Multi-Armed Bandit

Schwartz

August 31, 2016

The Multi-Armed Bandit

"Originally considered by Allied scientists in World War II, it proved so intractable that, according to Peter Whittle, the problem was proposed to be dropped over German so that German scientists could also waste their time on it."

– Erich Wellinger, commenting on "Discussion of Dr Gittins' paper", presented in the Journal of the Royal Statistical Society, Series B, issue 41, 1979.

The Multi-Armed Bandit

The Multi-Armed Bandit

Objectives

1. Understand A/B testing inside and out
2. From frequentist *and* Bayesian perspectives
   - Stone cold mastery of hypothesis testing
   - Bayesian paradigm on lock
     - ability to wax eloquent if (not ad nauseam) about "priors"
3. Ability to deploy a multi-armed bandit solution to A/B testing at the drop of a hat
   - ϵ-greedy, softmax, UCB1, and the Bayesian Bandit served up like they were in the list of 31 flavors
4. Regret: all about it and having none of it

Let's start a little simpler: A/B Testing

Trial 1	
Trial 2	
Trial 3	
Trial 4	
⋮	
Trial $n_A + n_B$	
Total	$\hat{p}_A = \frac{\sum x_A^{(i)}}{n_A}$ $\hat{p}_B = \frac{\sum x_B^{(j)}}{n_B}$

A/B **Hypothesis Testing**

$$\hat{p}_A = \frac{\sum X_A^{(i)}}{n_A} \qquad \hat{p}_B = \frac{\sum X_B^{(j)}}{n_B}$$

$$X_A^{(i)} \sim \text{Bern}(\theta_A) = \text{Binomial}(\theta_A, N_A = 1)$$
$$X_B^{(j)} \sim \text{Bern}(\theta_B) = \text{Binomial}(\theta_B, N_B = 1)$$

IF $\theta_A = \theta_B$ $[H_0]$

THEN $Var\left(X_A^{(i)}\right) = Var\left(X_B^{(j)}\right) = \ ?$

SO $\hat{p}_A - \hat{p}_B \sim \ ?$ [By CLT]

AND what is a good estimator of $\theta = \theta_A = \theta_B$?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?
- How do we get it?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

  probability the new site is better than the old site (i.e. H_0 false)?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

  probability the new site is better than the old site (i.e. H_0 false)?

  probability both sites are the same (i.e. H_0 true)?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

  probability the new site is better than the old site (i.e. H_0 false)?

  probability both sites are the same (i.e. H_0 true)?

  probability that we correctly concluded the new is better than the old?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

  probability the new site is better than the old site (i.e. H_0 false)?

  probability both sites are the same (i.e. H_0 true)?

  probability that we correctly concluded the new is better than the old?

  probability that we incorrectly concluded the new is better than the old?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

- Can we say there's a
  $100 \cdot (1 - p)\%$ chance the new site is better?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

- Can we say there's a

  $100 \cdot (1 - p)\%$ chance the new site is better?

  $100 \cdot p\%$ chance that both sites are the same?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

- Can we say there's a

  $100 \cdot (1 - p)\%$ chance the new site is better?

  $100 \cdot p\%$ chance that both sites are the same?

  $100 \cdot (1 - \alpha)\%$ chance the new site is better?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

- Can we say there's a

  $100 \cdot (1 - p)\%$ chance the new site is better?

  $100 \cdot p\%$ chance that both sites are the same?

  $100 \cdot (1 - \alpha)\%$ chance the new site is better?

  $100 \cdot \alpha\%$ chance that both sites are the same?

A/B **Hypothesis Testing** *questions*

Suppose we think a new site is better than an old site

- What's the α significance level we're testing at?

- We observed a very small p-value, p... what's the...

- Can we say there's a

  $100 \cdot (1 - p)\%$ chance the new site is better?

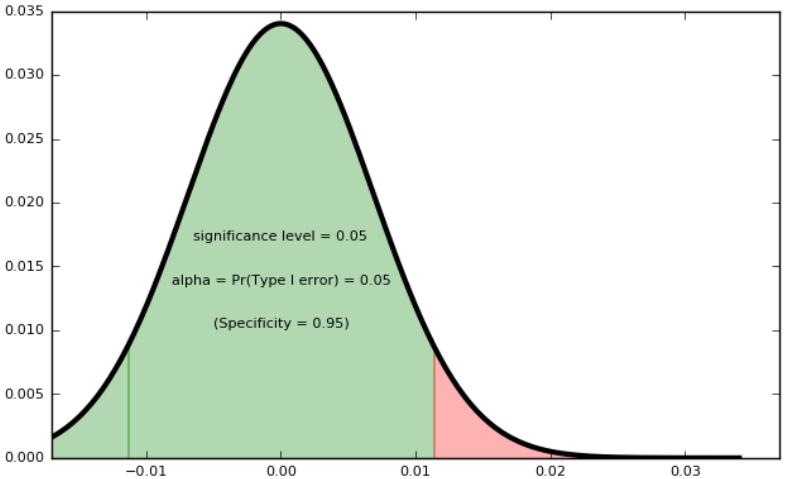
  $100 \cdot p\%$ chance that both sites are the same?

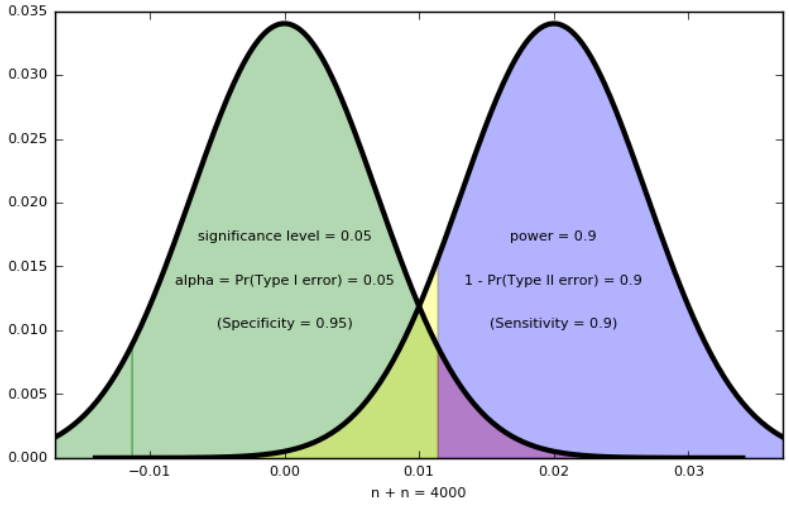
  $100 \cdot (1 - \alpha)\%$ chance the new site is better?

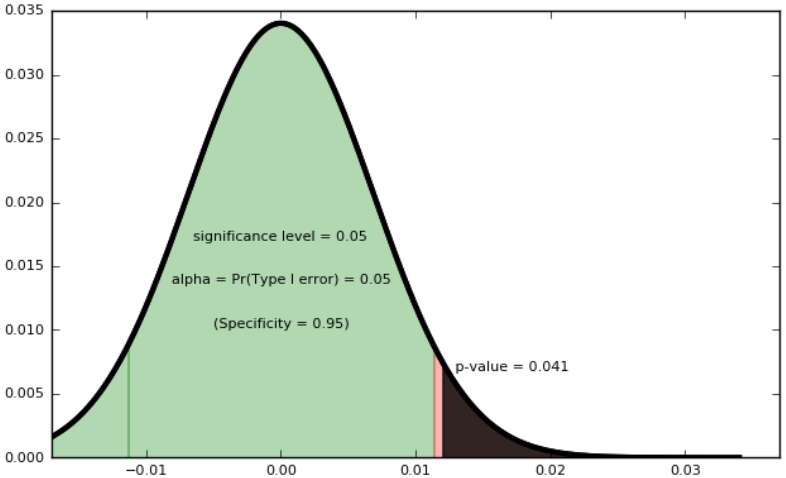
  $100 \cdot \alpha\%$ chance that both sites are the same?

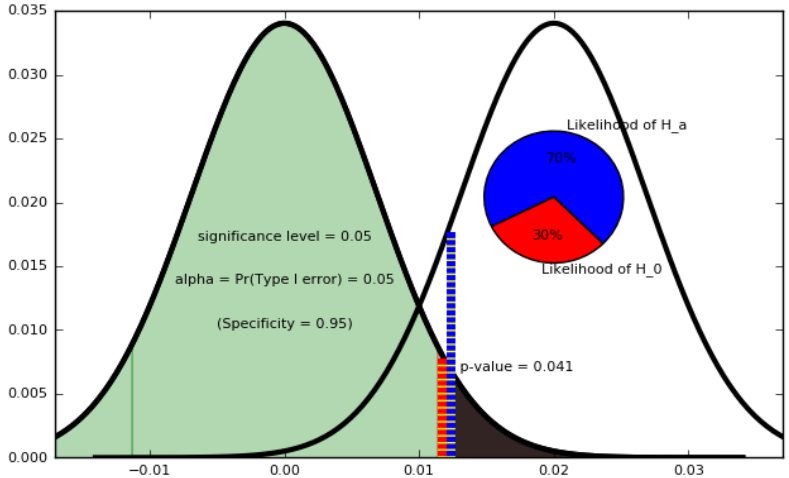
  $100 \cdot \alpha\%$ chance that we incorrectly concluded the sites are different?

A/B **Hypothesis Testing** *significance*



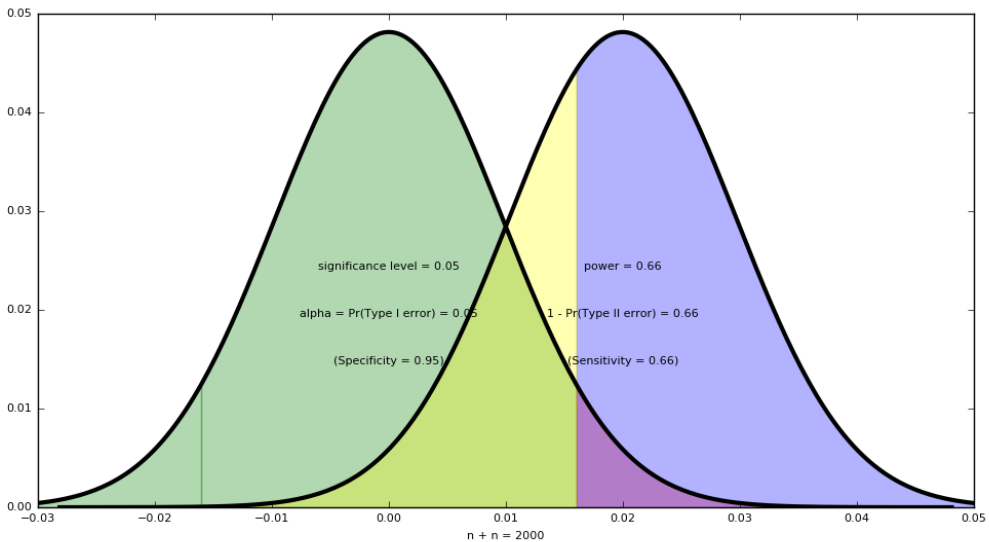


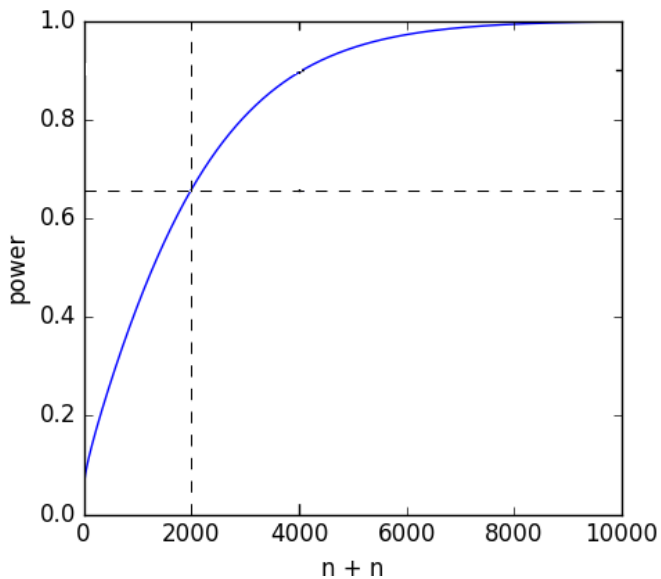




A/B **Hypothesis Testing** *power calculations*

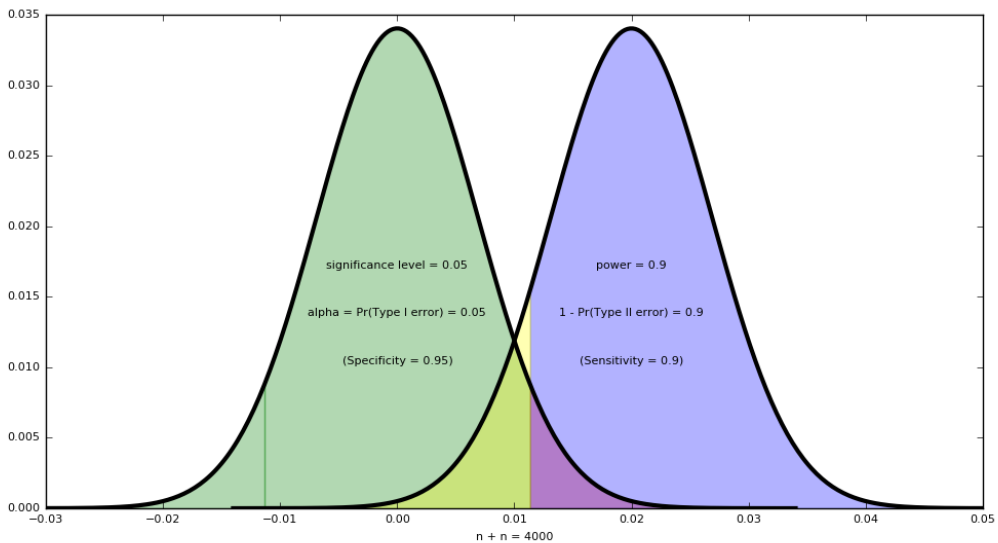
- Effect size is $0.06 - 0.04$
- With $2n$ alternating trials
  $\hat{p} = 0.05$
- $se = \sqrt{2\frac{\hat{p}(1-\hat{p})}{n}}$

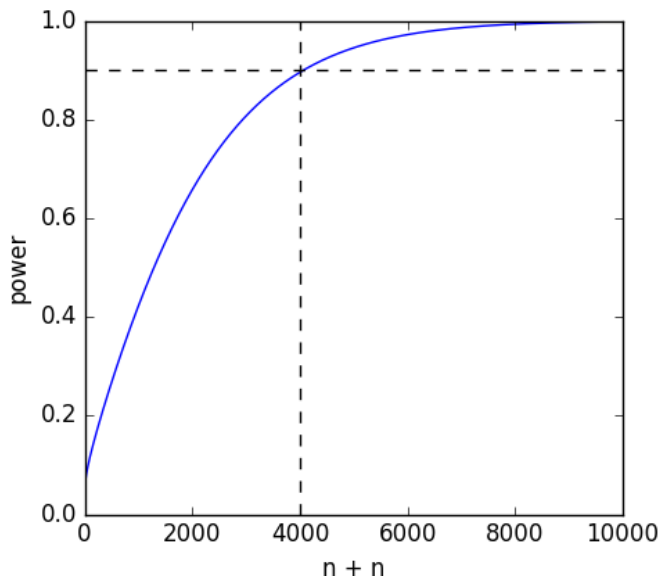




A/B **Hypothesis Testing** *power calculations*

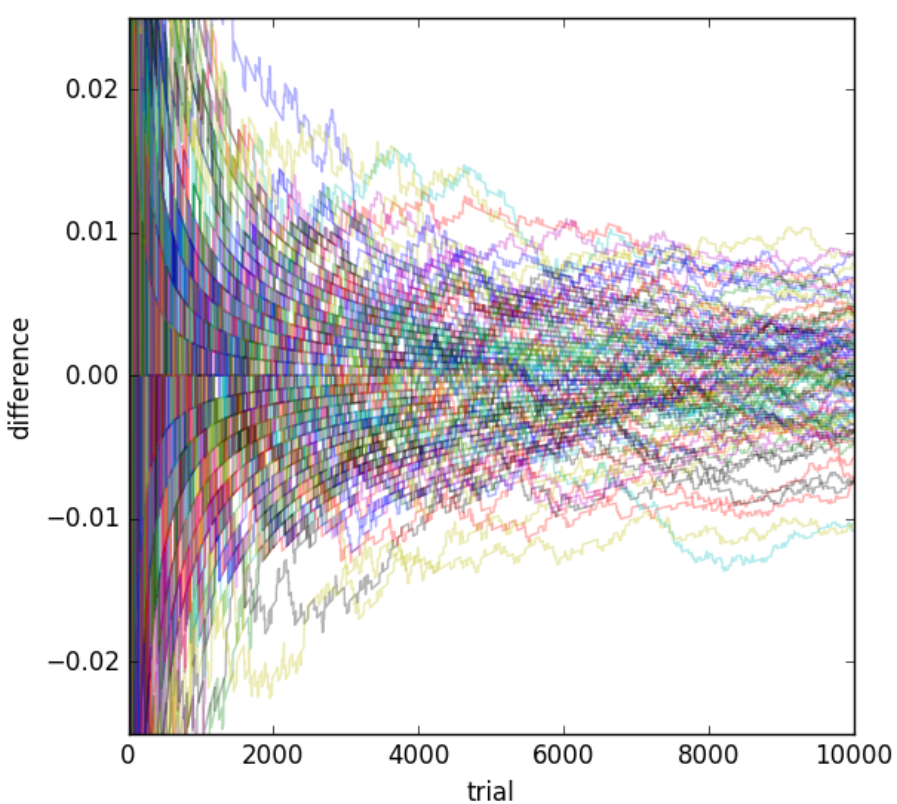
- Effect size is $0.06 - 0.04$
- With $2n$ alternating trials
  $\hat{p} = 0.05$
- $se = \sqrt{2\frac{\hat{p}(1-\hat{p})}{n}}$

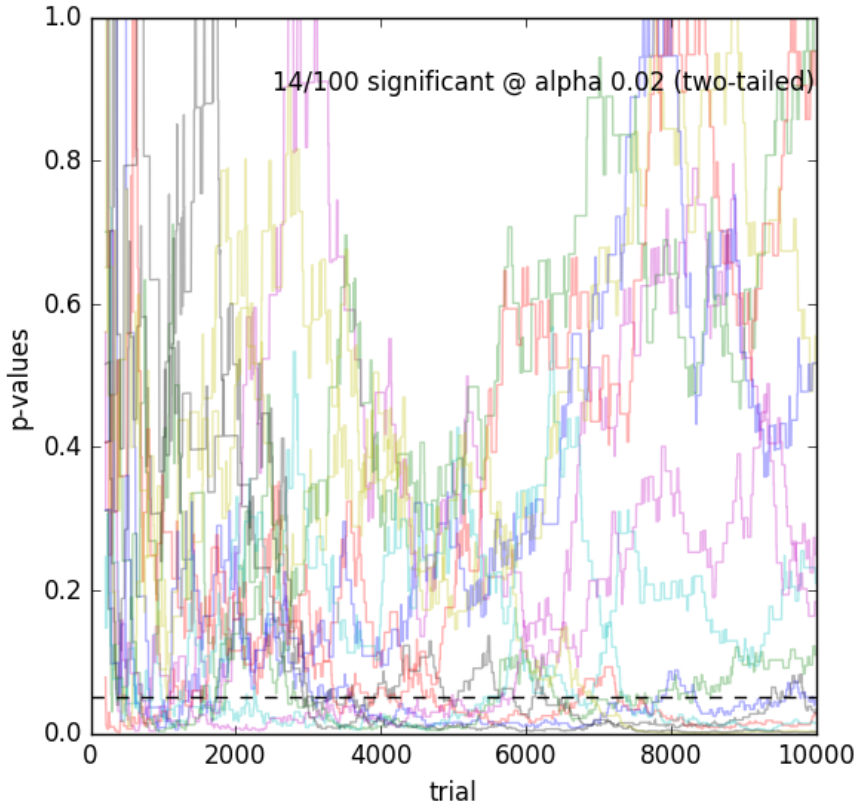




Multiple A/B Testing

- There is no difference in conversion rates in these simulations





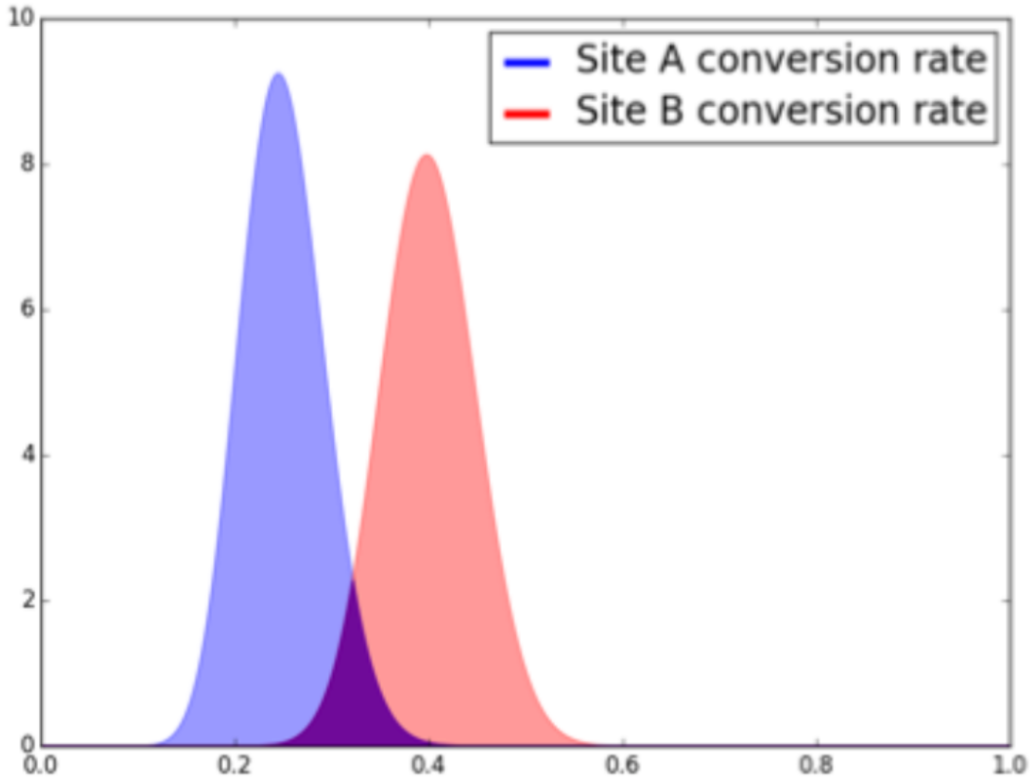
- Continuous (multiple) testing does not achieve α-significance

Bayesian *philosophy*

- Assume θ_A and θ_B have distributions $f(\theta_A)$ and $f(\theta_B)$
  E.g., each time a customer arrives at landing page $[A, B]$
  the probability of conversion is drawn from $[f(\theta_A), f(\theta_B)]$

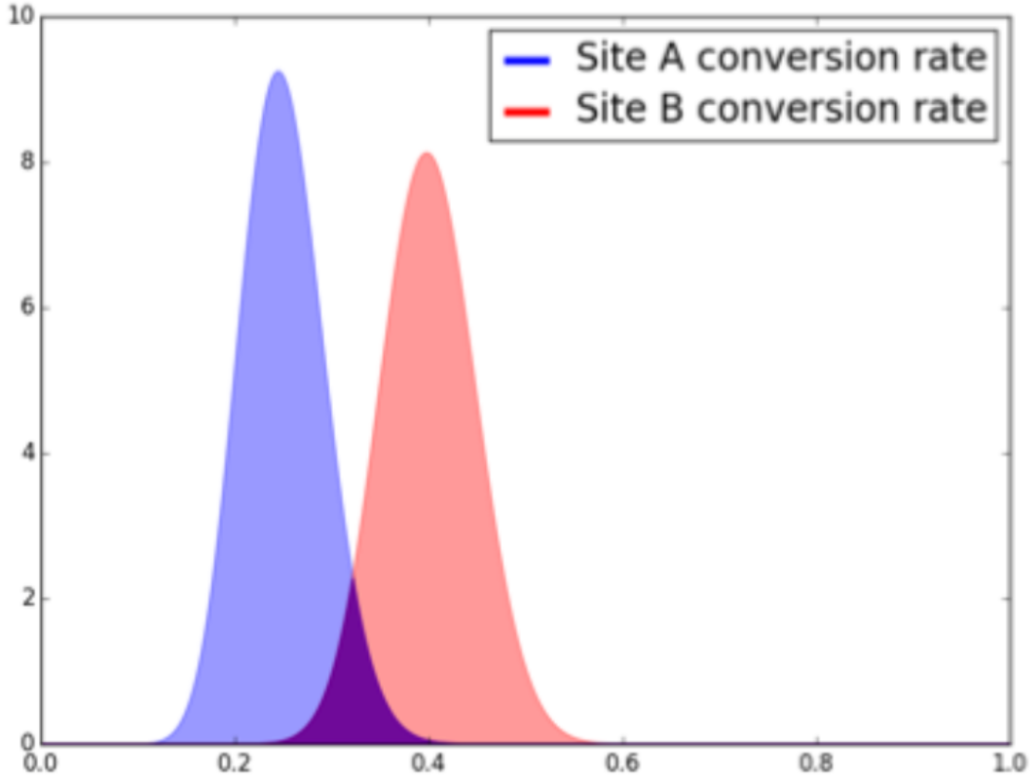
Bayesian *philosophy*

- Assume θ_A and θ_B have distributions $f(\theta_A)$ and $f(\theta_B)$
  E.g., each time a customer arrives at landing page $[A, B]$ the probability of conversion is drawn from $[f(\theta_A), f(\theta_B)]$



Bayesian *philosophy*

- Assume θ_A and θ_B have distributions $f(\theta_A)$ and $f(\theta_B)$
  E.g., each time a customer arrives at landing page $[A, B]$ the probability of conversion is drawn from $[f(\theta_A), f(\theta_B)]$



- Interest is in which distribution produces more conversions
  E.g., the distributions of $\theta_A > \theta_B$, or $\theta_A - \theta_B$, or $\frac{\theta_A}{\theta_B}$

Bayesian *transformations*

- Distributions of variable transformations can be approximated by performing the transformation on samples of the variables

$$f(x) = \int x df(x) \neq \int x f(x) dx$$

$$\approx \text{return } x\text{'s for } x\text{'s sampled according to } f(x)$$

$$f(g(x)) = \int g(x) df(x)$$

$$\approx \text{return } g(x)\text{'s for } x\text{'s sampled according to } f(x)$$

Bayesian *transformations*

- Distributions of variable transformations can be approximated by performing the transformation on samples of the variables

$$f(x) = \int x df(x) \neq \int x f(x) dx$$

$$\approx \text{return } x\text{'s for } x\text{'s sampled according to } f(x)$$

$$f(g(x)) = \int g(x) df(x)$$

$$\approx \text{return } g(x)\text{'s for } x\text{'s sampled according to } f(x)$$

- Thus, for $g(\theta_A, \theta_B) = \left[\theta_A > \theta_B \text{ or } \theta_A - \theta_B \text{ or } \frac{\theta_A}{\theta_B}\right]$

$$f(g(\theta_A, \theta_B)) = \int g(\theta_A, \theta_B) df(\theta_A, \theta_B)$$

$$= \int \int g(\theta_A, \theta_B) df(\theta_A) df(\theta_B)$$

This Slide Contains Complicated Bayesian Analogs to Hypothesis Testing Interpretation Difficulties & Weirdness

There are none

Bayesian analysis is just distribution probability

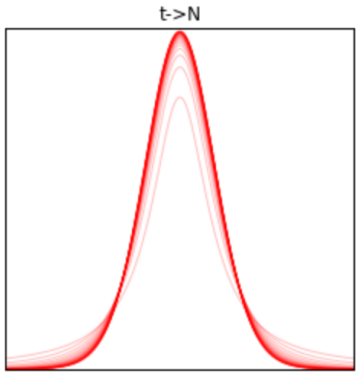
This Slide Contains Complicated Bayesian Analogs to Hypothesis Testing Interpretation Difficulties & Weirdness

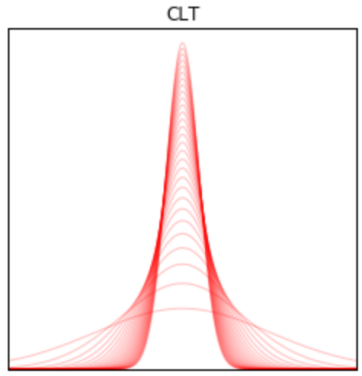
But even that's too much effort for a good Bayesian...

To validate probability statements we just infinity sample posteriors

A brief synopsis of the world

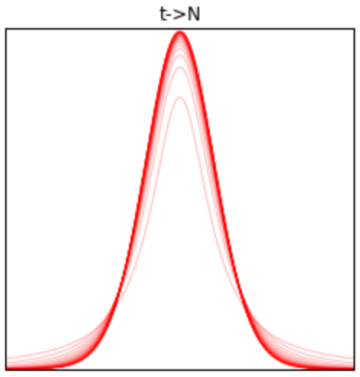
A brief synopsis of the world

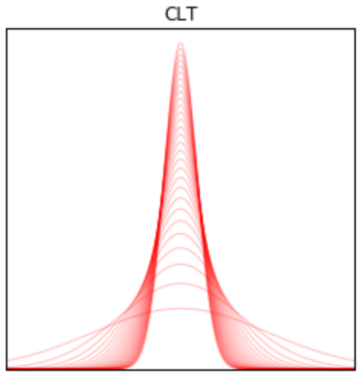


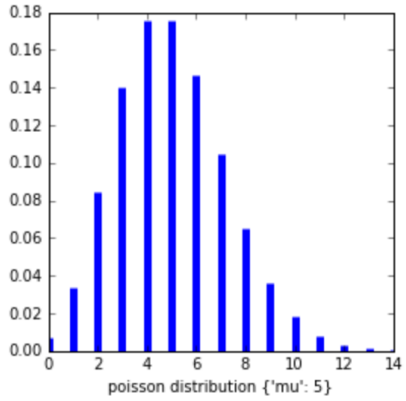


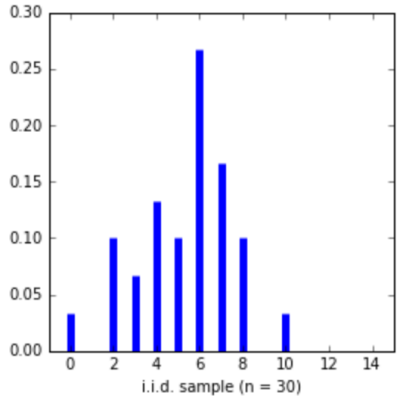
$$n \longrightarrow \infty$$

A brief synopsis of the world



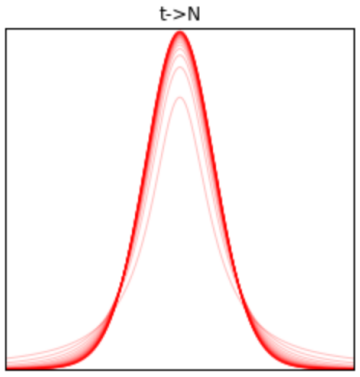


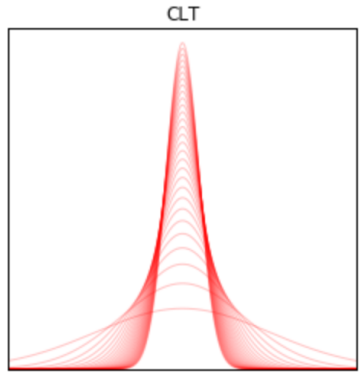


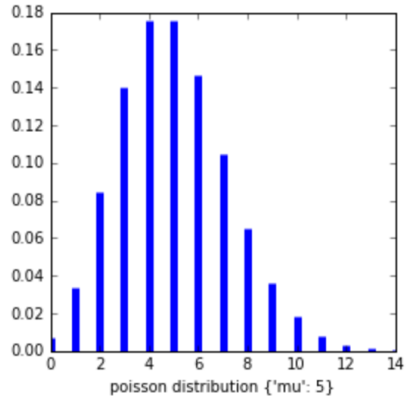


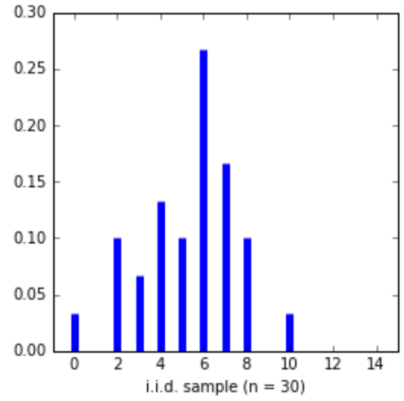
$$n \longrightarrow \infty$$

A brief synopsis of the world

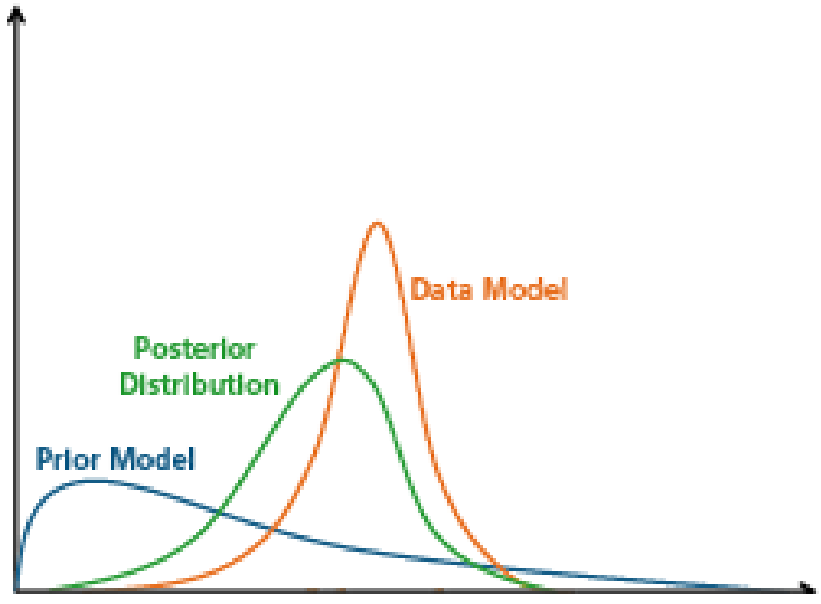


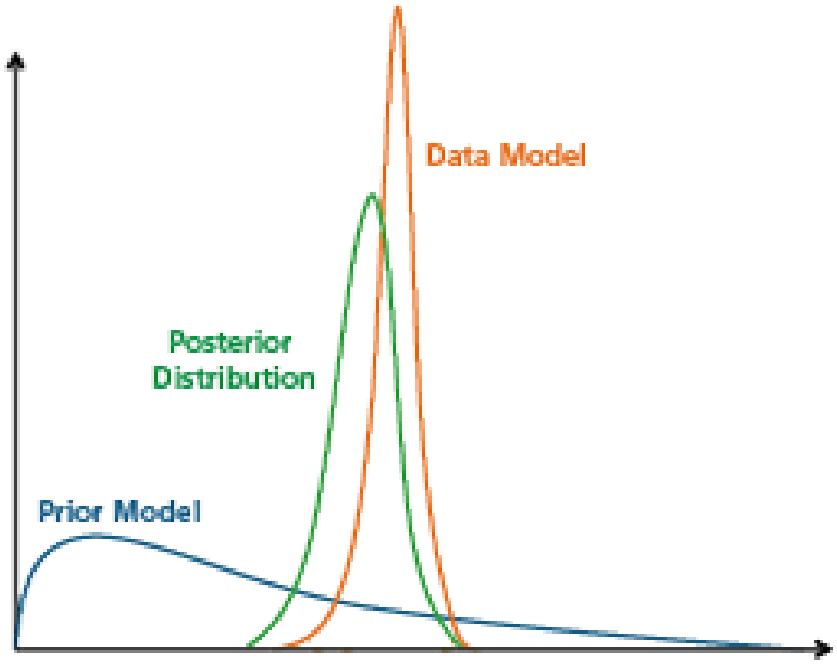




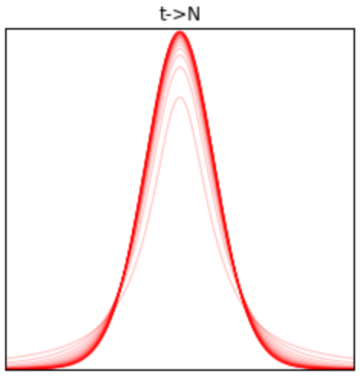


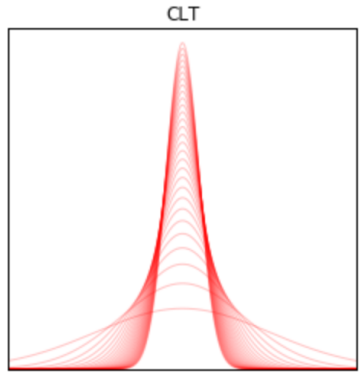
$$n \longrightarrow \infty$$

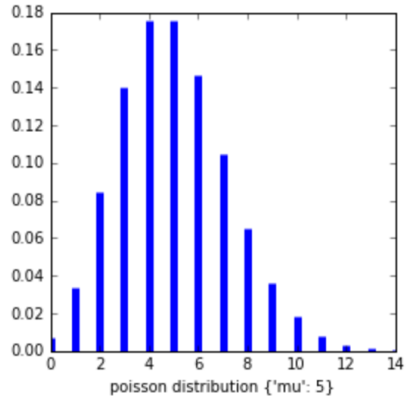


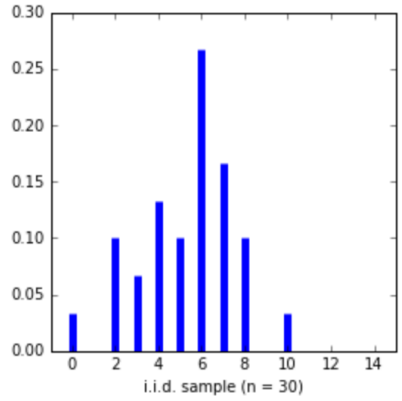


A brief synopsis of the world





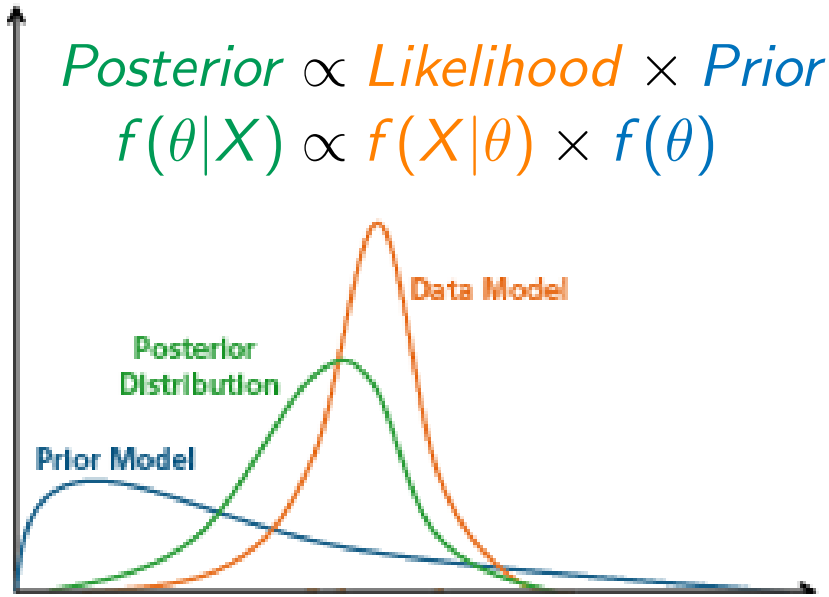


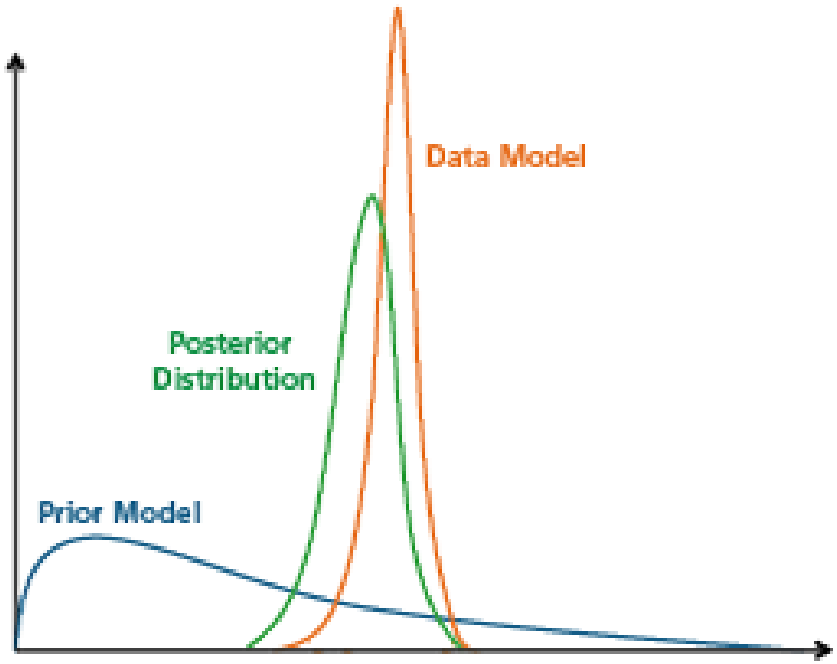


$$n \longrightarrow \infty$$

$$Posterior \propto Likelihood \times Prior$$

$$f(\theta|X) \propto f(X|\theta) \times f(\theta)$$





Parametric Versus Nonparametric

We might characterize an analysis framework as parametric if

1. Results are buttressed or bolstered by modeling assumptions

Parametric Versus Nonparametric

We might characterize an analysis framework as parametric if

1. Results are buttressed or bolstered by modeling assumptions
   E.g., leveraging the structure of normality via a t-test increases power but comes at a cost of loss of robustness compared to nonparametric tests free of distributional assumptions.

Parametric Versus Nonparametric

We might characterize an analysis framework as parametric if

1. Results are buttressed or bolstered by modeling assumptions
   E.g., leveraging the structure of normality via a t-test increases power but comes at a cost of loss of robustness compared to nonparametric tests free of distributional assumptions.
2. Predicted values are based on "parameters."

Parametric Versus Nonparametric

We might characterize an analysis framework as parametric if

1. Results are buttressed or bolstered by modeling assumptions
   E.g., leveraging the structure of normality via a t-test increases power but comes at a cost of loss of robustness compared to nonparametric tests free of distributional assumptions.
2. Predicted values are based on "parameters."
   E.g., the β coefficients in linear regression.

Parametric Versus Nonparametric

We might characterize an analysis framework as parametric if

1. Results are buttressed or bolstered by modeling assumptions
   E.g., leveraging the structure of normality via a t-test increases power but comes at a cost of loss of robustness compared to nonparametric tests free of distributional assumptions.
2. Predicted values are based on "parameters."
   E.g., the β coefficients in linear regression.
3. Parameter estimation determines the specific instance of a model within a "model class" defined by those parameters.

Parametric Versus Nonparametric

We might characterize an analysis framework as parametric if

1. Results are buttressed or bolstered by modeling assumptions
   E.g., leveraging the structure of normality via a t-test increases power but comes at a cost of loss of robustness compared to nonparametric tests free of distributional assumptions.
2. Predicted values are based on "parameters."
   E.g., the β coefficients in linear regression.
3. Parameter estimation determines the specific instance of a model within a "model class" defined by those parameters.
   E.g., the CLT guarantees a normal distribution which is determined by estimating μ and $\sigma^2/\sqrt{n}$ with n large

Parametric Versus Nonparametric

We might characterize an analysis framework as parametric if

1. Results are buttressed or bolstered by modeling assumptions
   E.g., leveraging the structure of normality via a t-test increases power but comes at a cost of loss of robustness compared to nonparametric tests free of distributional assumptions.
2. Predicted values are based on "parameters."
   E.g., the β coefficients in linear regression.
3. Parameter estimation determines the specific instance of a model within a "model class" defined by those parameters.
   E.g., the CLT guarantees a normal distribution which is determined by estimating μ and $\sigma^2/\sqrt{n}$ with n large
4. The complexity of the model does grows as data size n grows.

Parametric Versus Nonparametric

We might characterize an analysis framework as parametric if

1. Results are buttressed or bolstered by modeling assumptions
   E.g., leveraging the structure of normality via a t-test increases power but comes at a cost of loss of robustness compared to nonparametric tests free of distributional assumptions.
2. Predicted values are based on "parameters."
   E.g., the β coefficients in linear regression.
3. Parameter estimation determines the specific instance of a model within a "model class" defined by those parameters.
   E.g., the CLT guarantees a normal distribution which is determined by estimating μ and $\sigma^2/\sqrt{n}$ with n large
4. The complexity of the model does grows as data size *n* grows.
   E.g., trees grow in complexity as data becomes richer while a normal distribution is defined by μ and σ^2 regardless of n.

Questions snarky Bayesians ask Frequentists

Questions snarky Bayesians ask Frequentists

1. Can you say "there's 95% probability that A beats B"?

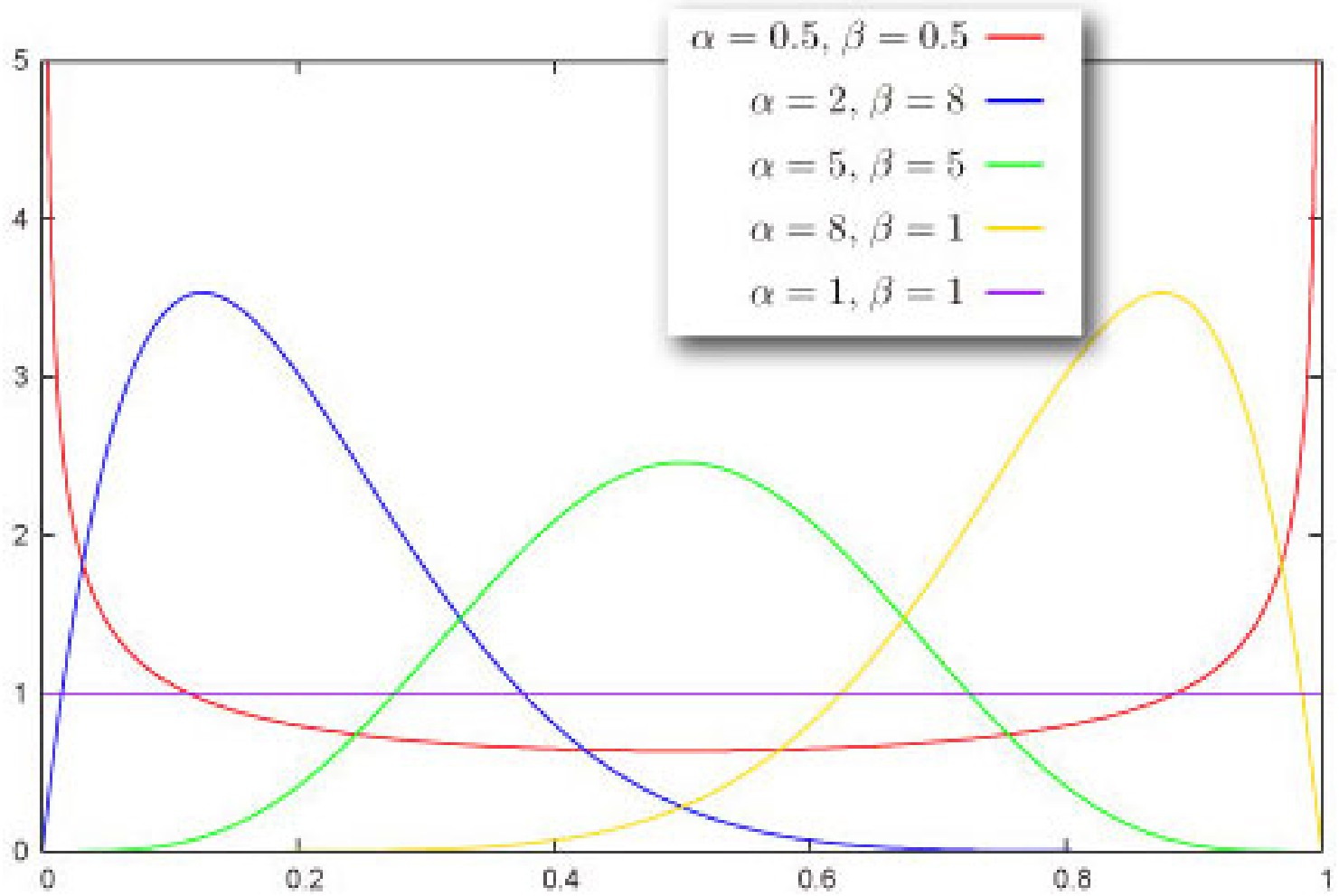
Questions snarky Bayesians ask Frequentists

1. Can you say "there's 95% probability that A beats B"?
2. Can you stop the test early based on surprising results?

Questions snarky Bayesians ask Frequentists

1. Can you say "there's 95% probability that A beats B"?
2. Can you stop the test early based on surprising results?
3. Can you update the test parameters while it's running?

Beta Distribution



$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Bayesian *theory*

- Data X that informs our knowledge of θ_A and θ_B allows us to "update" our beliefs about θ_A and θ_B using *Bayes' Theorem*

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)}$$

Bayesian *theory*

- Data X that informs our knowledge of θ_A and θ_B allows us to "update" our beliefs about θ_A and θ_B using *Bayes' Theorem*

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)}$$

- Bayes' Theorem derives the posterior as a function of the likelihood, the prior and the marginal likelihood

Bayesian *theory*

- Data X that informs our knowledge of θ_A and θ_B allows us to "update" our beliefs about θ_A and θ_B using *Bayes' Theorem*

$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)}$$

- Bayes' Theorem derives the posterior as a function of the likelihood, the prior and the marginal likelihood
- The posterior is the "distribution of theta *given* the data"

Bayesian *theory*

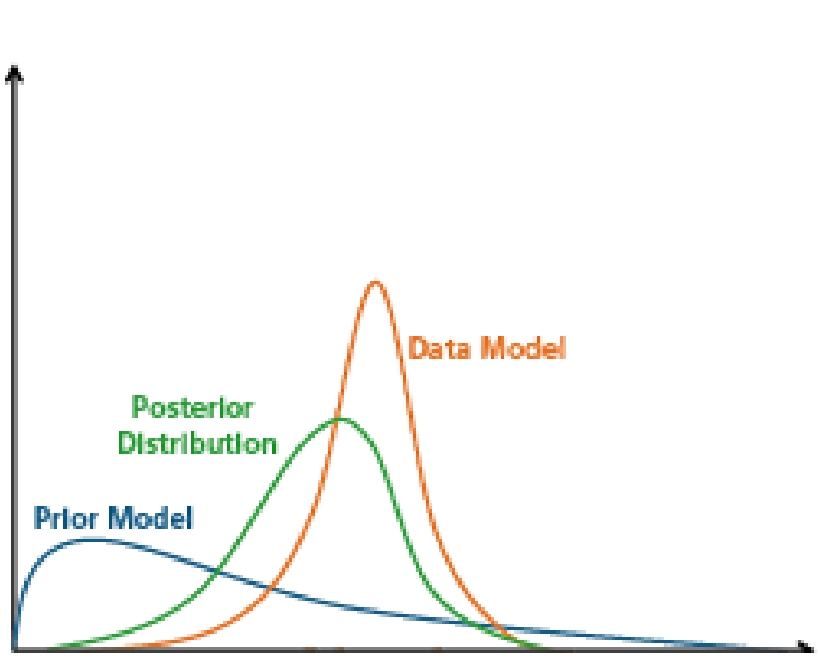
- Data X that informs our knowledge of θ_A and θ_B allows us to "update" our beliefs about θ_A and θ_B using *Bayes' Theorem*

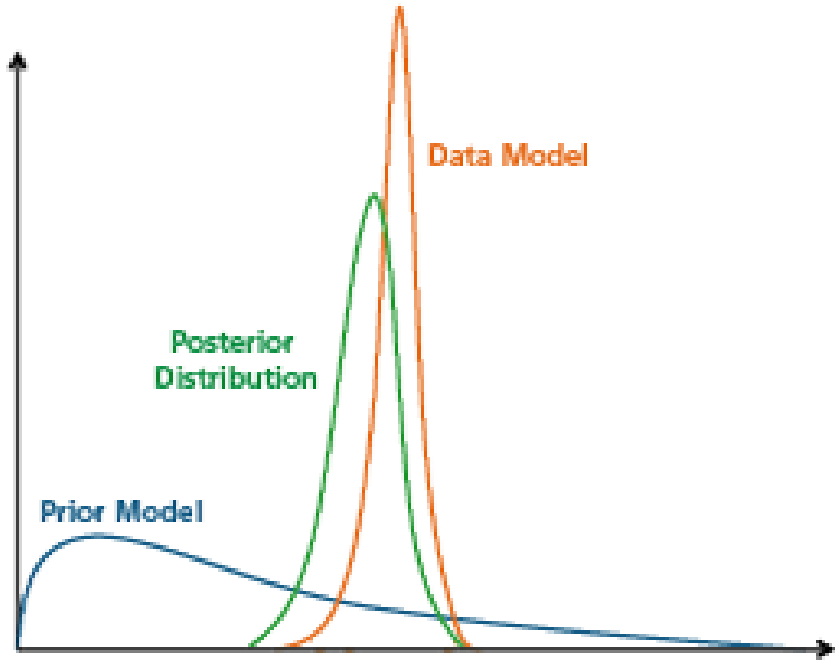
$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)}$$

- Bayes' Theorem derives the posterior as a function of the likelihood, the prior and the marginal likelihood
- The posterior is the "distribution of theta *given* the data"
- When we can recognize posterior distributions without the marginal likelihood, we abbreviate this as "proportional to"

$$f(\theta|X) \propto f(X|\theta)f(\theta)$$

Bayesian *theory*





Bayesian A/B Testing

- Likelihood

$$
\begin{aligned}
f(X_1, X_2, \cdots, X_n|\theta) &= f(X_1|\theta)f(X_2|\theta)\cdots f(X_n|\theta) \\
&= \prod \theta^{X_i}(1-\theta)^{1-X_i} \\
&= \theta^{\sum X_i}(1-\theta)^{n-\sum X_i}
\end{aligned}
$$

Bayesian A/B Testing

- Likelihood

$$\begin{aligned} f(X_1, X_2, \cdots, X_n|\theta) &= f(X_1|\theta)f(X_2|\theta)\cdots f(X_n|\theta) \\ &= \prod \theta^{X_i}(1-\theta)^{1-X_i} \\ &= \theta^{\sum X_i}(1-\theta)^{n-\sum X_i} \end{aligned}$$

- Conjugate* Prior

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \text{[beta]}$$

Bayesian A/B Testing

- Likelihood

$$\begin{aligned} f(X_1, X_2, \cdots, X_n|\theta) &= f(X_1|\theta)f(X_2|\theta)\cdots f(X_n|\theta) \\ &= \prod \theta^{X_i}(1-\theta)^{1-X_i} \\ &= \theta^{\sum X_i}(1-\theta)^{n-\sum X_i} \end{aligned}$$

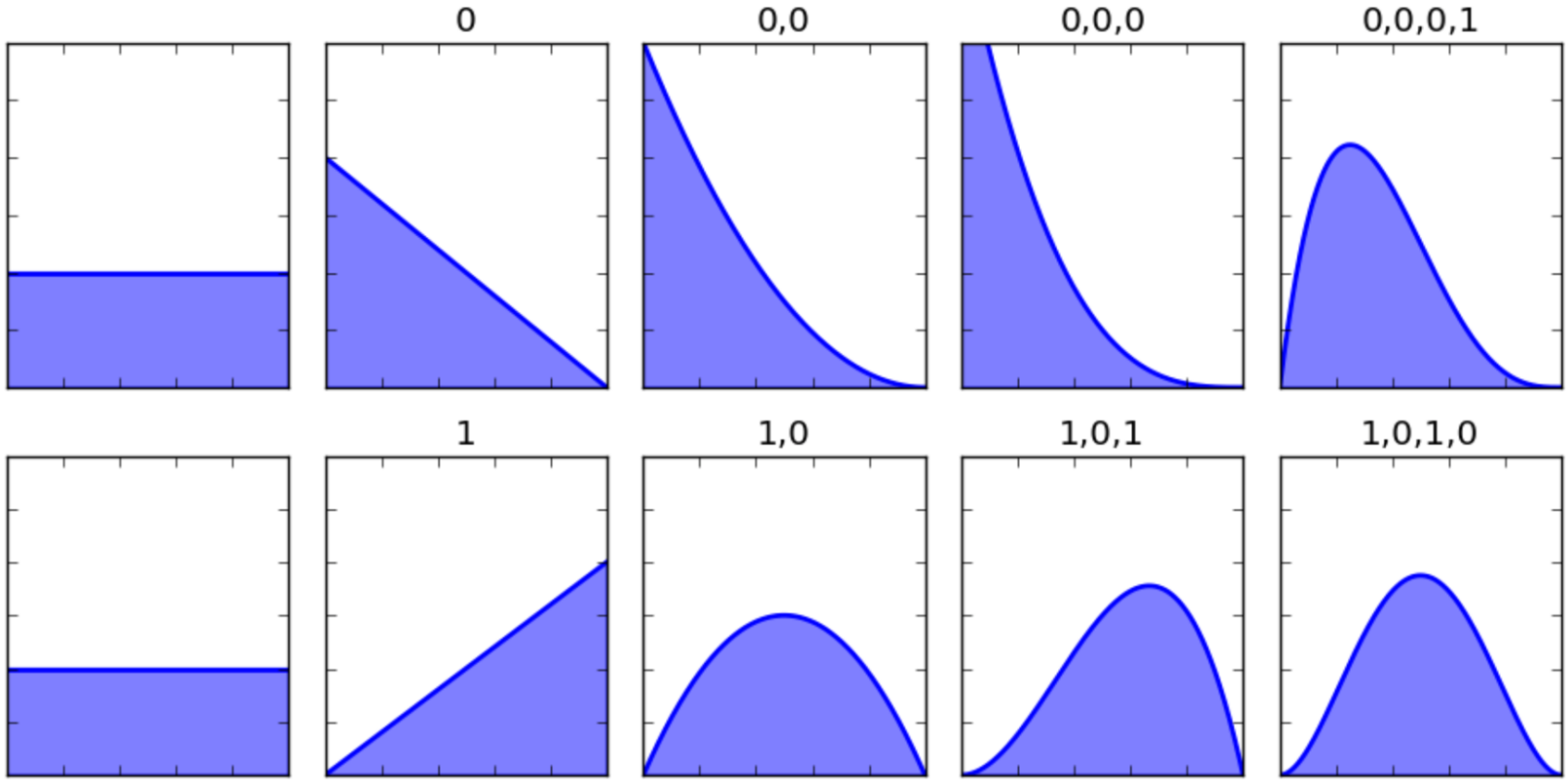
- Conjugate* Prior

$$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \text{[beta]}$$

- Posterior

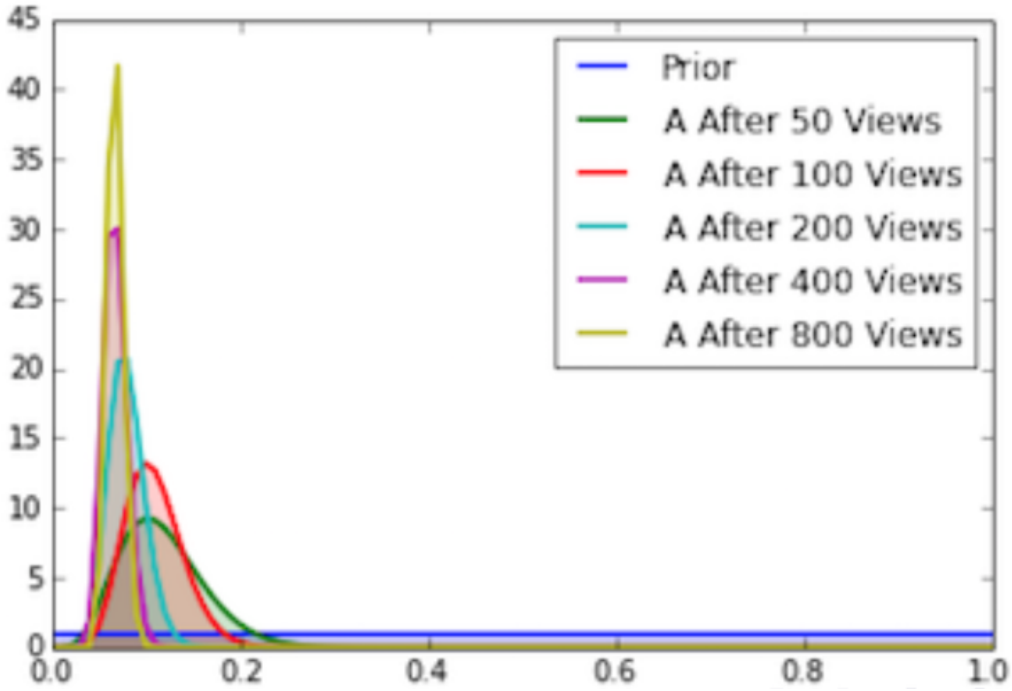
$$\begin{aligned} f(\theta|X) &\propto \theta^{\sum X_i}(1-\theta)^{n-\sum X_i}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+\sum X_i-1}(1-\theta)^{\beta+n-\sum X_i-1} \qquad \text{[beta]} \end{aligned}$$

Beta Distribution Updates



$$\text{Beta}\left(\alpha + \sum X_i, \beta + n - \sum X_i\right)$$

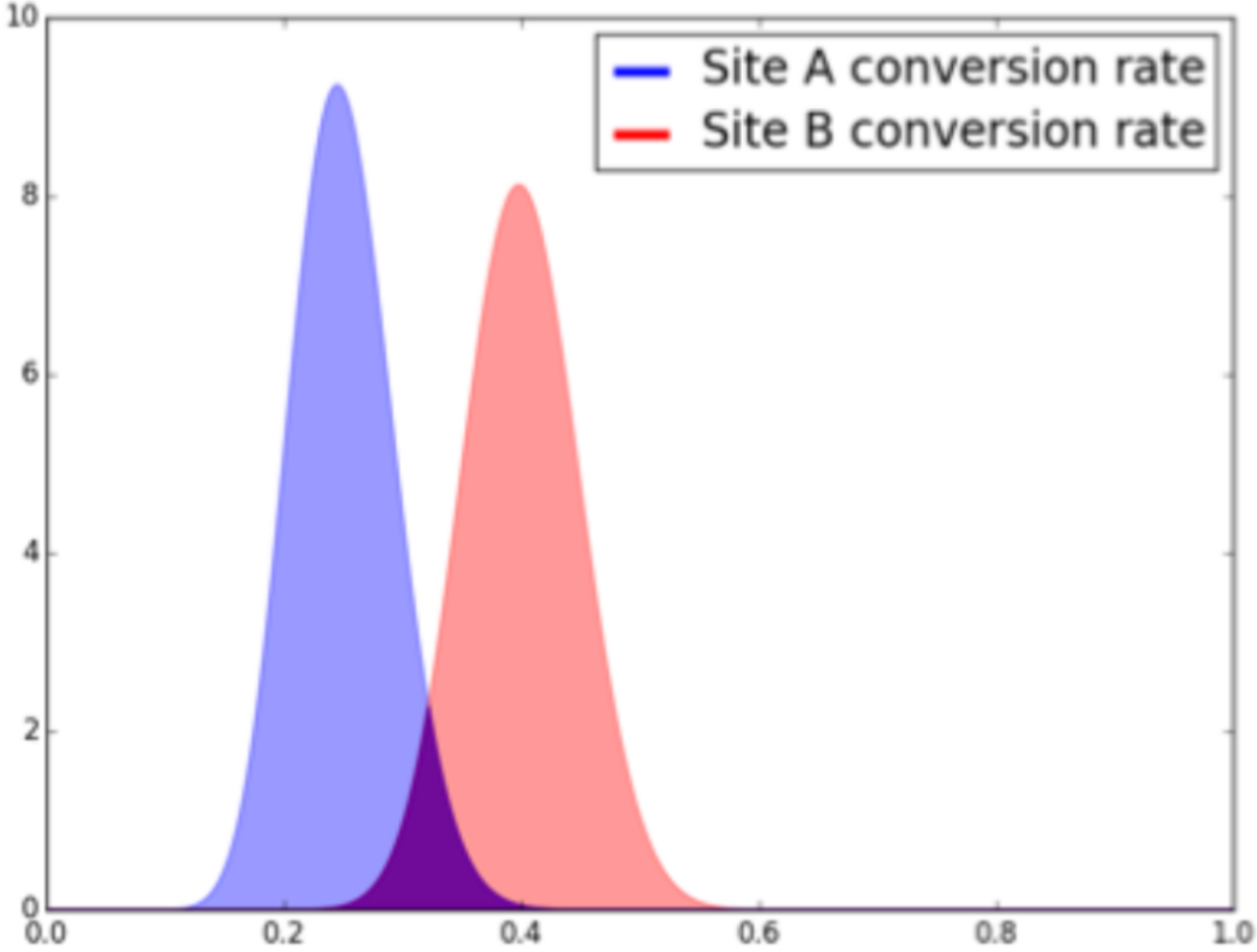
Beta Distribution Updates



Rev. Thomas Bayes (1701 – 1761)

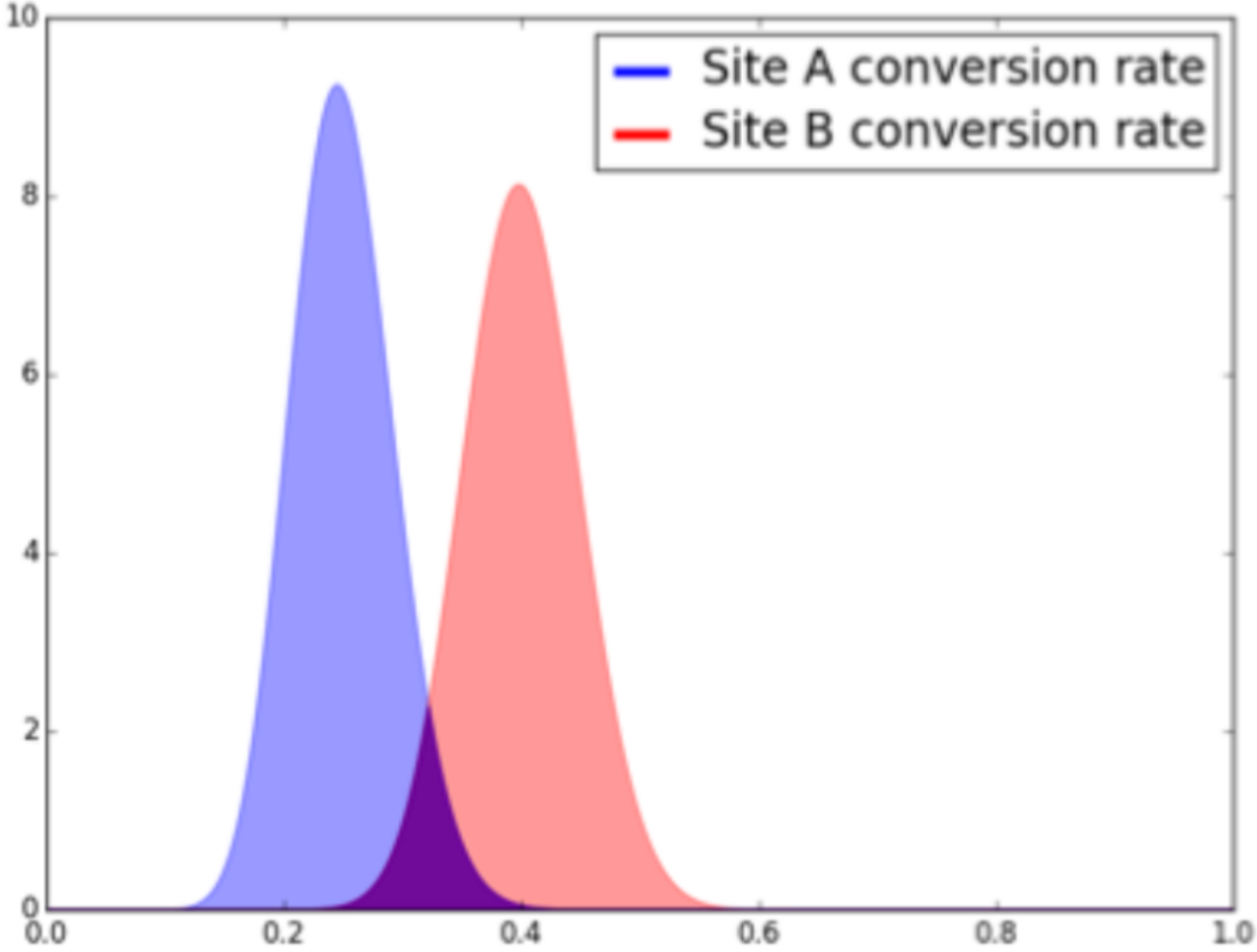
$$f(\theta|X) = \frac{f(X|\theta)f(\theta)}{f(X)} = \text{Beta}\left(\alpha + \sum X_i, \beta + n - \sum X_i\right)$$

Answers snarky Bayesians remind Frequentists of



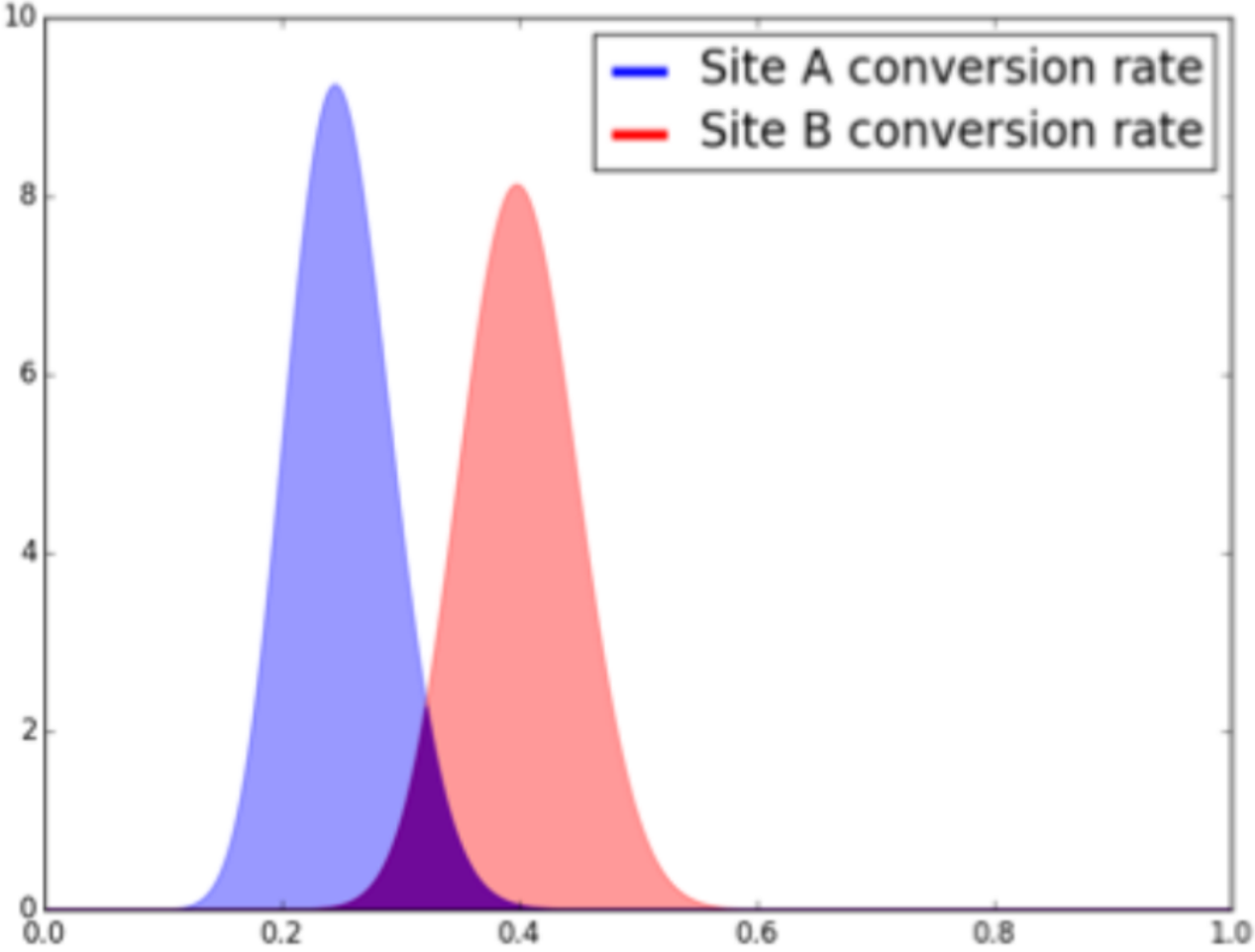
- Can you say "there's 95% probability that A beats B"?

Answers snarky Bayesians remind Frequentists of



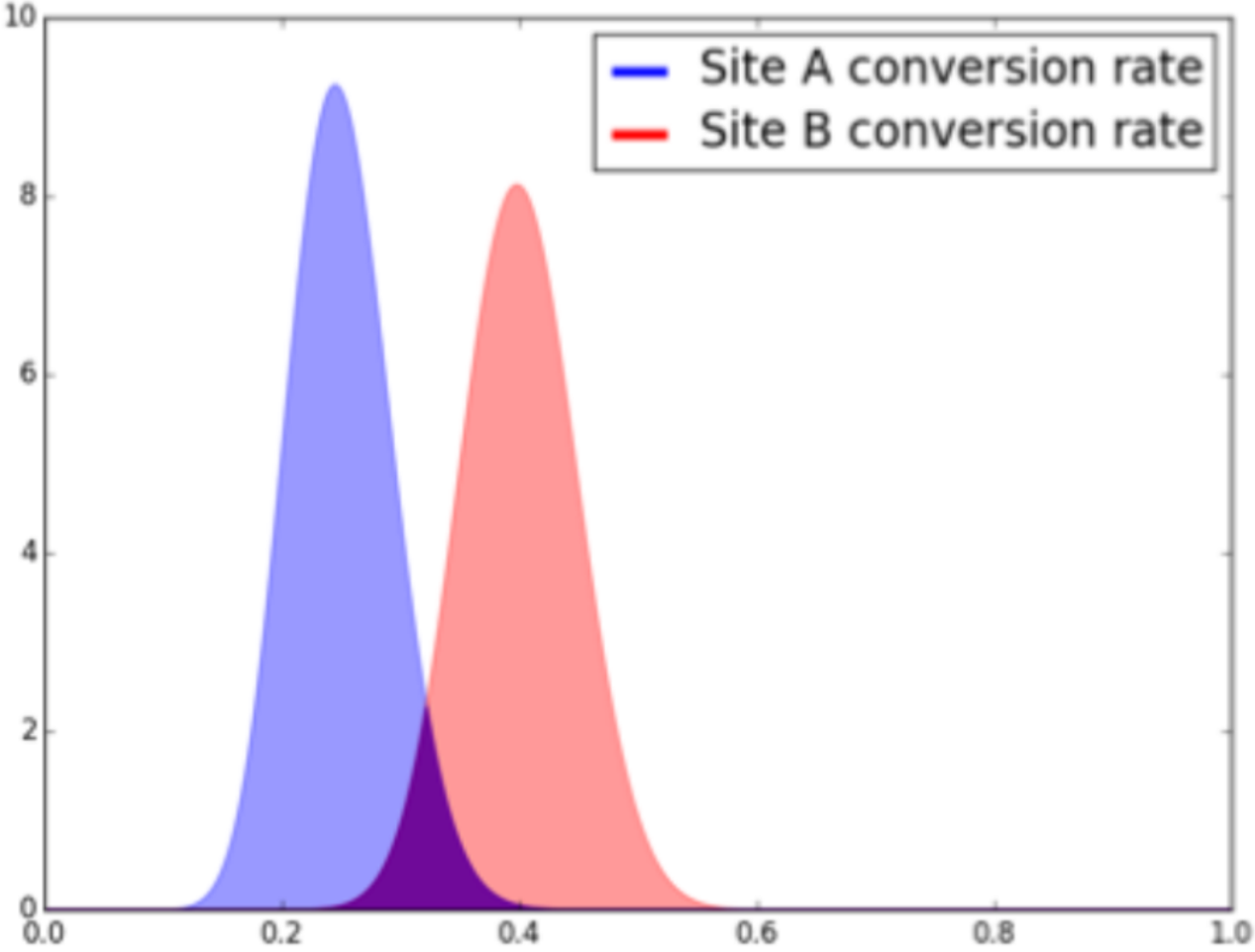
- Can you say "there's 95% probability that A beats B"? Yep.

Answers snarky Bayesians remind Frequentists of



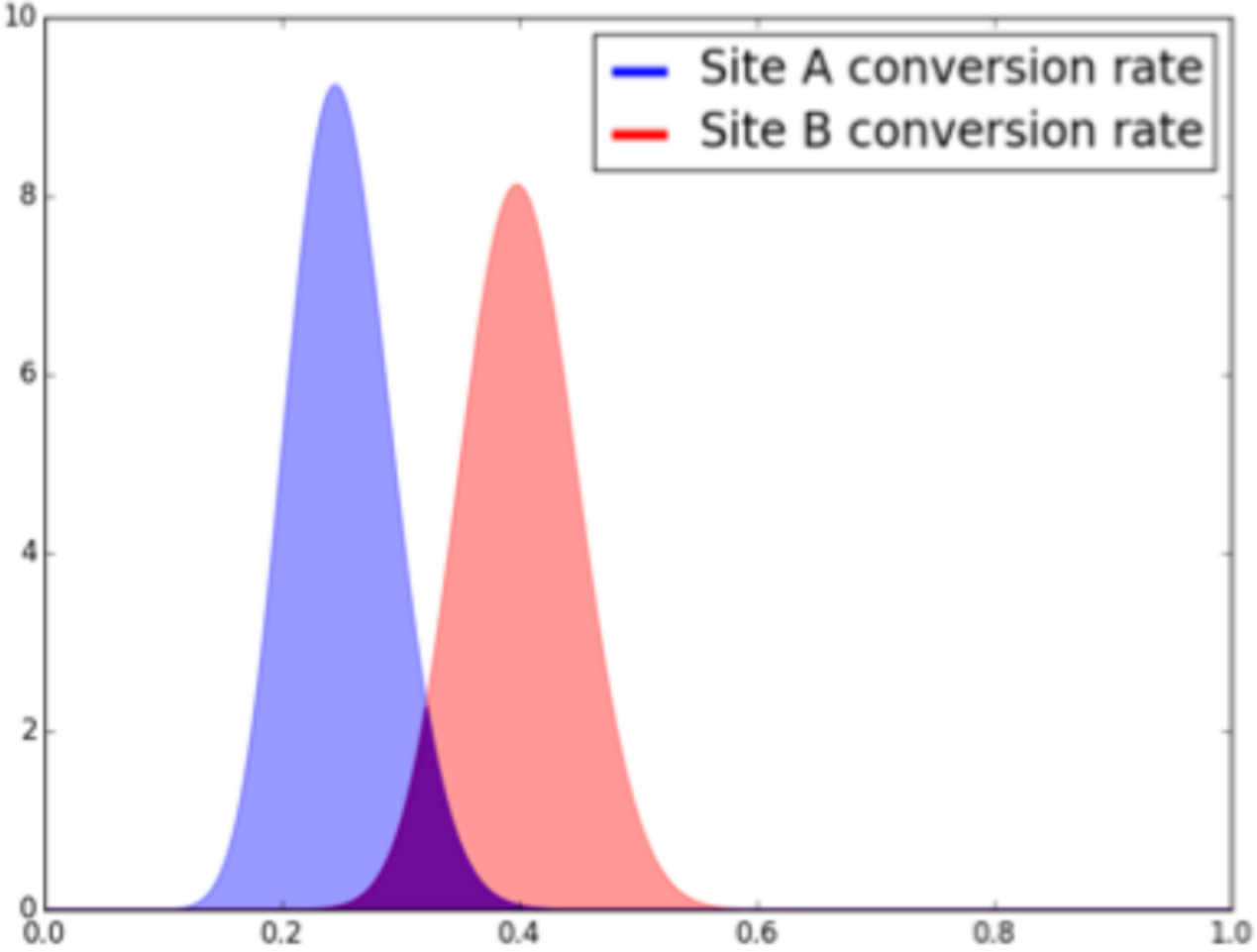
- Can you say "there's 95% probability that A beats B"? Yep.
- Can you stop the test early based on surprising results?

Answers snarky Bayesians remind Frequentists of



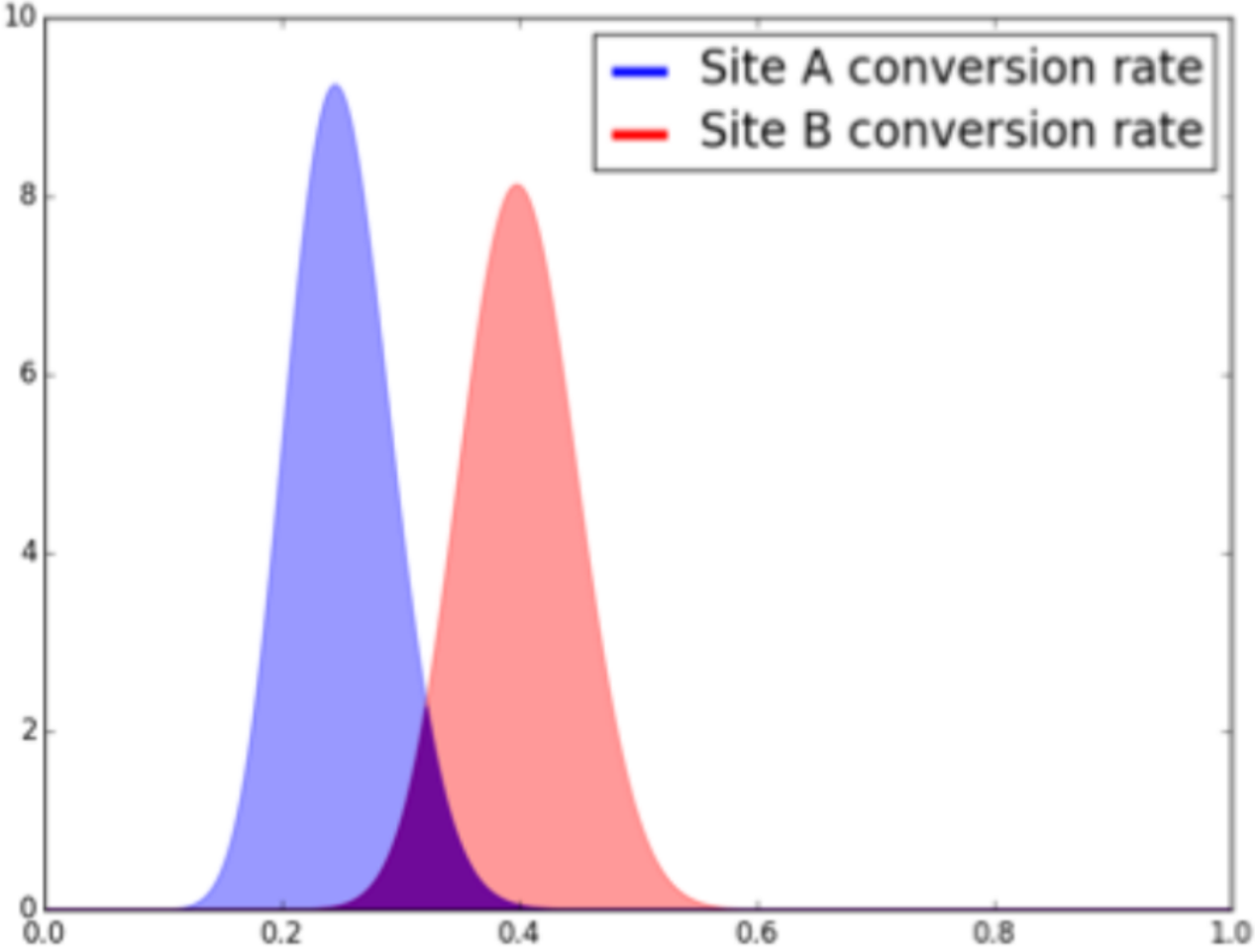
- Can you say "there's 95% probability that A beats B"? Yep.
- Can you stop the test early based on surprising results? Yep.

Answers snarky Bayesians remind Frequentists of



- Can you say "there's 95% probability that A beats B"? Yep.
- Can you stop the test early based on surprising results? Yep.
- Can you update the test parameters while it's running?

Answers snarky Bayesians remind Frequentists of



- Can you say "there's 95% probability that A beats B"? Yep.
- Can you stop the test early based on surprising results? Yep.
- Can you update the test parameters while it's running? Yep.

The Multi-Armed Bandit

The Multi-Armed Bandit

Truth / Trial	0.01	0.02	0.04	0.03	0.05	0.04	0.02
1							
2							
⋮							

The Multi-Armed Bandit

Truth / Trial	0.01	0.02	0.04	0.03	0.05	0.04	0.02
1							
2							
⋮							

Exploration & Exploitation

The Multi-Armed Bandit and *Regret*

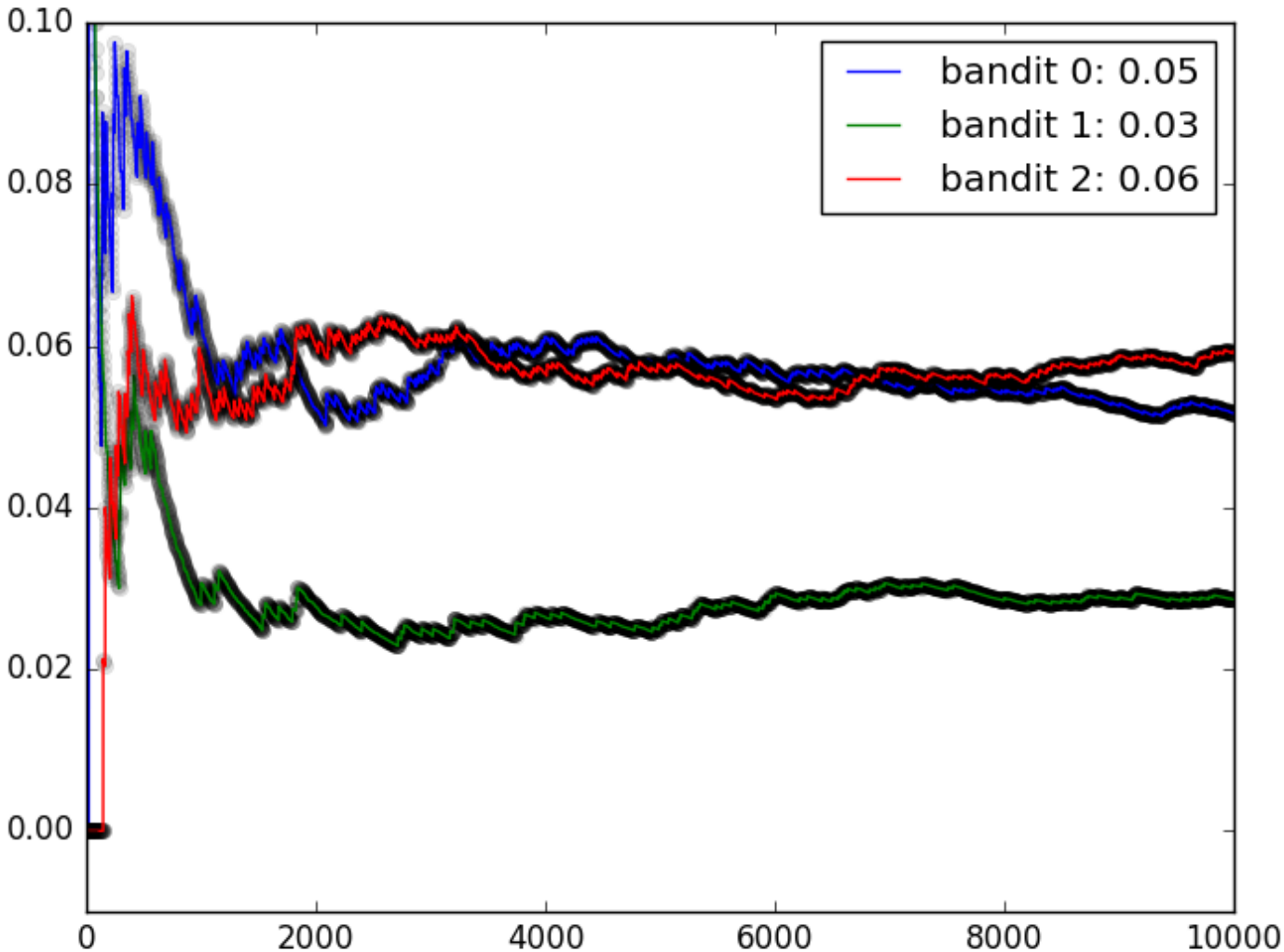
$$\text{regret} = \sum_{t}^{T} \max_{k} \theta_k - \theta^{(t)}$$

$$= T \cdot \max_{k} \theta_k - \sum_{t}^{T} \theta^{(t)}$$

The Multi-Armed *random* Bandit

Random: *A/B[/C/D/E/F/G]-test*

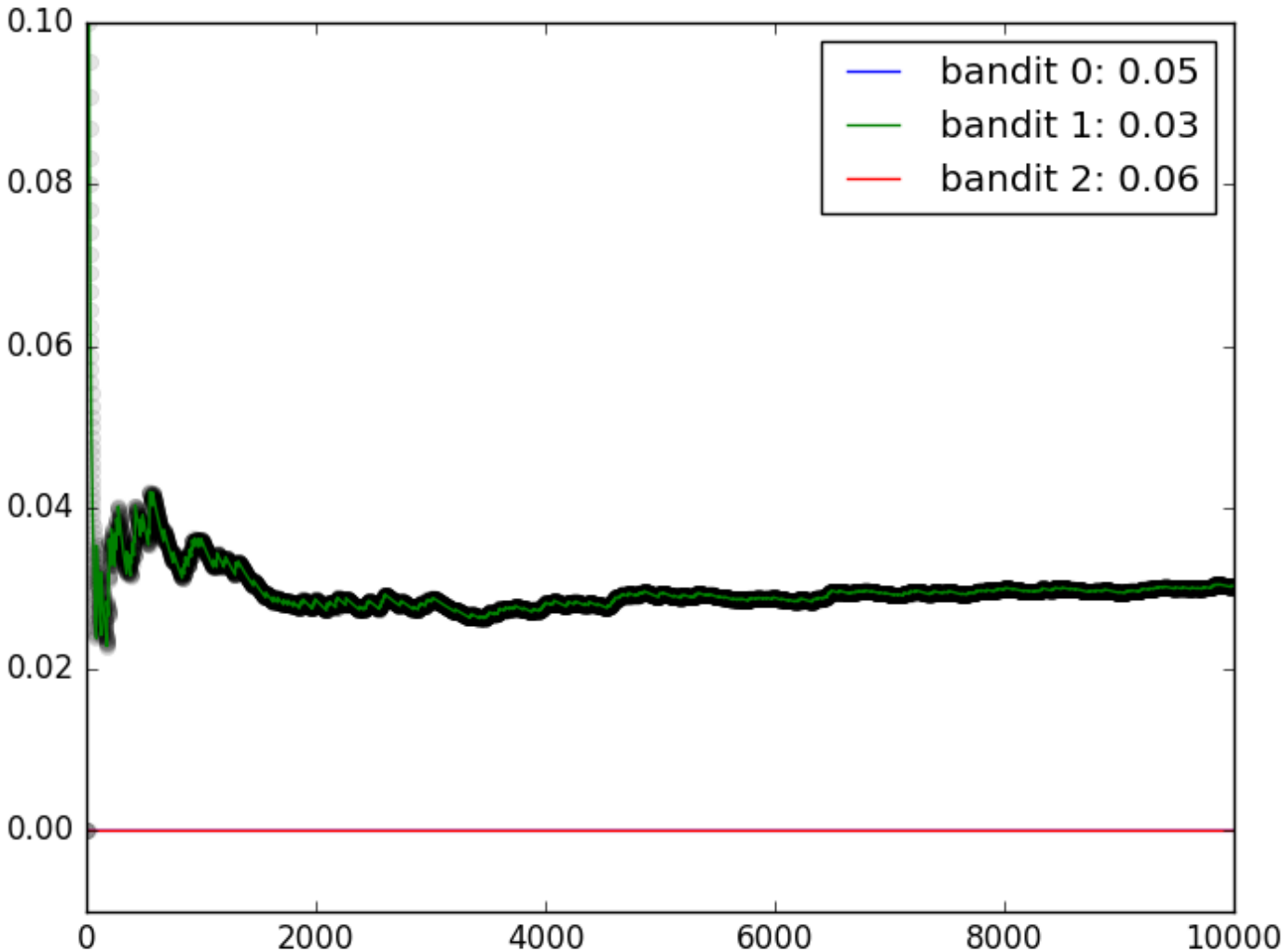
[when should you test?]



The Multi-Armed *max* Bandit

Max: *use the (currently) best performing option*

[perhaps after a burn-in?]

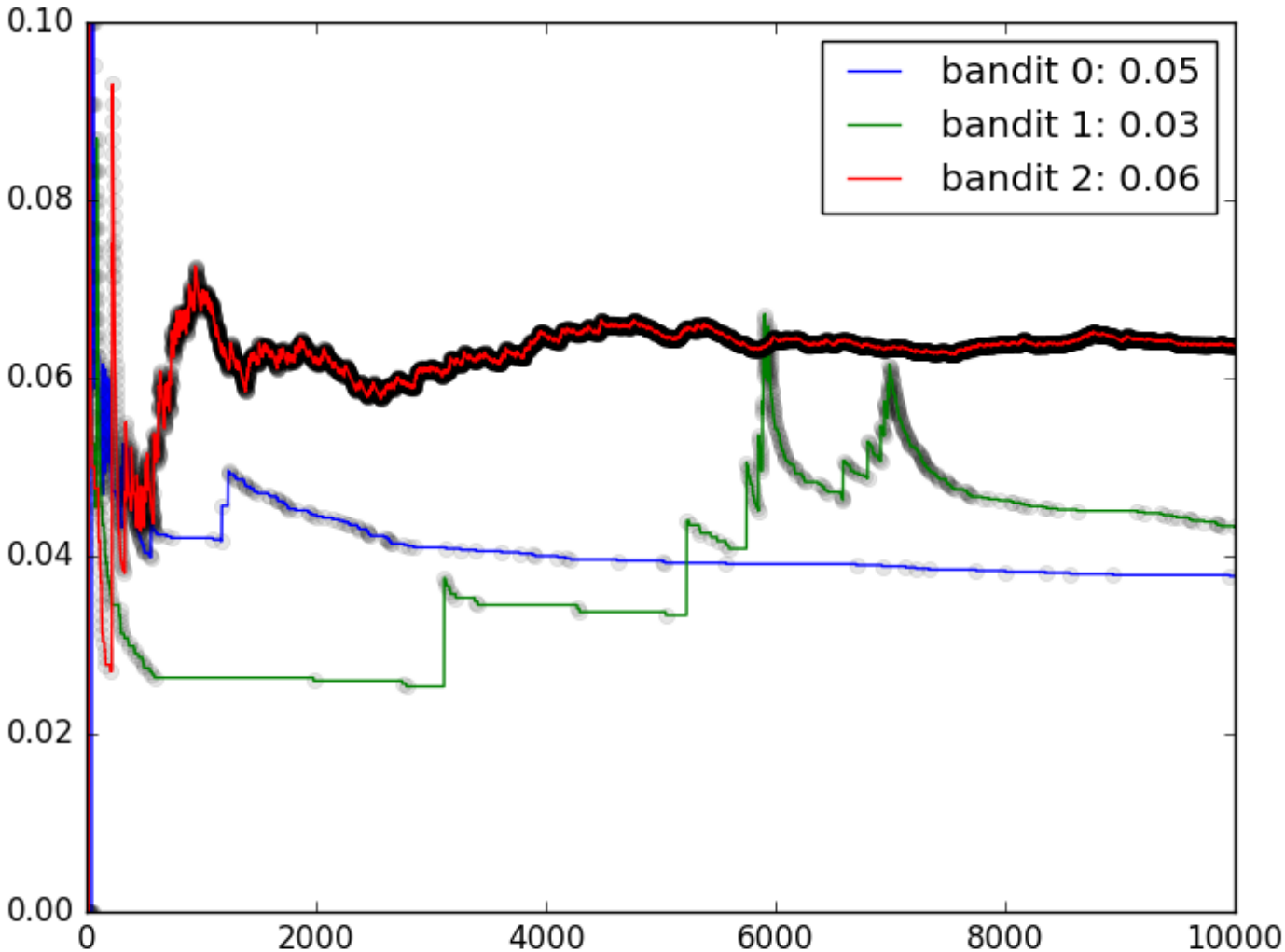


The Multi-Armed *softmax* Bandit

Softmax: *Select proportionally to the currently estimated success rates*

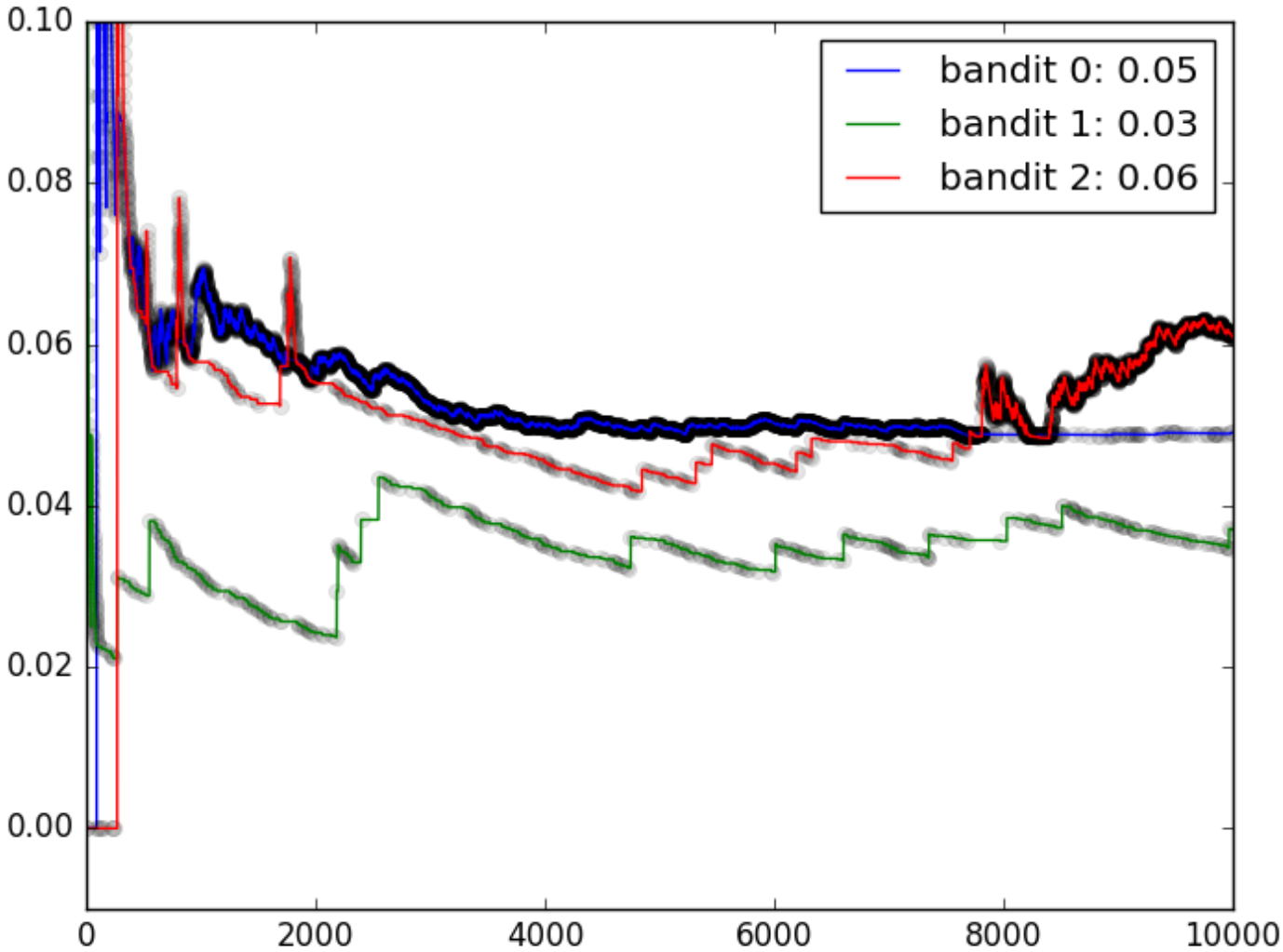
$$\Pr(\text{select } k) \propto \exp(\hat{p}_k/\tau), \text{ i.e.,}$$

$$\Pr(\text{select } k) = \frac{\exp(\hat{p}_k/\tau)}{\sum \exp(\hat{p}_k/\tau)}$$



The Multi-Armed ϵ-*greedy* Bandit

ϵ-greedy: *Select randomly with probability* ϵ
otherwise use the current best option

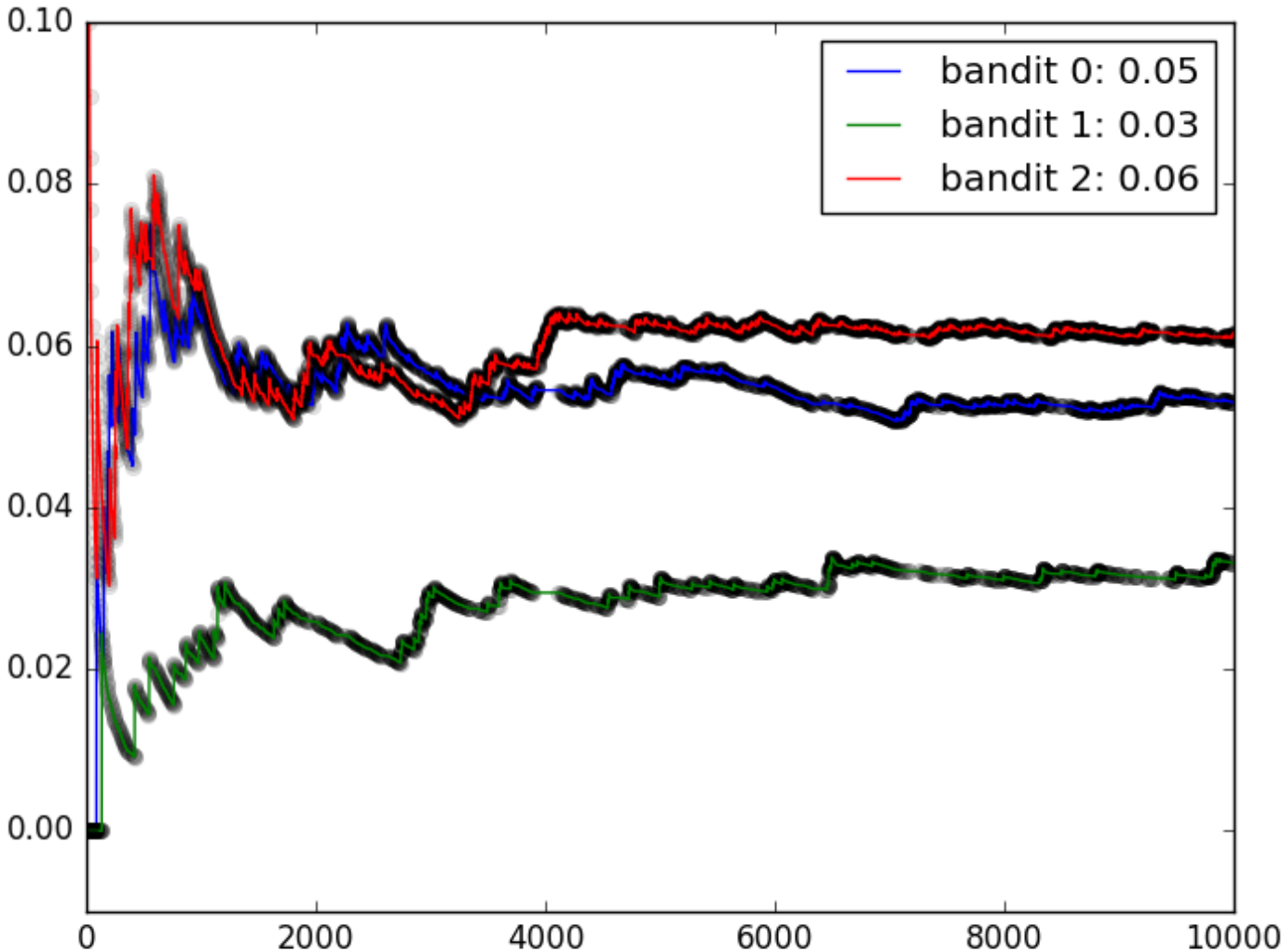


The Multi-Armed *UCB1* Bandit

UCB1: *Select the option which has the highest* **possible** *conversion potential*

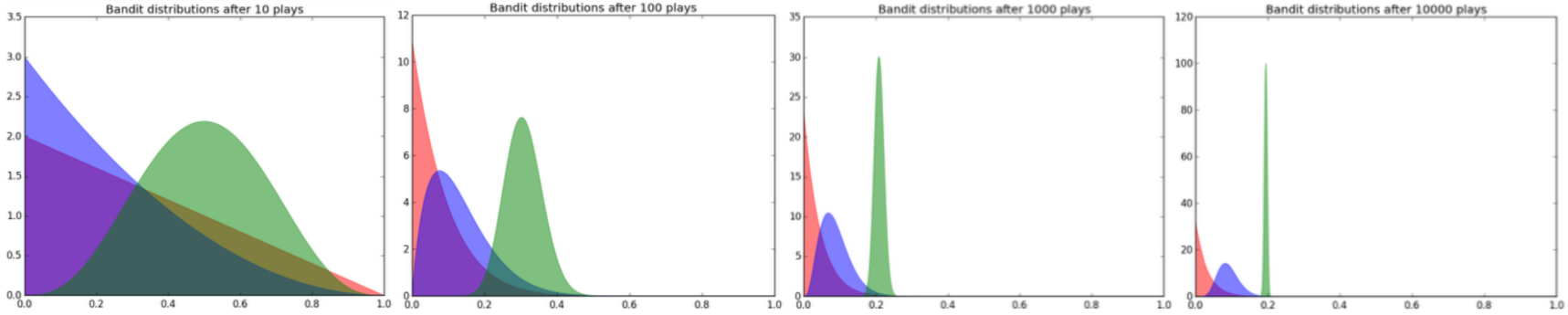
$$\max_k \hat{p}_k + \sqrt{\frac{2log\sum n_k}{n_k}}$$

with n_k trials for bandit k



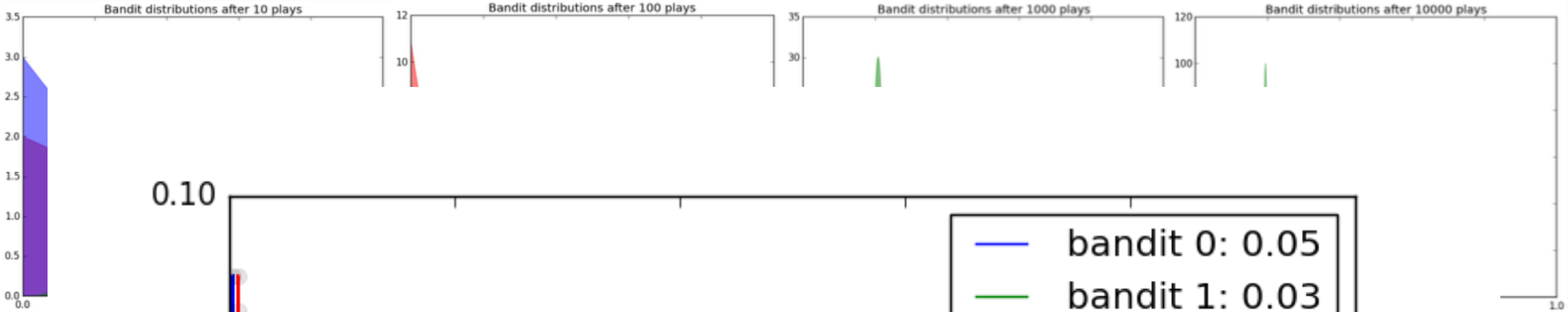
The Multi-Armed *Bayesian* Bandit

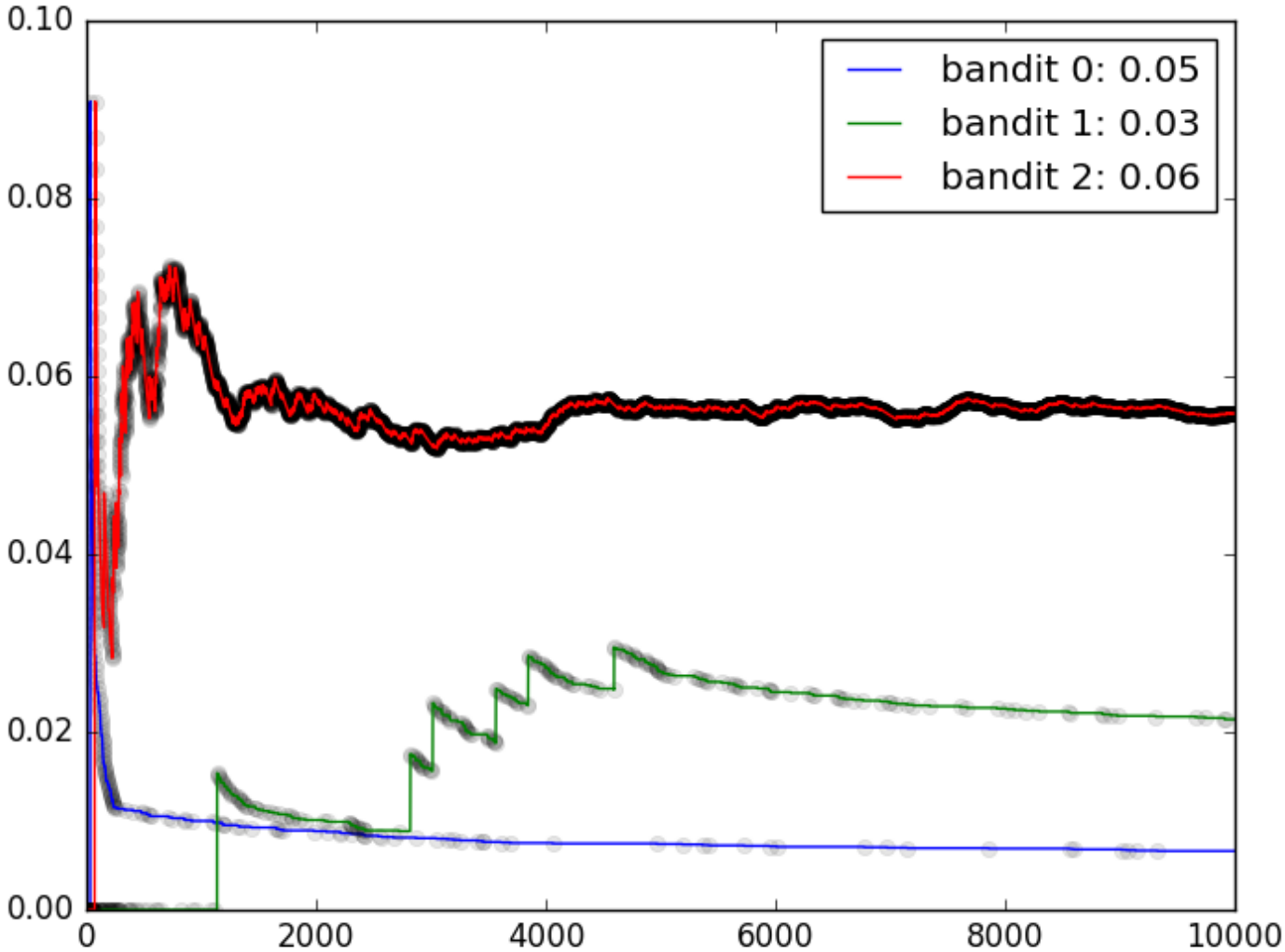
- Bayesian: *sample from the current posteriors* $\longrightarrow$ *then use the best*



The Multi-Armed *Bayesian* Bandit

- Bayesian: *sample from the current posteriors* $\longrightarrow$ *then use the best*





The Multi-Armed Bandit

Exploration & Exploitation