USC
School of Business

Data Driven Decision Making: A/B Hypothesis Testing

GSBA 545, Fall 2021

Professor Dawn Porter

A/B Hypothesis Testing

- Definition and Purpose
- Examples
  - Highrise Signup Page
  - Amazon’s Pre-Checkout Screen
- Metrics to Test
- Hypothesis Testing Calculation

What is A/B Testing?

Procedure for deciding which of two alternatives (“A” or “B”) is “better”

- Can also be used with more than two alternatives (covered in ANOVA session)
- Based on two-sample hypothesis testing

General Procedure:

- Randomly assign some number of cases to each of the options
- Track the outcomes, compare key metrics, and choose the “better” one

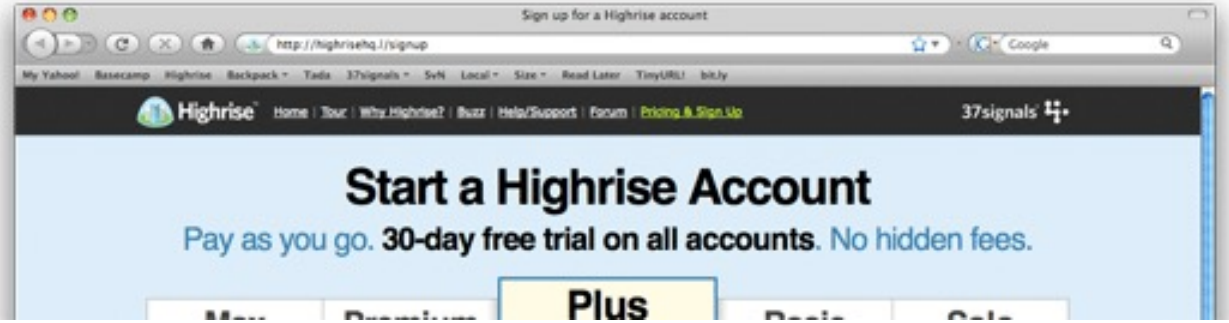
A/B testing is indispensable in many industries

- Simple and intuitive
- Possible wherever experimentation is cheap and large volumes data

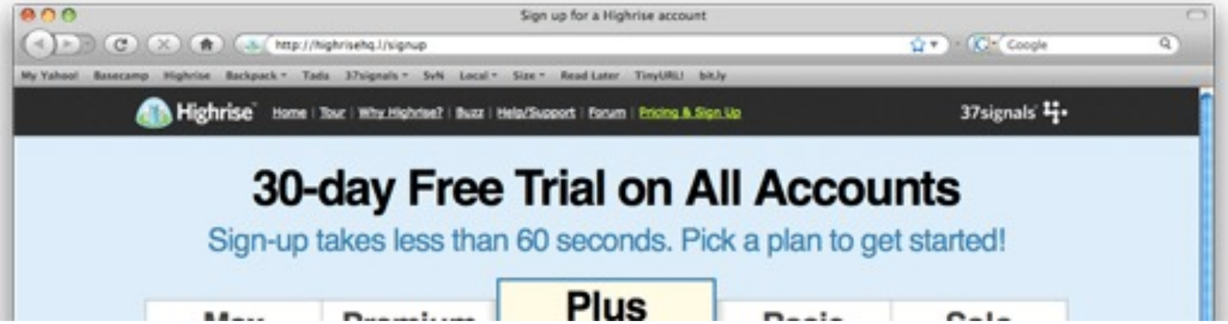
Wired article: Inside the Technology that's Changing the Rules of Business

Ex: Increasing Customer Sign-ups

Original



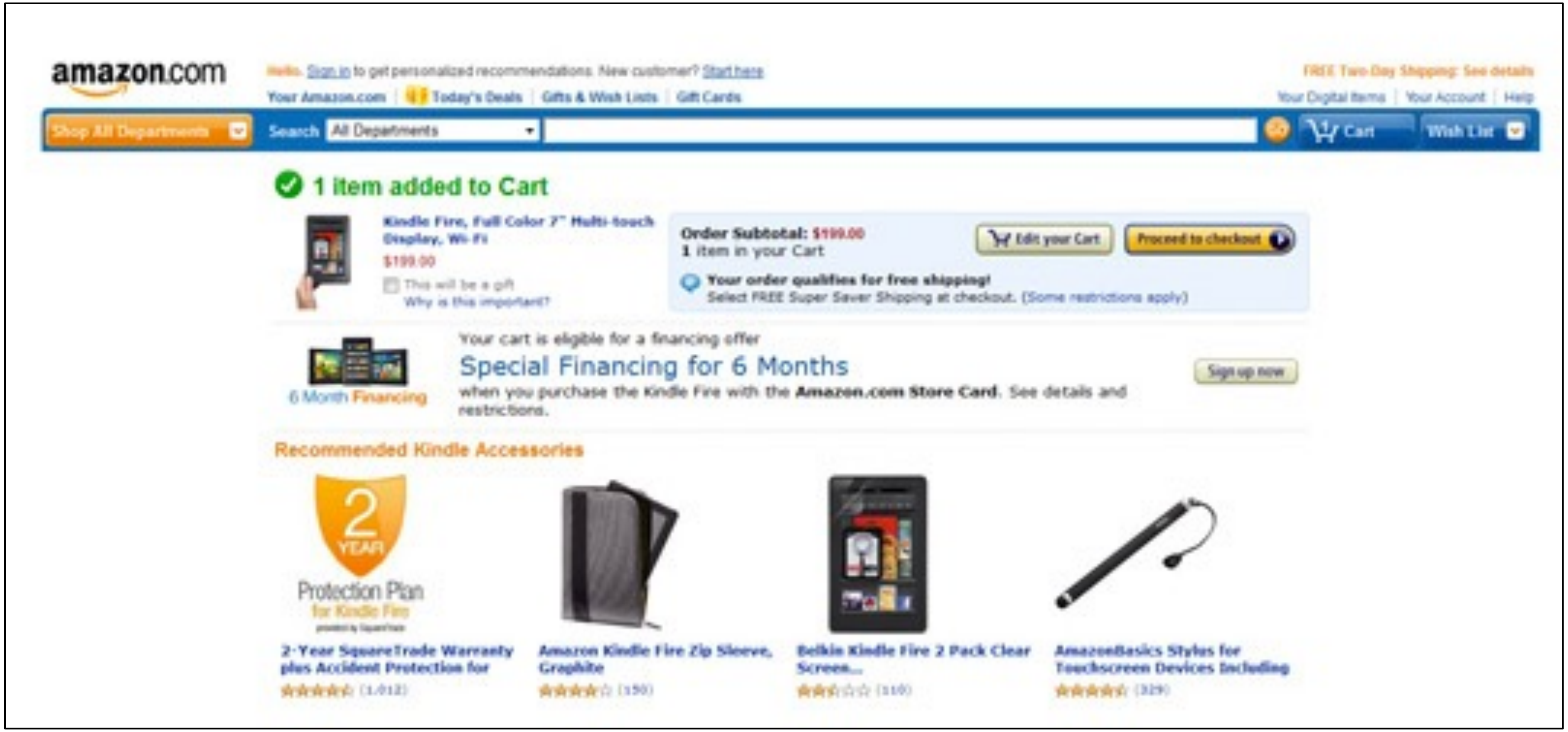
Alternate



+30%

* https://signalvnoise.com/posts/1525-writing-decisions-headline-tests-on-the-highrise-signup-page

Ex: Amazon's Pre-Checkout Screen

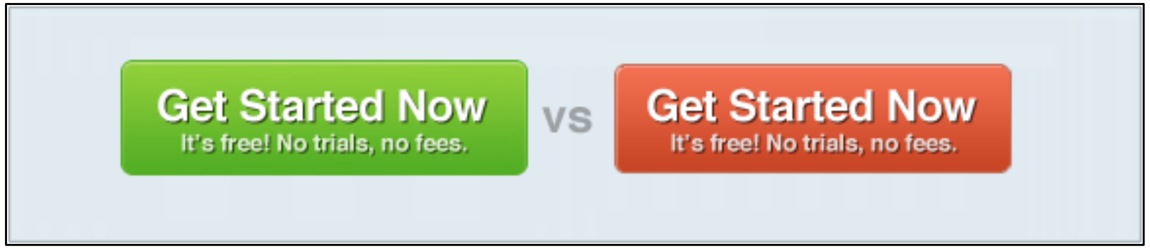


What should we be measuring?

Defining "better" is not always easy

- What if alternative "A" is better on one metric and worse on another?

Example: Deciding between two different sign-up buttons



What should we measure:

- Click rate? (what percentage of people click each button?)
- Purchase rate? (what percentage of people actually buy a product?)
- Purchase amount? (If you buy a product, how much do you spend?)

Intuitions for A/B Testing

Mean performance helps understand which is better.

The higher the variability, the less confident we are in the result.

The more data we have, the more confident we are in the result.

Comparing Two Groups

To compare Alternative (Method 2) to Default (Method 1), compute:

Test statistic: $$t = \frac{(\bar{x}_1 - \bar{x}_2 - D_0)}{se(\bar{x}_1 - \bar{x}_2)}$$

Bigger t → the more confident you are that 2 is better than 1.

Suppose you pick Method 2: *p*-value gives you the probability that you are wrong.

- Small *p*-values suggest that Method 2 really is better than Method 1
- Large *p*-values suggest Method 1 is just as good or better.

Confounding/Lurking Variables

Sometimes customers have subgroups that might respond differently to the two variants.

Ex: Green Button vs. Red Button

Men prefer Green Button. Click through about 2% more often.

Women dislike Green Button. Click through about 2% less often.

Overall, two effects wash-out, looks like there's no effect.

Even worse: Sometimes the effect could be exactly the opposite.

Bias in Berkeley Graduate admissions: https://setosa.io/simpsons/

What is going on here???

If possible, split the data by subpopulation and do *separate* analyses.