Bayesian AB Testing

Max Kochurov

Moscow State University

Lecture 3

Agenda

How it is done, Classic

"if your p-value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true"

1. p-value is used in thousands of research papers
2. p-value is extremely popular for its easy interpretation
3. easy to calculate confidence intervals

How it is done, Classic

"if your p-value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true"

1. p-value is used in thousands of research papers
2. p-value is extremely popular for its easy interpretation
3. easy to calculate confidence intervals

Are you sure?

Do you understand the nature of the p-value?

Do you understand p-values?

Express test, what is true from this?

1. P - the probability that the null hypothesis is true.
2. 1 minus the value of P - this is the probability that the alternative hypothesis is true.
3. A statistically significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.
4. The value $P > 0.05$ means that no effect was observed.

Do you understand p-values?

Express test, what is true from this?

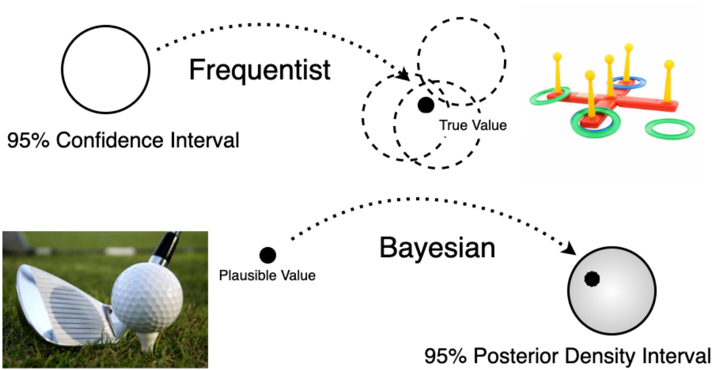
1. P - the probability that the null hypothesis is true.
2. 1 minus the value of P - this is the probability that the alternative hypothesis is true.
3. A statistically significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.
4. The value $P > 0.05$ means that no effect was observed.

Is using p-value bad?

I do not urge you to give up p-values, but I urge you to add more understanding.

Interpreting p-values

Greatest insights into p-values:



Hypothesis Testing in H0, H1 framework

You should know what is hypothesis testing, t-test, p-values.

- 1 sample mean test $t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\widehat{\sigma}/\sqrt{n}}$
- 2 sample mean test $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{2}{n}}}, \quad s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}, \dots$
- 2 sample not equal variances, now equal sample sizes test $\dots, s_p = \sqrt{\frac{(n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}}$

Too Complicated

The less assumptions we have, the more complicated is math and implementation

What is Bayes like?

Be careful with p-value interpretation

- Frequentist confidence intervals are not the most probable values
- p-value - not the probability of "no effect"

Bayesian approach is about interpretation:

Good

- Easier to explain
- Easier to turn into actions

Bad

- You have to understand the domain problem

Bayesian Visualization Tools

1. Highest Density Interval
2. Region of Practical Equivalence
3. Bayes Factor
4. Custom

Highest Density Interval

HDI The most popular way to interpret the posterior

1. Represents a range of most probable values
2. Easy to interpret and calculate
3. Easy to visualize

Example

- With 95% probability effect size in range [A, B]
- Range [A, B] represents 95% of most probable effect sizes

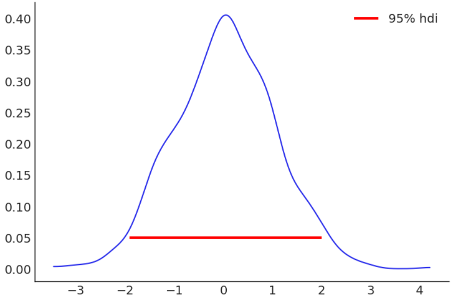


Figure: Highest Density Interval

Region of Practical Equivalence

RoPE is a common way to say if a parameter estimate is "significant". The use case:

1. You do not care if the effect size is less than 0.1
2. Plot the region overlapping with the posterior
3. Decide

Example

The effect size "E" is out of the region of practical equivalence so we treat it as a significant one

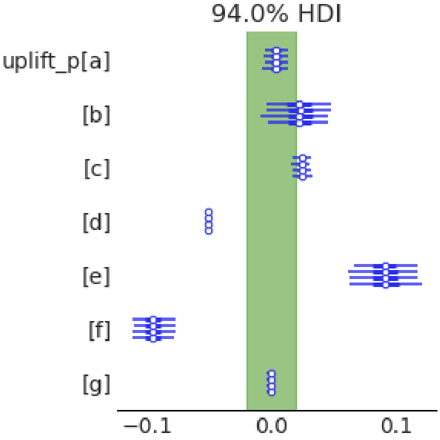


Figure: Rope Plot

Bayes Factor

IMO the most complicated to explain statistic.

1. Similar to the Frequentist p-value
2. Harder to interpret and explain to people
3. Checks H0 vs H1 for x_0

Definition

Bayes Factor is defined as the ratio of the likelihood of one particular hypothesis to the likelihood of another hypothesis

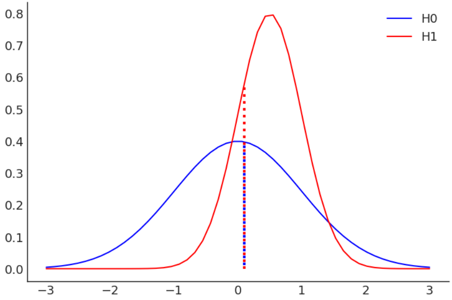


Figure: $\mathrm{BF} = \frac{\mathrm{pdf}_{H1}(x_0)}{\mathrm{pdf}_{H0}(x_0)}$

Custom Queries

You can do much more!

1. $P(A < 0)$
2. $P(A > B)$
3. $P(\max(A) > \max(B))$
4. $P(A = \arg\max(A, B, C, D))$
5. $P(\text{profit}(X, \Theta) > \$100)$
6. Quantiles - $Q_{0.05}(\text{profit}(X, \Theta))$

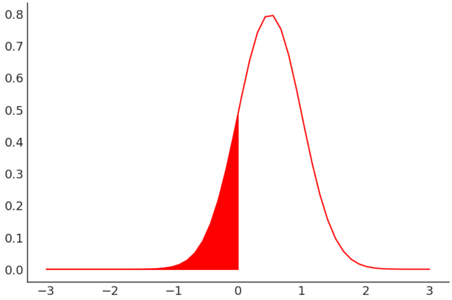


Figure: $P(A < 0)$

Takeouts

Bayesians have a Swiss Knife for Hypothesis Checking

1. Numerous ways to interpret results
2. Not a Yes/No answer
3. Uncertainty is obviously represented
4. Flexibility in analysis
5. Easy to implement
6. Easy to interpret

Figure: Bayesian Hypothesis Testing

Types of Problems

Bayesian AB testing is widely applicable

1. Discrete Observations (views and clicks)
2. Continuous Observations (read time, spent amount)
3. With Context Predictors (CUPED[1])
4. With Hierarchies (Regions)

Types of Problems

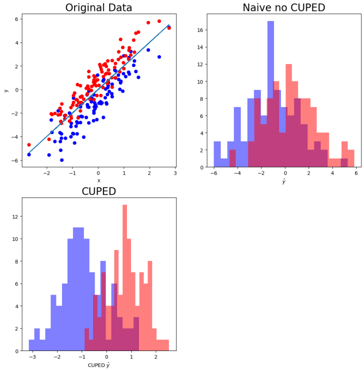
Bayesian AB testing is widely applicable

1. Discrete Observations (views and clicks)
2. Continuous Observations (read time, spent amount)
3. With Context Predictors (CUPED[1])
4. With Hierarchies (Regions)

Types of Problems

Bayesian AB testing is widely applicable

1. Discrete Observations (views and clicks)
2. Continuous Observations (read time, spent amount)
3. With Context Predictors (CUPED[1])
4. With Hierarchies (Regions)



Types of Problems

Bayesian AB testing is widely applicable

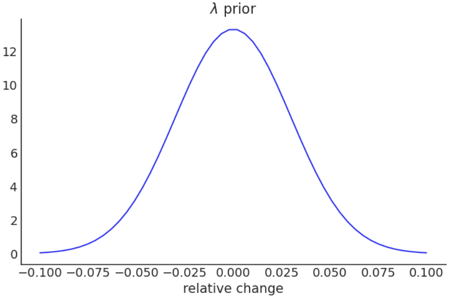
1. Discrete Observations (views and clicks)
2. Continuous Observations (read time, spent amount)
3. With Context Predictors (CUPED[1])
4. With Hierarchies (Regions)

Approaching Priors

Uplift λ

Relative change to the baseline

When you start the experiment, don't you know anything about the set of possible outcomes?



Setting priors for Uplift

You are in the preparation to run an experiment B vs holdout A. You might be interested in increasing the mean of statistics (average bill)

Setting priors for Uplift

You are in the preparation to run an experiment B vs holdout A. You might be interested in increasing the mean of statistics (average bill)

- Do you expect you have a 1000% increase? Very sure No

Relative or Absolute change?

Make it clear if the change is relative or absolute!

Setting priors for Uplift

You are in the preparation to run an experiment B vs holdout A. You might be interested in increasing the mean of statistics (average bill)

- Do you expect you have a 1000% increase? Very sure No
- Do you expect you have a 100% increase? Very sure No

Relative or Absolute change?

Make it clear if the change is relative or absolute!

Setting priors for Uplift

You are in the preparation to run an experiment B vs holdout A. You might be interested in increasing the mean of statistics (average bill)

- Do you expect you have a 1000% increase? Very sure No
- Do you expect you have a 100% increase? Very sure No
- Do you expect you have a 10% increase? Unlikely

Relative or Absolute change?

Make it clear if the change is relative or absolute!

Setting priors for Uplift

You are in the preparation to run an experiment B vs holdout A. You might be interested in increasing the mean of statistics (average bill)

- Do you expect you have a 1000% increase? Very sure No
- Do you expect you have a 100% increase? Very sure No
- Do you expect you have a 10% increase? Unlikely
- Do you expect you have a 3% increase? Maybe

Relative or Absolute change?

Make it clear if the change is relative or absolute!

Setting priors for Uplift

You are in the preparation to run an experiment B vs holdout A. You might be interested in increasing the mean of statistics (average bill)

- Do you expect you have a 1000% increase? Very sure No
- Do you expect you have a 100% increase? Very sure No
- Do you expect you have a 10% increase? Unlikely
- Do you expect you have a 3% increase? Maybe
- Do you expect you have a 3% decrease? Maybe

Relative or Absolute change?

Make it clear if the change is relative or absolute!

Setting priors for Uplift

You are in the preparation to run an experiment B vs holdout A. You might be interested in increasing the mean of statistics (average bill)

- Do you expect you have a 1000% increase? Very sure No
- Do you expect you have a 100% increase? Very sure No
- Do you expect you have a 10% increase? Unlikely
- Do you expect you have a 3% increase? Maybe
- Do you expect you have a 3% decrease? Maybe
- Do you expect you have an X% decrease? Your answer

Relative or Absolute change?

Make it clear if the change is relative or absolute!

Example Workflow

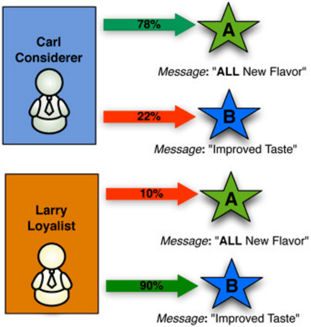
- How to set up the experiment?
- How to plan the execution?
- How to interpred the results?

Binomial Model Example

- The example is binary Yes/No choice
- Observations follow the Bernoulli likelihood

$$x_i^A \sim \text{Bernoulli}(p_A)$$
$$x_i^B \sim \text{Bernoulli}(p_B)$$

Do we have additional information?



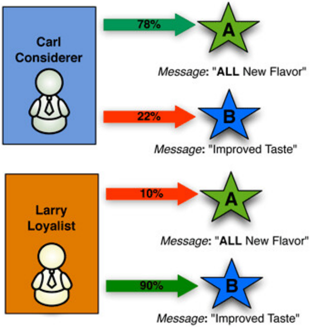
Binomial Model Example

- The example is binary Yes/No choice
- Observations follow the Bernoulli likelihood

$$x_i^A \sim \text{Bernoulli}(p_A)$$
$$x_i^B \sim \text{Bernoulli}(p_B)$$

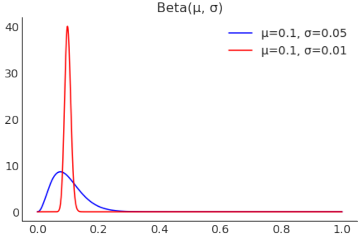
Do we have additional information?

- Historical $\bar{p}$
- Expected improvement $\pm\bar{\sigma}\%$ (e.g. $\pm 0.01\%$)



Adding Additional Information

We can parametrize Beta distribution in a special way



$$
\begin{aligned}
G &\in \{A, B\} \\
x_i^G &\sim \text{Bernoulli}(p_G) \\
p_G &\sim \text{Beta}(\alpha_G, \beta_G)\ s.t. \\
\mathbb{E}p_G &= \bar{p}, \\
\text{Var}\, p_G &= \bar{\sigma}^2
\end{aligned}
$$

Adding Additional Information

We can parametrize Beta distribution in a special way

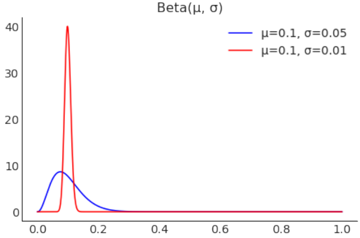
$$X \sim \text{Beta}(\alpha, \beta)$$

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\sigma = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$X \sim \text{Beta}(\mu, \sigma) \Rightarrow$$

$$\Rightarrow \begin{cases} \alpha & = \mu\kappa \\ \beta & = (1-\mu)\kappa \\ \text{where} & \kappa = \frac{\mu(1-\mu)}{\sigma^2} - 1 \end{cases}$$



$$G \in \{A, B\}$$

$$x_i^G \sim \text{Bernoulli}(p_G)$$

$$p_G \sim \text{Beta}(\alpha_G, \beta_G) \; s.t.$$

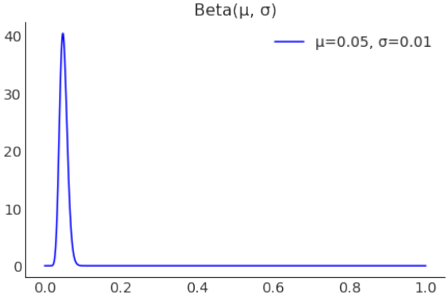
$$\mathbb{E} p_G = \bar{p},$$

$$\text{Var}\, p_G = \bar{\sigma}^2$$

Prior Specification

Case Study

Our historical levels of conversion are about 5% (and fixed). We expect about 1% **absolute** change ($\bar{\sigma}$) after implementing the solution. Or, similarly, 20% **relative** change ($\bar{\delta}$).

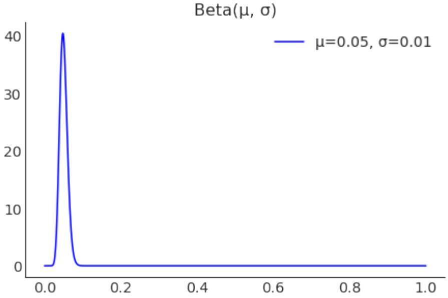


$$
\begin{aligned}
\bar{p} &= 0.05 \\
\bar{\sigma} &= 0.01 = \bar{\delta} \cdot 0.05 \\
G &\in \{A, B\} \\
p_G &\sim \text{Beta}(\mu = \bar{p}, \sigma = \bar{\sigma})
\end{aligned}
$$

Prior Specification

Case Study

Our historical levels of conversion are about 5% (and fixed). We expect about 1% **absolute** change ($\bar{\sigma}$) after implementing the solution. Or, similarly, 20% **relative** change ($\bar{\delta}$).



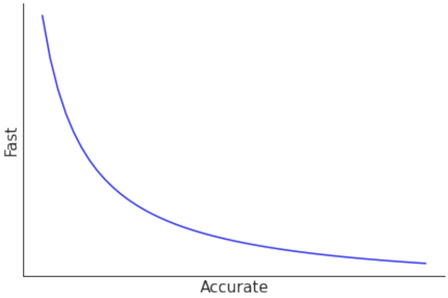
$$\begin{aligned} \bar{p} &= 0.05 \\ \bar{\sigma} &= 0.01 = \bar{\delta} \cdot 0.05 \\ G &\in \{A, B\} \\ p_G &\sim \text{Beta}(\mu = \bar{p}, \sigma = \bar{\sigma}) \end{aligned}$$

Takeout

Special Beta parametrization leads to more interpretable priors

Key questions be for you start

- How much time can be allocated for the test?
    - How accurate is the decision then?
- How accurate should be the decision?
    - How much time will be allocated for the test?

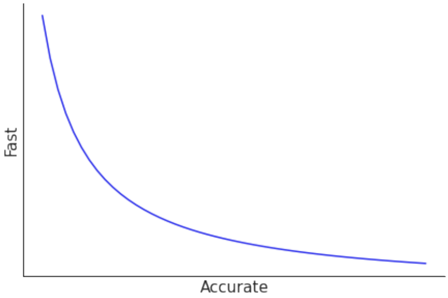


Key questions be for you start

- How much time can be allocated for the test?
  - How accurate is the decision then?
- How accurate should be the decision?
  - How much time will be allocated for the test?

Impossibility

You can't be fast in data collection and accurate at the same time



Parameter Recovery Study

Parameter recovery is a simulated experiment to know your model better.

1. Generate data from a model configuration
2. Pretend you do not know the true values
3. Run inference for your model
4. Compare estimated parameters and ground truth ones

Given the results

- How well can you infer the model state?
- How does data size affects the results?
- Are there unidentifiable parameters?

Suggested Reading

Chapter 4 in Bayesian Workflow

Parameter Recovery Study

Parameter recovery is a simulated experiment to know your model better.

1. Generate data from a model configuration
2. Pretend you do not know the true values
3. Run inference for your model
4. Compare estimated parameters and ground truth ones

Given the results

- How well can you infer the model state?
- How does data size affects the results?
- Are there unidentifiable parameters?

Suggested Reading

Chapter 4 in Bayesian Workflow

Parameter Recovery for AB testing

Given:

- Effect is significant if $|p - \bar{p}| > \bar{\sigma}$

Recall the model

$$
\begin{aligned}
i &\in 1 \ldots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Parameter Recovery for AB testing

Given:

- Effect is significant if $|p - \bar{p}| > \bar{\sigma}$
- Ignore effects $|p - \bar{p}| < \bar{\sigma}$

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Parameter Recovery for AB testing

Given:

- Effect is significant if $|p - \bar{p}| > \bar{\sigma}$
- Ignore effects $|p - \bar{p}| < \bar{\sigma}$
- How large should be N to decide if the effect is significant?

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Parameter Recovery for AB testing

Given:

- Effect is significant if $|p - \bar{p}| > \bar{\sigma}$
- Ignore effects $|p - \bar{p}| < \bar{\sigma}$
- How large should be N to decide if the effect is significant?
- $N = 0$, $N = 1000$, $N = 100000$?

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Parameter Recovery for AB testing

Given:

- Effect is significant if $|p - \bar{p}| > \bar{\sigma}$
- Ignore effects $|p - \bar{p}| < \bar{\sigma}$
- How large should be N to decide if the effect is significant?
- $N = 0$, $N = 1000$, $N = 100000$?
- What metric to use to evaluate detect effectiveness?

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Parameter Recovery for AB testing

Given:

- Effect is significant if $|p - \bar{p}| > \bar{\sigma}$
- Ignore effects $|p - \bar{p}| < \bar{\sigma}$
- How large should be N to decide if the effect is significant?
- $N = 0$, $N = 1000$, $N = 100000$?
- What metric to use to evaluate detect effectiveness?

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Key observation

Effect detection is a classification problem. E.g. negative, neutral, positive effects. We can use ROC-AUC for multiclass

AB Testing as classification

Some definitions of our classification setup

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Posterior $p(p \mid X_{1:N})$

AB Testing as classification

Some definitions of our classification setup

1. Target $\hat{p}$, used for data generation

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Posterior $p(p \mid X_{1:N})$

AB Testing as classification

Some definitions of our classification setup

1. Target $\hat{p}$, used for data generation
2. Labels
   - "0" is $\hat{p} < \bar{p} - \bar{\sigma}$, negative
   - "1" is $\bar{p} - \bar{\sigma} < \hat{p} < \bar{p} + \bar{\sigma}$, neutral
   - "2" is $\hat{p} > \bar{p} + \bar{\sigma}$, positive

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Posterior $p(p \mid X_{1:N})$

AB Testing as classification

Some definitions of our classification setup

1. Target $\hat{p}$, used for data generation
2. Labels
   - "0" is $\hat{p} < \bar{p} - \bar{\sigma}$, negative
   - "1" is $\bar{p} - \bar{\sigma} < \hat{p} < \bar{p} + \bar{\sigma}$, neutral
   - "2" is $\hat{p} > \bar{p} + \bar{\sigma}$, positive
3. Predictions (probabilities using the posterior):
   - $P(\text{p is negative} \mid X_{1:N})$
   - $P(\text{p is neutral} \mid X_{1:N})$
   - $P(\text{p is positive} \mid X_{1:N})$

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Posterior $p(p \mid X_{1:N})$

AB Testing as classification

Some definitions of our classification setup

1. Target $\hat{p}$, used for data generation
2. Labels
   - "0" is $\hat{p} < \bar{p} - \bar{\sigma}$, negative
   - "1" is $\bar{p} - \bar{\sigma} < \hat{p} < \bar{p} + \bar{\sigma}$, neutral
   - "2" is $\hat{p} > \bar{p} + \bar{\sigma}$, positive
3. Predictions (probabilities using the posterior):
   - $P(\text{p is negative} \mid X_{1:N})$
   - $P(\text{p is neutral} \mid X_{1:N})$
   - $P(\text{p is positive} \mid X_{1:N})$

Recall the model

$$
\begin{aligned}
i &\in 1 \dots N \\
x_i &\sim \text{Bernoulli}(p) \\
p &\sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})
\end{aligned}
$$

Posterior $p(p \mid X_{1:N})$

Run the simulation study

1. for $\hat{p} \in \dots$, for $N \in \dots$ get $p(p \mid X_{1:N})$
2. for $N \in \dots$ calculate ROC-AUC

ROC-AUC in Action

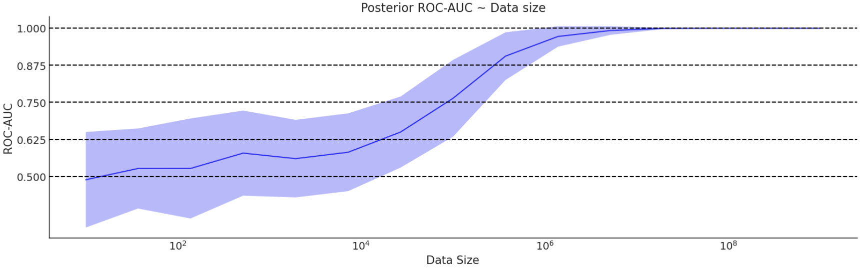


Figure: ROC-AUC increases as you get more data

Time is constraint:

1. Discuss maximum affordable time
2. Consult the plot for the expected ROC-AUC in decision

ROC-AUC is constraint:

1. Discuss minimum required ROC-AUC
2. Consult the plot for the expected data size

After the Inference

Situation: you've run the test for the aforehand specified duration.
Key questions:

1. Which alternative to choose?
2. What is the comparison criterion?
3. Is the criterion connected to the real life?

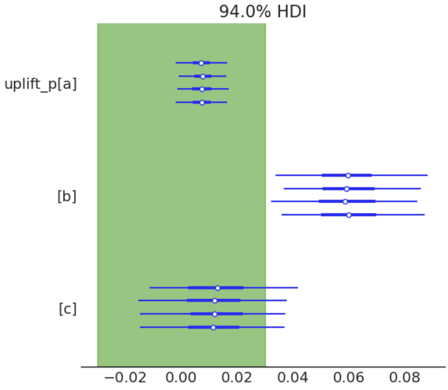


Figure: Example ROPE plot

After the Inference

Situation: you've run the test for the aforehand specified duration. Key questions:

1. Which alternative to choose?
2. What is the comparison criterion?
3. Is the criterion connected to the real life?

A better metric

A good metric is the one that is connected to expected profit.

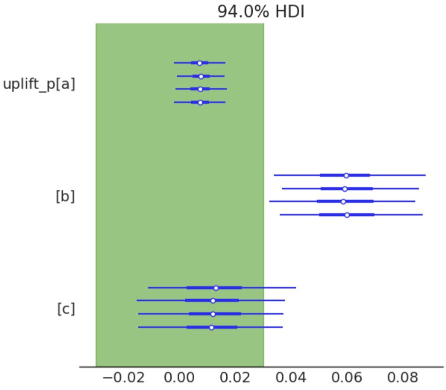


Figure: Example ROPE plot

Interpreting the Posterior

How can we calculate a better mectic?

It could look like this:

Interpreting the Posterior

How can we calculate a better mectic?

- Connect the conversion rate p_A or p_B to the company size audience

It could look like this:

Interpreting the Posterior

How can we calculate a better mectic?

- Connect the conversion rate p_A or p_B to the company size audience
- Use "Customer Value" as a proxy for money effect

It could look like this:

Interpreting the Posterior

How can we calculate a better mectic?

- Connect the conversion rate p_A or p_B to the company size audience
- Use "Customer Value" as a proxy for money effect

It could look like this:

$$\text{Monetization}_A = \\ (\text{Per User Value}) \times (\text{Num Users}) \times \Delta p_A - (\text{Implementation Cost})$$

Interpreting the Posterior

How can we calculate a better mectic?

- Connect the conversion rate p_A or p_B to the company size audience
- Use "Customer Value" as a proxy for money effect

It could look like this:

$$\text{Monetization}_A = \\ (\text{Per User Value}) \times (\text{Num Users}) \times \Delta p_A - (\text{Implementation Cost})$$

Use the posterior

We can calculate $p(\text{Monetization}_A \mid X_A)$ out of $p(p_A \mid X_A)$

Monetization Posterior

$$(\text{Per User Value}) \times (\text{Num Users}) \times \Delta p_A - (\text{Implementation Cost})$$

- Implementation cost might differ
- Per User Value might have scenarios
- You connect the experiment with business
- Compare outcomes with uncertainty

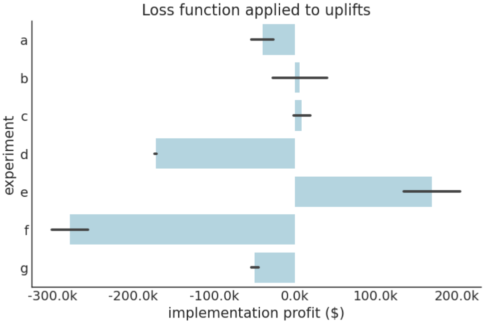


Figure: $p(\text{Monetization}_G \mid X_G)$

Takeouts

Real Life AB testing is full of challenges. Bayesian tools can do much more to turn data into action.

1. Framing the statistical test
   - Setting priors
   - Setting likelihood
2. Planning the experiment
   - Parameter recovery study
3. Bayesian decision making to take action
   - Loss functions
   - Scenario testing

References I

R. Kohavi, A. Deng, Y. Xu, and T. Walker.
In *Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data*, 02 2013.