

**ТПОЭ - 23/24**

**Лекция 7**

**Нерсес Багиян**

# Смотрим на три типа метрик

## Все свели к нормальному распределению

### Средняя

- User Average Metrics (ARPU/  
ARPPU/etc)

### Ratio

- User-level Conversion Metrics  
(Retention / etc)
- Page-level Conversion Metrics  
(Global CTR / etc)

### Квантиль

- Ну тут просто квантиль (.99 latency / перцентиль чека)

### Абсолюты

- Метрики (GMV / Выручка /  
Просмотры)

# Растим ARPU с помощью рекомендаций аксессуаров

## 1. Формулировка гипотезы с сформулированным ожидаемым размером эффекта

Предложение мыла с подборкой популярных аксессуаров на основе анализа предпочтений покупателей и данных о самых продаваемых товарах увеличит средний доход на пользователя (ARPU) на 20%.

## 2. Описание аудитории

Покупатели онлайн-магазина мыла, включая как новых, так и возвращающихся пользователей.

## 3. Описание вариантов с размером каждой группы

**Контрольная группа (A):** Покупателям предлагается стандартный ассортимент без акцентов на комплекты.

**Экспериментальная группа (B):** Покупателям активно предлагаются комплекты мыла с популярными ароматами на главной странице и в разделе рекомендаций.

Размер каждой группы составляет 50% от общего числа посетителей в период эксперимента.

## 4. Ожидаемые исходы и метрики

**Основная метрика:** Увеличение **среднего чека, .75 квантиля, .1 квантиля**

**Контрметрика:** Не падение **конверсии в покупку**

## 5. Продолжительность

Эксперимент продлится 4 недели, чтобы собрать достаточно данных для статистически значимых результатов, учитывая недельные колебания трафика и поведения покупателей.

## 6. Результаты

TBD

# Берем нашу любимую формулу и считаем

$$n \geq \frac{2(F^{-1}(1 - \frac{\alpha}{2}) - F^{-1}(\beta))^2 s^2}{MDE^2}$$

**Тест запустили! Ура! Расходимся?**

**Вопрос аудитории**

# Давайте знакомиться

Это Даниил



Он был заказчиком А/Б теста и получил ваш дизайн эксперимента с длительностью эксперимента

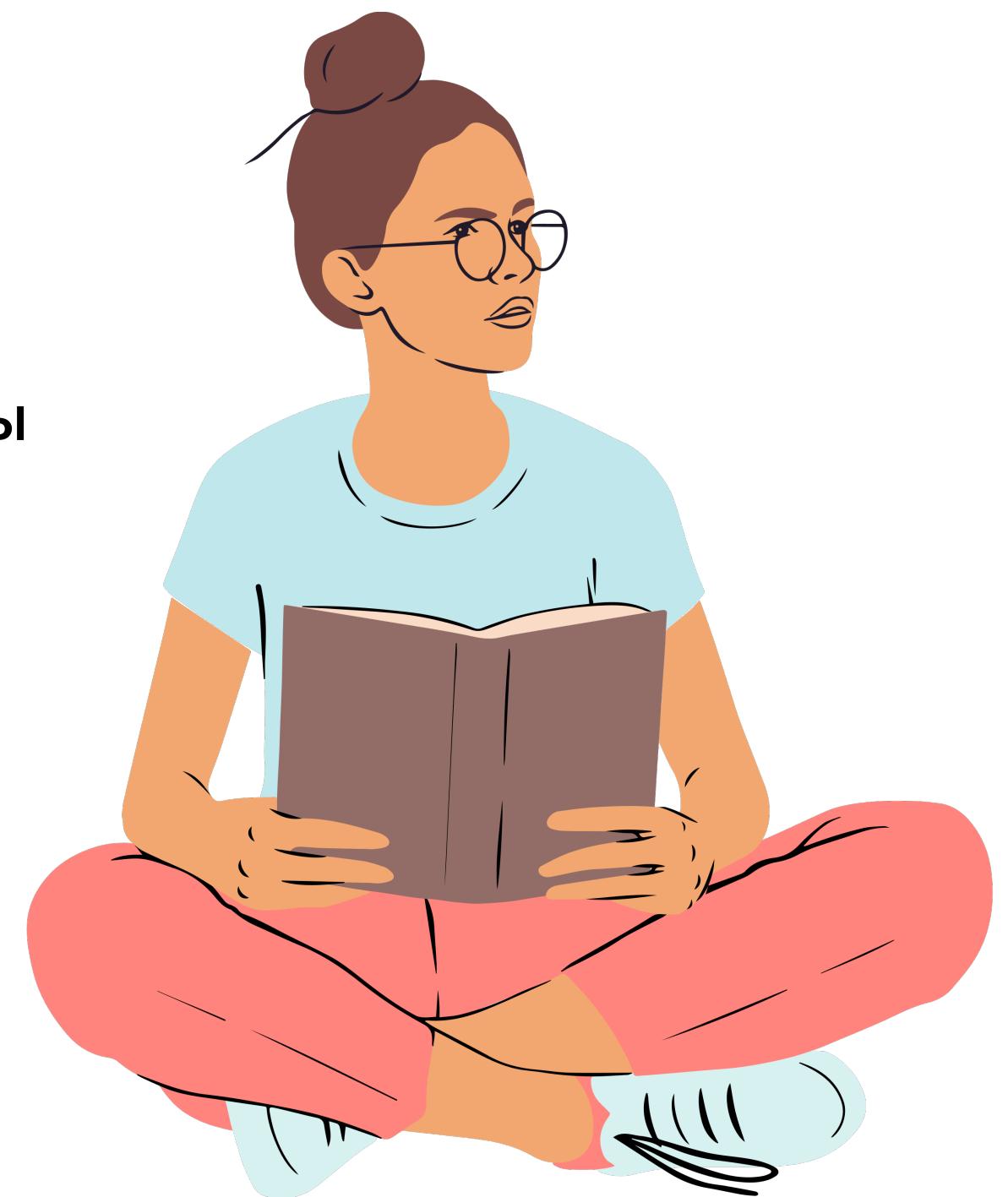
Он очень хочет по скорее получить результаты, потому что скоро ревью и нужна премия

# У Даниила есть аналитик

Аналитика зовут Елизавета



Вместе они работают над экспериментом и все вопросы  
про эксперимент Даниил задает Лизе



# И решил написать команде про дизайн

## Решив, что сделать это в 21:53 хорошая идея



The screenshot shows a Slack interface for the '#anal-team' channel. The sidebar on the left lists various channels and apps, with '# anal-team' currently selected. The main pane displays a conversation between three users: Daniil, Liza Analitika, and Elena Nowak. The messages are in Russian. The timestamp 'Friday, June 14th' is visible at the bottom of the message list. A message input field at the bottom is pre-filled with 'Message #dev-team'.

# anal-team Track and coordinate w/ analytics

Project brief Resources To do Jira board +

Дань, я добавил события, все на проде!

Даниил 18:03 Кайф! Настя, ну что там? Где на онбординге у нас основной затык?

Лиза Аналитика 18:04 Эмммм, а я не могу сказать - события же только начали накапливаться, а по тому количеству что есть сейчас я выводов сделать не могу, надо подождать пока накопятся, думаю это недели две

Даниил 18:04 ...ясно, ждем

Friday, June 14th

Даниил 21:53 Лиза, привет! Я посмотрел дизайн эскра и там написано, что его надо держать 3-4 недели. А можно ли как-то побыстрее? Хочу скорее увидеть результаты

Лиза Аналитика 21:59 Слушай, ну с текущим инструментарием можно только так

Даниил 22:01 А как же вот эти вот модные способы уменьшать дисперсию

Лиза Аналитика 22:01 Да, это первое о чем я подумала, но... это не так уж и просто и не всегда работает

Даниил 22:02 Ок, давай запускать эксперимент, но попробуем изучать эти методы

Message #dev-team

# **За счет чего можно снизить дисперсию?**

## **Вопрос аудитории**

$$n \geq \frac{2(F^{-1}(1 - \frac{\alpha}{2}) - F^{-1}(\beta))^2 s^2}{MDE^2}$$

# Внимательно смотрим на нашу любимую формулу

По большому счету, мы в этой формуле влияем только на дисперсию, так как другие слагаемые мы фиксируем при расчете формулы

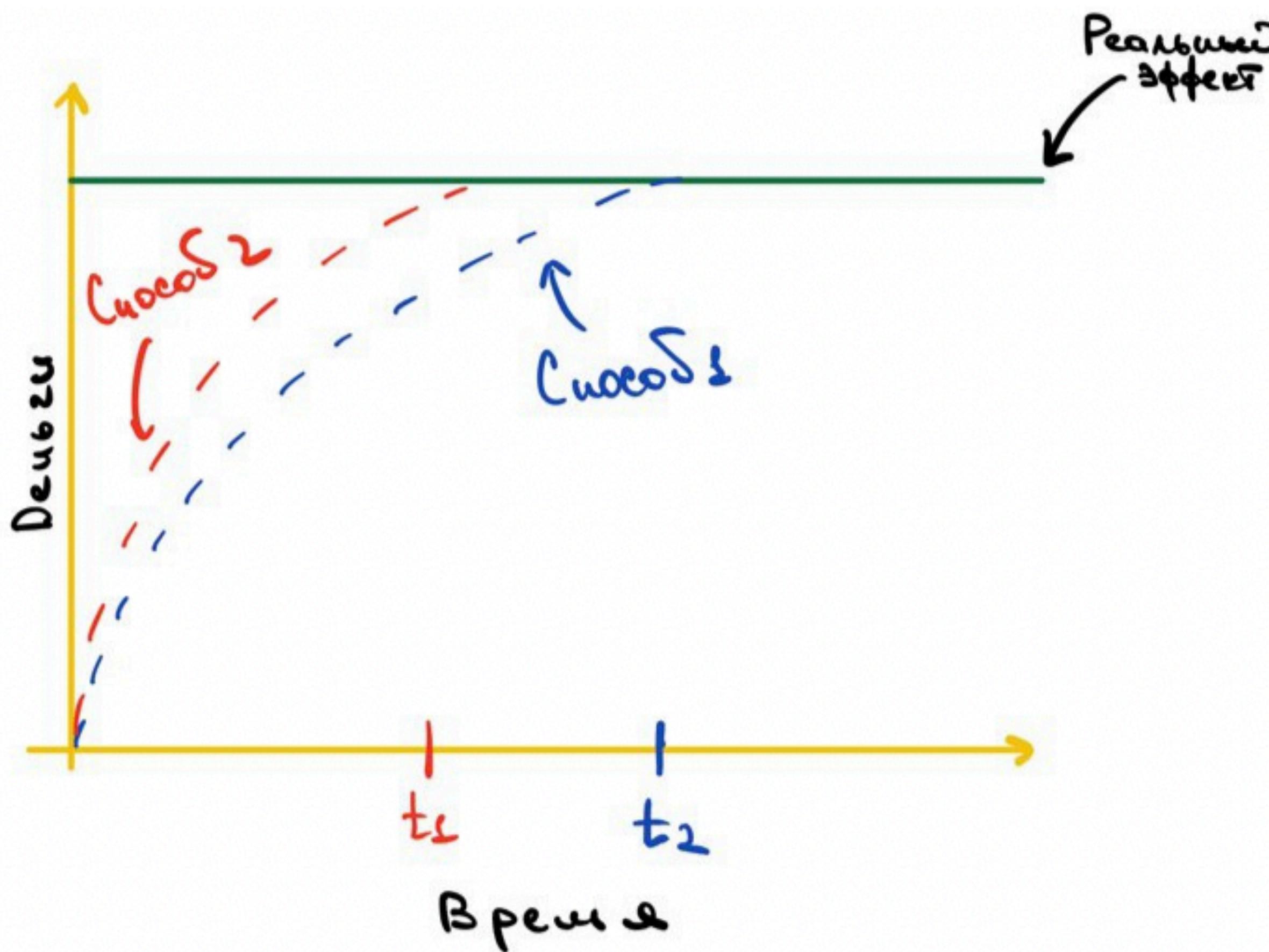
$$n \geq \frac{2(F^{-1}(1 - \frac{\alpha}{2}) - F^{-1}(\beta))^2 s^2}{MDE^2}$$

**Как думаете почему нам вообще важно провести эксперимент быстрее?**

**Вопрос аудитории**

# Ускоряем тестирование

Зачем это нужно?

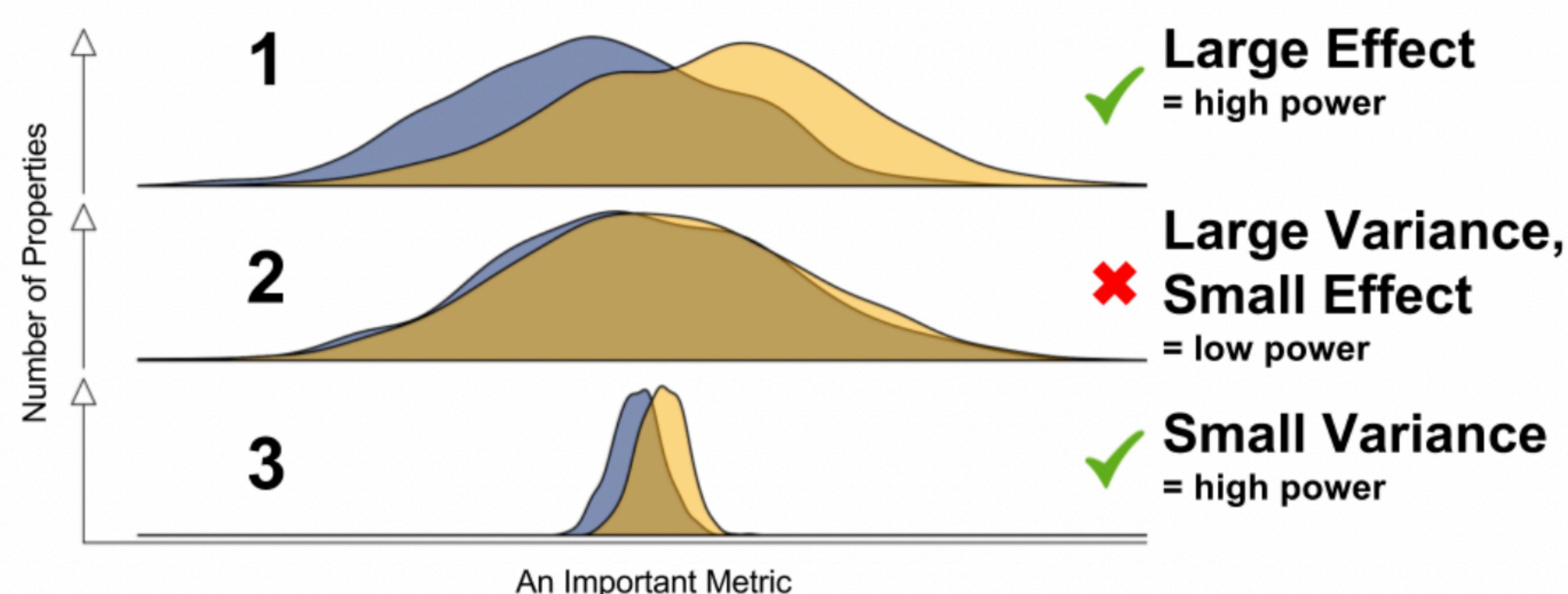


1. Плохие эксперименты => останавливаем раньше => меньше денег теряем
2. Хорошие эксперименты => запускаем раньше => больше денег зарабатываем

# Каждый эксперимент можно свести к одному из 3 случаев

Дисперсия показывает, насколько данные в выборке разбросаны относительно их среднего значения. Чем меньше разброс, тем больше наша уверенность в оценке нашего среднего.

**Вопрос: как мы можем этот разброс уменьшить и перейти из 2 случая к 3?**



# Можно использовать исторические данные

**CUPED (Controlled-experiment Using Pre-Experiment Data)** - преобразование, использующее данные до эксперимента, чтобы снизить дисперсию метрики

Пример:

Мы в нашем магазине мыла проводим эксперимент, чтобы понять увеличивают ли наши рекомендации средней чек. У нас есть разные юзеры, некоторые из них могут купить несколько раз за эксперимент. Как бы работал CUPED?

# Пусть у нас есть некоторый эксперимент

Данные по группам. Обычная линия - группа А, пунктир - группа Б

920 рублей

1200 рублей

1050 рублей

1100 рублей

День N: эксперимент

# Пусть у нас есть некоторый эксперимент

## Может посмотреть на разницу относительно их предыдущего поведения?

Данные по группам. Обычная линия - группа А, пунктир - группа Б

Стандартное отклонение:  
100.8 рублей

920 рублей

920 рублей

920 рублей

1000 рублей

1000 рублей

1000 рублей

1050 рублей

1050 рублей

1050 рублей

1000 рублей

1000 рублей

1000 рублей

День N-3: нет экспа

День N-2: нет экспа

День N-1: нет экспа

920 рублей

1200 рублей

1050 рублей

1100 рублей

День N: эксперимент

# Пусть у нас есть некоторый эксперимент

## Может посмотреть на разницу относительно их предыдущего поведения?

Данные по группам. Обычная линия - группа А, пунктир - группа Б

Стандартное отклонение:  
82.9 рублей

0 рублей

День N-3: нет экспа

День N-2: нет экспа

День N-1: нет экспа

0 рублей

200 рублей

0 рублей

100 рублей

День N: эксперимент

# CUPED

**CUPED (Controlled-experiment Using Pre-Experiment Data) - преобразование, использующее данные до эксперимента, чтобы снизить дисперсию метрики**

Пусть у нас есть метрика  $\bar{X}$  посчитанная в эксперименте, являющаяся поузерным средним. И есть метрика  $\bar{Y}$  являющаяся поузерным средним до эксперимента. Тогда, преобразование будет выглядеть следующим образом:

$$\bar{X}_{\text{cuped}} = \bar{X} - w\bar{Y} + wE[Y]$$

# Как найти параметр $w$ ?

**Наша цель снизить дисперсию, давайте посмотрим на дисперсию новой метрики:**

$$V[\bar{X}_{\text{cuped}}] = V[\bar{X} - w\bar{Y} + wE[Y]] = \frac{V[X - wY]}{n} = \frac{V[X] - 2w \operatorname{cov}(X, Y) + w^2V[Y]}{n}$$

**Можно решить задачу оптимизации  $\min_w V[\bar{X}_{\text{cuped}}]$  и получим:**

$$w = \frac{\operatorname{cov}(X, Y)}{V[Y]}$$

# Алгоритм применения CUPED

## Пишем пошагово

Пусть у нас есть метрика  $\bar{X}$  посчитанная в эксперименте, являющаяся поюзерным средним. И есть метрика  $\bar{Y}$  являющаяся поюзерным средним до эксперимента. Тогда, алгоритм будет выглядеть следующим образом:

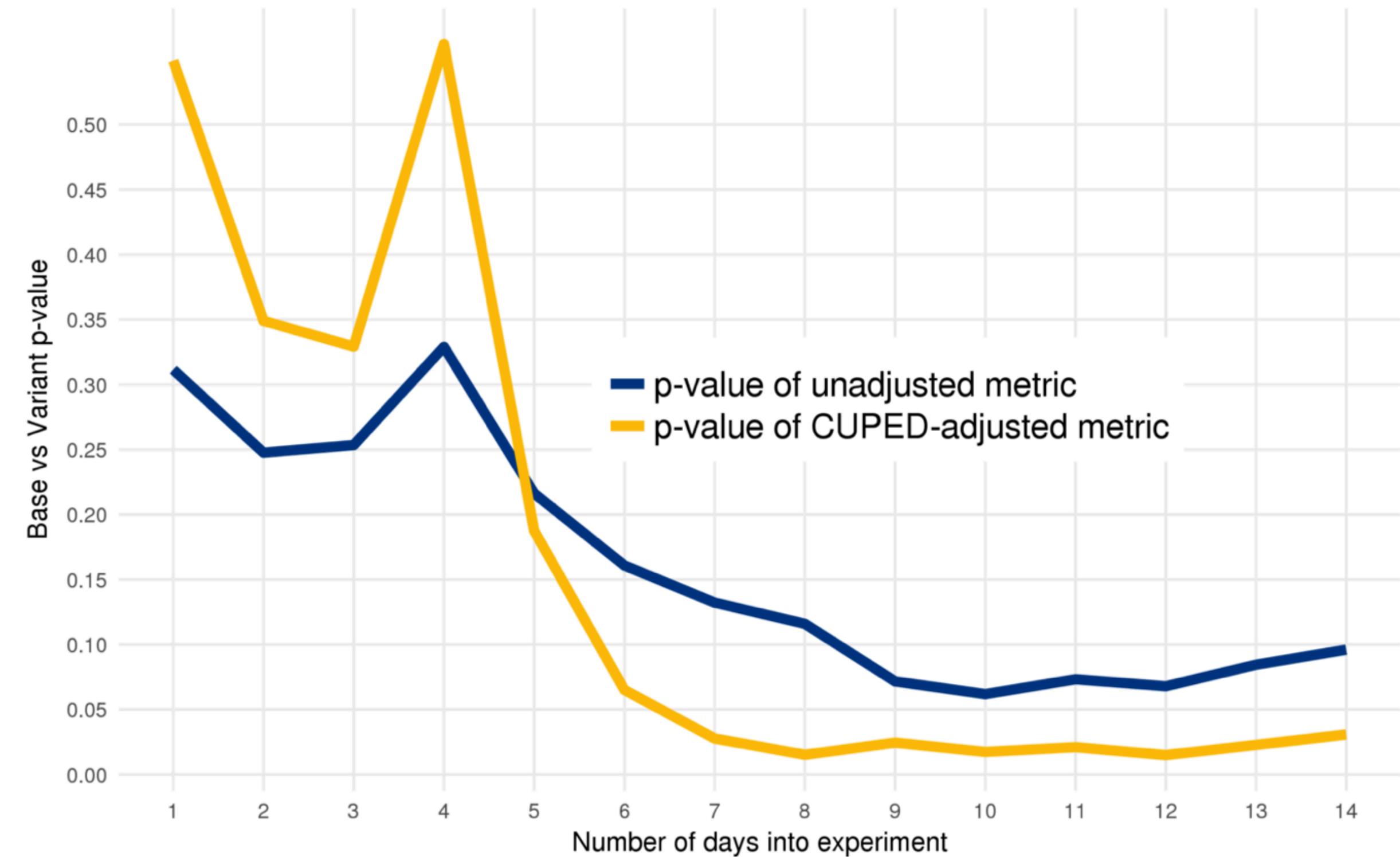
1. Считаем поюзерно метрику в эксперименте и до него
2. Считаем по формуле параметр  $w$
3. Считаем поюзерную купед метрику
4. Применяем статистический тест

**Как думаете, в чем плюс линеаризации?**

**Вопрос аудитории**

# Пример Airbnb

На данном примере мы видим, что применение cuped предобра



# Продолжаем нашу историю

## Следим за перепиской в слаке

A screenshot of a Slack application window titled "Our Product Team". The sidebar shows various channels and apps. The main area is a conversation in the "#anal-team" channel. The messages are as follows:

# anal-team ▾ Track and coordinate w/ analytics

- Project brief Resources To do Jira board +
- Слушай, ну с текущим инструментарием можно только так
- Даниил 22:01 А как же вот эти вот модные способы уменьшать дисперсию
- Лиза Аналитика 22:01 Да, это первое о чем я подумала, но... это не так уж и просто и не всегда работает
- Даниил 22:02 Ok, давай запускать эксперимент, но попробуем поизучать эти методы

Friday, June 26th

Даниил 22:53 Лиза, привет! Видел, что мы прикрутили CUPED и эксперимент уже крутится. Давай проверим вдруг уже можно остановить?

Saturday, June 27th

Даниил 22:53 Лиза, доброе утро! Давай сегодня тоже глянем?

Sunday, June 28th

Даниил 22:53 Лизаааа, привет! А сегодня тоже можем?

Message #dev-team

Message input field with rich text editor icons: bold, italic, underline, etc.

# Даниил стал жертвой проблемы подглядывания



Он хочет остановить эксперимент, как только увидит стат. значимую разницу. Это приведет к проблеме подглядывания. Давайте разберемся откуда она появляется

# Давайте еще раз посмотрим на p-value

Есть два способа дать такое определение:

1. Общее распределение

$$p = P(T \geq t_{obs} | H_0)$$

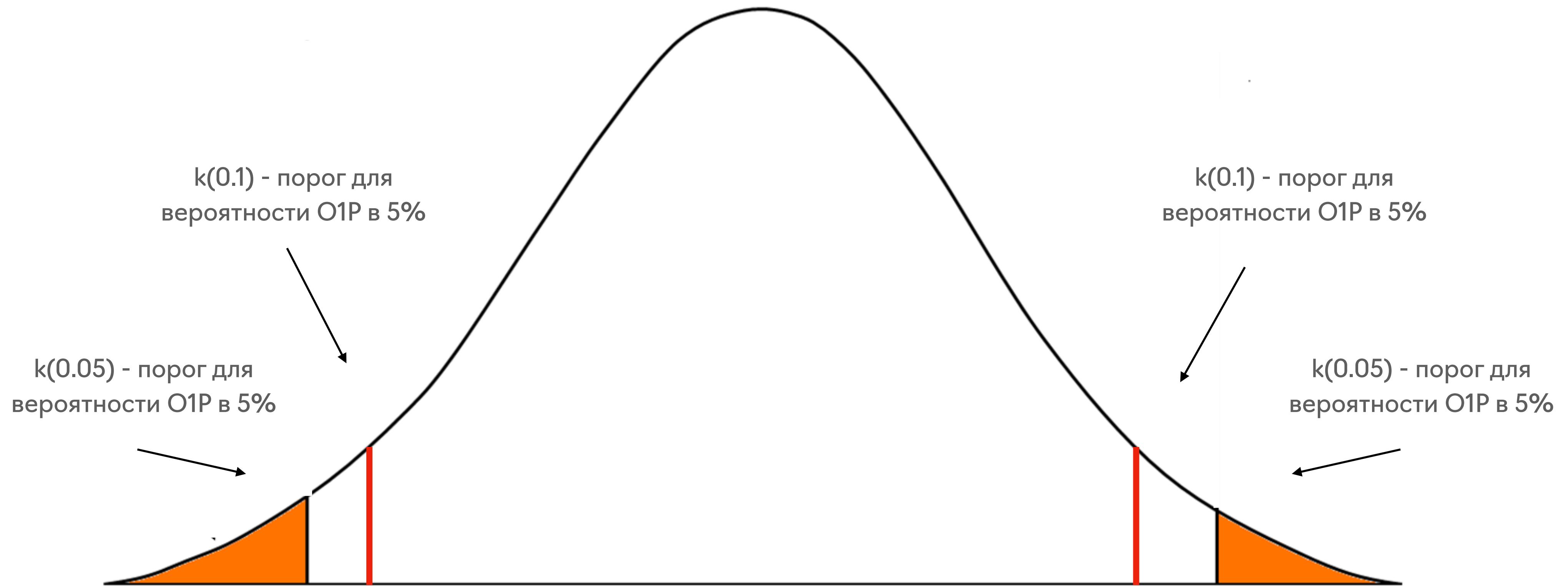
где  $t_{obs}$  - статистика нашего тест

2. Альтернативное распределение

$$p = \inf\{\alpha : t_{obs} \geq k(\alpha)\}$$

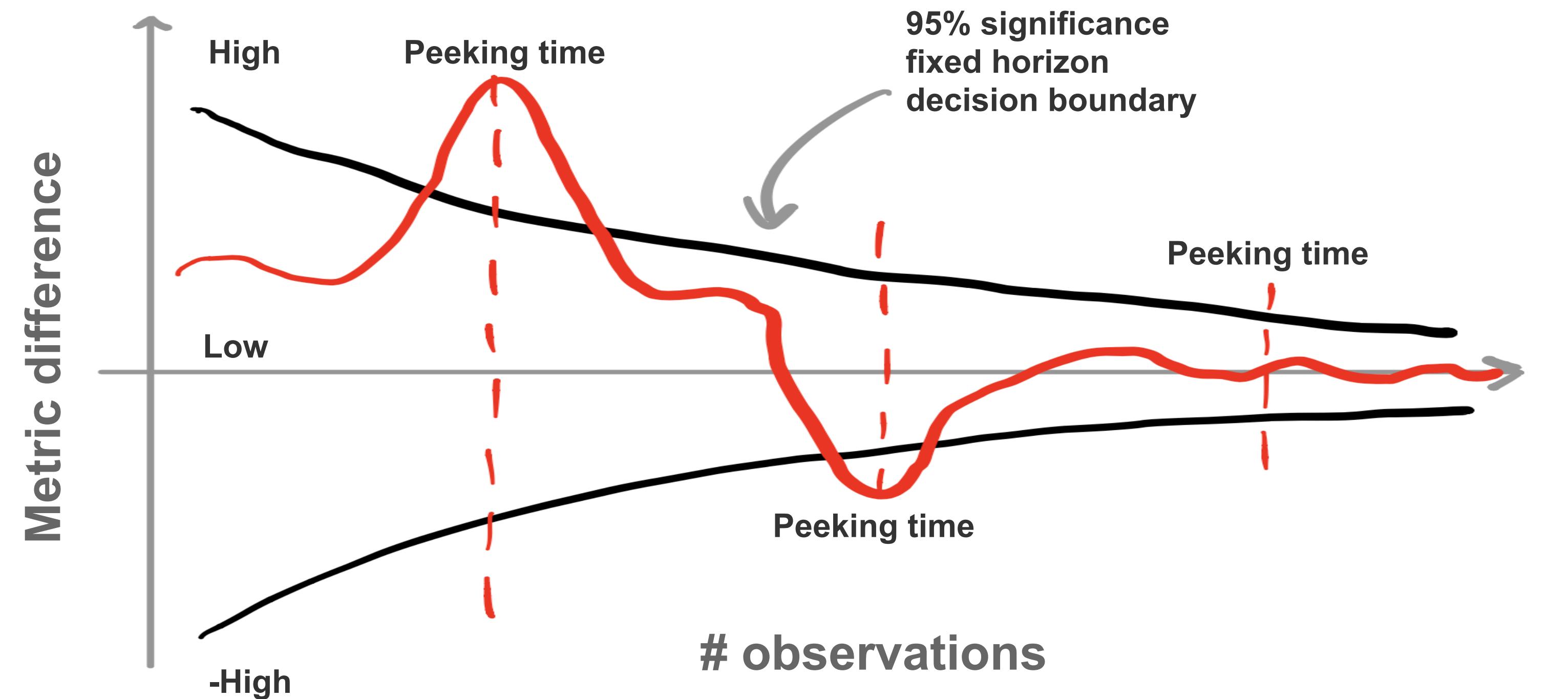
где  $k(\alpha)$  порог параметризованный вероятностью ошибки первого рода. Этот тест отвергает нулевую гипотезу, когда  $t_{obs}$  превосходит  $k(\alpha)$

# Посмотрим на примере какой-то статистики теста



# Decision boundaries

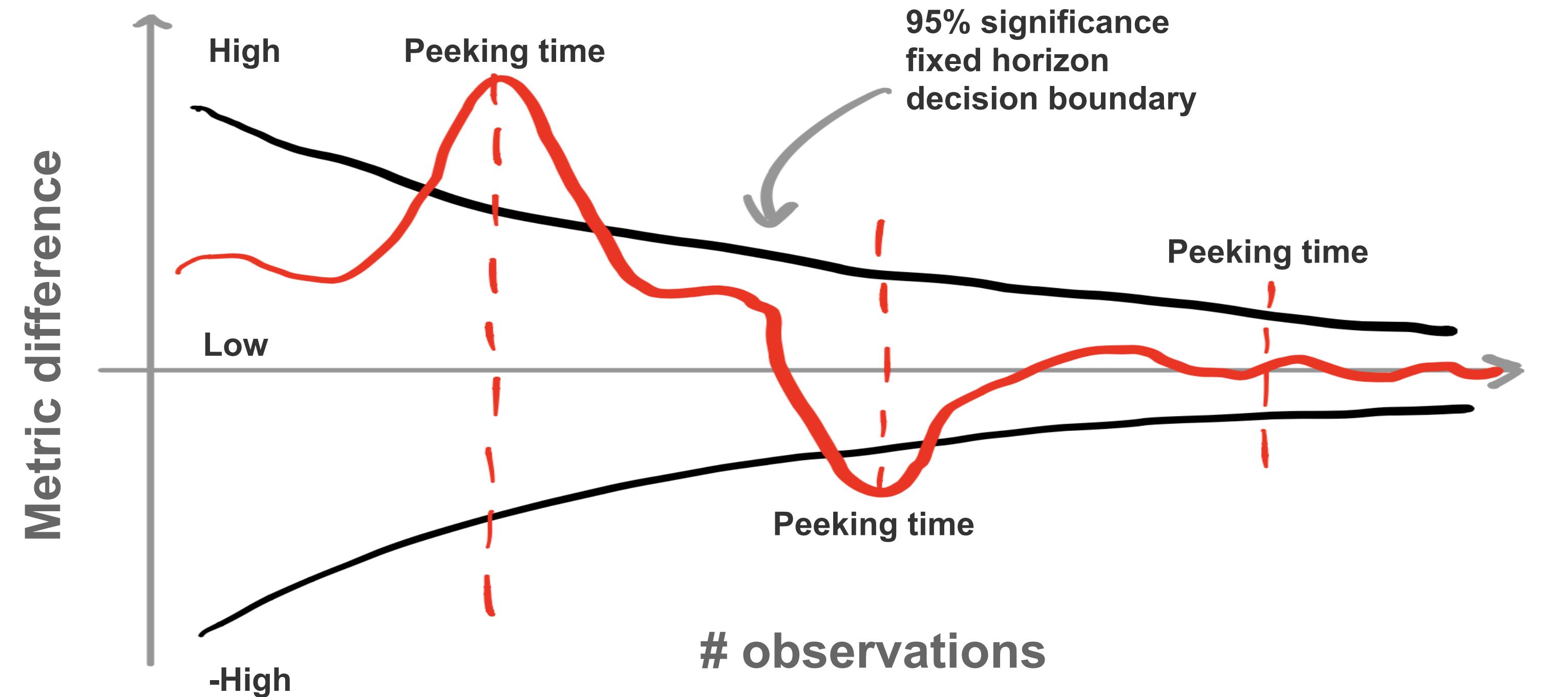
Как выглядит процесс работы с такими порогами



Строго говоря, пороги задаваемые такой параметризацией пропорциональны  $Const * \sqrt{\frac{\log n}{n}}$ .

# Decision boundaries

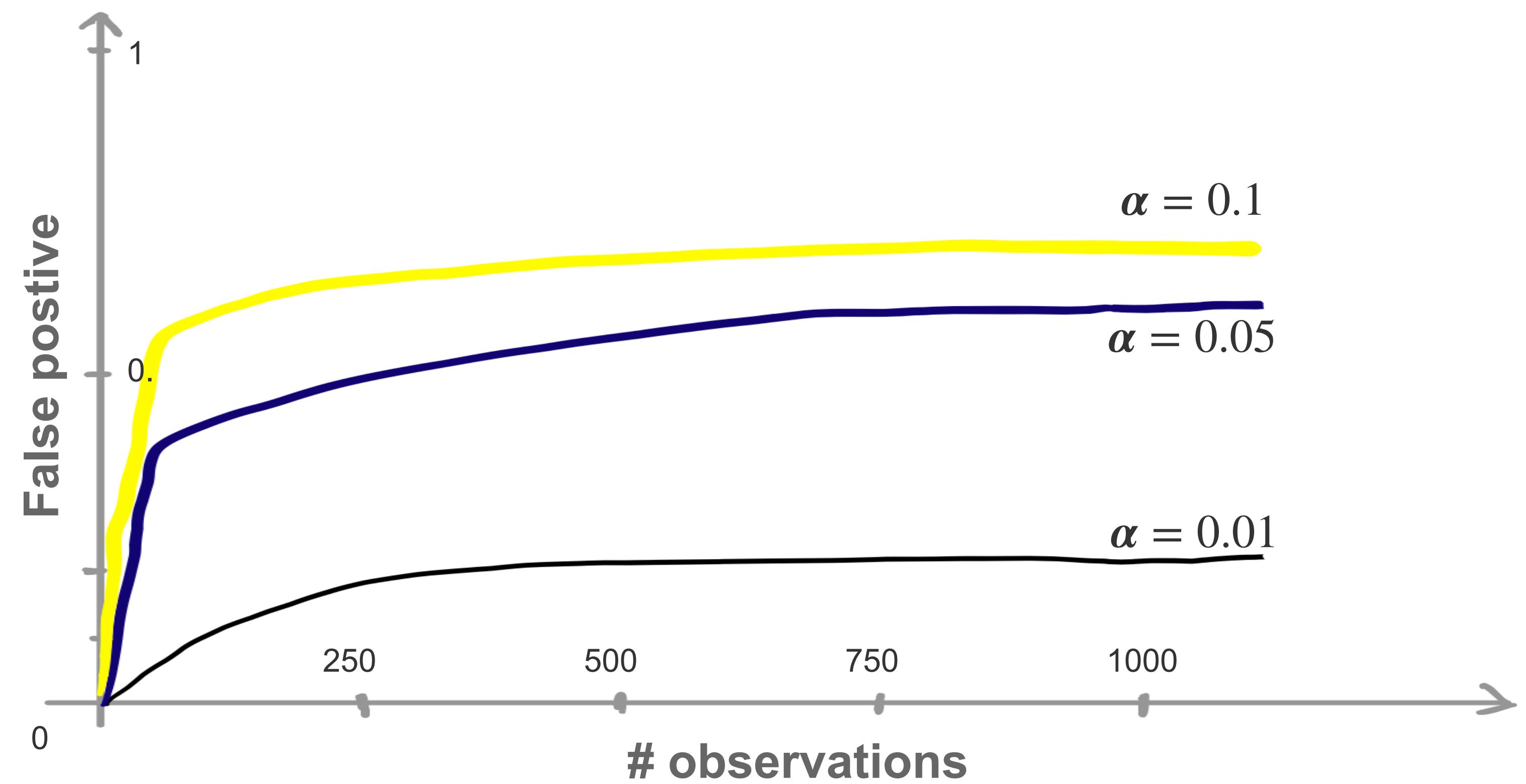
Как выглядит процесс работы с такими порогами



Данные границы в обычном случае выбираются так, чтобы при достижении определенного размера выборки давать False Positive результат с определенной вероятностью

# Но если смотреть постоянно

То мы начинаем реагировать на случайные пересечения этих границ и принимать неверные решения. Как следствие, мы перестаем контролировать вероятность ошибки 1-ого рода



# **Откуда взялось последовательное тестирование?**

## **Вопрос аудитории**

# Откуда взялось последовательное тестирование?

У ВМФ есть две альтернативные конструкции снаряда (скажем, А и Б). Они хотят определить, какая из них лучше. Для этого они совершают серию парных выстрелов. В каждом раунде они присваивают значение 1 или 0 для более производительной группы. Собственно, у ВМФ возник вопрос: "А как интерпретировать результаты?"

Абрахам Вальд нашел ответ на это так, чтобы после каждого раунда выстрелов можно было проверять результаты.



# Sequential Probability Ratio Test

**Формулируем гипотезу:**

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

**В основе лежит тест отношения правдоподобия:**

$$\Lambda_k := \prod_{i=1}^k \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}, \quad k = 1, 2, \dots$$

# Sequential Probability Ratio Test

В основе лежит тест отношения правдоподобия:

$$\Lambda_k := \prod_{i=1}^k \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}, \quad k = 1, 2, \dots$$

Наша задача найти такие пороги  $\gamma_0$  и  $\gamma_1$ , так чтобы при получении нового наблюдения  $k$ :

- Если  $\Lambda_k \geq \gamma_1$ , то мы отвергаем  $H_0$  и останавливаемся
- Если  $\Lambda_k \leq \gamma_0$ , то мы принимаем  $H_0$  и останавливаемся

При этом эти пороги должны быть установлены так, чтобы контролировать ошибку первого рода и мощность

# Sequential Probability Ratio Test

**Пусть у нас есть некоторые набор наблюдение  $x = (x_1, \dots, x_k)$  и область, где мы отвергаем нулевую гипотезу  $R_1 = \{x : \Lambda_k \geq \gamma_1\}$ , тогда мы можем записать мощность как:**

$$1 - \beta = \int_{R_1} p_{\theta_1}(x)dx = \int_{R_1} \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} p_{\theta_0}(x)dx = \int_{R_1} \Lambda_k p_{\theta_0}(x)dx \geq \gamma_1 \int_{R_1} p_{\theta_0}(x)dx = \gamma_1 \alpha,$$

**Теперь распишем вероятность ошибки первого рода определив  $R_0$  как  $R_0 = \{x : \Lambda_k \leq \gamma_0\}$ :**

$$1 - \alpha = 1 - \int_{R_1} p_{\theta_0}(x)dx = \int_{R_0} \frac{p_{\theta_0}(x)}{p_{\theta_1}(x)} p_{\theta_1}(x)dx = \int_{R_0} \Lambda_k^{-1} p_{\theta_1}(x)dx \geq \gamma_0^{-1} \int_{R_0} p_{\theta_1}(x)dx = \gamma_0^{-1} \beta.$$

# Sequential Probability Ratio Test

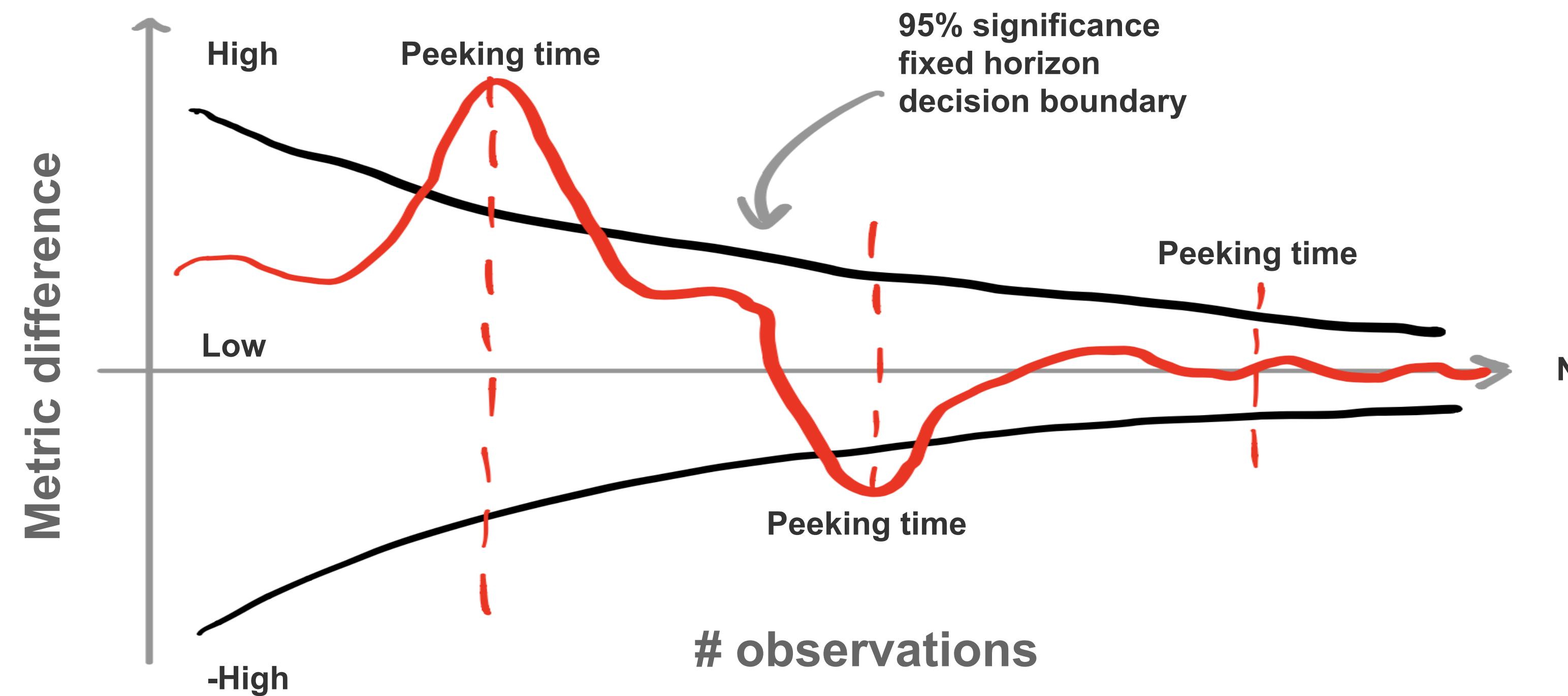
**Тогда решив систему мы можем записать:**

$$\gamma_1 \leq \frac{1 - \beta}{\alpha}$$

$$\gamma_0 \geq \frac{\beta}{1 - \alpha}$$

# Decision boundaries

Как бы выглядело для SPRT? Вопрос аудитории



# **Какая проблема применять SPRT на практике?**

## **Вопрос аудитории**

# mSPRT (mixture Sequential Probability Ratio test)

У нас есть проблема с формулировкой альтернативной гипотезы, поэтому нам тяжело использовать SPRT. Мы можем посмотреть на другую статистику, которая бы учитывала в себе разновероятные альтернативные гипотезы:

$$\Lambda_n^H = \int_{\Theta} \left( \frac{f_{\theta}(\theta_n)}{f_{\theta_0}(\theta_n)} \right)^n h(\theta) d\theta$$

где  $\Lambda_n^H$  смесь likelihood ratio взвешенная на априорную вероятность  $h(\theta)$  эффекта теста.

# mSPRT (mixture Sequential Probability Ratio test)

$$\Lambda_n^H = \int_{\Theta} \left( \frac{f_{\theta}(\theta_n)}{f_{\theta_0}(\theta_n)} \right)^n h(\theta) d\theta$$

Имея такую статистику можно определить p-value:

$$p_n = \inf\{\alpha : T(\alpha) \leq n, \delta(\alpha) = 1\},$$

где:

- $T(\alpha)$  - размер выборки при остановке теста.
- $\delta(\alpha)$  - индикатор того,  $H_0$  отвергли

В нашем случае они рассчитываются так:

$$T^H(\alpha) = \inf\{n : \Lambda_n^H \geq \alpha^{-1}\}$$
$$\delta^H(\alpha) = \mathbf{1}\{T^H(\alpha) < \infty\}$$

# Алгоритм применения mSPRT

1. Статистика  $\Lambda_n^H$  пересчитывается каждый раз, когда мы получаем новое наблюдение.  $\Lambda_n^H$  представляет собой evidence против  $H_0$  в пользу смеси альтернативных гипотез.
2. Чтобы выбрать  $f_\theta$  и  $h(\theta)$  - мы используем исторические данные
3. Наша статистика представляет собой upgrade likelihood probability ratio (LPR) теста с помощью “байесовского мышления”: мы добавили априорное распределение  $h(\theta)$  на наш эффект и смотрим теперь на смесь распределений
4. На основе нашего определения мы можем посчитать p-value
5. Если оно меньше 0.05 - останавливаем тест

**Как думаете, может ли тест идти бесконечно?**

**Вопрос аудитории**

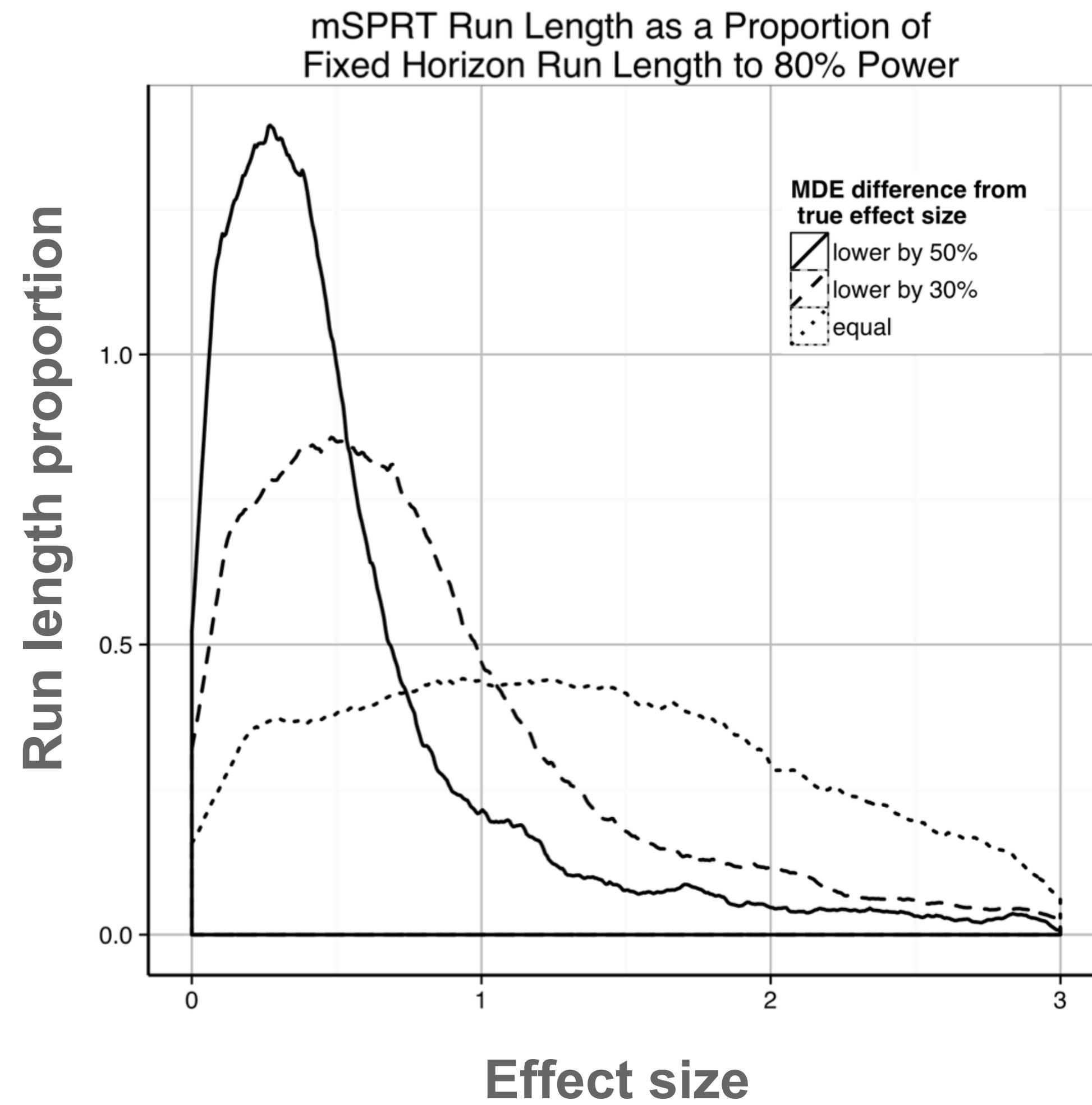
# Алгоритм применения mSPRT

**Нужно не забыть добавить максимальное время остановки**

1. Статистика  $\Lambda_n^H$  пересчитывается каждый раз, когда мы получаем новое наблюдение.  $\Lambda_n^H$  представляет собой evidence против  $H_0$  в пользу смеси альтернативных гипотез.
2. Чтобы выбрать  $f_\theta$  и  $h(\theta)$  - мы используем исторические данные
3. Наша статистика представляет собой upgrade likelihood probability ratio (LPR) теста с помощью “байесовского мышления”: мы добавили априорное распределение  $h(\theta)$  на наш эффект и смотрим теперь на смесь распределений
4. На основе нашего определения мы можем посчитать p-value
5. Если оно меньше 0.05 - останавливаем тест
6. Если тест идет слишком долго - просто принимаем  $H_0$

# mSPRT: empirical results

**Результаты показывают, что при правильной оценке эффекта mSPRT может работать  
дольше, но давать возможность подглядывать. Но дьявол в деталях...**



# Итого

1. Узнали, почему появляются продвинутые методы в А/Б тестах
2. Поняли какие проблемы доставляет Даниил Лизе
3. Разобрались с одним из методов уменьшения дисперсии
4. И поняли как давать менеджерам подглядывать А/В теста