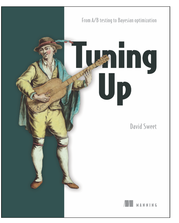
Practical matters in A/B testing

Selected material from:
Tuning up: From A/B testing to Bayesian Optimization
Manning Publications, 2021 (summer, estimated)

David Sweet
linkedin.com/in/dsweet99/
@phinance99



Audience

- ML/AI engineers
- Quantiative traders, "quants"
- Software engineers

A/B Test

- A: The current system
- B: A good idea, meant to improve the system
- Test: An experiment

Good ideas aren't that good

- How many experiments improve metrics?
  - Amazon: 50%
  - Microsoft: 33.3%
  - Netflix: 10%

Engineer's workflow

A/B test basics

- Randomization
- Replication
- Limit false positives (5%) and false negatives (20%)
  - ex: one A/B test every two weeks for a year, 33.3% accepted, <1 f.p., ~2 f.n.

Holdout test

- Many A/B tests over 6 months
- Holdout
  - A: System at start of 6 months
  - B: System at end of 6 months
- Net improvement < sum of individual improvments
  - 5% f.p, nonstationarity

Business Metric

- Immediate reward
  - Click-through rate
  - Markout profit
  - Engagement: like, retweet, comment, skip song, etc.
- Daily aggregates
  - Revenue, pnl, trading volume
  - Time spent on app, number of songs streamed
  - Active users
- Long-term
  - Monthly active users
  - Pnl/trade with multi-day hold time
  - Will an ad view lead to a purchase later?
  - User activity over next D days

Multiple business metrics

- Don't usually care about just one
- Maybe trade off: more revenue, less time spent
- Maybe "guardrail": higher CTR, but only if revenue and engagement don't drop

Deciding to accept or reject

- Acceptance / rejection a group discussion
- Higher stakes ==> larger discussion
- Sanity check surprising/dramatic results
  - Could there be an error in the experiment?
  - Did you learn something new? Dig deeper to understand
- Carefully weigh tradeoffs of multiple metrics

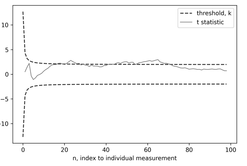
Running an A/B test

- Risks
  - deploying bugs
  - reducing business metrics
  - wasting time

Running an A/B test

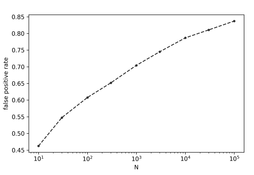
- A/A test, small: Does experimentation alon code change metrics?
- A/B test, small: Does new code have bugs or dramatically change metrics?
- A/B test, large: The full experiment

Early stopping



- Idea: "To save time, if t-stat crosses threshold, I'll stop" **NO**
- Can generate false positives

Early stopping



- False positive rate can be *very* high
- Much higher than 5%, for which A/B test is (usually) designed

p - hacking

- "cherry-picking"
- 5% f.p. is 1/20
- p-Hack: Run an experiment 20 times
- p-Hack: Run an experiment and examine 20 metrics

Transient effects

- Short-lived, goes away
- Ex: Users engage with your new feature b/c it's novel, then abandon it
- Fix: Drop first K samples or days of data

- learn K for your system by running many different experiments

Tuning up: From A/B testing to Bayesian Optimization
Manning Publications, 2021 (summer, estimated)

David Sweet
linkedin.com/in/dsweet99/
@phinance99