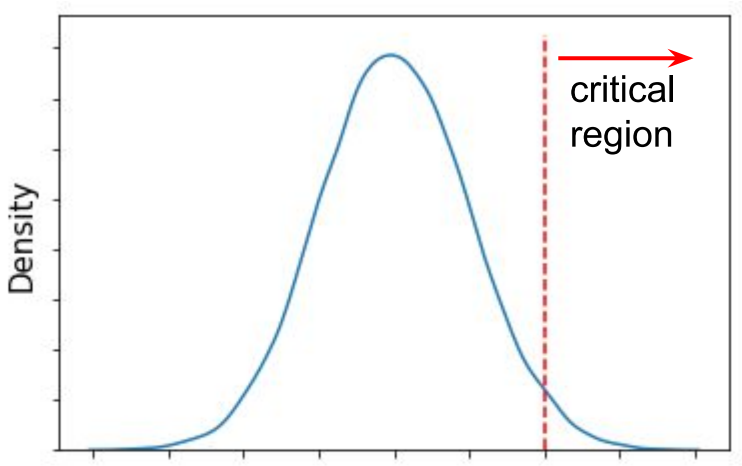
Making decisions

➢ We need to define a **critical region** of values in our test for which we are going to **reject** null hypothesis.



Making decisions

➢ We need to define a **critical region** of values in our test for which we are going to **reject** null hypothesis.



Making decisions

- We need to define a **critical region** of values in our test for which we are going to **reject** null hypothesis.

Making decisions

- We need to define a **critical region** of values in our test for which we are going to **reject** null hypothesis.

- As we have seen, the critical region consists of the most extreme values → the tail of the distribution.

Making decisions

- We need to define a **critical region** of values in our test for which we are going to **reject** null hypothesis

- As we have seen, the critical region consists of the most extreme values → the tail of the distribution.

- Everything within the critical region is considered unlikely under the null hypothesis and therefore we are sufficiently confident to reject it →
→ We claim to observe a **significant** result!

We might make mistakes!!!

Any time that we make a decision about whether to trust the null hypothesis or not, we are subject to committing errors!

- Type I error: If we reject H_0 (accept H_a) when in fact H_0 is true.
- Type II error: if we accept H_0 (reject H_a) when in fact H_a is true.

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

We might make mistakes!!!

Any time that we make a decision about whether to trust the null hypothesis or not, we are subject to committing errors!

- <u>Type I error</u>: If we reject H_0 (accept H_a) when in fact H_0 is true.
- <u>Type II error</u>: if we accept H_0 (reject H_a) when in fact H_a is true.

unbiasedresearch.blogspot.com

We might make mistakes!!!

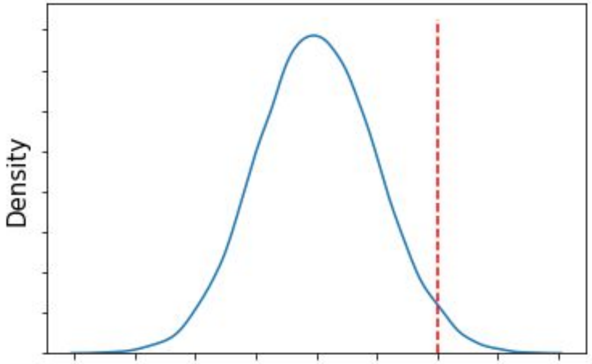
- Any time that we make a decision about whether to trust the null hypothesis or not, we are subject to committing errors!

- **Hypothesis testing** particularly focuses on keeping **Type I errors** rate as **low** as possible, but low **Type II errors are also important** (We'll see this later).

Quantifying evidence: p-values

The p-value is the **probability** that the test statistic would take a value as extreme or more extreme than that actually observed assuming H_0 is **true**.

Distribution under the null
(e.g. here with a particular mean)



Quantifying evidence: p-values

The p-value is the **probability** that the test statistic would take a value as extreme or more extreme than that actually observed assuming H_0 is **true**.

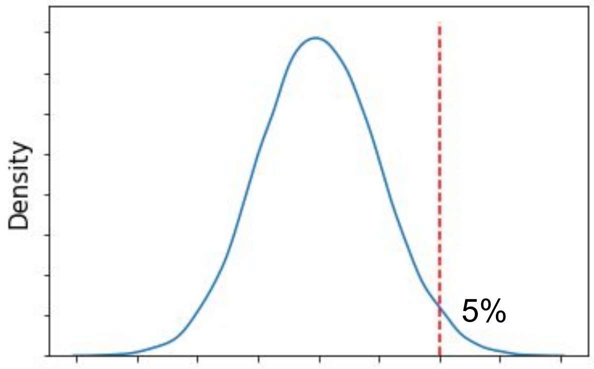
Two possible interpretations:

1. The smaller the p-value, the stronger the evidence AGAINST H_0 provided by the data. (FISHER's)

2. p-values represent the smallest Type I error rate (α) that one is willing to tolerate if you want to reject the null hypothesis. (NEYMAN's)

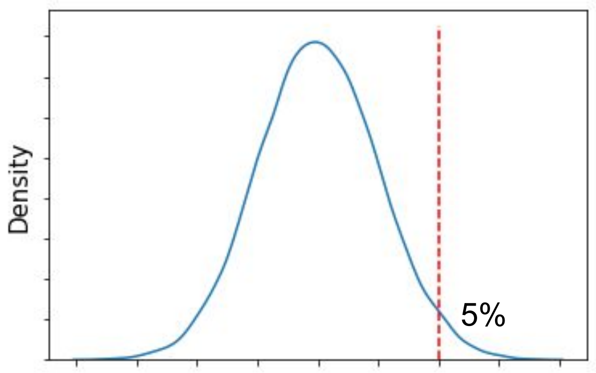
Quantifying evidence: rejecting the null

- In general, any result whose p-value is **below** the threshold α=**0.05** is claimed to be **significant.**



Quantifying evidence: rejecting the null

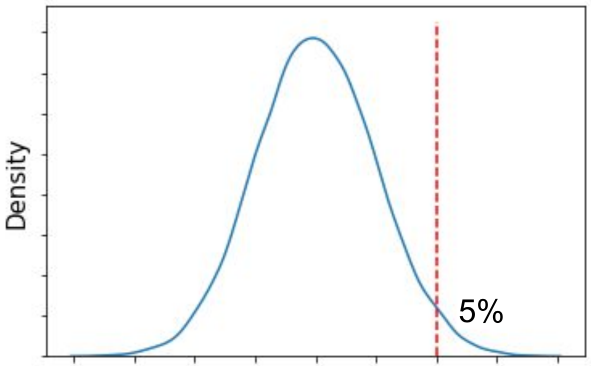
➢ In general, any result whose p-value is **below** the threshold α=**0.05** is claimed to be **significant.**



We believe it is sufficiently small to be an evidence against the null (Fisher’s), or it’s the maximum type I error that we are decided to commit if the null was true (Neyman’s).

Quantifying evidence: rejecting the null

➢ In general, any result whose p-value is **below** the threshold α=**0.05** is claimed to be **significant.**



0.05 is just a **SUBJECTIVE** threshold; it does not have any special meaning!!!!

We believe it is sufficiently small to be an evidence against the null (Fisher's), or it's the maximum type I error that we are decided to commit if the null was true (Neyman's).

Quantifying evidence: p-values

Let’s discuss: Would you say that a result $p=0.5$ means that the null hypothesis is ten times more likely than $p=0.05$?

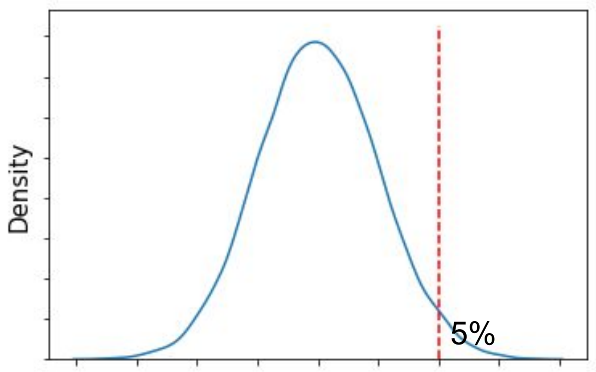
Quantifying evidence: p-values

<u>Let's discuss:</u> Would you say that a result p=0.5 means that the null hypothesis is ten times more likely than p=0.05?

- **NOOOOOOOOOO** and **NEVER** interpret p-values this way!
- P-values are just a probability about the observed data under an assumed model; it does not quantify the probability of a hypothesis!

Quantifying evidence: p-values

➢ As we said earlier, the critical region represents the tail area of a distribution.



Quantifying evidence: p-values

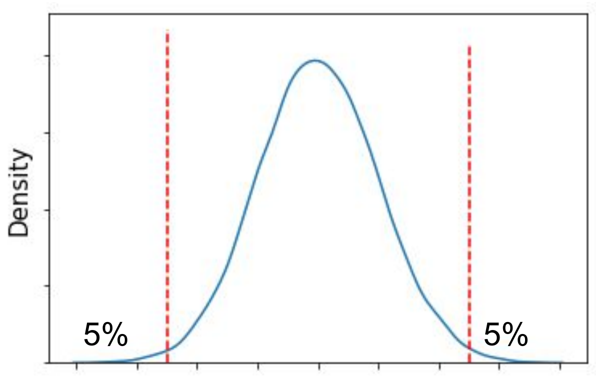
- As we said earlier, the critical region represents the tail area of a distribution.

- But distributions have two tails (tests have **directionality**!).

e.g. Is a certain parameter different from zero?

Quantifying evidence: p-values

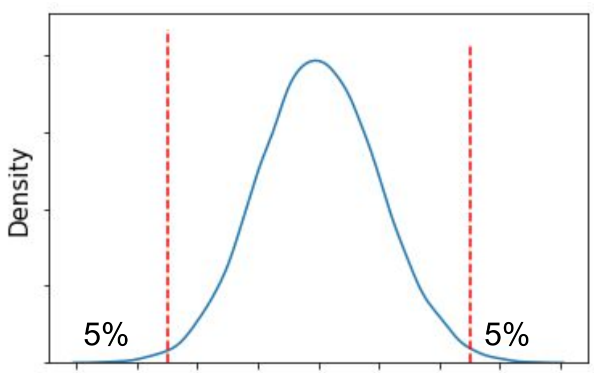
- As we said earlier, the critical region represents the tail area of a distribution.
- But distributions have two tails (tests have **directionality**!).



Quantifying evidence: p-values

- ➢ As we said earlier, the critical region represents the tail area of a distribution.

- ➢ But distributions have two tails (tests have **directionality**!).



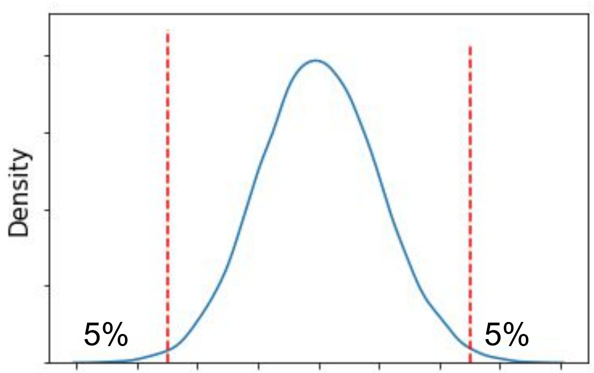
This increases our Type I Error rate!

We now reject twice as many null hypotheses!

Quantifying evidence: p-values

➢ As we said earlier, the critical region represents the tail area of a distribution.

➢ But distributions have two tails (tests have **directionality**!). We need to **adjust** the amount of critical region depending on whether we include one or both tails.



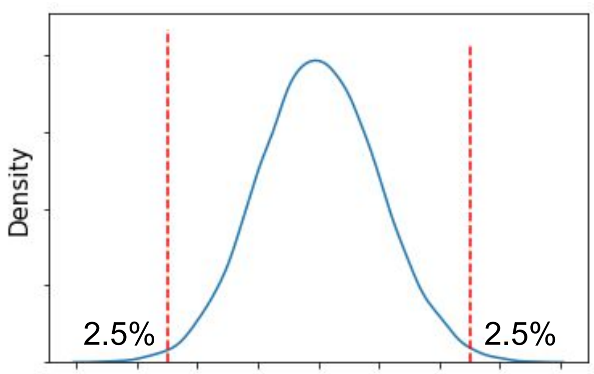
This increases our Type I Error rate!

We now reject twice as many null hypotheses!

Quantifying evidence: p-values

- As we said earlier, the critical region represents the tail area of a distribution.

- But distributions have two tails (tests have **directionality**!). We need to **adjust** the amount of critical region depending on whether we include one or both tails.



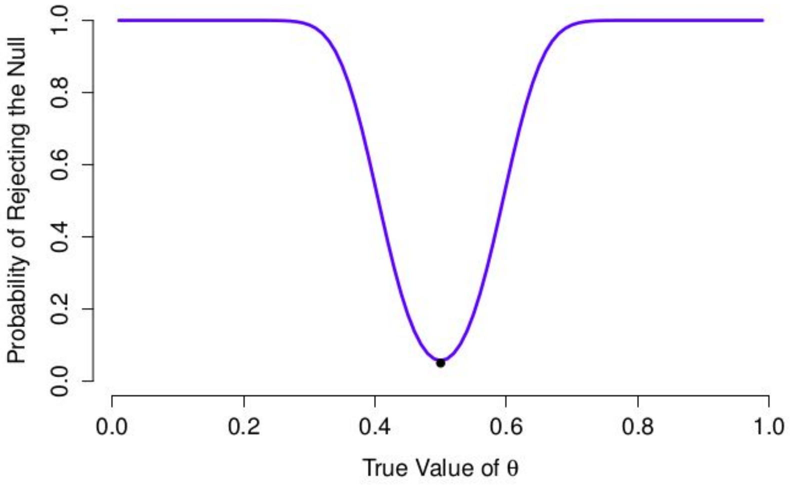
This keeps our total Type I error rate the same.

Controlling type II errors

- A secondary goal of hypothesis testing is to try to minimise the Type II error rate (β) or, maximise the **power** of the test (1-β).
- How can we increase the power of the test?

<u>Big effect sizes.</u>

- **The bigger the effect,**
  the easier to reject the null.
- This is usually something
  out of our control.



Controlling type II errors

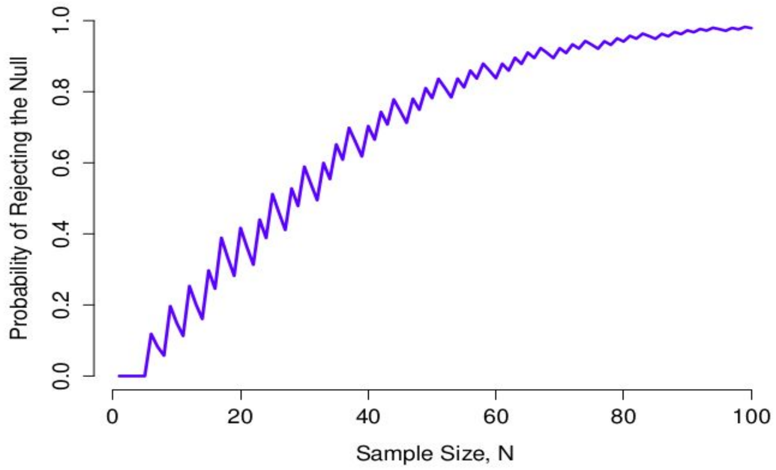
➢ A secondary goal of hypothesis testing is to try to minimise the Type II error rate (β) or, maximise the **power** of the test (1-β).

➢ How can we increase the power of the test?

<u>Big sample sizes.</u>

- **The more data, the easier to pick up any difference.**

- This is something that **we can control when designing the experiment.**



Recap

- Null hypothesis testing is a framework for **quantifying evidence**.
- We quantify this evidence to be able to **make a decision about the null**!
- We may need to **adjust** our degree of decision based on our tested alternative hypothesis.
- We **generally** talk about **Type I error**, but **Type II** errors are also **important**, particularly in **designing a study**!