

Последовательное тестирование

АЛЕКСАНДР САХНОВ

linkedin.com/in/amsakhnov

Staff MLE at Alibaba Group

2 сентября 2021 г.

Оглавление

- 1 Почему мы любим подглядывать?
- 2 Peeking problem — Проблема подглядывания
- 3 Последовательное тестирование. Новые схемы принятия решений
- 4 Критерий Вальда
 - Сходимость
 - Выбор границ
 - Тождество Вальда
 - Среднее количество испытаний
- 5 Сложные гипотезы

Классическое тестирование

Как раньше проводили АВ-тесты

1. Подготавливаем проведение эксперимента
2. На основе заранее определенных уровней значимости и необходимого бизнес-эффекта определяем размер тестовой и контрольной групп
3. Вычисляем продолжительность эксперимента и на этот период запускаем
4. После окончания собираем и обрабатываем данные
5. Выносим окончательный вердикт

Классическое тестирование

Как раньше проводили АВ-тесты

1. Подготавливаем проведение эксперимента
2. На основе заранее определенных уровней значимости и необходимого бизнес-эффекта определяем размер тестовой и контрольной групп
3. Вычисляем продолжительность эксперимента и на этот период запускаем
4. После окончания собираем и обрабатываем данные
5. Выносим окончательный вердикт

Всё меняется в online экспериментах

Современные технологии дают почти мгновенный доступ к данным:

- Мы можем в любой день забрать логи запущенного эксперимента
- Автоматизированные системы готовы сразу же выдать вердикт

Это имеет как свои преимущества, так и скрытые проблемы.

Типичные разговоры

Предварительные данные показывают успех

Босс: "Вариант Б показывает прекрасные результаты! Раскатим на всех!"

Сотрудник: "Вероятность ошибки велика. Нужно подождать еще."

Начало эксперимента пошло плохо

Босс: "Всё ужасно. Мы теряем деньги. Эксперимент нужно прекратить!"

Сотрудник: "Вероятность ошибки велика. Нужно подождать еще."

Банальное любопытство

Босс: "Эксперимент идет уже две недели. Что же лучше, А или Б?"

Сотрудник: "Пока мы не можем обнаружить существенной разницы."

Иногда результат очень хочется найти

Босс: "Нам просто не хватило данных чтобы обнаружить результат."

Давайте продлим эксперимент!"

Сотрудник: "Это некорректно. Нужен другой эксперимент."

Почему мы подглядываем?

Время — деньги

Проведение эксперимента затратно как само по себе, так и имеет косвенные затраты:

- Сопутствующие затраты. Если мы проводим какую-нибудь промо-акцию, то на её организацию требуются дополнительные ресурсы.
- Может отпугнуть пользователей. Если эксперимент вредит пользователям, то мы потеряем их лояльность.
- Упущеная выгода. Если есть положительный эффект, то каждый день пока мы откладываем внедрение создает упущенную выгода.

Естественное желание: завершить эксперимент как можно раньше. Это нужно в том числе для того, чтобы мы могли запустить новые эксперименты.

Нельзя просесть в уровне значимости

Конечно, можно "ускорить" эксперименты, снизив требования на уровнях значимости. Если мы будем допускать более высокие уровни ошибок I и II рода, то эксперименты пойдут быстрее.

Но это приведет к принятию неверных решений. Нам бы очень хотелось **заранее** подсмотреть результаты без изменения уровня ошибок.

Наблюдаем за доверительным интервалом

Определение доверительных интервалов

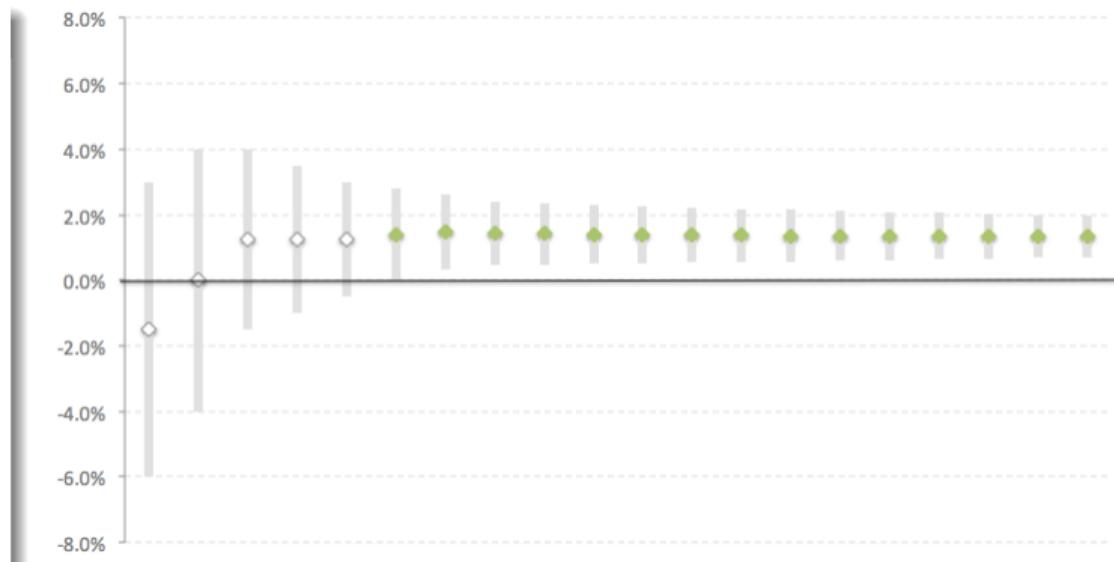
Каждый день мы можем рассчитывать доверительные интервалы на основе реально проведенного числа наблюдений. Если доверительный интервал не содержит нуля, то изменения статистически значимы.

Ход эксперимента

Для эксперимента было запланировано 20 дней.

Первые 5 дней мы не получали статистически значимый результат.

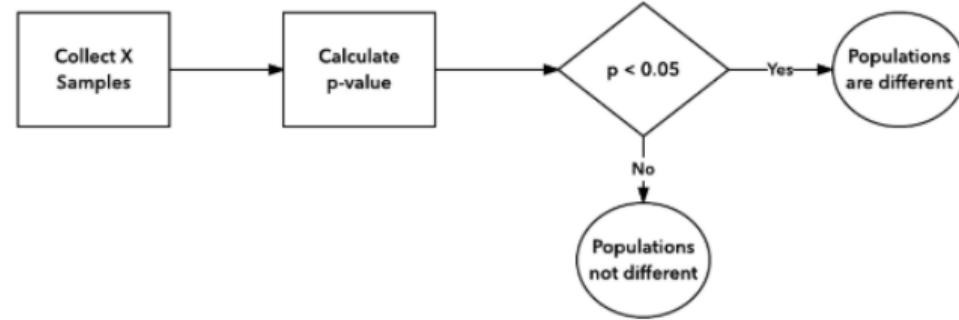
Начиная с 6-го дня и до конца эксперимента накопилось достаточно данных для статистически значимого результата.



Процедура принятия решения

Классический дизайн

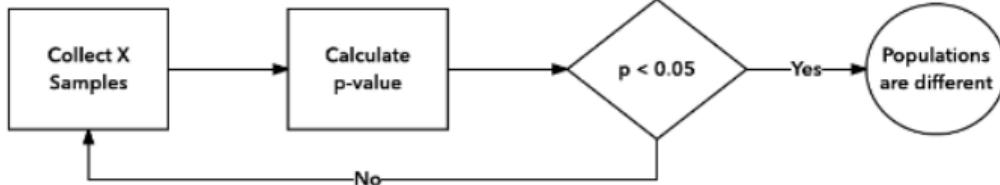
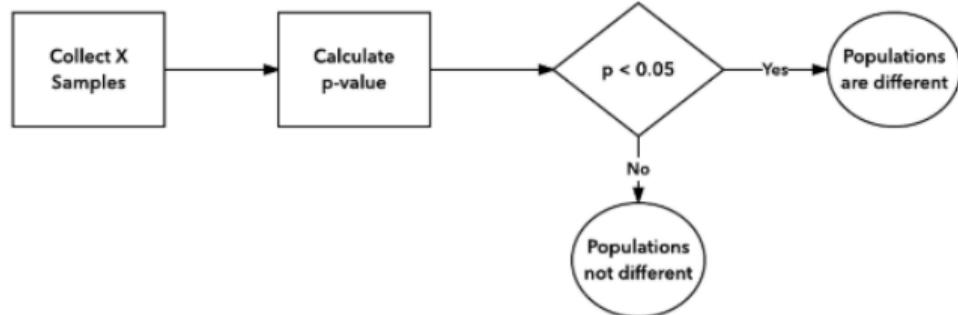
Надо помнить, что все наши тесты ориентированы на то, что мы **один раз** принимаем решение. Это решение в дальнейшем не изменяется.



Процедура принятия решения

Классический дизайн

Надо помнить, что все наши тесты ориентированы на то, что мы **один раз** принимаем решение. Это решение в дальнейшем не изменяется.



Последовательные решения

Если мы принимаем решение о приостановке или продолжении эксперимента каждый день, то схема принятия решений перестает соответствовать дизайну наших тестов. Это приводит к ошибкам.

Peeking problem — ошибаемся по-умному

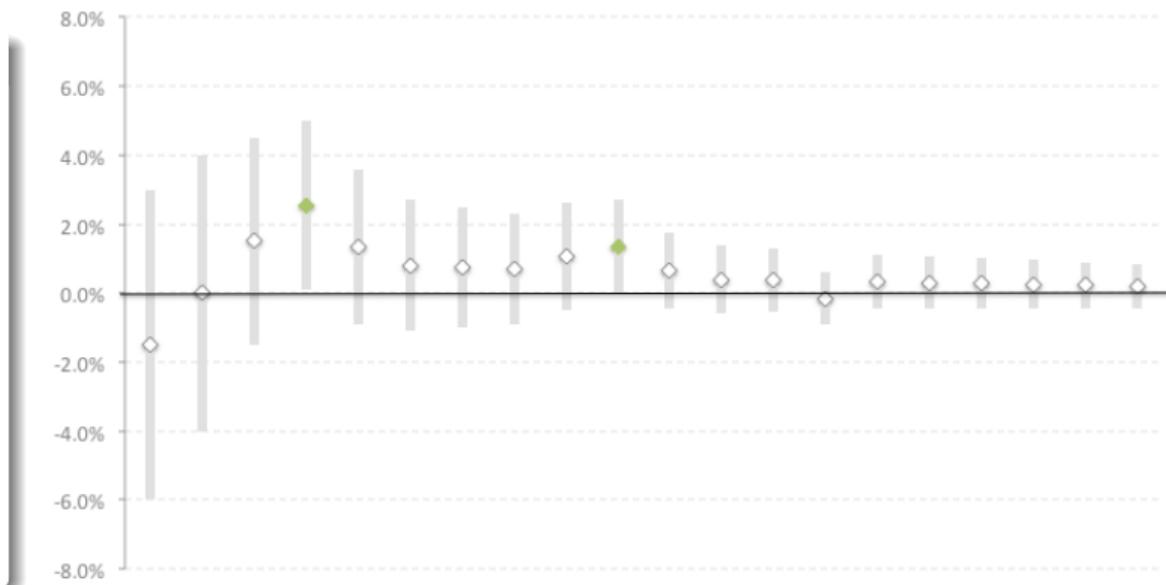
Разрешено смотреть. Принимать решения запрещено!

Ежедневные доверительные интервалы могут служить лишь справочной информацией о ходе эксперимента. Принимать на основе их значений решение до окончания эксперимента запрещено!

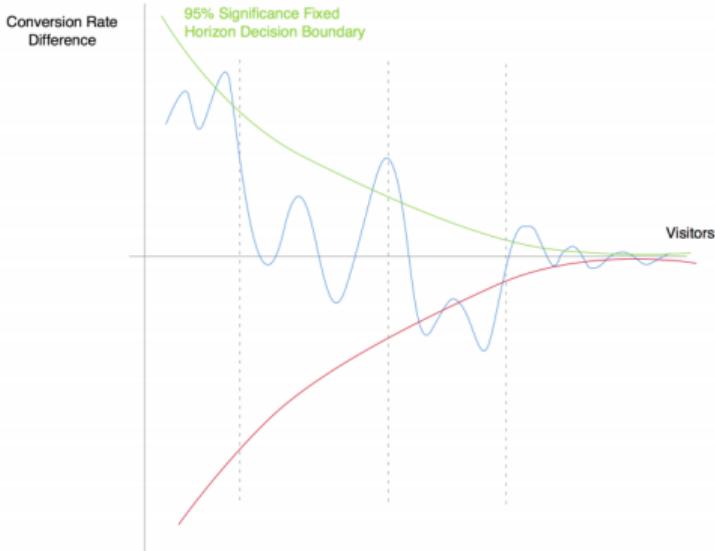
Интерпретация результата

Как бы нам ни хотелось сэкономить несколько дней, надо помнить, что дизайн эксперимента подразумевает его **проведение до конца**.

После наблюдения значимых отличий во время эксперимента, к концу эксперимента различия могут оказаться незначимыми.



Подглядывание увеличивает вероятность ошибки 1 рода



Доверительный интервал

Все доверительные интервалы посчитаны для однократного тестирования. В каждом сечении получаем **95%** шанс оказаться внутри интервала и **5%** шанс оказаться за его пределами при отсутствии эффекта.

Множественное тестирование

На проблему подглядывания можно смотреть как на задачу множественного тестирования. Последовательные принятия решений — это новые тесты. При этом результаты таких тестов связаны.

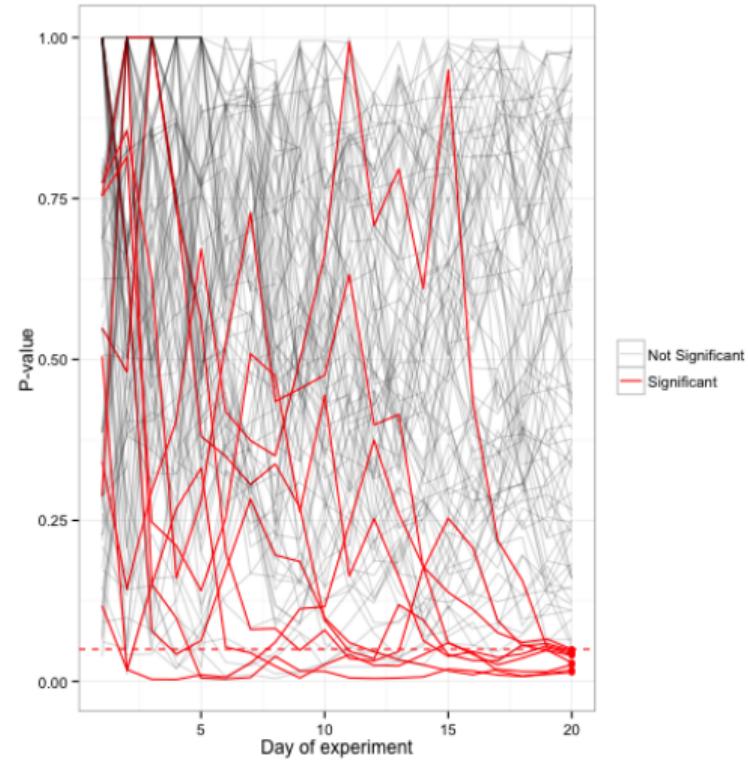
Вероятность хотя бы одной ошибки первого рода будет выше выбранного p -value.

Изменение p-value в процессе эксперимента

Остановка в конце эксперимента

Часть траекторий уходят ниже выбранного уровня в 0.05 и затем возвращаются.

В конце эксперимента мы получаем **5.34%** ложноположительных срабатываний.



Изменение p-value в процессе эксперимента

Остановка в конце эксперимента

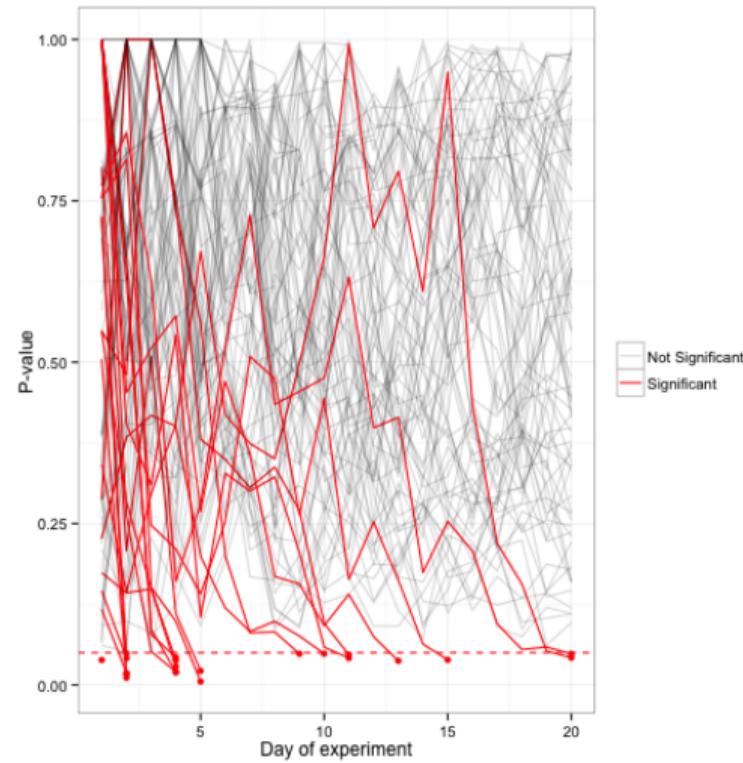
Часть траекторий уходят ниже выбранного уровня в 0.05 и затем возвращаются.

В конце эксперимента мы получаем **5.34%** ложноположительных срабатываний.

Остановка в процессе эксперимента

Если те же траектории обрывать как только значение p-value оказывается меньше 0.05, то получаем совсем другую картину.

Число ошибок первого рода равно **22.68%**.



Ошибка первого рода

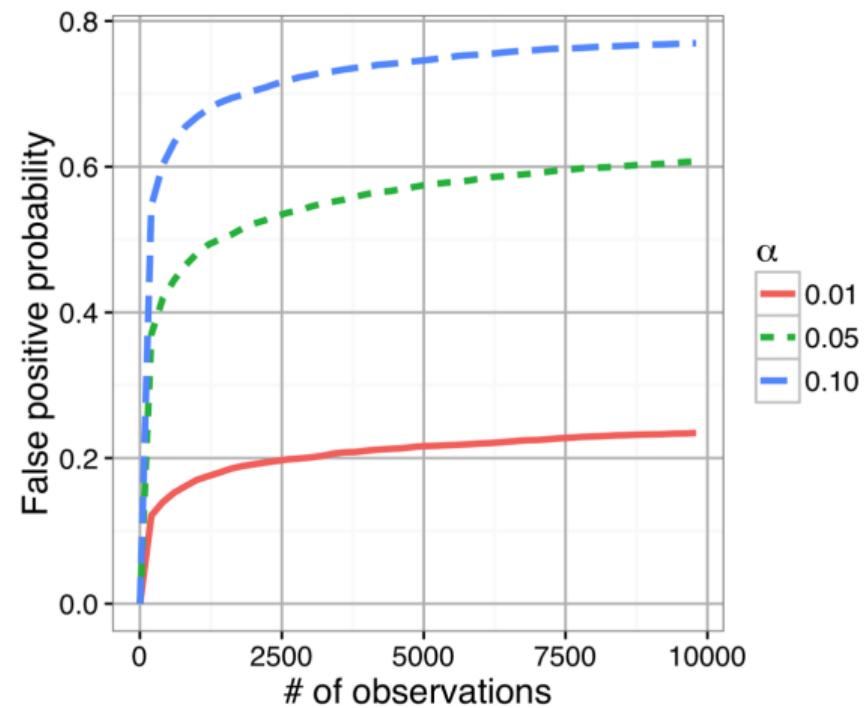
В случае тестирования множества независимых гипотез мы могли бы предъявить простую формулу для величины ошибки I рода.

Тут так же будет идти рост, но по более сложному закону, поскольку последовательные измерения связаны друг с другом.

Изменение уровней значимости

Заранее зная сколько раз мы планируем подглядывать, мы можем подсчитать истинное значение ошибки I рода. И произвести пересчет доверительных интервалов.

Необходимо изменение схемы принятия решений!



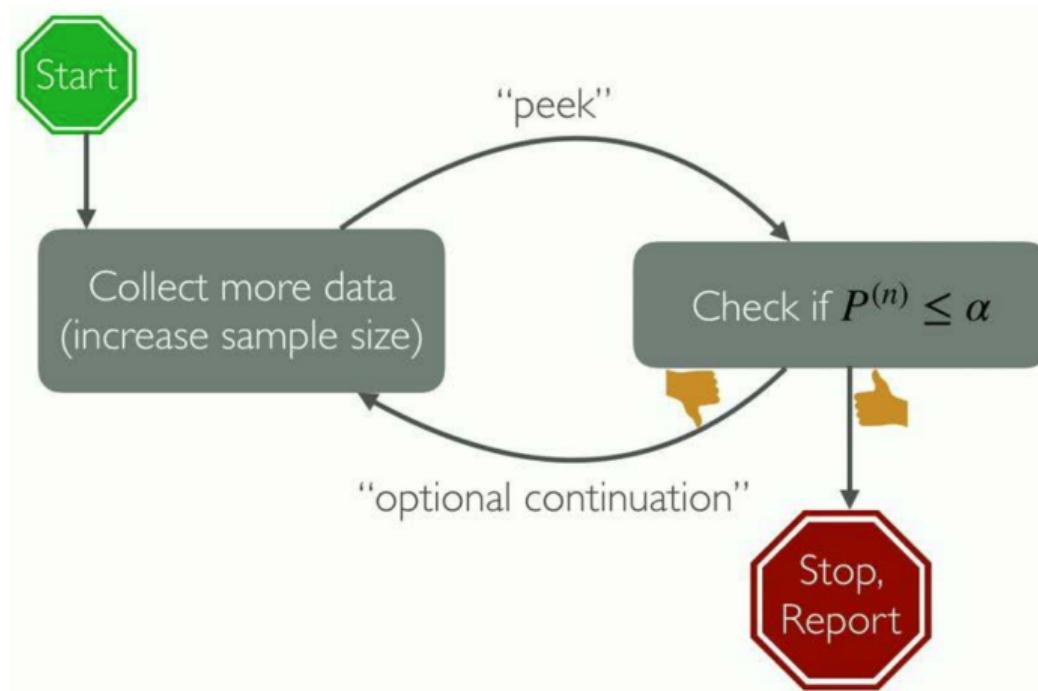
Последовательное тестирование

Новые схемы принятия решений

Подглядывать можно. И полезно.

Но такое желание должно быть до начала тестов учтено в дизайне эксперимента и в системе принятия решений.

До начала эксперимента мы должны определить сколько раз хотим подглядывать. И выбрать правила принятия решений так, чтобы общий уровень ошибки I рода соответствовал определенному значению (например, 5%).



Медицинская статистика даёт ответы

Методы последовательного тестирования в медицине

- [1] Pocock, S.J. 1977. "Group sequential methods in the design and analysis of clinical trials." *Biometrika* 64: 191–199. doi:10.2307/2335684.
- [2] O'Brien, P.C., and T.R. Fleming. 1979. "A Multiple Testing Procedure for Clinical Trials." *Biometrics* 35: 549–556. doi:10.2307/2530245.
- [3] Lan, K.K.G., and D.L. DeMets. 1983. "Discrete Sequential Boundaries for Clinical Trials." *Biometrika* 70: 659–663. doi:10.2307/2336502.

Biometrika (1977), 64, 2, pp. 191–9

Printed in Great Britain

191

Group sequential methods in the design and analysis of clinical trials

By STUART J. POCOCK

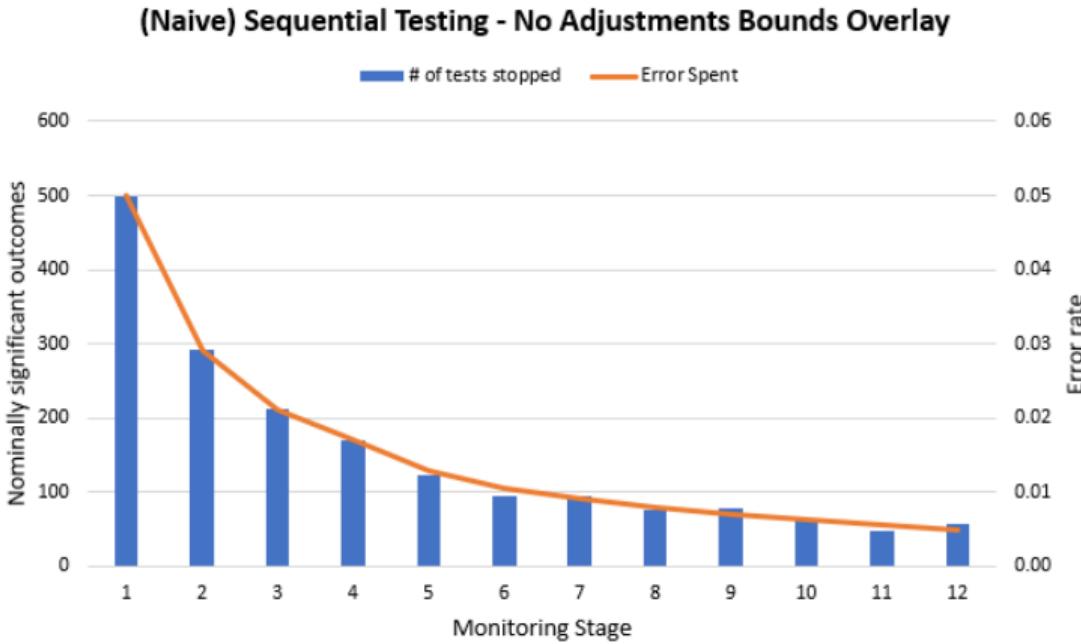
*Medical Computing and Statistics Group,
Medical School, University of Edinburgh*

Наивное тестирование

Величина ошибки

При наивном тестировании ошибка первого рода составит **18.04%**. Превышение в 3.6 раза.

Stage	z	p-value
1	1.64	.05000
2	1.64	.05000
3	1.64	.05000
4	1.64	.05000
5	1.64	.05000
6	1.64	.05000
7	1.64	.05000
8	1.64	.05000
9	1.64	.05000
10	1.64	.05000
11	1.64	.05000
12	1.64	.05000

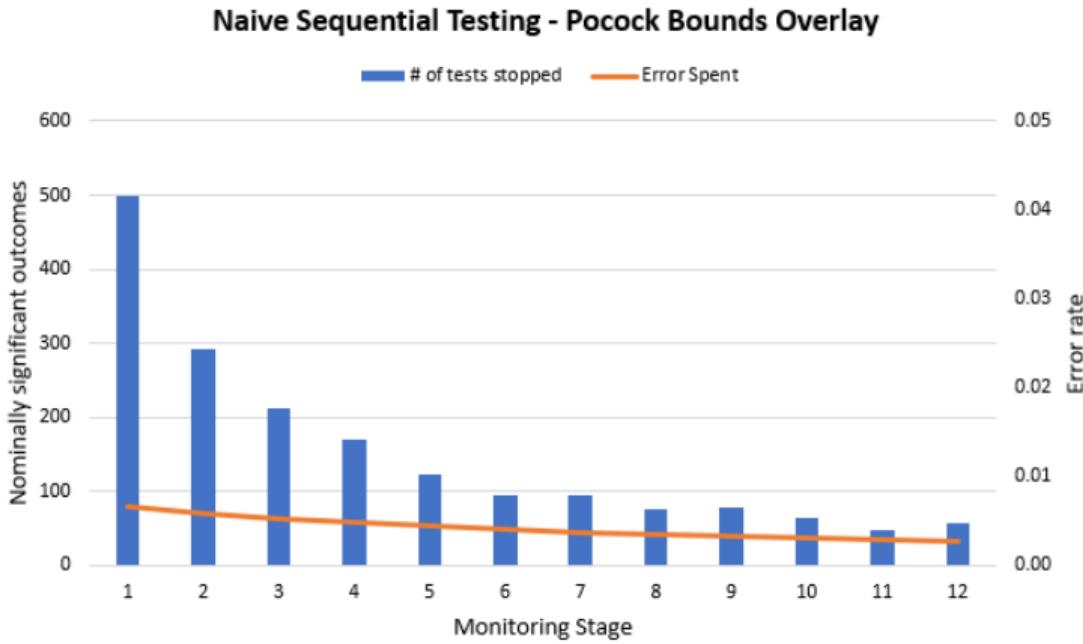


Pocock Bounds

Особенность метода

Выбирается одинаковое значение p-value. Большинство ошибок на начальных этапах.

Stage	z	p-value
1	2.30	.01070
2	2.30	.01070
3	2.30	.01070
4	2.30	.01070
5	2.30	.01070
6	2.30	.01070
7	2.30	.01070
8	2.30	.01070
9	2.30	.01070
10	2.30	.01070
11	2.30	.01070
12	2.30	.01070

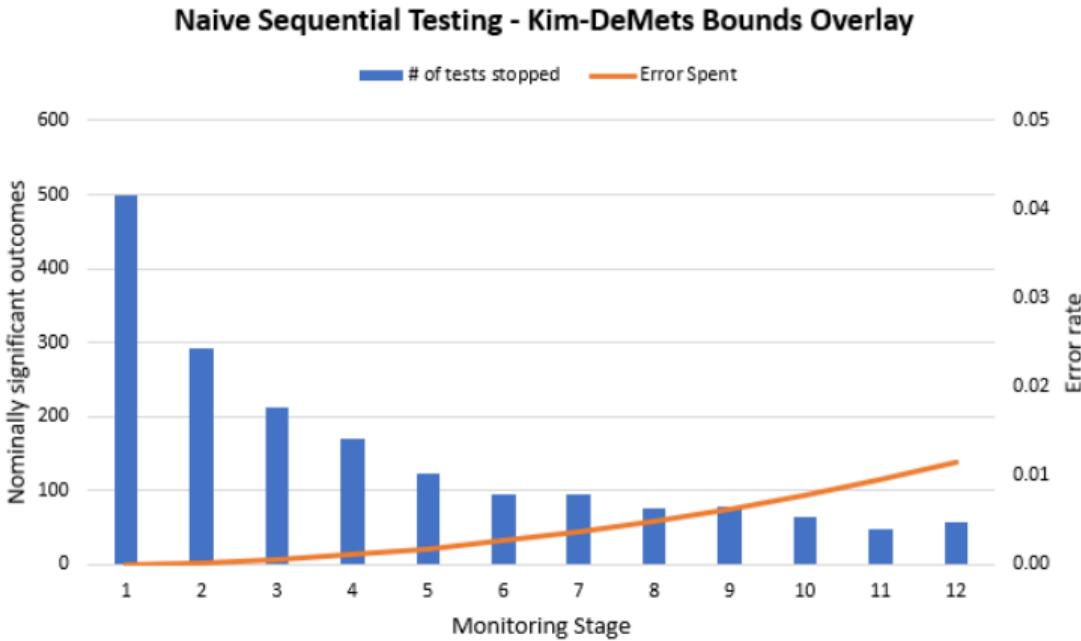


Kim-DeMets Bounds

Особенность метода

Растущие значения p-value. Большинство ошибок первого рода в конце.

Stage	z	p-value
1	4.02	.00003
2	3.53	.00021
3	3.22	.00064
4	2.98	.00144
5	2.78	.00272
6	2.61	.00453
7	2.45	.00714
8	2.3	.01072
9	2.16	.01539
10	2.03	.02118
11	1.9	.02872
12	1.78	.03754



Недостатки коррекции p-value

Что мы ускоряем?

Какой бы вариант коррекции p-value мы не придумали, ранняя остановка возможна только если мы отвергаем нулевую гипотезу.

То есть раньше времени мы можем получить только подтверждение статистической значимости эффекта или ошибку первого рода. Вряд ли мы стремимся к тому, чтобы раньше получать ошибки.

Если же эффекта нет, то нам придется ждать до конца.

Что можно улучшить?

Мы хотим иметь возможность раньше остановить тест и сказать, что эффекта нет. Если значения в двух группах совсем не различаются, то зачем ждать до конца?

Две границы

Выбор подходящей статистики

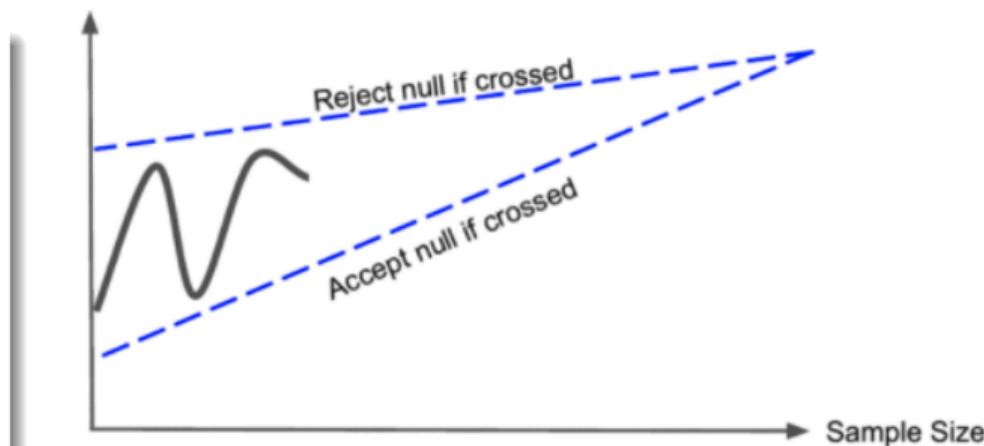
Нам нужно выбрать статистику, для которой большое значение будет свидетельствовать о необходимости отклонить нулевую гипотезу, а маленькое о том, что её не нужно отклонять.

Обычно в такой роли выступает логарифмическое правдоподобие.

Правило принятия решения

- **Не пересекли границы:**

Продолжаем тест.



Две границы

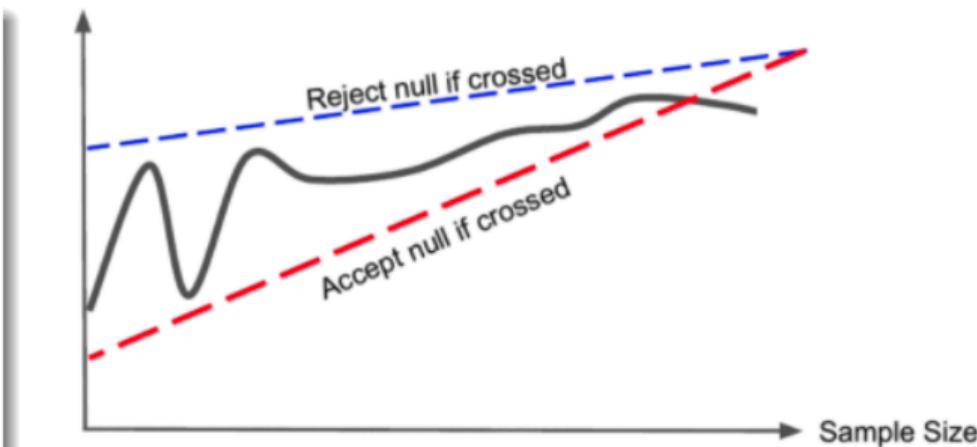
Выбор подходящей статистики

Нам нужно выбрать статистику, для которой большое значение будет свидетельствовать о необходимости отклонить нулевую гипотезу, а маленькое о том, что её не нужно отклонять.

Обычно в такой роли выступает логарифмическое правдоподобие.

Правило принятия решения

- **Не пересекли границы:**
Продолжаем тест.
- **Пересекли верхнюю границу:**
Завершаем тест. Отклоняем нулевую гипотезу.
- **Пересекли нижнюю границу:**
Завершаем тест. Не отклоняем нулевую гипотезу.



Критерий Вальда

Пусть X_i - СВ $X \sim F(x|\theta)$ в i -ом испытании, $i = 1, 2, \dots$

Гипотезы: $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$

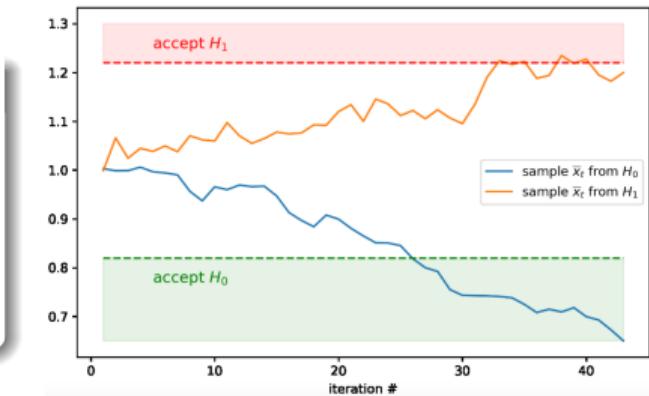
Отношение правдоподобий для первых n испытаний

$$\Lambda_T = \frac{L(X_1, \dots, X_n; \theta_1)}{L(X_1, \dots, X_n; \theta_0)} = \frac{\prod_{i=1}^n f(X_i | \theta_1)}{\prod_{i=1}^n f(X_i | \theta_0)}$$

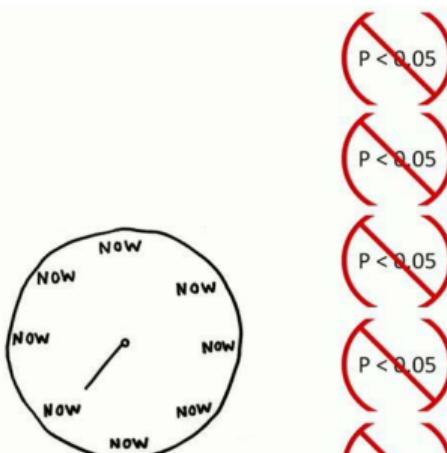
Definition (Критерий Вальда)

Зададим положительные константы $A < 1 < B$.

- если $\Lambda_T > B$, отклоняем H_0 и останавливаемся;
- если $\Lambda_T < A$, отклоняем H_1 и останавливаемся;
- иначе продолжаем собирать данные.



Долгое ожидание



After 10 people

After 284 people

After 1214 people

After 2398 people

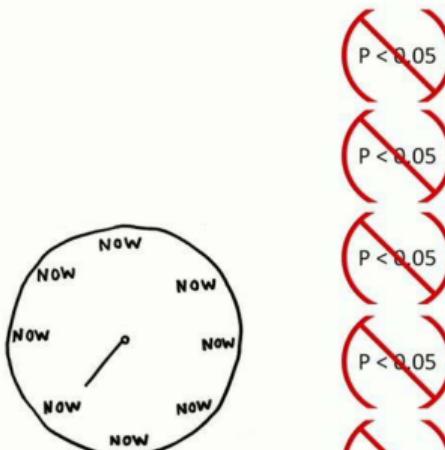
After 7224 people



After 11,219 people, STOP!

Не может ли блуждание Λ_T между A и B продолжаться бесконечно долго?

Долгое ожидание



After 10 people

After 284 people

After 1214 people

After 2398 people

After 7224 people



After 11,219 people, STOP!

Не может ли блуждание Λ_T между A и B продолжаться бесконечно долго?

Theorem

Критерий Вальда с вероятностью 1 заканчивается за конечное число шагов, то есть

$$\lim_{n \rightarrow \infty} \mathbb{P}(\nu > n | \theta_i) = 0, \quad i = \{0, 1\}$$

где ν — количество испытаний до момента остановки.

Теорема о сходимости

Пусть $Z = \ln(f_1(X)/f_0(X))$, причём $\mathbb{E}(Z|\theta_i) \neq 0$ и $\mathbb{V}(Z|\theta_i) = \sigma^2(\theta_i) > 0$, $i = \{0, 1\}$. Тогда границы критерия Вальда: $\ln A < z_1 + \dots + z_n < \ln B$, $A < 1 < B$.

Зафиксируем некоторое число r и введём СВ

$$\eta_1 = Z_1 + \dots + Z_r, \quad \eta_2 = Z_{r+1} + \dots + Z_{2r}, \quad \dots$$

Тогда событие $\{\nu > rk\}$, эквивалентное событию $\{\ln A < Z_1 + \dots + Z_i < \ln B, i \leq rk\}$, включается в событие $\{\ln A < \eta_1 + \dots + \eta_j < \ln B, j \leq k\}$, которое включается в событие $\{|\eta_j| < \Delta = \ln B - \ln A, j \leq k\}$.

СВ η_j независимы и одинаково распределены, поэтому

$$\mathbb{P}(\nu > rk|\theta) \leq \mathbb{P}(|\eta_j| < \Delta, j \leq k|\theta) = \mathbb{P}(|\eta_1| < \Delta|\theta)^k = \mathbb{P}(\eta_1^2 < \Delta^2|\theta)^k$$

Но $\mathbb{E}(\eta_1^2|\theta) \geq \mathbb{V}(\eta_1|\theta) = \sum_{i=1}^r \mathbb{V}(Z_i|\theta) = r\sigma^2(\theta)$, что можно сделать больше, чем Δ^2 , если выбрать $r > \max(\Delta^2/\sigma^2(\theta_0), \Delta^2/\sigma^2(\theta_1))$. Выбором r можно обеспечить $\mathbb{P}(|\eta_1| < \Delta|\theta) < 1$. Получаем

$$\lim_{n \rightarrow \infty} \mathbb{P}(\nu > n|\theta) = \lim_{k \rightarrow \infty} \mathbb{P}(\nu > rk|\theta) = 0$$



Выбор границ

Theorem

Границы A и B критерия Вальда силы (α, β) удовлетворяют неравенствам

$$A \geq A^* = \frac{\beta}{1 - \alpha}, \quad B \leq B^* = \frac{1 - \beta}{\alpha}$$

при этом, если границы A и B заменить их оценками A^* и B^* , то сила полученного критерия будет равна (α^*, β^*) , где

$$\alpha^* \leq \frac{\alpha}{1 - \beta}, \quad \beta^* \leq \frac{\beta}{1 - \alpha} \quad \text{и} \quad \alpha^* + \beta^* \leq \alpha + \beta.$$

Выбор границ

Обозначим \mathfrak{X}_{0n} (\mathfrak{X}_{1n}) множество тех результатов наблюдений (x_1, \dots, x_n) , для которых процедура заканчивается на n -ом шаге принятием H_0 (соответственно H_1), например

$$\mathfrak{X}_{0n} = \left\{ (x_1, \dots, x_n) : A < \frac{L_{1k}}{L_{0k}} < B, k = 1, \dots, n-1, \frac{L_{1n}}{L_{0n}} \leq A \right\}.$$

Заметим

$$\sum_{n=1}^{\infty} \mathbb{P}(\nu = n | \theta) = \sum_{n=1}^{\infty} \mathbb{P}(\mathfrak{X}_{0n} | \theta) + \sum_{n=1}^{\infty} \mathbb{P}(\mathfrak{X}_{1n} | \theta) = 1$$

В точках множества \mathfrak{X}_{1n} выполняется $L_{0n} \leq L_{1n}/B$. Тогда

$$\alpha = \mathbb{P}(H_1 | H_0) = \sum_{i=1}^{\infty} \mathbb{P}(\mathfrak{X}_{1n} | \theta_0) \leq \frac{1}{B} \sum_{i=1}^{\infty} \mathbb{P}(\mathfrak{X}_{1n} | \theta_1) = \frac{1}{B} (1 - \mathbb{P}(H_0 | H_1)) = \frac{1 - \beta}{B}$$

Аналогично получаем

$$\beta = \mathbb{P}(H_0 | H_1) = \sum_{i=1}^{\infty} \mathbb{P}(\mathfrak{X}_{0n} | \theta_1) \leq A \sum_{i=1}^{\infty} \mathbb{P}(\mathfrak{X}_{0n} | \theta_0) = A(1 - \mathbb{P}(H_1 | H_0)) = A(1 - \alpha)$$

Выбор границ

Первая пара неравенств доказана.

Рассмотрим критерий Вальда с границами A^* и B^* , и пусть α^* и β^* — его ошибки. Тогда на основании доказанного должны выполняться неравенства

$$\frac{\beta}{1-\alpha} \geq \frac{\beta^*}{1-\alpha^*}, \quad \frac{1-\beta}{\alpha} \leq \frac{1-\beta^*}{\alpha^*}$$

Отсюда имеем

$$\beta^* \leq \frac{(1-\alpha^*)\beta}{1-\alpha} \leq \frac{\beta}{1-\alpha}, \quad \alpha^* \leq \frac{(1-\beta^*)\alpha}{1-\beta} \leq \frac{\alpha}{1-\beta}$$

Складывая неравенства

$$\beta^*(1-\alpha) \leq \beta(1-\alpha^*) \quad \text{и} \quad -\alpha(1-\beta^*) \leq -\alpha^*(1-\beta)$$

получаем, что

$$\beta^* - \alpha \leq \beta - \alpha^* \quad \text{или} \quad \alpha^* + \beta^* \leq \alpha + \beta$$



Границы для гипотезы о равенстве средних

$X_1, \dots, X_n \sim N(\mu_1, \sigma^2)$, $Y_1, \dots, Y_n \sim N(\mu_2, \sigma^2)$.

Гипотезы: $H_0 : \mu_2 - \mu_1 = 0$, $H_1 : \mu_2 - \mu_1 = \theta_1 > 0$.

Разница случайных величин: $Y - X \sim N(\theta, 2\sigma^2)$, где $\theta = \mu_2 - \mu_1$.

Логарифм отношения правдоподобий

$$\ln \Lambda = \ln \frac{\prod_{i=1}^n \exp\left(-\frac{(y_i - x_i - \theta_1)^2}{4\sigma^2}\right)}{\prod_{i=1}^n \exp\left(-\frac{(y_i - x_i)^2}{4\sigma^2}\right)} = \sum_{i=1}^n \frac{2\theta_1(y_i - x_i) - \theta_1^2}{4\sigma^2} = \frac{n\theta_1}{2\sigma^2} \left((\bar{Y}^n - \bar{X}^n) - \frac{\theta_1}{2} \right)$$

Границы критерия Вальда: $\ln A < \ln \Lambda < \ln B$.

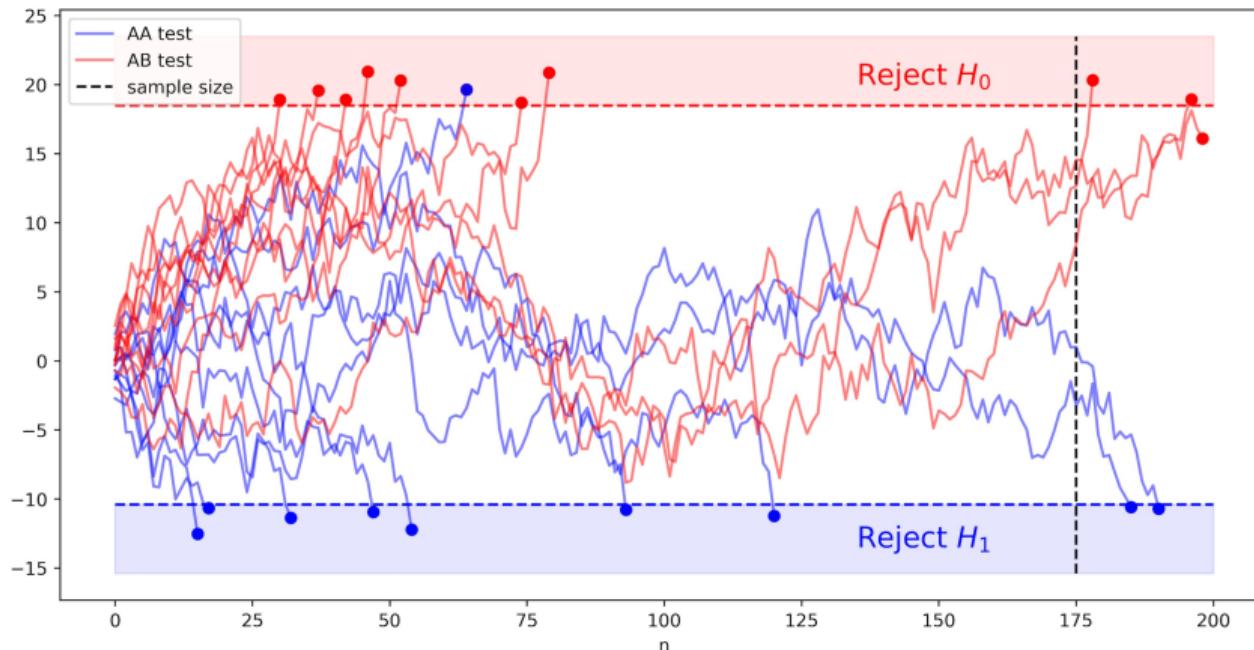
Обозначим $\hat{\theta} = \bar{Y}^n - \bar{X}^n$, тогда

$$\frac{2\sigma^2}{n\theta_1} \ln A < \hat{\theta} - \frac{\theta_1}{2} < \frac{2\sigma^2}{n\theta_1} \ln B$$

где $A \approx \frac{\beta}{1 - \alpha}$, $B \approx \frac{1 - \beta}{\alpha}$.

Пример. Проверка гипотезы о равенстве средних

$$\frac{2\sigma^2}{\theta_1} \ln \left(\frac{\beta}{1-\alpha} \right) < n \left(\hat{\theta} - \frac{\theta_1}{2} \right) < \frac{2\sigma^2}{\theta_1} \ln \left(\frac{1-\beta}{\alpha} \right)$$



Тождество Вальда

Theorem

Пусть $S_\nu = Z_1 + \dots + Z_\nu$ - ордината бружающей частицы в момент остановки, где $Z = \ln(f_1(X)/f_0(X))$. Тогда

$$\mathbb{E}(S_\nu | \theta) = \mathbb{E}(Z | \theta) \mathbb{E}(\nu | \theta)$$

Определим СВ Y_1, Y_2, \dots , где $Y_n = 1$, если решение не принято до n -го шага, и $Y_n = 0$ в противном случае. Y_n есть функция от Z_1, \dots, Z_{n-1} , и Y_n не зависит от Z_n . Тогда, можно записать

$$S_\nu = \sum_{i=1}^{\infty} Y_i Z_i$$

Учитывая независимость Y_n и Z_n , получаем

$$\mathbb{E}(S_\nu | \theta) = \sum_{i=1}^{\infty} \mathbb{E}(Y_i Z_i | \theta) = \mathbb{E}(Z | \theta) \sum_{i=1}^{\infty} \mathbb{E}(Y_i | \theta) = \mathbb{E}(Z | \theta) \sum_{i=1}^{\infty} \mathbb{P}(Y_i = 1 | \theta)$$

Событие $\{Y_n = 1\}$ эквивалентно событию $\{\nu \geq n\}$. Для любой целочисленной положительной СВ κ верно $\mathbb{E}\kappa = \sum_{i=1}^{\infty} \mathbb{P}(\kappa \geq i)$. Получаем требуемое утверждение. \square

Среднее количество испытаний

Пусть заданы значения ошибок α и β .

При малых α и β границы можно положить равными: $A = \frac{\beta}{1-\alpha}$, $B = \frac{1-\beta}{\alpha}$.

При малых α и β ширина интервала $(\ln A, \ln B)$ велика, а величина $\mathbb{E}(Z|\theta)$ от α и β не зависит и конечна. Поэтому можно положить $S_\nu \approx \ln a$, если отклоняется гипотеза H_1 . Аналогично $S_\nu \approx \ln b$, если отклоняется гипотеза H_0 .

Тогда

$$\mathbb{E}(Z|\theta_0) \mathbb{E}(\nu|\theta_0) = \mathbb{E}(S_\nu|\theta_0) \approx \ln A \mathbb{P}(H_0|H_0) + \ln B \mathbb{P}(H_1|H_0) = (1 - \alpha) \ln A + \alpha \ln B$$

Аналогично

$$\mathbb{E}(Z|\theta_1) \mathbb{E}(\nu|\theta_1) = \mathbb{E}(S_\nu|\theta_1) \approx \ln A \mathbb{P}(H_0|H_1) + \ln B \mathbb{P}(H_1|H_1) = \beta \ln A + (1 - \beta) \ln B$$

Таким образом, $\mathbb{E}_j(\nu) = \mathbb{E}(\nu|\theta_j)$ можно вычислить по следующим приближенным формулам:

$$\mathbb{E}_0(\nu) \approx \frac{(1 - \alpha) \ln \left(\frac{\beta}{1 - \alpha} \right) + \alpha \ln \left(\frac{1 - \beta}{\alpha} \right)}{\mathbb{E}_0(Z)}, \quad \mathbb{E}_1(\nu) \approx \frac{\beta \ln \left(\frac{\beta}{1 - \alpha} \right) + (1 - \beta) \ln \left(\frac{1 - \beta}{\alpha} \right)}{\mathbb{E}_1(Z)}$$

Среднее количество испытаний

Параметры эксперимента:

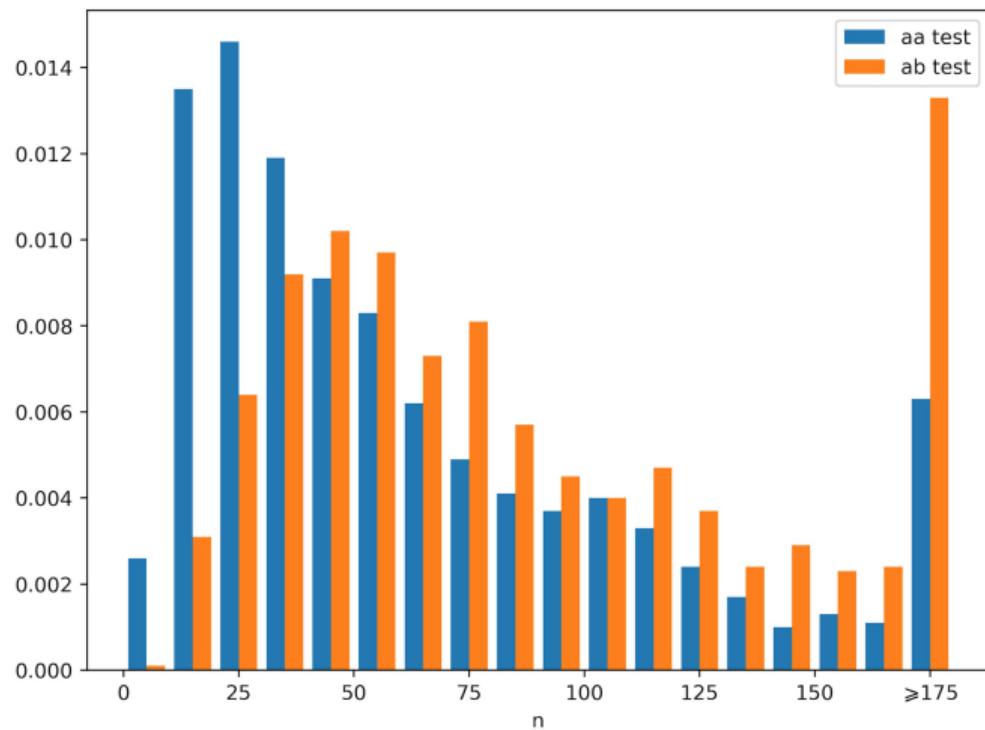
- $\alpha = 0.05$
- $\beta = 0.2$
- $\varepsilon = 3\%$

Sample Size = 175.

91% тестов остановились раньше.

Среднее кол-во итераций: 80.

Критерий Вальда является оптимальным критерием. Он требует в среднем меньше испытаний, чем критерий Неймана-Пирсона с такими же вероятностями ошибок (α, β).



Сложные гипотезы

SPRT — Sequential Probability Ratio Test

Weighted SPRT

$$\Lambda_n = \frac{\int_{\Theta_1} w_1(\theta) \prod_{i=1}^n p_\theta(X_i) d\theta}{\int_{\Theta_0} w_0(\theta) \prod_{i=1}^n p_\theta(X_i) d\theta}$$

Generalized SPRT

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_1} \prod_{i=1}^n p_\theta(X_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n p_\theta(X_i)}$$

Mixture SPRT

$$\Lambda_n = \int_{\Theta} w(\theta) \prod_{i=1}^n \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} d\theta$$

Резюме

- Мы познакомились с подходом последовательного тестирования
- Узнали об опасностях подглядывания. При неумелом подглядывании значительно растет вероятность ошибки первого рода.
- Познакомились с тем, как можно спроектировать дизайн эксперимента, чтобы сохранить преимущество ранней остановки и при этом не сломать АВ тестирование.

Дополнительные материалы

Ссылки для самостоятельного изучения

1. The Fatal Flaw of A/B Tests: Peeking
2. Unlocking Peeking in AB-Tests
3. Peeking problem – the fatal mistake in A/B testing and experimentation
4. Is Bayesian A/B Testing Immune to Peeking? Not Exactly
5. AB Testing: effect of early peeking and what to do about it
6. Simple Sequential A/B Testing
7. Tempted to Peek? Why Sequential Testing May Help
8. Peeking at AB-Tests: Why it matters, and what to do about it (Johari et al.)
9. Always Valid Inference: Continuous Monitoring of A/B Tests
10. Error Spending in Sequential Testing Explained