Виды тестирования

Fixed Horizon

Проще в реализации

Проблема подглядывания

- Фиксированный временной горизонт проведения эксперимента
- Решение принимается один раз в конце

Sequential testing

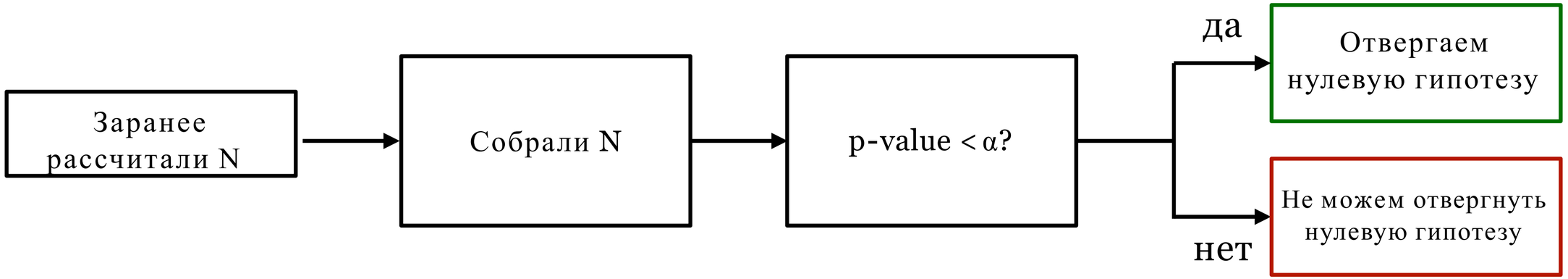
Сложнее в реализации

Always Valid Inference (AVI) – решение проблемы подглядывания

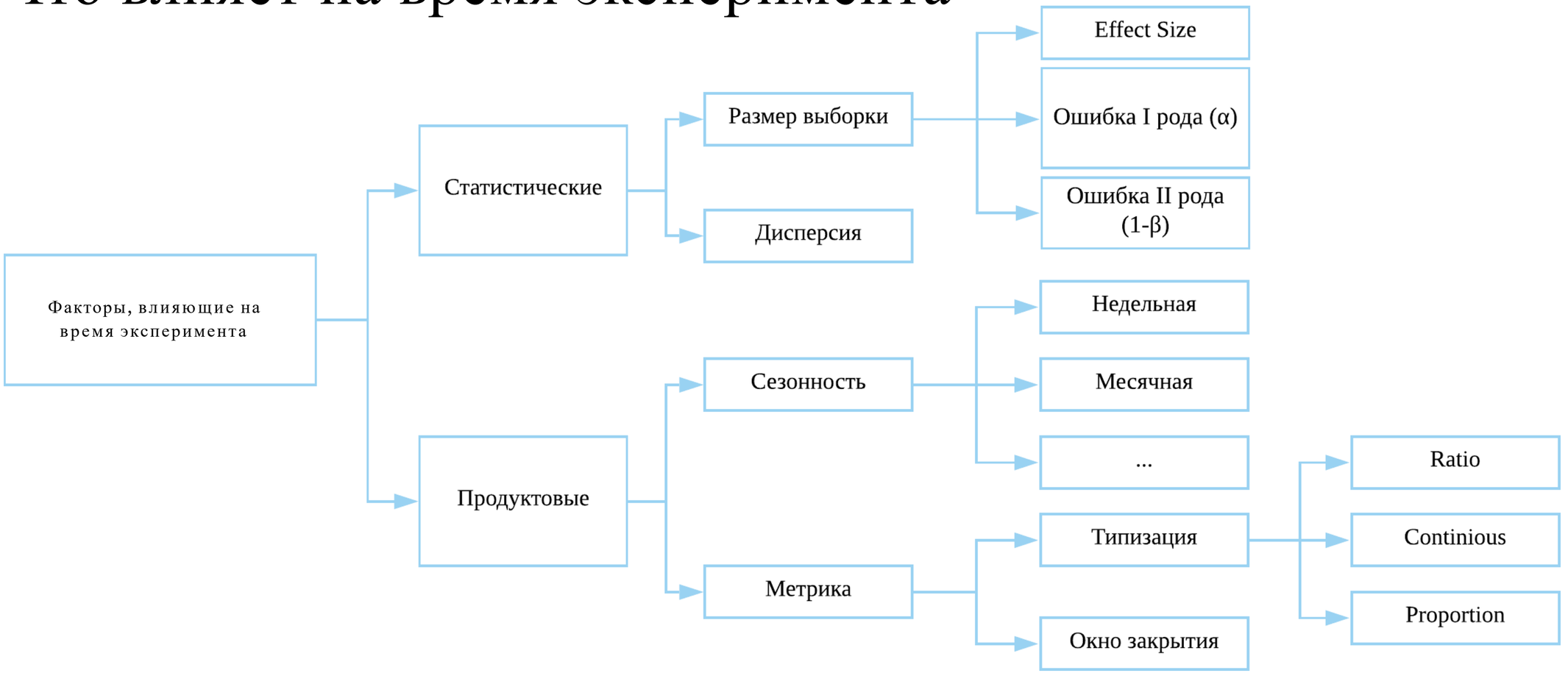
- Решение об остановке теста принимается real-time
- Значимость накопительно рассчитывается по дополнительным порогам принятия решения

Fixed horizon

Этапы



Что влияет на время эксперимента



Мощность

Способность увидеть значимые различия в метрике там, где они на самом деле есть, называется *мощностью* (то же самое, что и чувствительность)

Высокая мощность метрики позволяет:

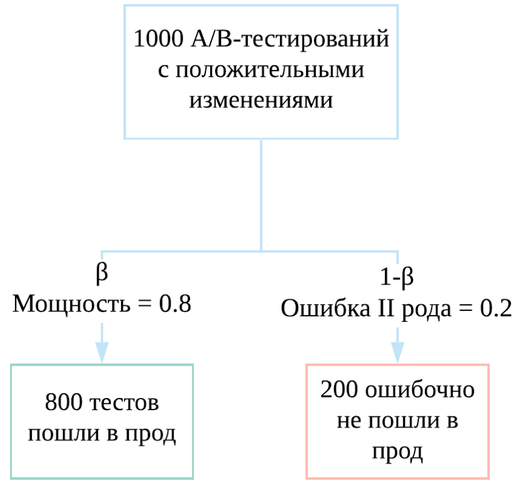
- видеть достаточно маленькие изменения
- или использовать меньше данных

Мощность и ошибка II рода

Мощность является важным параметром, потому что никому не хотелось бы выкидывать эксперименты с реальными эффектами.

Пример:

Допустим, мы берем уровень мощности в 80% как минимальный допустимый порог для экспериментов, то из 1000 экспериментов с реальным приростом в метрику, в 800 мы были бы уверенны, что прирост есть. Остаются 200, которые будут выкинуты в мусорку зря, потому что остается False Negative (ошибка II) = 0.2.



Мощность и ошибка II рода

Вывод

Мощность нужно максимизировать на столько, на сколько позволяют возможности продукта.

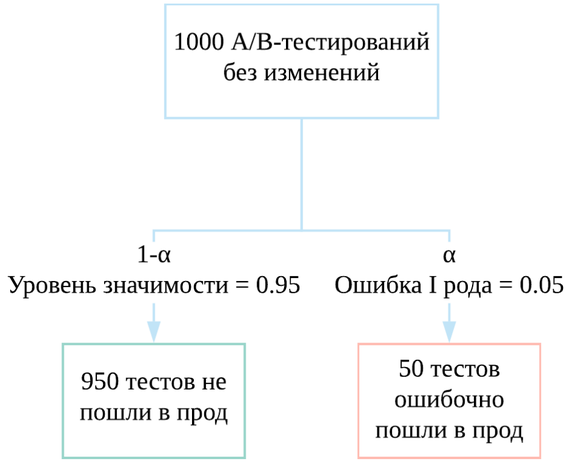
Чем выше уровень мощности, тем меньше хороших экспериментов будут ошибочно забыты

Уровень значимости и ошибка I рода

Когда принимается решение о результатах эксперимента, считается, что в первую очередь нужно смотреть на p-value (ошибка I рода или false positive): вероятность отклонить нулевую гипотезу при условии, что она верна

Пример:

Допустим, мы берем уровень значимости в 95% как минимальный допустимый порог для экспериментов, то из 1000 безуспешных экспериментов (нет эффекта) в 950 мы были бы уверены, что эффекта реально нет. Но 50 ошибочно выкатили бы в продакшен, хотя в этом нет никакого смысла, потому что в них мы наблюдаем случайность, а не реальную закономерность.

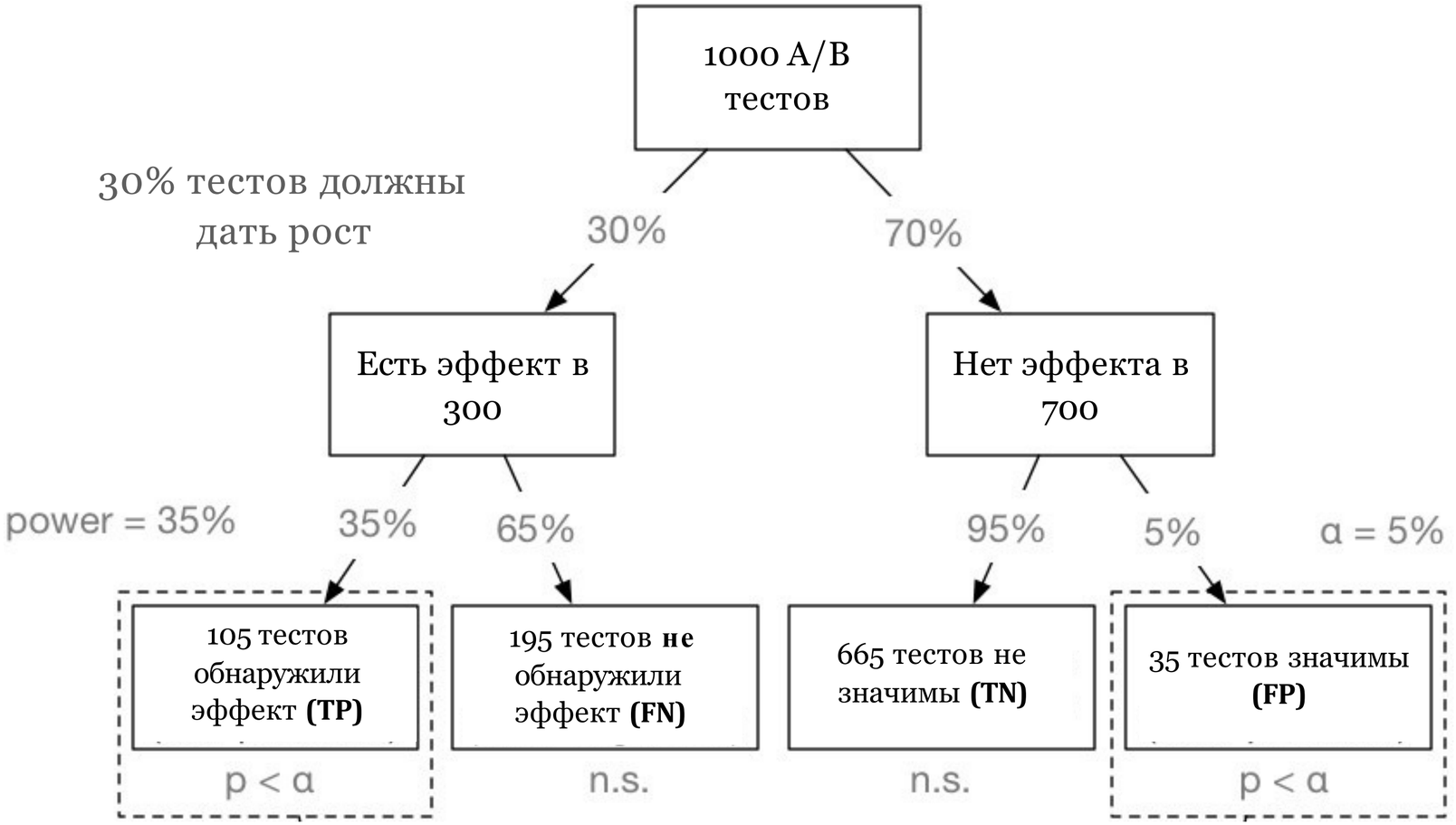


Уровень значимости и ошибка I рода

Вывод

Так же как и мощность, уровень значимости необходимо максимизировать по мере возможностей.

Чем выше уровень значимости, тем меньше бесполезных экспериментов будут выкатываться в продакшн.


1000 A/B тестов
30% тестов должны дать рост
30%
70%
Есть эффект в 300
Нет эффекта в 700
power = 35%
35%
65%
95%
5%
α = 5%
105 тестов обнаружили эффект (TP)
195 тестов не обнаружили эффект (FN)
665 тестов не значимы (TN)
35 тестов значимы (FP)
p < α
n.s.
n.s.
p < α

Формула размера выборки

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

Далее будем обозначать MDE так:

$$\text{MDE} = (\mu_1 - \mu_2)^2$$

Если ее переписать под two sample t-test, то получится следующая формула:

$$H_0: \mu_1 = \mu_2$$

$$n_g \geq \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \frac{\sigma_1^2 + \sigma_2^2}{(\mu_1 - \mu_2)^2}$$

Предполагая равные дисперсии получим:

$$n_g \geq 2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{\sigma}{\mu_1 - \mu_2}\right)^2$$

Это размер 1 группы. То есть для всей выборки (A и B группы)

$$\text{N} = 2 * n_g$$

Пример

Пусть

$\alpha = 0.05$, то есть $Z_{1-\alpha/2} = 1.96$

$\beta = 0.20$ so $Z_{1-\beta} = 0.8416$

$$n_g = 2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{\sigma}{\mu_1 - \mu_2}\right)^2$$

$$= 2\times(1.960 + 0.8416)^2 \left(\frac{\sigma}{\mu_1 - \mu_2}\right)^2$$

$$= 2\times 7.849 \left(\frac{\sigma}{\mu_1 - \mu_2}\right)^2$$

Округлив 7.849 до 8 можно получить быстро формулу для двустороннего т-теста:

$$n_g = 16 * \left(\frac{\sigma}{\mu_1 - \mu_2}\right)^2$$

MDE

Минимальный ожидаемый эффект – это наименьший **детектируемый** эффект полученный от изменений, который с уверенностью сможет обнаружить статистический критерий.

Более формально: это наименьший истинный эффект полученный от изменений, который имеет определенный уровень статистической мощности для определенного уровня статистической значимости, учитывая конкретный статистический тест.

Минимальный ожидаемый эффект (MDE)

Минимальный ожидаемый эффект – это наименьший **детектируемый** эффект полученный от изменений, который с уверенностью сможет обнаружить статистический критерий.

Более формально: это наименьший истинный эффект полученный от изменений, который имеет определенный уровень статистической мощности для определенного уровня статистической значимости, учитывая конкретный статистический тест.

... детектируемый... - если p-value > alpha (уровень значимости), то это не значит, что нет эффекта. Эффект может быть, но не больше, чем MDE с заданной O1P, мощностью и посчитанной дисперсией

Минимальный ожидаемый эффект (MDE)

Минимальный ожидаемый эффект – это наименьший детектируемый эффект полученный от изменений, который с уверенностью сможет обнаружить статистический критерий.

Более формально: это наименьший истинный эффект полученный от изменений, который имеет определенный уровень статистической мощности для определенного уровня статистической значимости, учитывая конкретный статистический тест.

... уровень статистической мощности для определенного уровня статистической значимости... - MDE считается для конкретных альфа, бетты, размера выборки и дисперсии. То есть MDE может варьироваться в зависимости от заданных вами альфы и бетты.

Минимальный ожидаемый эффект (MDE)

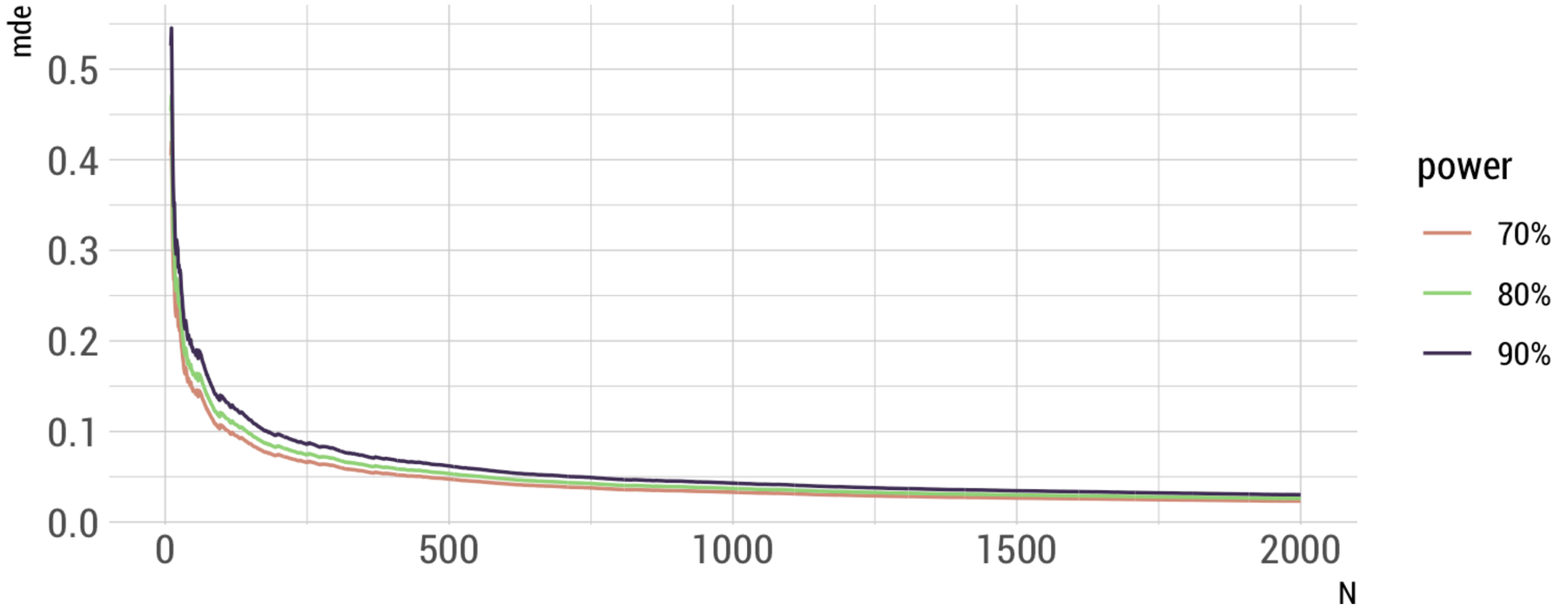
Минимальный ожидаемый эффект – это наименьший детектируемый эффект полученный от изменений, который с уверенностью сможет обнаружить статистический критерий.

Более формально: это наименьший истинный эффект полученный от изменений, который имеет определенный уровень статистической мощности для определенного уровня статистической значимости, учитывая конкретный статистический тест.

... статистический тес... - в формуле MDE используется конкретное распределение метрики. MDE будет считаться по-разному для разных критериев: t-критерий, хи-квадрат и т.д.

Чем меньший MDE мы хотели бы получить, тем больше наблюдений потребуется для его обнаружения. То есть больше наблюдений – выше точность.

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



Как принимать решение на основе MDE

Если p-value выше уровня α, то это не означает отсутствие эффекта. Эффект может и есть, но он точно не больше, чем MDE для α, β и дисперсии.

Например, вы наблюдаете 14-й день эксперимента, MDE на уровне 1%, p-value выше уровня α. Это означает, что если вы продолжите эксперимент и увидите значимые результаты эксперимента, то только для эффекта равный или больше 1%.

Принимать решение продолжать эксперимент или нет – можно, обращая внимание на MDE.

Примеры

- Прирост метрики = 2%, MDE = 1%: говорить о том, что мы видим реальный прирост в 2% можно
- Прирост метрики = **2%**, MDE = **3%**: говорить о том, что мы видим реальный прирост в 2% нельзя
- Прирост метрики = **-2%**, MDE = **3%**: говорить о том, что мы видим реальный прирост в -2% нельзя