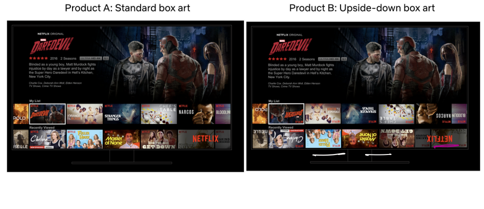
A/B testing

Thursday, 17 August 2023 19:53

Lecture Objective:

- Business sense towards A/B testing.
- How to launch A/B testing to get "valid" and "reliable" results.
  - Developing hypothesis
  - Designing A/B tests
  - Evaluating test results
  - Making decisions
  - Launch recommendations and pitfalls with A/B testing set ups.

Example Cases: A/B testing Use Cases



500 (15 to 30 days)

100m → 50009 → objective

Cost ↓ → Strategy

→ certain Group

Before (10 min) After (15 min)

Group 1 – old, Group 2 – New → [Random] (Active users) Geo

Population → filter → Population with filter → sample → A, B (RS) → Control [Normal], Test [upside down] — 30 days

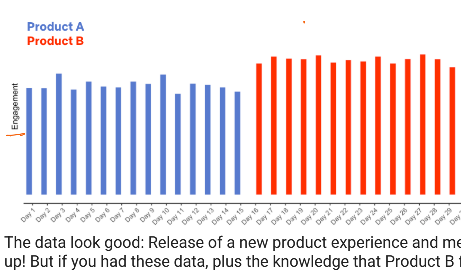
sample 1, sample 2; Group 40 to 60 → A → 40-60 1, 40-60 2; Group 80-100 B

What is an A/B Test?

An A/B test is a simple controlled experiment. Let's say — (Not ever close to reality) Let say we want to learn if a new product experience that flips all of the boxart upside down in the TV UI is good for Netflix members.

Without A/B Test

Let's say Experiment was implemented , and flipped the switch to the Upside Down experience on the 16th day of a month. How would you act if we gathered the following data



The data look good: Release of a new product experience and member engagement goes way up! But if you had these data, plus the knowledge that Product B flips all the box art in the UI upside down, how confident would you be that the new product experience really is good for our members?

Can we really know that the new product experience is what caused the increase in engagement? What other explanations are possible?

What if you also knew that Netflix released a hit title, like a new season of Stranger Things on the same day as the (hypothetical) roll out of the new Upside Down product experience?

The key point is that we don't know if the new product experience caused the increase in engagement

What if instead we'd run an A/B test with the Upside Down box art product experience, with one group of members receiving the current product ("A") and another group the Upside Down product ("B") over the entire month

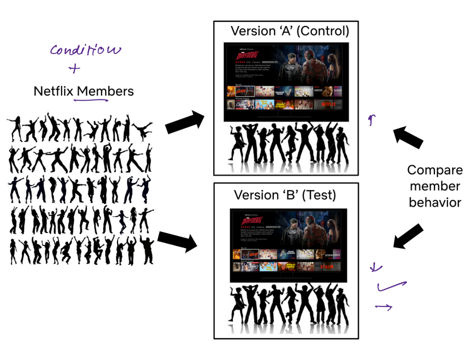
To run the experiment, Lets take a subset of members, usually a simple random sample, and then use random assignment to evenly split that sample into two groups.

- Group "A," often called the "control group," continues to receive the base Netflix UI experience,
- Group "B," often called the "treatment group", receives a different experience. Here, Group B receives the Upside Down box art.

Then compare the values of a variety of metrics from Group A to those from Group B. Some metrics will be specific to the given hypothesis.

- For a UI experiment, we can look at engagement with different variants of the new feature.
- For an experiment that aims to deliver more relevant results in the search experience, we can measure if members are finding more things to watch through search.
- In other types of experiments, we might focus on more technical metrics, such as the time it takes the app to load, or the quality of video we are able to provide under different network conditions.

Condition +



walmart

coke → lays ?

coke → water

coke! ↓ coke! Basket Analysis

→ wines ↑ sales ↳ A/B ↓ fruit section → Grapes

4P's: Product, Price → ↑ sales ↓ Profit Margin; Place → Anywhere in store ?; Promotion → ↑ sales ↑ cost ↓ Profits; discount (cost)

US ↓ walmart → similar stores ↳ No of Product ↳ Demographics ↳ Sales History ← City

2 similar store Before / After Together

A ↓ wines Alcohol — B ↓ fruit

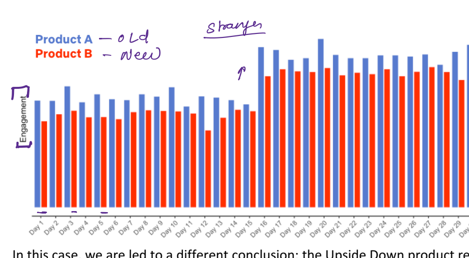
N 100 → D 120 ← +2.5% ↑ ↓ 150 → 220 ? (sales)

2 things FP, FN ↓ 400 → Phase 1 → 10, 2 → 100, 3 → 1000, 4 → 4000

Holding everything else constant

- Because there are control ("A") and treatment ("B") groups using random assignment, individuals in the two groups are, on average, balanced on all dimensions that may be meaningful to the test.
- Random assignment ensures, for example, that the average length of Netflix membership is not markedly different between the control and treatment groups, nor are content preferences, primary language selections, and so forth.
- The only remaining difference between the groups is the new experience we are testing, ensuring our estimate of the impact of the new experience is not biased in any way.

A/B test with the Upside Down box art product experience, with one group of members receiving the current product ("A") and another group the Upside Down product ("B") over the entire month, and gathered the following data:



In this case, we are led to a different conclusion: the Upside Down product results in generally lower engagement (not surprisingly!), and both groups see an increase in engagement concurrent with the launch of the big title

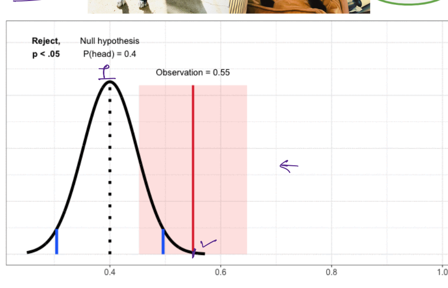
A/B tests let us make causal statements. We've introduced the Upside Down product experience to Group B only, and because we've randomly assigned members to groups A and B, everything else is held constant between the two groups. We can therefore conclude with high probability that the Upside Down product caused the reduction in engagement

Using a framework based on hypothesis generation, A/B testing, and statistical analysis allows us to carefully quantify uncertainties, and understand the probabilities of making different types of mistakes.

There are two types of mistakes we can make in acting on test results.

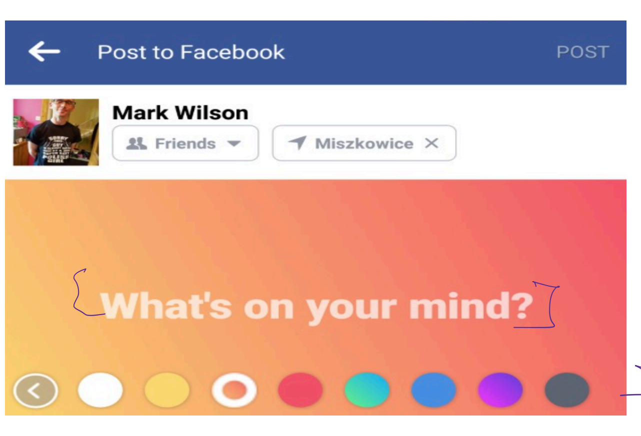
A false positive* (also called a Type I error)
occurs when the data from the test indicates a meaningful difference between the control and treatment experiences, but in truth there is no difference. This scenario is like having a medical test come back as positive for a disease when you are healthy

A false negative (also called a Type II error),
occurs when the data do not indicate a meaningful difference between treatment and control, but in truth there is a difference.



Case Study:

How would you test if Facebook incorporating colored backgrounds to statuses improves user engagement?



1. Clarify the goal
↓
2. Discuss the Metric
3. Experimentation

Objective ? Purpose → 1. Really Required 2. Metric → Measure

→ Understand goal

1. If any other changes are in progress
2. which group are we target ↳ Population ↓ Population - filter

2. Define the Metric
→ Engagement → daily Active user
→ Active user ↓ Engagement % ↑ ↳ Likes ↳ shares etc

3. Hypothesis
Null 1. No significant change in engagement
Alternate → significant increase in engagement

4. Test :- 2 Groups (Ratio) ↓ Z - propotion test

5. find out Control group / Test group
P ↓ P - filter out (Target group) + web / App ↓ Random sample ← size; RS 1, RS 2

Sample size ?

EDA → Engagement - 45%

significance level — 5%

95% ↓ Mean + σ + 2σ 2 x std

50,000 > 100, too small

→ Duration of Experiment ↑ (15) 1m, 2m, 3m ↓ 100 TP ↓ 1L

Economies of scale ↓ Economies of speed

50 K ↳ 5000 user (Avg) 10 days 5 days

→ Primacy / Novelty ↓ → ↳ Recurrency → 2 week Increase our Exp duration

Novelty ↳ Excited (Threads)

1 week → increase during

→ Type of change

Group effect ↳ Adding Music to stories

Network — Clubhouse — Live discussion ↓ (By invite only)

5. Recomendation

Lift of 0.1% → $2m ($250B) ↓ Sample

→ Advantages

1. Cost effective
2. Less Risky
3. No user troubles

→ Test → MCQ (Notes)

→ Business Case → 1 session

Disadvantage → Dont over do it !

→ No interview

Guestimates ↓ saturday (Next sunday)

2 case studies ↓ 1 class (sunday) ? + 1 extra — Coming ↳ 10-12, 12-2, [2-4], 4-6