

Ошибки, что мы совершаем

| | H_0 верна | H_0 неверна | |
|----------------------|-------------|---------------|------------------|
| H_0 не отвергается | <i>ok</i> | β | ошибка 2 рода |
| H_0 отвергается | α | <i>ok</i> | |

ошибка 1
рода

$$\alpha = \mathbb{P}(H_0 \text{ отвергнута} \mid H_0 \text{ верна})$$

$$\beta = \mathbb{P}(H_0 \text{ не отвергнута} \mid H_0 \text{ не верна})$$

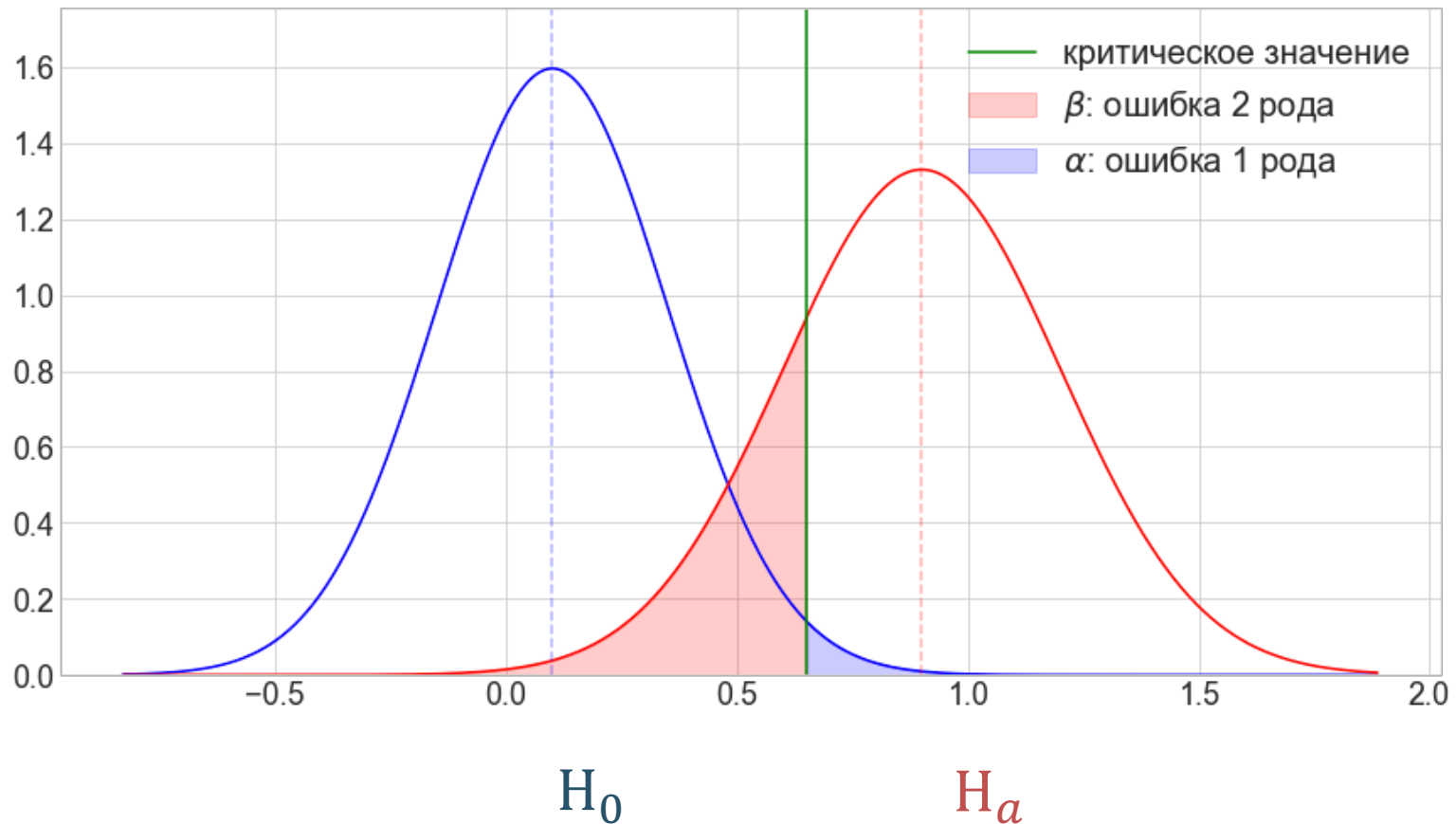
Величину $1 - \beta$ называют **мощностью** критерия

Ошибки, что мы совершаем

$$H_0: p = p_0$$

$$H_a: p = p_a$$

Ошибки первого и второго рода неравнозначны:
мы перед экспериментом фиксируем α ,
а β минимизируется по остаточному принципу

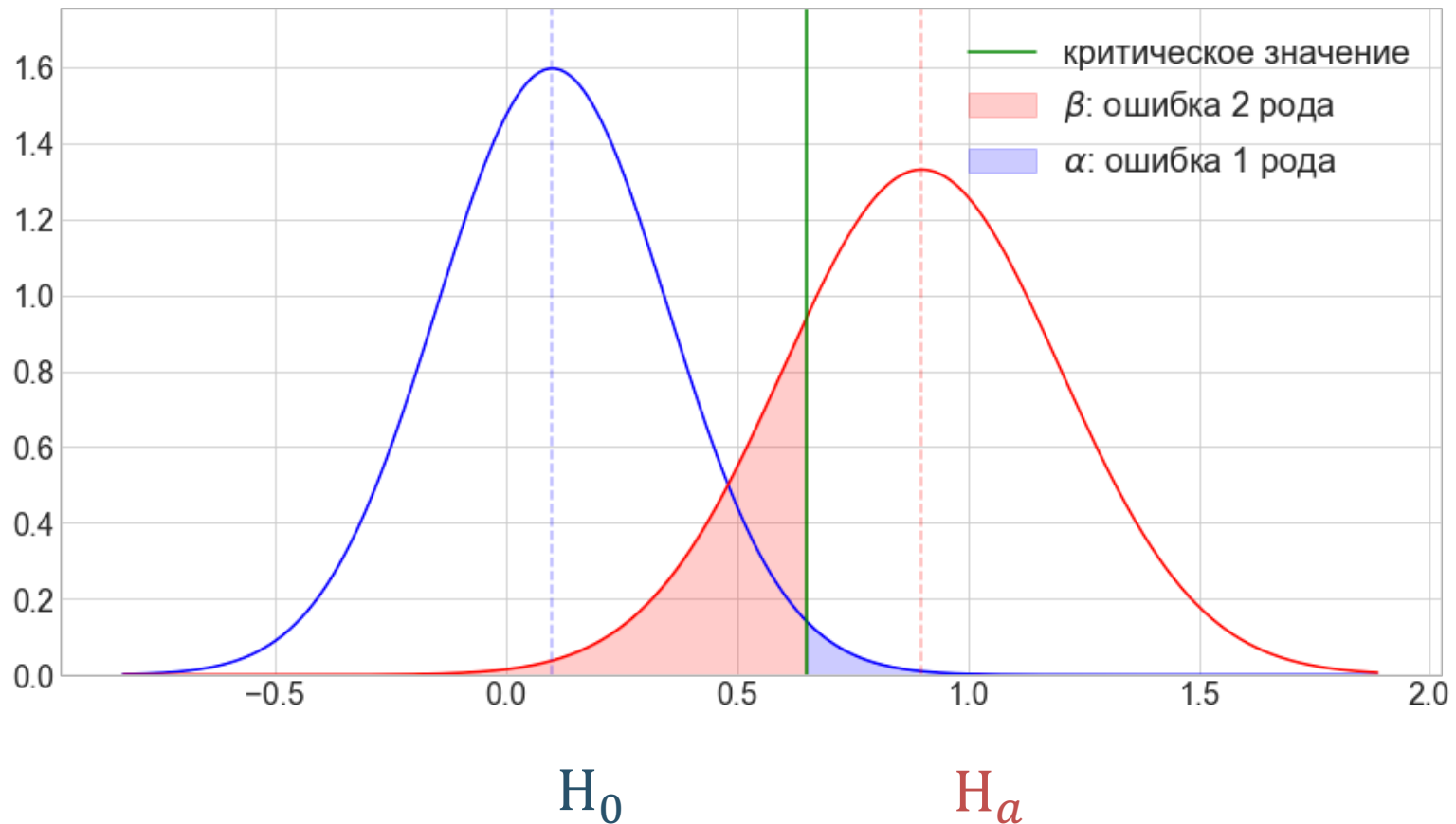


Ошибки, что мы совершаем

$$H_0: p = p_0$$

$$H_a: p = p_a$$

❗ При уменьшении ошибки первого рода всегда возрастает ошибка второго рода










Ошибки, что мы совершаем

H_0 : нет беременности

H_a : есть беременность

H_0 верна

H_0 неверна

| | | |
|-------------------------|--|--|
| H_0 не отвергается | <div></div> <p>Вы не беременны</p> | <div></div> <p>Вы не беременны</p> |
| H_0 отвергается | <div></div> <p>Вы беременны</p> | <div></div> <p>Вы беременны</p> |

Аналогия с классификацией

| | $y = 1$ | $y = 0$ |
|---------------|---------|---------|
| $\hat{y} = 1$ | TP | FP |
| $\hat{y} = 0$ | FN | TN |

ошибка 2
рода

ошибка 1
рода

Пример: хотим, чтобы классификатор удалял спам и задел минимум хороших документов

Подбор порога: зафиксировать $FPR = \frac{FP}{FP+TN} \leq 0.05$

(доля зря удалённых), а дальше максимизировать полноту

$$Recall = TPR = \frac{TP}{TP+FN} = 1 - \frac{FN}{TP+FN}$$

Наиболее мощный критерий

- **Статистический критерий** – способ посчитать расстояние между наблюдаемым значением и предполагаемым
- Подобные расстояния можно считать разными способами
- Хочется выбрать такой способ, который при фиксированном размере выборки и фиксированной ошибке первого рода будет давать наименьшую ошибку второго рода
- Такой критерий называется **наиболее мощным**

Сколько надо наблюдений

Ошибки, что мы совершаем

| | H_0 верна | H_0 неверна | |
|----------------------|-------------|---------------|------------------|
| H_0 не отвергается | <i>ok</i> | β | ошибка 2 рода |
| H_0 отвергается | α | <i>ok</i> | |

ошибка
1 рода

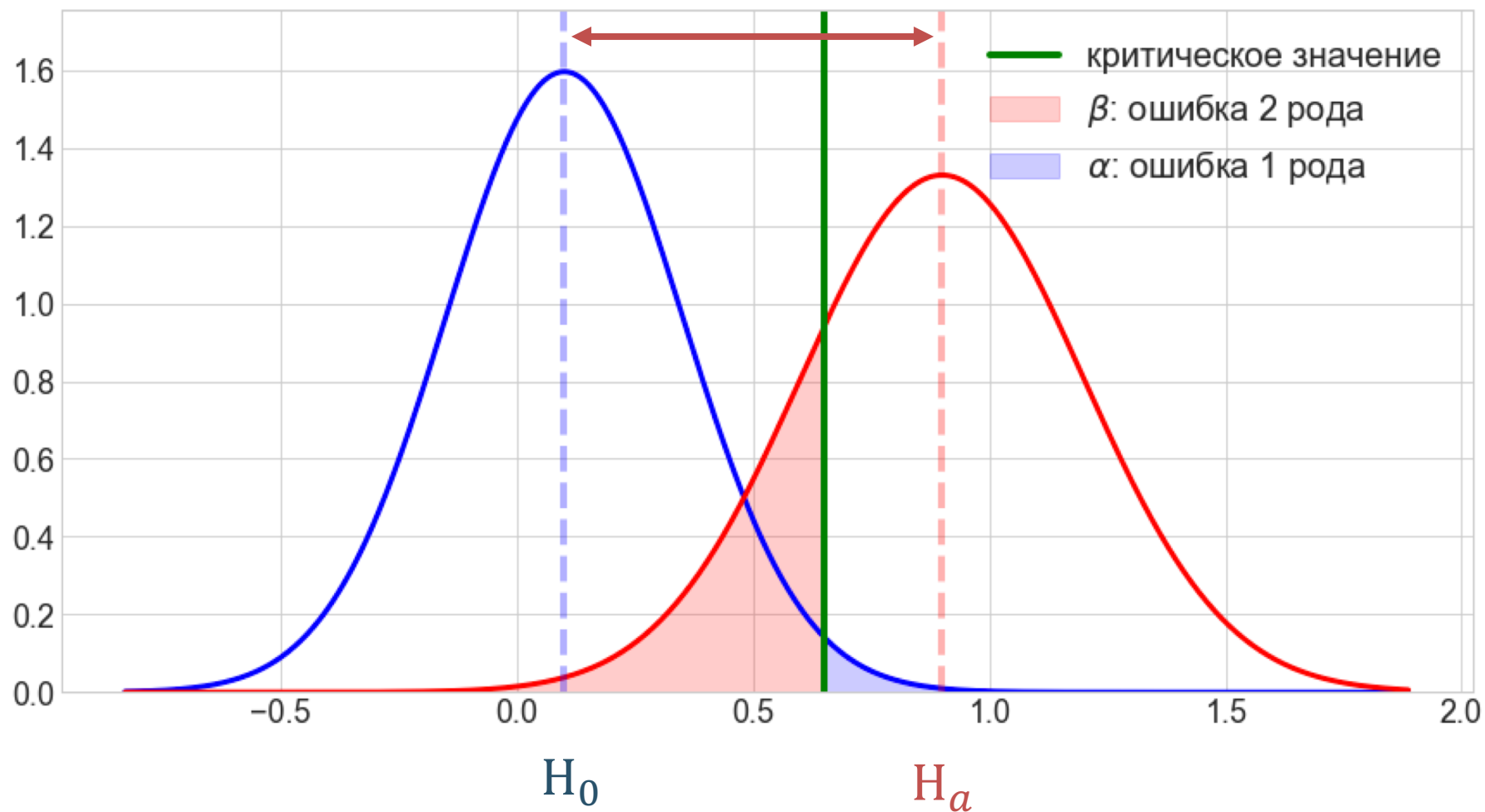
$$\alpha = \mathbb{P}(H_0 \text{ отвергнута} \mid H_0 \text{ верна})$$

$$\beta = \mathbb{P}(H_0 \text{ не отвергнута} \mid H_0 \text{ не верна})$$

Величину $1 - \beta$ называют **мощностью** критерия

Размер эффекта

Размер
эффекта



Сколько нужно наблюдений

- Необходимое количество наблюдений зависит от размеров ошибок первого и второго рода, а также от размера эффекта
- Фиксируем уровень значимости (ошибку 1 рода), на которую мы согласны
- Подбираем соотношение между минимальным размером эффекта, желаемой мощностью и объёмом выборки
- В выборе соотношении помогает заказчик эксперимента, у него обычно есть ограничения, с которыми нам придётся работать (количество магазинов, длительность АБ-теста и т.п.)

Таблица эффекта-ошибки

| | | Ошибка 1/2 рода $\alpha = \beta$ | | | |
|-------------------|------|----------------------------------|----|----|----------------|
| | | 0.1% | 1% | 5% | 10% |
| размер эффекта | 1% | много данных | | | |
| | 1.5% | | | | |
| | 3% | | | | |
| | 5% | | | | |
| | 10% | | | | мало данных |

- ❗ Совокупность этих трёх параметров (ошибка 1/2 рода, размер эффекта) позволяют рассчитать необходимый для эксперимента объём выборки.

Сколько нужно наблюдений

Пример: проверяем равенство конверсий до и после нововведений

$$H_0: p_0 = p_a$$

$$H_a: p_0 \neq p_a$$

Используем асимптотически-нормальный тест:

$$z = \frac{p_a - p_0}{\sqrt{P(1 - P) \cdot \left(\frac{1}{n} + \frac{1}{n}\right)}} \underset{H_0}{\overset{asy}{\rightsquigarrow}} N(0, 1)$$

**размер
эффекта**

Сколько нужно наблюдений

Ошибка второго рода:

$$\beta = \Phi \left(\frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_a(1-p_a)}} \cdot z_{1-\alpha} + \frac{p_0 - p_a}{\sqrt{\frac{p_a(1-p_a)}{n}}} \right)$$

Число наблюдений:

$$n = \left(\frac{z_{1-\alpha} \cdot \sqrt{p_0(1-p_0)} + z_{1-\beta} \cdot \sqrt{p_a(1-p_a)}}{p_a - p_0} \right)^2$$

**размер
эффекта**

Анализ мощности

До эксперимента:

- Какой нужен объём выборки, чтобы найти различия с разумной степенью уверенности
- Различия какой величины мы можем найти, если известен объём выборки

После эксперимента:

- смогли бы мы найти различия с помощью нашего эксперимента, если бы величина эффекта была равна Δ

Резюме

- Для многих критериев можно вывести формулу для расчёта необходимого числа наблюдений
- Число наблюдений зависит от ошибок $\frac{1}{2}$ рода и минимального размера эффекта, который мы хотим уловить
- Перед экспериментом необходимое число наблюдений определяют исходя из пожеланий заказчика и физических возможностей