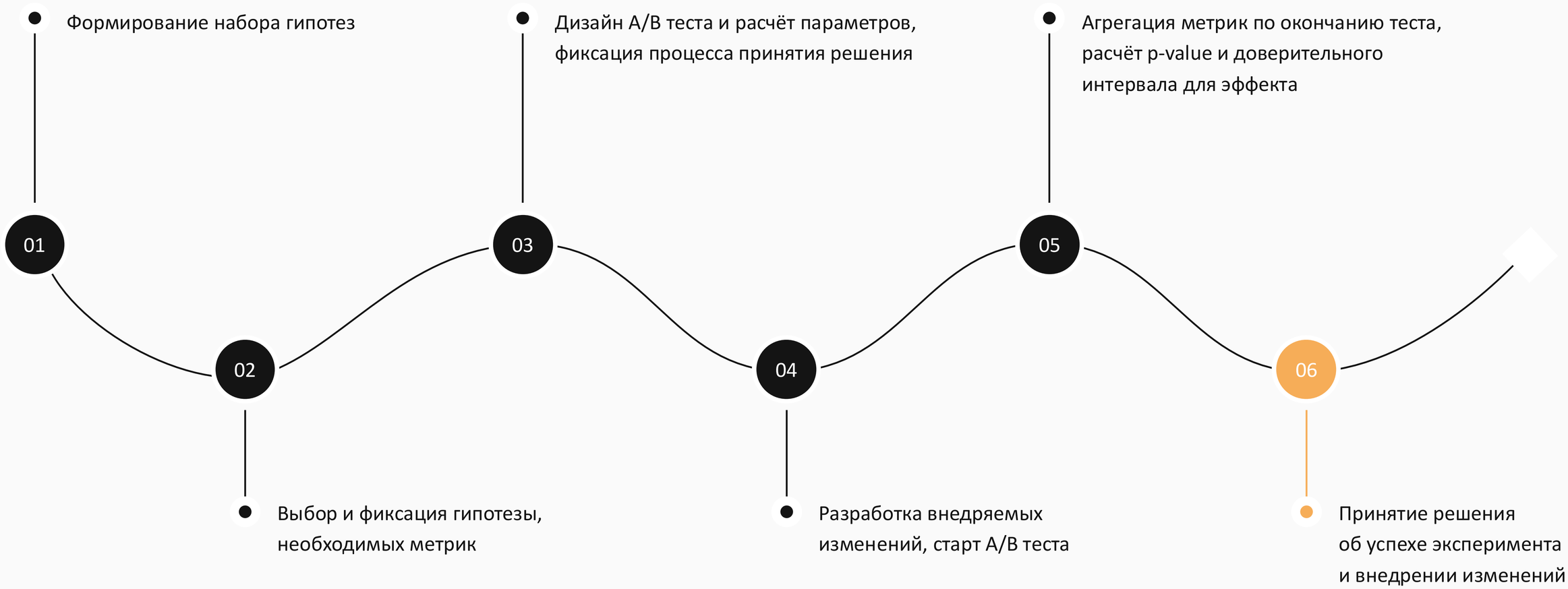
Лекция 4

Байесовский подход в A/B тестировании

План

- Схема классического частотного A/B теста
- Недостатки частотных подходов
- Введение в байесовское тестирование
- Сравнение байесовского и частотного подходов
- Интеграция байесовского подхода в процессы A/B
- Сравнительный анализ подходов на примере сценариев МТС Big Data

Схема классического частотного A/B теста



Два эксперимента

Эксперимент #1

p-value: **0.049**

effect: 2.4%

interval: **(0.05%, 5%)**

Вывод: тест удачный, внедряем изменения!

Эксперимент #2

p-value: **0.055**

effect: 4.3%

interval: **(0.05%, 5%)**

Вывод: значимых результатов нет, отказ от внедрения изменений.

Возникающие проблемы:

→ Сложное объяснение p-value, доверительного интервала и процесса принятия решения

→ P-value во втором случае отличается на доли, эффект в 2 раза выше, однако изменения не были внедрены

Недостатки частотных подходов

- Низкая интерпретируемость p-value и доверительного интервала.
- Тяжело объяснить результаты – нет вероятностей, рисков, прибылей и потерь
- Подход нацелен на контроль ошибки I рода.
- Механизмы проверки гипотез негибкие к цене ошибок и работают по заданному порогу
- Мониторинг и ранняя остановка завышают целевую ошибку I рода

Байесовский вывод

Возникающие проблемы:

→ Неизвестный эффект – случайная величина, хотим восстановить распределение.
В классическом подходе эффект – константа, которую пытаемся оценить точечно /
с помощью доверительного интервала.

→ На основе априорного знания о целевой метрике и поступающих в ходе эксперимента данных формируем апостериорное распределение.

Вероятность того, что событие B истинно, если событие A истинно

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

← Вероятность того, что событие A истинно

← Вероятность того, что событие B истинно

Вероятность того, что событие A истинно, если событие B истинно

Байесовское тестирование. Априорное распределение

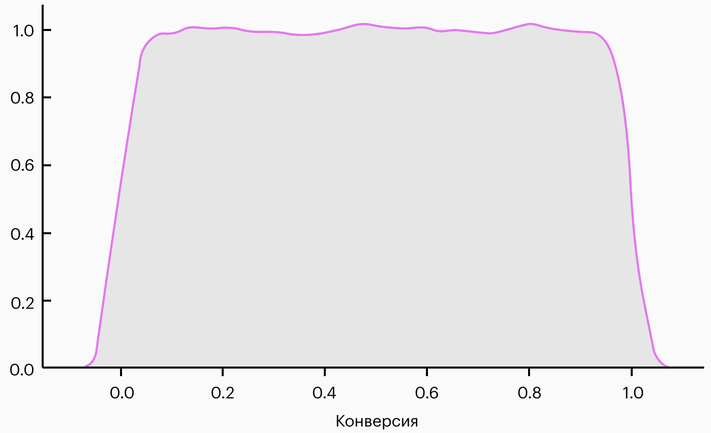
- Априорное распределение помогает описать наше незнание реальности.
- Хотим отбросить заведомо неверные значения.
- Нельзя подстраивать его под данные.

Априорное распределение выбираем, исходя из здравого смысла или на основе исторической информации.

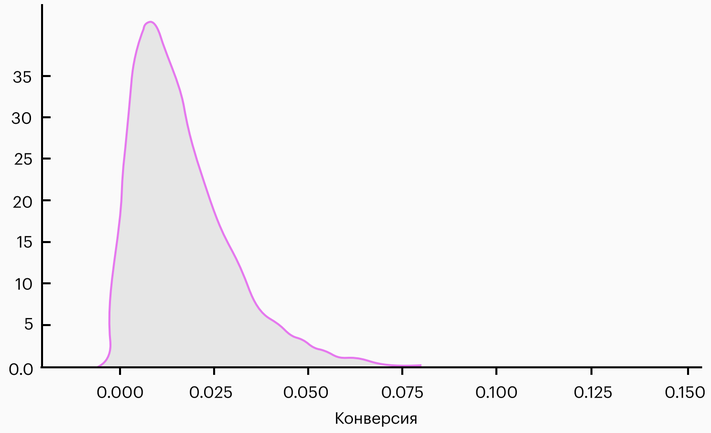
Пример. Конверсия может быть задана равномерным распределением от 0 до 1, если по ней нет никакой истории. Если же мы знаем, что конверсия колеблется в районе 2%,
то эту информацию можно использовать для уточнения априорного распределения, взяв бета-распределение
с параметрами 2 и 98.

Хорошее априорное знание может помочь ускорить тест!

Априорное распределение B (1, 1)



Априорное распределение B (2, 98)



Байесовское тестирование. Априорное распределение

Чаще всего используют два подхода к получению апостериорных распределений:

Аналитический (сопряженные распределения) Хотим отбросить заведомо неверные значения.

Численный (чаще всего методы Монте-Карло)

Аналитический подход хорошо работает для ограниченного числа случаев (см. знаменатель в формуле Байеса). Поэтому часто используются методы Монте-Карло.

Работая с конверсиями, мы можем обойтись аналитическими расчетами, но для порядка сравнить результаты аналитических расчетов с результатами работы MCMC (Markov Chain Monte Karlo).

$$P_a \sim Beta(\alpha, \beta)$$
$$X_a|P_a \sim Bin(n_a, p_a)$$
$$\Rightarrow P_a|X_a \sim Beta(\alpha + x_a, \beta + n_a - x_a)$$

$$D_1 = (P_b - P_a)|X$$
$$D_1 \sim N(E[P_b|X_b] - E[P_a|X_a], D[P_a|X_a] + D[P_b|X_b])$$
$$\alpha_A = \alpha + x_A$$
$$\beta_A = \beta + n_A - x_A$$
$$E[P_a|X_a] = \frac{\alpha_A}{\alpha_A + \beta_A}$$
$$D[P_a|X_a] = \frac{\alpha_A \beta_A}{(\alpha_A + \beta_A)^3}$$

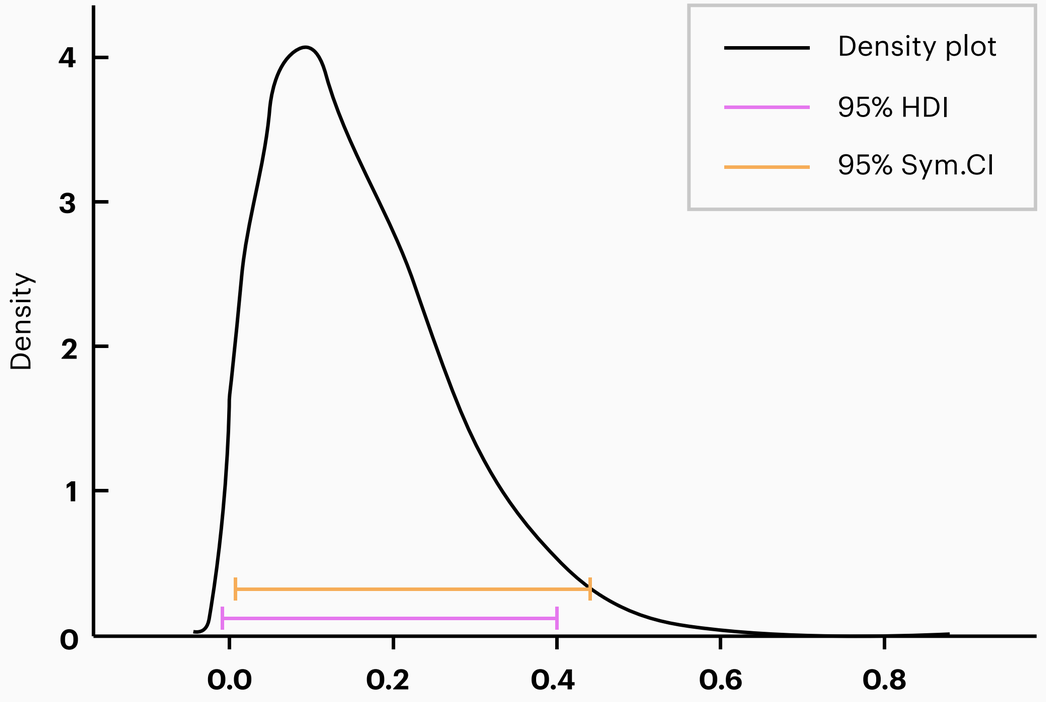
Пример аналитического вывода для конверсии на основе бета-биномиальной модели

Байесовское тестирование. Априорное распределение

В частотном подходе мы строили доверительные интервалы.
В байесовском подходе интервальные оценки делаются с помощью байесовских (bayesian или credible) интервалов

Доверительный интервал:
истинный параметр – константа
границы интервала – случайные величины
с вероятностью 0.95 границы покрывают истинное значение.

Байесовский интервал:
истинный параметр – случайная величина;
истиный параметр лежит с вероятностью 0.95 в интервале.
Самый короткий байесовский интервал - HPD [HDI] (highest probability density interval)



Байесовское тестирование. Зачем и как?

Используя байесовский подход для проверки гипотез мы хотим:

01 Улучшить интерпретируемость результатов

02 Ускорить A/B тесты

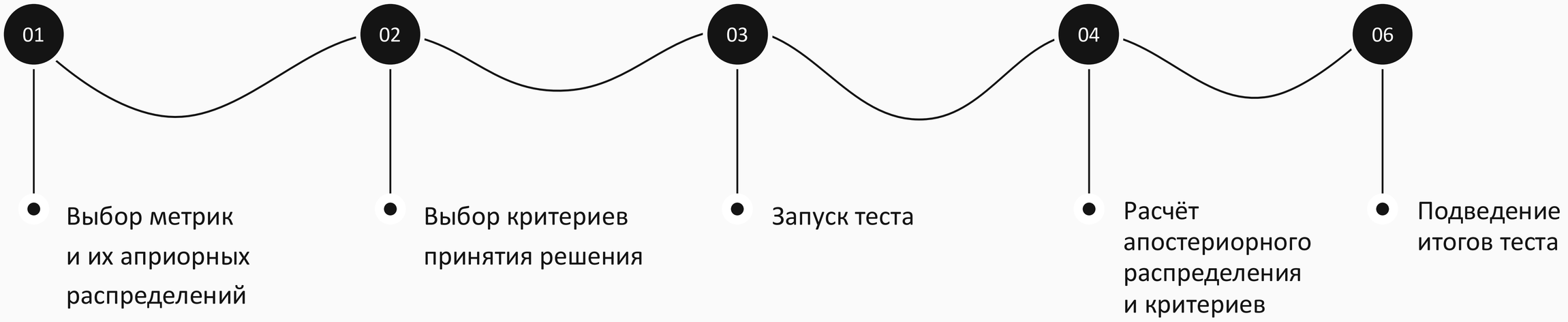
03 Оценить потери

Основные критерии принятия решения:

01 Вероятность, что группа B лучше группы A по сравниваемой метрике (пороговые значения)

02 Ожидаемые потери (можем дополнительно заложить дельту, учитывая расходы)

Байесовское тестирование. Жизненный цикл

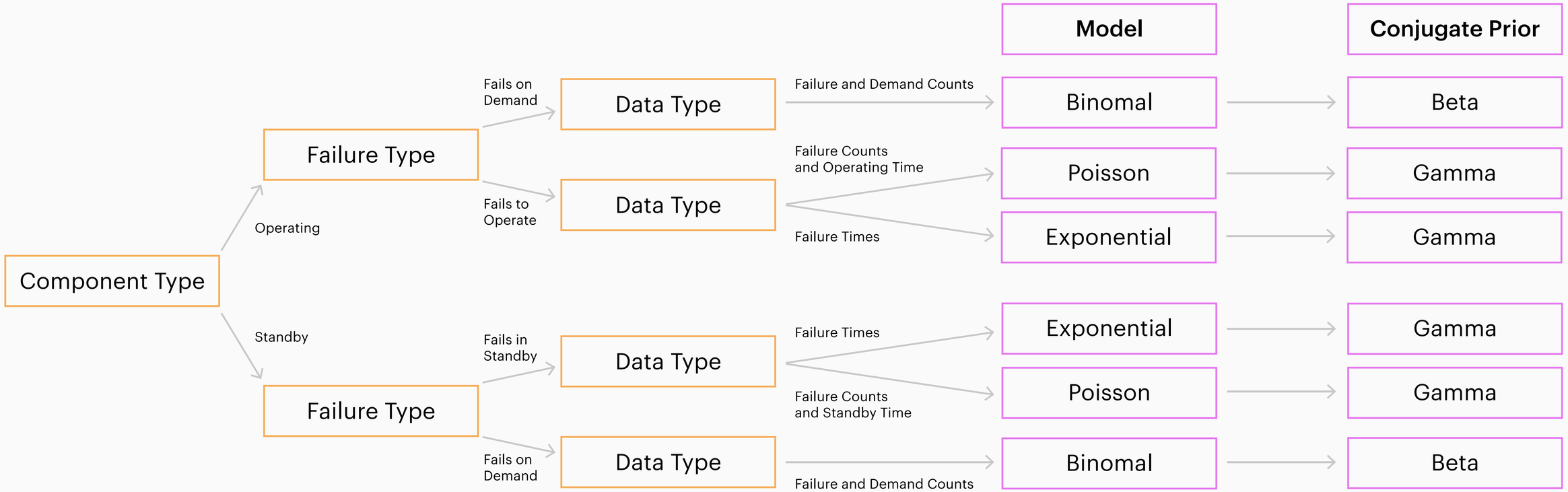


Байесовское тестирование. Аналитический метод

Аналитический метод заключается в выводе формул для апостериорной вероятности на основе теорем и предположений. Вывод красивых формул возможен, только если мы используем определенные удобные распределения.

Например, в случае конверсии это будет бета-распределение. Если априорные вероятности являются бета-распределениями, то и апостериорные вероятности будут являться бета-распределениями от измененных параметров.

Рекомендации по выбору сопряженных априорных распределений



Байесовское тестирование. Априорное распределение

Предположим, что мы измеряем бинарный исход (успех или неудача)

Пример. Была ли интеграция просмотра рекламы в конверсию

$x_A \sim \text{Binomial}(n_A, \theta_A)$, $x_B \sim \text{Binomial}(n_B, \theta_B)$

где x_A, x_B — число успешных событий

n_A, n_B — общее число наблюдений.

θ_A, θ_B — истинные вероятности успеха

Для параметров вероятности θ_A и θ_B выбираем априорное распределение

$\theta_A \sim \text{Beta}(\alpha_A, \beta_A)$

$\theta_B \sim \text{Beta}(\alpha_B, \beta_B)$

Как аналитически определить параметры α и β априорного Бета-распределения на основе предварительной информации или предположений?

Вариант 1

Если у нас есть предварительная информация об ожидаемой вероятности успеха μ

$E[\theta] = \alpha / \alpha + \beta = \mu$

степени уверенности s в этой оценке

$s = \alpha + \beta$

Тогда

$\alpha = \mu s, \beta = (1 - \mu)s$

Вариант 2

Если мы хотим использовать неинформативное априорное распределение, которое минимально влияет на апостериорный вывод:

$\theta \sim Beta(1, 1)$

Здесь $\alpha = 1$ и $\beta = 1$, что соответствует равномерному распределению на интервале [0, 1].

Вариант 3

Если у нас есть исторические данные, в которых было k_o успехов из n_o наблюдений, мы можем задать:

$\alpha = k_o + \alpha_o$

$\beta = n_o - k_o + \beta_o$

где α_o и β_o — гиперпараметры, отражающие нашу уверенность в исторических данных. Если мы полностью доверяем историческим данным, можем взять $\alpha_o = 0$ и $\beta_o = 0$.

Частотный подход

$$\hat{p}_A = x_A/n_A, \hat{p}_B = x_B/n_B$$

$$SE = \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$

$$(\hat{p}_B - \hat{p}_A) \pm z_{\alpha/2} \times SE$$

Байесовский подход

$$\theta_A \mid D \sim \text{Beta}\,(\alpha_A + x_A, \beta_A + n_A - x_A)$$

$$\theta_B \mid D \sim \text{Beta}\,(\alpha_B + x_B, \beta_B + n_B - x_B)$$

$$\Delta = \theta_B - \theta_A$$

Credible interval

$$P\,(\Delta_{left} \leq \Delta \leq \Delta_{right} \mid D) = 1 - \alpha$$

Highest Posterior Density Interval

$$HPD(1-\alpha) = \Big(\Delta \mid p\,(\Delta \mid D) \geq k\Big)$$

$$\int_{\Delta : p(\Delta \mid D) \geq k} p(\Delta \mid D) d\Delta = 1 - \alpha$$

Байесовское тестирование. Ожидаемые потери

Рассмотрим ожидаемые потери на примере конверсий в двух группах.

[Ожидаемые потери в конверсии с одного юзера при выборе одной из групп]

$$group = \text{A или B}$$

$$\mathrm{E}\,[\mathrm{L}]\,(group) = \int_0^1 \int_0^1 \mathrm{L}\,(\lambda_A, \lambda_B, group)\,\mathrm{P}\,(\lambda_A, \lambda_B)\,d\,\lambda_B \lambda_A$$

01 Конверсии в группах A/B: λ_A, λ_B

02 Совместная апостериорная плотность: $\mathrm{P}\,(\lambda_A, \lambda_B)$

03 Функция потерь. Она равна 0, если конверсия в группе выше, чем в другой, либо показывает разницу в конверсиях, если конверсия ниже:

$$\mathrm{L}\,(\lambda_A, \lambda_B, \mathrm{A}) = max\,(\lambda_B - \lambda_A, 0)$$

$$\mathrm{L}\,(\lambda_A, \lambda_B, \mathrm{A}) = max\,(\lambda_A - \lambda_B, 0)$$

Байесовское тестирование. Аналитический метод

Преимущества

- Явные формулы для апостериорного распределения, что упрощает вычисления и интерпретацию результатов.
- Аналитический метод предоставляет более прямой и понятный путь для апостериорной оценки параметров.

Недостатки

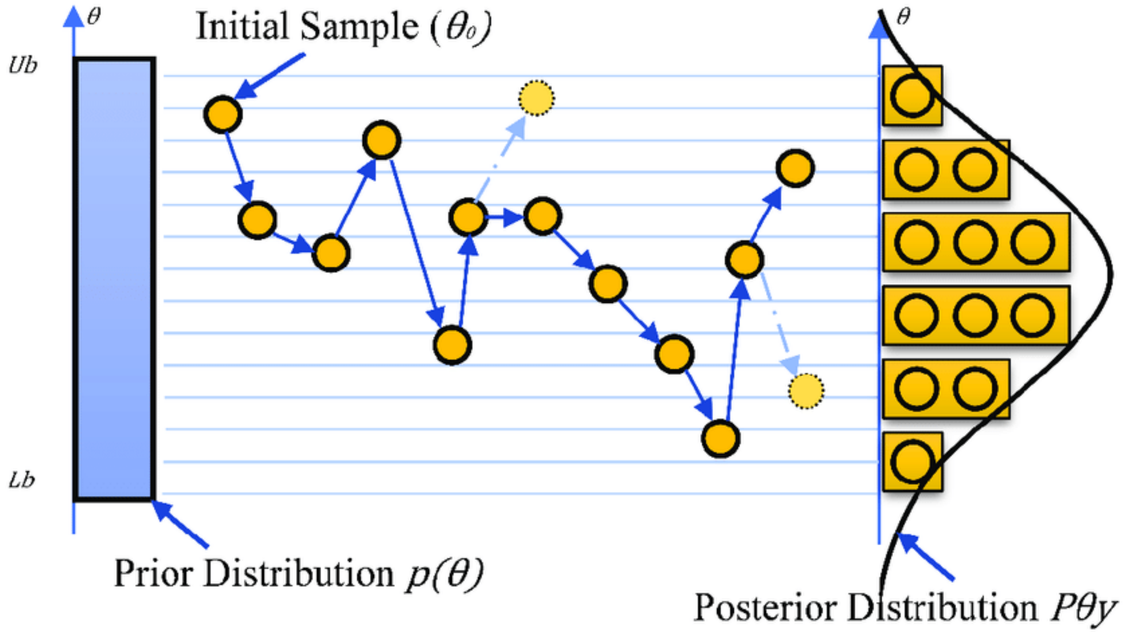
- Априорное распределение должно быть сопряжено с апостериорным. Это ограничивает выбор распределений и может быть недостаточно гибким для некоторых моделей или сложных сценариев.
- Предварительный выбор и спецификация априорного распределения. Ошибка в выборе априорного распределения может привести к неправильным выводам, поэтому требуется определенный уровень экспертизы и знаний.

Байесовское тестирование. Численный метод

Численный метод — более гибкий подход, чем аналитический. Он заключается в итеративном формировании апостериорного распределения.

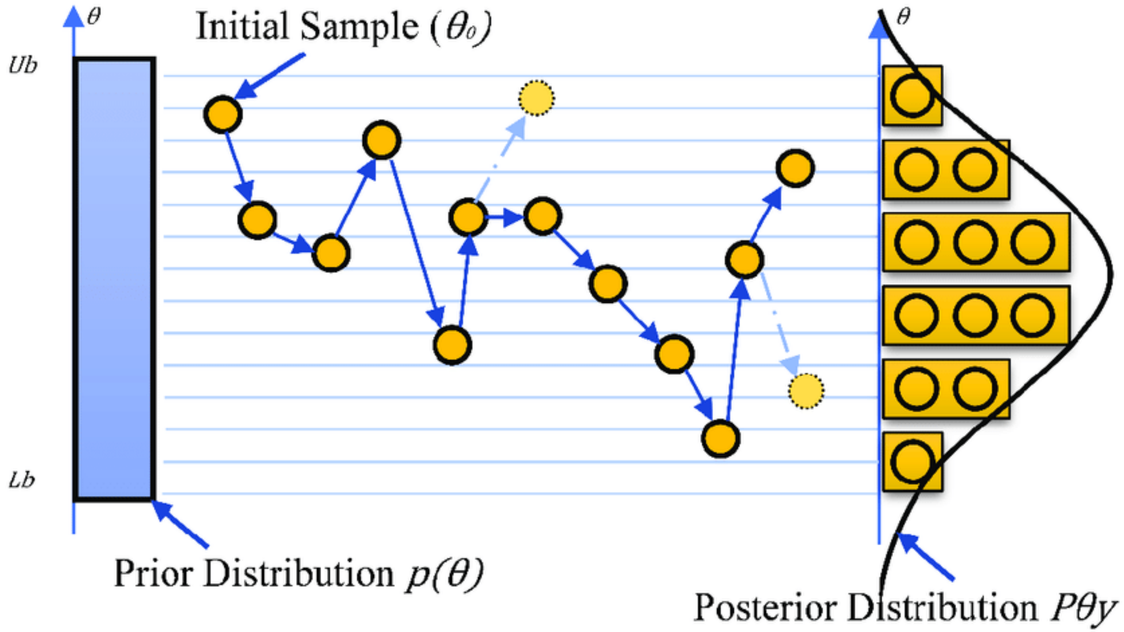
Методы Монте-Карло для марковских цепей (MCMC) — это методы для получения выборок
из вероятностных распределений, которые известны только до некоторой нормировочной константы.

Марковская цепь — это случайная последовательность состояний, где вероятность выбора определенного состояния зависит только
от текущего состояния цепи.



Байесовское тестирование. Численный метод

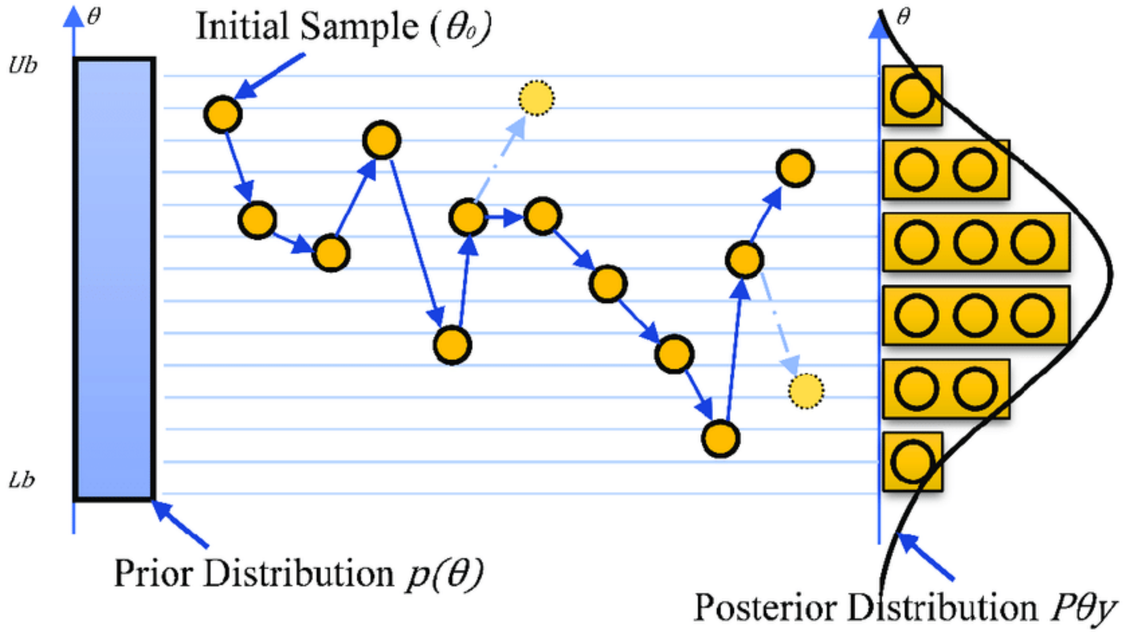
Алгоритм MCMC создает и имитирует марковскую цепь. Изначально имеется
априорное распределение и накопленные
данные эксперимента. Алгоритм выбирает случайное
значение из исходной выборки
и по определенному правилу принимает
или отвергает это значение в зависимости
от его соответствия данным и априорному
распределению.



Этот процесс повторяется много раз. Первые несколько тысяч шагов алгоритм «прогревается», после чего обычно сходится к апостериорному распределению. На выходе получается большая выборка из распределения.

Байесовское тестирование. Численный метод

Алгоритм MCMC создает и имитирует марковскую цепь. Изначально имеется
априорное распределение и накопленные
данные эксперимента.
Алгоритм выбирает случайное значение из исходной выборки и по определенному правилу принимает или отвергает это значение в зависимости
от его соответствия данным и априорному распределению.



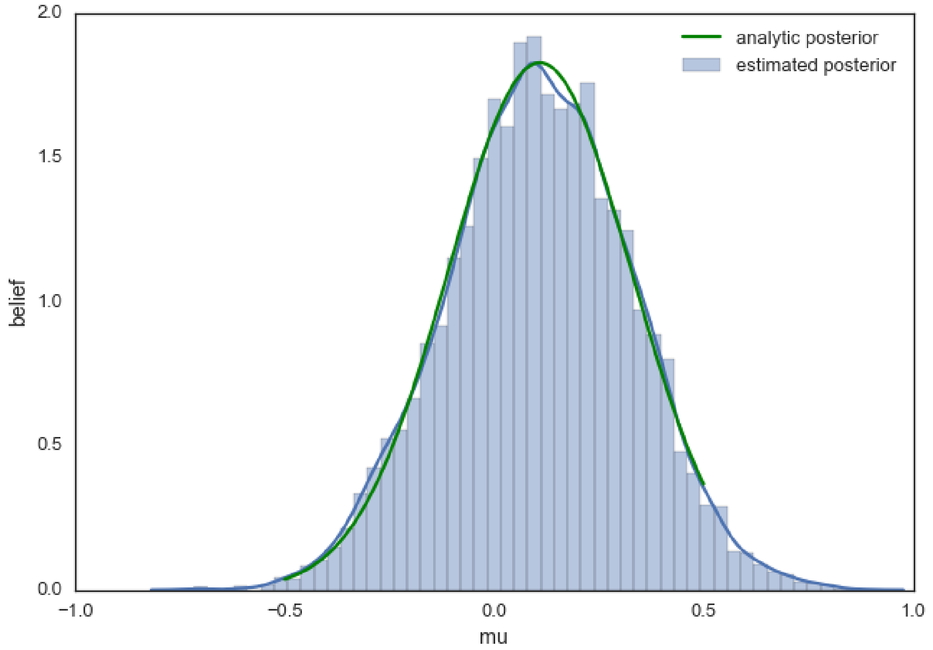
Этот процесс повторяется много раз. Первые несколько тысяч шагов алгоритм «прогревается», после чего обычно сходится к апостериорному распределению. На выходе получается большая выборка из распределения.

Общая идея MCMC

Получим выборки из сложного вероятностного распределения p(θ) , где θ - вектор параметров

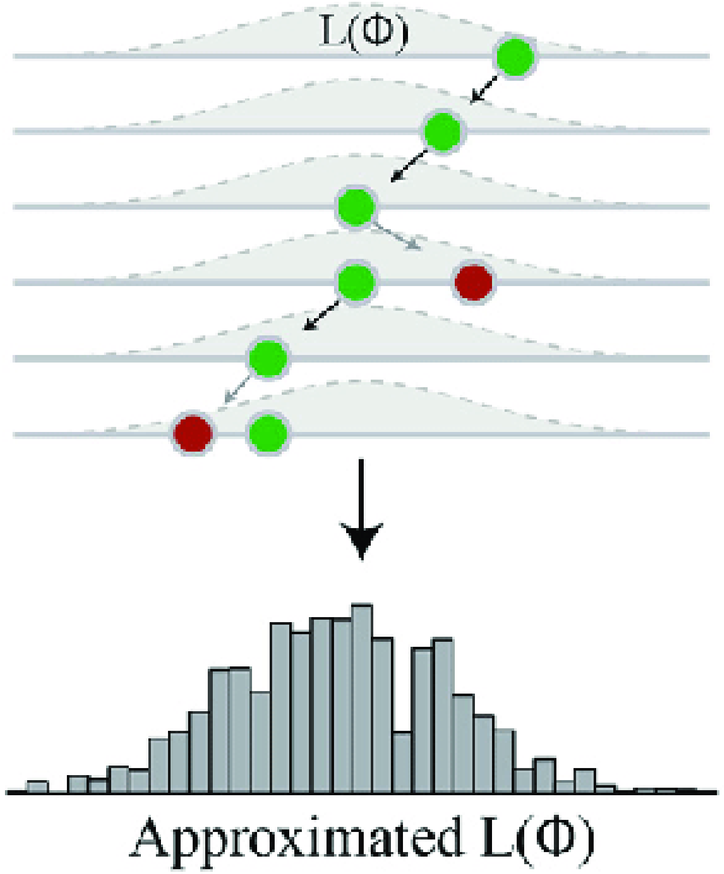
Построим цепь Маркова, где состояние стремится к целевому распределению p(θ)

После достаточного количества шагов Burn-in выборки из цепи можно считать приближенными выборками из p(θ)



MCMC. Алгоритм Метрополиса-Гастингса

1. Выбираем начальное значение параметра $\theta^{(0)}$
2. Для t из [1, T]
   - Proposal step: Сгенерируем θ^* из предложенного распределения $q(\theta^* \mid \theta^{(t-1)})$.
   - Вычислением вероятность принятия θ^*
     Для этого считаем отношение вероятностей:
     $\alpha = \min((1, p(\theta^*)\, q(\theta^{(t-1)} \mid \theta^*) / p(\theta^{(t-1)})\, q(\theta^* \mid \theta^{(t-1)})))$
   - Принять/отклонить θ^*
     Сгенерируйте $u \sim \text{Uniform}(0, 1)$.
     Если $u < \alpha$, то $\theta^{(t)} = \theta^*$ (принимаем)
     Иначе $\theta^{(t)} = \theta^{(t-1)}$ (отклоняем)

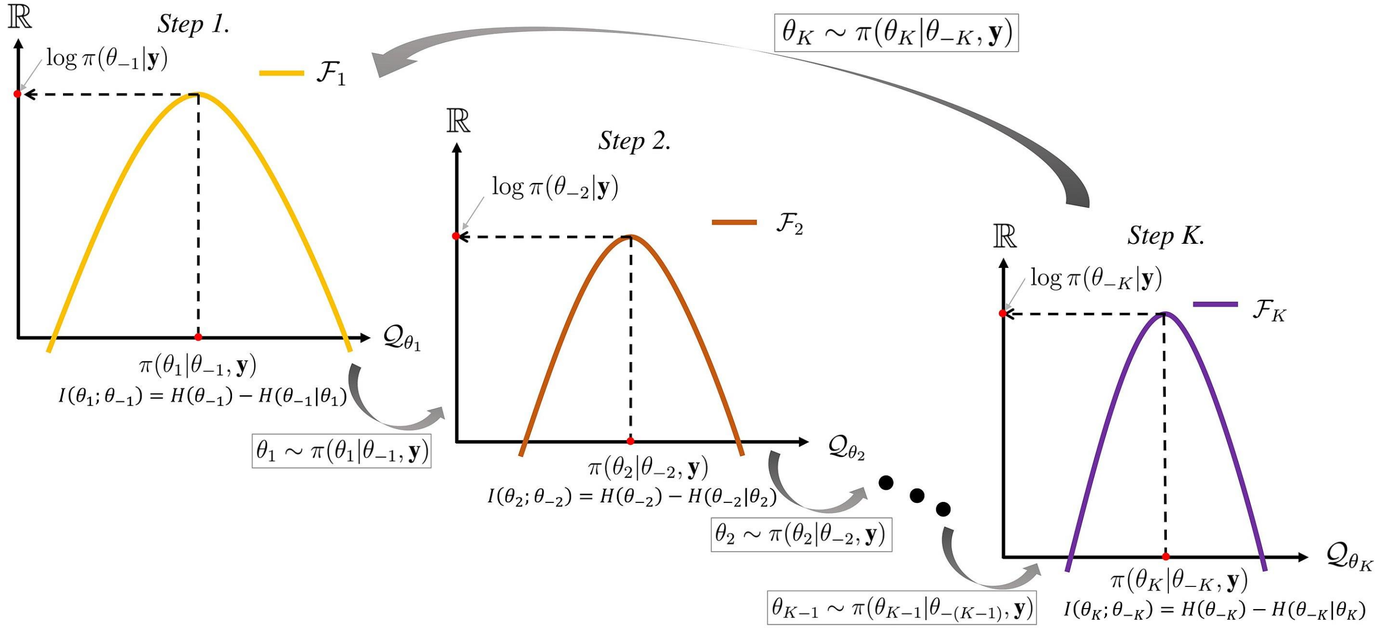


1) Draw new parameter Φ′ close to the old Φ
2) Calculate L(Φ′)
3) Jump proportional to L(Φ′)/L(Φ)

MCMC. Gibbs Sampling

Частный случай Метрополиса-Гастингса, где выборки берутся из условных распределений.

1. Задаем начальные значения для всех параметров $\theta = (\theta_1, \theta_2, ..., \theta)$
2. Для t из [1, T]
   - Для каждого параметра θ_i сгенерируем $\theta_i^{(t)}$ из условного распределения
     $p(\theta_i \mid \theta_{-i}^{(t-1)}, data)$, где θ_{-i} — все параметры, кроме θ_i



MCMC. Hamiltonian Monte Carlo, HMC

1. Инициализируем начальное значение вектора параметров θ_0 (произвольное значение/априорное)
2. Введение переменных импульса:

   Для каждого параметра θ_i вводится соответствующая переменная импульса p_i:

   $p \sim \mathcal{N}(0, M)$, где M — метрика, равная единичной матрице
3. Определим гамильтониан H(θ, p):

   H(θ, p) = U(θ) + K(p), где

   U(θ)- потенциальная энергия, U(θ) = -ln p(θ)

   K(p) - Кинетическая энергия, K(p) = $1\ /2\ \mathrm{p}^{\mathrm{T}}\mathrm{M}^{-1}\ \mathrm{p}\ ]$
4. Симулируем гамильтонову динамику с помощью интегрирования численных уравнений

Leapfrog integration с параметрами ε и L

- Шаг интегрирования ε - малая положительная величина
- Количество шагов L - как далеко перемещается система по траектории.

По итогу получаем новые значения (θ*, p*).

5. Вычислим изменения гамильтониана. Найдем разницу в полной энергии между начальным и предложенным состояниями:

$\Delta H = H(\theta^*, p^*) - H(\theta_t, p_t)$

MCMC. Hamiltonian Monte Carlo, HMC

6. Принимаем/отклоняем новое состояние

Принимаем новое состояние с вероятностью:

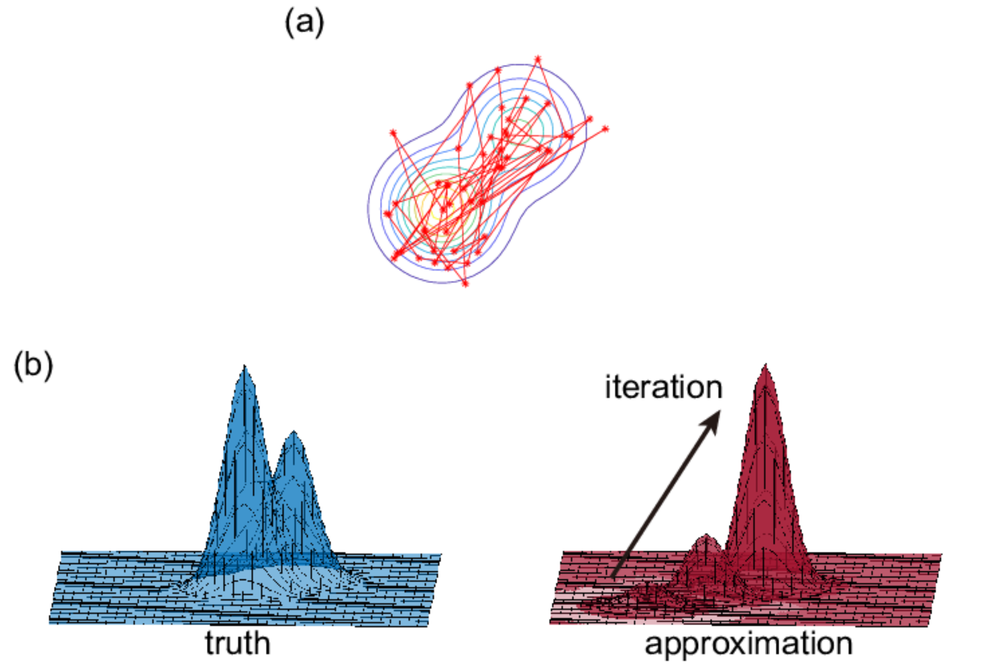
$\alpha = \min((1, e^{(-\Delta H)}))$

Генерируем $u \sim \text{Uniform}(0, 1)$.

Если $u < \alpha$, принимаем $\theta\{t+1\} = \theta^*$.

Иначе, отклоняем $\theta\{t+1\} = \theta t$.

Вернемся к шагу 2 и повторяйте процесс необходимое количество раз или до достижения сходимости.



7. Burn-in и Thinning

Разогрев цепи: отбросим первые N_burnin выборок, чтобы устранить влияние начальных условий.

Тонкая выборка: выберем каждый k-й элемент цепи, чтобы уменьшить автокорреляцию.

MCMC. Hamiltonian Monte Carlo, HMC

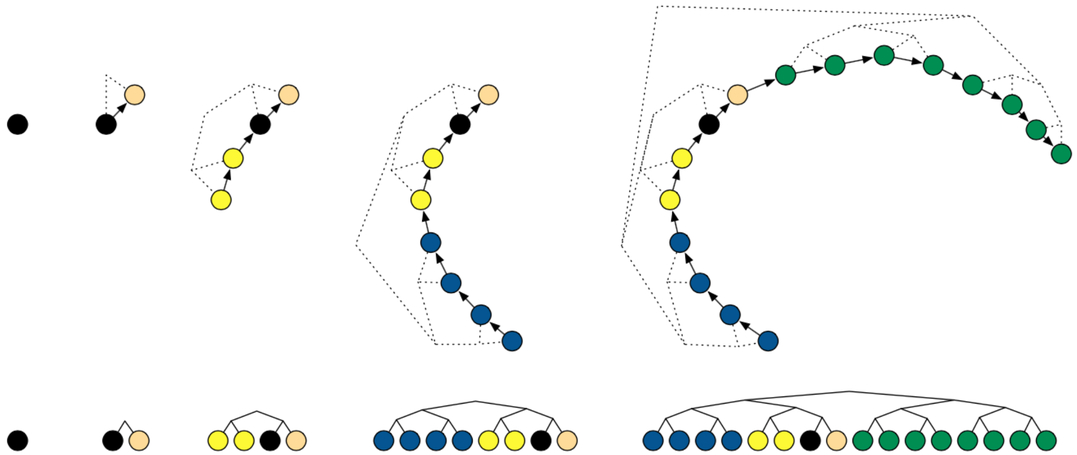
Выбор ε:

- Слишком большой шаг - неточность
- Слишком маленький шаг - увеличение вычислительных затрат

Выбор L:

- Слишком большое – вернемся в исходную точку
- Слишком маленькое – увеличение вычислительных затрат

Автоматическая настройка параметров - "No-U-Turn Sampler" (NUTS), который автоматически определяет оптимальные значения ε и L



MCMC. Сравнение методов

	Алгоритм Метрополиса-Гастингса	**Gibbs Sampling**	**Hamiltonian Monte Carlo**
Преимущества	Применяем к любому распределению Простота применения Не требует знания условных распределений	Не требует выбора распределения Эффективность при наличии сопряженных априорных распределений Просто реализовать при известных условных распределениях	Эффективность в высоких размерностях Низкая автокорреляция между выборками Генерирует более независимые выборки
Недостатки	Зависимость от выбранного распределения Последовательные выборки могут быть зависимы Выбор подходящего распределения и его параметров может быть сложным Эффективность снижается с увеличением числа параметров.	Неприменим, если условные распределения сложно получить Медленная сходимость Выборки могут быть зависимы, поэтому нужно делать большое число итераций	Необходимо вычисления градиентов и более сложной настройки Выбор параметров Высокая вычислительная сложность
Применимость	Модели со сложными апостериорными распределениями Требует выбора распределения и мониторинга сходимости Может быть вычислительно затратным при большом объеме данных	Эффективен для моделей с сопряженными распределениями (например, бета-биномиальное) Малоприменим для моделей со сложными или неявными условными распределениями.	Подходит для сложных моделей и больших наборов данных Реализация упрощается с помощью современных библиотек (например, Stan, PyMC3)

Компания проводит A/B-тест для сравнения эффективности двух разных интерфейсов мобильного приложения (варианты A и B) с точки зрения удержания пользователей.

Данные
Для каждого пользователя измеряется время (в днях) до того момента, когда он перестает использовать приложение. Некоторые пользователи все еще активны на момент окончания эксперимента, поэтому их время до оттока является цензурированным.

Модель
Survival analysis для моделирования времени до оттока.

Функция риска
Предполагается, что время до оттока следует произвольному сложному распределению без аналитического выражения функции плотности

Возможна индивидуальная гетерогенность между пользователями, моделируемая с помощью случайных эффектов или смешанных моделей.

Цель - Оценить апостериорные распределения параметров модели для каждого варианта (A и B) и определить, какой интерфейс лучше с точки зрения удержания пользователей.

Алгоритм Метрополиса-Гастингса	**Gibbs Sampling**	**Hamiltonian Monte Carlo**
+	неприменима из-за отсутствия доступных условных распределений	неприменима из-за невозможности вычисления необходимых градиентов

Компания хочет сравнить эффективность двух вариантов рекламного баннера на своем веб-сайте (варианты A и B) с точки зрения кликабельности (Click-Through Rate, CTR).

Данные
Для каждого варианта измеряется количество раз, когда баннер был показан пользователям.
Для каждого варианта измеряется количество кликов по баннеру.

Биномиальная модель
Полагаем, что количество кликов по баннеру для каждого варианта следует биномиальному распределению.

Используются бета-распределения в качестве априорных для параметров CTR каждого варианта.

Цель - Оценить апостериорные распределения параметров CTR для каждого варианта (A и B) и определить, какой баннер лучше

Алгоритм Метрополиса-Гастингса	**Gibbs Sampling**	**Hamiltonian Monte Carlo**
требует настройки и приводит к дополнительным вычислительным затратам без необходимости.	+	не приносит преимуществ в простых моделях с известными апостериорными распределениями и требует вычисления градиентов, что излишне усложняет задачу.

Компания хочет оценить влияние двух вариантов дизайна веб-страницы (A и B) на время, проведенное пользователем на странице (Time on Page). Время на странице считается важным показателем вовлеченности пользователя.

Данные
Для каждого варианта дизайна собирается информация о времени, которое пользователь провел на странице до перехода или закрытия.

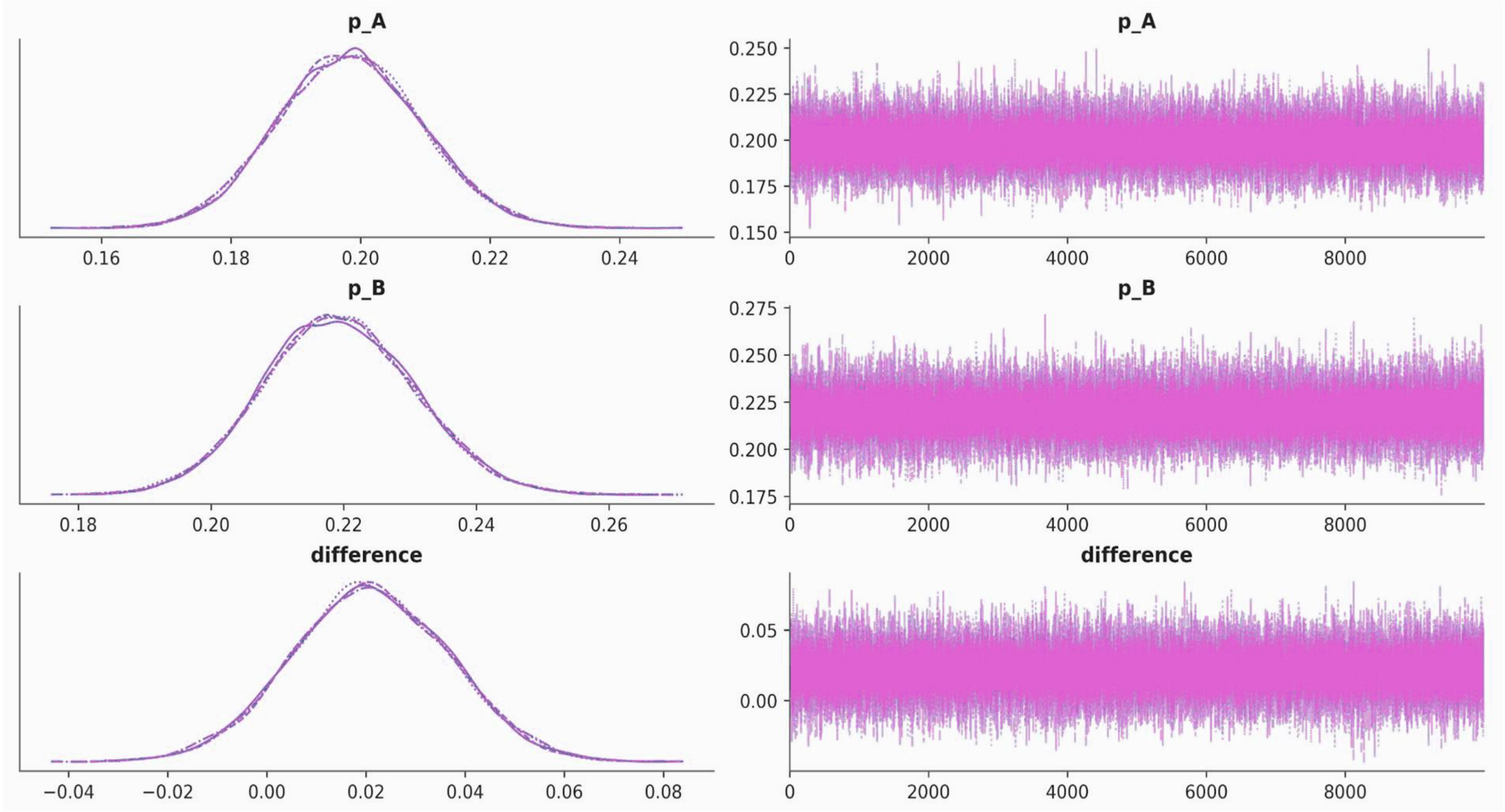
Предполагается, что время на странице для каждого варианта следует логнормальному распределению

Учёт возможной гетерогенности среди пользователей и потенциальной корреляции между параметрами модели.

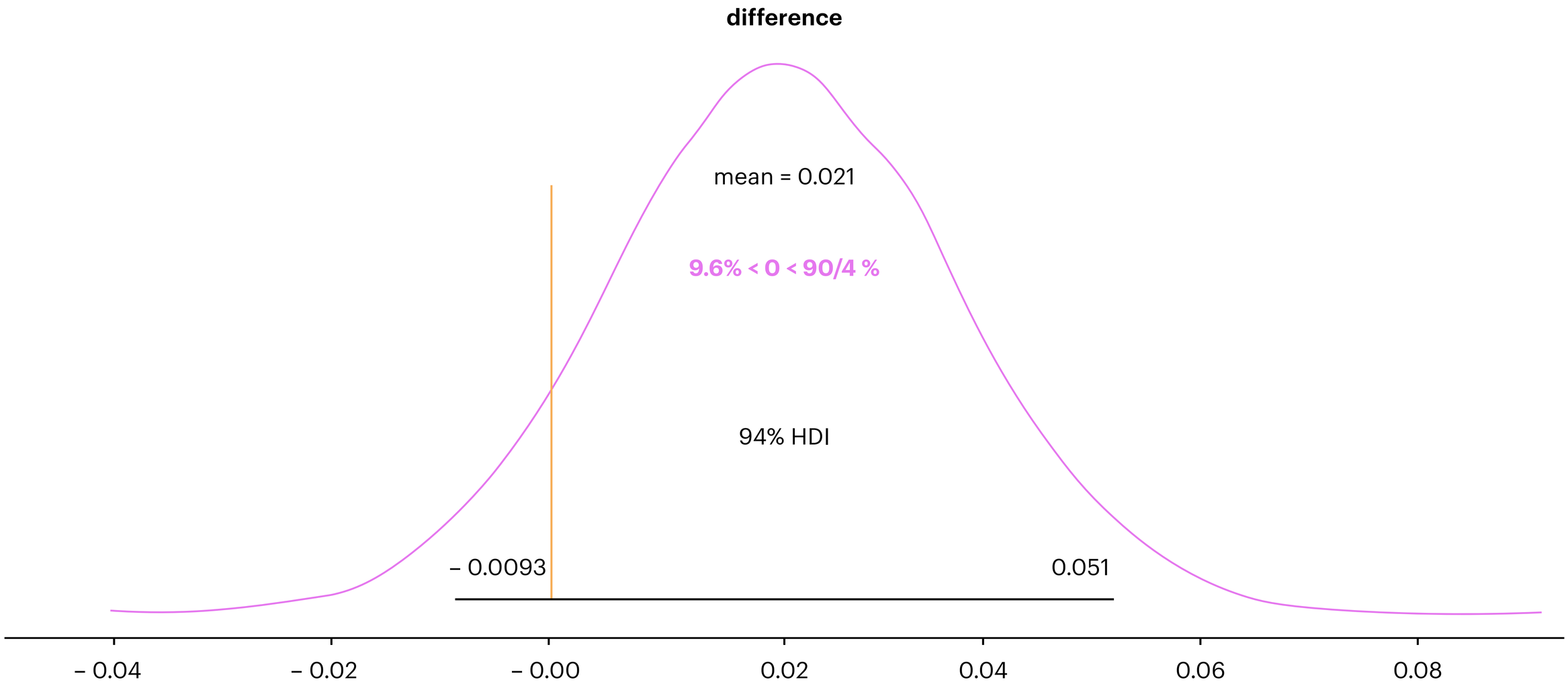
Цель: оценить апостериорные распределения параметров для вариантов A и B и определить, какой дизайн способствует большему времени на странице.

Алгоритм Метрополиса-Гастингса	**Gibbs Sampling**	**Hamiltonian Monte Carlo**
менее эффективен из-за низкой скорости сходимости и сложности настройки в высокоразмерных пространствах с коррелированными параметрами	неприменим из-за отсутствия простых условных распределений и невозможности прямой выборки	неприменимо из-за невозможности вычисления необходимых градиентов и требований дифференцируемости апостериорной плотности

Байесовское тестирование. Результаты симуляций MCMC



Байесовское тестирование. Результаты симуляций MCMC



Байесовское тестирование. Численный метод

Преимущества

- Более гибкое моделирование апостериорного распределения. Не требуем сопряженности распределений. Это дает возможность использовать более сложные модели и априорные распределения.
- Можем проводить симуляции на основе большого числа сгенерированных значений, что может дать более точные и надежные оценки параметров.
- Возможность оценки неопределенности, позволяя получить интервальные оценки.

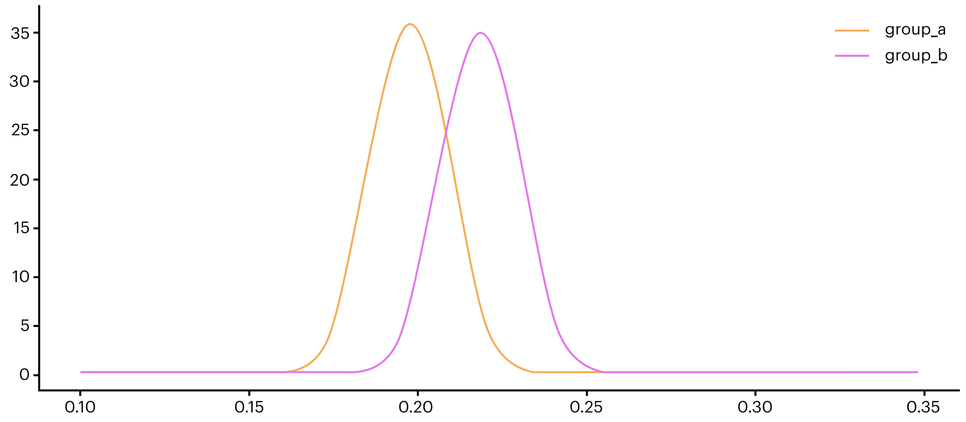
Недостатки

- MCMC могут быть вычислительно сложными и требовать значительных вычислительных ресурсов и времени.
- Для применения численных методов необходимо настроить алгоритмы и выбрать подходящие параметры симуляции. Неправильная настройка может привести к ненадежным оценкам или неэффективным вычислениям.

Байесовское тестирование. Сравнительный анализ

Аналитический метод

Плотность бета-распределения для групп A/B



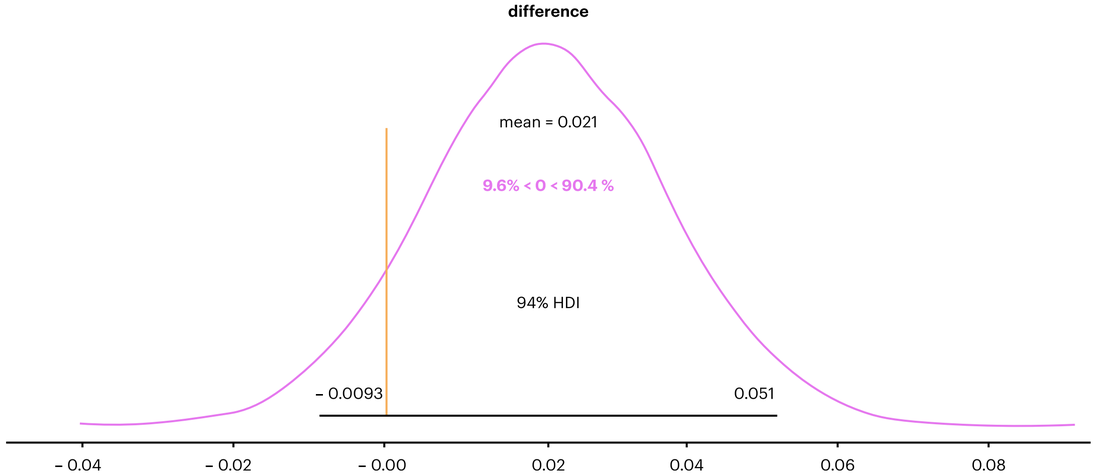

```
# 1 подход
n = 20_000_000
b_group = beta(alpha_b, beta_b).rvs(n)
a_group = beta(alpha_a, beta_a).rvs(n)

print(f'Группа B лучше, чем группа A с вер-тью {(b_group > a_group).mean()*100:.2f}%')
```

```
Группа B лучше, чем группа A с вер-тью 90.40%
```

```
# 2 подход
d1_beta = norm(
    loc=beta.mean(alpha_b, beta_b) - beta.mean(alpha_a, beta_a),
    scale=(beta.var(alpha_a, beta_a) + beta.var(alpha_b, beta_b))**0.5
)

print(f'Группа B лучше, чем группа A с вер-тью {d1_beta.sf(0)*100:.2f}%')
```

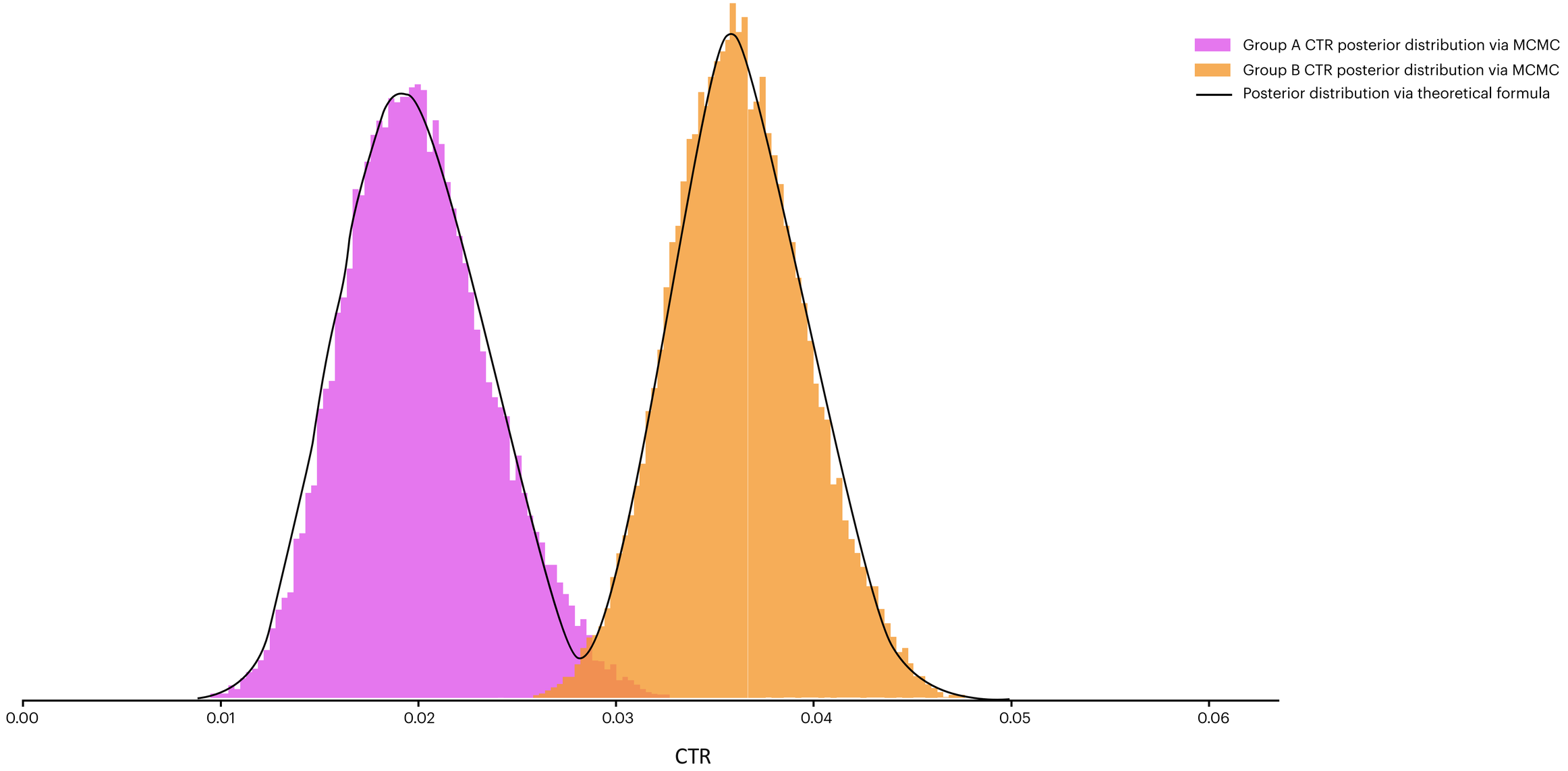
```
Группа B лучше, чем группа A с вер-тью 90.41%
```

Численный метод

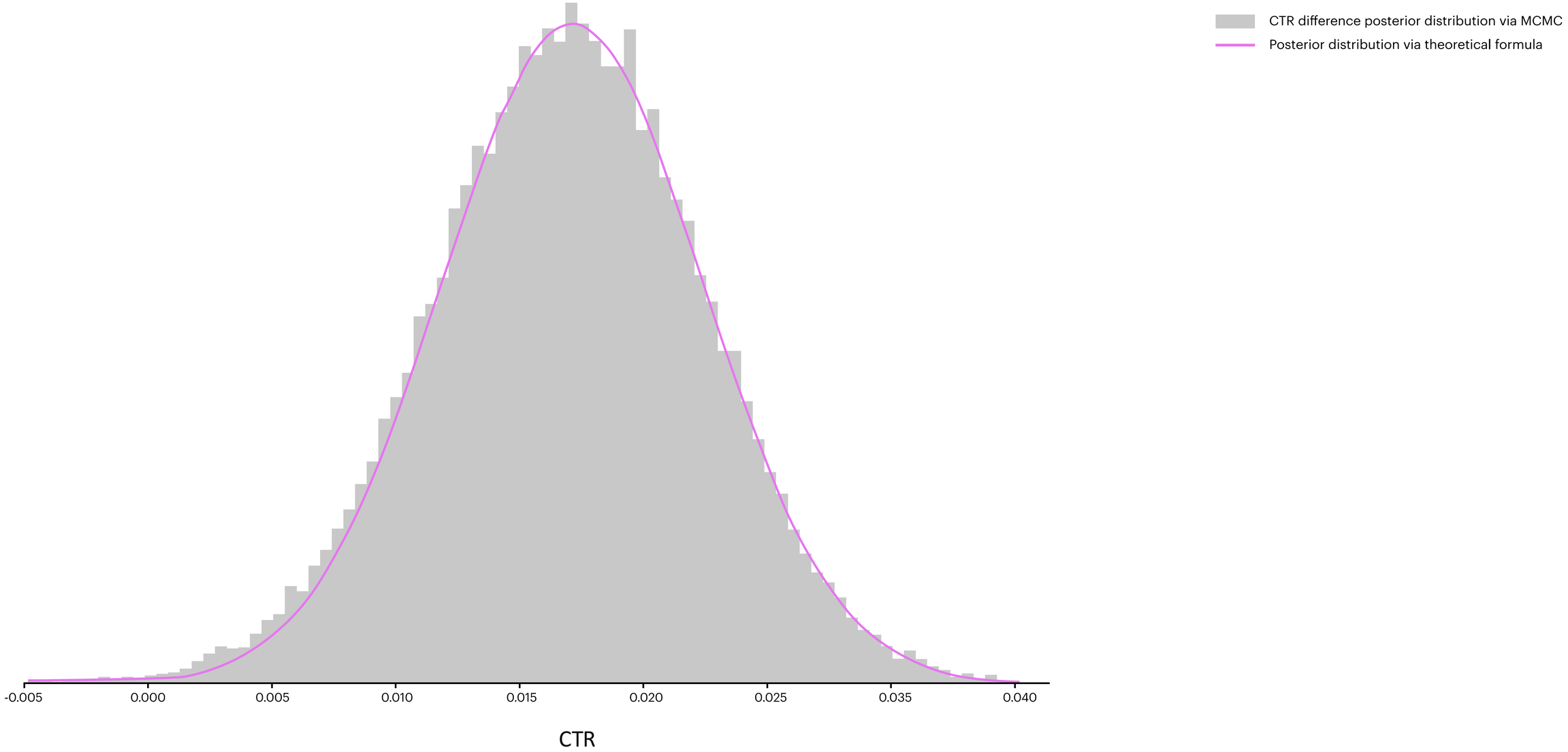



```
with pm.Model() as model:
    p_A = pm.Beta('p_A', alpha=1, beta=1)
    p_B = pm.Beta('p_B', alpha=1, beta=1)

    obs_A = pm.Binomial('obs_A', n=n_a, p=p_A, observed=x_a)
    obs_B = pm.Binomial('obs_B', n=n_b, p=p_B, observed=x_b)

    pm.Deterministic('difference', p_B - p_A)
    trace = pm.sample(draws=50000, progressbar=True)
```

Теория vs MCMC для CTR



Теория vs MCMC для разницы CTR



Мониторинг и ранняя остановка теста

Тезис: Байесовский подход устойчив к мониторингу и ранней остановке теста
в контексте False Discovery Rate

Апостериорные шансы для гипотез после получения данных:

$$\frac{P(H_1|Data)}{P(H_0|Data)} = \frac{P(H_1)}{P(H_0)} \times \frac{P(Data|H_1)}{P(Data|H_0)} \quad \rightarrow \quad \text{Posterior Odds} = \text{Prior Odds} \times \text{Bayes Factor}.$$

При известных апостериорных шансах K, получим следующие вероятности:

$$\mathrm{P}(\mathrm{H}_0|Data) = \frac{1}{1+\mathrm{K}} \qquad \mathrm{P}(\mathrm{H}_1|Data) = \frac{\mathrm{K}}{1+\mathrm{K}}$$

Мониторинг и ранняя остановка теста

$$\frac{P(H_0)}{P(H_1)} = 1{:}1$$

$$H_0 : X \sim N(0,1)$$

$$H_1 : X \sim N(\delta,1)$$

$$\frac{P(\overline{X}|H_1)}{P(\overline{X}|H_0)} = \frac{\exp\left(-(\overline{X}-\delta)^2/(2/N)\right)}{\exp\left(-(\overline{X})^2/(2/N)\right)} = \exp\left(\frac{N}{2}\delta\ (2\overline{X}-\delta)\right)$$

Мониторинг и ранняя остановка теста

Fixed Horzion Test

10000 выборок:

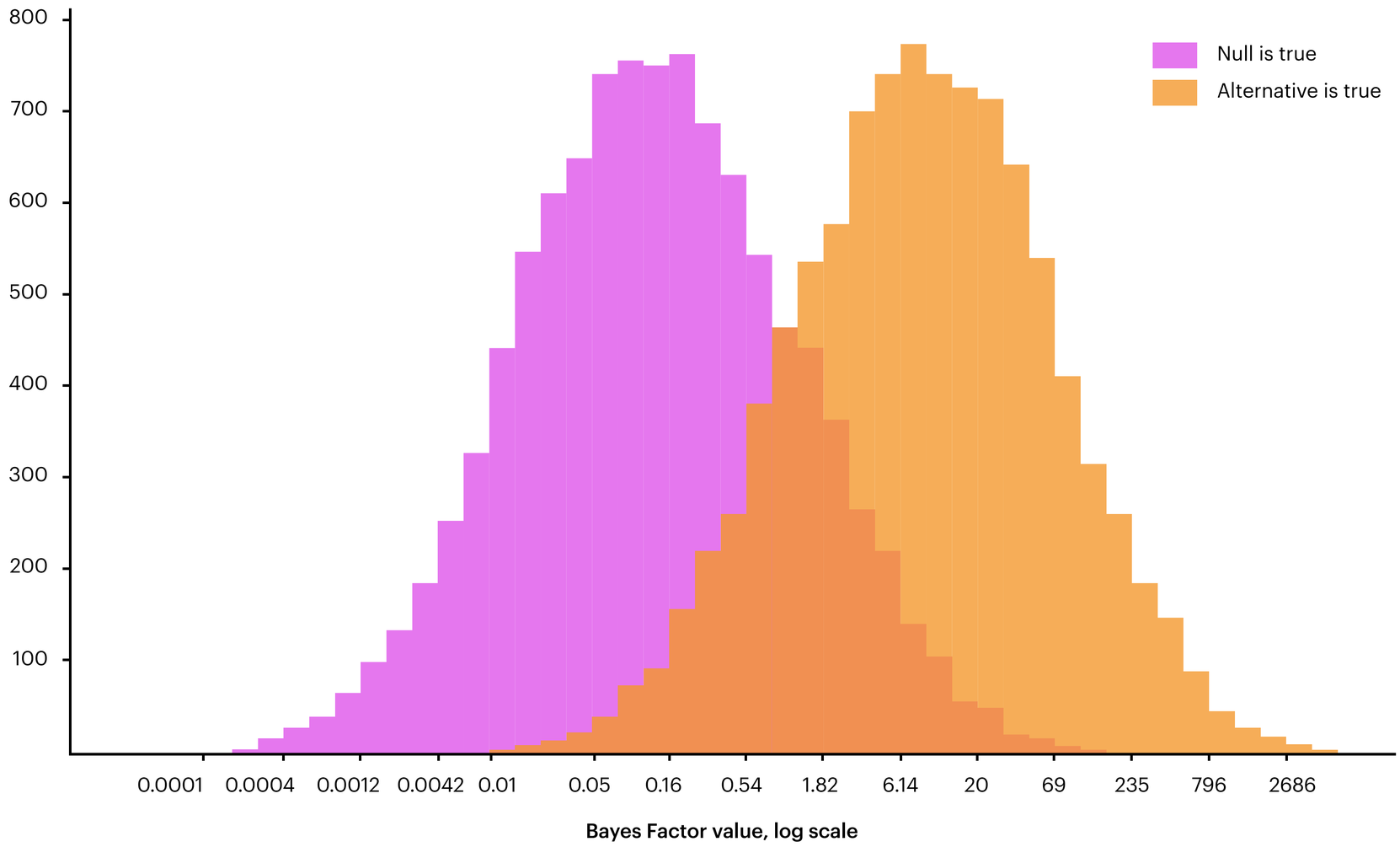
$X_i \sim N(0, 1)$, i=1,..,M

M=100

10000 выборок:

$X_i \sim N(0.2, 1)$, i=1,..,M

M=100



Мониторинг и ранняя остановка теста

Тест с остановкой,
если BF > 10

10000 выборок:

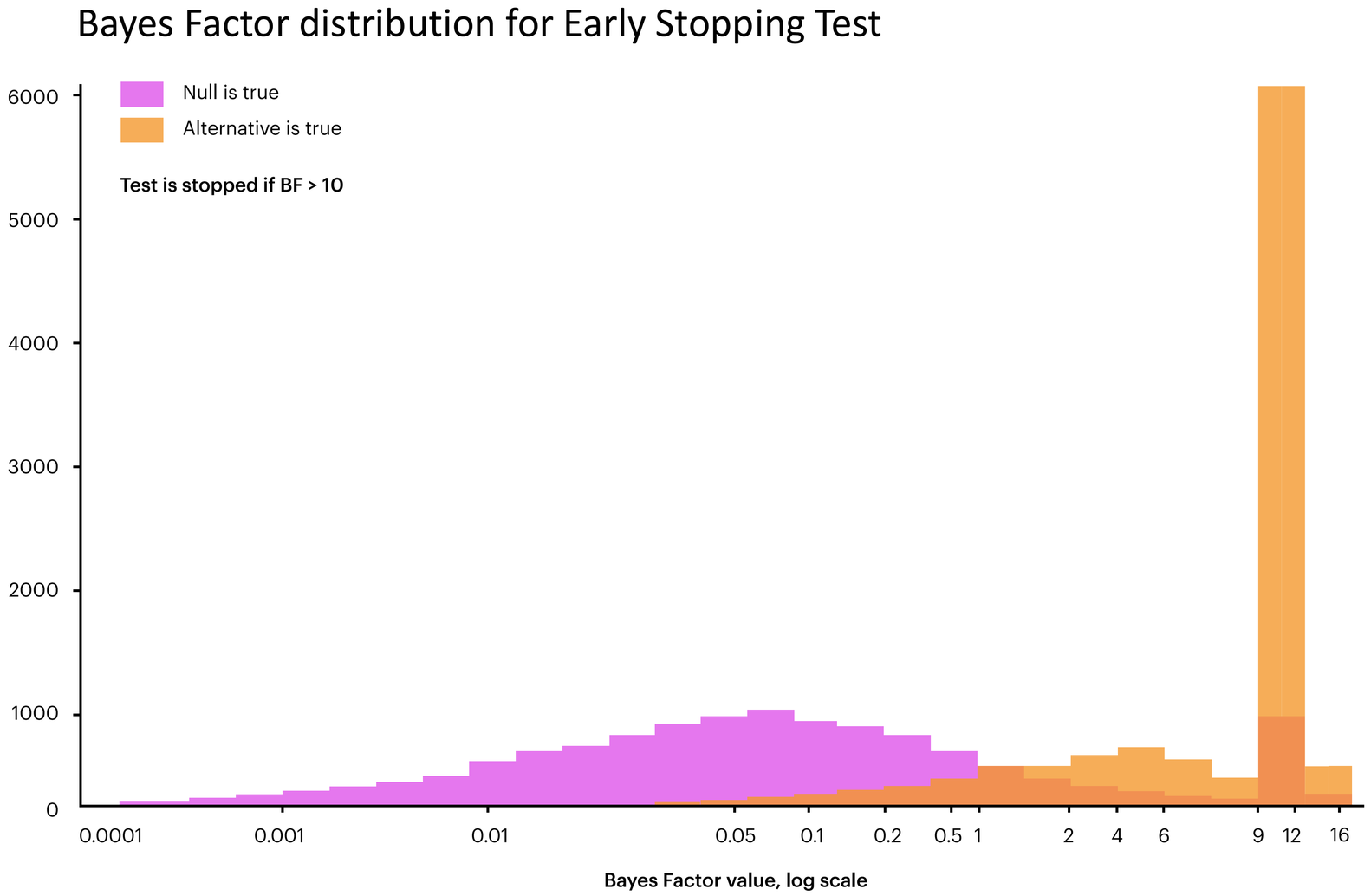
$X_i \sim N(0, 1)$, i=1,..,M

$M=\min(100, M_{stopped})$

10000 выборок:

$X_i \sim N(0.2, 1)$, i=1,..,M

$M=\min(100, M_{stopped})$



Мониторинг и ранняя остановка теста

Тест с остановкой,
если BF > 10 или BF < 1/10

10000 выборок:

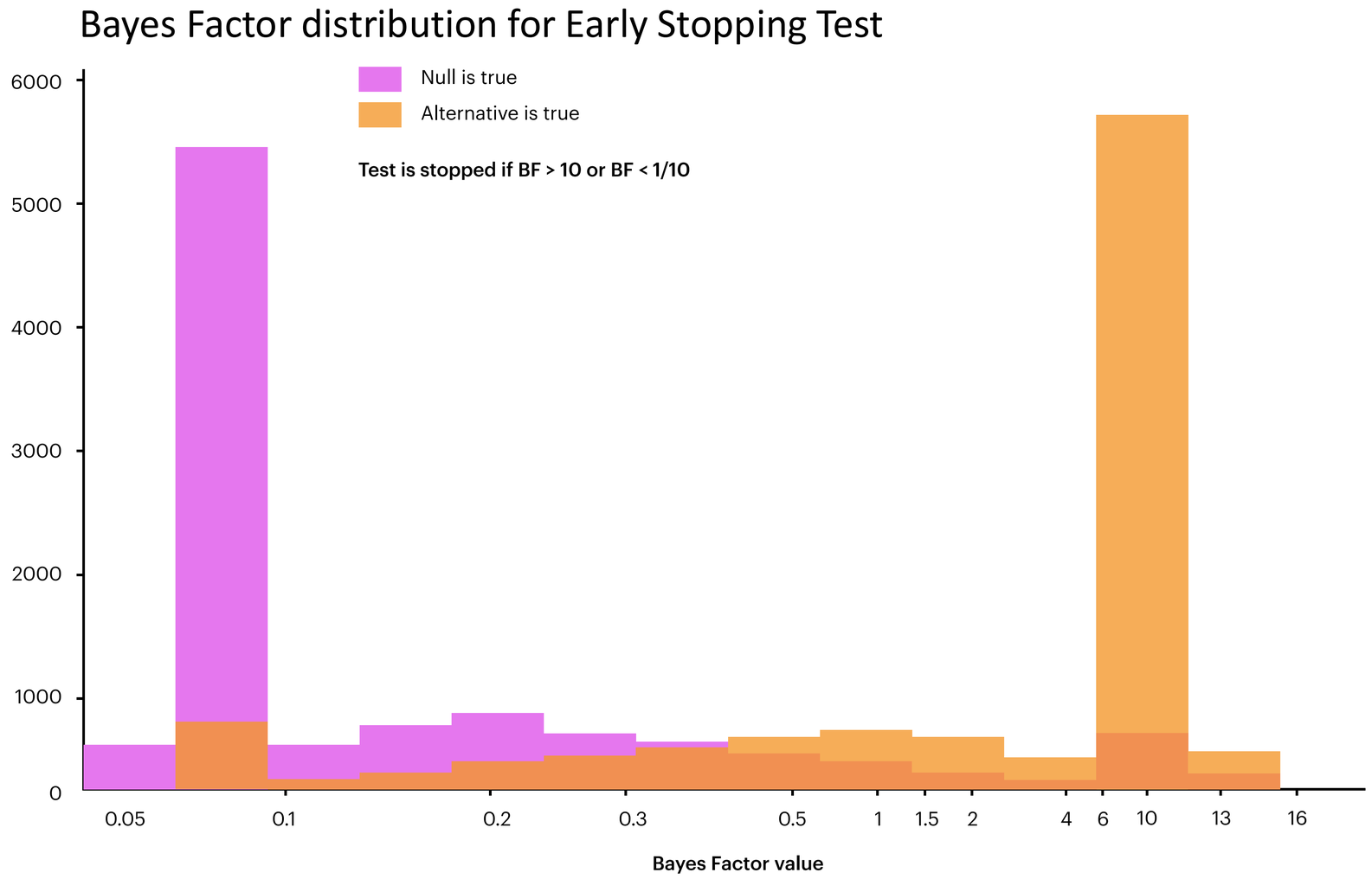
$X_i \sim N(0, 1)$, i=1,..,M

$M=\min(100, M_{stopped})$

10000 выборок:

$X_i \sim N(0.2, 1)$, i=1,..,M

$M=\min(100, M_{stopped})$



Сравнение частотного и байесовского подходов

	Частотный подход	Байесовский подход
Оцениваемый параметр	Неизвестный и фиксированный	Неизвестный и имеющий распределение
Результат на выходе	P-value, доверительный интервал	Апостериорная вероятность для параметра
Механизм расчета	Статистическая проверка H_0 против H_1	Формула Байеса, необходимо априорное знание о параметре
Интерпретируемость	Низкая, контр-фактуальная логика	Высокая, расчет произвольных оценок на основе апостериорного распределения
Контролируемая ошибка при множественном сравнении	FWER	FDR
Непрерывный мониторинг и ранняя остановка теста	Увеличивает FWER, необходимы методы последовательного тестирования	Не влияет на FDR, не требует коррекции
Наиболее подходящая область применения	Там где цена ошибки I рода высока: научные исследования, клинические испытания	Там где важна быстрота принятия большого количества решений, интерпретируемость: **внедрение улучшений в бизнесе**

Почему же Байес не такой популярный как частотный?

Надо понять почему его не так часто используют (вдруг есть принципиальные косяки и мы дилетанты)

Почему байесовский подход не так широко используется в А/Б-тестах?

- Дело привычки. P-value давно и часто используется в А/Б-тестах.
- Меньше решений на рынке, чем для частотного подхода. Хотя уже есть реализации в крупных продуктах для АБ-тестов: Growthbook, VWO.
- Требует больших усилий, чем частотный подход: сложнее вычисляется, методика А/Б-теста менее проработана.»

Список зарубежных компаний и платформ которые используют его – на западе это популярно

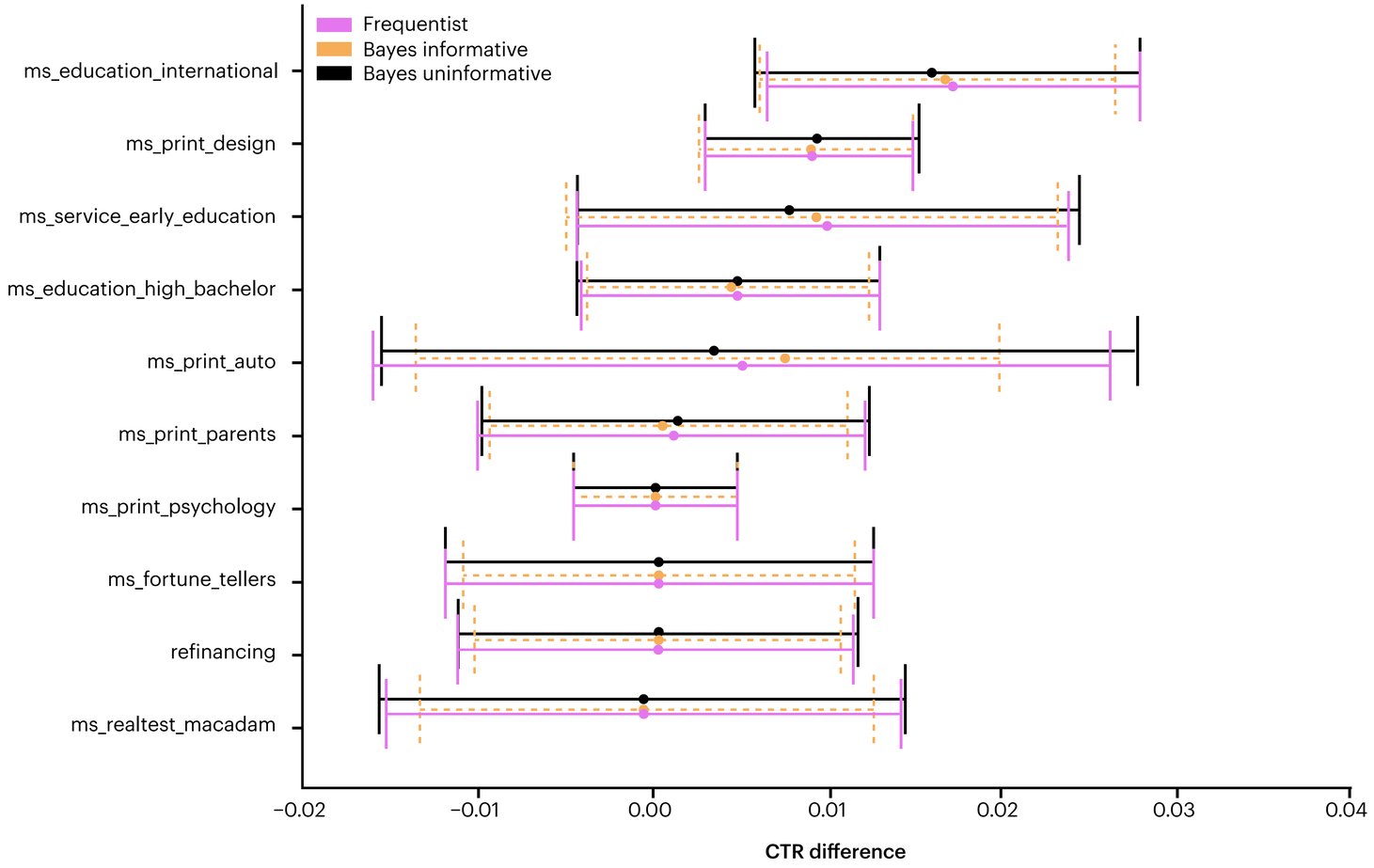
VWO, Growthbook, Optimizely, Lyft, etc

МТС Маркетолог

Рассмотрим результаты различных тестов

Сравним в 10 тестах ширины доверительных интервалов построенных разными подходами

Байесовские интервалы на основе **информативного априорного** распределения **самые узкие**!



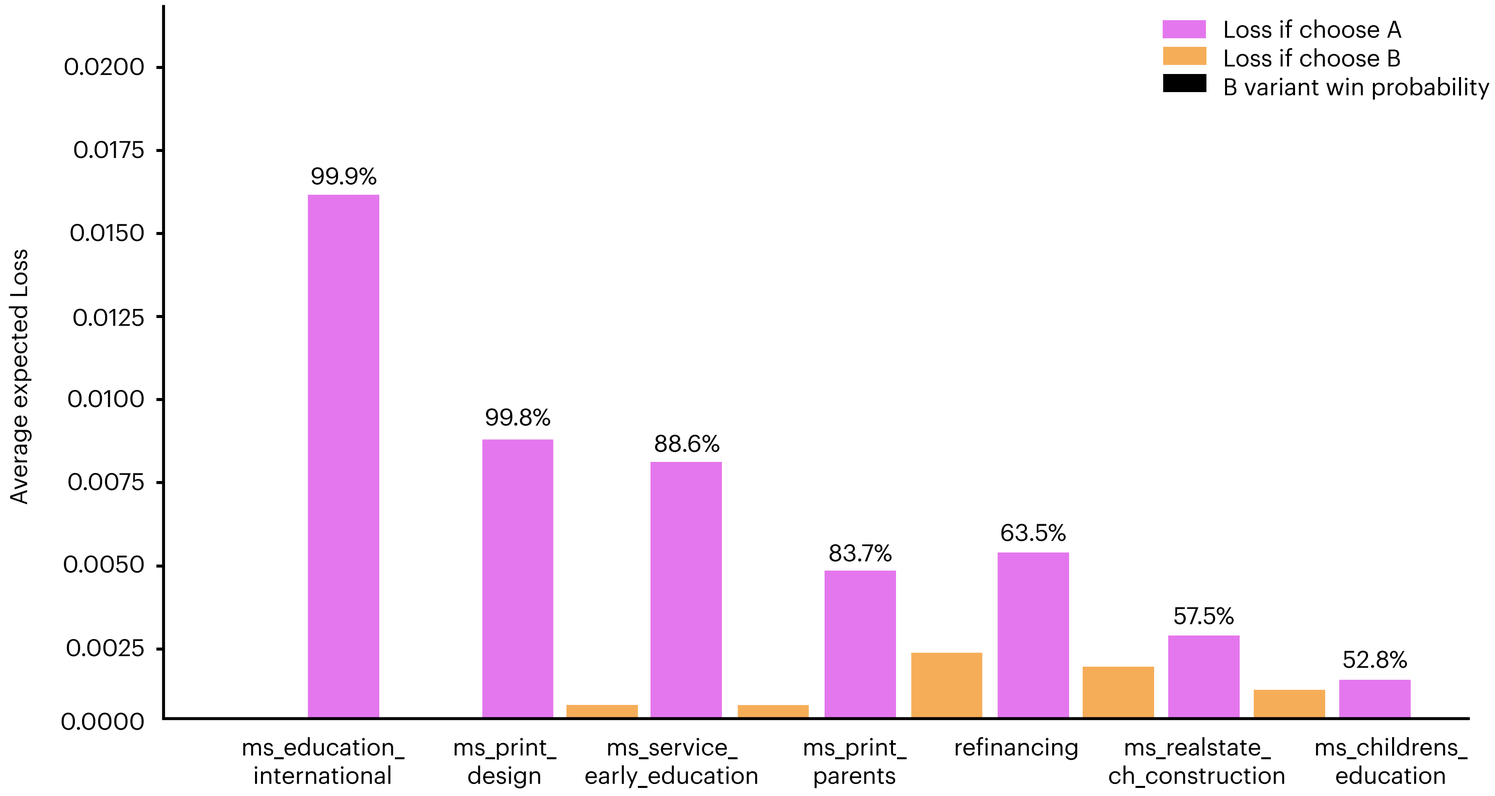
МТС Маркетолог

Рассмотрим результаты различных тестов

Вероятность успеха и оценка потенциальных потерь позволяют гибко принимать решение о внедрении новых изменений

Можно включить запас δ в Loss функцию для контроля расходов при переходе на вариант *B*:

$$L(\lambda_A, \lambda_B, B) = \max(\lambda_A - (\lambda_B - \delta), 0)$$



МТС Кион

Результаты непрерывного мониторинга Эксперимента #1

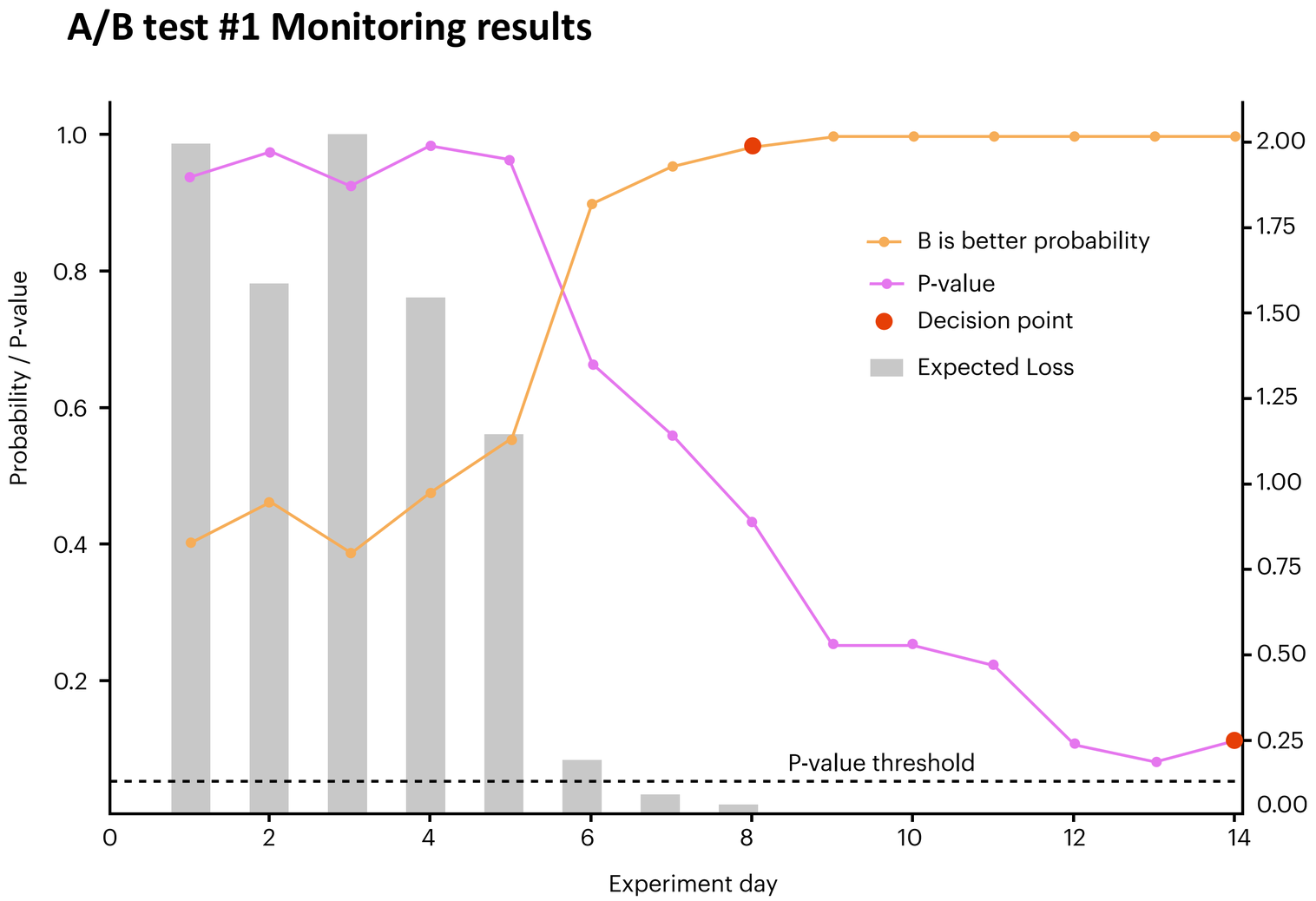
Использование байесовского подхода и априорных знаний позволяют досрочно принять решение о результате теста

Байесовский подход:

Тест успешно остановлен на 8 день

Частотный подход:

Тест не прокрасился в конце срока (14 день)



МТС Кион

Результаты непрерывного мониторинга Эксперимента #2

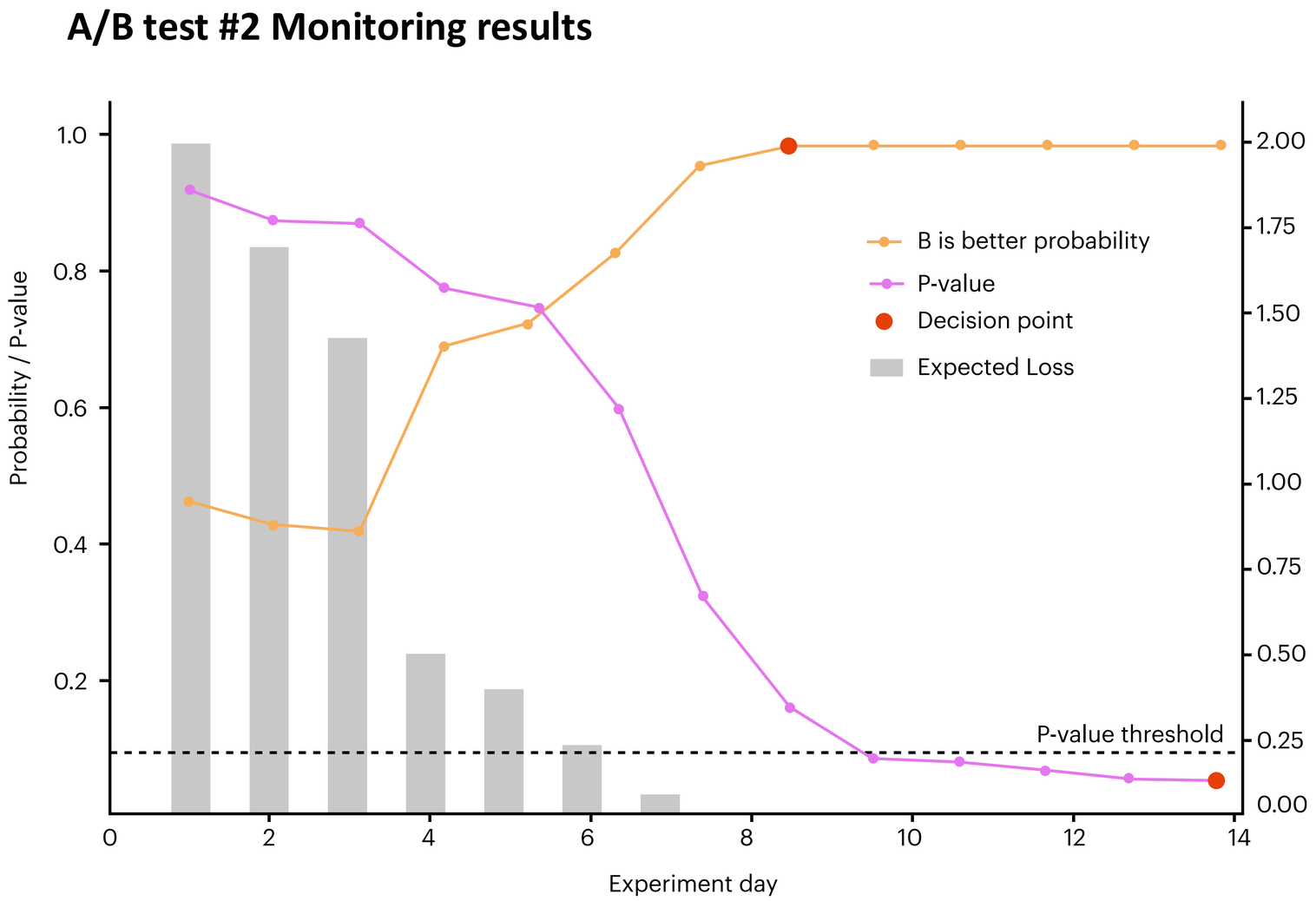
Использование байесовского подхода и априорных знаний позволяют досрочно принять решение о результате теста

Байесовский подход:

Тест успешно остановлен на 8 день

Частотный подход:

Тест прокрасился в конце срока (13 день)



МТС Кион

Результаты непрерывного мониторинга Эксперимента #3

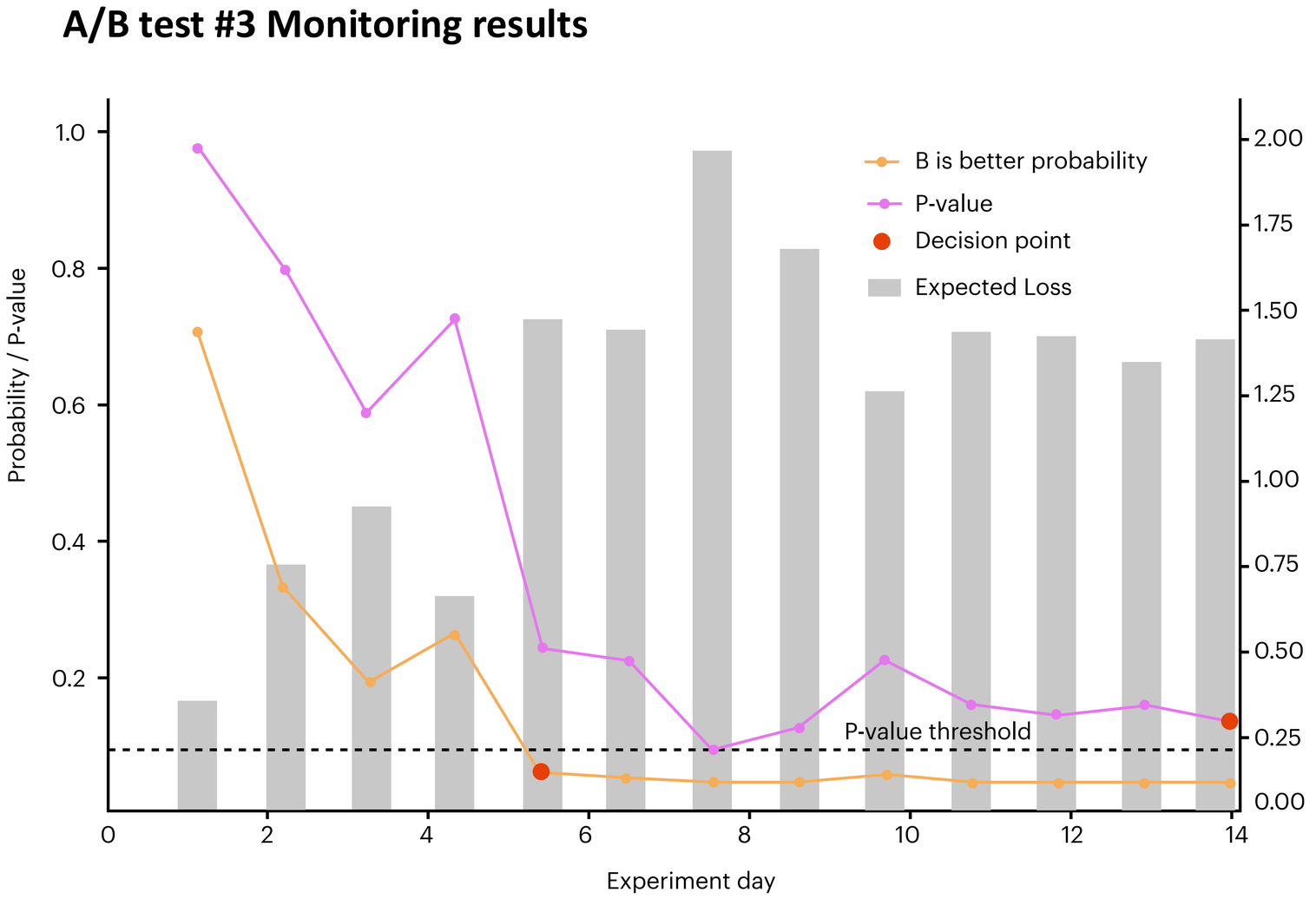
Использование байесовского подхода и априорных знаний позволяют вовремя остановить неудачный тест

Байесовский подход:

Тест остановлен на 5 день
из-за низких результатов
в группе B

Частотный подход:

Тест не прокрасился
в конце срока (14 день)



Почему же Байес не такой популярный как частотный?

Историческая предыдущая практика. Частотные методы имеют более долгую историю использования в A/B тестировании и являются широко распространенными в индустрии. Многие компании уже имеют установленные процедуры для проведения частотных A/B тестов, и им необходимо дополнительное время и усилия для перехода на байесовские методы.

Распространенные инструменты. Множество статистических инструментов и библиотек предоставляют готовые функции и процедуры для проведения частотных тестов. Это делает их удобными и доступными для использования в практике.

Простота применения. Частотные методы обладают простыми математическими выражениями. Это делает их более доступными для широкого круга пользователей, включая неспециалистов в статистике. Методика байесовского подхода к A/B тестированию проработана куда меньше.