DeepMind

Causality in A/B Testing

Alan Malek (DeepMind, formerly Optimizely)

25/05/2022

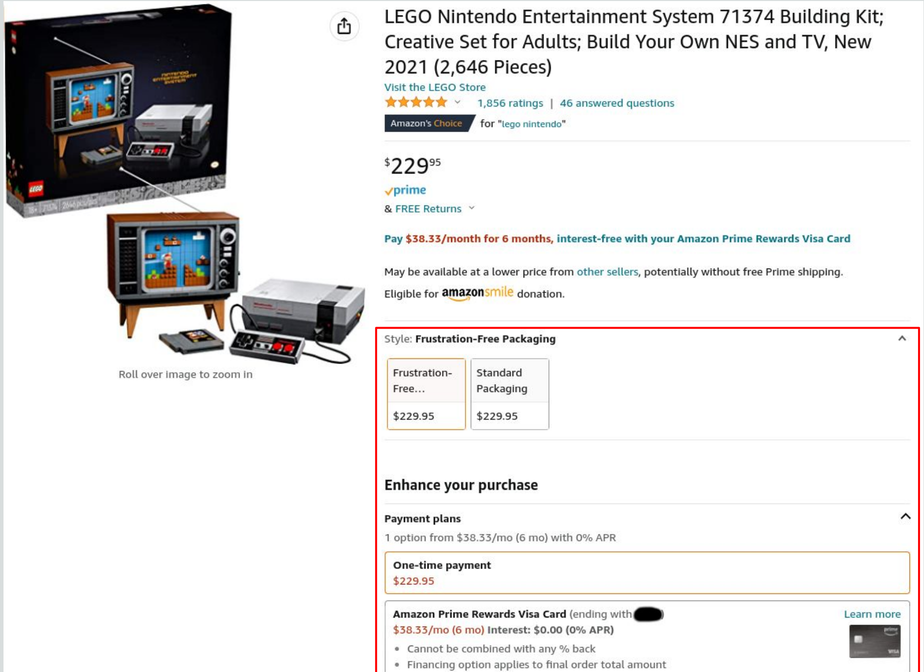
DeepMind

1 Causal Inference?

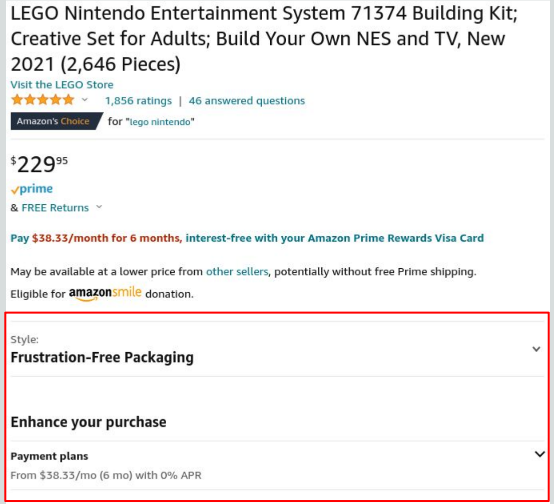
Example: advanced options

- New check-out flow
  - Present “advanced options”
  - Want to measure impact on spend
- Users can opt-in to beta, which shows “advanced options” by default

- Two user types:
  - Regular users
  - Power users
    - More likely to opt-in
    - Love options

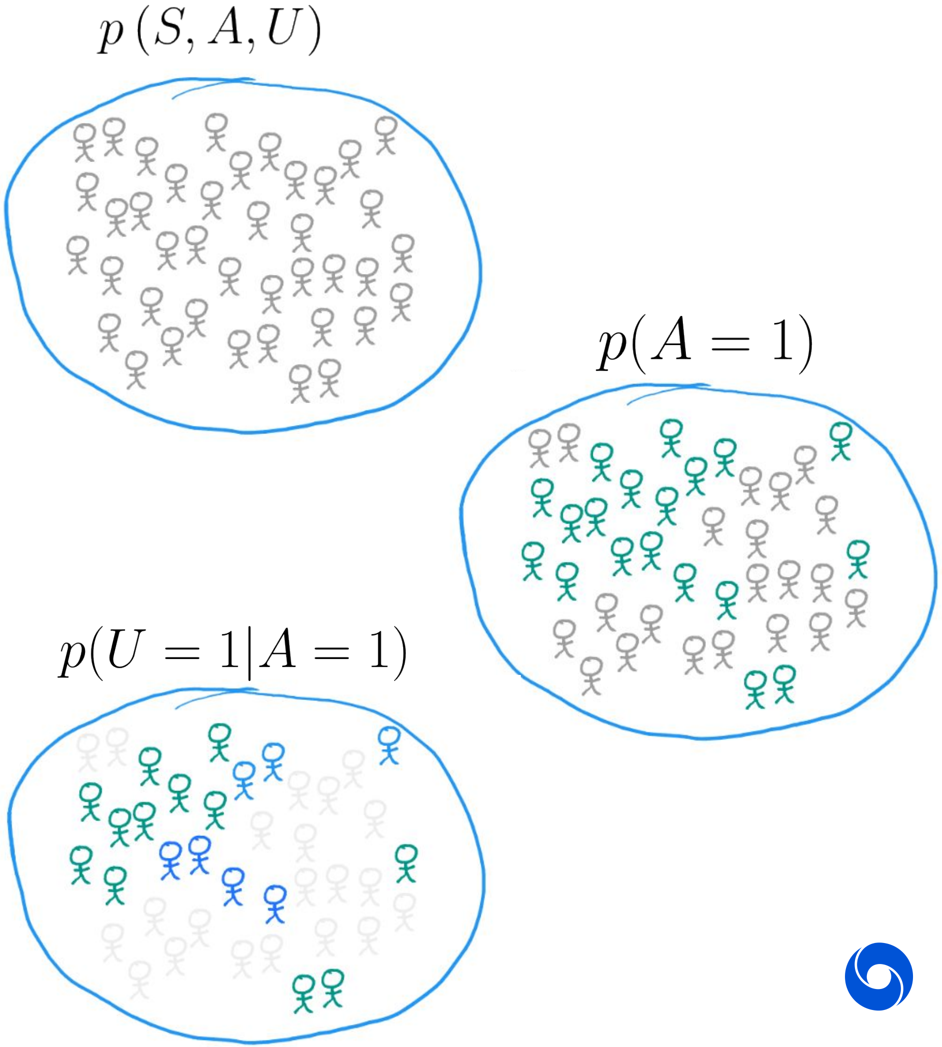


VS



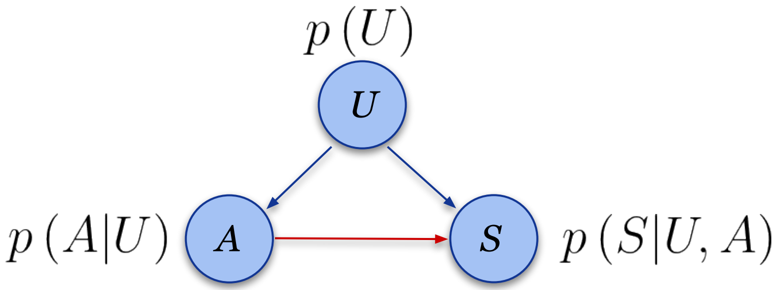
Prelim I: Probability

- All users in the world: the population
- Model attributes
  - U: power user {0, 1}
  - A: advanced options {0, 1}
  - S: difference in spend $
- Joint distribution $p(S, A, U)$ describes user demographics



Prelim II: Causal Model

- All users in the world: the population
- Model attributes
  - *U*: power user {0, 1}
  - *A*: advanced options {0, 1}
  - *S*: difference in spend $
- Joint distribution $p(S, A, U)$ describes user demographics
- Causal model describes causal relationships between attributes
  - If an attribute changed, which other attributes would?



Constructing an example

- Power users less common

$$p(U=1)=\frac{1}{3}, p(U=0)=\frac{2}{3}$$

- Power users love new features

$$p(A=1|U=1)=\frac{8}{12}$$

- Regular users do not

$$p(A=1|U=0)=\frac{5}{12}$$

- Power user love options

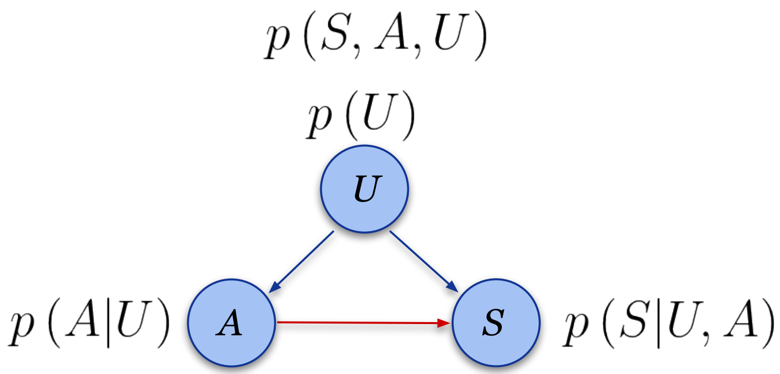
$$\mathbb{E}[S|A=1,U=1]=\$45$$

- Regular users are get confused easily

$$\mathbb{E}[S|A=1,U=0]=-\$27$$

- Status Quo

$$\mathbb{E}[S|A=0,U=1]=\mathbb{E}[S|A=0,U=0]=\$0$$



$$p(A=1)=p(A=1|U=1)p(U=1) + p(A=1|U=0)p(U=0) = \frac{8}{12}\frac{1}{3}+\frac{5}{12}\frac{2}{3}=\frac{1}{2}$$

$$p(U=1|A=1)=\frac{p(A=1|U=1)p(U=1)}{p(A=1)}=\frac{\frac{8}{12}\frac{1}{3}}{\frac{1}{2}}=\frac{4}{9}$$

Are advanced options good?

- Idea: We have observational data

$$(a_1, u_1, s_1), \ldots, (a_n, u_n, s_n)$$

- Look at:

$$\mathbb{E}_n[S|A=1] - \mathbb{E}_n[S|A=0]$$
$$= \frac{\sum_{i=1}^n 1_{\{a_i=1\}} s_i}{\#_{\{a_i=1\}}} - \frac{\sum_{i=1}^n 1_{\{a_i=0\}} s_i}{\#_{\{a_i=0\}}}$$

- Calculate:

$$\begin{aligned}\mathbb{E}[S|A=1] &= \mathbb{E}[S|A=1, U=1]p(U=1|A=1) \\ &\quad + \mathbb{E}[S|A=1, U=0]p(U=0|A=1) \\ &= \frac{4}{9}\$45 - \frac{5}{9}\$27 = \$5\end{aligned}$$

$$\mathbb{E}[S|A=0] = \$0$$

- Indicates that we should add options!

What will happen if *we* set A=1?

- We only looked at correlations in the data: found that higher spend appeared when additional options are displayed
- What do you think will happen
- If we change *A=1* for everybody?
- Poll:
  a. We will see a $5 increase
  b. The increase will be more than $5
  c. The increase will be less than $5
  d. The spend will actually decrease

- Power users less common

$$p(U=1)=\frac{1}{3}, p(U=0)=\frac{2}{3}$$

- Power users love new features

$$p(A|U=1)=\frac{8}{12}$$

- Regular users do not

$$p(A|U=0)=\frac{5}{12}$$

- Power user love options

$$\mathbb{E}[S|A=1,U=1]=\$45$$

- Regular users are get confused easily

$$\mathbb{E}[S|A=1,U=0]=-\$27$$

- Status Quo

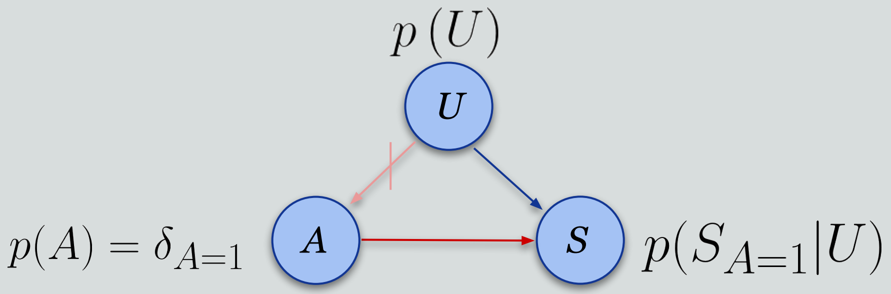
$$\mathbb{E}[S|A=0,U=1]=\mathbb{E}[S|A=0,U=0]=\$0$$

Mismatch

- Answer: d) The spend will decrease!

$$\mathbb{E}[S_{A=1}] = \mathbb{E}[S|A=1, U=1]p(U=1) + \mathbb{E}[S|A=1, U=0]p(U=0) = \frac{1}{3}\$45 - \frac{2}{3}\$27 = -\$3$$

Set *A=1* by *intervention*



- Power users less common

$$p(U=1) = \frac{1}{3}, p(U=0) = \frac{2}{3}$$

- Power users love new features

$$p(A|U=1) = \frac{8}{12}$$

- Regular users do not

$$p(A|U=0) = \frac{5}{12}$$

- Power user love options

$$\mathbb{E}[S|A=1, U=1] = \$45$$

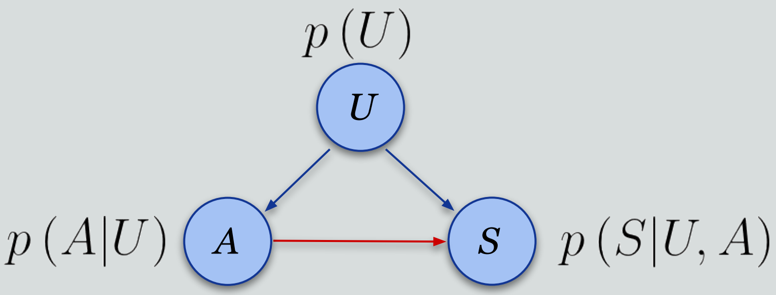
- Regular users are get confused easily

$$\mathbb{E}[S|A=1, U=0] = -\$27$$

- Status Quo

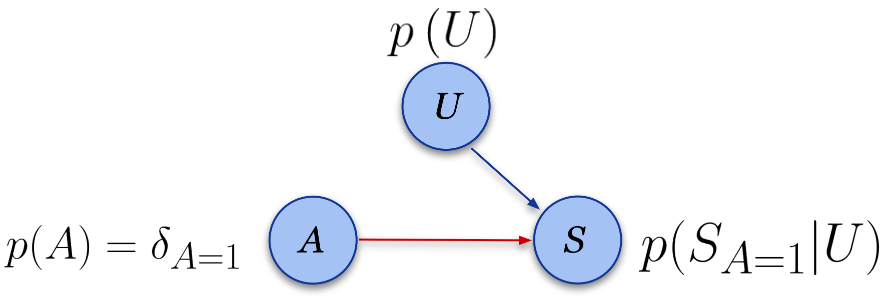
$$\mathbb{E}[S|A=0, U=1] = \mathbb{E}[S|A=0, U=0] = \$0$$

Conditional Distribution (Correlation)



$\neq$

Causal Effect (Causation)



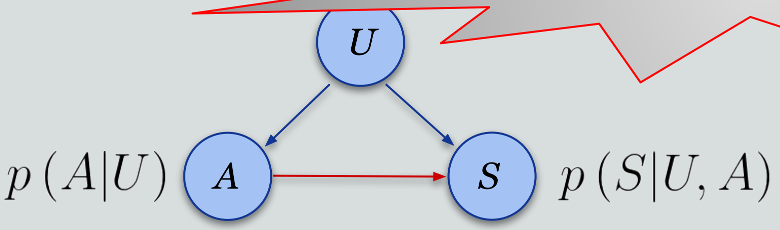
$$\mathbb{E}[S|A=1]$$
$$= \sum_u \mathbb{E}[S|A=1, U=u]p(U=u|A=1)$$

In our data, *U=1|A=1* was greatly overrepresented

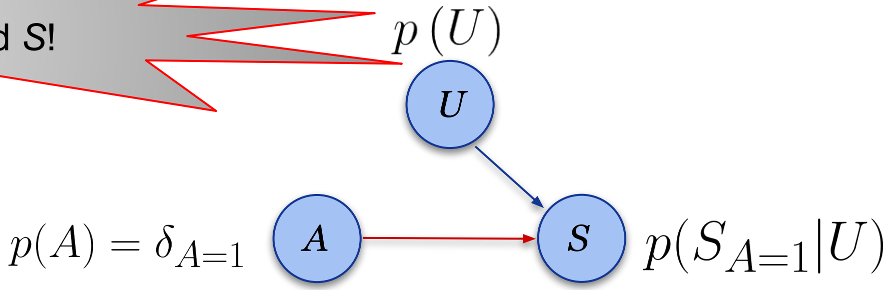
$$\mathbb{E}[S_{A=1}]$$
$$= \sum_u \mathbb{E}[S|A=1, U=u]p(U=u)$$

Conditional Distribution (Correlation)　　　　Causal Effect (Causation)

Have confounding between *A* and *S*!



$\neq$



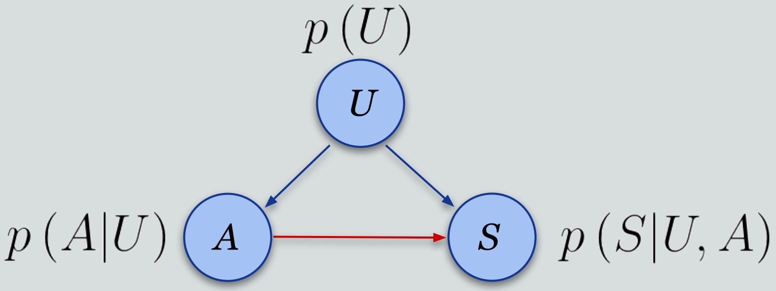
$$\mathbb{E}[S|A=1] = \sum_u \mathbb{E}[S|A=1, U=u]p(U=u|A=1)$$

In our data, *U=1|A=1* was greatly overrepresented

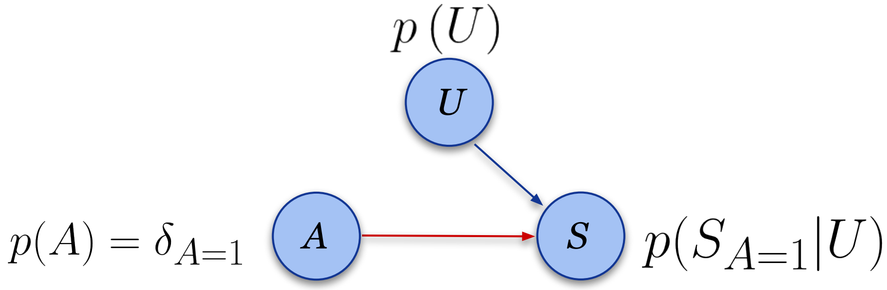
$$\mathbb{E}[S_{A=1}] = \sum_u \mathbb{E}[S|A=1, U=u]p(U=u)$$

- Confounding: a common cause of *A* and *S*
- If we see *A* and *S* correlate in the data, don't know whether
  - It was caused directly (red arrow)
  - Indirectly (through mutual correlation with *U*)

First solution: estimate causal effect from data



$\neq$



$$\mathbb{E}[S|A=1]$$
$$= \sum_u \mathbb{E}[S|A=1, U=u]p(U=u|A=1)$$
$$\approx \frac{\sum_{i=1}^n 1_{\{a_i=1\}} s_i}{\#_{\{a_i=1\}}}$$

$$\mathbb{E}[S_{A=1}]$$
$$= \sum_u \mathbb{E}[S|A=1, U=u]p(U=u)$$
$$= \sum_u \mathbb{E}\left[\frac{1_{\{A=1\}} S}{p(A=1|U=u)}\right] p(U=u)$$
$$\approx \frac{1}{n} \sum_{i=1}^n \frac{1_{\{a_i=1\}} s_i}{p(A=1|U=u_i)}$$

Inverse Propensity Weight (IPW) estimator

Observational data

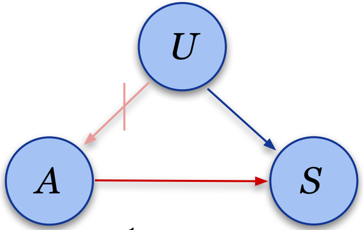
Second solution: randomized control trial (experimentation)

- Alter the environment to break the correlation between *U* and *A*
- Replace $p(A|U)$ with a coin flip
- This is why experimentation works

$$ATE = \mathbb{E}[S_{RCT}|A=1] - \mathbb{E}[S_{RCT}|A=0]$$
$$\approx \frac{\sum_{i=1}^{n} 1_{\{a_i=1\}} s_i}{\#_{\{a_i=1\}}} - \frac{\sum_{i=1}^{n} 1_{\{a_i=0\}} s_i}{\#_{\{a_i=0\}}}$$

Data collected by RCT

$p(U)$

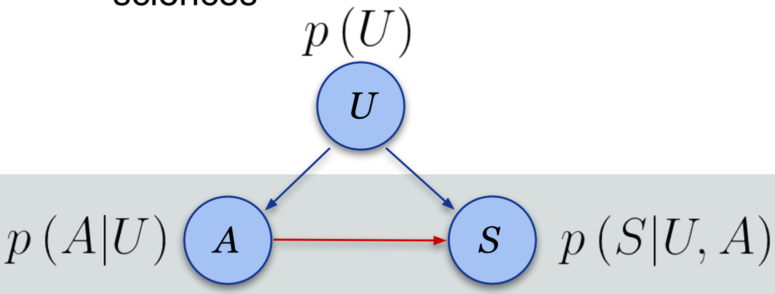


$p_{RCT}(A=1) = \frac{1}{2}$

$p(S_{RCT}|A, U)$

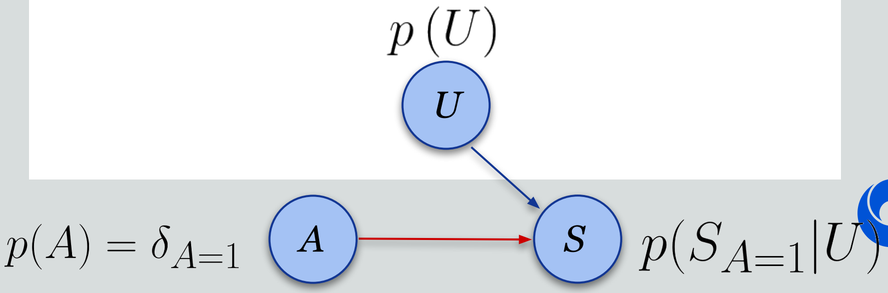
Causal Inference

- Observational data is easy to collect
  - No additional infrastructure
  - Experiments can be impossible/unethical
- Often requires strong assumptions on the causal model
  - Ignorability: $S_{A=a} \perp\!\!\!\perp A|U$
  - *U* blocks “backdoor paths”
- Cannot learn causal model from observational data
- Communities: econometrics, social sciences



Experimentation

- Experiments are costly
  - Requires infrastructure
  - Expensive (opportunity cost)
  - Easy to abuse
- No assumptions on causal model: we break the correlation through intervention
- Handles *unobserved* confounders
- RCT: “gold standard” in establishing causation

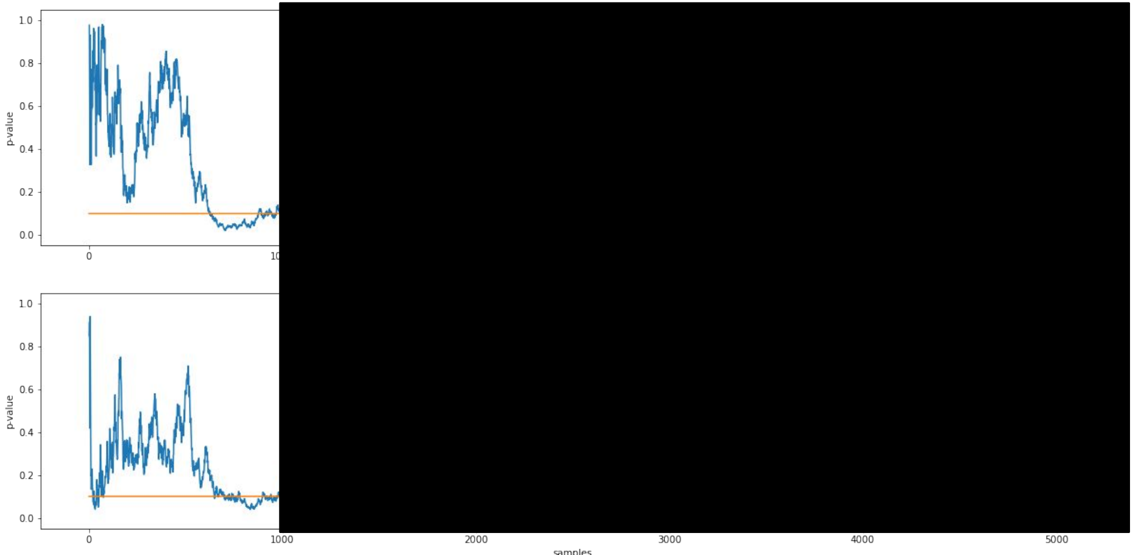


2

3 Pitfalls

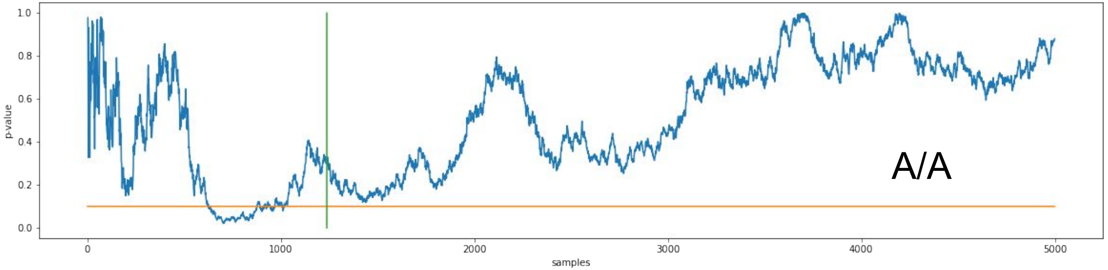
Pitfall I: peeking

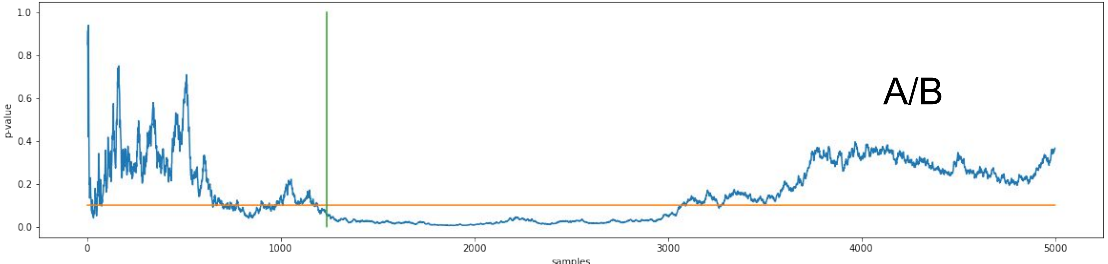
- Peeking: looking at the test results multiple times
- The t-test is a fixed-sample-size test
  - False positives (finding a difference when there is none) are only controlled for a *single* view of the data
  - Misconception: a “more significant test” (where the effect is much smaller than the MDE) allows you to stop early
- Pop quiz: Below is one A/A and one A/B test. Can you tell them apart?



Pitfall I: peeking

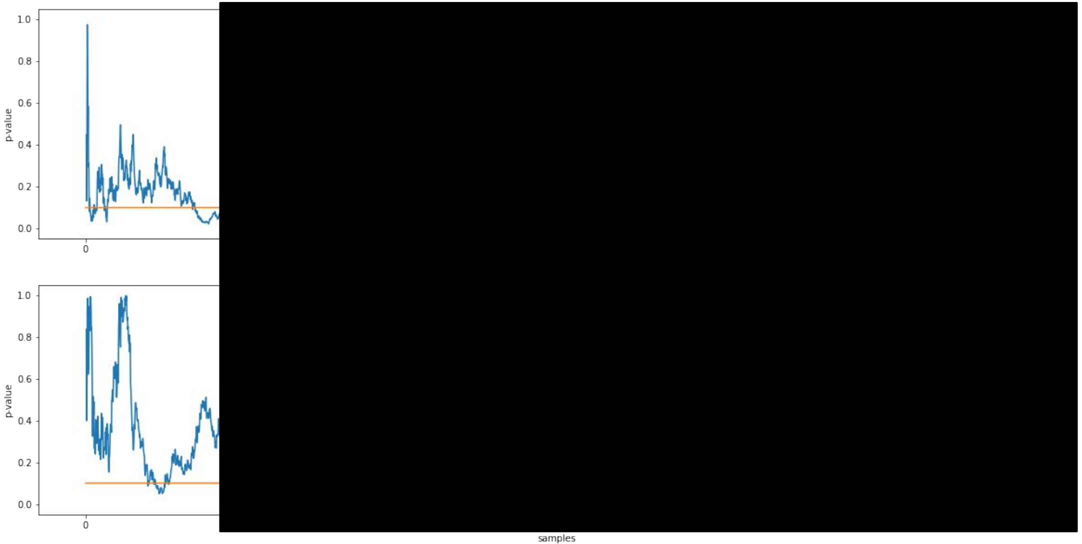
- Peeking: looking at the test results multiple times
- The t-test is a fixed-sample-size test
    - False positives (finding a difference when there is none) are only controlled for a *single* view of the data
    - Misconception: a “more significant test” (where the effect is much smaller than the MDE) allows you to stop early
- Pop quiz: Below is one A/A and one A/B test. Can you tell them apart?





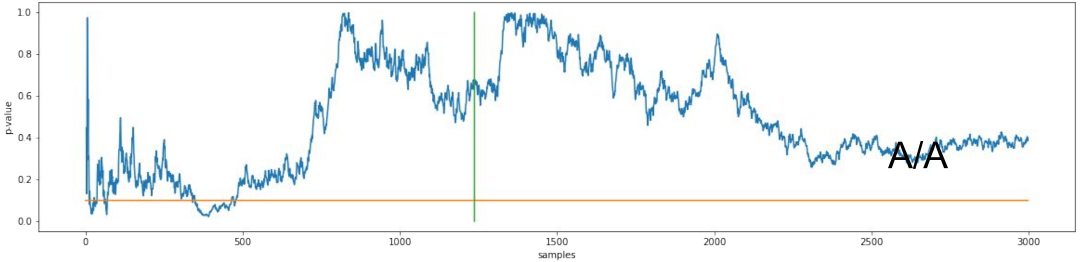
Pitfall I: peeking

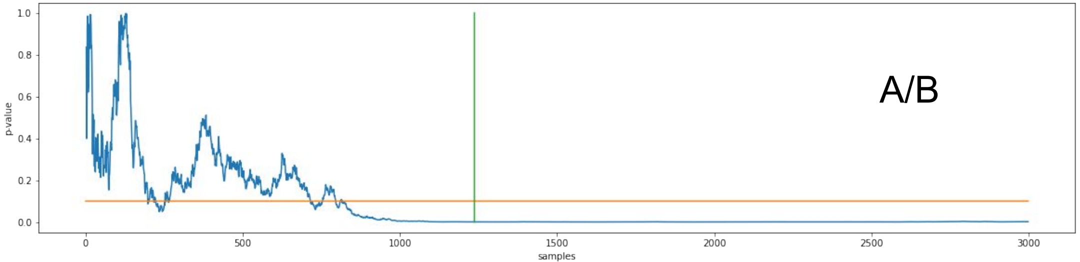
- The t-test is a fixed-sample-size test
  - False positives (finding a difference when there is none) are only controlled for a *single* view of the data
  - Misconception: a “more significant test” (where the effect is much smaller than the MDE) allows you to stop early
- Pop quiz: Below is one A/A and one A/B test. Can you tell them apart?



Pitfall I: peeking

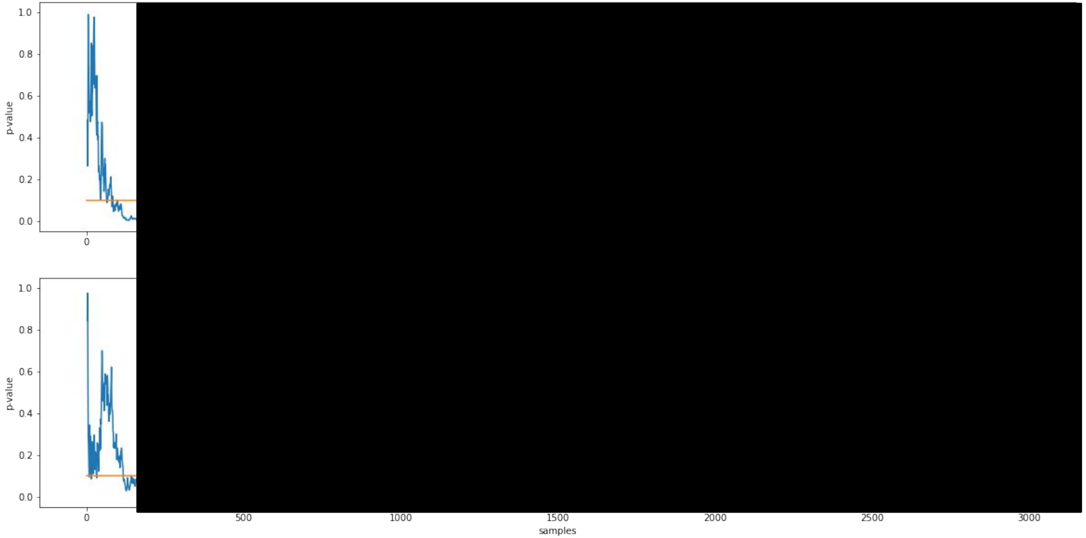
- The t-test is a fixed-sample-size test
  - False positives (finding a difference when there is none) are only controlled for a *single* view of the data
  - Misconception: a “more significant test” (where the effect is much smaller than the MDE) allows you to stop early
- Pop quiz: Below is one A/A and one A/B test. Can you tell them apart?





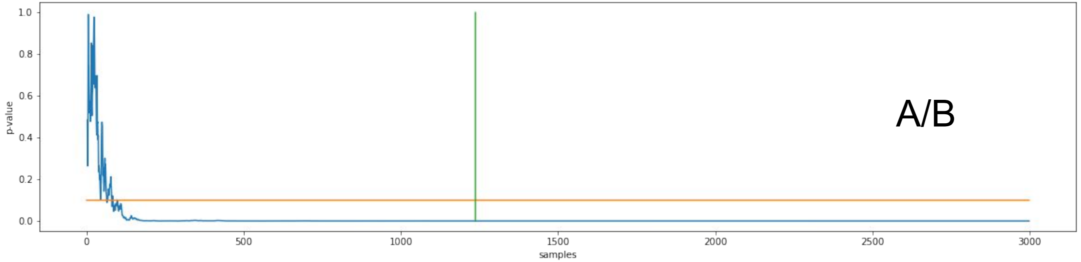
Pitfall I: peeking

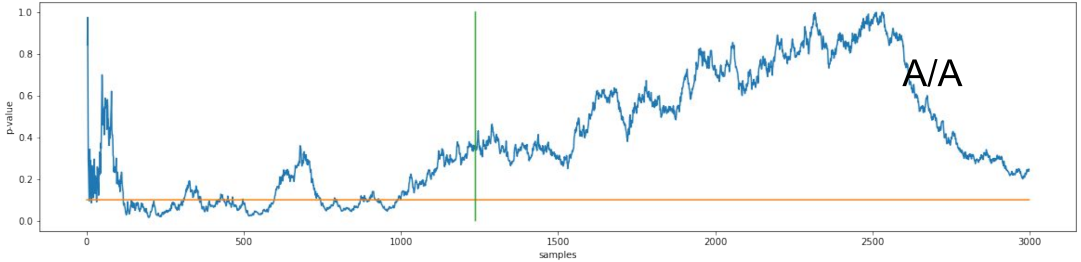
- The t-test is a fixed-sample-size test
  - False positives (finding a difference when there is none) are only controlled for a *single* view of the data
  - Misconception: a “more significant test” (where the effect is much smaller than the MDE) allows you to stop early
- Pop quiz: Below is one A/A and one A/B test. Can you tell them apart?



Pitfall I: peeking

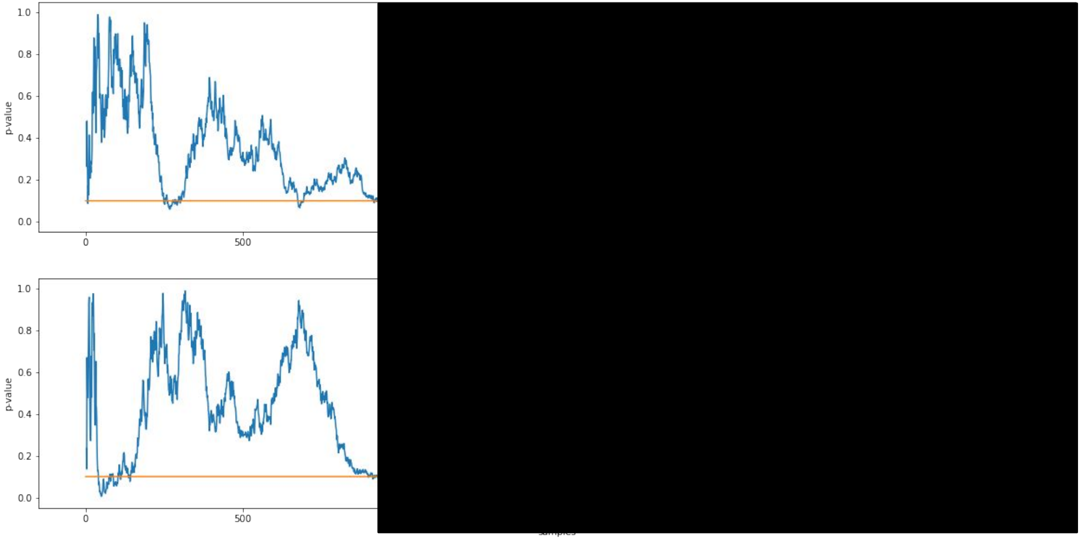
- The t-test is a fixed-sample-size test
    - False positives (finding a difference when there is none) are only controlled for a *single* view of the data
    - Misconception: a “more significant test” (where the effect is much smaller than the MDE) allows you to stop early
- Pop quiz: Below is one A/A and one A/B test. Can you tell them apart?





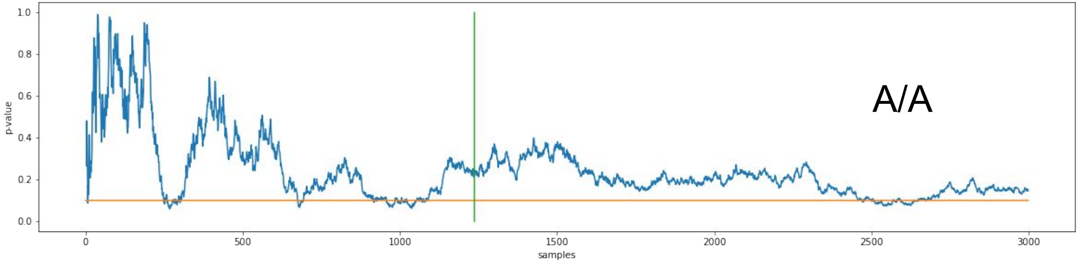
Pitfall I: peeking

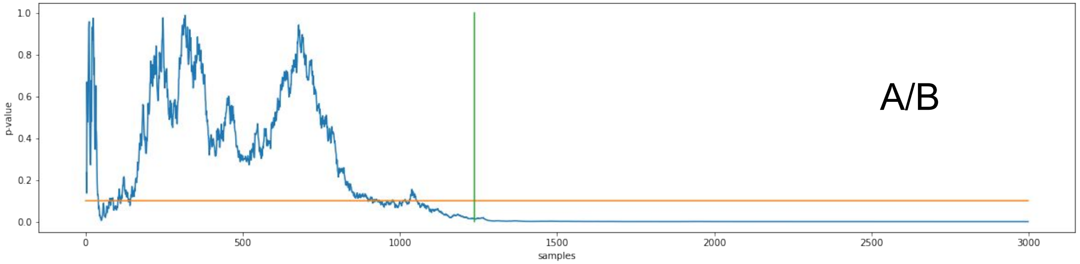
- The t-test is a fixed-sample-size test
  - False positives (finding a difference when there is none) are only controlled for a *single* view of the data
  - Misconception: a “more significant test” (where the effect is much smaller than the MDE) allows you to stop early
- Pop quiz: Below is one A/A and one A/B test. Can you tell them apart?



Pitfall I: peeking

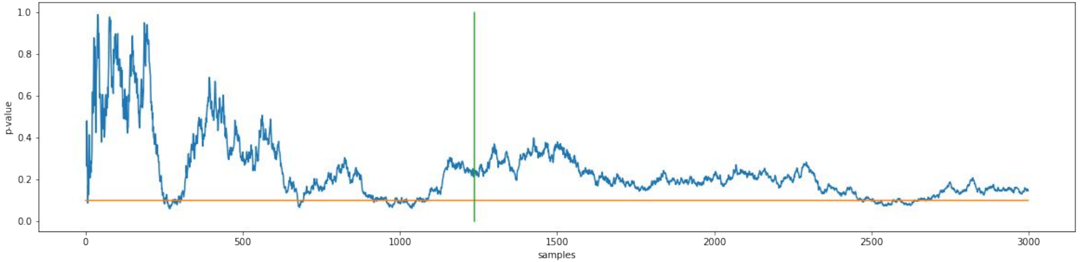
- The t-test is a fixed-sample-size test
  - False positives (finding a difference when there is none) are only controlled for a *single* view of the data
  - Misconception: a “more significant test” (where the effect is much smaller than the MDE) allows you to stop early
- Pop quiz: Below is one A/A and one A/B test. Can you tell them apart?

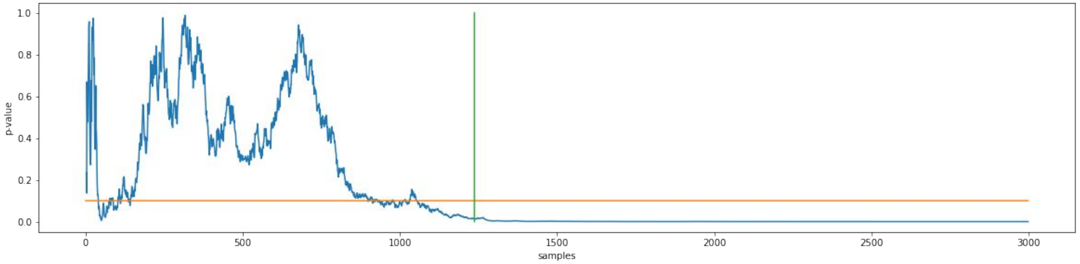




Pitfall I: peeking

- Conclusion: stopping early can really blow up your false positive rate
- Either use sequential methods, or don’t ignore your sample size calculator
- Examples weren’t that contrived (took the most egregious 4 out of the first 20)
- Code after the end; try it yourself!





Pitfall II: Not correcting for multiplicity

- Need to adjust α when running multiple hypotheses
- Examples:
  - A/B/n tests
  - Looking at sub-populations/segments of the data
- Ways to adjust: Bonferroni, False Discovery Rate (FDR)
- Significant test ⇏ significant result on sub-population
  - OK: using sub-population data to form a hypothesis test which becomes the subject of a follow up experiment
  - Not ok: concluding anything statistical

Pitfall III: using the wrong paradigm

- Multi-armed bandits:
  - Have multiple options, want to funnel users to the best performing one
  - Objective: most users to best option, quickly
  - No Type I error guarantees, but can guarantee low regret
  - E.g. which headline to show on today's front page?

- When hypothesis testing appropriate?
  - When you really need false positive control
  - Results used to decide on long-term changes
  - Results used to steer development / future testing efforts
  - E.g. should we invest more in better descriptions or better pictures

- When are multi-Armed bandits appropriate?
  - When knowledge of the best option has little effect on future decisions
  - There is lots of temporal variation / change in actions
  - E.g. population distribution today and tomorrow are different

DeepMind

The end and thank you

Code

```
import numpy as np
import scipy
import matplotlib.pyplot as plt
from statsmodels.stats.power import tt_ind_solve_power

n = 3000
min_sample_size = tt_ind_solve_power(effect_size=.1, alpha=0.1, power=0.8, ratio=1)
c_samples = np.random.normal(loc=0,scale=cov, size=(n,))
c2_samples = np.random.normal(loc=0,scale=cov, size=(n,))
t_samples = np.random.normal(loc=.1, scale=cov, size=(n,))
AA_p_values = [scipy.stats.ttest_ind(c2_samples[:pos], c_samples[:pos]).pvalue for pos in range(n)]
AB_p_values = [scipy.stats.ttest_ind(t_samples[:pos], c_samples[:pos]).pvalue for pos in range(n)]
fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(20, 10)
ax1.plot(AA_values)
ax1.plot([.1]* n)
ax1.plot([min_sample_size] * n, np.linspace(0,1,n))
ax2.plot(AB_values)
ax2.plot([.1]* n)
ax2.plot([min_sample_size] * n, np.linspace(0,1,n))
```