Статистический последовательный анализ

Fixed-horizon решения

Первое приходящее на ум решение - использовать классический подход, при котором необходимое число наблюдений оценивается заранее, а эксперимент проводится до момента сбора необходимого числа наблюдений.

Но, в нашем случае, есть очевидное желание ускорить этот процесс…

На прошлом занятии обсуждалось ускорение через манипуляцию с данными. Сегодня эта проблема будет рассмотрена с другой стороны.

Причины ускорить проведение теста

Иногда может возникнуть желание получить результат с меньшим числом показов.

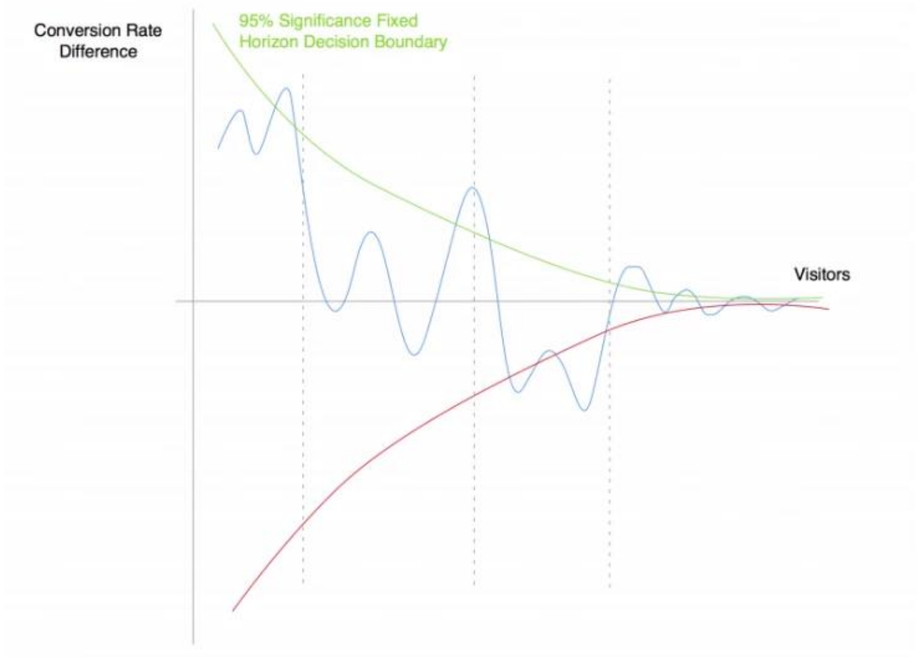
Причины могут быть следующие:

1. Каждое новое наблюдение может требовать затрат ресурсов.
2. Требуется много времени, чтобы получить нужное количество наблюдений.
3. Если эксперимент неудачный - хочется сразу его остановить, чтобы перейти к новому тесту (с новым дизайном), а если удачный - скорее применить это ко всем пользователям.

Из-за этого может возникнуть проблема подсматривания.

Проблема подсматривания

Если посчитать p-value на выборке, чей размер меньше необходимого, то результат может показаться статистически значимым, даже если на самом деле никакого эффекта нет (так называемая “peeking problem”).



Что хочется сделать?

Перед нами возникают две цели:

1. остановиться как можно раньше,
2. получить корректный результат

Возможно, мы готовы немного пожертвовать статистической мощностью теста в обмен на существенное ускорение метода.

Правдоподобие выборки

Пусть имеется выборка X = {x_1, x_2, …, x_n}, где каждый объект взят из некоторого распределения f(x, θ).

Правдоподобие параметра θ_0 - вероятность генерации выборки X распределением f(x, θ) с параметром θ = θ 0:

$$\mathcal{L}(X|\theta_0) = \prod_{i=1}^{n} f(x_i, \theta_0)$$

Часто также используют значение log-правдоподобия, поскольку его проще дифференцировать:

$$\log \mathcal{L}(X|\theta_0) = \sum_{i=1}^{n} \log f(x_i, \theta_0)$$

Тест отношения правдоподобия

Тест отношения правдоподобия - статистический тест, помогающий проверить гипотезу о значении параметра θ, использованного для генерации выборки, через значения функции правдоподобия.

Пусть имеется выборка X, сгенерированная распределением f(x, θ) с неизвестным значением параметра θ. Выдвинем гипотезу H_0: θ = θ_0 против альтернативы H_1: θ != θ_0. Статистика теста будет считаться как

$$-2\ln\Lambda(X) = -2\ln\frac{\mathcal{L}(X|\theta_0)}{\max_\theta \mathcal{L}(X|\theta)}$$

т иметь распределение хи квадрат.

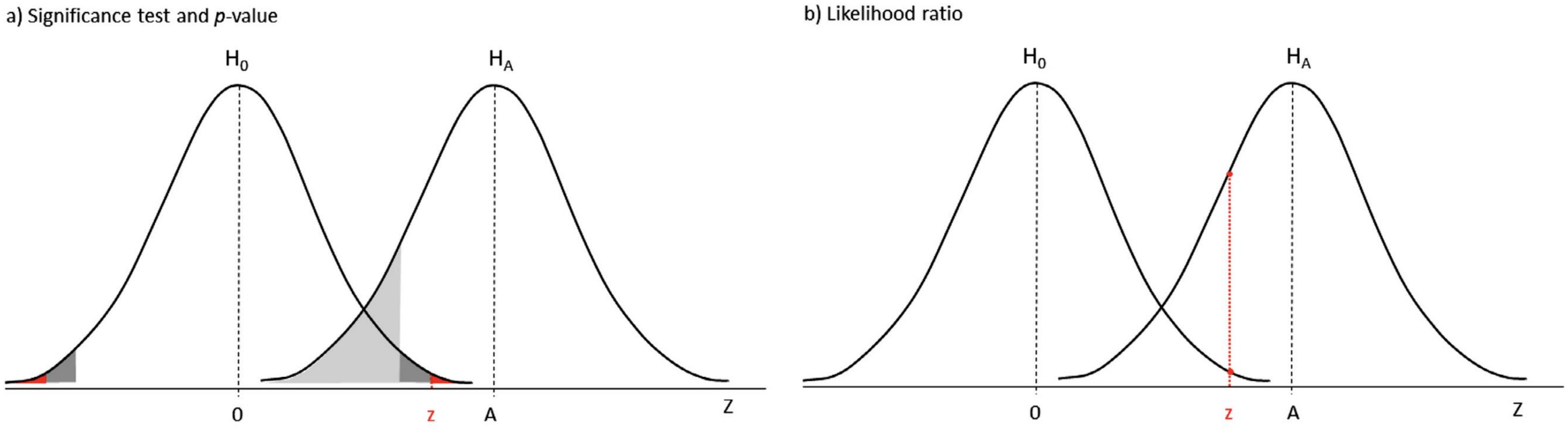
Если гипотеза H_1 простая, то в знаменателе используется L(X|θ_1).

Лемма Неймана-Пирсона

Лемма Неймана-Пирсона гласит, что при проверке простой гипотезы против простой альтернативы, методы, основанные на отношении правдоподобия, являются наиболее мощными.

Что такое мощность?

Сравнение методов



Источник: https://www.jclinepi.com/action/showPdf?pii=S0895-4356%2821%2900132-3

Статистический последовательный анализ Вальда

В своей работе 1945 года Вальд с коллегами сформулировал метод, названный статистическим последовательным анализом (Sequential Analysis) и предложенный как инструмент для более эффективного контроля промышленного качества во время Второй мировой войны.

Идея метода состоит в последовательном анализе наблюдений, в ходе которого на каждом шаге либо отвергается одна из гипотез, либо запрашивается следующее наблюдение.

Математическая постановка

1. Рассмотрим H_0: θ = θ_0 против альтернативы H_1: θ = θ_1.
2. Установим желаемые значения ошибок первого и второго рода α и β соответственно.
3. Находим пороговые значения статистики ln $\Lambda(x)$ ечивающие необходимые α и β. Назовем их ln A и ln B.
4. С каждым новым x_n-ым наблюдением:
    - Считаем значение статистики ln Λn на текущей выборке
    - Останавливаемся и принимаем H1, если ln Λn >= B
    - Останавливаемся и принимаем H0, если ln Λn <= A
    - Просим новое наблюдение x_{n+1}, если A < ln Λn < B

Нахождение граничных значений A и B

Плохая новость: точное нахождение пороговых значений A и B, обеспечивающих α и β, весьма затруднительно даже для простых распределений.

Хорошая новость: существует способ приближенного вычисления значений A и B, обеспечивающих очень близкие значения α и β.

Недостатки последовательного анализа

Если всё так хорошо и мы можем получить результат с меньшим числом наблюдений, почему этот метод не используется всегда?

1. Распределение не всегда задано заранее, что делает невозможным оценку правдоподобия.
2. Ошибка в априорных знаниях может привести к ошибочному результату, когда при анализе большой выборки эта ошибка стала бы очевидной.
3. При использовании с множественным тестированием, методы с фиксированной выборкой дают более мощный результат.
4. Требуется достаточно точно сформулированное альтернативное распределение.
5. Метод вычислительно сложный.
6. Аномальная последовательность может быстро привести к неверному результату.

Однако преимущества метода часто перевешивают его недостатки.

Двухвыборочный последовательный анализ

В случае A/B тестирования обычно приходится иметь дело с двумя выборками: тестовыми и контрольными пользователями. Оценка правдоподобия генерации двух выборок с заданной разностью средних является нетривиальной задачей. Поэтому часто выбирают какую-нибудь статистику, оценивающую разность средних, и оценивают её правдоподобие при различных средний выборок.

Пример с t-тестом

К примеру, можно воспользоваться все тем же t-тестом, чья статистика зависит от средних.

$$\Lambda = \frac{p(t^2|H_1)}{p(t^2|H_0)}$$

Источник: https://www.jstor.org/stable/2333131?origin=crossref

T-SPRT

Предположим, что мы проверяем H0: θ = 0.5 против альтернативы H1: θ = 0.7, когда истинное значение параметра равно 0.6. В этом случае SPRT может выйти из под контроля, ибо обе гипотезы одинаково вероятны.

Truncated SPRT позволяет решить проблему. В самом простом случае решением является простое ограничение сверху максимального числа наблюдений. Также можно использовать более сложные эвристики.

Источник: https://www.sciencedirect.com/science/article/abs/pii/0020025577900500

Mixture SPRT

Модификация SPRT, используемая на многих платформах для A/B тестирования и гигантами вроде убер/нетфликс.

Предположим, что имеется априорная информация о распределении параметра \theta, то есть заранее известно, что одни значения параметра более вероятны, чем другие. В этом случае значение статистики находится как

$$\Lambda_n^{H,\theta_0} = \int_\Theta \prod_{i=1}^{n} \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} h(\theta) d\theta.$$

Always Valid p-value

Метод SPRT легализует подглядывание, но лишает нас возможности объяснить наблюдаемые значения. Если p-value имеет понятную интерпретацию, то ни пороговые значения SPRT, ни сама статистика так легко не объясняются.

Для решения этой проблемы на основе mSPRT был разработан метод, который может дать интерпретируемое значение p-value.

Источник: https://arxiv.org/abs/1512.04922

Always Valid p-value

Пусть имеется две выборки: контрольные пользователи X и тестовые пользователи Y. Оба множества состоят из наблюдений, взятых из разных распределений Бернулли.

Наша цель — сравнить средние значения этих выборок и сказать, есть ли между ними существенные различия.

Always Valid p-value

Без сложной математики, только результат.

На каждом шаге n необходимо, чтобы |X_n| = |Y_n| = n, то есть чтобы число наблюдений в обеих выборках было равным.

Статистика считается по следующей формуле:

$$\Lambda_n = \sqrt{\frac{\sigma_n^2}{\sigma_n^2 + n\tau^2}} \exp\left(\frac{n^2\tau^2(\overline{Y}_n - \overline{X}_n)^2}{2\sigma_n^2(\sigma_n^2 + n\tau^2)}\right)$$

В случае распределения Бернулли: $$\sigma_n^2 = \overline{Y}_n(1 - \overline{Y}_n) + \overline{X}_n(1 - \overline{X}_n)$$

\tau - это параметр, заслуживающий отдельного исследования.

Always Valid p-value

Значения p-value вычисляются по следующей формуле:

$$p_0 = 1, p_n = \min\{p_{n-1}, 1/\Lambda_n\}$$

Эксперимент останавливается если:

1. Значение p_n < \alpha
2. Мы достигли заранее заданного числа наблюдений T.

Значение T может быть произвольным на основе наших желаний и возможностей, но рекомендуется брать T равным числу наблюдений, необходимом для Z-теста.

А можно ли как-то попроще?

Предположим, что мы держим две погнутые монетки. В ходе эксперимента мы случайным образом подбрасываем одну из монеток. Для каждой монетки мы считаем число выпавших орлов. Если на одной из монеток выпадает существенно больше орлов, чем на другой — значит монетки разные. Какая разница в числе выпавших орлов должна быть существенной для нас?

Последовательный анализ для Бернулли

Алгоритм проверки гипотезы следующий:

1. Выдвигаем гипотезу о базовом значении конверсии для контрольной группы пользователей, а также задаем минимальный наблюдаемый эффект.
2. Для выбранных параметров, находим необходимое число наблюдений N.
3. Подбрасываем монетки по очереди. Пусть T - число успехов в тестовой группе, а C — в контрольной.
4. Если T - C >= 2*sqrt(N), останавливаем тест и принимаем гипотезу о том, что конверсия в тестовой группе выросла
5. Если T + C = N, останавливаем эксперимент и говорим об отсутствии эффекта в тестовой группе.

Замечание про последовательное тестирование

Иногда результаты тестирования появляются с задержкой:

- Пользователь добавил рекомендуемый фильм в закладки и посмотрел позже.
- Скидочные купоны можно применять в течение продолжительного времени.
- Лояльность клиентов проявляется со временем.

В этом случае рекомендуется использовать агрегированные по времени данные на уровне отдельного пользователя.

Итоги занятия

Сегодня мы обсудили:

- Что такое последовательный анализ?
- Как можно значительно ускорить получение результата?
- Как считать статистику, за которой можно следить в ходе всего теста?