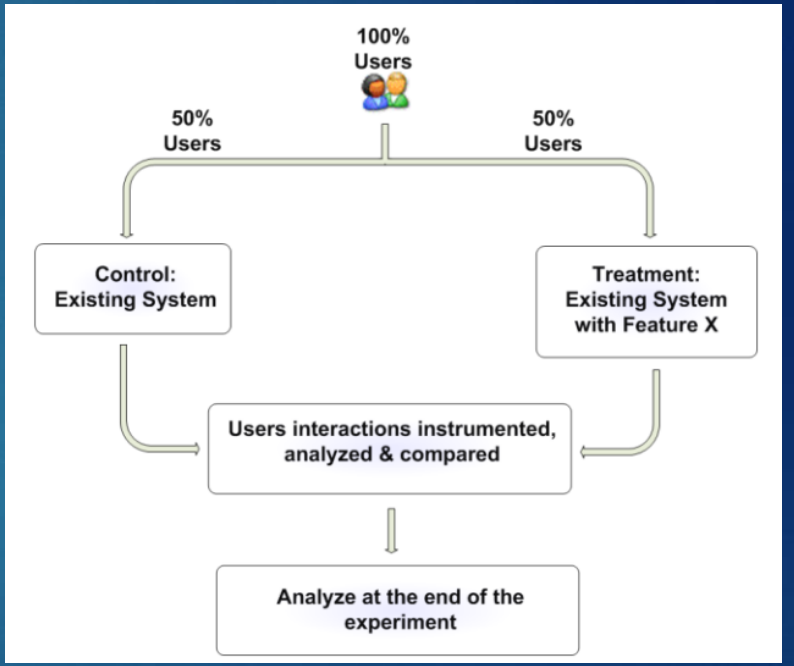
A/B Testing for the Next Decade

CHALLENGES, COMPETITIONS AND OPPORTUNITIES

ALEX DENG @ MICROSOFT 2015

A/B/n Tests in One Slide

- Randomly split traffic between two (or more) versions
  - A (Control)
  - B (Treatment(s))
- Collect data and analyze, typically in a form of two -sample hypothesis testing of the shift in mean/percentile



- Online Controlled Experiment
- Best scientific way to establish causality
  - Observational data analyses hard and error prone
  - Conceptually simple -> powerful and robust/"model free"

Five Challenging Problems

- Ronny Kohavi summarized 5 challenging problems in his recent follow-up talk after KDD and codecon@MIT(last week)
- Five challenges (paraphrase)
  1. Metric sensitivity (**)
  2. Problems with NHST and p-value (**)
  3. Beyond Population Average Treatment Effect (*)
  4. Novelty Effects and Experiment Duration
  5. Leakage/Violations of SUTVA(stable unit treatment value assumption)
- Will talk about (**), and (*) if time permitted

Challenge 2: NHST and p-value

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

Archive > Volume 519 > Issue 7541 > Research Highlights: Social Selection > Article

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

PDF | Rights & Permissions

A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research[1].

Authors are still free to submit papers to *BASP* with *P* values and other statistical measures that form part of 'null hypothesis significance testing' (NHST), but the numbers will be removed before publication. Nerisa Dozo, a PhD student in psychology at the University of Queensland in Brisbane, Australia, tweeted:

- Many published research findings found not reproducible.
  - Notable/Surprising results even more so

  "The fluctuating female vote: Politics, religion, and the ovulatory cycle"
  - Many seemingly good results we observed won't survive confirmation run
- P-value hack
  - Multiple Testing: Team keep making minor changes to a feature and keep running A/B testings until they get small p-value
  - Optional stopping/continuous monitoring: stop the test once the p-value is "statistically significant"
- In ExP we end up requiring confirmation/certification run for any experiment that seems super successful. Also only treat p-value<0.01 more seriously

Problems of NHST

- Null and Alternative is asymmetric.
  - Test only try to reject null, and gather evidence against the null
  - Even with infinite data, will never accept the null with 100% confidence
- Multiple testing/Optional Stopping/Early stopping : because of the asymmetry, multiple testing/check-point can only favor the alternative.
- "Genuine" Prior information not used
  - Researchers motivated to publish counter-intuitive results, which are more often not reproducible
  - Twyman's law: any piece of data that looks interesting or unusual is probably wrong
- P-value often misunderstood: Goodman's a dirty dozen

Objective for A/B Testing

- Feature owner: want to know whether a feature is good or bad with high accuracy (ideally 0% Type-I and Type-II error)
- Organization:
  - There are always mistakes as long as there exist noises. 0% Type-I error -> 0% statistical power
  - In the **long run**, we want **majority** of features shipped are good for our users
- Long run could be a month in a company like Microsoft, where different teams conducting test in a weekly bases
- "Majority" should be quantified and controlled, is 51% enough?

Opportunity #1: Beauty of A/B testing at scale, we benefit from law of large number and don't necessarily need near perfect decision making

P(H0 | Data) and P(H1 | Data), not P(Data | H0)

As an organization, with A/B testing at scale, we should focus on P(H1 | Data)!

- If we only ship a feature if P(H1 | Data) > x% (P(H0 | Data) < 1-x%), then we know x% of the ship decisions are correct
- P(H1 | Data) is the Bayesian posterior belief of the alternative hypothesis, it is closely related to the concept of FDR(False discovery Rate)
- Many people misunderstood p-value as P(H0 | Data), and therefore treat 1-pvalue as "confidence" of a correct ship decision
  - A few commercial A/B testing tools use "confidence" instead of p-value
- P(H1 | Data) can only be reasoned within Bayesian framework, probability of H1 or H0 make no sense in frequentist NHST

Frequentist NHST vs Bayesian: Two Trial Systems

- Frequentist:
  - One group of jury, with presumption of innocence, reckoning evidence of the defendant being guilty
- Bayesian:
  - Two groups of jury, one reckon the evidence of being guilty, the other reckon the evidence of being innocent
  - Judge make final decision based on decisions of both jury, together with prior belief
- Benefit of two jury system
  - Symmetry: gather evidence for innocence, not just fail to convict
  - Balanced design: Think optional stopping, both two groups of jury will both benefit from multiple check-points and the judge can still make a balanced call

Bayesian Two Sample Hypothesis Testing

1. H0 and H1, with prior odds

$$PriorOdds = \frac{P(H1)}{P(H0)}$$

2. Given observations, likelihood ratio

$$LR = \frac{P(Data|H1)}{P(Data|H0)}$$

3. Bayes Rule

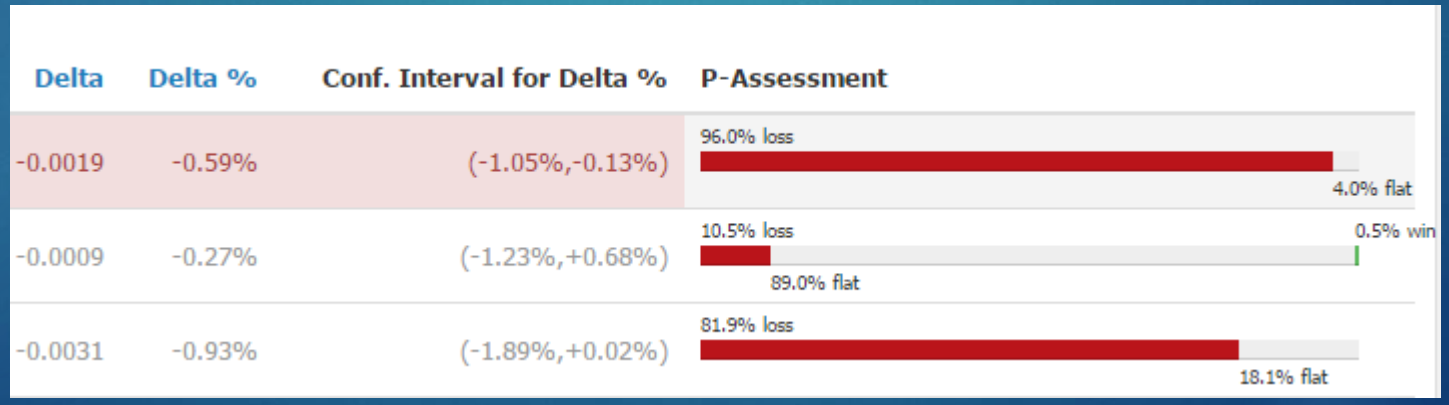
$$\frac{P(H1|Data)}{P(H0|Data)} = PriorOdds \times LR = \frac{P(H1)}{P(H0)} \times \frac{P(Data|H1)}{P(Data|H0)}$$

Bayesian Advantages

- More intuitive result
- Accepting the Null: ship based on no harm
- Supports optional stopping/early stopping
- Solves many(not all) multiple testing issues
- Useful Prior information
  - Prior should be learned objectively, not set subjectively!
- Meta Analysis: combine results from different studies
  - Very useful if you run same experiment multiple times
- P(Metric A moved | Observations of a set of metrics)

P-Assessment

- Bayesian Framework allows us to estimate posterior of negative, positive, and flat movement
- We call it P-Assessment: [PNeg, PFlat, PPos]
- Use PNeg and PFlat for feature shipping
- Use PFlat for shipping with no harm



Objective Prior Learning

- p = P(H1)
- Observe $\Delta_i \sim N(\mu_i, \sigma_i^2)$, where σ_i assumed to be known
- $\mu = 0$ if H0, $\mu_i \sim F$ if H1
- Problem: given a dataset of historical observations Δ_i, how can we estimate p and distribution F
- If we know ground truth of whether H1 or H0 for each historical experiment, we can get p easily
- For a parametric F (e.g. normal or doubly exponential), EM algorithm can help to fit p and F (Deng 2015, Johnstone and Silverman 2004)
  - EM: initialize a guess of p and parameters of F, calculate posterior P(H1 | Data) for each experiment, use this as soft label, iterate until converge
- For general nonparametric F, active research area

- Bayesian hypothesis testing has a long history, many articles online
- Most of them uses a subjective prior/conjugate prior or "non-informative" prior
- There is no prior that is non-informative prior (Lindley's paradox)
- Many advantages of Bayesian results subjective to knowing the true prior, e.g. multiple testing adjustment

Opportunity #2: Empirical Bayes procedure allows us to learn prior objectively, with enough historical data

- Active research areas: prior for a metric might depend on product area, types of experiments, e.g. UX, backend algo, etc. Also it might change over time

Optimizely Stats Engine

- Stats Engine is based on a frequentist method (SPRT)
- SPRT method reject based on Likelihood Ratio, there is no prior information used
- LR requires the knowledge of F, i.e. distribution of effect under alternative. EM algorithm here can be used
- Stats Engine still controlled Type-I error, not P(H1 | Data)
- SPRT requires observations being i.i.d., as in most sequential analysis method. Can be weaken but need formal proof (Likelihood Ratio need to be a martingale)

Challenge 1: Metric Sensitivity

- P(Detect a Movement) = P(Detect Movement | Movement) (1)
  × P(Movement) (2)
- Statistical Power mainly concerns (1)
- P(Movement) = P(H1) which can be learned objectively using EM algorithm
- P(Movement) is more fundamental. If your ideas don't work, don't expect ***trustworthy*** analysis to save you
- If your OEC don't move, understand which part is the bottleneck

Case 1: Statistical Power

- Increase traffic/"power up", limited to capacity
- Run longer, won't always work (Ronny's Sessions/UU example)
- Variance Reduction (CUPED/Doubly Robust Estimation)
  - improve your statistical test, same metric but pure power increase
  - $\Delta^* = \Delta_{\mathrm{AB}} - \theta^* \Delta_{\mathrm{AA}}$
  - Doubly Robust Estimation (with known propensity, just modeling counterfactuals)
- Transformation and capping of highly-skewed metrics
  - Transformation changed metric definition: how to reason about geometric mean?
  - Capping: interpret with caution

Case 2: P(Movement) is low

- A perfectly designed metric is not actionable if you can not move it
- What should I do?
  - Re-engineer your metric. You need different OEC at different stages of your product. DAU(daily active users) easy to move for a new product, but harder for matured sites
  - Session Success Rate and Time To First Success moves a lot more than Sessions/UU
  - Define a new surrogate metric as a function of metrics with higher P(Movement) and reasonably statistical power. Calibrate the function form so that the surrogate metric aligns with your OEC
    - Linear combination is easy to work with
    - Optimization problem: maximize metric sensitivity given constraint of alignment

Metric Calibration

- Metric Y is your target metric, but many other metrics X1, ..
- Define Y*:= f(X1, ..) such that
  - Movement of Y* align well with Y
  - Y* is more sensitive than Y
- Existence: CUPED is an example such Y* exist
- X: metrics that focus on
- Functional form of f need to be calibrated and optimized
- Schuth, et. al. calibrated interleaving metrics with A/B testing

Opportunity #3: Data driven metric development

Bing Results

Metric	P(H0)	P(H1)
X1	97.63%	2.37%
X2	99.80%	0.20%
X3	90.60%	9.40%
X4	98.77%	1.23%
X5	78.55%	21.45%
X6	97.41%	2.59%
X7	97.75%	2.25%
X8	35.50%	64.50%
X9	85.73%	14.27%
X10	98.35%	1.65%
X11	89.25%	10.75%
X12	81.02%	18.98%
X13	73.79%	26.21%
X14	65.57%	34.43%
X15	71.18%	28.82%
X16	66.74%	33.26%
X17	68.12%	31.88%

Device	Metric	PFlat
Mobile	X	66.07%
Desktop	X	81.02%
Mobile	X(Capped)	61.85%
Desktop	X(Capped)	75.19%

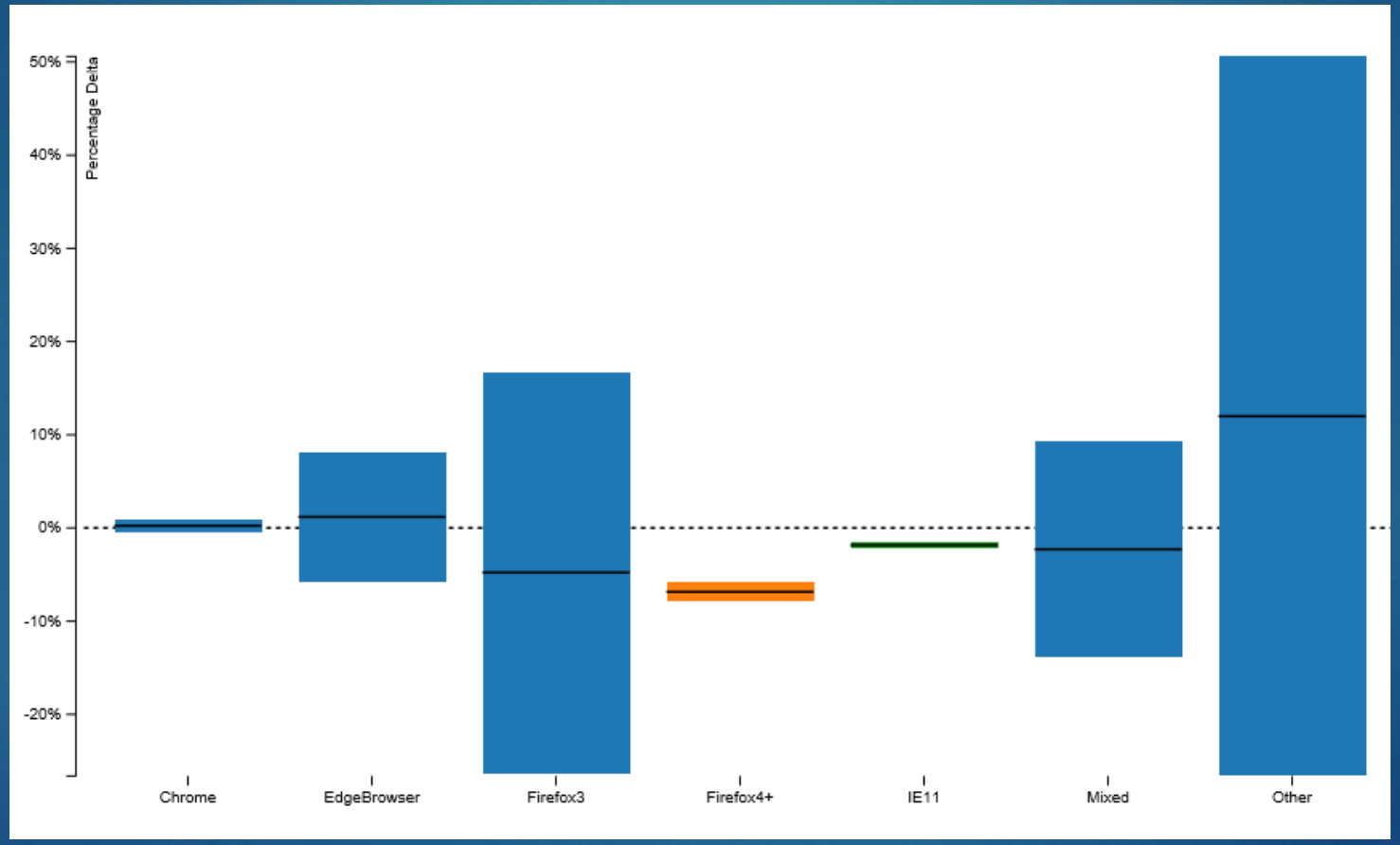
- User Engagement Metrics harder to move, e.g. active days per user, visits per user
- Revenue easier to move than engagement
- Signals on a module or part of page much easier to move than whole page
- Capping metrics for highly skewed distribution increased sensitivity (KDD 2013, Online Controlled Experiments at Large Scale) by increasing power
- Variance Reduction method helps (CUPED, WSDM 2013) by increasing power
- Different devices, product areas have different priors

Challenge 3: Beyond Population Average Treatment Effect

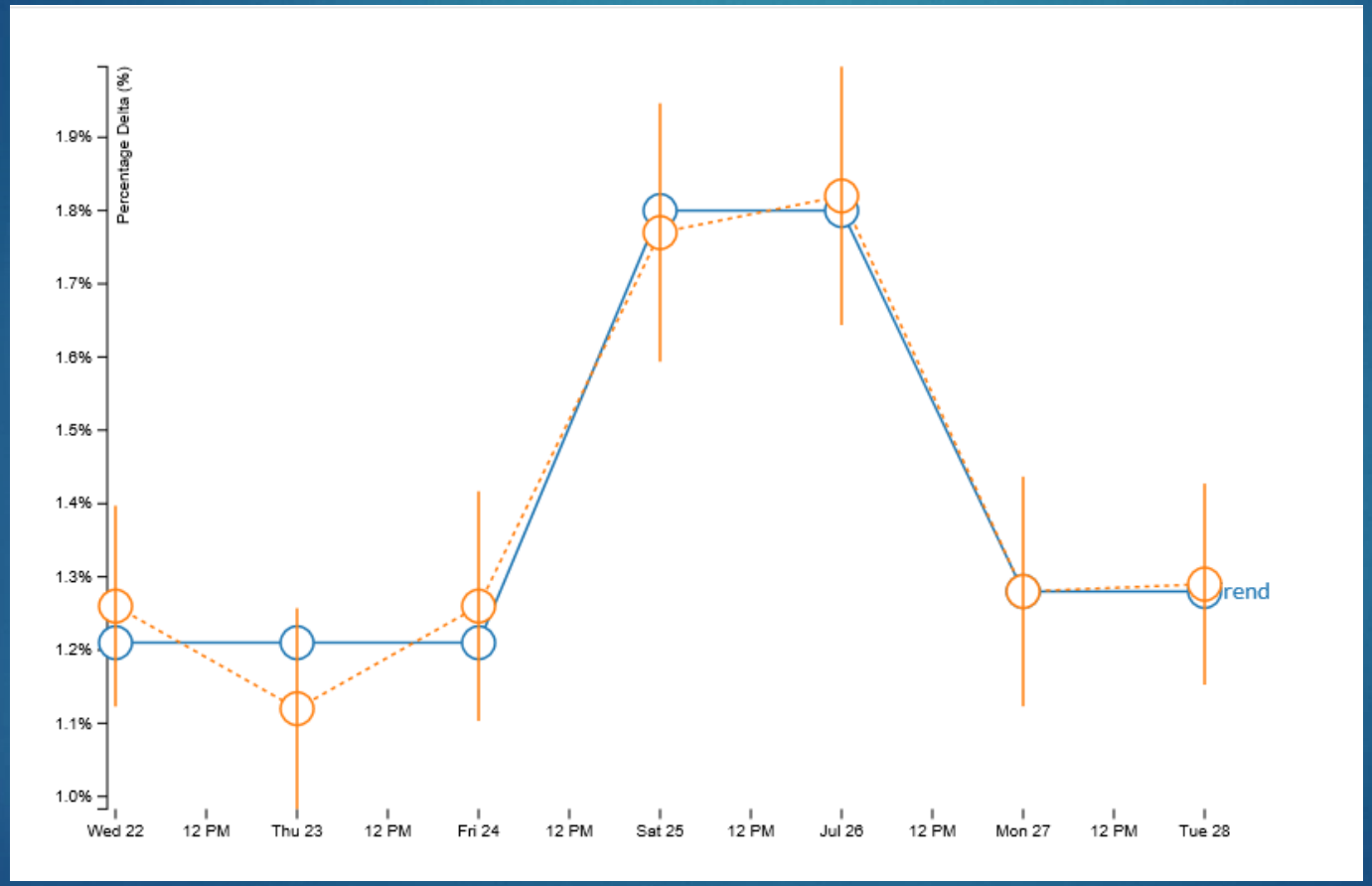
- When we say "treatment effect" most cases we refer to Population Average Treatment Effect (ATE or PATE)
- We know treatment effect differs from person to person
  - A feature might not be popular in some markets -> improvement
  - A feature might be broken on one browser -> bug
- There could be many micro-structure in subpopulations, where treatment effect varies, or even flip sign!
- Heterogeneous Treatment Effect (HTE): Hot topic in Economics/Policy Evaluation, Personalized Treatment/Drug, etc.

- ExP creates different segments and provide segmented scorecard
  - Date: daily metric and effect time series
  - Browsers
  - Markets
  - …
- Challenge
  - we need to find HTE for people, not expecting them to look for it
  - Segment only provides first order HTE, what about higher order effect such as Browser X on weekends

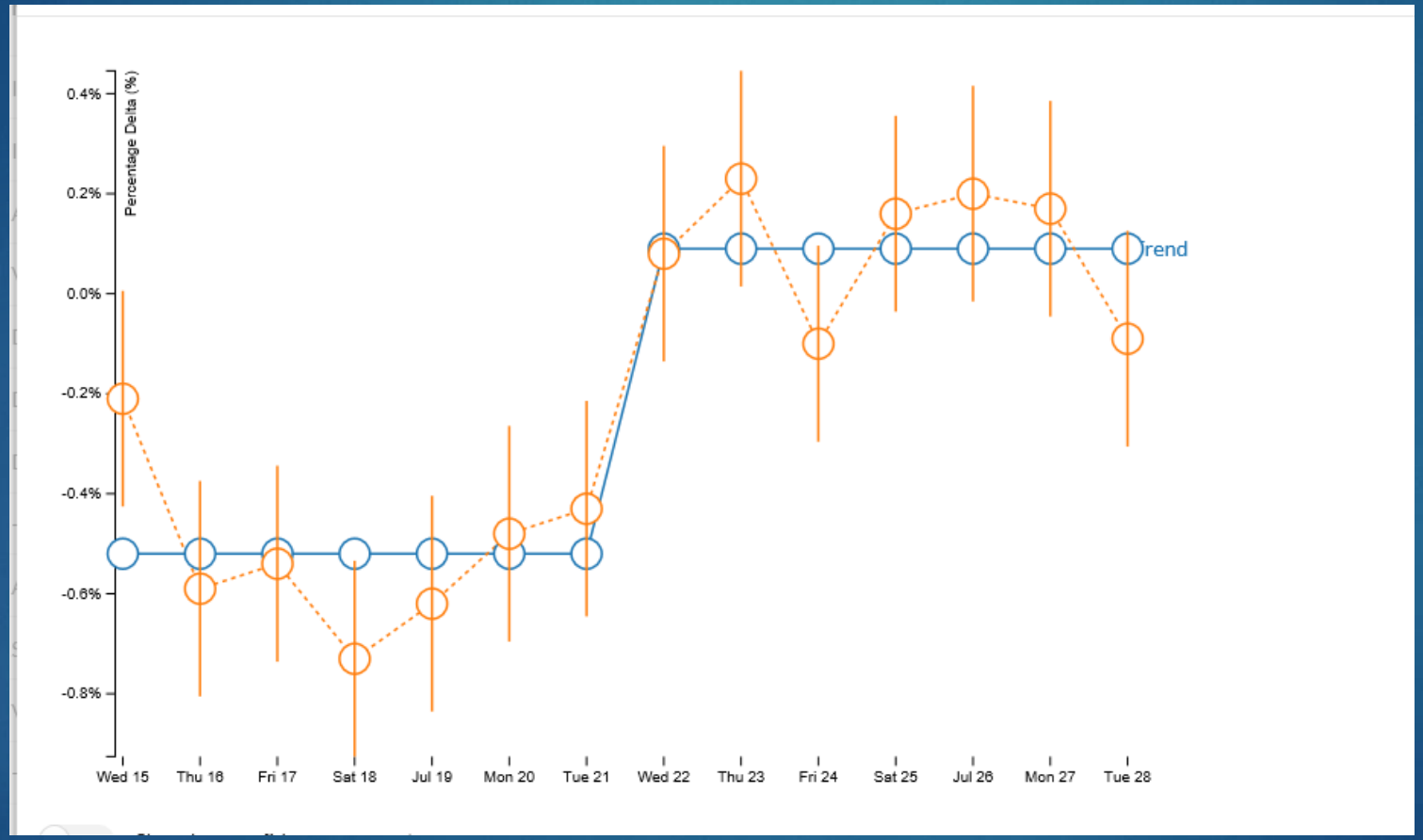
Browser difference



Weekend vs weekday



Shift


Percentage Delta (%)
0.4%
0.2%
0.0%
-0.2%
-0.4%
-0.6%
-0.8%
Wed 15
Thu 16
Fri 17
Sat 18
Jul 19
Mon 20
Tue 21
Wed 22
Thu 23
Fri 24
Sat 25
Jul 26
Mon 27
Tue 28
rend

Machine Learning Framework

- Recall $\tau = Y(1) - Y(0)$ (difference of potential outcomes/counterfactual)
- Given covariates X, we want to learn $\tau(X) = E(\tau|X)$, i.e. regression of τ on X. **WLOG, we assume τ is a function of X, i.e. don't distinguish τ and $E(\tau|X)$**
- If we observe (τ, X), this would be a typical supervised learning task. Find a predictor $\hat{\tau}(X)$ according to optimization criteria
- Theoretical loss: $E((\tau - \hat{\tau})^2)$
- Empirical loss function: $\frac{1}{\mathrm{N}} \times \sum(\tau_i - \hat{\tau}(X_i))^2$
- But we don't observe τ!

Modified/Transformed Outcome

- $T = \pm 1$, 1 for Treatment and -1 for control
- Define $Y^* = 2YT$
- Observation: $E(Y^*|X) = \tau(X)$!
- Loss function

$$E\left(\left(Y^* - \hat{\tau}(X)\right)^2\right) = E\left((Y^* - \tau)^2\right) + E\left(\left(\tau - \hat{\tau}(X)\right)^2\right) + E\left((Y^* - \tau)\left(\tau - \hat{\tau}(X)\right)\right)$$

Note that

- $E\left((Y^* - \tau)\left(\tau - \hat{\tau}(X)\right)\right) = 0$ (first condition on X, $E(Y^* - \tau|X) = 0$)
- $E\left((Y^* - \tau)^2\right)$ is constant, i.e. does not depend on $\hat{\tau}$

Minimize $E\left(\left(Y^* - \hat{\tau}(X)\right)^2\right)$ is equivalent to minimize $E\left(\left(\tau - \hat{\tau}(X)\right)^2\right)$!

Empirical loss: $\frac{1}{N}\sum\left(Y_i^* - \hat{\tau}(X_i)\right)^2$

- Now we can employ different machine learning algorithms
- Different algorithms cover different spaces of $\hat{\tau}$
- Athey and Imbens: use regression tree and cross-validation for tree pruning
  - No problem with nested segments/multiple covariates X
  - Good for categorical covariate as well as continuous covariate
- Athey and Imbens also changed loss function slightly for training(in sample) and testing(out of sample) and demonstrate improvement
  - They called their winning method Causal Tree (CT)

Issues and Limitations of Tree

- Works only on absolute delta, not percent delta
  - Log transformation-> percent delta of geometric mean, not good enough many cases
  - This is due to the function search space of the tree algorithm. Need some custom modifications for percent delta
- Overfit
  - Cross validation only controls the tree size, not the structure
  - Tree structure very unstable/high variance
- Each split reduces sample size, and make subsequent signal weaker
- Could be hard to interpret
  - Tree splitting are hard to follow.
  - You need to summarize from all leaves, and number of leaves could be large

Linear Models + Lasso

- τ = GlobalEffect + FirstOrder Effect + SecondOrder Effect
- Global Effect is intercept
- First Order Effect is the main effect of each segments: WeekEnd is a% more than WeekDay, etc.
- Second Order Effect is the interaction effect between segments
- Tian, et. al. (with Tibshirani) used this together with Transformed Covariate (Equivalent to Transformed Outcome in our case)
- Pros: Easy to interpret.
- Cons:
  - Still high False positive (40% when 50 covariates)
  - Lasso on categorical variables need special care, when a variable have too many categories, putting coefficient on each dummy is not parsimonious
  - Continuous variable could have nonlinear effect

Work in Progress

Main Ideas

- Adaptive to nonlinear effect easily as in tree
- An additive model + regularization for simple and parsimonious interpretation
- Restrict to only kth order interaction effect, k = 2 for most need
  - First order effect is called Segment of Interest, most important
  - Second order effect useful to uncover additional interesting insights

Opportunities #4: New Machine learning algorithms to help automatically uncover insights with focus on interpretation

Competitions

- Many "A/B testing alternatives"
  - A/B testing is a hot area and a growing community
  - More people feels the pain of different challenging problems
  - Some people even go further saying we should discard current A/B testing framework
- Improve sensitivity, test more variants more efficiently
  - Multi-armed Bandit and adaptive traffic reallocation
  - Contextual multi-armed bandit/Multi-verse experiment
  - Metric Optimization Engine (MOE)

Multi-armed bandit

- When we have a large number of treatments to test, e.g. parameter sweeping, we aim to only ship the best one
- Pure exploration bandit: we care about whether we pick up the best treatment. Opportunity cost(regret) in the experiment are not our main concern
- Traditional A/B testing fix the traffic splitting
- Simple idea: can we reduce traffic or early stop some treatments and reallocate traffic to more promising ones?
  - Decades of academic research
  - Simple algorithms: greedy, Thompson sampling, UCB, racing, etc.

Issues

- Fundamental Assumption: metric distribution doesn't change over time (most theory assumes i.i.d. observations)
- Simpson's paradox
  - Two treatments have the same revenue per search, but naïve comparison fails

average revenue	1st day $0.1	2nd day $0.2	Revenue Per Search?
treatment 1	50 searches	75 searches	**0.16**
treatment 2	50 searches	25 searches	**0.133**

- In traditional A/B test, everything, including time effect, are controlled(fixed splitting)
- Remedy: weighted average -> contextual MAB

- Other issues:
  - Observations are not always independent: randomize by user, but same user multiple visits
  - Assumes instant feedback/effect, e.g. no left over
  - Treatment effect also has time effect: treatment 1 might be good in the weekend but bad in the weekdays (Heterogeneous Treatment Effect)

Contextual MAB

- Observational causal inference: unbiased estimate by inverse-propensity-weighting
- Propensity = P(Assigned to treatment i | context/uncontrolled confounders)

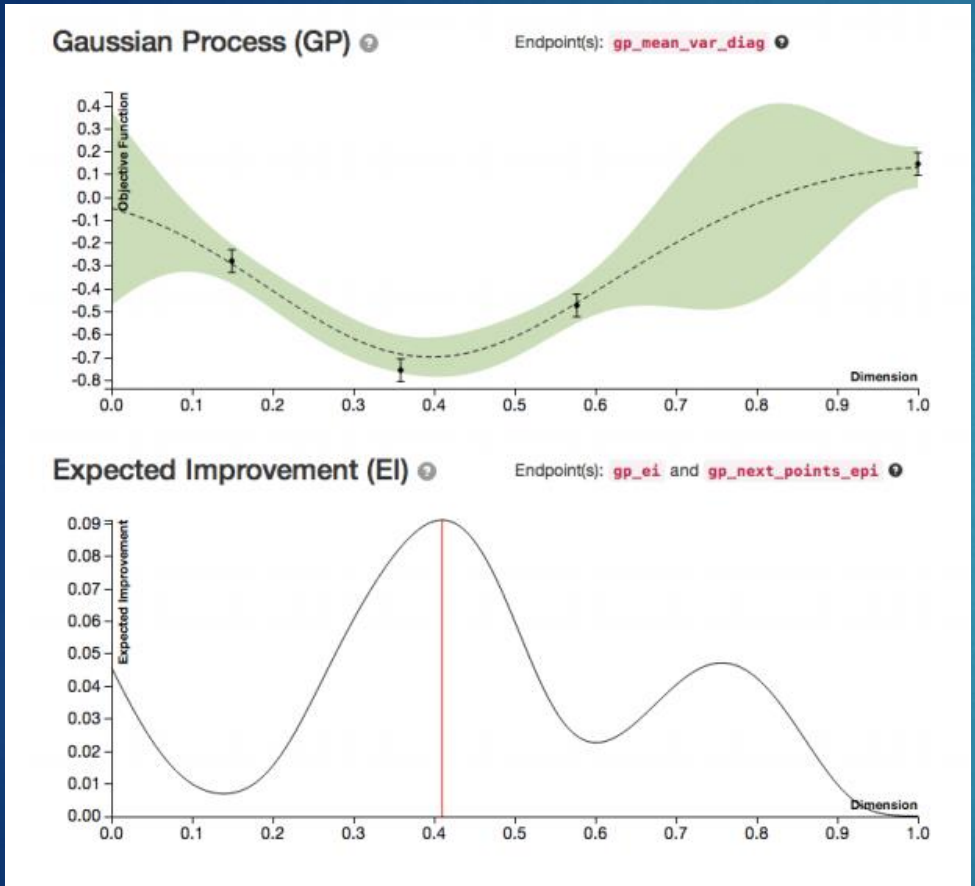
average revenue	1st day $0.1	2nd day $0.2	IPW-Average
treatment 1	50 searches(1/2)	75 searches (3/4)	$\frac{50 \times \frac{0.1}{\frac{1}{2}} + 75 \times \frac{0.2}{\frac{3}{4}}}{\frac{50}{\frac{1}{2}} + \frac{75}{\frac{3}{4}}} = 0.15$
treatment 2	50 searches(1/2)	25 searches (1/4)	$\frac{50 \times \frac{0.1}{\frac{1}{2}} + 25 \times \frac{0.2}{\frac{1}{4}}}{\frac{50}{\frac{1}{2}} + \frac{25}{\frac{1}{4}}} = 0.15$

Contextual MAB

- Time is a special case of context, in general, any ML algorithm has its input as context
- Goal of Contextual MAB is therefore find the best "policy" that maps context to action → optimize ML algorithm over a large set of policy
- Multi-verse experimentation:
  - Randomize at context-> action level
  - Can simultaneously compare infinite number of policies, as long as for each candidate policy, every possible (context, action) pair has support(data point)
- My take:
  - User feedback label: you use live traffic feedback as your labeled data, and with the correct weight adjustment(inverse propensity) .
  - Active Learning: You also choose how to collect labeled data
  - Evaluation: can unbiasedly evaluate different ML algorithms on the labeld data

- Theory guarantees unbiased policy/treatment evaluation
- In practice, variance is the main problem, each (context, action) pair need to have a “not so small” propensity, otherwise the variance will explode
- Therefore even though in theory we can evaluate infinite policies, in practice we still need to have a few candidate policies and design randomization scheme accordingly, and likely don’t adaptively change traffic allocation
- Many other issues similar to MAB applies here too: left over effect, independent observations, etc.

MOE: Metric Optimization Engine



- In MAB, we assume we don't put any assumptions on how the treatment effects of different treatment are related
  - Each treatment independently compete with each other
  - Traffic allocated to treatment A has no value for evaluation of treatment B
- Actually, in parameter sweeping scenario, the unknown "reward" of treatments have certain functional form->Bayesian Global Optimization
- Smoothing/Variance-Bias trade-off, a Bayesian MAB algorithm
- Algorithm for choosing where to sample next
- Practically limited to low dimension parameter space and a less dynamic environment(lab experiments, etc.)

www.exp-platform.com
alexdeng.github.io