Rediscovering P-values

*but not as you know it*

P-values are ubiquitous in the reporting of health research - you will notice that every statistical test performed has an associated p-value given. The p-value is the most popular, but at the same time, **most misused**, **misunderstood**, and **misinterpreted** statistical tool in all of biomedical research. This often leads to **gross overestimation** of strength of evidence, **compromising** clinical **decision making** and practical **application of data**. In fact, 7 out of 10 doctors routinely make the wrong conclusion when reading research results; most medical educators and researchers make the same mistake; and many are confident in their misunderstanding!

In this workshop, you will navigate common statistical misconceptions, learn strategies on how to avoid them in future practice, and consolidate your correct understanding of p-values & statistical significance. You will no longer fall for the blind pursuit of a p<0.05.

**Learning Outcomes:**

* Understand purpose of and the systematic approach to hypothesis testing
* Correctly define and interpret commonly misinterpreted concepts:
  + null hypothesis, alternative hypothesis
  + type 1 & 2 errors
  + level of significance, p-value & statistical significance
* Test understanding of p-values & significance by navigating common misconceptions
* Discuss the negative impact of these prevalent misconceptions and learn simple strategies to overcome them
* Develop enhanced skills in basic interpretation of research findings for better decision making & actual ‘evidence-based’ practice

**TEST YOUR CURRENT UNDERSTANDING**

*Jot down your current thoughts on what p-values & statistical significance are in research. Compare this existing interpretation with your modified understanding post-workshop!*

|  |
| --- |
| *→ What are p-values? What do they tell us?*    *→ What does a statistically significant result mean?* |

A blue rectangular frame with white background

Description automatically generated**HYPOTHESIS TESTING**

**Hypothesis testing** involves comparing a null hypothesis to an alternate hypothesis. It is the most common framework for **statistical inference** (process of extrapolating findings from a sample to the wider population from which it came).

Null hypothesis **(H0)** → always predicts no true effect, no relationship between variables, or no difference between groups.

Alternative hypothesis **(H1)** → states your main prediction of a true effect, a relationship between variables, or a difference between groups.

It is easier to disprove a theory, than to prove one (not feasible to test every single conceivable case or the entire population). Thus, the null hypothesis is assumed true until proven otherwise. If the weight of evidence leads us to believe that the null hypothesis is highly unlikely (determined by the p-value), then we have a statistical basis upon which we may reject the null hypothesis. Thus, the data collected from a sample is used to **gather evidence *AGAINST* the null,** as opposed to *FOR* the alternative.

**TWO CRITICAL ASPECTS ABOUT THE H0 TESTING FRAMEWORK:**

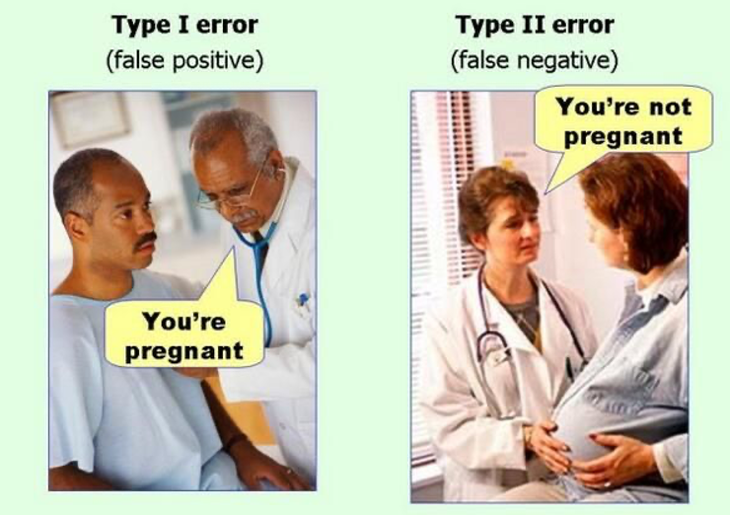
* You cannot ‘accept’ the null hypothesis. You can only reject or fail to reject it.
* You can never prove the alternative hypothesis to be true, because we are gathering evidence against the null, as opposed to for the alternative.

All you can do is conclude that your alternative hypothesis is **VALID** *(not ‘proven’)* by showing that the opposite hypothesis (H0) is highly unlikely given your collected data

**TYPE 1 & 2 ERRORS**

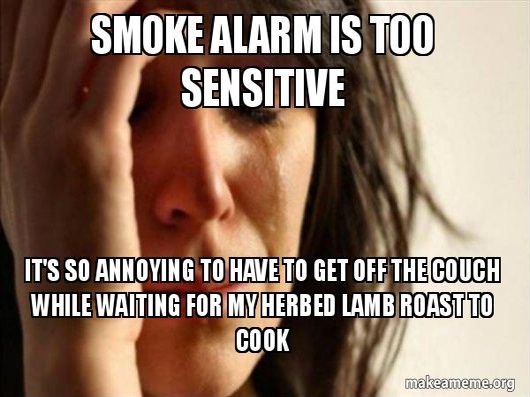
In the null hypothesis framework, it is always possible to draw an incorrect conclusion based upon sampled data. There are two kinds of error we can make:

**→ Type I error:** rejecting H0 when H0 is really true (false positive)

**→ Type II error**: failing to reject H0 when H0 is really false (false negative)

A blue and white frame with a white border

Description automatically generatedWhile both of these mean we were wrong, sometimes we may prefer and therefore want to minimize one type of error over the other. For example, in the context of smoke alarms, which error would you prefer? While the sound of the smoke alarm going off is annoying, there's not a lot of cost to having a false positive (type 1 error) – all u have to do is press a button to reset it. There is however a huge risk if your smoke alarm does not go off when there really is a fire (type 2 error)! For this reason, fire alarms tend to favour having type 1 errors over type 2 errors.



**>**

**# NULL HYPOTHESIS MODEL ANALOGY**

Our legal system uses the idea of innocence until proven guilty. With statistics that innocence is the null hypothesis - it's the idea we assume to be true until evidence is presented to show us that it's not. With enough evidence the jury rejects the idea of innocence. If there is no evidence, the accused cannot be declared guilty, but this does not mean that innocence has been proven.

*So then… how do we decide how much evidence is enough to reject the null?*

**P-VALUES**

The p-value is used to determine if enough evidence exists to reject the null hypothesis in favor of the alternative.

It indicates the **degree of compatibility** of your collected data within the null hypothesis model. In other words, it measures how **rare** or **extreme** or **ridiculous** your sample data would be if it really did come from the world of the null, a world where there is no relationship between variables or difference between groups.

P-value is **NOT** the probability that the alternate or null hypothesis is true

A blue rectangular frame with white background

Description automatically generated*In light of this definition… how do you interpret a:*

|  |
| --- |
| 1. **large p-value?**      1. **small p-value?** |

**“STATISTICAL SIGNIFICANCE”**

This is one of those terms we often hear without really understanding.

“Statistical significance” **DOES NOT** mean “proof” or “meaningfulness”, rather it signals that your results are worthy of notice, warranting further experimentation.

If a result is **statistically significant**, that means it’s unlikely to be explained solely by chance or random factors. In other words, a statistically significant result has a very low chance of occurring (*small p-value*) if there were no true effect in a research study (*in the world of the null hypothesis*) → the p-value (probability of such an extreme result) is so small, it looks too extreme or incompatible to fit the null hypothesis model → we therefore reject the null hypothesis.

*How small does p-value need to be to reach significance?*

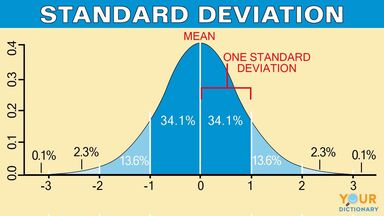
|  |
| --- |
|  |

A blue and white frame with a white border

Description automatically generated**Self-note: NORMAL DISTRIBUTION**

**Normal distribution:** a smooth, symmetric bell-shaped frequency distribution, defined by the mean (centre of the data) & standard deviation (spread of the data).

* helpful in visual representation of how data is distributed
  + 68% of data within ± 1 standard deviation of the mean
  + 95% of data within ± 2 standard deviations of the mean
  + 99.7% of data within ± 3 standard deviations of the mean
* most statistical tests applied work under the assumption of normality



Not all variables however, will naturally follow a normal distribution - some may be skewed to the left or right → Central Limit Theorem (CLT) can fix this.

**SIMULATION: Replace p<0.05 with understanding**

*Space to write notes during coin-flipping simulation incorporating concepts covered so far.*

|  |
| --- |
|  |

A blue rectangular frame with white background

Description automatically generated**PRACTICE: INTERPRETING P-VALUES**

You want to test if a drug is effective at reducing heart disease. It was found that fewer people suffered from heart disease after taking the drug (intervention group) than compared to those who were given a placebo (control group).

Assume a significance threshold of 0.01 (p<0.01) was set for this study

|  |
| --- |
| H0:  H1:  How would you interpret a p-value of:     * **p = 0.60?** * **p = 0.001?** |

**MISCONCEPTIONS: what the p-value is not!**

*Discussion notes*

|  |
| --- |
| ***The P-value is the probability that the null hypothesis is true i.e. If P<.05, the null hypothesis has only a 5% chance of being true***  True? or False?  Explanation: |

|  |
| --- |
|  |

A blue and white frame with a white border

Description automatically generated

|  |
| --- |
| ***p-value says nothing about the size of the effect***  True? or False?  Explanation: |

**ASIDE:** CONFIDENCE INTERVALS

A 95% confidence interval (CI) tells you that if you repeated your experiment many times, you would expect your constructed confidence intervals to “trap” the true parameter value (observed statistic) 95% of the time.

* easier to interpret
* more informative than p-values (shows statistical **AND** clinical significance)

A graph of values and values

Description automatically generated

If CI does not

contain 0 → statistically significant

If CI is higher than

clinical significance threshold

→ stronger result

A blue rectangular frame with white background

Description automatically generated

***A large p-value (> 0.05) means that there is no significant effect; A small p-value (< 0.05) means that there is a significant effect***

True? or False?

Explanation:

***A statistically significant finding is clinically important***

True? or False?

Explanation:

**Statistical Significance:** shows whether an effect exists in a study (using p-values)

**Practical Significance:** shows the magnitude of the effect ie. is it large enough to be meaningful in the real world (using effect size & confidence intervals)

**P-HACKING**

*What is p-hacking?!? What are some examples?*

|  |
| --- |
|  |

A blue and white frame with a white border

Description automatically generated**APPLICATION: Pinpoint areas of potential research misconduct**

**Context:**

2003 → antiretroviral drugs are still controversial for treating HIV (virus that attacks the immune system). A russian scientist invented a device that aims to destroy the HIV virus particle using electromagnetic waves (at the right frequency). Once it was found safe in animal tests, the technology was sold to a company that wanted to profit from ‘curing’ HIV.

You are the biostatistician, part of a team testing this device in a clinical trial. Together you design a double blinded dose-ranging sham controlled trial. Those randomised to the sham group sat in front of the device when it was turned off. Over a year, the outcomes monitored included: HIV viral loads, mortality, morbidity, quality of life, and biochemical measures.

**Results:**

25% decrease in HIV viral load in the treatment group compared to the sham group. Statistically significant at the 0.001 level.

*How would you interpret these results?*

|  |
| --- |
|  |

*What are some misinterpretations to avoid when interpreting these results?*

|  |
| --- |
|  |

Virologists on your team reveal that the statistically significant difference (p<0.001) was NOT clinically significant as a change of 1 log or more is required for antiretroviral treatment to be considered clinically effective.

The company and their funders are not satisfied with the results, and therefore order a series of subgroup analysis based on trends seen in the collected data. For example:

→ “just look at men over 50”

→ “just include individuals with severe disease”

A blue rectangular frame with white background

Description automatically generatedYou (biostatistician) are about to comply when you realise that this is a classic example of?

* 1. P-hacking: **ad hoc sample size selection**
  2. P-hacking: **cherry-picking**
  3. P-hacking: **harking**
  4. No p-hacking

**THE TAKE HOME**

* P-value is **NOT** the probability that our alternate or null hypothesis is true, but rather it tells us how incompatible **rare**, **extreme** or **ridiculous** our sample data would be if it really did come from the world of the null, a world where there is no relationship between variables or difference b/w groups
* ‘Statistical significance’ does not necessarily mean a ‘substantial or important finding’, rather it just means the finding was ‘mathematically unusual’
* Rejecting the null hypothesis does not mean the alternate hypothesis is true - this depends on critically evaluating the research design

Note 4 strategies to avoid and 4 strategies to incorporate moving forward with research analysis endeavours.

|  |  |
| --- | --- |
| **WHAT TO DO** | **WHAT NOT TO DO** |
|  |  |
|  |  |
|  |  |
|  |  |

**“P values are not ‘badges of truths’ and p < 0.05 is not a line that separates real results from false ones. It is simply one piece of the puzzle that should be considered in the context of other evidence“**