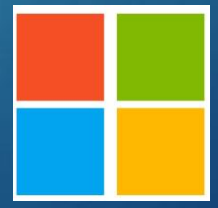
Objective Bayesian Hypothesis Testing

ALEX DENG @ MICROSOFT 2015

THE 1ST WORKSHOP ON OFFLINE AND ONLINE EVALUATION OF WEB-BASED SERVICES



nature International weekly journal of science

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

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:

Crazy? Not Entirely

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

  "The fluctuating female vote: Politics, religion, and the ovulatory cycle"
- P-value hack
  - Multiple testing: different testing methods used by different groups of researchers repeatedly on the same data
  - Optional stopping: stop recruiting subjects once the test is "statistically significant"
- File Drawer Effect and Publication Bias

Pathology of **N**ull **H**ypothesis **S**tatistical **T**esting

- 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
- "Genuine" Prior information not used
  - Researchers motivated to publish counter-intuitive results, which are more often not reproducible

Frequentist NHST's philosophy is **opportunistic**

-Brad Efron, A 250-YEAR ARGUMENT: BELIEF, BEHAVIOR, AND THE BOOTSTRAP

Frequentist vs Bayesian: Two Trial Systems

- Frequentist:
  - One group of jury, with **presumption of innocence**, reckoning evidence of 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
  - Principled, not opportunistic anymore. Think multiple testing, both two groups of jury will share the same multiple testing dividend 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(Bayes Factor)

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

3. Bayes Rule

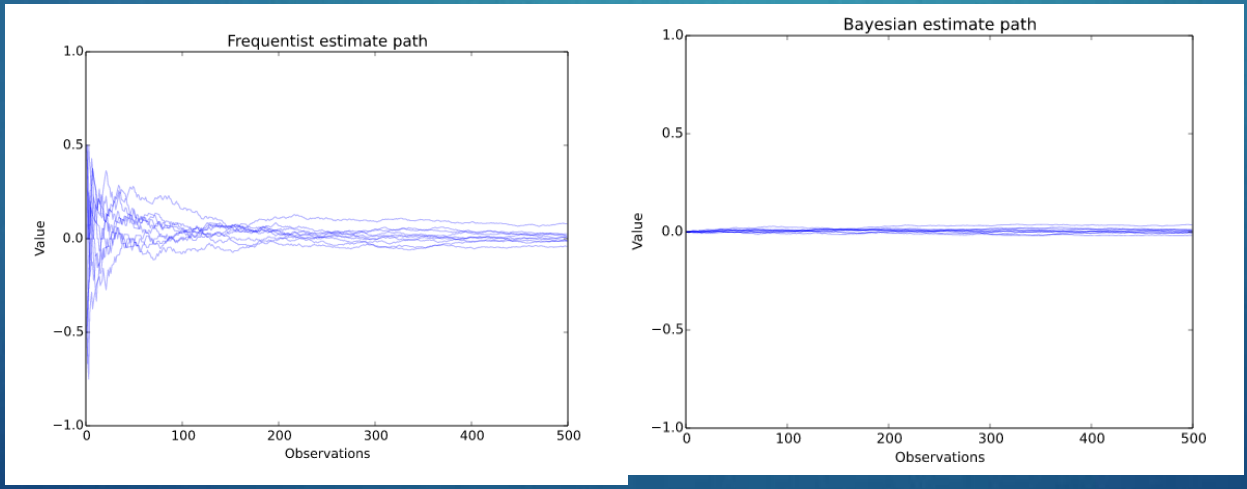
$$\frac{P(H1|Data)}{P(H0|Data)} = PriorOdds \times LR = \frac{P(H1)}{P(H0)} \times \frac{P(Data|H1)}{P(Data|H0)}$$

Bayesian Advantages

- Multiple testing
- Optional stopping/early stopping
  - Smoothing /Regularization
- Useful Prior information
  - Twyman's law "**Any piece of data or evidence that looks interesting or unusual is probably wrong!**"
- More intuitive result
  - P-value often misunderstood
  - Business executives and engineers naturally understand P(H1 | Data)
- Accepting the Null
- Meta Analysis: combine results from different studies

Why Not Everyone Is a Bayesian?

- Prior is often subjective, Conjugate prior often used for closed formula
- So called "non-informative" priors are never truly informative
  - Lindley's Paradox: uniform prior carries a lot of information
- Many above Bayesian Advantages only applies when we know the **true prior(genuine prior)**



Common Ground: the Twin Problem

- A pregnant physicist knows she is having twin boys
- Identical twin or fraternal twin?
- Dr. says based on birth data, 1/3 twins

are identical and 2/3 are fraternal

Bayes Rule:

PriorOdds: P(Identical)/P(Fraternal) = ½

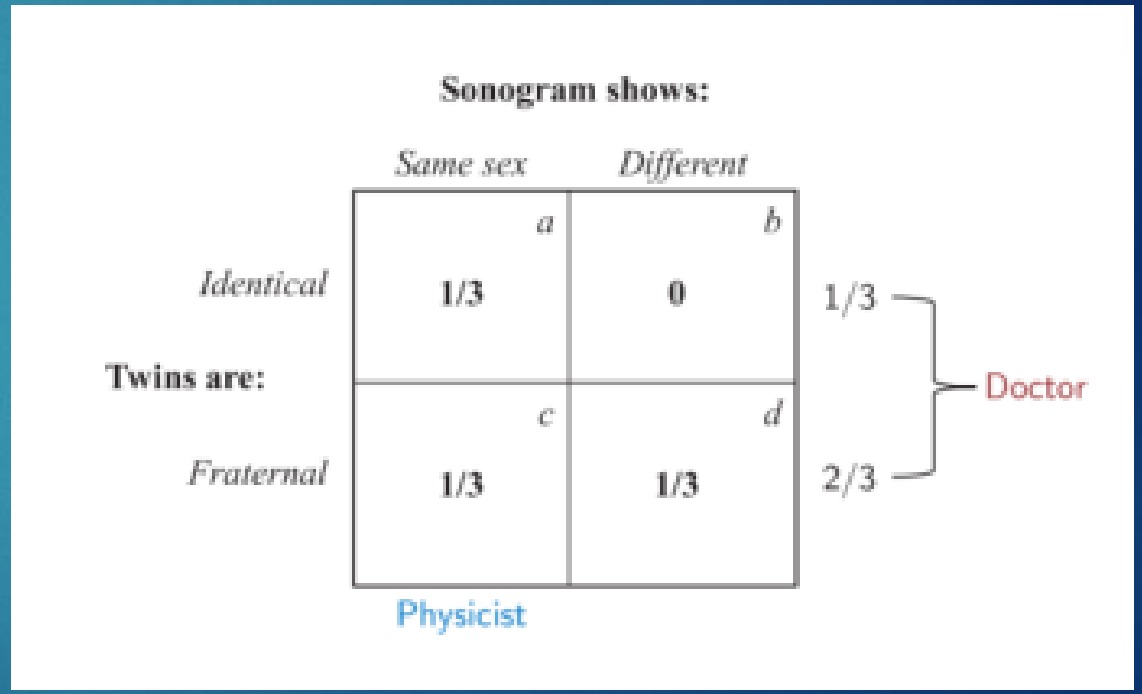
LikelihoodRatio:

P(Data | Identical)/P(Data | Fraternal) = 1/(1/2) = 2

Posterior Odds:

PriorOdds * LikelihoodRatio = 1

=> P(Identical | Data) = P(Fraternal | Data) = 1/2



Compare to Hypothesis Testing

- Similarities
  - Both are testing two hypotheses
  - Both have some data observed
- Dissimilarities
  - Twins: the variable of interest can be observed
  - Testing: we never observe the variable of interest (Null or Alternative)
    - we only observe metric movements, a noisy version of it
- The Dr's input is critical in twin's problem, it provides an objective prior assessment
- Do we have similar input in AB Testing?

Learning Prior Objectively

How does the doctor know the prior?

- Historical Birth Data!
- Estimate the prior using frequentist methods, e.g. MLE, confidence interval, etc.

If we have historical experiments with oracle label, i.e. Null or Alternative, we can easily do the same thing to know prior P(Null) and P(Alternative)

Reality: we don't have label, and also we don't know the distribution of treatment effect

Notation

- $tstat = \frac{\Delta}{\sqrt{\frac{\sigma_t^2}{N_T}+\frac{\sigma_c^2}{N_C}}}$, NEff (effective sample size): $\frac{1}{\frac{1}{N_T}+\frac{1}{NC}}$
- Sigma (Pooled SD): $\sqrt{\frac{\sigma_t^2}{N_T}+\frac{\sigma_c^2}{N_C}}/\sqrt{NEff}$, δ (Effect Size) : $\frac{\Delta}{Sigma}$
- **tstat = $\delta/\sqrt{\frac{1}{NEff}}$** *turn two sample into one sample problem*
- E(δ) = μ (treatment effect scaled by Sigma)

Make more sense to put prior on effect size since it is scale-invariant

Two Group Model

- Prior: Any Feature has
  - P(H1) = p to have an effect
  - P(H0) = 1-p to be flat
- Under H0, $\mu = 0$
- Under H1, $\mu \sim N(0, V^2)$ (normal for practical simplicity, could be any distribution)
- We observe: $\delta = \frac{\Delta}{\text{Sigma}} \sim N(\mu, \frac{1}{NEff})$ (given μ)
- Things to inference:
  - Based on observation δ, what is $\mathrm{P}(H1|Data)$ and $\mathrm{P}(H0|Data)$
  - What is the distribution of μ given the observation?

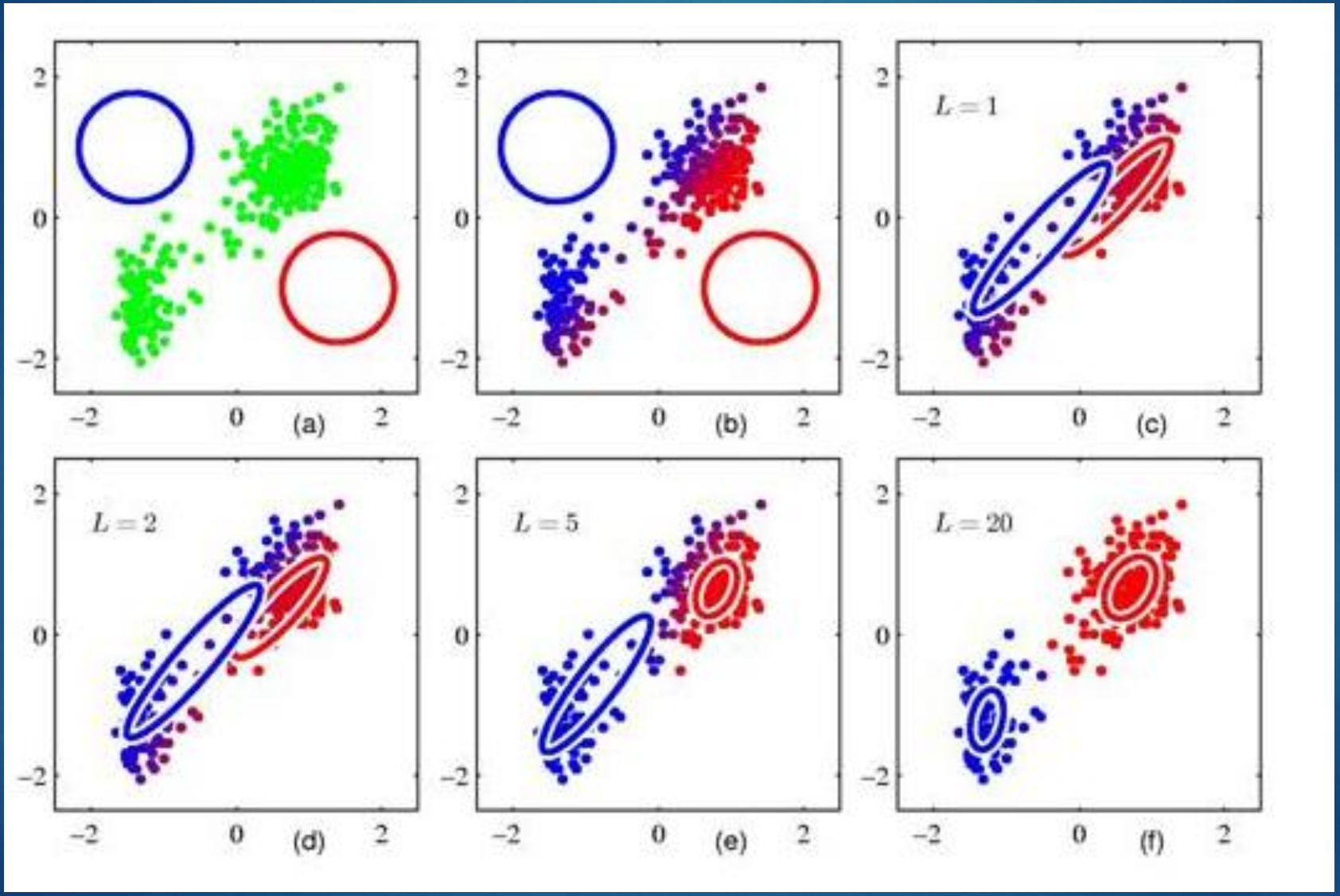
Model Fit

Given a set of historical experiment data

1. If we know the label(H0 or H1), we can estimate p and V
2. If we know p and V, we can estimate P(H1 | Data) and P(H0 | Data) for each historical experiment
3. P(H1 | Data) and P(H0 | Data) are like Soft-Label/Fuzzy-Label in step 1. We can iterate between 1 and 2 until convergence!

- This is classic Expectation-Maximization!
- Converge to MLE of p and V
  - Called MLE-II(MLE of hyper-parameter) or Empirical Bayes

Soft K-means(Bishop)



Does it work?

1. If I randomly simulate data from H0, can this algorithm converges to P(H0) = 100%?
2. If I randomly simulate x% from H0 and 1-x% from H1 with a given V, can this algorithm converge to P(H1)=x% and V = V?

- Answer in general is yes if we have more than 1000 historical data points. Estimation is also reasonable for more than 200 historical data points
  - This is properties of MLE, as this algorithm estimates MLE and MLE is consistent
  - For 1, we need to bound V away from 0. otherwise H1 = H0 and there is no way to separate these two

Simulation Results

- Common set up: NEff (effective sample size) = 1E6
- P(H0) = 100%

N	100	200	1,000	2,000
$\widehat{P(H_0)}$	0.987(0.040)	1.000(0.0007)	1.000(0.004)	1.000(0.0005)

  - N here is the number of historical data points
- P(H0) and P(H1) mixed
  - Varying V, the larger the V, the easier the problem
  - We vary V by changing k where V = k*1/sqrt(NEff), see later for intuition

N=2000		k=4(V=4E-3)	k=8(V=8E-3)	k=10(V=1E-2)
$P(H_0) = 95\%$	$\widehat{P(H_0)}$	0.962(0.008)	0.953(0.006)	0.953(0.006)
	$\widehat{V}$	4.23E-3(0.55E-3)	7.76E-3(0.82E-3)	9.66E-3(0.98E-3)
$P(H_0) = 90\%$	$\widehat{P(H_0)}$	0.909(0.012)	0.903(0.009)	0.903(0.008)
	$\widehat{V}$	3.81E-3(0.31E-3)	7.42E-3(0.48E-3)	9.28E-3(0.58E-3)
$P(H_0) = 80\%$	$\widehat{P(H_0)}$	0.798(0.015)	0.802(0.011)	0.802(0.011)
	$\widehat{V}$	3.89E-3(0.18E-3)	7.83E-3(0.31E-3)	9.80E-3(0.37E-3)
$P(H_0) = 50\%$	$\widehat{P(H_0)}$	0.487(0.020)	0.491(0.015)	0.492(0.014)
	$\widehat{V}$	3.88E-3(0.11E-3)	7.77E-3(0.19E-3)	9.73E-3(0.23E-3)

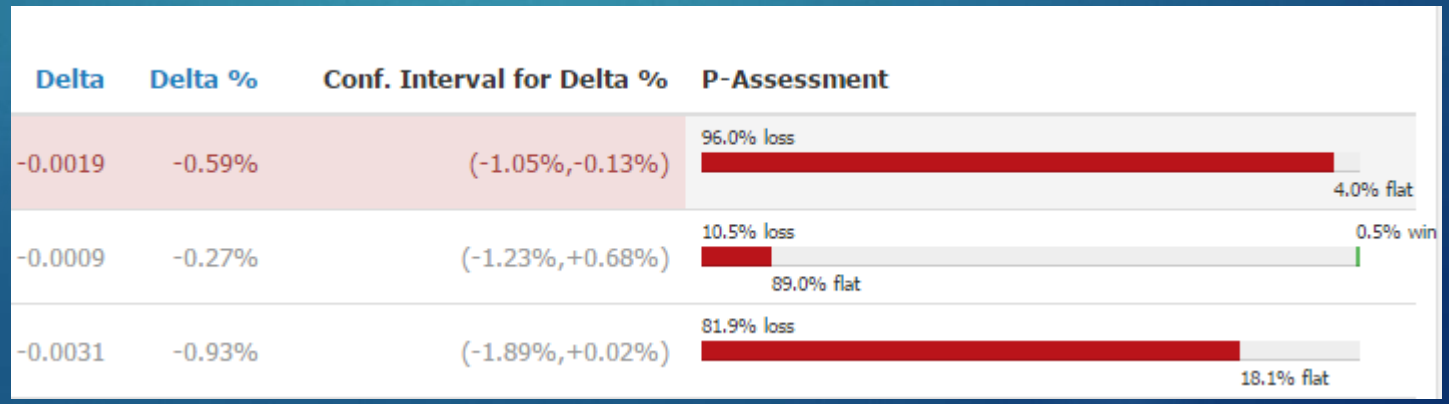
N=200		k=4(V=4E-3)	k=8(V=8E-3)	k=10(V=1E-2)
$P(H_0) = 95\%$	$\widehat{P(H_0)}$	0.965(0.019)	0.963(0.016)	0.962(0.016)
	$\widehat{V}$	4.44E-3(1.04E-3)	8.67E-3(2.04E-3)	1.07E-2(0.25E-2)
$P(H_0) = 90\%$	$\widehat{P(H_0)}$	0.925(0.034)	0.907(0.026)	0.908(0.025)
	$\widehat{V}$	3.62E-3(0.75E-3)	6.80E-3(1.12E-3)	8.55E-3(1.45E-3)
$P(H_0) = 80\%$	$\widehat{P(H_0)}$	0.869(0.042)	0.843(0.033)	0.835(0.033)
	$\widehat{V}$	3.94E-3(0.69E-3)	7.40E-3(1.10)E-3	9.05E-3(1.33E-3)
$P(H_0) = 50\%$	$\widehat{P(H_0)}$	0.594(0.067)	0.518(0.051)	0.506(0.047)
	$\widehat{V}$	3.44E-3(0.37E-3)	6.42E-3(0.59E-3)	7.94E-3(0.71E-3)

Bing Results/Presentation

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)
- Variance Reduction method helps (CUPED, WSDM 2013)
- Different devices, product areas have different priors



FAQ

How to pick a historical experiment corpus?

- Ideally, you want a corpus that represents the type of experiment you are running
- This is like matching in observational data causal inference. Practically, we can just use product area and type of treatments, e.g. UX change, Algo change or Perf change

Why should I believe my experiment now is "like" those from a year ago?

- Even for the same product area, your success rate might change. So some kind of time-dependent weighting might be needed in areas where a lot of changes are going on.

Any other distributions for effect size beyond normal?

- Maybe the real distribution has heavier tail. In theory you can use any parametric model and learn parameters. But more parameters mean you need more historical data to get a good estimation.

FAQ

What if my historical data is limited

- Classic cold start problem. One solution is to use full Bayesian, put a prior on prior. EM -> Variational Bayes
- Intuitively this is like start with “population mean” and gradually converge to the true “subgroup mean”

Conclusion

- Bayesian framework provides a unified framework that solves many pathologies of Frequentist NHST
  - Multiple Testing, optional stopping
- Choice of prior is critical
- For online A/B Testing at scale, we are in a unique position where we can unify Bayesian and Frequentist method by learning prior objectively using historical data

Question?

- Full paper available at alexdeng.github.io

Appendix

Learn p and V from historical data

- For each metric, put historical data together into a dataset, compute $NEff, \delta$ from $\Delta, N_T, N_C \; and \; tstat$
- Initial guess: p=0.5, V = 1
- For each data point
  1. Calculate P(data_i | H1) and P(data_i | H0) using V
  2. Calculate P(H1 | data_i) using Bayes Rule and p
  3. P<- average P(H1 | data_i) from 2
  4. Estimate V using weight P(data_i | H1)
     1. $Var(\delta|H1) = E\left(\frac{1}{NEff}|H1\right) + V^2$
     2. $Var(\delta|H1) \; and \; E\left(\frac{1}{NEff}|H1\right)$ estimated from data with weight
- Iterate until converge