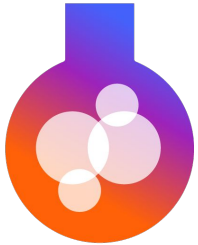
Bayesian Experimentation with Spike-and-Slab Priors

Niloy Biswas
niloy@meta.com
Research Scientist, Central Applied Science, Meta



Central Applied Science

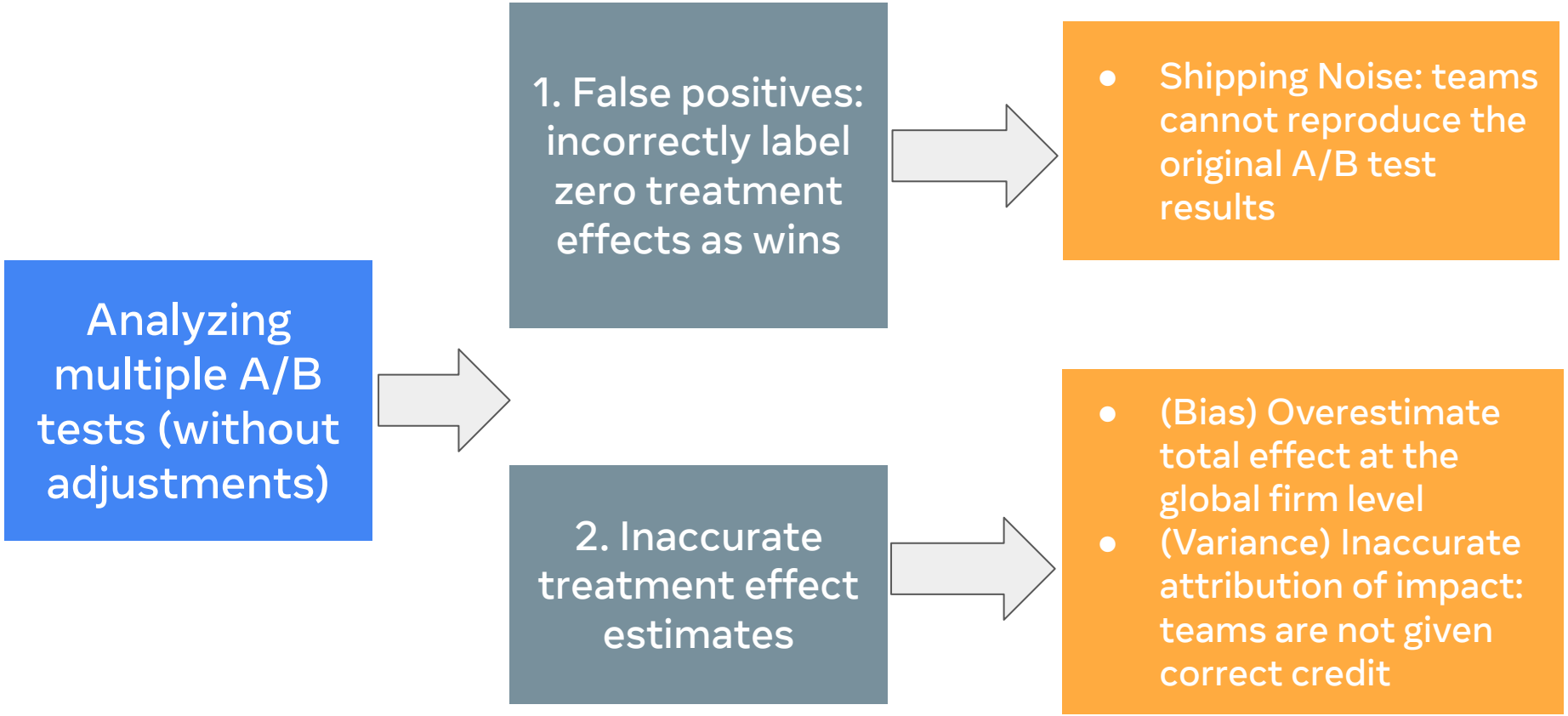
Meta

Large scale experimentation at Meta

- Multiple A/B tests are reviewed daily at Meta
- The tests can be advertiser level, user level, varying duration, switchback …

- **Our goal**: develop a Bayesian model for
  a. Selecting which experiments to launch
  b. Accurately estimating treatment effects of A/B tests

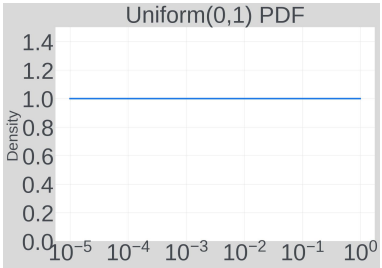
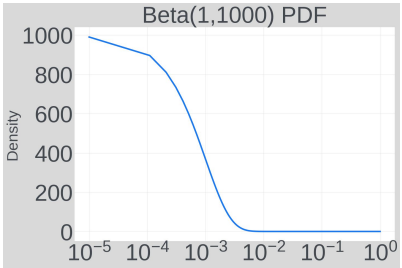
Challenges of large scale experimentation



1. False positives: why do we get false positives with large scale experimentation?

Suppose we ran N=1000 A/A tests and obtained N p-values.

- What is the distribution of the smallest (“winning”) p-value?

	Naive (Un-adjusted)	Correct (Adjusted for Multiple Testing)
P-value Distribution	Uniform(0,1)	Min of N Uniform(0,1) ~ Beta(1,N) for N=1000.
Density		
5% significance threshold	5% quantile of Uniform(0,1) = 0.05	5% quantile of Beta(1,1000)=**0.0000513 (<<<0.05)**

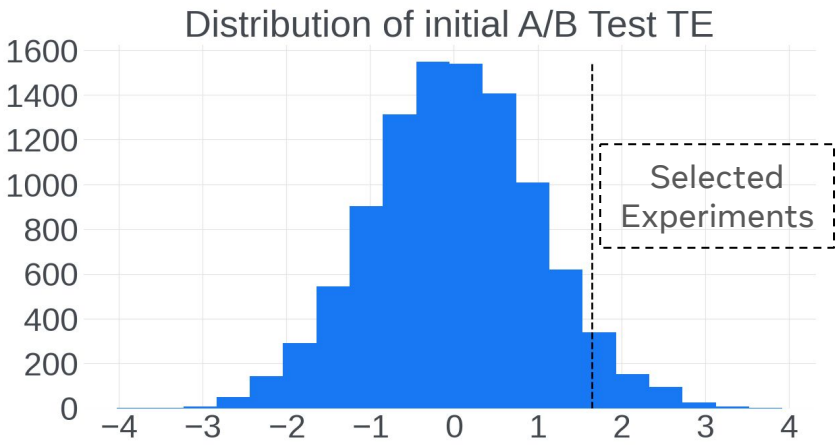
- At 5% significance threshold, we would select N*0.05=50 *false discoveries*
- *We need to account for the total number of experiments run*

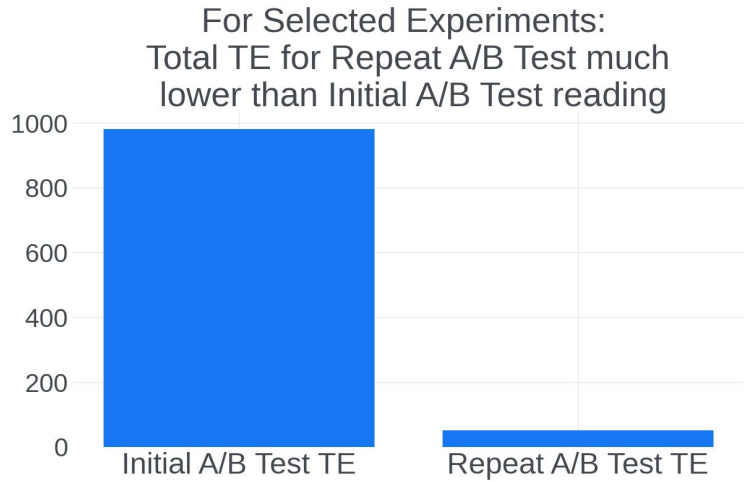
Not controlling false positive rate can incentivize teams to run as many experiments as possible

- Under any fixed significance threshold: the more experiments we run, the more false discoveries are discovered
- If team performance is linked to launches and significant treatment effects from A/B tests, this creates incentives to:
  - Re-run the same experiment repeatedly until significance
  - Record the daily experiment readings repeatedly until significance
  - Track multiple metrics and report only the significant ones
- *We need to account for the total number of experiments run and quality of experiments run by different teams*

2. Inaccurate TE estimates: bias from false discoveries

Suppose we ran multiple A/B tests where the true TEs are zero, and selected positive TE experiments with significant p-values



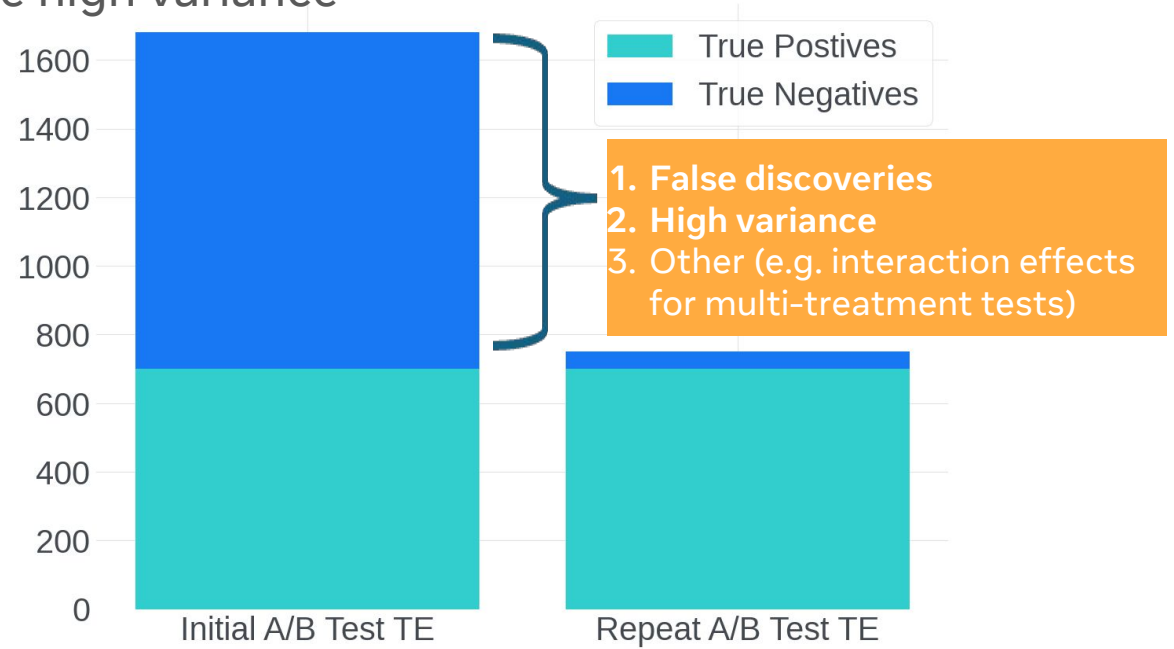


- The selected experiments' TE estimates from the initial A/B tests overestimates the actual TEs post-launch, which leads to overestimation of total TE at firm level

Figures based on synthetic data

2. Inaccurate TE estimates: high variance

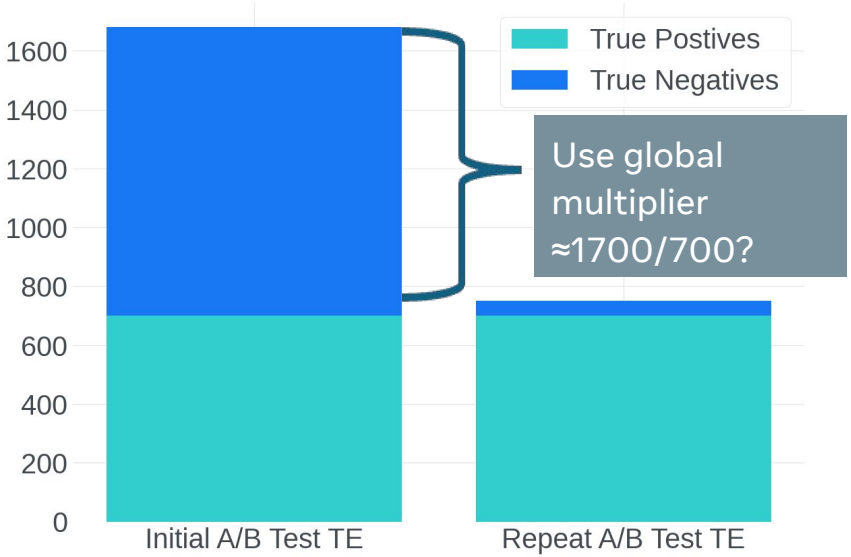
- Even for changes which are *true positives*, the unadjusted TE estimates can have high variance



- We need more accurate TE estimates

Figures based on synthetic data

Inaccurate TE estimates: why not use a global haircut?

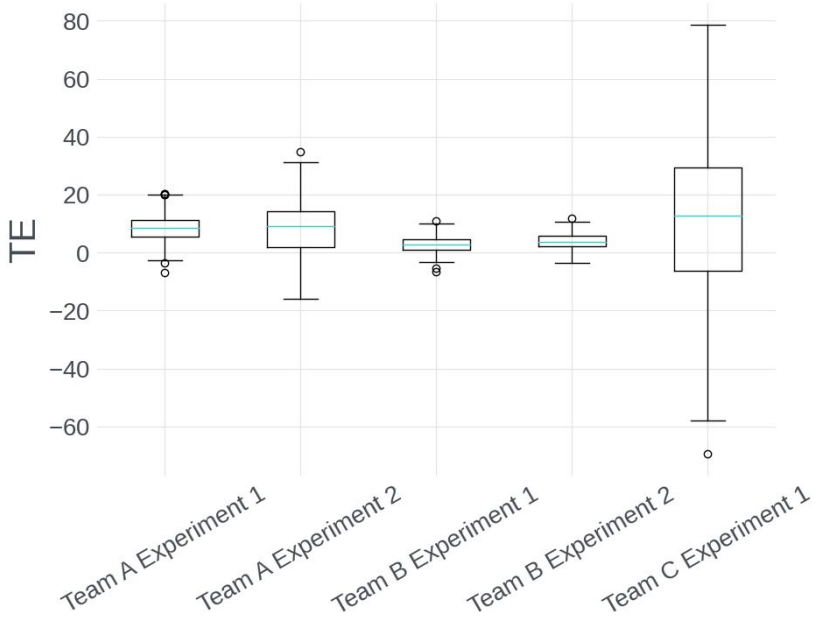


- A global haircut will:
  - Over/under-penalize teams which have high/low-quality experiments
  - Teams will still be incentivized to run as many (potentially low quality) experiments as possible
- We need *local adjustments* at the experiment level

Figures based on synthetic data

Improving TE estimates by incorporating uncertainty

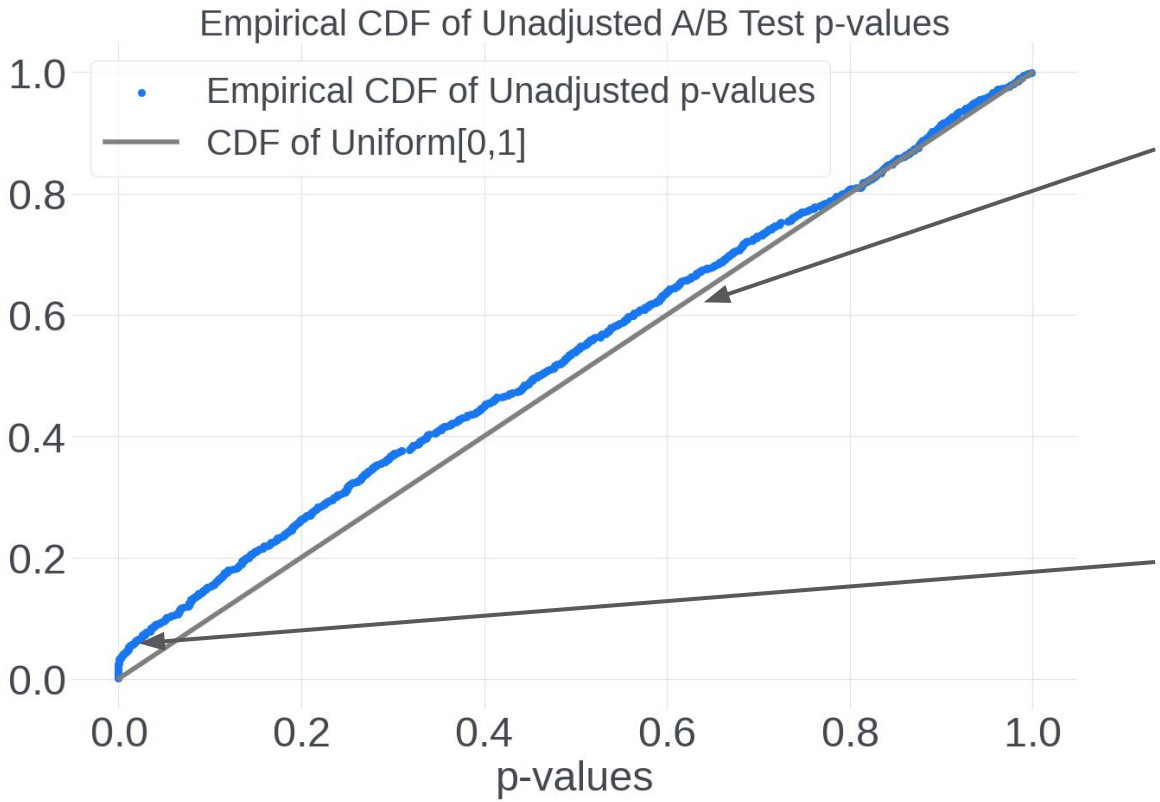
Two experiments with same TE reading but different standard errors are not created equal



Figures based on synthetic data

- For the same TE, we are more confident about experiments which have smaller standard errors.
- *Shrink TE estimates based on their standard errors:*
  - *Shrinkage towards what? Zero? Non-zero global mean? Team-level mean?*

How to model large-scale experimentation?

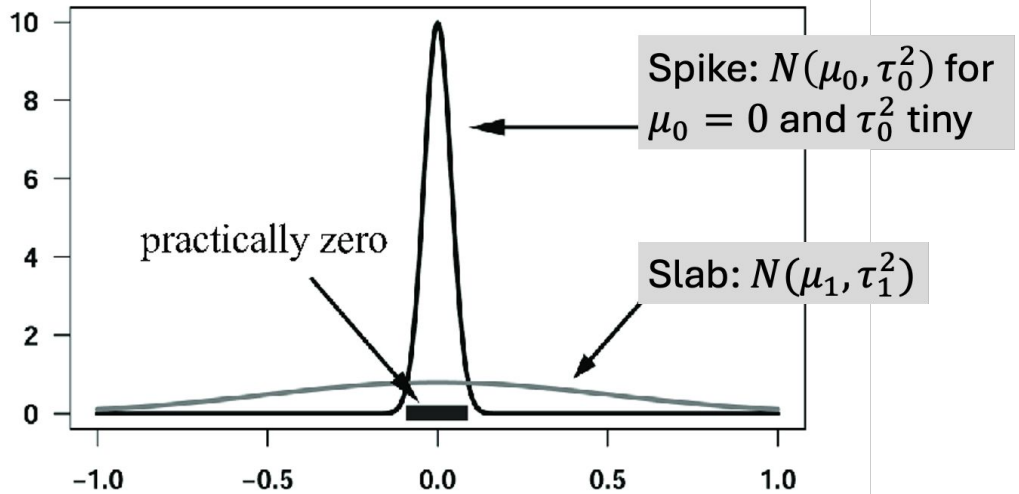


If all experiments were noise, we would expect them to closely follow this 45 degree line

Pulling away from the 45 degree line indicates more non-null experiments

A Bayesian Model for A/B Tests: Spike-and-Slab Prior

- *Spike-and-Slab Prior*: (1-q) $N(\mu_0, \tau_0^2)$+ q $N(\mu_1, \tau_1^2)$

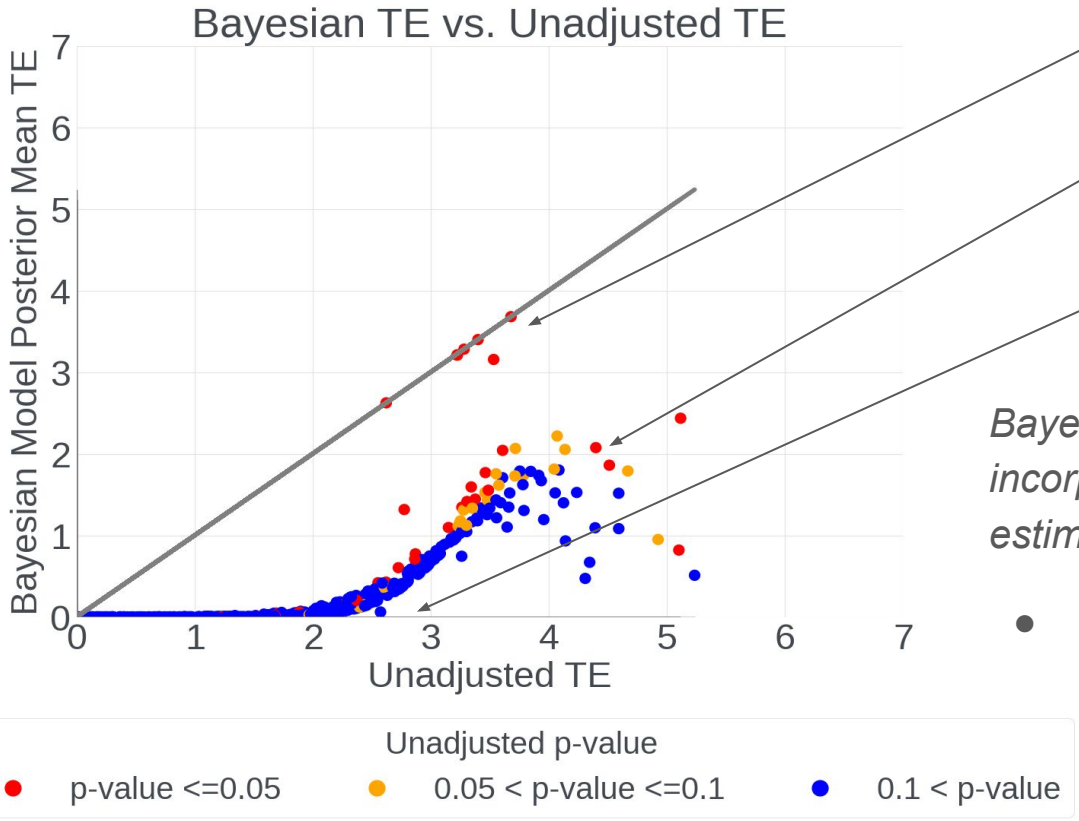


Prior Hyperparameters:

- q: proportion of non-null experiments (spike)
- (μ_0=0,τ_0^2): (mean, variance) of the spike distribution
- (μ_1, τ_1^2): (mean, variance) of the slab distribution

George and McCulloch, JASA 1993
Tadesse and Vannucci, Handbook of Bayesian Variable Selection 2022

Bayesian Spike-and-Slab Model for A/B Tests



Almost no shrinkage for non-null

Medium shrinkage for non-null with medium probability

Aggressive shrinkage for null

Bayesian estimators gives a principled way to incorporate uncertainty of the unadjusted TE estimates.

- *Local level shrinkage: reward higher-quality experiments (with smaller confidence intervals) by shrinking less*

Figures based on synthetic data

Bayesian Model for A/B Tests: operational considerations

- How to choose the prior hyperparameters (i.e. how much to shrink)?
  a. Cross-validation or experiment splitting [1]
  b. Aggregated global shrinkage level guidance based on business needs
  c. Refresh the prior hyperparameters semi-regularly (e.g. biannually, or synchronized with internal release cycles) as a compromise between stale vs noisy model estimates for TE

- Should experiment metadata be used?
  a. Team/ product level information could be difficult to operationalize because of: (i) lack of data for new teams/changing orgs (ii) behavioural externalities (e.g. excessive shrinkage may affect team motivation)

[1] Coey and Cunningham, Meta Research & WWW 2019

Appendices

Posterior Inference for Spike-and-Slab prior

Spike-and-Slab Prior:

- *Latent null/non-null variables*: $z_i \sim Bernoulli(q)$ i.i.d.
- True treatment effects θ_i given z_i: $\theta_i \mid z_i \sim (1-z_i)\, N(0, \tau_0^2) + z_i\, N(\mu_1, \tau_1^2)$ i.i.d.,

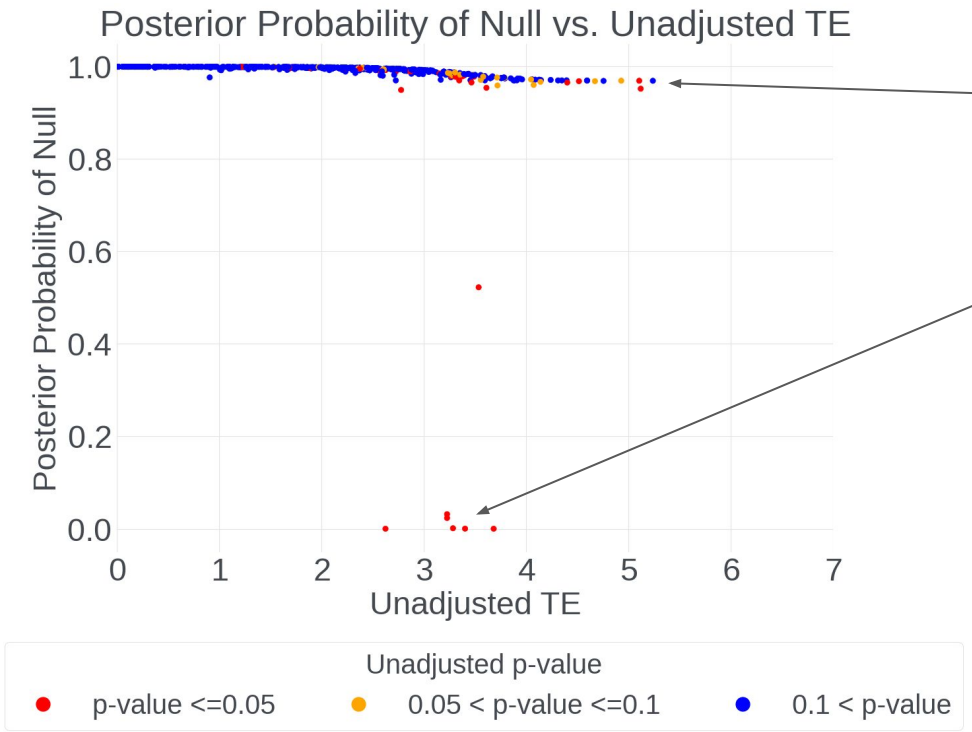
  where $q, \tau_1, \sigma_0, \sigma_1$ are fixed constants with $\sigma_0 << \sigma_1$.

Inference and interpretation of the posterior:

- *Posterior null probabilities $P(z_i=0 \mid y_i)$: use to select experiments*
- *Posterior TE mean $E[\theta_i \mid y_i]$: Bayesian iRev estimates*

Posterior Computation: for this particular model, posterior is analytically tractable

Bayesian Spike-and-Slab Model for A/B Tests



Bayesian model identifies these experiments as null

Bayesian model identifies these experiments as non-null

Recall prior: $z_i \sim$ ***Bernoulli(q)***

Smaller value of global hyperparameter ***q*** corresponds to stronger shrinkage towards the null.