Статистическая мощность: мифы и реальность

А. М. Шитова, к.ф.-м.н.

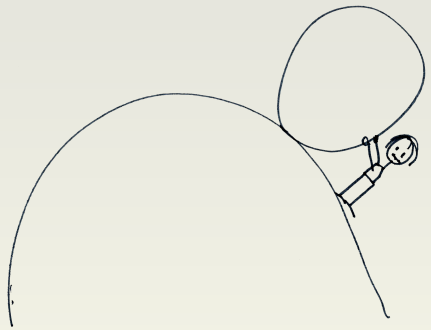
Актуальные вопросы доклинических и клинических исследований лекарственных средств, биомедицинских клеточных продуктов и клинических испытаний медицинских изделий

Санкт-Петербург, 2018

Содержание

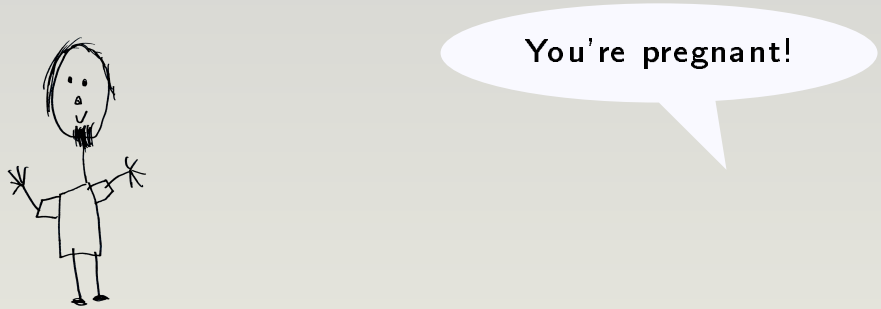
Мифы созданы для того, чтобы привлекать наше воображение

А. Камю, Миф о сизифе



Ошибки I и II рода

- Ошибка I рода — ложноположительное решение (нулевую гипотезу отвергают, когда она истинна);



- Ошибка II рода — ложноотрицательное решение (нулевую гипотезу не отвергают, когда она не является истинной);

Нулевая гипотеза H^0	верная	ложная
отклоняется	ошибка I рода	верно
не отклоняется	верно	ошибка II рода

Ошибки I и II рода

α — вероятность ошибки I рода (уровень значимости);
β — вероятность ошибки II рода (мощность критерия — $1 - \beta$);
В исследованиях биоэквивалентности нулевая гипотеза H^0 — гипотеза о небиоэквивалентности.

Решение	препараты неБЭ	препараты БЭ
H^0 отклонена	ошибка I рода	верно
H^0 отклонить не удалось	верно	ошибка II рода

Ошибка I рода связана с риском потребителя, ошибка II рода — с риском производителя.

Статистические гипотезы: noninferiority&superiority

Для непрерывных переменных[1] (Blackwelder, 1982):

H^0 : $\mu_T - \mu_R \leq \delta$

H^A : $\mu_T - \mu_R > \delta$

Для частоты достижения эффекта:

H^0 : $p_T - p_R \leq \delta$

H^A : $p_T - p_R > \delta$

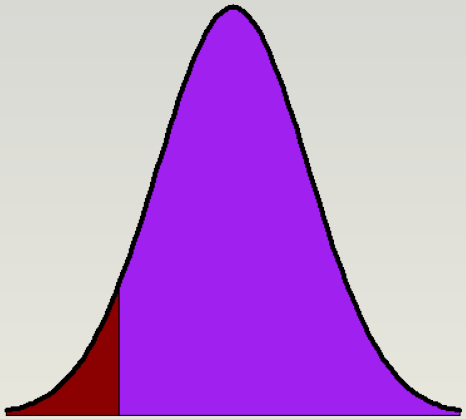
Препарат T **не менее эффективен**, чем препарат R, если нижняя граница 95%-ного ДИ для разности значений показателей эффективности больше δ ($\delta < 0$!).

Препарат T **превосходит по эффективности** препарат R, если нижняя граница 95%-ного ДИ для разности значений показателей эффективности больше δ ($\delta > 0$!).

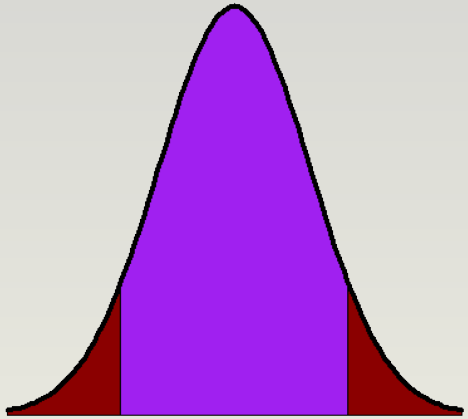
При условии доказанной не меньшей эффективности возможен переход к доказательству превосходства без необходимости введения поправок (Marcus et al., 1976).

[1] "чем больше — тем лучше"

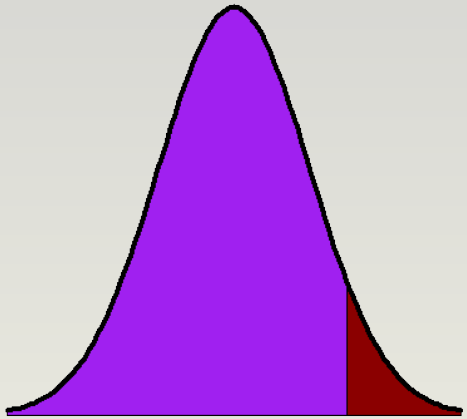
Односторонняя и двусторонняя гипотезы



ДИ: 95%, нижняя граница, односторонний



ДИ: 90%, двусторонний



ДИ: 95%, верхняя граница, односторонний

Односторонний нижний 95%-ный ДИ совпадает с нижней границей двустороннего 90% ДИ. Аналогично односторонний 97,5% ДИ совпадает с нижней границей двустороннего 95% ДИ.

Golden standard trial

Нулевые статистические гипотезы в плацебо-контролируемых исследованиях не меньшей эффективности:

$$H^{01}: \quad \mu_T - \mu_P \leq 0, \tag{1}$$

$$H^{02}: \quad \mu_T - \mu_R \leq \delta. \tag{2}$$

Effect retention approach (Pigeot at al., 2003):

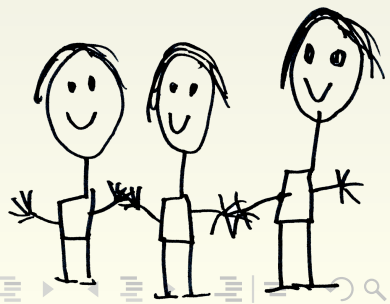
$$H^{01}: \quad \mu_R - \mu_P \leq 0, \tag{3}$$

$$H^{02}: \quad \frac{\mu_T - \mu_P}{\mu_R - \mu_P} \leq f, \tag{4}$$

где $0 \leq f \leq 1$ – установленная константа.

При $\delta = -(1-f)(\mu_R - \mu_P)$ $\qquad H^{02}: \quad \mu_T - \mu_R \leq \delta.$

Для исследований в нескольких группах необходимо введение поправок на множественные сравнения.



Статистические гипотезы: equivalence

Гипотеза эквивалентности:

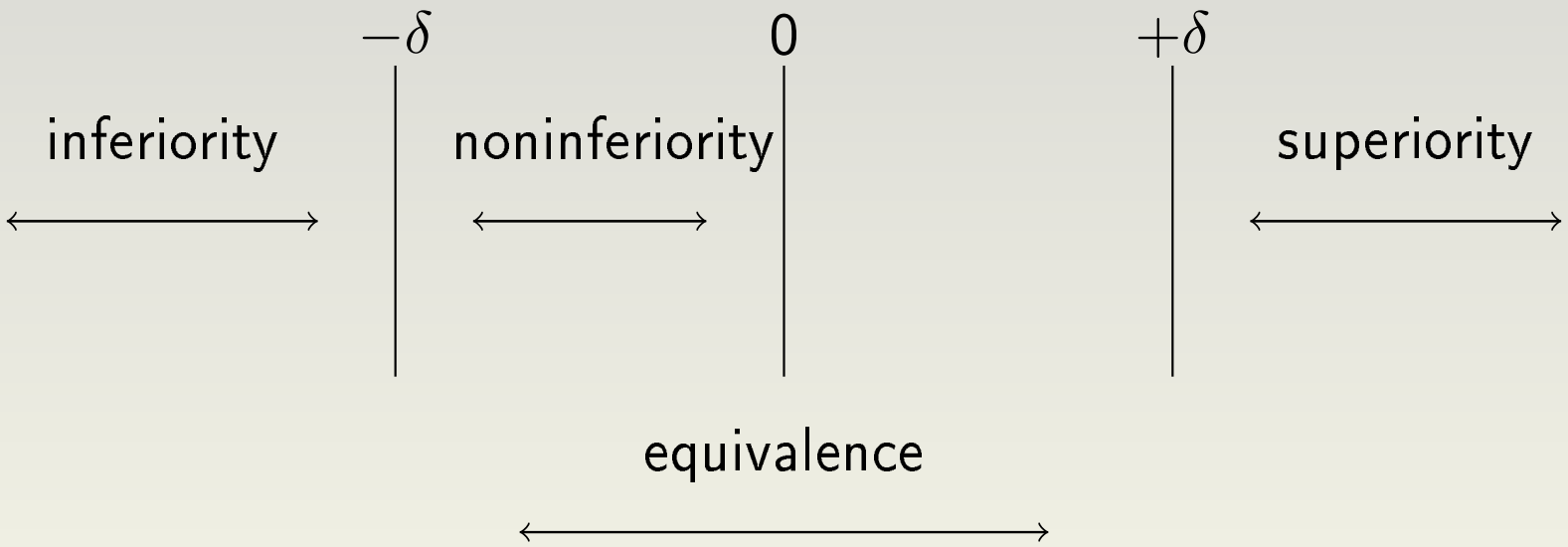
$$H^0: \quad |\mu_T - \mu_R| \leq \delta$$
$$H^A: \quad |\mu_T - \mu_R| > \delta$$

Биоэквивалентность: процедура TOST (Schuirmann, 1987), две односторонние гипотезы:

$$H^{01}: \quad \mu_T - \mu_R \leq Q_1 \qquad H^{02}: \quad \mu_T - \mu_R \geq Q_2$$
$$H^{A1}: \quad \mu_T - \mu_R > Q_1 \qquad H^{A2}: \quad \mu_T - \mu_R < Q_2$$

Модификации: scaled approach (для высоковариабельных препаратов и веществ, а также веществ с узким терапевтическим диапазоном)
Модификации дизайна: адаптивный, Group-sequential.
В некоторых случаях возможна инфляция ошибки первого рода (TIE inflation).

Статистические гипотезы



Миф 1: "обосновать можно любой размер выборки".

Проведено исследование не меньшей эффективности. Размер выборки рассчитан на основании заниженного значения SD.
Итог: среднее значение показателя эффективности тестируемого препарата **превысило** соответствующее значение референтного препарата, однако нижняя граница доверительного интервала оказалась **ниже** границы не меньшей эффективности.

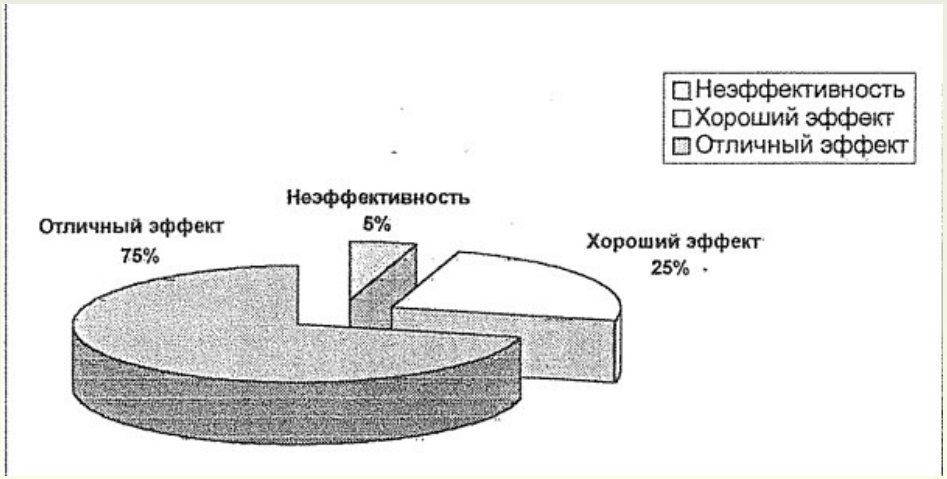
Размер выборки

Исследовалась профилактическая эффективность препарата X:
27 — группа исследуемого препарата; 22 — группа препарата контроля
Заболело 3 человека в основной группе и 7 в группе сравнения:

$$p_T = 11,11\%$$

$$p_R = 31,82\%$$

Препарат в три раза эффективнее?
95%-ный ДИ: $[-43,50; 2,08]\%$ С поправкой на непрерывность:
$[-47,62; 6,21]\%$ ДИ содержит ноль!



Размер выборки

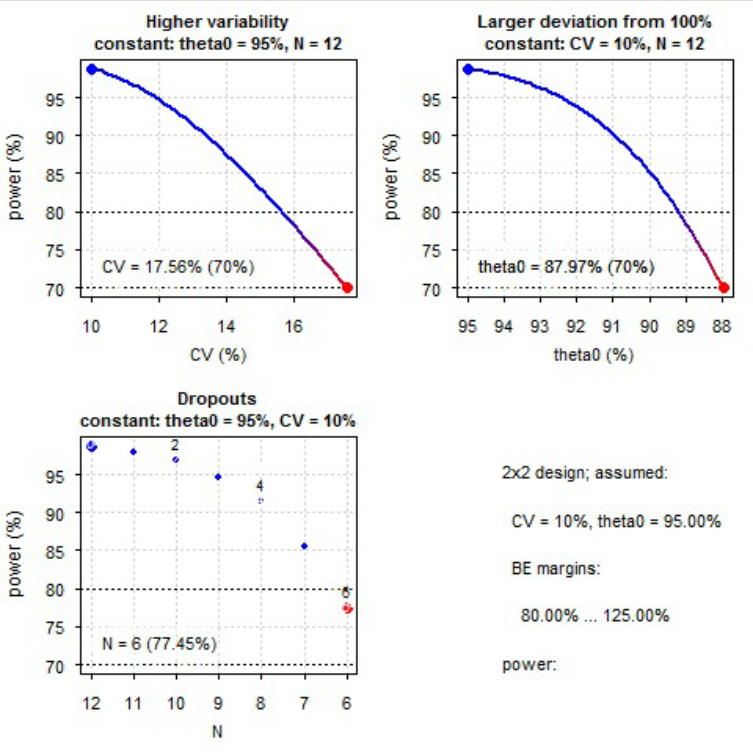
Проведено исследование первой фазы, несколько групп, размер группы 4 человека. В протоколе: оценка нормальности распределения. Спонсор требует оценить нормальность и провести сравнения в каждой группе.

Underpowered study

Коэффициент внутрииндивидуальной вариабельности C_{max} сорафениба составляет около 60%. Если в исследовании 2x2x2 будет принимать участие 24 человека, T/R=0,95, мощность составит 1%!
Для выборки в 50 человек — 24%.
Однако выборка в 24 добровольца будет достаточна для обеспечения 80%-ной мощности в исследовании 2x2x4 по методу SABE (T/R=0,95, при условии гомоскедастичности).

Overpowered study

Коэффициент внутрииндивидуальной вариабельности Cmax диеногеста около 10%. Если в исследовании 2x2x2 будут принимать участие 24 человека, а T/R=0,95, мощность составит 99,99%!



Overpowered study: форсированная биоэквивалентность

Проведено повторное исследование БЭ, первое неудачное. Точечная оценка 85%, Спонсор вводит в исследование более 100 добровольцев и доказывает БЭ.

Миф 2: "статистически=клинически"

Проведено исследование I фазы, сравнение препарата с плацебо. Доверительный интервал не включает ноль, однако величина эффекта менее 1 балла из 15 по непрерывной сложно интерпретируемой шкале. Спонсор предлагает взять большую выборку и показать эффективность препарата для $\delta = 0,5$.

Миф 3:"ретроспективная мощность"

ЕАЭС: "статистический отчёт ... анализ мощности исследования..."

Апостериорная мощность не даёт новой информации об уже проведённом исследовании. Мощность следует использовать **только** при планировании следующих исследований.

"Whereas the utility of prospective power analysis in experimental design is universally accepted, post hoc power analysis is fundamentally flawed".

Power Approach Paradox: higher observed power does not imply stronger evidence for a null hypothesis that is not rejected.

В некоторых статистических пакетах приводится значение наблюдаемой мощности, хотя оно бесполезно.

"Ретроспективная мощность"

"Rule of thumb": вероятность получить ретроспективную мощность большую или меньшую, чем целевое значение — 50%.
Пример: симуляции, $n = 10^5$ исследований ($CV = 20\%$, T/R=0.95), среднее значение 83%.

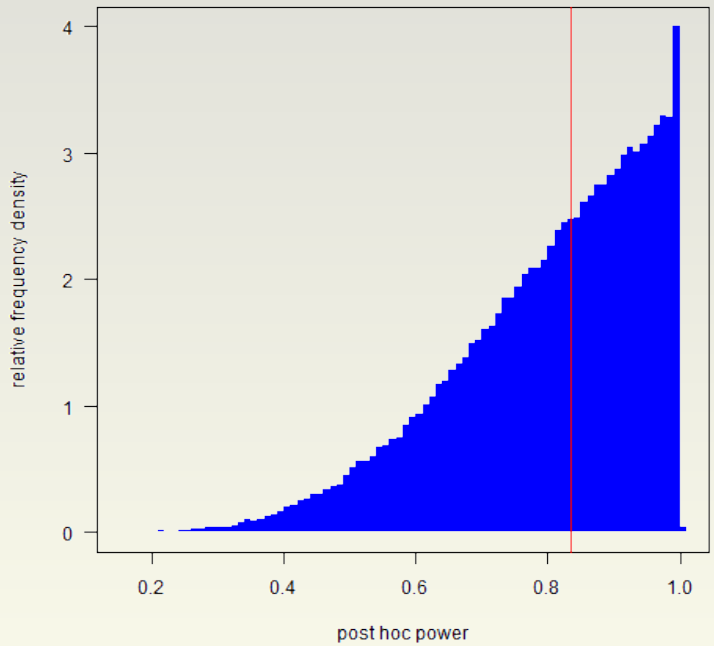
$\geq$ target : 57.60%

$\geq$ achieved : 49.41%

$\geq$ 0.90 : 31.95%

[0.95, 0.99]: 12.94%

$\geq$ 0.95 : 16.97%



в 42% исследований ретроспективная мощность оказалась меньше 80%.

"Ретроспективная мощность"

Анализ 50 реальных исследований: среднее значение 86%; в 26% ретроспективная мощность оказалась менее 80%.

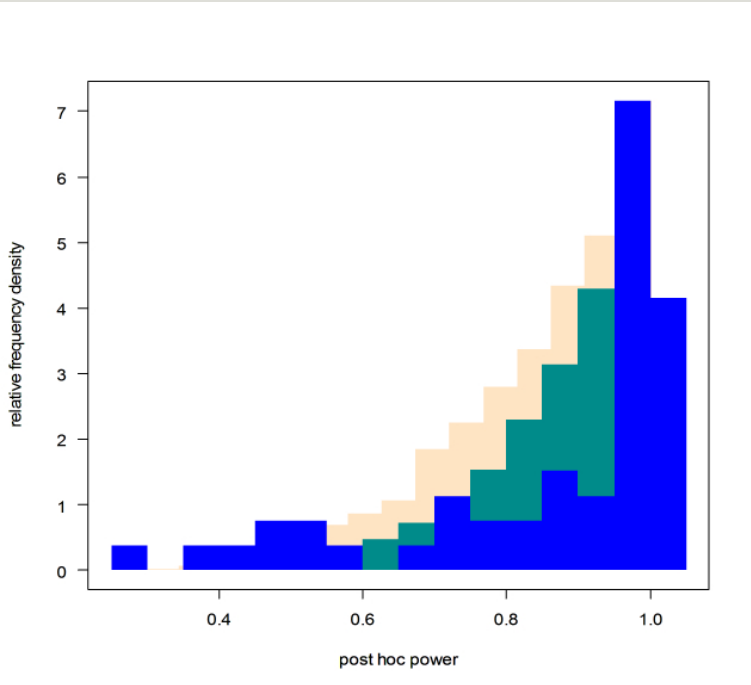
$\geq$ target : 73.58%

$\geq$ achieved : 67.92%

$\geq$ 0.90 : 62.26%

[0.95, 0.99]: 16.98%

$\geq$ 0.95 : 56.60%



A posteriori Power vs Interim Analysis

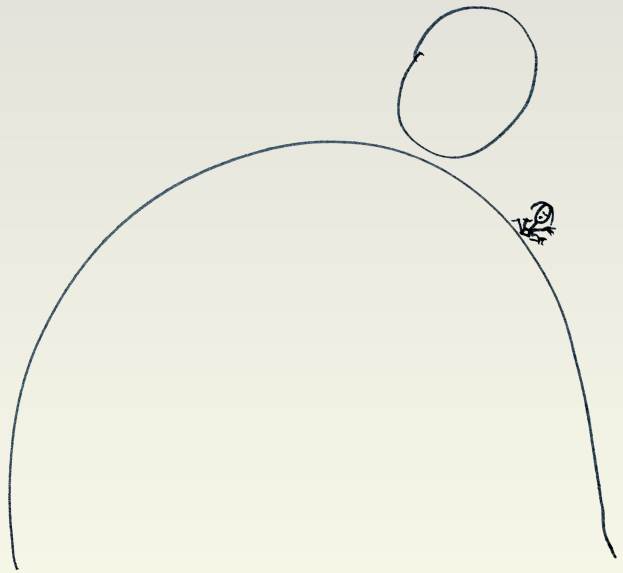
Промежуточная оценка мощности, используемая в адаптивных дизайнах и дизайнах с последовательным включением групп (group-sequential), не является апостериорной.

Пример: метод C, Potvin

По результатам 1го этапа исследования, в котором приняли участие n_1 добровольцев, промежуточная оценка мощности составила менее 80%. Был произведён перерасчёт выборки на основании значений $\alpha = 0,0294$, $1 - \beta = 0,8$ и точечной оценки 0,95. Из полученного значения N вычиталось число добровольцев с 1го этапа, итоговый объём выборки для второго этапа исследования составил: $n_2 = N - n_1$.

Заключение

Оценка размера выборки исходя из заданной мощности — не формальное действие, а краеугольный камень планирования успешного исследования.
Ретроспективная оценка мощности является статистическим нонсенсом.



Благодарю за внимание!

a.shitova@qayar.ru

Дополнение I

Представленная на слайдах информация отражает личное мнение автора. Автор доклада выражает благодарность участникам форума BEBAC (http://forum.bebac.at/)


Chow S., Shao J., Wang H.
Sample Size Calculations in Clinical Research. 2nd Ed. Chapman Hall/CRC Biostatistics Series, 2008


Sarfaraz K. Niazi
Handbook of Bioequivalence Testing, Second Edition, CRC Press, 2015


Julious S.A.
Samples Sizes for Clinical Trials . Chapman Hall/CRC, 2009


Blackwelder W.C.
'Proving the null hypothesis' in clinical trials. Controlled Clinical Trials 3: 345–353, 1982


Endrenyi L., Tothfalusi L.
Sample Sizes for Designing Bioequivalence Studies for Highly Variable Drugs, J Pharm Pharm Sci, 15 (1): 73-84, 2012


Hoenig G.M., Heisey D.M.
The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis, The American Statistician, Vol. 55, No. 1, 2001


Kupzyk K.A.

The Effects of Simplifying Assumptions in Power Analysis, Public Access Theses and Dissertations from the College of Education and Human Sciences. 106, 2011


Labes D, Schütz H.

Inflation of Type I Error in the Evaluation of Scaled Average Bioequivalence, and a Method for its Control. Submitted to Pharm Res., 2016


Marcus R. et al.

On closed testing procedures with special reference to ordered analysis of variance. Biometrika, 63:655–660, 1976


Newcombe R.G.

Interval estimation for the difference between independent proportions: comparison of eleven methods, Statist. Med. 17, 873890, 1998


Pigeot I. et al.

Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. Statistics in Medicine, 22(6):883– 899, 2003

Pocock S. J.
Group sequential methods in the design and analysis of clinical trials. Biometrika, 64(2):191–199, 1977

Potvin D. et al.,
Sequential design approaches for bioequivalence studies with crossover designs, Pharmaceut. Statist., 2007