PAPER: RON KOHAVI ET AL.

TALK: NEIL MENNE

TRUSTWORTHY ONLINE CONTROLLED EXPERIMENTS

BACKGROUND AND TERMINOLOGY

KEY TERMINOLOGY

- Overall Evaluation Criterion (OEC): quantitative measure of the experiment's objective.
- Experimental Unit: the entity randomly assigned to the control or treatment
- Null Hypothesis: the hypothesis that the OECs for the variants are not different and that any observed differences during the experiment are due to random fluctuations
- Confidence Level: the probability that we will fail to reject the null hypothesis if it is true
- Power: the probability of correctly rejecting the null hypothesis
- Primacy: changes negatively affect the productivity of experienced users
- Novelty: changes are interesting enough that users return for the experience

BING ME WHEN YOU GET A CHANCE

THE MYSTERIOUS CASE OF THE UNUSABLE SEARCH ENGINE

WHEN BING HAD A BUG IN AN EXPERIMENT, WHICH RESULTED IN VERY POOR RESULTS BEING SHOWN TO USERS, TWO KEY ORGANIZATIONAL METRICS IMPROVED SIGNIFICANTLY . . .

PICKING A GOOD OEC MAKES ALL THE DIFFERENCE

- Choosing OECs that were in direct conflict with the user's goal led to "positive" results
- OECs that focus on longer term objectives would not have been as susceptible to the bug

KNOW THY USERS…

A STUDY IN CLICK TRACKING

THIS SLOWED DOWN THE USER EXPERIENCE SLIGHTLY, YET THE EXPERIMENT SHOWED USERS WERE CLICKING MORE!

… AND THEIR IDIOSYNCRASIES

- One page apps don’t have users navigating away from the page that wants to capture an event
- Open in new tab can be useful in mitigating this as well
- Understanding that sometimes the effect can be *too good*

WHAT'S TRENDING?

LOOKING FOR PATTERNS WHERE THEY DON'T EXIST

MY FEATURE IS OBVIOUSLY GREAT, BUT IT JUST TAKES TIME FOR USERS TO GET USED TO IT.

UNDERSTAND HOW EXPERIMENTS PLAY OUT

- Initial reactions to a feature are based on smaller sets of users
- We tend to see trends where there are none
- Confirmation bias is real

POWER OVER TIME

WAITING FOR THE WORLD TO CHANGE

FOR SOME OF OUR KEY METRICS . . . THE CONFIDENCE INTERVAL OF THE PERCENT EFFECT DOES NOT SHRINK OVER TIME.

INDEPENDENCE MIGHT ONLY EXIST ON PAPER

- OECs based on users can vary per user
- OECs can vary for the same user over time
- When the variance is too high, more time won’t cut it.

DEALING WITH A BAD CARRYOVER

JUST A DISGRUNTLED DROP IN A BUCKET

AN EXPERIMENT RAN AND THE RESULTS WERE VERY SURPRISING...METRICS UNRELATED TO THE CHANGE MOVED...AND THE EFFECTS WERE HIGHLY STATISTICALLY SIGNIFICANT.

POSITIVE AND NEGATIVE CARRYOVER EFFECTS EXIST

- Tracking users before and after an experiment can be used to identify when it's happening
- Independently considering users for any experiment they're in mitigates the carryover effect

QUESTIONS?