P-hacking

Shrey Patel
Shrey Nair

Testing efficacy of different drugs

Drug Candidates

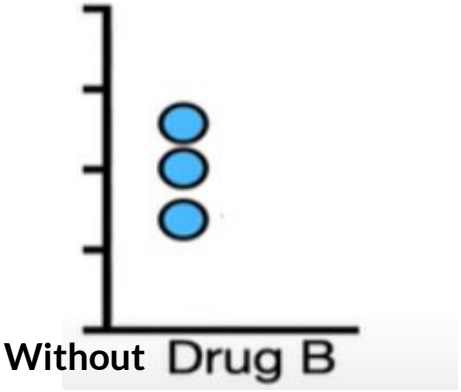
VS

Virus

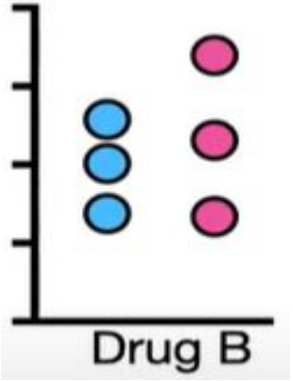
Experiment Time!

(A)

(B)

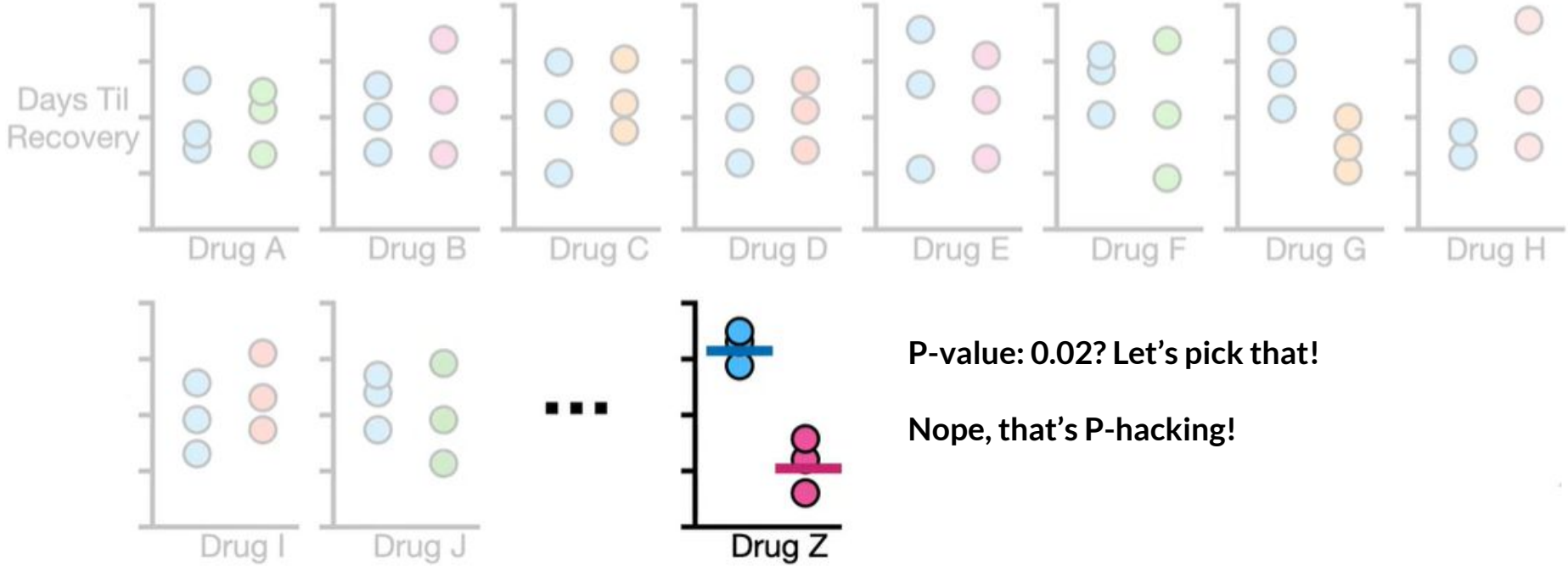


(C)



(D)

Collect data of recovery time, calculate mean and then calculate p-values by comparing mean amongst the groups

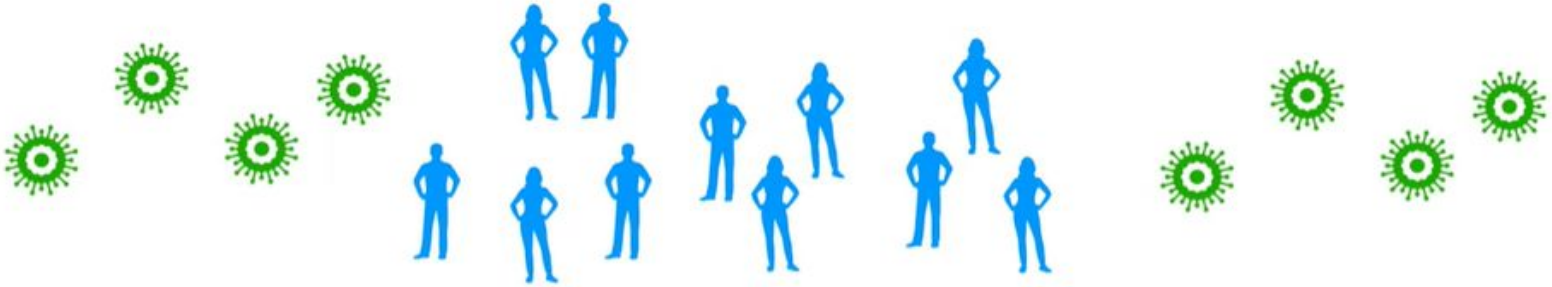


P-value: 0.02? Let's pick that!

Nope, that's P-hacking!

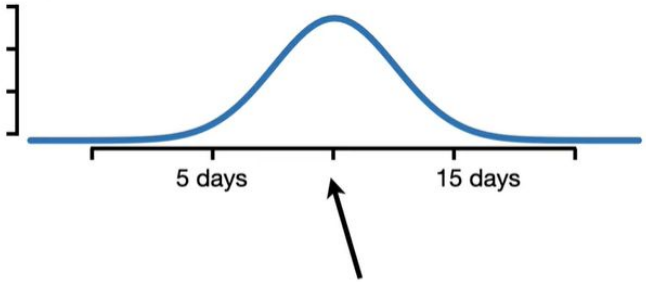
P-hacking

- Misuse and abuse of analysis techniques resulting in being tricked by false positives

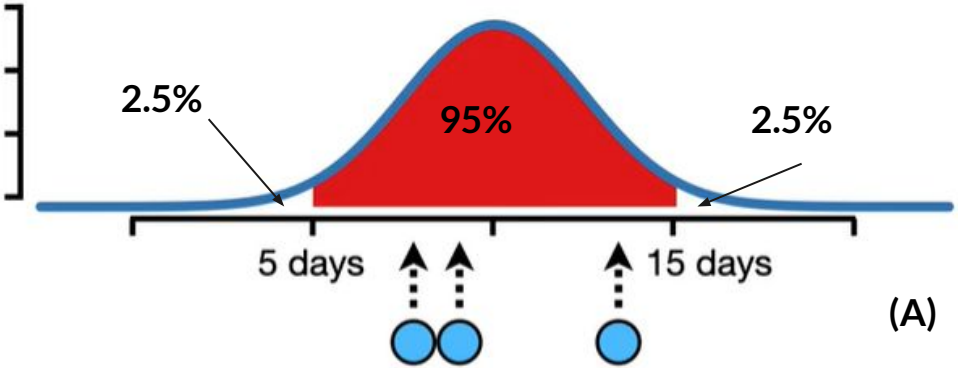


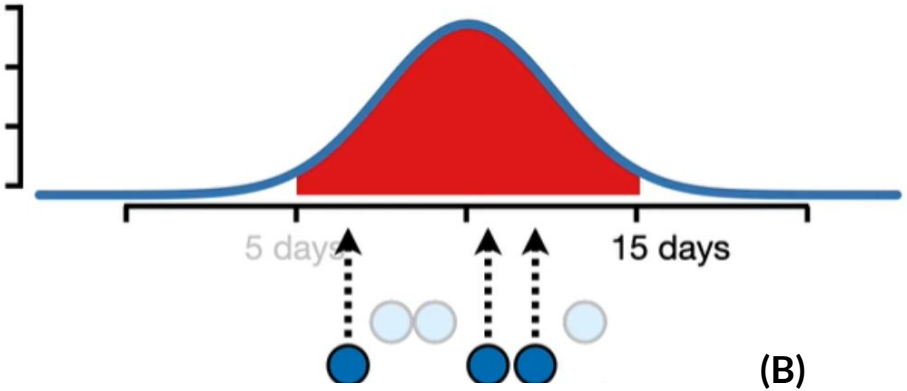
Imagine we measured recovery times for a whole lot of people who did not take any drugs to fight the virus.

Normal Distribution of Patient Recovery

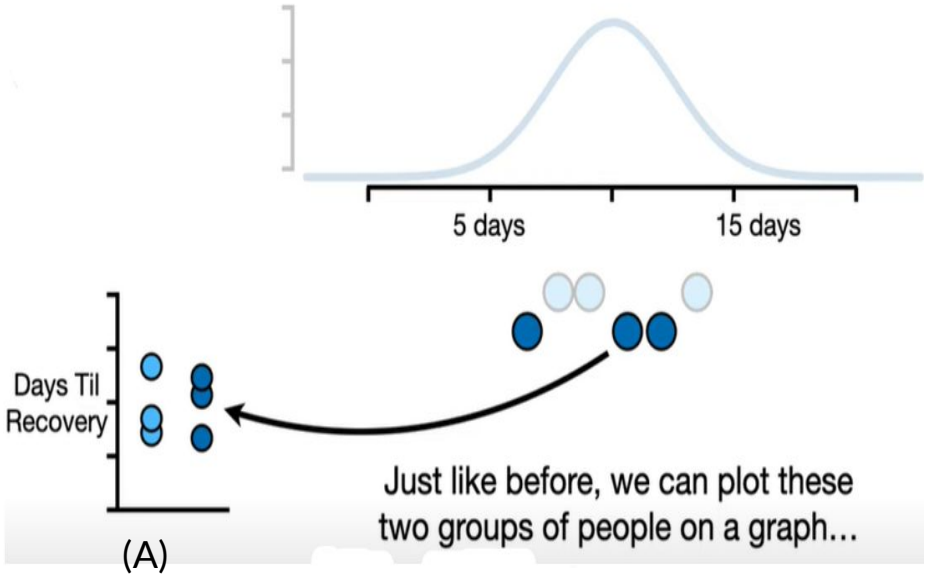


And then we fit a normal distribution to all of the recovery times.

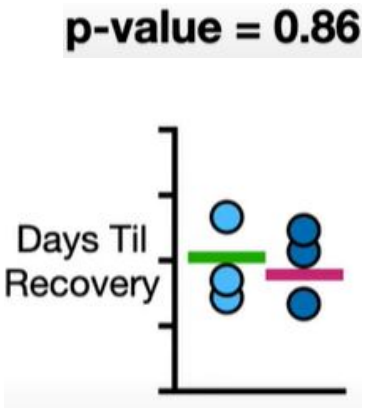




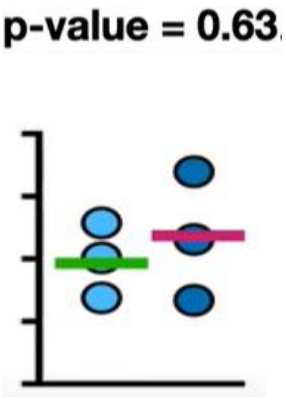
Calculation and Comparison of p-values



(A)

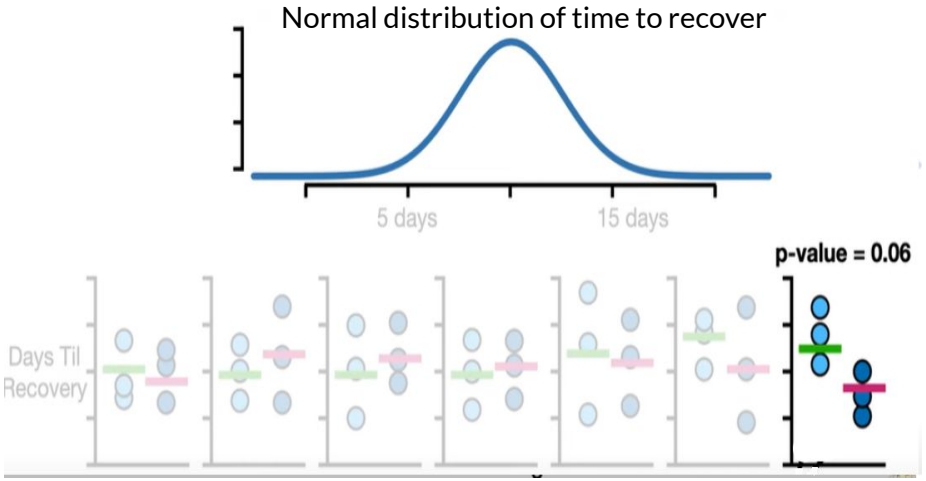


(A)



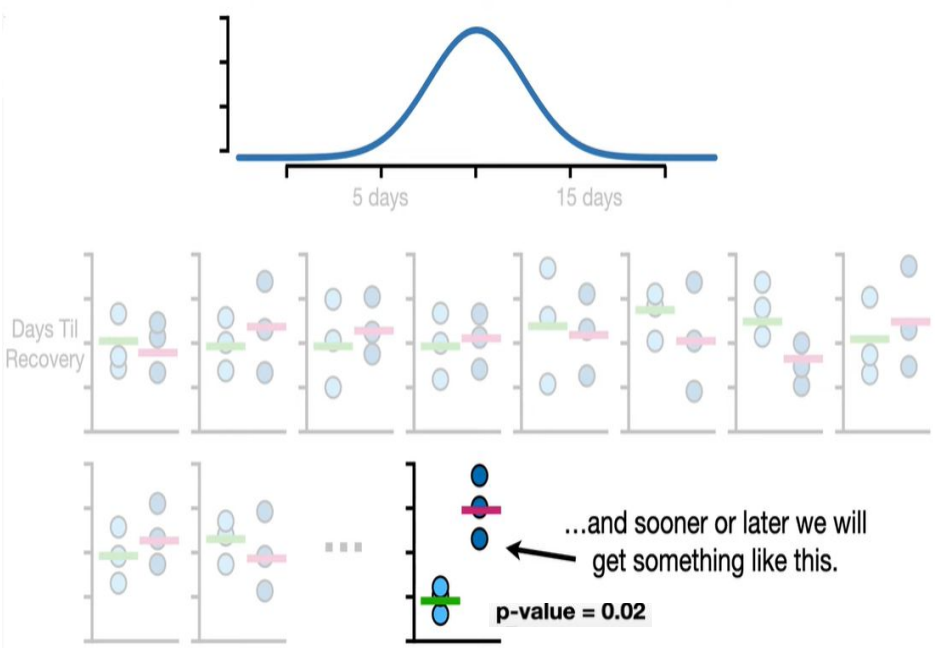
(B)

Multiple testing problem



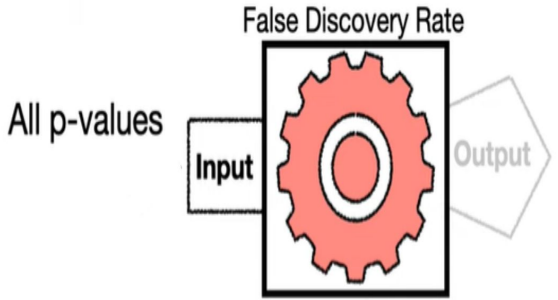
(C)

Collect data of recovery time, calculate mean and then p-values by comparing means of two groups



False Discovery Rate

- Input the p-values for every single comparison
- Outcome are adjusted p-values that are larger than the original p-values
- A way to filter out False positives
- In order for false discovery rates to work properly, all the p-values for all tests have to be included and not the ones that have a small p-value



How bad is p-hacking really?

- It's pretty bad.
- P-hacking is also known as data dredging, data fishing, data snooping, data butchery, significance chasing, significance questing, and selective inference.
- P-hacked analyses are misleading.
- But is p-hacking always malicious or could it just be a byproduct of techniques used in our analysis?
    - It could be a researcher's honest mistake
    - Or arise from a gap in their statistical knowledge
    - Or it could be their well-intentioned belief in a specific scientific theory

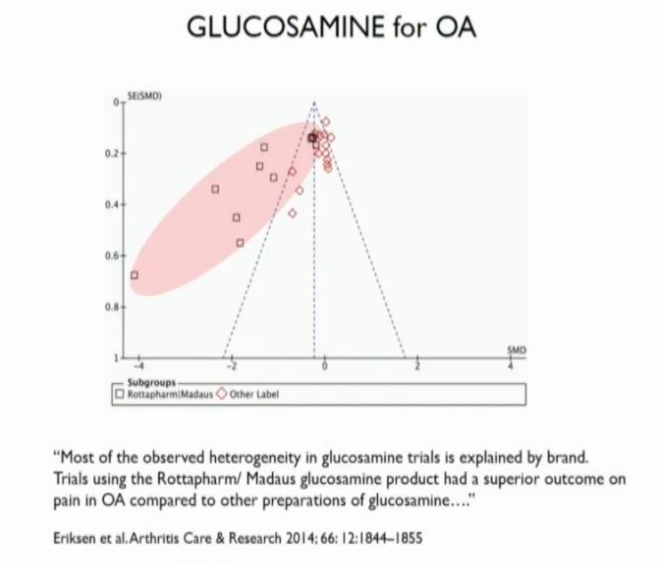
Why does it happen?

Usually to "secure the bag"

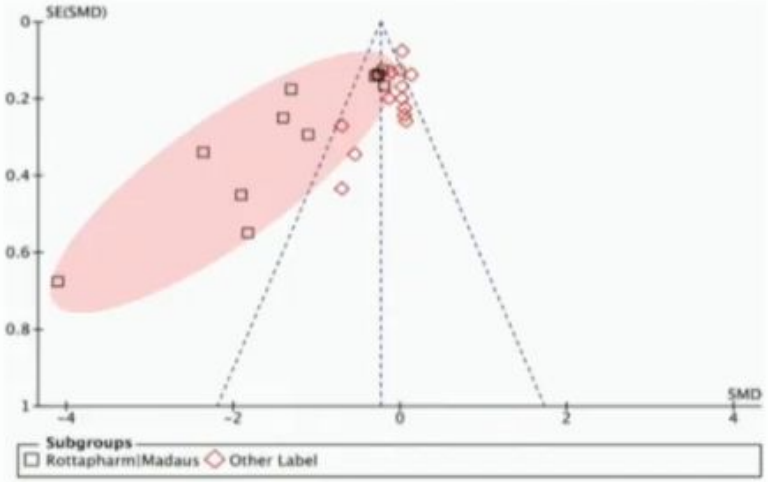
- Why do scientists p-hack? Or why does it even happen?
    - What is the main objective of an analysis?
        - Is it to reasonably test a hypothesis? OR
        - Is it to hunt for a statistically significant result?
- Encountering statically insignificant results can tempt researchers into pursuing analytical methods that could lead to multiple testing problem.
- Running into a p-value that is slightly more than the threshold (for example, 0.06) can tempt researchers into slightly altering the sample size in order to lower the p-value.
- Both of the above cases lead to inflation in Type-I errors.

Publication Bias

- Many argue that current scientific practices create strong incentives to publish statistically significant (i.e., “positive”) results
- There is also good evidence that journals, especially prestigious ones with higher impact factors, disproportionately publish statistically significant results
- This leads to **researcher-driven publication bias** - meaning, it creates an incentive for researchers to selectively pursue and selectively attempt to publish only the significant findings.
    - Selective reporting (p-hacking)



GLUCOSAMINE for OA



"Most of the observed heterogeneity in glucosamine trials is explained by brand. Trials using the Rottapharm/ Madaus glucosamine product had a superior outcome on pain in OA compared to other preparations of glucosamine…."

Eriksen et al. Arthritis Care & Research 2014; 66: 12:1844–1855

Where's the rest of the plot? :/

Consequences of p-hacking

- If the published data are bias, meta-analysis might lead to flawed conclusions.
  - Meta-analysis: A statistical method of pooling data together from many clinical trials to get a quantitative estimate of the overall effect of a particular intervention or variable on a defined outcome (For example, combining effect sizes from multiple studies to get an estimate of the overall effect size)
  - Results from meta-analysis might inherit bias from the individual studies - leading to analytical sloppiness
  - Lack of understanding of basic issues
  - Failure to consider major covariates
  - Overstating the strength or precision of the results

How to prevent p-hacking?

What to **not** do?

- Conducting analyses midway through the experiment to decide whether to continue collecting data.
- Recording many response variables and deciding which to report post-analysis.
- Deciding whether to include or drop outliers post-analyses.
- Excluding, splitting, or combining treatment groups post-analysis.
- Including or excluding covariates post-analysis.
- Stopping data exploration if an analysis yields a significant p-value.

References

- “Data dredging” https://en.wikipedia.org/wiki/Data_dredging
- “StatQuest with Josh Starmer” https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw
- “Science isn’t broken” https://fivethirtyeight.com/features/science-isnt-broken/
- “P-Hacking: Crash Course Statistics #30” https://www.youtube.com/watch?v=Gx0fAjNHb1M
- “The Extent and Consequences of P-hacking in Science” https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106
- “Risk of Bias and Brand Explain the Observed Inconsistency in Trials on Glucosamine for Symptomatic Relief of Osteoarthritis: A Meta-Analysis of Placebo-Controlled Trials” https://onlinelibrary.wiley.com/doi/full/10.1002/acr.22376

Thank you and stay true to yourself!