DR. XITZEL SÁNCHEZ CASTRO

SCIENTIFIC ANALYSIS IN BUSINESS

XVIII Mexican Workshop of Particles and Fields 2022

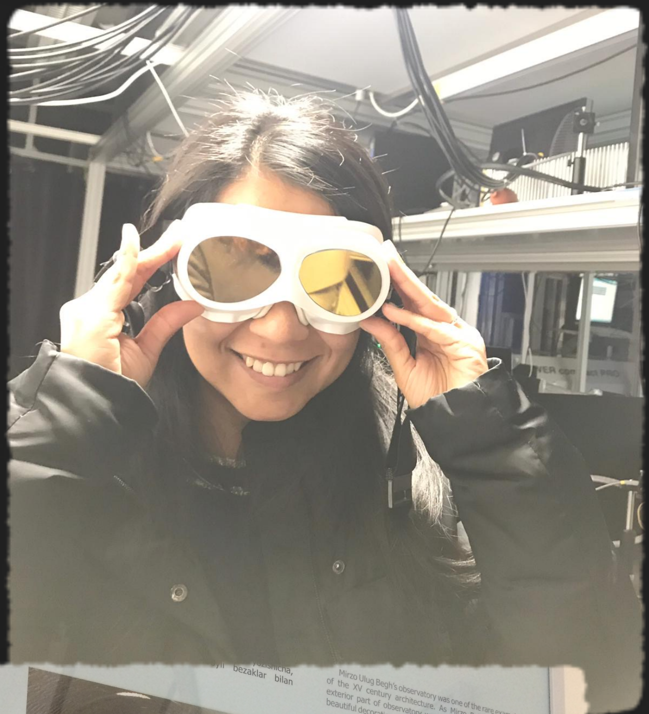
November 25, 2022

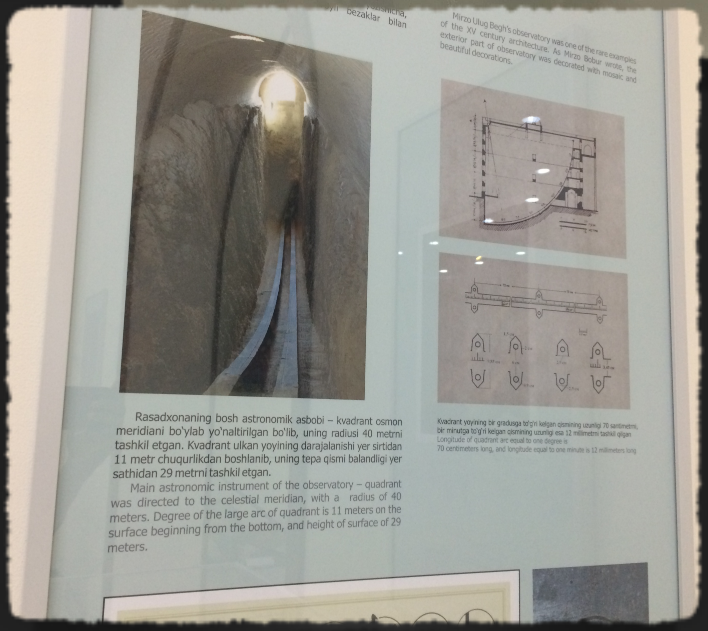
OUTLINE

- About me
- Statistical experiments
  - How to?
  - Further considerations
- Q&A

ABOUT ME

- 2009 - Bachelor of Physics. Thesis: *"Study of the strange and charm production in relativistic heavy-ion collisions in the threshold model"*, UNAM
- 2011 - Master of Science (physics). Thesis: *"Two-hadron correlations with strangeness in proton-proton collisions at 7 TeV in ALICE"*, UNAM
- 2015 - Ph.D. in physics. Thesis: *"K^0_S and Λ production associated to high-p_T charged hadrons in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 2.76 TeV with ALICE"*, UDS/IPHC/CNRS/CERN
- 2015-now - Business experience as Data Scientist / Master Data Developer / Portfolio Analyst / Scrum Master across industries in international and multicultural environments
- Hobbies: sports, traveling, reading and learning languages





TALES ABOUT CORRELATION AND CAUSATION

In 2020, the correlation between police spending and crime was examined at a Washington Post article [1]:

"A review of spending on state and local police over the past 60 years… shows no correlation nationally between spending and crime rates."

However, causal research has shown that more police lead fo a reduction in crime.

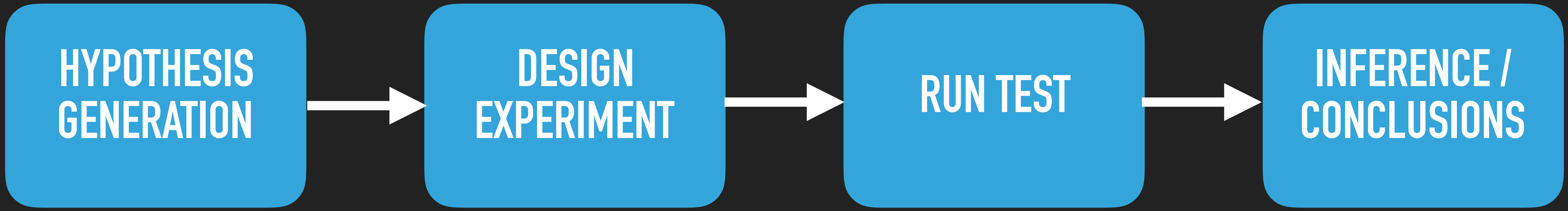
STATISTICAL EXPERIMENTS

- Testing a new option (B) versus a default option (A)
  - Has an intervention had a positive/negative impact?
- Physics
  - Search for new physics by using likelihood ratio tests [2,3]
- Social or medical sciences [4]
  - How does income affect childhood brain development? [5]
  - Evaluate the affective psychology of value [6]
- Online controlled experiments [7]
  - Increasing sales by sending promotional emails/mails/messages that include a coupon code for discount
  - Increasing the number of new consumers that signed up for a service after a trial period
- Generally speaking, modern products rely on understanding user behaviour (*User Analytics*)

GOOD TO KNOW

- Statistical experiments are also known as statistical testing, randomised controlled trial, A/B testing
- Treatment is the action or feature to which a subject is exposed, a.k.a. variation
- Treatment group is a random group of subjects exposed to a specific treatment
- Control group is a random group of subjects which receive no treatment

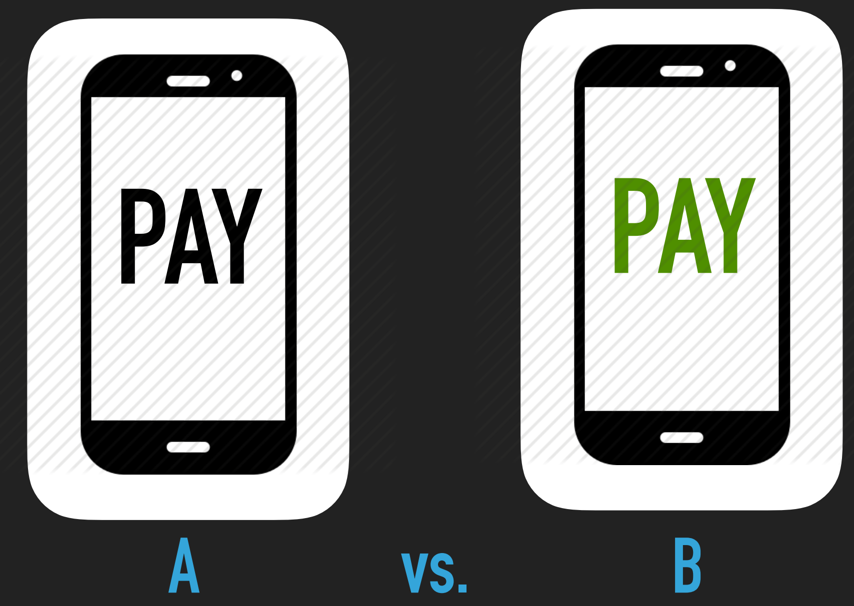
HOW TO?

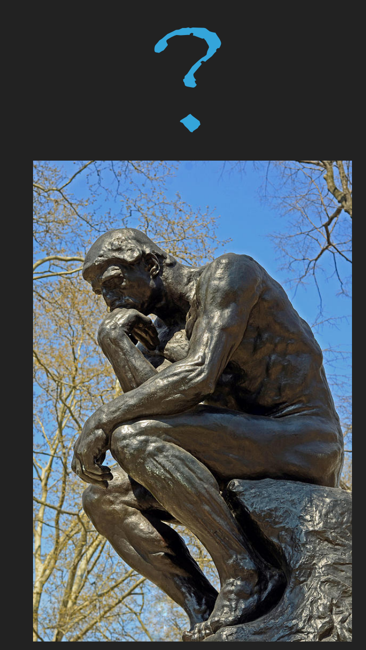


- In business, further considerations need to be considered before starting the experiment pipeline:
  - Are the technical systems in place?
  - Has the management agreed with the testings?

HYPOTHESIS GENERATION

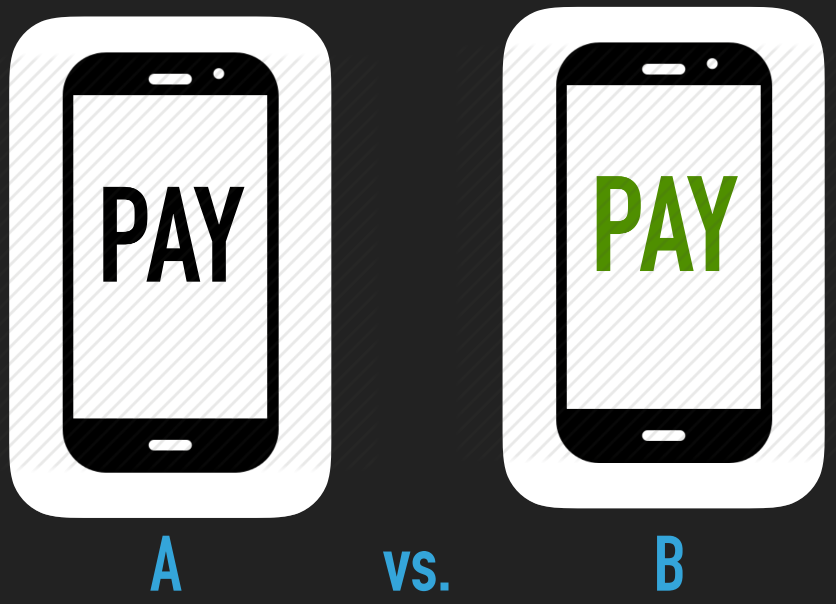
- A hypothesis is always needed and is used to understand whether random chance might be responsible for an observed effect
    - Example of one effect that we want to observe: *"Setting in green 'PAY' in the checkout page will increase revenue per customer"*
    - Treatment: the letters in green
- A true difference between groups A and B needs to be present
- … but what we really want is to reject the null hypothesis

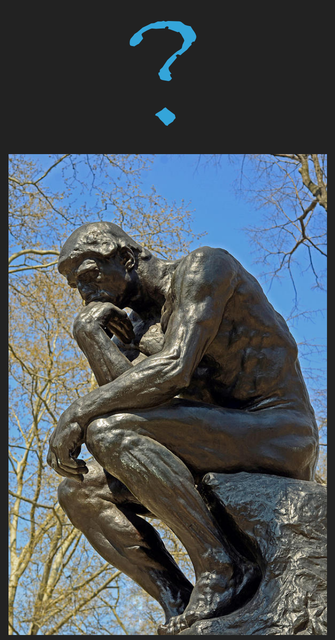




NULL HYPOTHESIS H_0

- H_0 is the baseline assumption that the treatments are equivalents, and any difference between groups is due to chance. Otherwise, the alternative hypothesis H_A holds
  - Ex. H_0 = *"The 'BUY' text in green will not increase revenue per customer"*
- It is necessary to understand the baseline value (mean or percentiles)
- It is equally important to know the distribution that follows, so we can apply later the right statistical test





TYPES OF ERRORS & POWER

- **Type I error** is when the null hypothesis is rejected and erroneously claiming H_A is true [10]
- **Type II error** is when no real difference is declared between treatment and control when there was one
- **Power** of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true:

  power = 1 - **Type II error**

  It is common in industry to choose 80% or 90% power

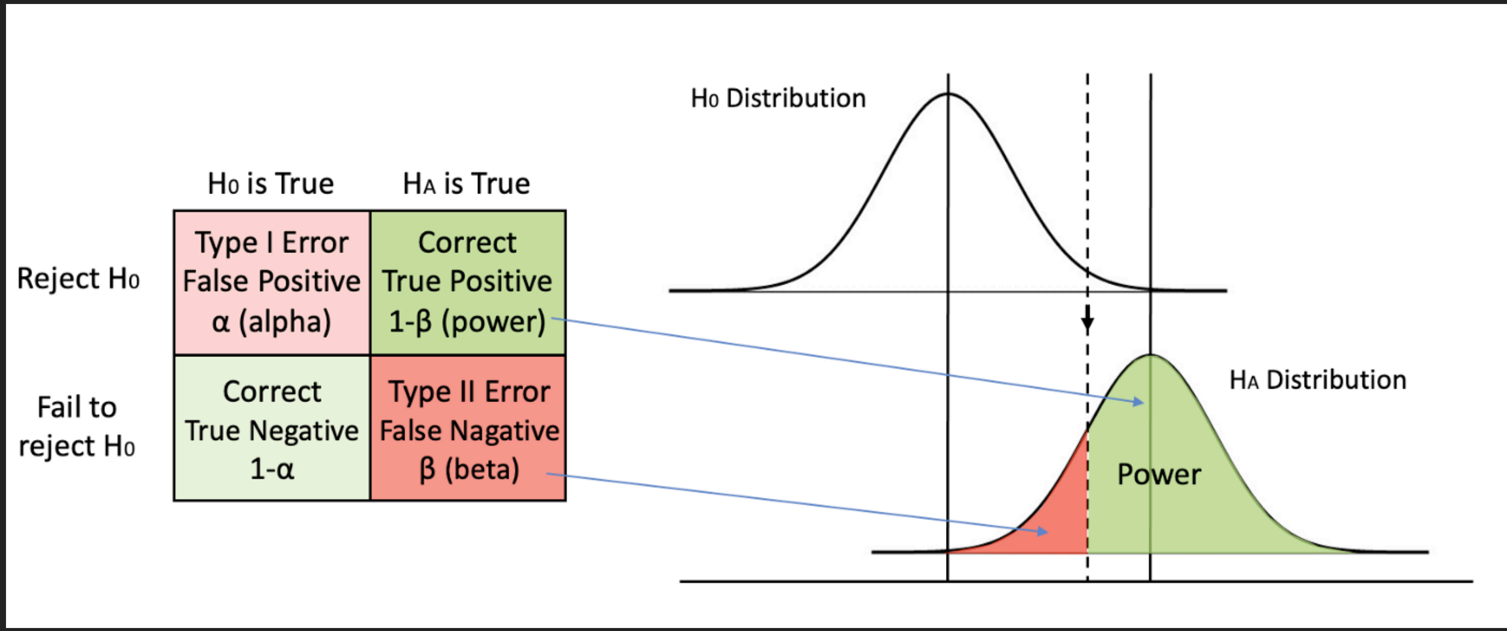
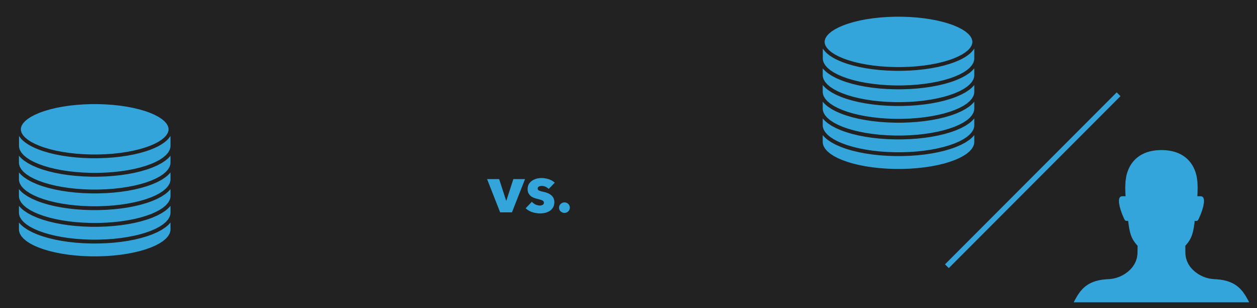


Image by A. Yuhan Yao [8]

DESIGN EXPERIMENTS

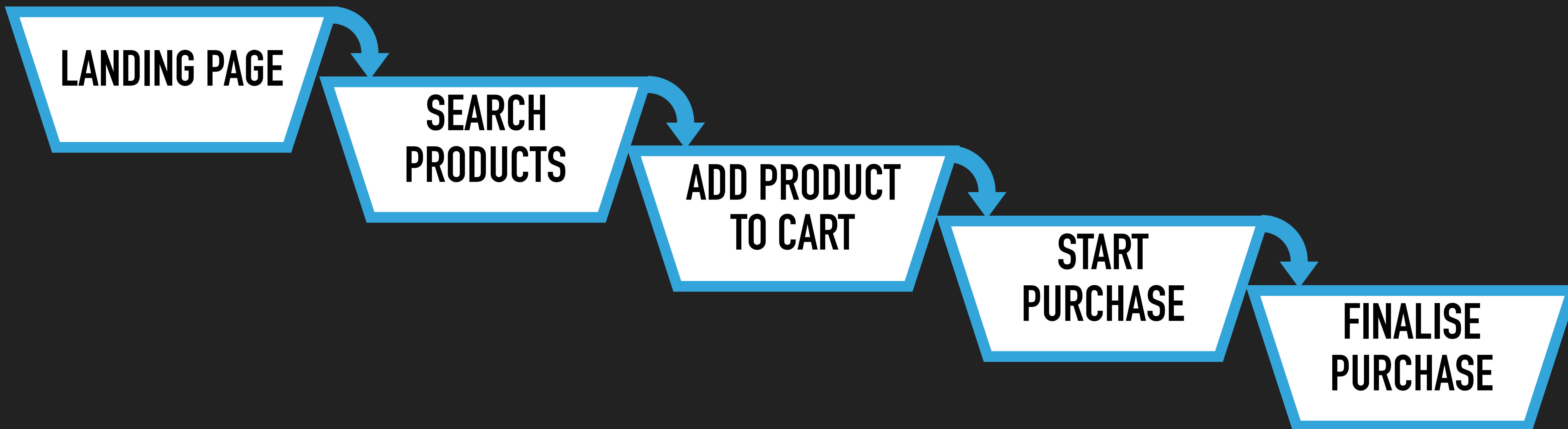
- Define success metrics to measure the impact of the change



Aggregated metric | Normalised metric

  - The success metric should be easy to measure
  - Other possible metrics:
    - Click-through rates, cost of acquisition, customer churn rate, conversion rate
  - From the business perspective, what size of impact is meaningful to detect?
    - +0.2% or a +10% change?
  - Sensitivity: the minimum level of change that we want to be able to detect in the test

- Define who are the users [7]



- At which step should we take the users to test them?

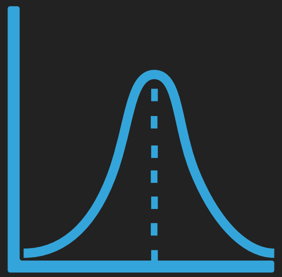
All users visiting the website

Users who started the purchase process

Users who completed the purchase process

- Which is the randomised mechanism?
  - users, user-day, session-level, etc.?
  - When and how will the users be put into randomised groups?
  - Randomisation helps eliminating selection bias

- Which population do we want to target?
  - Having a particular characteristic in common, ex. language setting, geographical location, platform, device type, user persona

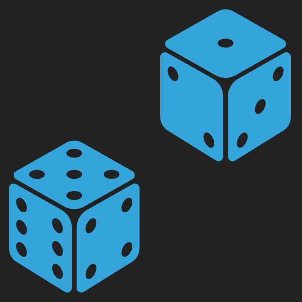


- How large does the sample size of our experiment need to be?
  - For a 80% power: $n \approx \dfrac{16\sigma^2}{\delta^2}$ [7] , with σ^2 is the base variance and δ is the sensitivity

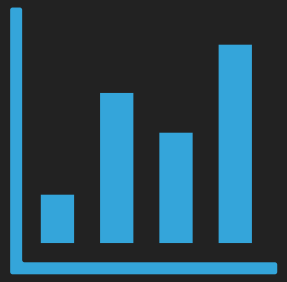
- How long do we need to run the experiment?
  - It's important to consider statistical power, effects due to day-of week, seasonality, novelty

RUN TEST

Things to check:



- Is randomisation working as expected?
- Is the data being collected?
- Walk yourself through the steps for treatment and control



INFERANCE / CONCLUSIONS

- It is important to evaluate the collected data
  - Does it look reasonable?

Groups	Size	Avg. Revenue	SD
Control	48 236	$25.3	$3.7
Treated	49 867	$26.1	$3.8

Observed difference of -$0.8

- We also need to calculate the treatment effect
  - Confidence interval
  - *p*-value

CONFIDENCE INTERVAL

- Does the confident interval of the observed difference overlap with zero?
- To get the range we apply [12]:

$$C.I. = (\mu_1 - \mu_2) \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n_1 + n_2}} \right)$$

with $z_{\alpha/2}$ is the $\alpha/2$ quantile of a normal distribution

- It is common to get the limits with a 95% probability of observing a true difference
- In our example, the confidence interval at 95% of our treated group is:

Groups	Size	Avg. Revenue	SD	Diff. Revenue w.r.t. Control	Confidence interval
Control	48 236	$25.3	$3.7	––	––
Treated	49 867	$26.1	$3.8	-$0.8	[-$0.85,-$0.75]

P-VALUE

- Usually, one does not report the outcome of the decision that is H_0 or H_A, but instead the *p*-value
- *p*-value is the probability of an equal or more extreme outcome occurs, when the null hypothesis is true
- It is standard to use the following thresholds to decide agains a null hypothesis [11]

  p-value <= 0.1 ⟷ weak evidence against H_0

  p-value <= 0.05 ⟷ increased evidence against H_0

  p-value <= 0.01 ⟷ strong evidence against H_0

- The *p*-value is derived through a statistical test

P-VALUE

- Since our metric *revenue per customer* follows a normal distribution and we want to compare the means, then we need to apply a *t*-test:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}}$$

- By getting the *p*-value from a Student's t-distribution table

Groups	Size	Avg. Revenue	SD	Diff. Revenue w.r.t. Control	Confidence interval	*p*-value
Control	48 236	$25.3	$3.7	––	––	––
Treated	49 867	$26.1	$3.8	-$0.8	[-$0.85, -$0.75]	< 0.0001

PROPOSAL

- Given the test results of our example, we can say that the alternative hypothesis is true within a confidence interval of [-\$0.85, -\$0.75] and *p*-value less than 0.0001

FURTHER CONSIDERATIONS

- Do we really need to go through the whole implementation of a feature to test it?
- We could also start with a simple A/A testing
  - In medicine, give the same placebo to all patients
  - In a digital product, for example, the field to enter a discount code can be visible before sending discounts codes to the users
- Why an A/A test? [7]
  - Ensure Type I errors are controlled
  - Assessing other pitfalls

MORE THAN JUST A & B

- Multi-arm bandit algorithm is a *smarter* version of the A/B testing that uses algorithms to dynamically allocate traffic to treatments that are performing well, while allocating less traffic to treatments that are underperforming
  - It allows explicit optimisation and faster decision making
  - For example, an algorithm to implement is the *epsilon-greedy algorithm* [13]:
    - Get a random number *x* from an uniform distribution between 0 and1
    - If *x* is between 0 and epsilon (epsilon between 0 and 1) , then flip a coin (50/50 probability) and:
      - Show A if the coin is heads, otherwise show B
    - If *x* is larger than epsilon, shows the offer than has the highest response rate to date
    - When epsilon is 1, we have the standard simple A/B experiment. When epsilon is zero, we have a purely *greedy* algorithm

Q&A

REFERENCES

[1] Luca, M. "Leaders: Stop Confusing Correlation with Causation" Harvard business review (2021)

[2] Algeri, Sara, et al. "Searching for new physics with profile likelihoods: Wilks and beyond." arXiv preprint arXiv:1911.10237 (2019).

[3] Cowan, Glen, et al. "Asymptotic formulae for likelihood-based tests of new physics." The European Physical Journal C 71.2 (2011): 1-19.

[4] Bhide A, Shah PS, Acharya G. "A simplified guide to randomized controlled trials." Acta Obstet Gynecol Scand. 2018 Apr;97(4): 380-387. doi: 10.1111/aogs.13309. Epub 2018 Feb 27. PMID: 29377058.

[5] Noble, K. G., MD, PhD. NPR Talks to Kim Noble About Study of Cash Gifts to Moms

[6] Hsee, Christopher K., and Yuval Rottenstreich. "Music, pandas, and muggers: on the affective psychology of value." Journal of Experimental Psychology: General 133.1 (2004): 23.

[7] Kohavi, Ron, Diane Tang, and Ya Xu. "Trustworthy online controlled experiments: A practical guide to a/b testing." Cambridge University Press, 2020.

[8] A. Yuhan Yao, 5 ways to Increase Statistical Power

[9] Joanne Rodrigues. "Product Analytics: Applied Data Science Techniques for Actionable Consumer Insights." (2020)

[10] Shreffler J, Huecker MR. "Type I and Type II Errors and Statistical Power." [Updated 2022 Mar 18]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557530/

[11] Kauermann, Göran, et al. "Statistical Foundations, Reasoning and Inference: For Science and Data Science." Schweiz, Springer International Publishing, 2021.

[12] G. Georgiev, ''Confidence Intervals: A Guide for A/B Testing

[13] Bruce, Peter, Andrew Bruce, and Peter Gedeck. "Practical statistics for data scientists: 50+ essential concepts using R and Python." O'Reilly Media, 2020.