Практическая эконометрика

Лекция 2.
1. Минимально различимый эффект (MDE)

2. Тесты для экспериментов в нестандартных ситуациях

Дисклеймер

- Слайды далеко не являются исчерпывающими. Сегодня очень многое будет на доске. Конспект доски повешу на онэкон по факту.
- Лекция основана на материалах:
  - Георгия Калашнова (часть 1),
  - Лаборатории JPAL (часть 1)
  - https://www.povertyactionlab.org/resource/power-calculations
  - Б.Б. Демешева (часть 2) https://www.youtube.com/@stats4mr174

Часть 2.1 Минимально различимый эффект (MDE)

- На практике «большие» эффекты редкость
- Если тест на значимость эффекта показал, что эффекта нет, то либо его на самом деле нет, либо он есть, но мы его не поймали.
- Если тест на значимость эффекта показал, что эффект есть, то либо он на самом деле есть, либо мы поймали какой-то шум (ложно-положительный результат).
- Увяжем между собой MDE, размер выборки, уровень значимости, мощность, долю наблюдений в тримент-группе и дисперсию признака.

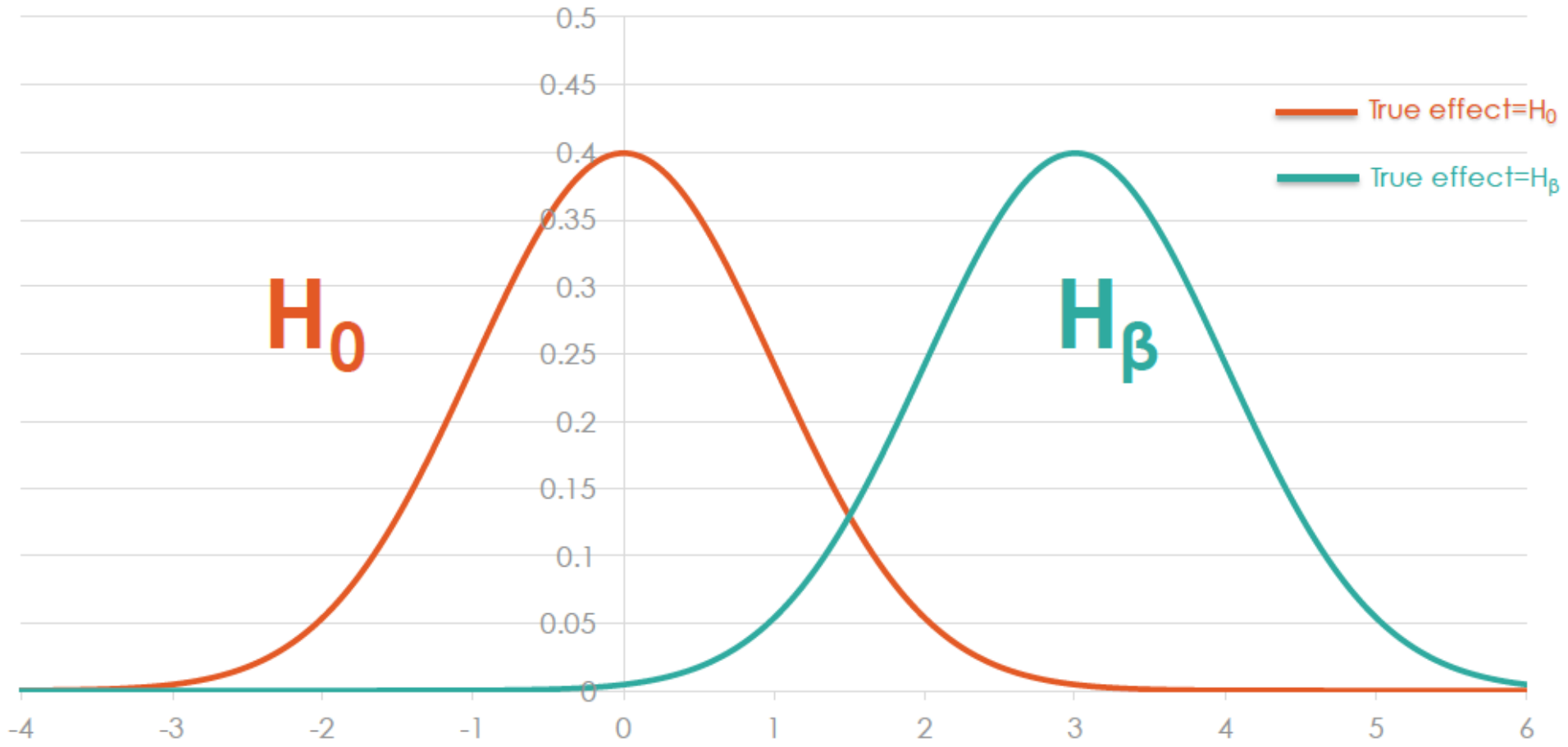
Неудобные вопросы:

- Что значит «достаточно большой объём выборки»?
- От чего зависит ответ?
- Как это связано с мощностью критерия?

Банальность, о которой часто забывают

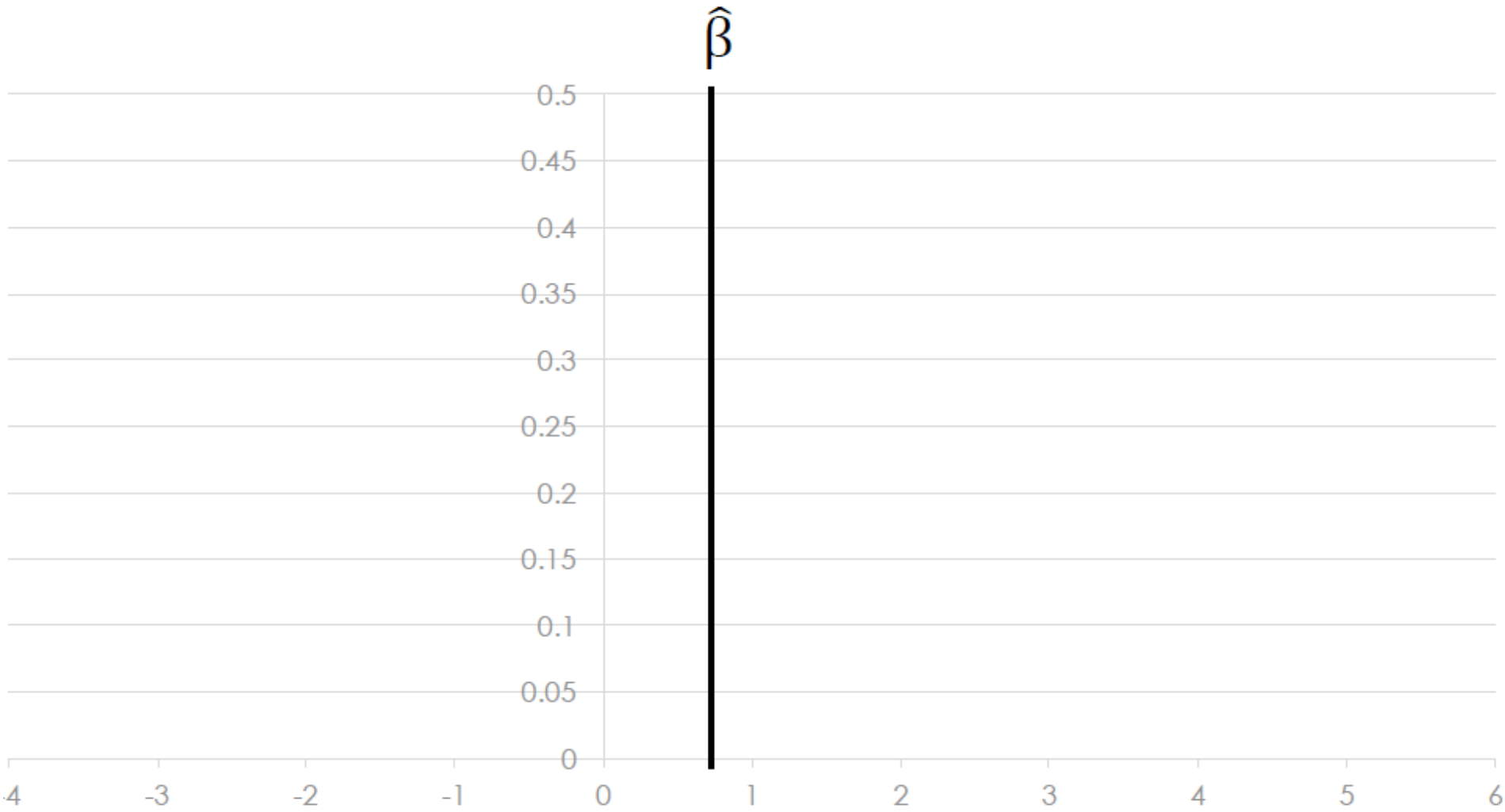
- Чем меньше предполагаемый размер эффекта, тем больше должна быть выборка!

Распределение эффекта при 2 разных гипотезах

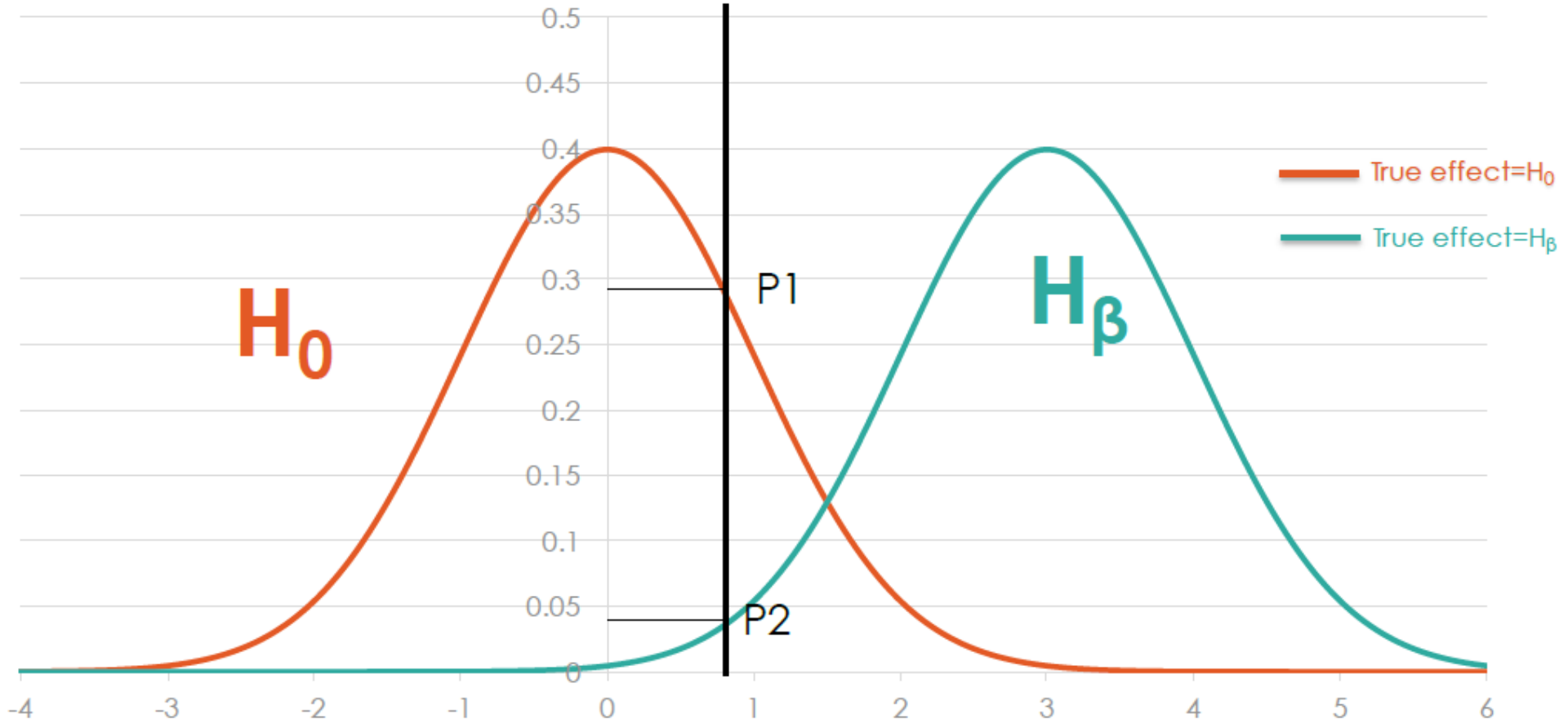


Beta – предполагаемый в альтернативной гипотезе размер эффекта

Проводим эксперимент 1 раз, видим 1 реализацию оценки

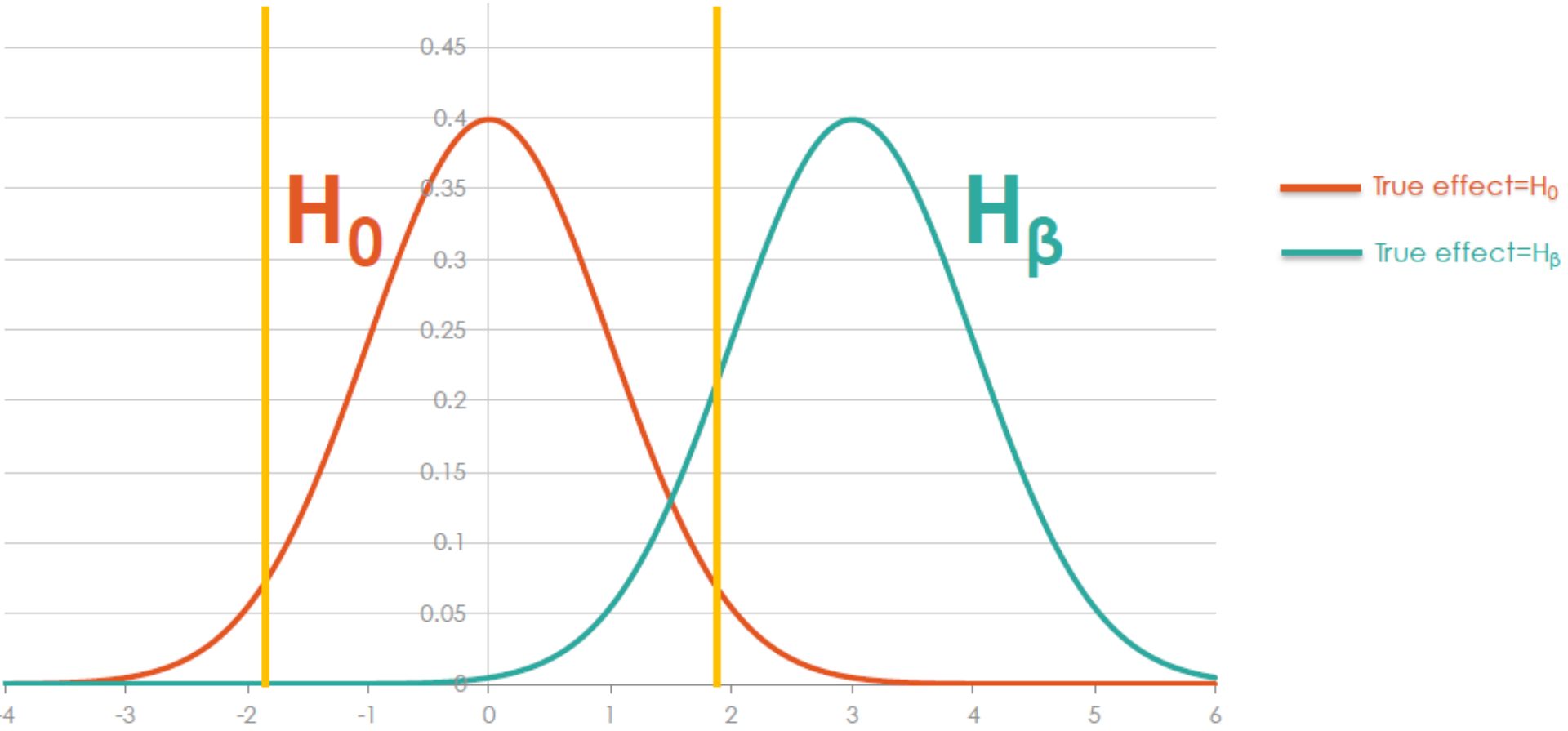


Из какого распределения наша оценка?

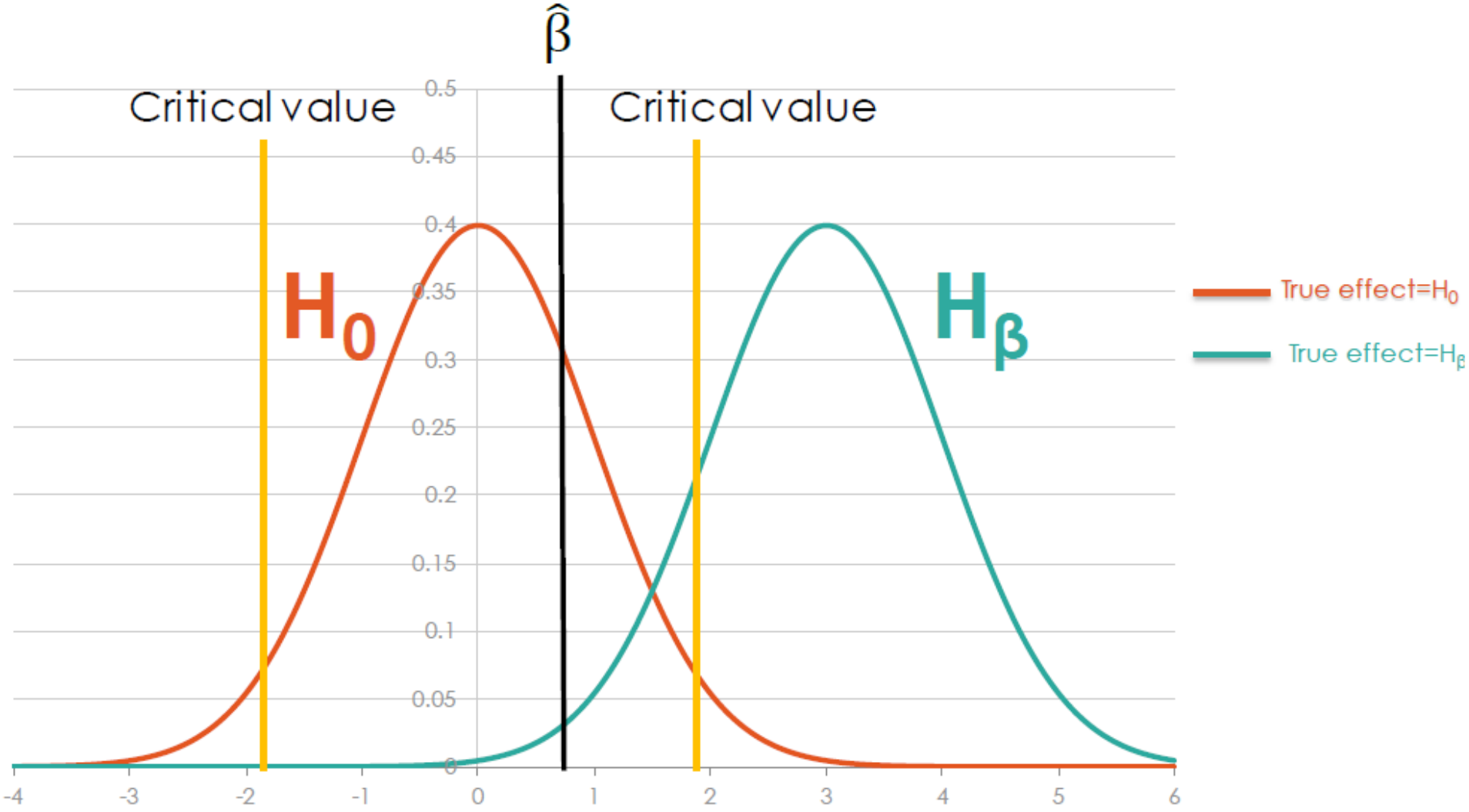


Скорее, из H0 – выше вероятность попасть туда

Критические значение 10% (по 5% с обоих «хвостов»)



Значима ли оценка? (двусторонняя гипотеза)
Размер эффекта ноль или больше нуля? (односторонняя)



«Презумпция нуля»

- «Бремя доказательства» лежит на исследователе: важно доказать, что эффект значим.
- Аналогично именно на исследователе лежит бремя доказательства, что оценка показывает именно заявленный эффект, что отсутствует эндогенностть и т.д.

Ещё несколько банальностей

Уровень значимости

- Уровень значимости 5% означает, вероятность случайно получить отличный от нуля результат составляет 5%

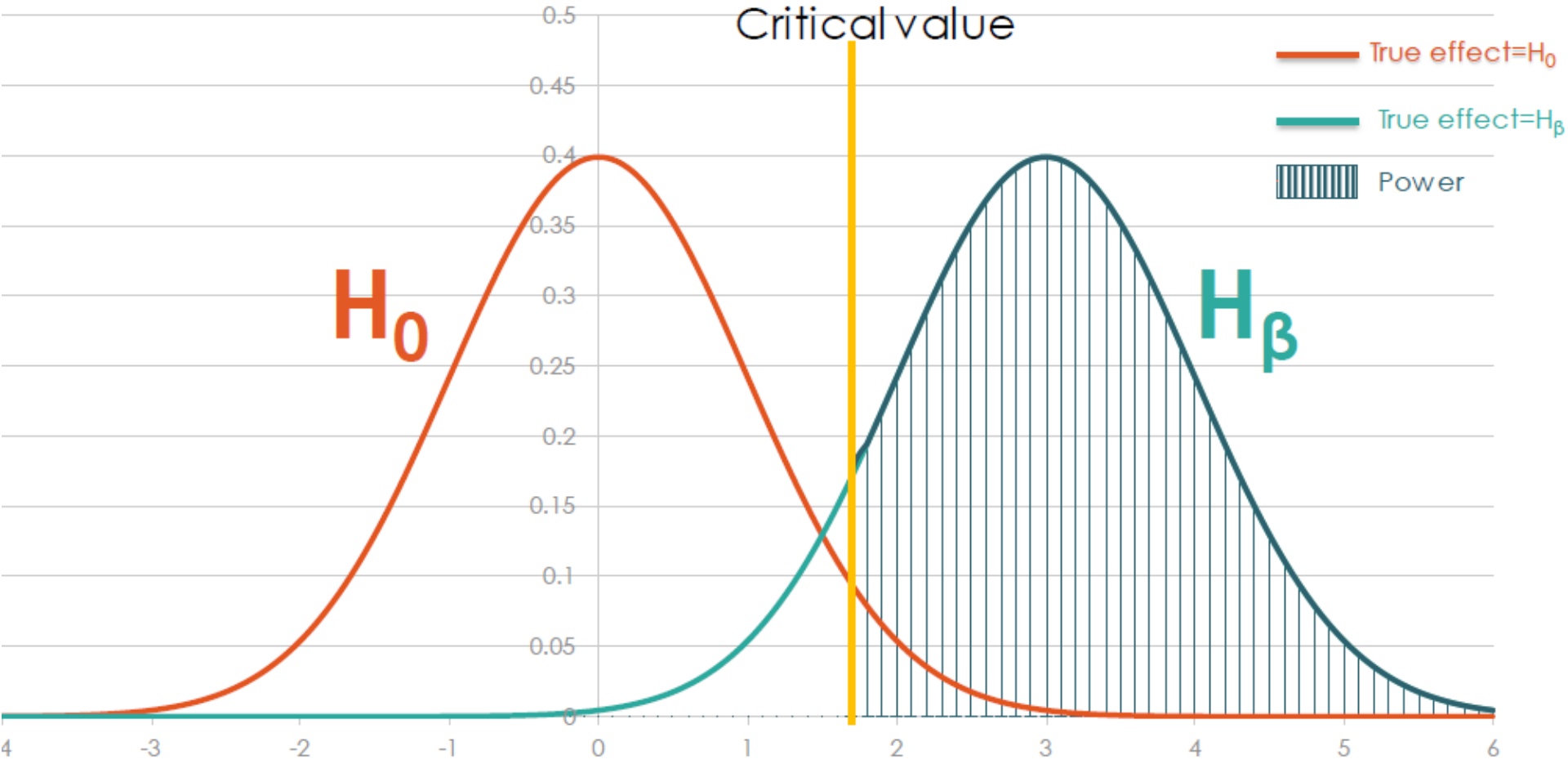
Ошибки первого и второго рода

	Тритмент эффект есть	**Тритмент эффекта нет**
Тест в пользу H-beta Оценка тритмент-эффекта значимая	Ок	Ошибка 1 рода (вероятность = альфа)
Тест в пользу H0 Оценка тритмент-эффекта незначимая	Ошибка 2 рода (вероятность = k)	Ок

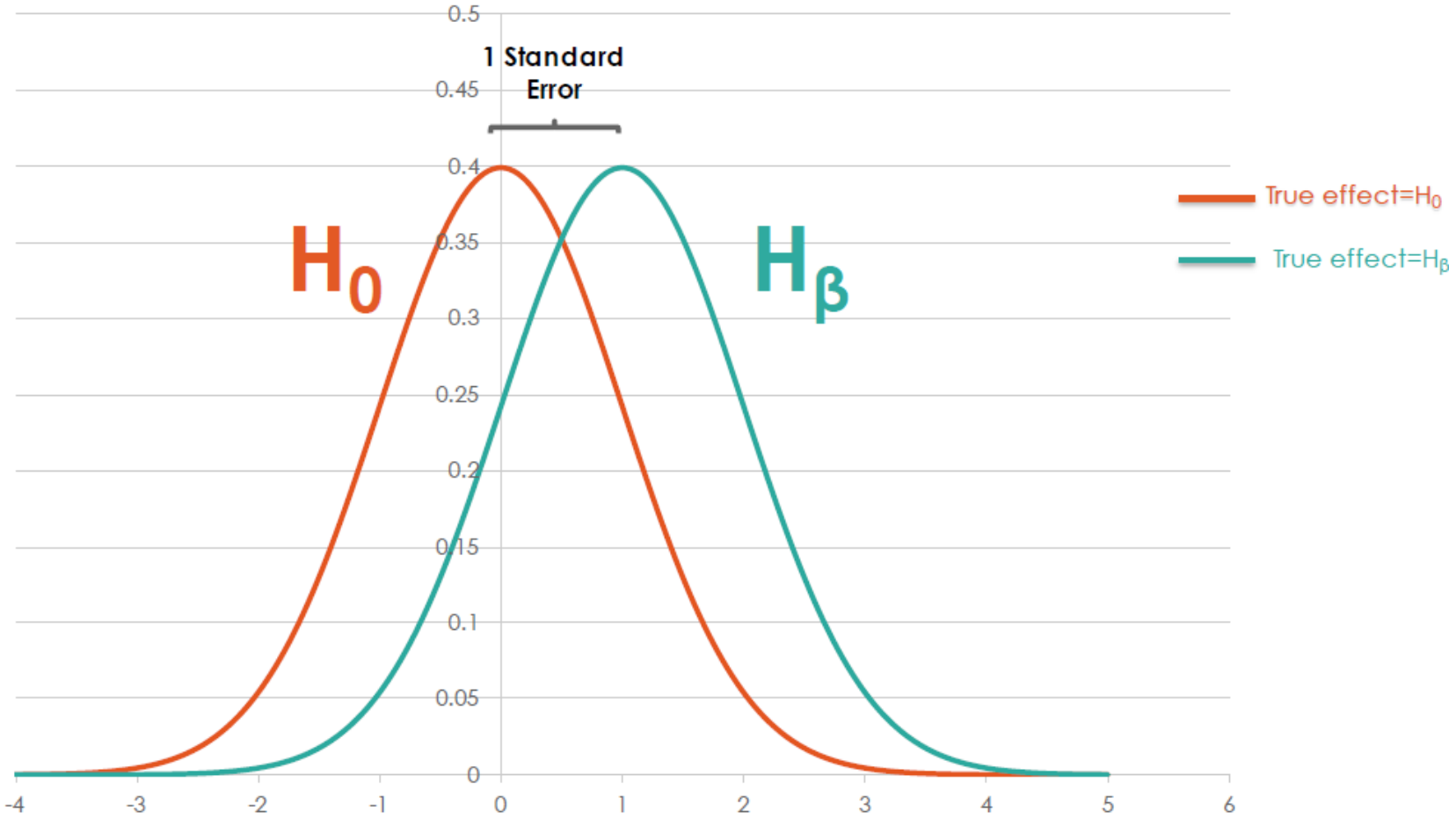
Мощность

- Это вероятность того, что если истинный эффект - размера b, то наш эксперимент будет в состоянии разграничить оценку этого эффекта и ноль.
- Это вероятность избежать ошибки 2 рода (= 1 - k).

Мощность

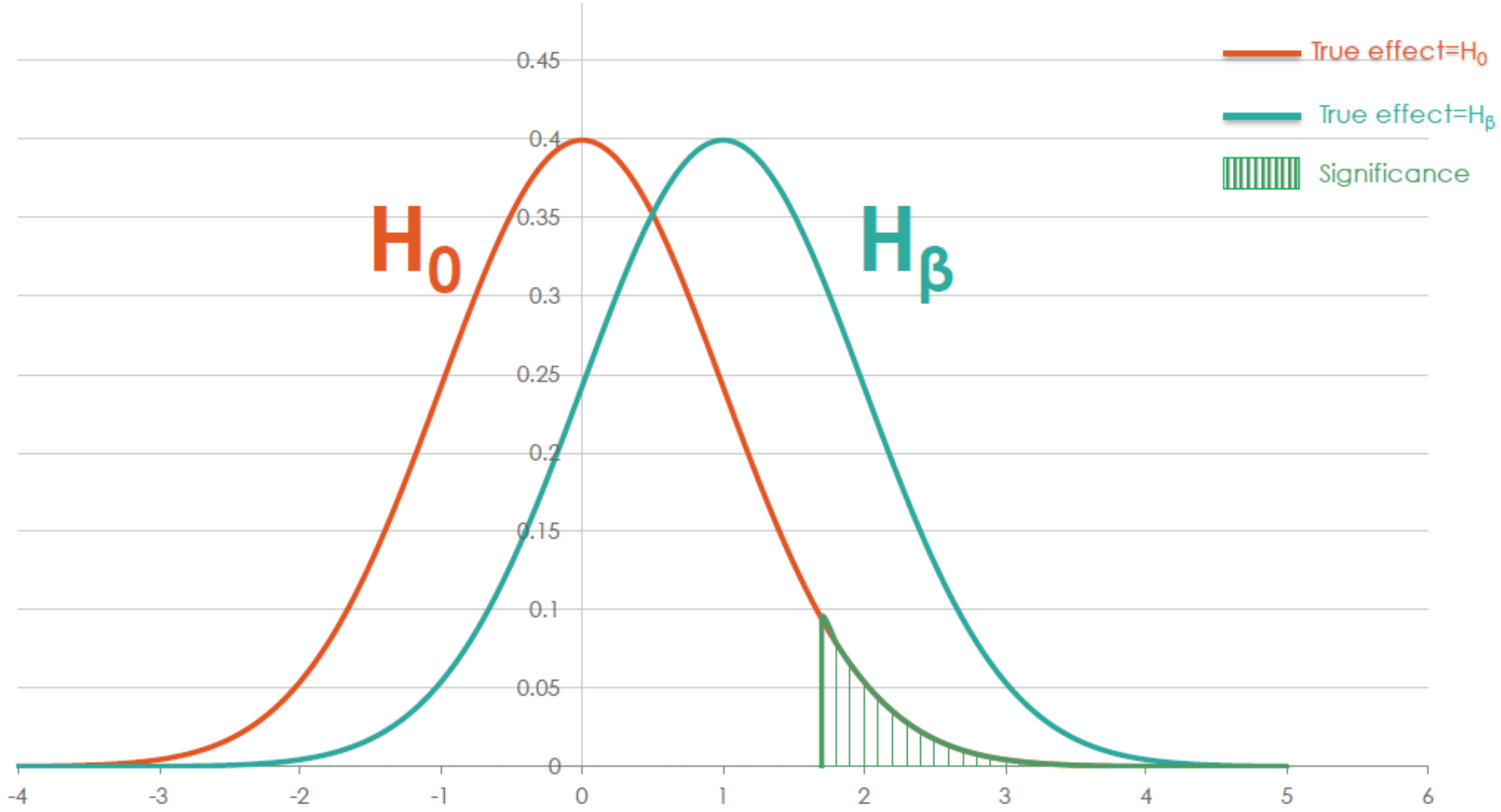

Critical value
True effect=H0
True effect=Hβ
Power
H0
Hβ
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
5
6

Мощность и предполагаемый размер эффекта

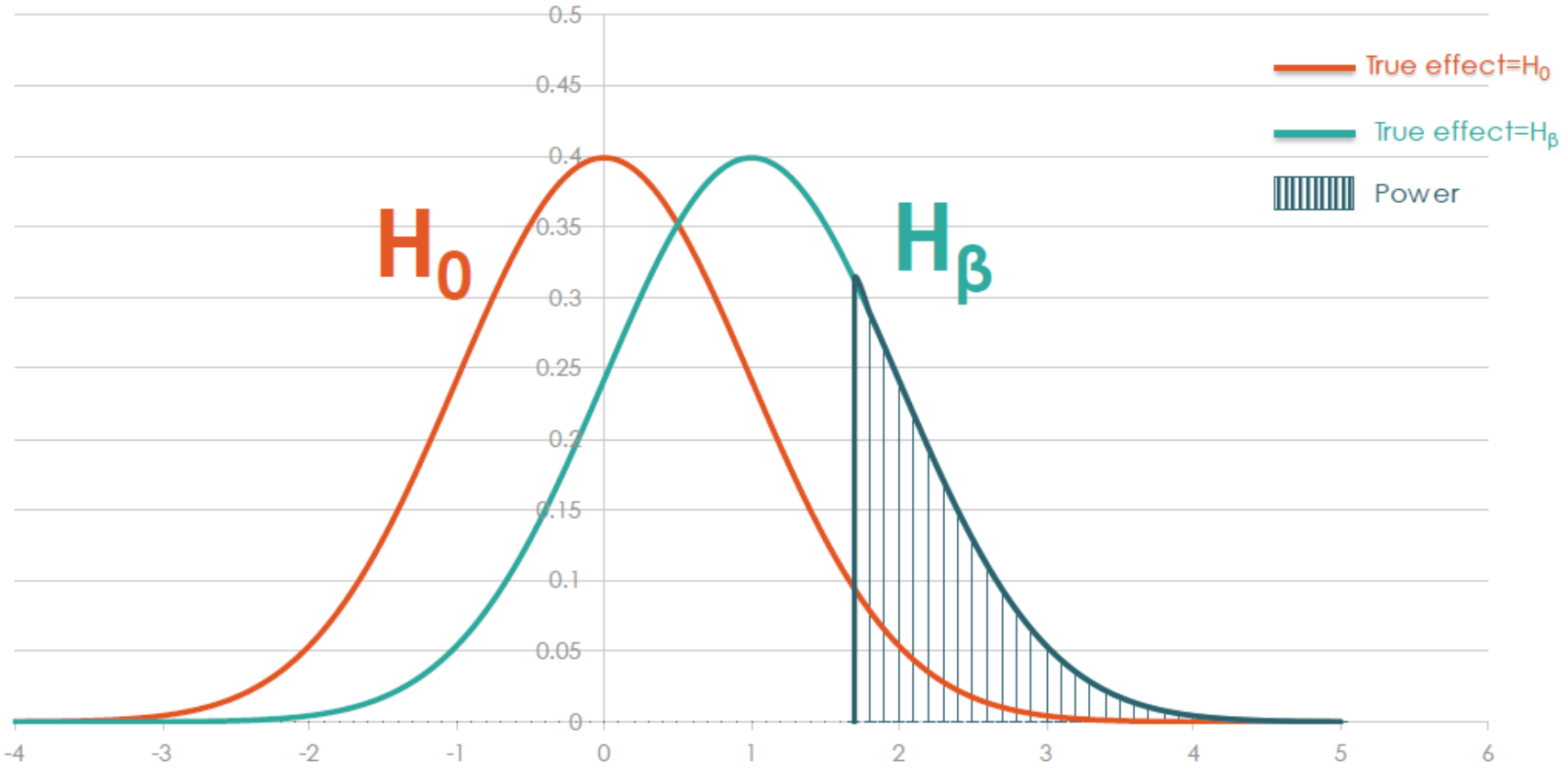


Предполагаемый в H1 размер эффекта = 1 s.e

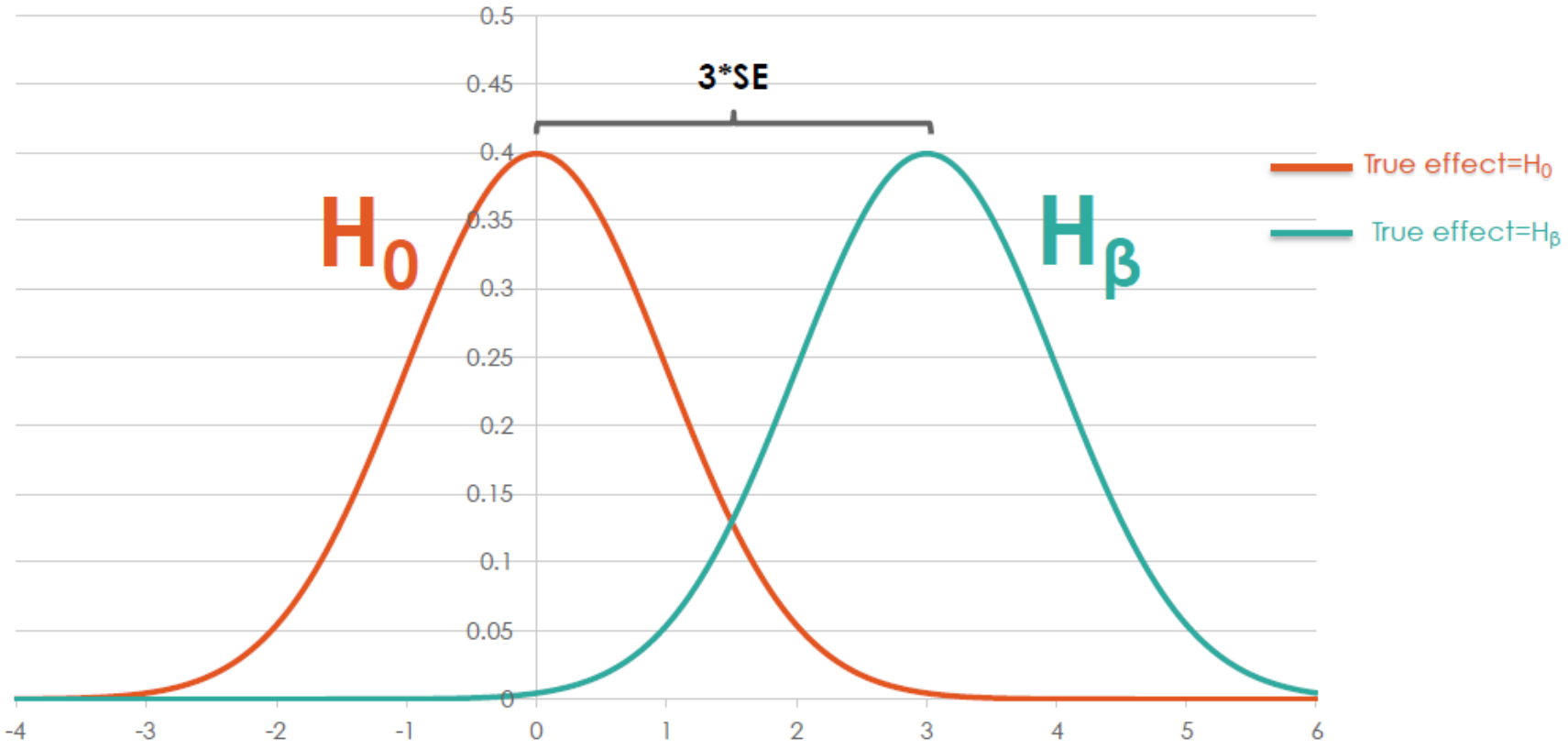
Мощность и предполагаемый размер эффекта



Мощность и предполагаемый размер эффекта

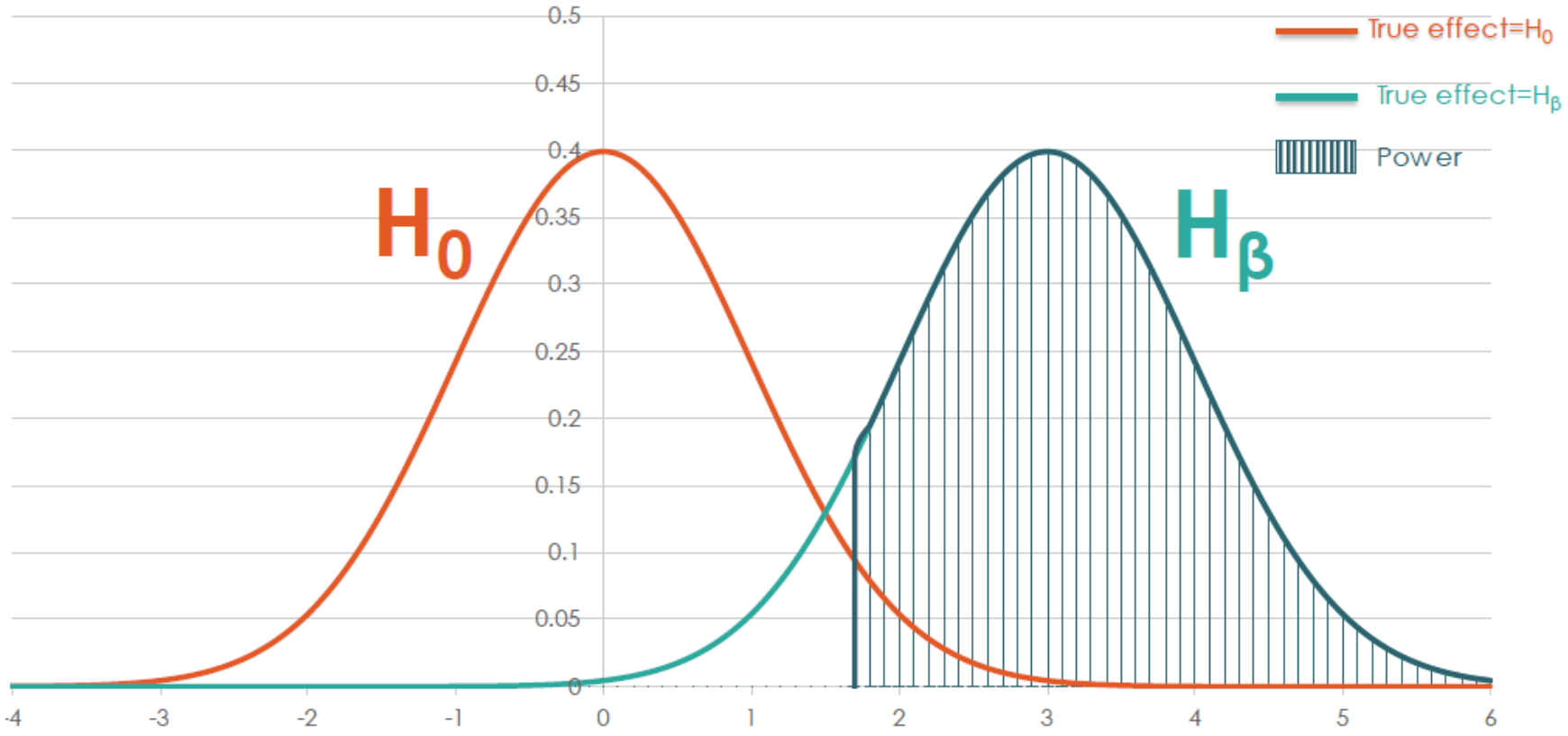


Мощность и предполагаемый размер эффекта



Предполагаемый в H1 размер эффекта = 3 s.e

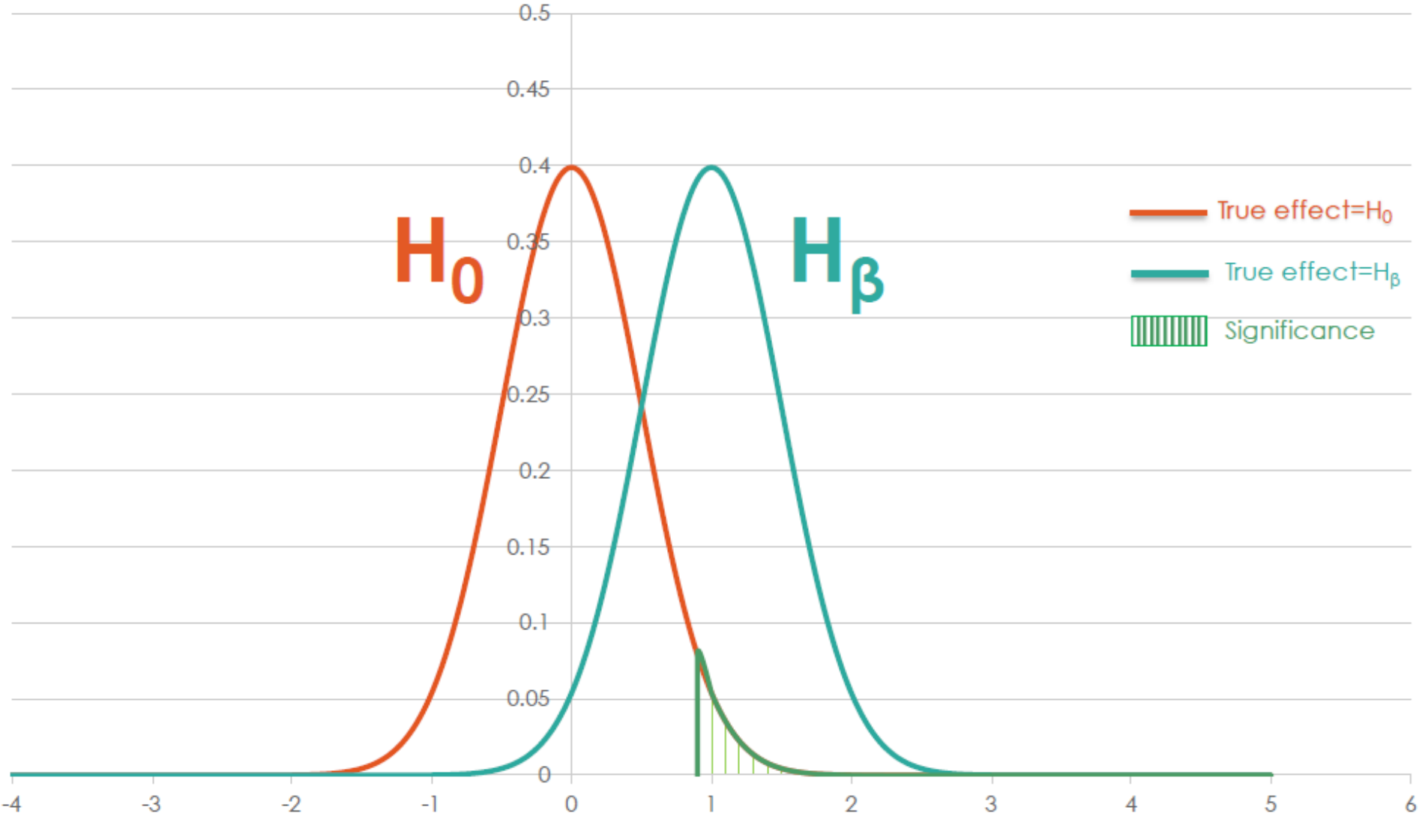
Мощность и предполагаемый размер эффекта



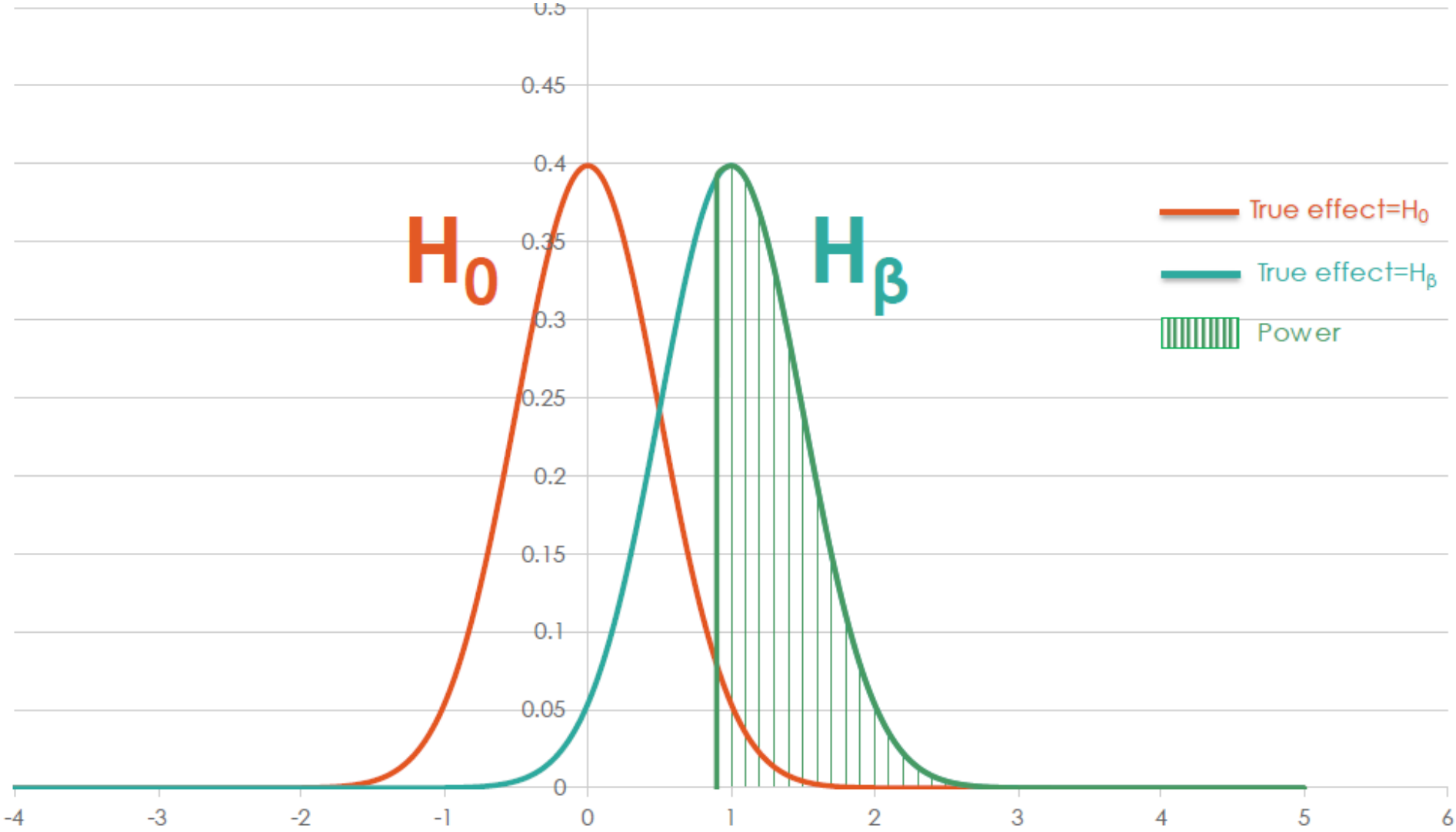
Мощность и предполагаемый размер эффекта

- Чем больше предполагаемый размер эффекта, тем больше мощность.
- Интуиция?
- Чем больше размер эффекта, тем «сложнее» его не отличить от нуля.

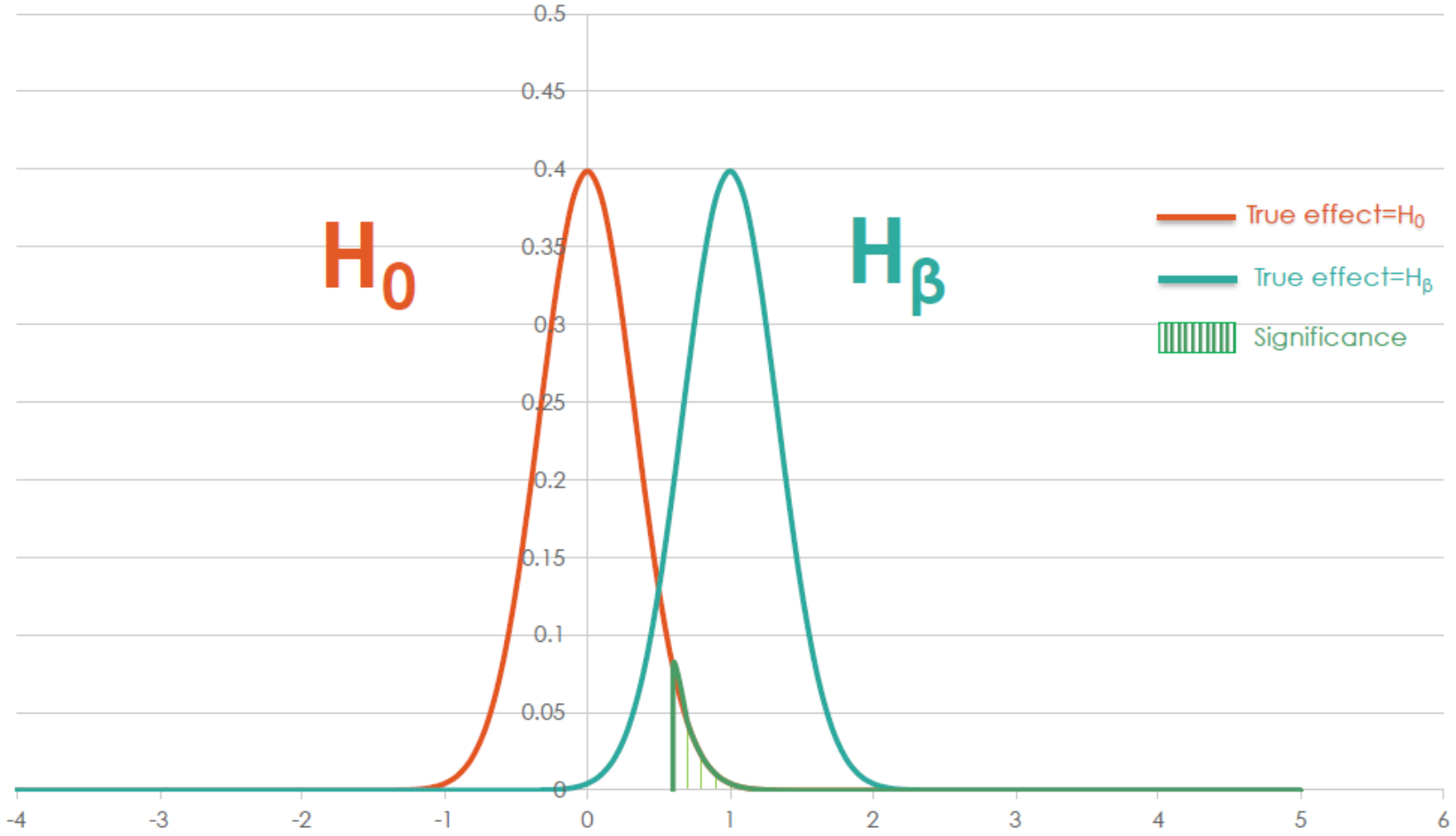
Значимость и размер выборки, n=4000



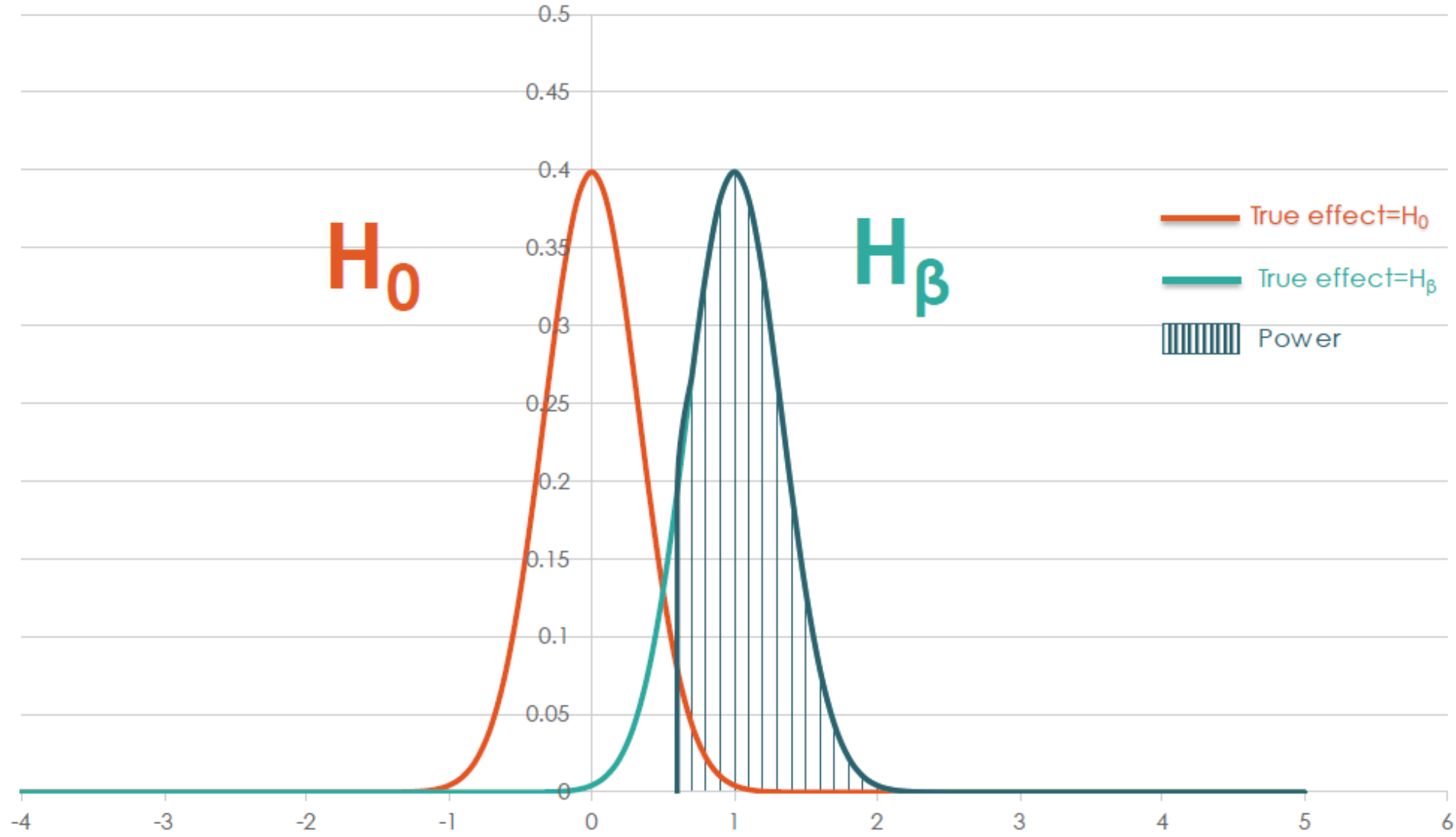
Мощность и размер выборки, n=4000



Мощность и размер выборки, n=9000



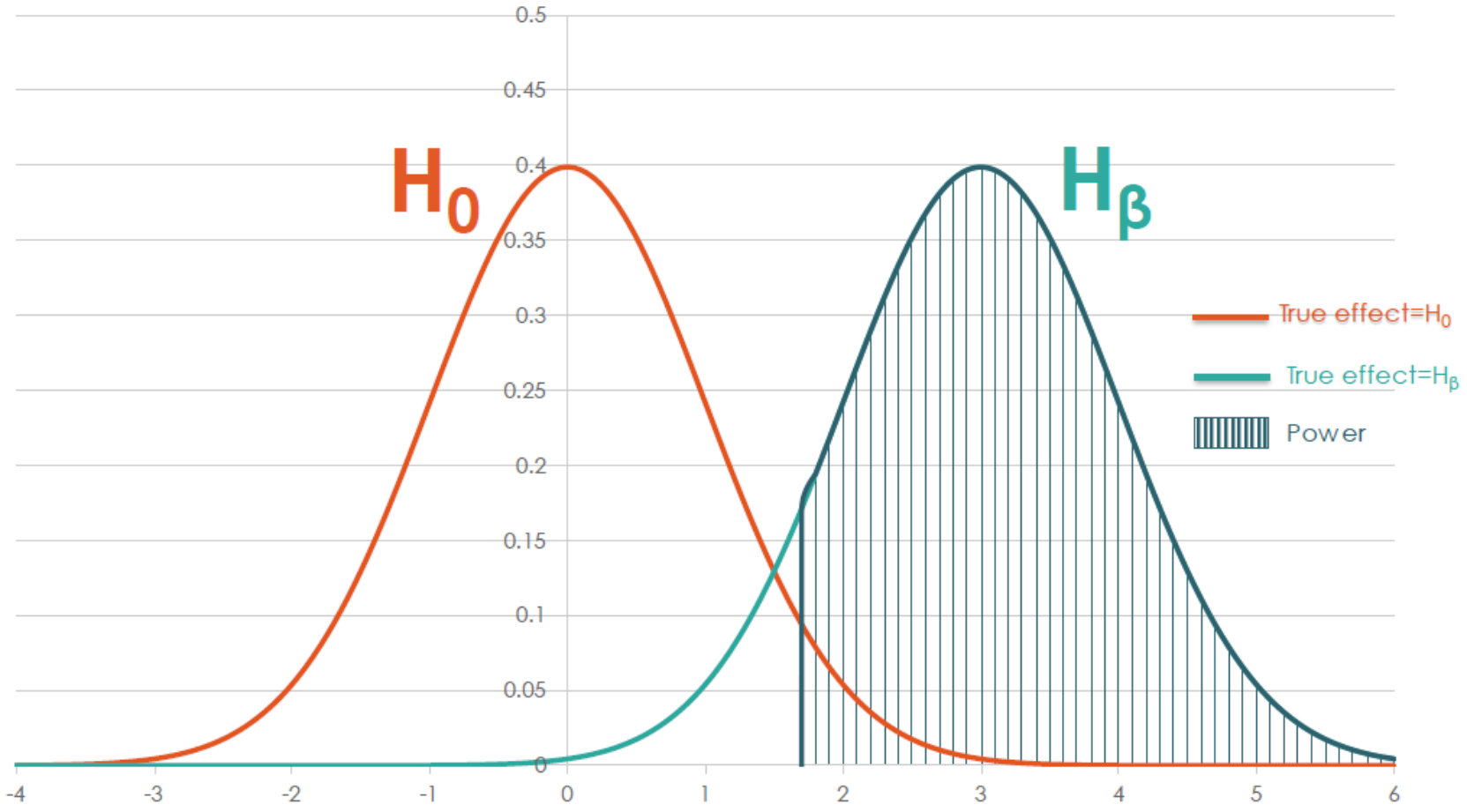
Значимость и размер выборки, n=9000



Мощность и размер выборки

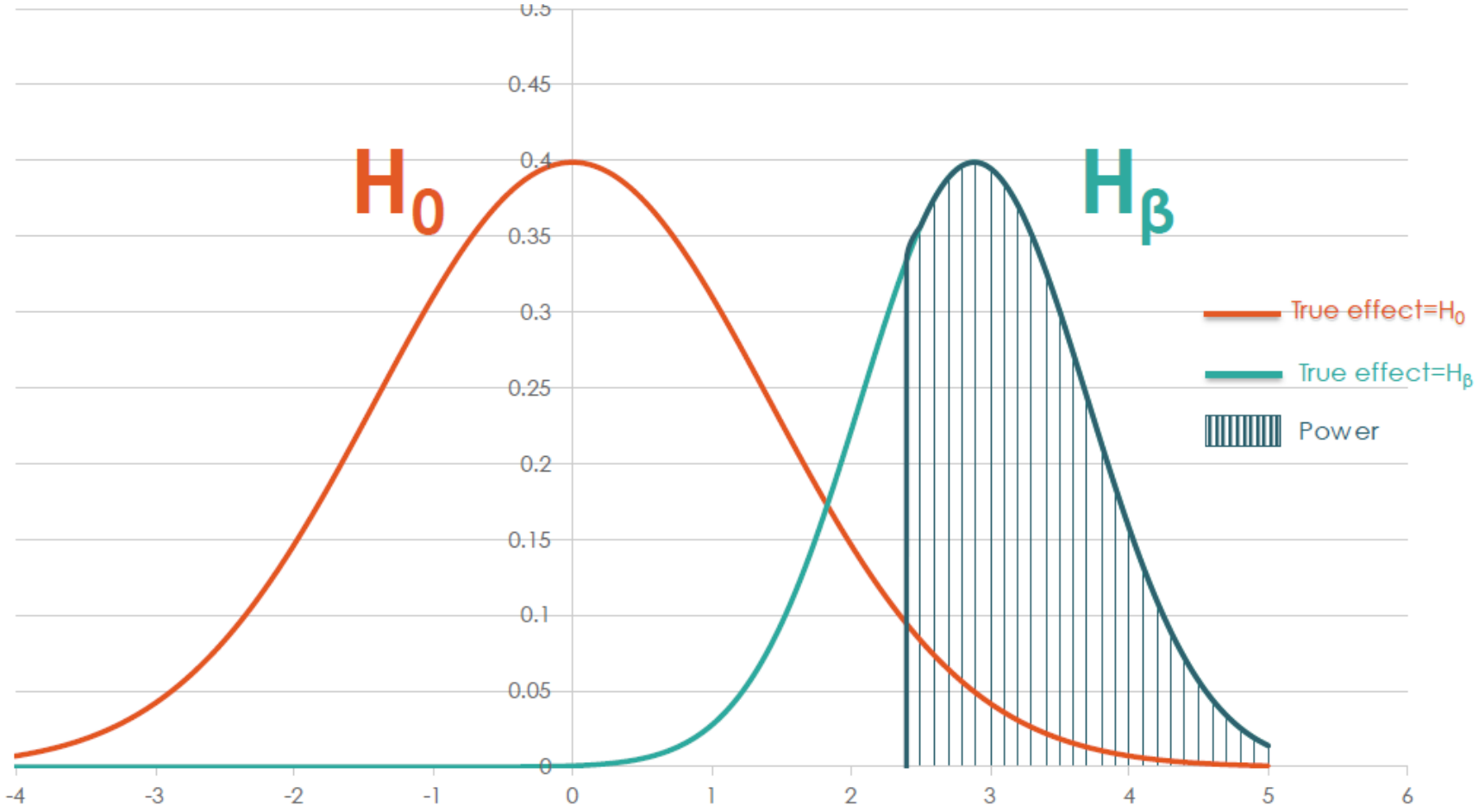
- Чем больше выборка, тем больше мощность.
- Интуиция?
- Принимая во внимание большее количество наблюдений, мы с меньшей вероятностью упустим эффект.

Мощность и пропорция между тритмент и контрольной гр.



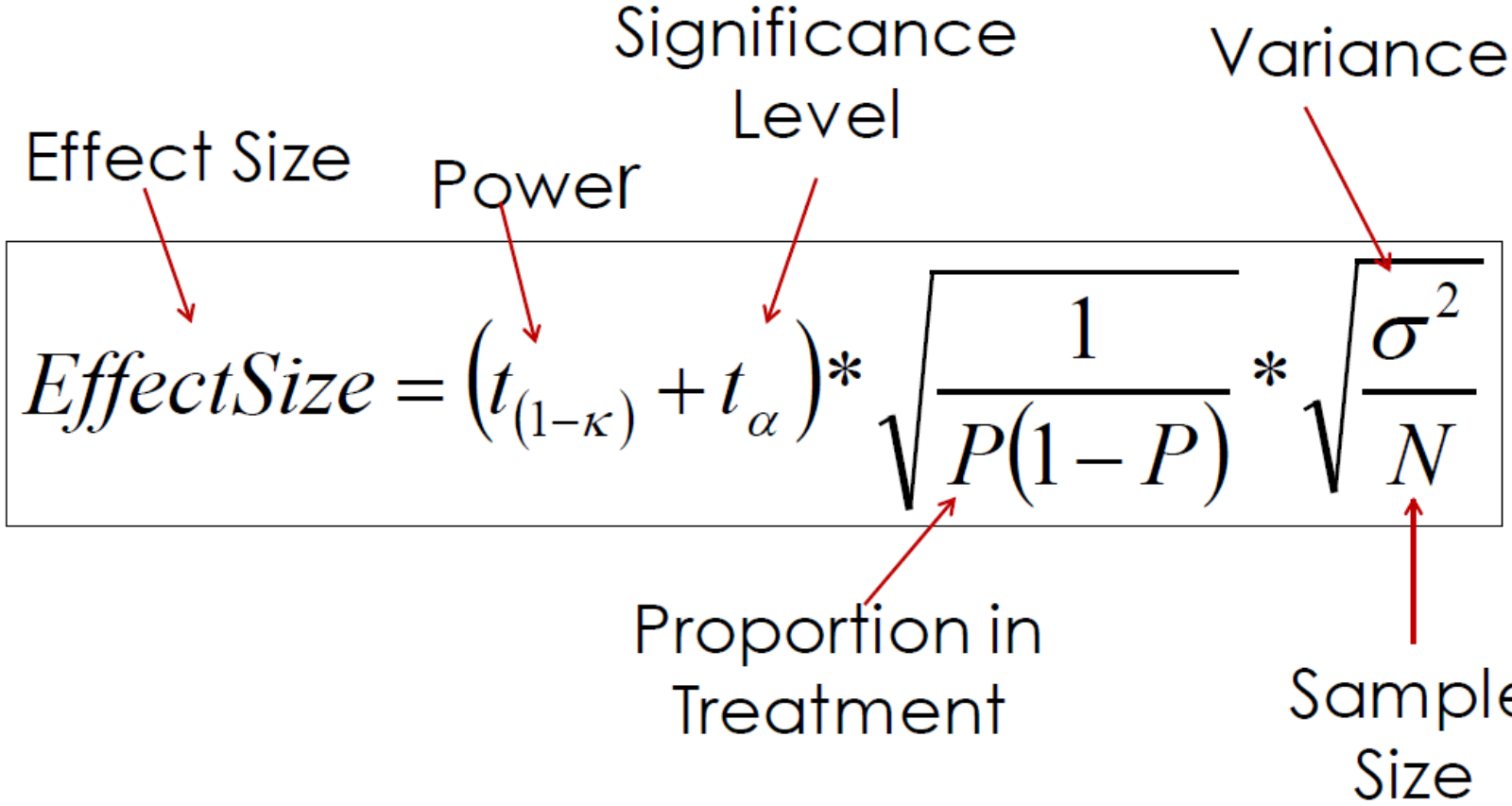
50% и 50%

Мощность и пропорция между тритмент и контрольной гр.



75% и 25%

MDE (minimum detectable effect) size



Источник: Glennerster, Takavarasha “Running randomized evaluations. A practical Guide”, ch 6.

Альфа, а не альфа/2, т.к. односторонний тест

MDE: аналогия

$$EffectSize = \left(t_{(1-\kappa)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}} * \sqrt{\frac{\sigma^2}{N}}$$

Стандартная ошибка оценки коэффициента

- Аналогия: расчётная t-статистика в тесте на значимость коэффициента показывает, на сколько стандартных ошибок оценка отличается от нуля.

Minimum detectable effect: аналогия

=>

Чтобы быть значимым на 5% уровне, коэффициент должен быть как минимум в “t_табличное” раз больше, чем его стандартная ошибка

=>

Минимально отличимый от нуля на alpha %-ном уровне коэффициент должен быть = t_alpha/2 * s.e(b^)

Minimum detectable effect: наш случай

- Чтобы быть отличным от нуля при заданной мощности (1-k) и уровне значимости (alpha), коэффициент должен быть как минимум в

  t_(1-k)+t_alpha

  раз больше, чем s.e.

$$EffectSize = \left(t_{(1-\kappa)} + t_{\alpha}\right) * \sqrt{\frac{1}{P(1-P)}} * \sqrt{\frac{\sigma^2}{N}}$$

Часть 2.2 Тесты для нестандартных ситуаций

- При каких(ой) предпосылках(е) t-статистка распределена стандартно нормально при верной нулевой гипотезе?
- Что делать, если выборка «маленькая»?
  - бутстрап («не совсем маленькая» – по 100 наблюдений в каждой группе)
  - t-тест Уэлча («совсем маленькая» – по 30 наблюдений в каждой группе)
- Что делать, если Y распределён не нормально?
  - бутстрап (надо думать о статистике)
  - U-критерий Манна-Уитни (часто используется, подходит для любого распределения Y и любого размера выборки)

Тест Уэлча на равенство средних в 2 выборках

- Подходит для малых выборок (по 30 наблюдений в каждой группе)
- t-статистика не успевает сойтись к N(0,1), но может быть, не сильно отклоняется от неё?
- Давайте заменим N(0,1) на t(d), но число степеней свободы посчитаем в зависимости от параметров выборки
- Основная идея в выкладках: подгоняем статистику под N(0,1), приравнивая матожидание и дисперсию к желаемым значениям
- В R это команда t.test(sample1, sample)
- На доске -- выкладки

U-критерий Манна-Уитни

- Подходит для любого распределения признака и любого размера выборки
- Игнорирует распределения признака, считает, сколько раз признак в тритмент-группе превосходит признак в контрольной группе
- На доске – игрушечный пример:
  - N(T=1)=3, N(T=0)=2
  - Значения при T=1: 2, 4, 7
  - Значения при T=0: 5, 6
  - Рассчитаем U-статистику, p-значение
- На доске – примеры тримент-эффектов, когда критерий работает «хорошо» или «плохо»
- В русскоязычной Википедии Манн-Уитни описан неверно!