

# Bayesian AB Testing

Max Kochurov

PyMC Labs

December 17, 2022

# Agenda

- ① Classic
  - Assumptions
- ② Hypothesis Testing
  - Highest density interval
  - Region of Practical Equivalence
  - Custom Hypothesis
- ③ AB Testing
  - Priors
- ④ Workflow
  - Prior
  - Preparing an experiment
  - Parameter Recovery
  - Posterior Simulations
- ⑤ Supplementary
  - Parameter Recovery

# p-value in $H_0$ , $H_1$ framework

"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

## p-value in $H_0$ , $H_1$ framework

"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

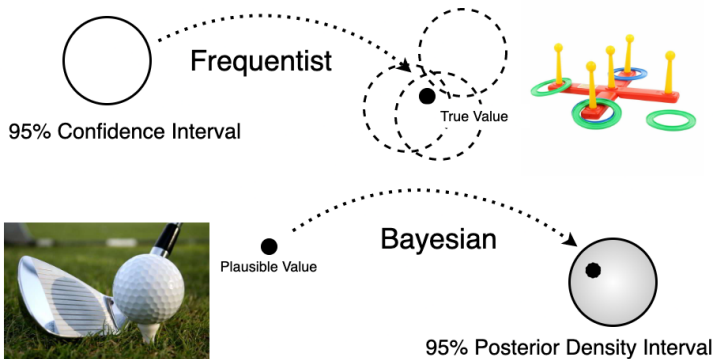
Are you sure?

Do you understand the nature of the p-value?

Disclaimer: I do not advocate against p-values, just know your tools.

# Interpreting p-values

Greatest insights into p-values:



Suggested Reading

Explanation of P-values by Joe Felsenstein

# Bayesian 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

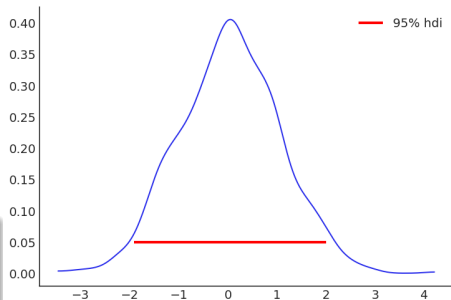


Figure: Highest Density Interval

# 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

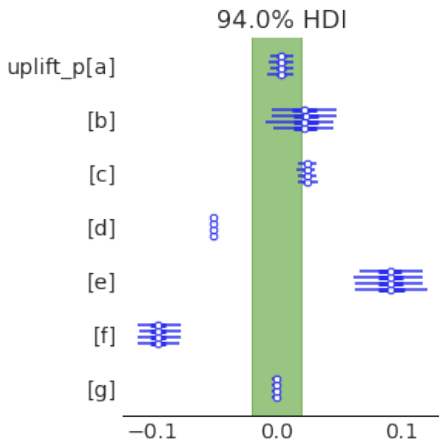


Figure: Rope Plot



# Custom Queries

You can do much more!

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

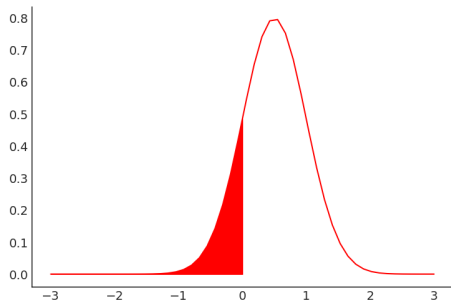


Figure:  $P(A < 0)$

# 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

# 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 after changes in B 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 after changes in B you have a 1000% increase? Very sure No
- Do you expect after changes in B 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 after changes in B you have a 1000% increase? Very sure No
- Do you expect after changes in B you have a 100% increase? Very sure No
- Do you expect after changes in B 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 after changes in B you have a 1000% increase? Very sure No
- Do you expect after changes in B you have a 100% increase? Very sure No
- Do you expect after changes in B you have a 10% increase? Unlikely
- Do you expect after changes in B 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 after changes in B you have a 1000% increase? Very sure No
- Do you expect after changes in B you have a 100% increase? Very sure No
- Do you expect after changes in B you have a 10% increase? Unlikely
- Do you expect after changes in B you have a 3% increase? Maybe
- Do you expect after changes in B 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)

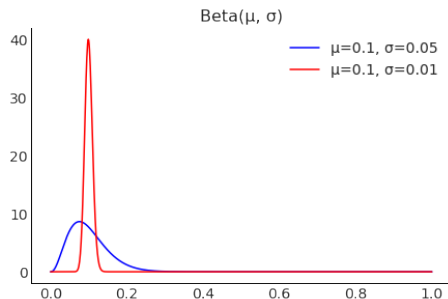
- Do you expect after changes in B you have a 1000% increase? Very sure No
- Do you expect after changes in B you have a 100% increase? Very sure No
- Do you expect after changes in B you have a 10% increase? Unlikely
- Do you expect after changes in B you have a 3% increase? Maybe
- Do you expect after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

Relative or Absolute change?

Make it clear if the change is relative or absolute!

# Adding Additional Information

We can parametrize Beta distribution in a special way



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

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

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

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

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

# Adding Additional Information

We can parametrize Beta distribution in a special way

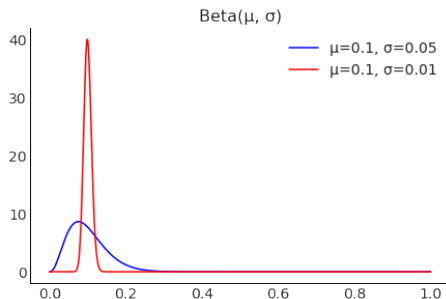
$$X \sim \text{Beta}(\alpha, \beta)$$

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

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

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

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



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

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

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

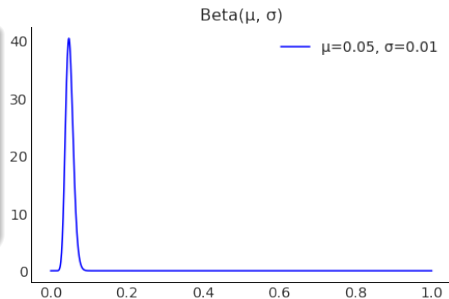
$$\mathbb{E}p_G = \bar{p},$$

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

# 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 ( $\bar{\delta}$ ).



$$\bar{p} = 0.05$$

$$\bar{\sigma} = 0.01 = \bar{\delta} \cdot 0.05$$

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

$$p_G \sim \text{Beta}(\mu = \bar{p}, \sigma = \bar{\sigma})$$

# Prior Specification

## Case Study

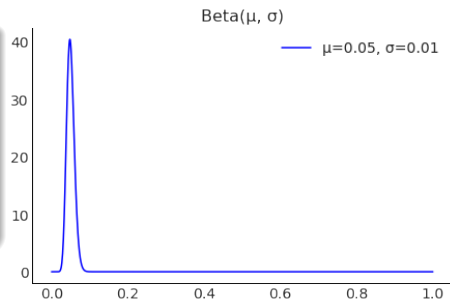
Our historical levels of conversion are about 5% (and fixed). We expect about 1% **absolute** change ( $\bar{\sigma}$ ) after implementing the solution. Or, similarly, 20% **relative** change ( $\bar{\delta}$ ).

$$\bar{p} = 0.05$$

$$\bar{\sigma} = 0.01 = \bar{\delta} \cdot 0.05$$

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

$$p_G \sim \text{Beta}(\mu = \bar{p}, \sigma = \bar{\sigma})$$

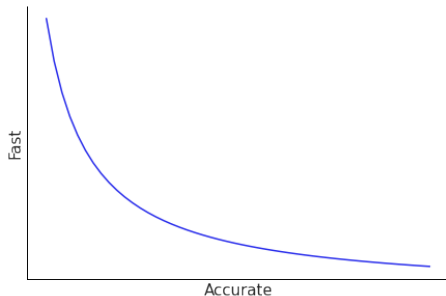


## Takeout

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?

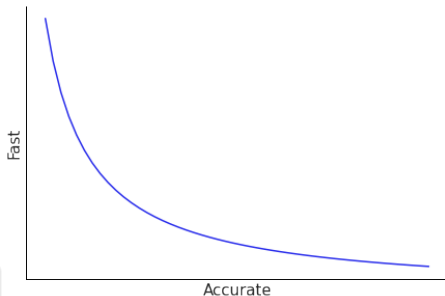


# Key questions be for you start

- 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](#)

# ROC-AUC in Action

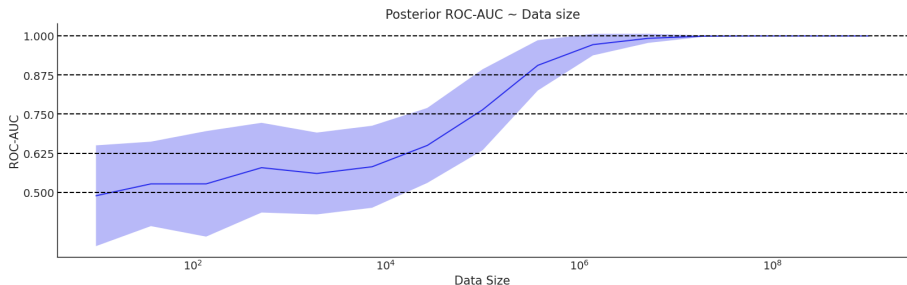


Figure: ROC-AUC increases as you get more data

## 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 beforehand specified duration.

Key questions:

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

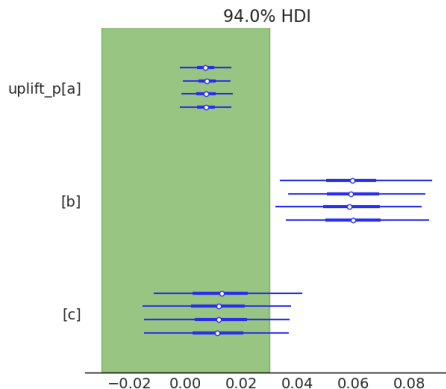


Figure: Example ROPE plot

# After the Inference

**Situation:** you've run the test for the beforehand 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.

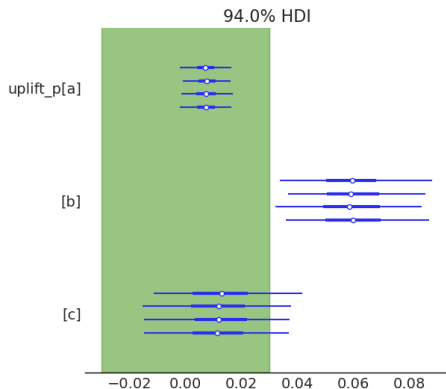


Figure: Example ROPE plot

# Interpreting the Posterior

How can we calculate a better metric?

It could look like this:

# Interpreting the Posterior

How can we calculate a better metric?

- 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 metric?

- 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 metric?

- 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 metric?

- 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
  - Positive
  - Negative
  - Average
- You connect the experiment with business
- Compare outcomes with uncertainty

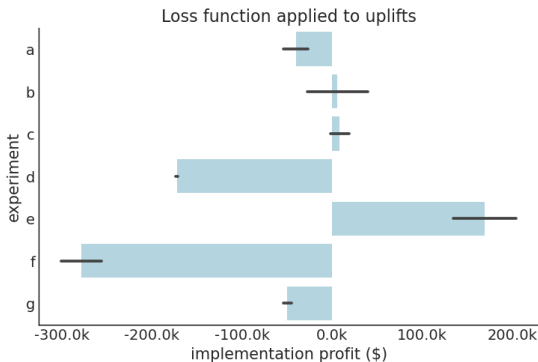


Figure:  $p(\text{Monetization}_G \mid X_G)$

# Takeouts

Real Life AB testing is full of challenges. Bayesian tools are still considered novel.

- 1 Framing the statistical test
  - Setting priors
  - Setting likelihood
- 2 Decision making before the test
  - Parameter recovery study
- 3 Bayesian decision making
  - Loss functions
  - Scenario testing



# Supplementary

# Parameter Recovery for AB testing

Given:

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

Recall the model

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

# 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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

# 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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

# 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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$



# 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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

# 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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

## 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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

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):
  - $P(p \text{ is negative} \mid X_{1:N})$
  - $P(p \text{ is neutral} \mid X_{1:N})$
  - $P(p \text{ is positive} \mid X_{1:N})$

Recall the model

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

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

$$i \in 1 \dots N$$

$$x_i \sim \text{Bernoulli}(p)$$

$$p \sim \text{Beta}(\mu = \bar{\mu}, \sigma = \bar{\sigma})$$

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