

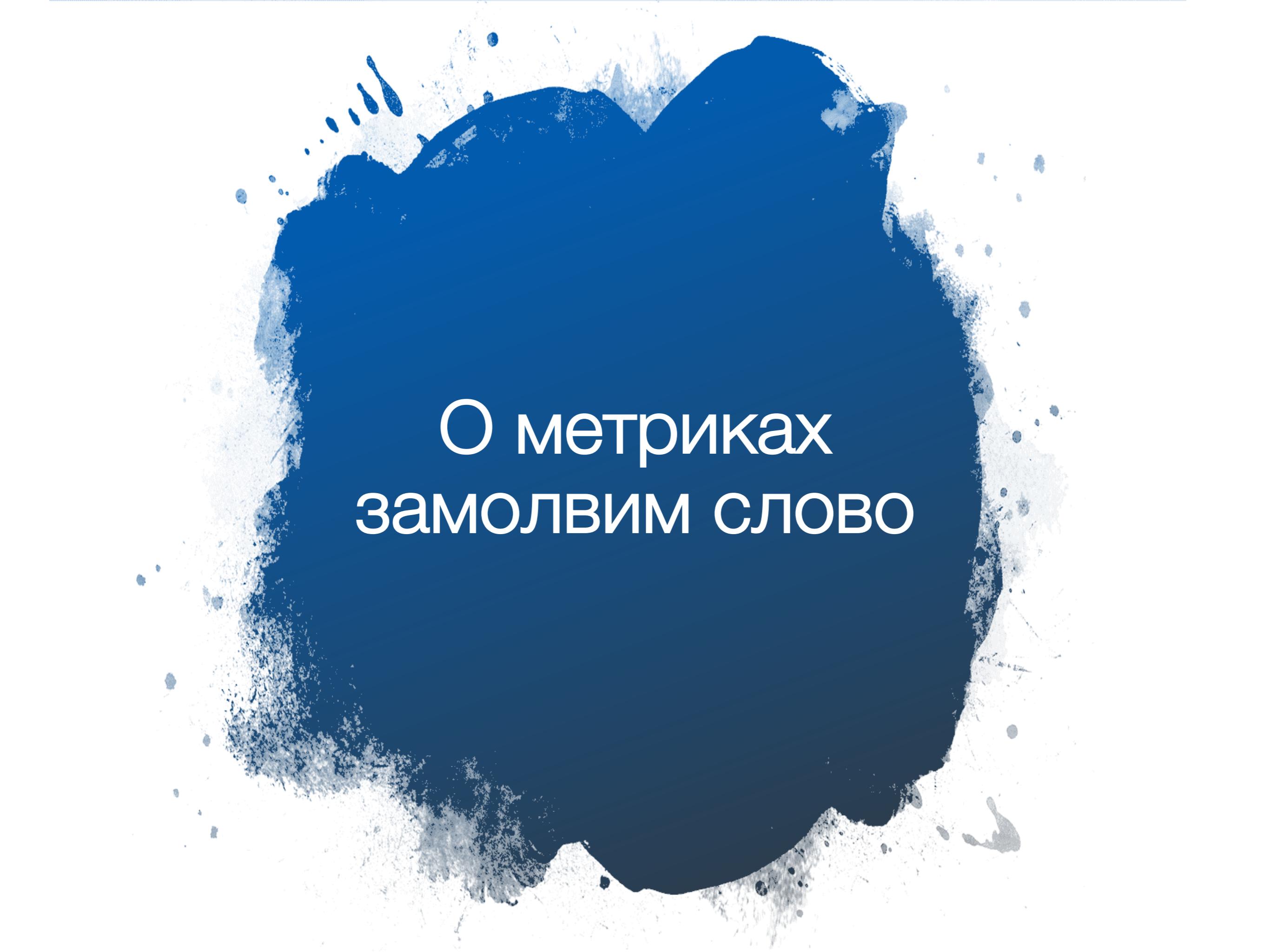


НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

# A/B-ТЕСТИРОВАНИЕ

Теванян Элен  
07.06.2019

Москва 2019



# О метриках замолвим слово

# МЕТРИКИ, КОТОРЫЕ МЫ ЗНАЕМ

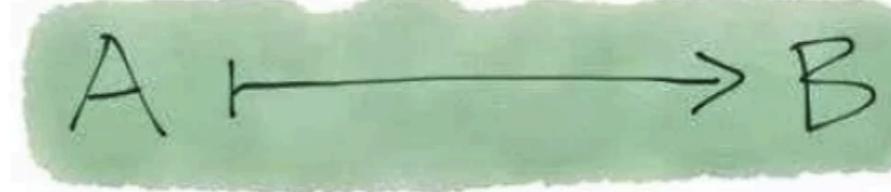
---

Метрики качества моделей машинного обучения:

- Регрессия:
  - ✓ MSE, RMSE
  - ✓ MAE
  - ✓ MAPE
  - ✓ R<sup>2</sup>
- Классификация
  - ✓ Accuracy
  - ✓ Recall
  - ✓ Precision
  - ✓ ROC-AUC

# ЧТО ВАЖНО ЗАКАЗЧИКУ?

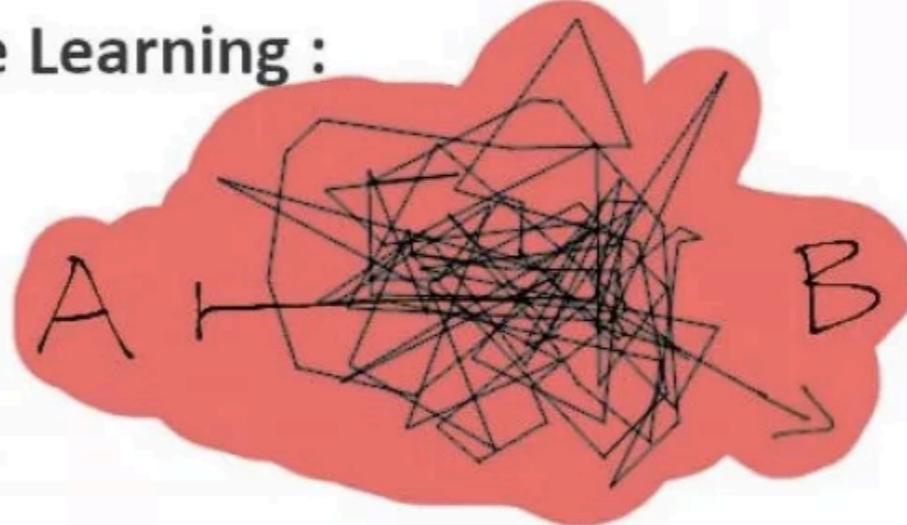
Theory:



Practice:



Machine Learning :



# БИЗНЕС ДУМАЕТ О ДРУГИХ МЕТРИКАХ

---

В целом, всегда о деньгах ☺



# ВОПРОСЫ, НА КОТОРЫЕ ИЩЕМ ОТВЕТЫ

---

- На сколько возрастает средний чек?
- На сколько процентов уменьшается текучка в организации?
- На сколько часов увеличивается среднее время безотказной работы сервиса?
- На сколько рублей увеличивается LTV (Lifetime Value) клиента?

# МЕТРИКИ БЫВАЮТ РАЗНЫЕ

---

- Бизнес-метрики
  - Метрики, понятные бизнесу. Например: ROI. Сложно измерить изменение.
- Онлайн-метрики
  - Метрики, измеряемые во время пилотирования, напрямую связаны с бизнес метrikами или влияют на них (прокси-метрики). Например: CTR. Вычисляются ~100 раз.
- Оффлайн-метрики
  - Должны коррелировать с онлайн-метриками, позволяют сравнивать модели до пилотирования (до онлайн теста). Например: AUC. Вычисляются ~1000 раз.

# МОРАЛЬ

---

- Модель оттока по ROC-AUC может быть и 0.99, но еще не факт, что она окажется полезной.
- Ее польза – это гипотеза, которую надо тестировать.

# Гипотезы и ошибки

# ВИДЫ ГИПОТЕЗ В СТАТИСТИКЕ (1/3)

---

- $H_0$  – нулевая гипотеза
- Это утверждение, веру в которое мы проверяем
- Суд: если присяжные выносят обвинительный вердикт в условии презумпции невиновности, подсудимый считается виновным:
  - $H_0$ : подсудимый не виновен
- Бизнес: новый график работы сотрудников увеличивает их продуктивность
  - $H_0$ : срок выдачи результатов не изменился

# ВИДЫ ГИПОТЕЗ В СТАТИСТИКЕ (2/3)

---

- $H_A$  – альтернативная гипотеза
- Это утверждение, в корректность которого мы надеемся 😊
- Суд: если присяжные выносят обвинительный вердикт в условии презумпции невиновности, подсудимый считается виновным:
  - $H_A$ : подсудимый виновен
  - Бизнес: новый график работы сотрудников увеличивает их продуктивность
  - $H_A$ : срок выдачи результатов сократился

# ВИДЫ ГИПОТЕЗ В СТАТИСТИКЕ (3/3)

---

- $H_A$  – альтернативная гипотеза
- Альтернативы бывают, как правило, трех видов:
  - $\neq$
  - $>$
  - $<$

# ОШИБКИ

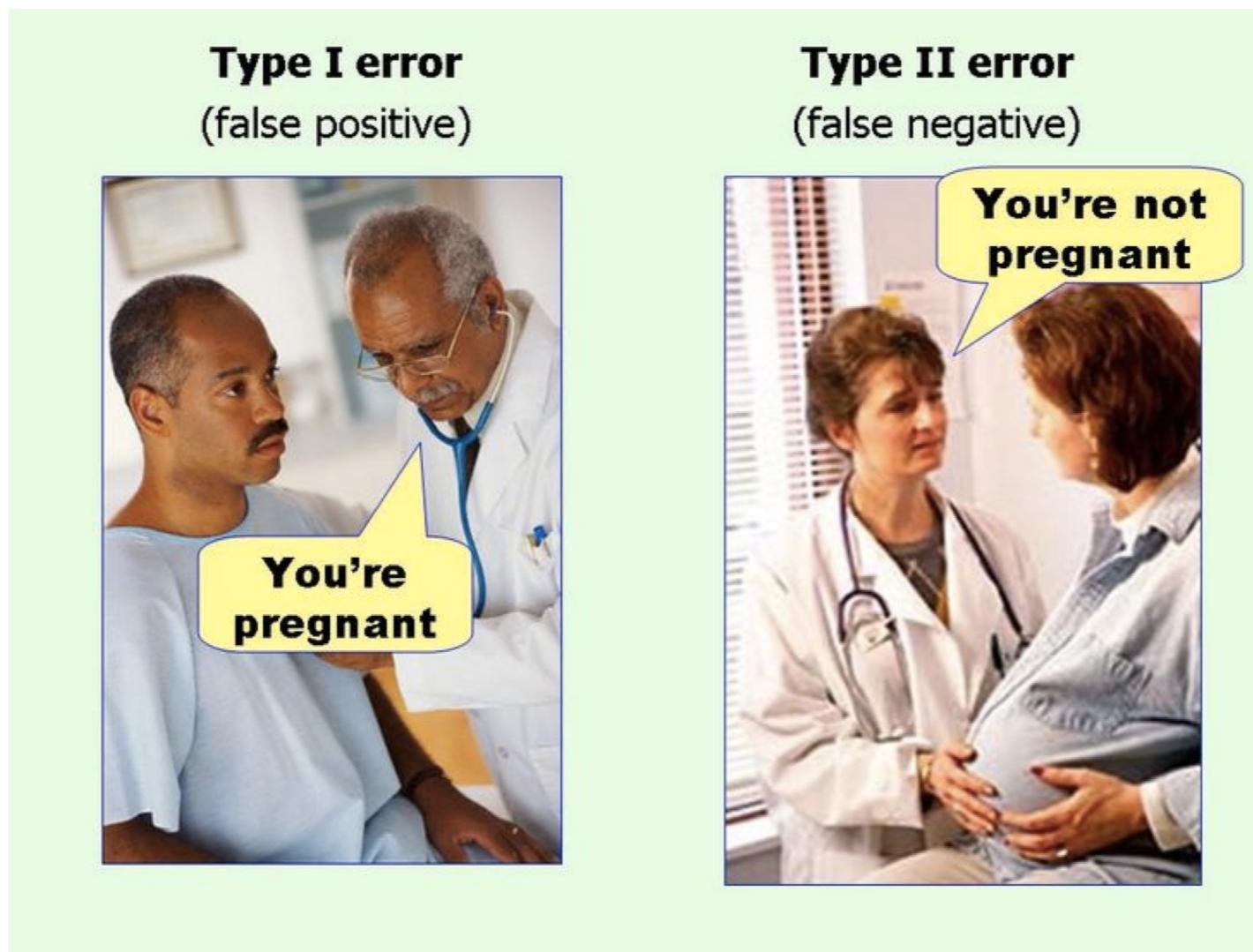
---

- $H_0$  – нулевая гипотеза
- $H_A$  – альтернативная гипотеза
- Ошибка первого рода ( $\alpha$ ) – вероятность отвергнуть нулевую гипотезу, когда она верна
- Ошибка второго рода ( $\beta$ ) – вероятность принять нулевую гипотезу, когда она неверна.

# ОШИБКИ

---

- $H_0$  – пациент не в положении
- $H_A$  – пациент в положении



# ТЕРМИНОЛОГИЯ ПРО ОШИБКИ

---

- $\alpha$  – уровень значимости – доля ошибок первого рода, которые мы готовы принять (увидели эффект, которого на самом деле нет)
- $\beta$  – ошибка второго рода, которые мы готовы принять (говорим, что эффекта нет, а он был)
- $1 - \alpha$  – уровень доверия (специфичность теста)
- $1 - \beta$  – мощность (чувствительность теста)

# WARNING

---

Мы не доказываем корректность  $H_0$

Мы либо находим достаточные основания, чтобы поверить ей (не отвергнуть) или отвергнуть в пользу веры в  $H_A$

# Тестирование гипотез

# ЦЕЛЬ

---

По маленькой выборке сделать выводы о всей совокупности

- Оценка величины, а также уверенность в этой оценке
  - пример: среднее и 95% доверительный интервал для него
- Статистическая проверка гипотез
  - пример: равенство средних значений в двух выборках
- Расчет размера выборки
  - пример: каков должен быть размер выборок, чтобы обнаружить различие в средних в 0.05% с уровнем надежности 95%?

# ВЗБОЛТАТЬ, НО НЕ СМЕШИВАТЬ

---

- Джеймс Бонд говорит, что предпочитает мартини взболтаным, но не смешанным
- Разбирается ли Джеймс Бонд в мартини или выбирает наугад?
- $H_0$ : Джеймс Бонд выбирает наугад
- $H_A$ : Джеймс Бонд выбирает не наугад



# ВЗБОЛТАТЬ, НО НЕ СМЕШИВАТЬ

---

- $H_0$ : Джеймс Бонд выбирает наугад
- $H_A$ : Джеймс Бонд выбирает не наугад
- Слепой тест: предложим ему  $n$  раз пару напитков и спросим, какой из двух он предпочитает

# ВЗБОЛТАТЬ, НО НЕ СМЕШИВАТЬ

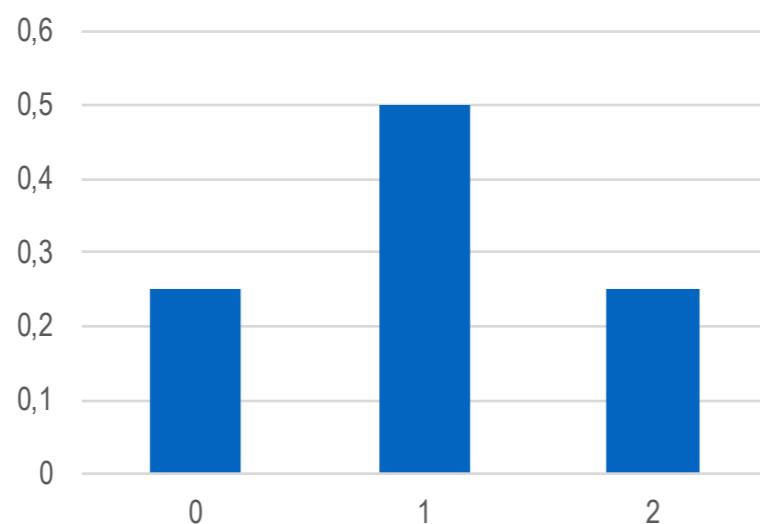
---

- $H_0$ : Джеймс Бонд выбирает наугад
- $H_A$ : Джеймс Бонд выбирает не наугад
- Слепой тест: предложим ему  $n$  раз пару напитков и спросим, какой из двух он предпочитает
- Выборка: вектор длины  $n$  из нулей и единиц (предпочёл взболтанный — 1, смешанный — 0)
- Статистика: число единиц в выборке

# ВЗБОЛТАТЬ, НО НЕ СМЕШИВАТЬ

---

0	0	0
0	1	1
1	0	1
1	1	2



Распределение  
статистики

# ВЗБОЛТАТЬ, НО НЕ СМЕШИВАТЬ

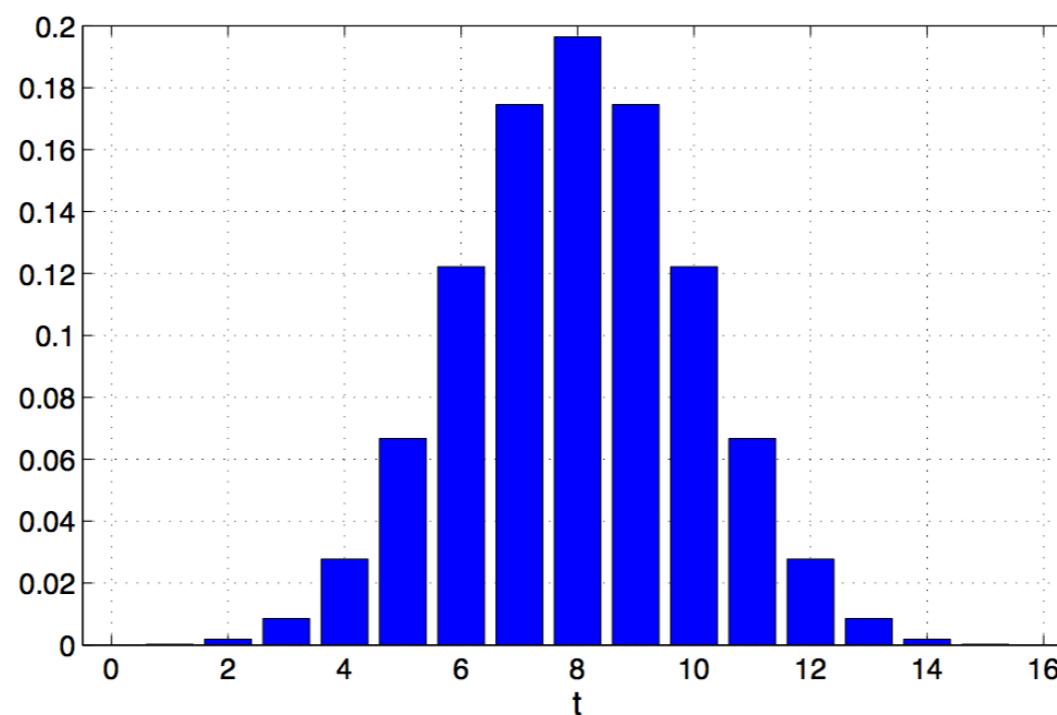
---

- Если нулевая гипотеза справедлива и Джеймс Бонд не различает напитки, то все исходы равновероятны
- Пусть  $n = 16$  — тогда существует 65536 различных бинарных векторов

# ВЗБОЛТАТЬ, НО НЕ СМЕШИВАТЬ

---

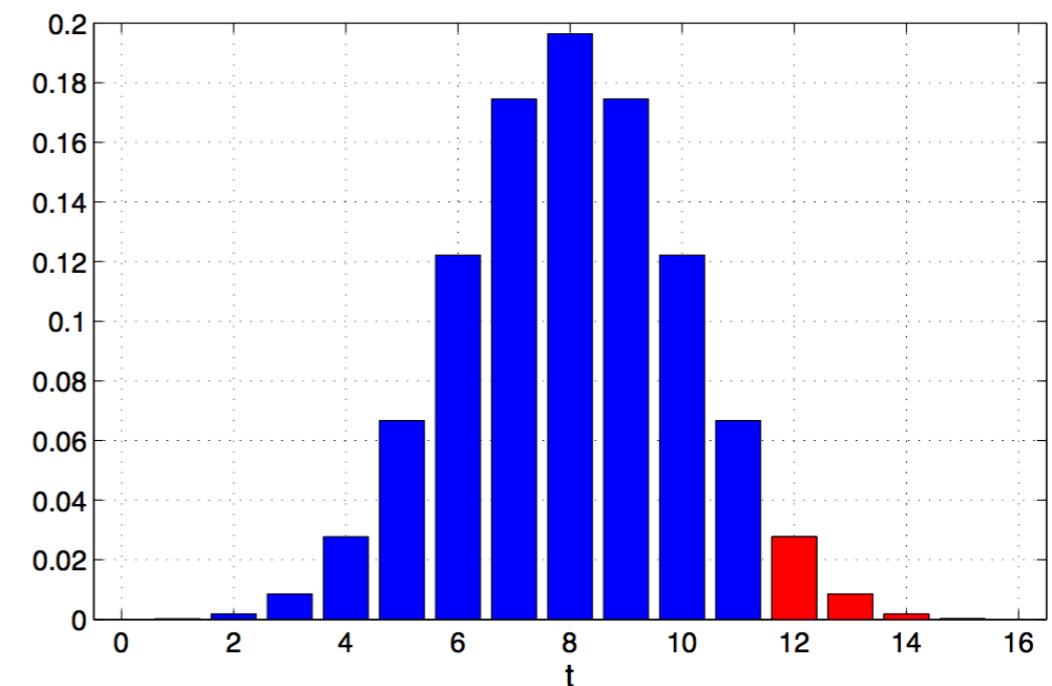
Если Бонд выбирает наугад, то вот так распределено количество «правильных» выборов:



# ВЗБОЛТАТЬ, НО НЕ СМЕШИВАТЬ

---

- Провели эксперимент — Джеймс Бонд выбрал взболтанный мартини в 12 из 16 раз
- Вероятность того, что он выберет взболтанный мартини 12 раз или больше при условии, что выбирает наугад:  $\frac{2512}{65536} \approx 0.0384$
- Отклоняем гипотезу о том, что Джеймс Бонд не разбирается в мартини



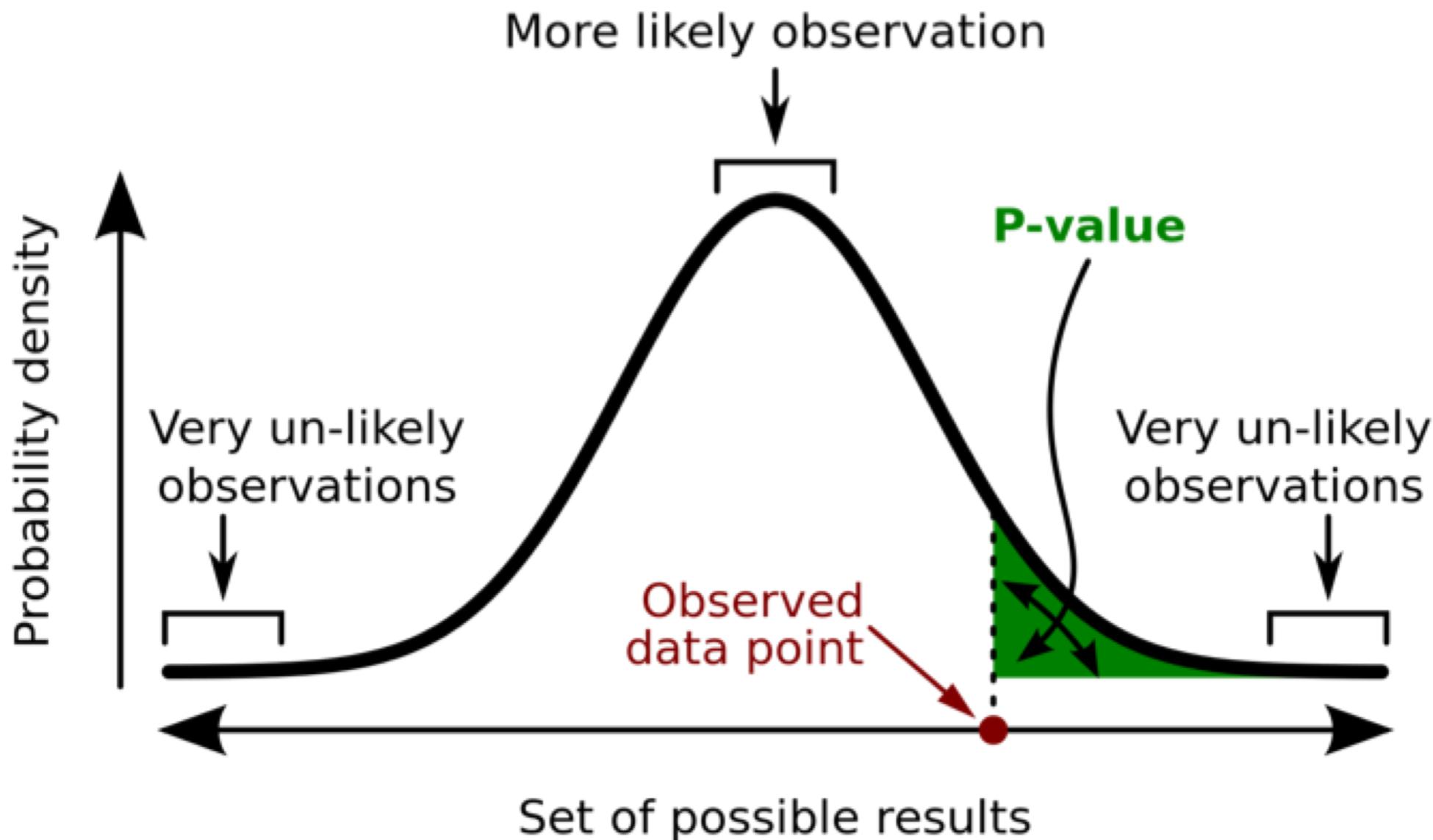
# КАК ПРОВЕРЯТЬ ГИПОТЕЗЫ

---

- Выдвинуть нулевую и альтернативные гипотезы
- В предположении о корректности нулевой вывести распределение статистики
- На основе распределения статистики определяется вероятность (p-value) про наблюдать такое значение или более критическое при верной  $H_0$
- Если p-value меньше наперед заданного порога  $\alpha$  (0.01, 0.05, ...), то  $H_0$  отвергается

# КАК ПРОВЕРЯТЬ ГИПОТЕЗЫ

---



# ОШИБКИ ПЕРВОГО РОДА

---

- Утверждается, что осьминог предсказывает результаты матчей с участием сборной Германии на чемпионате мира по футболу 2010 года, выбирая кормушку с флагом страны-победителя



# ОШИБКИ ПЕРВОГО РОДА

---

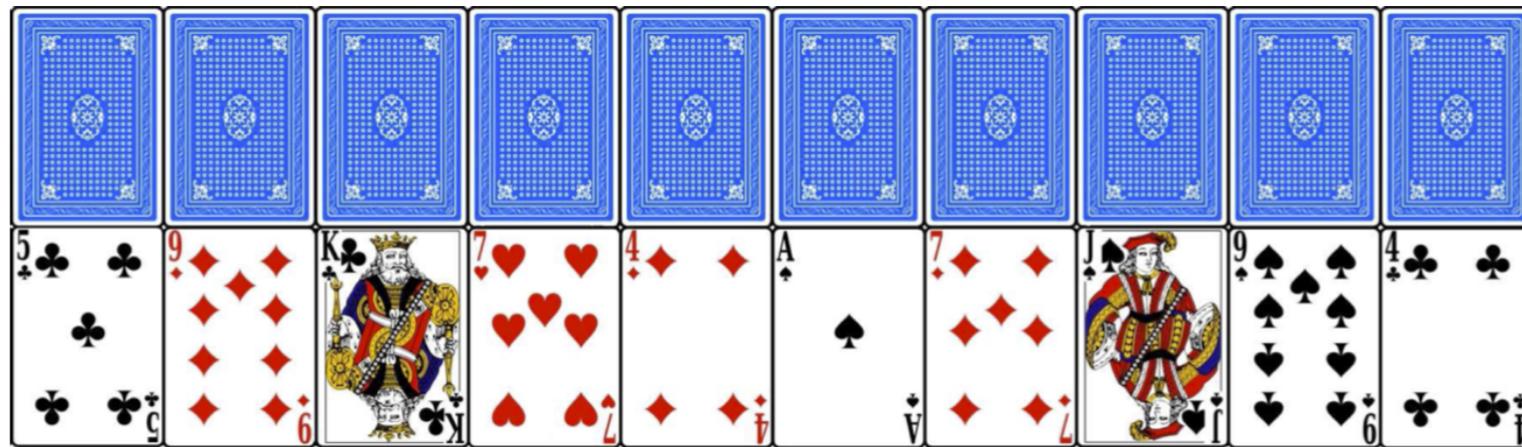
- По результатам 13 испытаний ему удается верно угадать результаты 11 матчей.
- Аналогичный предыдущему критерий даёт вероятность  $p \approx 0.0112$
- Настоящая вероятность того, что осьминог выбирает кормушку наугад, равна единице!

# МНОЖЕСТВЕННАЯ ПРОВЕРКА ГИПОТЕЗ

(Rhine, 1950): исследования возможности экстрасенсорного восприятия.

Первый этап — поиск экстрасенсов.

Испытуемому предлагается угадать цвет 10 карт.



$H_0$ : испытуемый выбирает ответ наугад.

$H_1$ : испытуемый может предсказывать цвета карт.

Статистика  $t$  — число карт, цвета которых угаданы.

$$P(t \geq 9 | H_0) = 11 \cdot \frac{1}{2}^{10} = 0.0107421875,$$

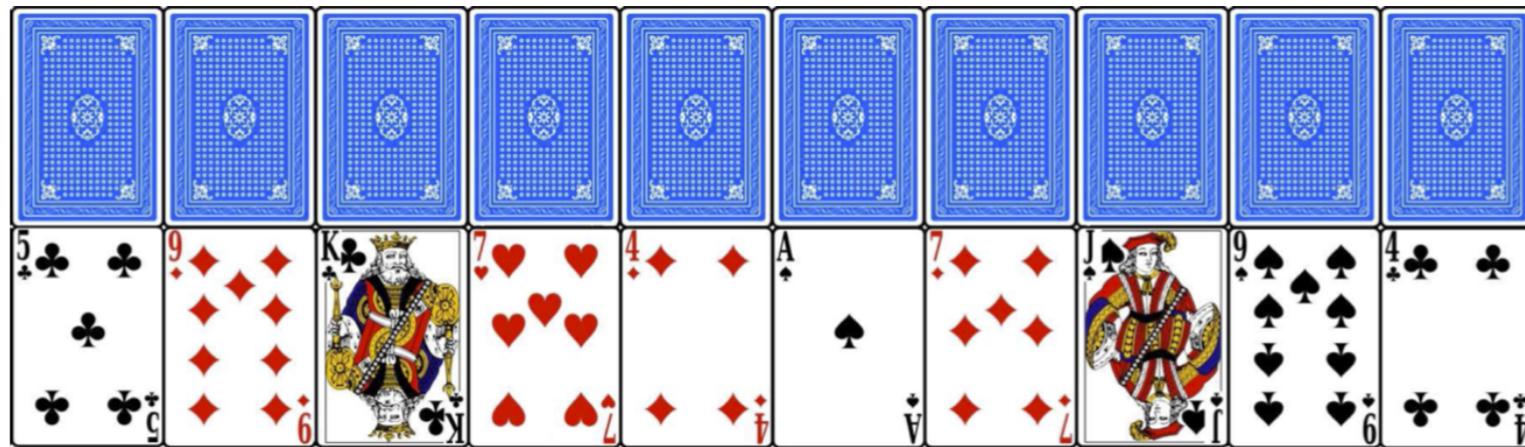
т. е. при  $t = 9$  получаем достигаемый уровень значимости  $p \approx 0.01$  — можно отклонять  $H_0$ .

# МНОЖЕСТВЕННАЯ ПРОВЕРКА ГИПОТЕЗ

(Rhine, 1950): исследования возможности экстрасенсорного восприятия.

Первый этап — поиск экстрасенсов.

Испытуемому предлагается угадать цвет 10 карт.



$H_0$ : испытуемый выбирает ответ наугад.

$H_1$ : испытуемый может предсказывать цвета карт.

Статистика  $t$  — число карт, цвета которых угаданы.

$$P(t \geq 9 | H_0) = 11 \cdot \frac{1}{2}^{10} = 0.0107421875,$$

т. е. при  $t = 9$  получаем достигаемый уровень значимости  $p \approx 0.01$  — можно отклонять  $H_0$ .

# МНОЖЕСТВЕННАЯ ПРОВЕРКА ГИПОТЕЗ

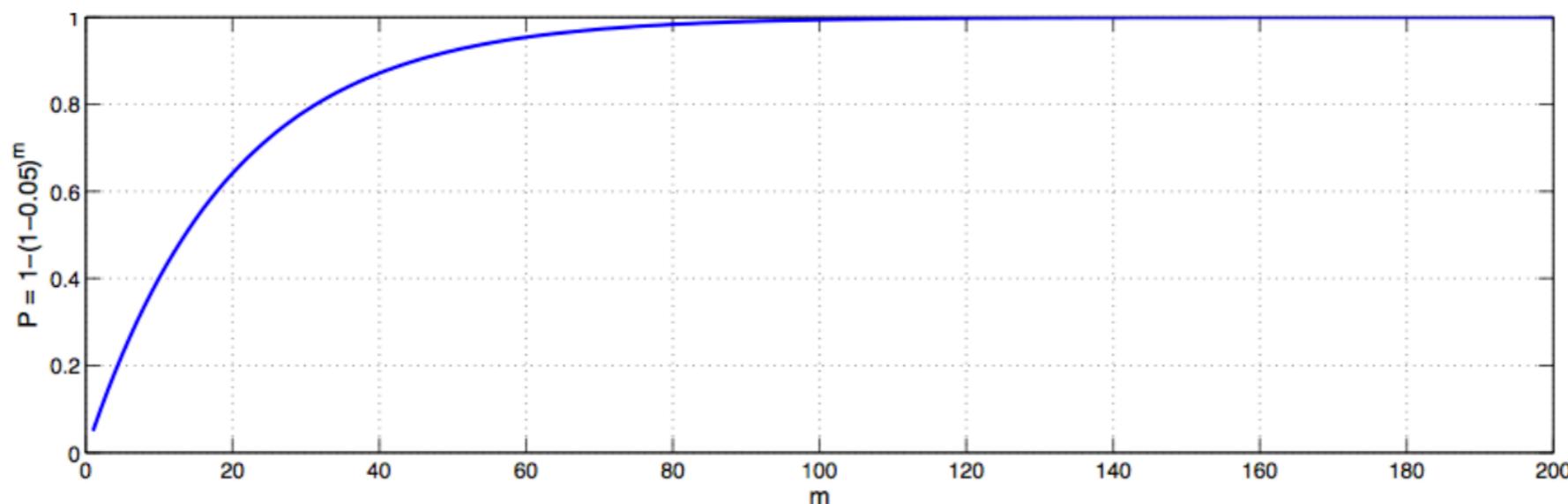
---

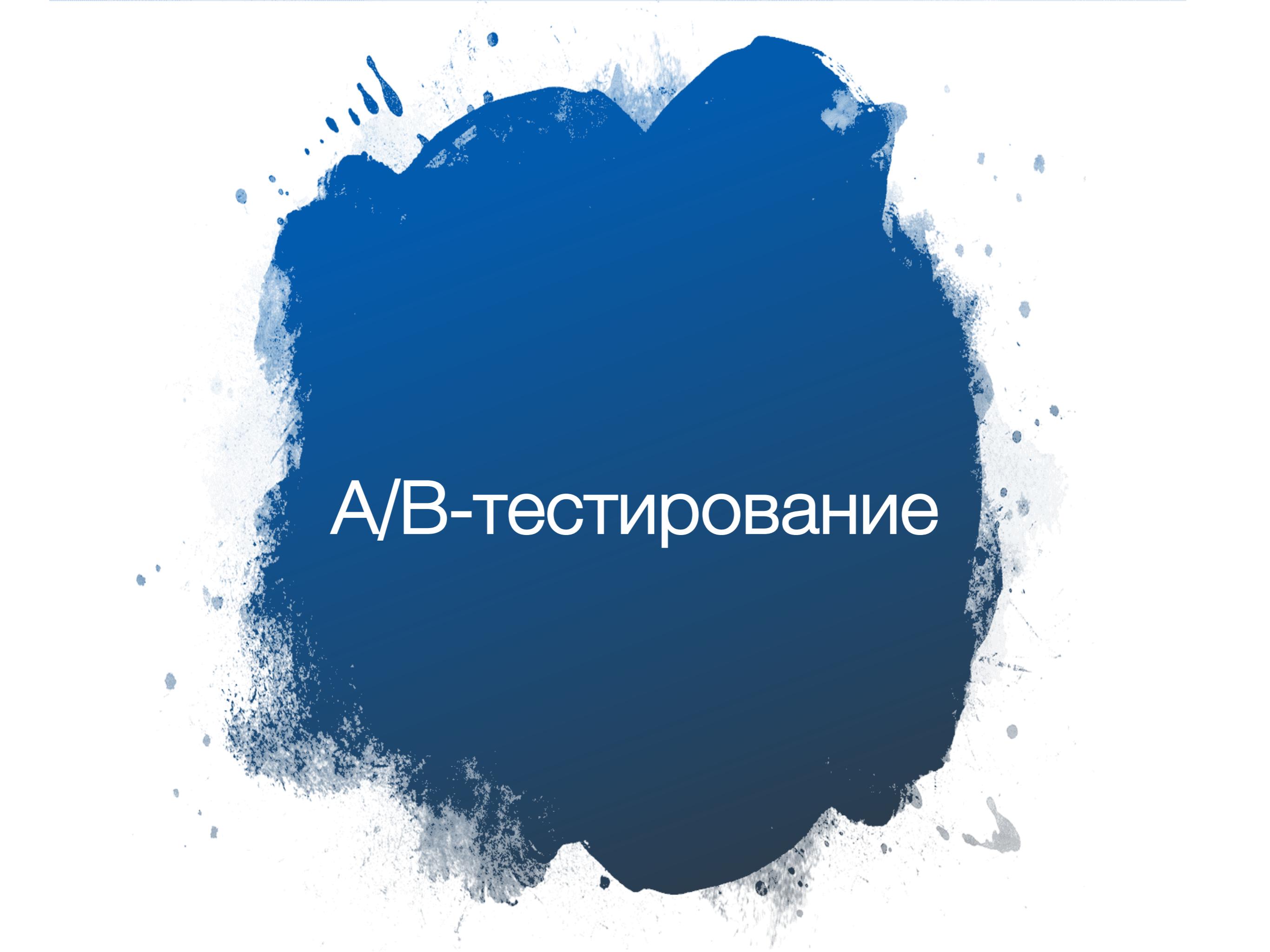
- Процедуру отбора прошли 1000 человек
- Девять из них угадали цвета 9 из 10 карт, двое — цвета всех 10 карт
- Ни один в последующих экспериментах не подтвердил своих способностей

# МНОЖЕСТВЕННАЯ ПРОВЕРКА ГИПОТЕЗ

---

Вероятность того, что из 1000 человек хотя бы один случайно угадает цвета 9 или 10 из 10 карт:  $1 - \left(1 - 11 \cdot \frac{1}{2}^{10}\right)^{1000} \approx 0.9999796.$



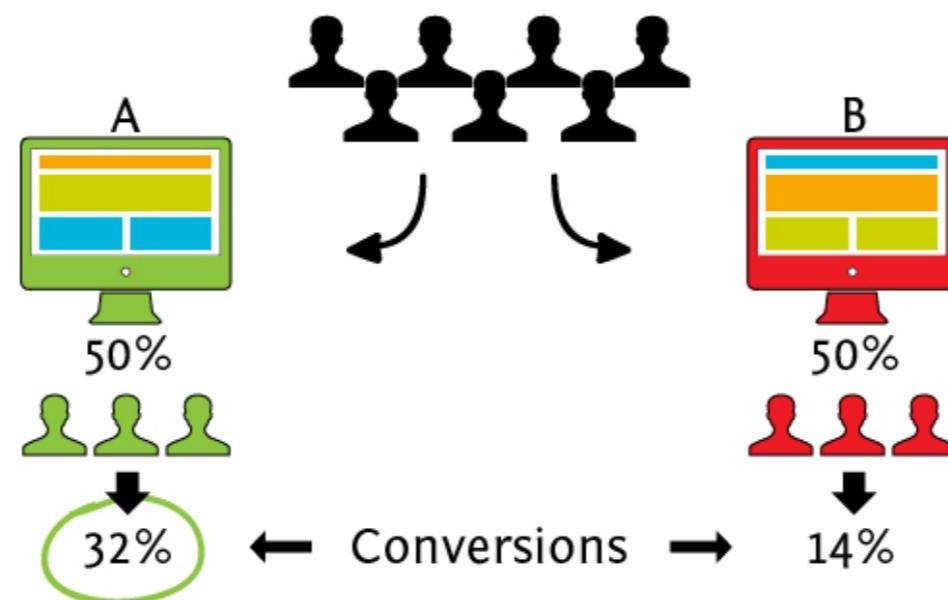


A/B-тестирование

# ПРОВЕРКА ЭФФЕКТИВНОСТИ ИЗМЕНЕНИЙ

---

- Группа А видит одну версию сайта. Группа В видит другую версию сайта (с одним изменением).
- Есть ли различие в конверсии между группами А и В?



# ПРОВЕРКА ЭФФЕКТИВНОСТИ ИЗМЕНЕНИЙ

---

Вероятностная модель:

- ✓ Версии А и В характеризуются вероятностями перехода  $p_A$  и  $p_B$
- ✓ Эти вероятности мы не знаем
- ✓ Нулевая гипотеза:  $p_A = p_B$

# A/B-ТЕСТИРОВАНИЕ

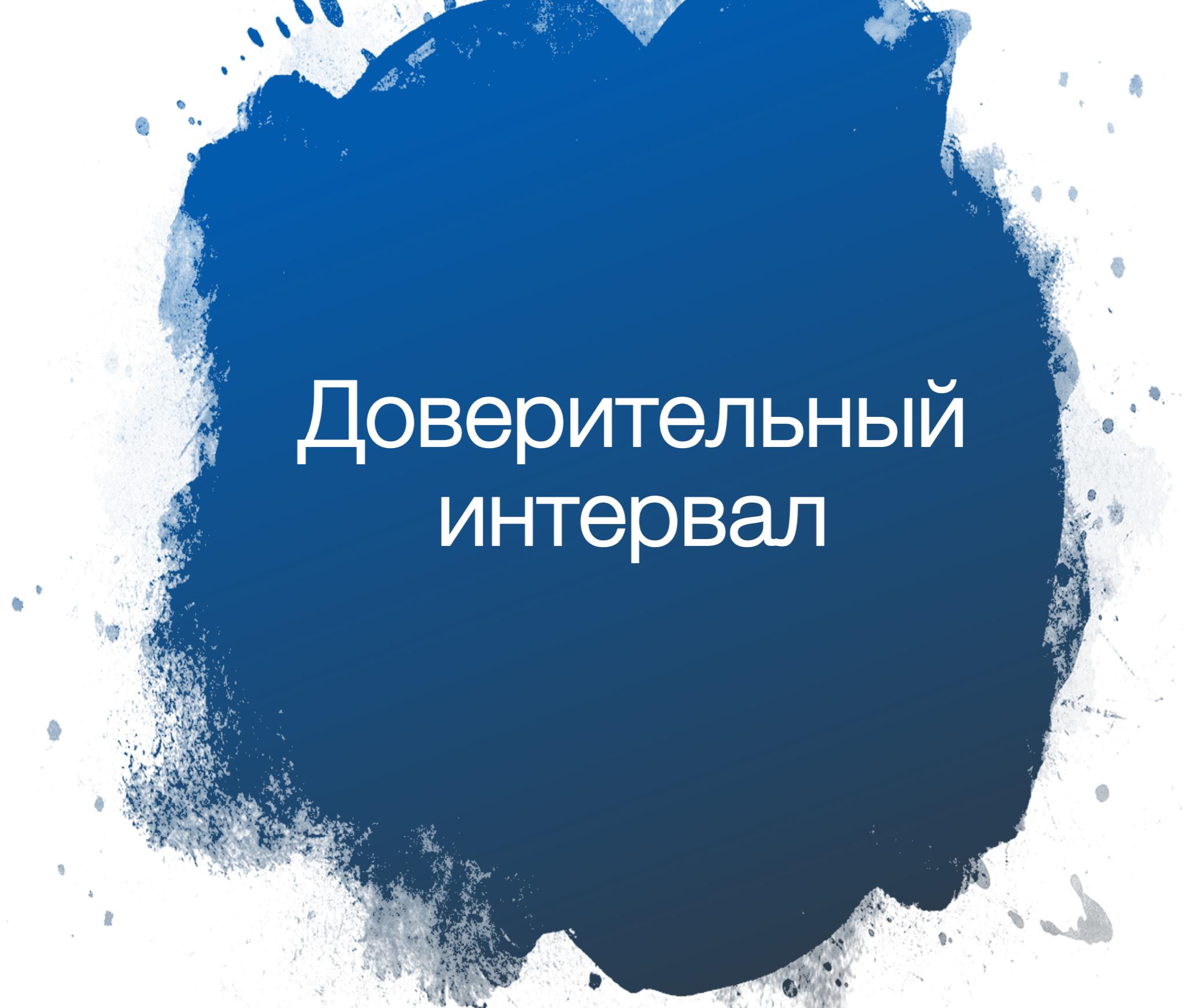
---

- Процедура, с помощью которой измеряем влияние изменений на целевой показатель
- Изменения могут произойти с чем угодно
  - Цветом кнопки на сайте
  - Порядком расположения элементов на сайте
  - Скриптом телефонного звонка
  - Внедрением нового алгоритма рекомендаций
- Группа А видит оригинальный вариант
- Группа В видит измененный вариант

# ЧТО ВАЖНО В АБ-ТЕСТАХ?

---

- Эффект, который мы хотим замерить
- Уровень значимости
- Ошибка второго рода
- Размер выборки



# Доверительный интервал

# ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ

---

- Некоторый неизвестный параметр  $a$
- Мы не знаем его настоящее значение, но у нас есть выборка  $X$
- По ней мы можем рассчитать эмпирическое значение  $\hat{a}$  (оценить его)
- И построить интервал, где с какой-то вероятностью находится настоящее значение величины  $a$

# ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ

---

- Чаще всего строят 95%-ный ДИ
- Строится по формуле:

$$a = \hat{a} \pm z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$$

где

$\hat{a}$  – эмпирическое значение параметра, посчитанное по выборке

$\hat{\sigma}$  – несмещеннное выборочное среднее

$n$  – количество элементов выборке

$z_{\frac{\alpha}{2}}$  – коэффициент погрешности (берем равным 1.96)

$\alpha$  – уровень значимости (равный 5%)



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ