Последовательный анализ

Людмила Коновалова
Senior Data Analyst
Yandex GPT

План лекции

- Частотный подход к проведению экспериментов
- Немного о Байесовских алгоритмах в А/Б
- Последовательный анализ: метод Вальда
- Последовательный анализ: always valid p-values

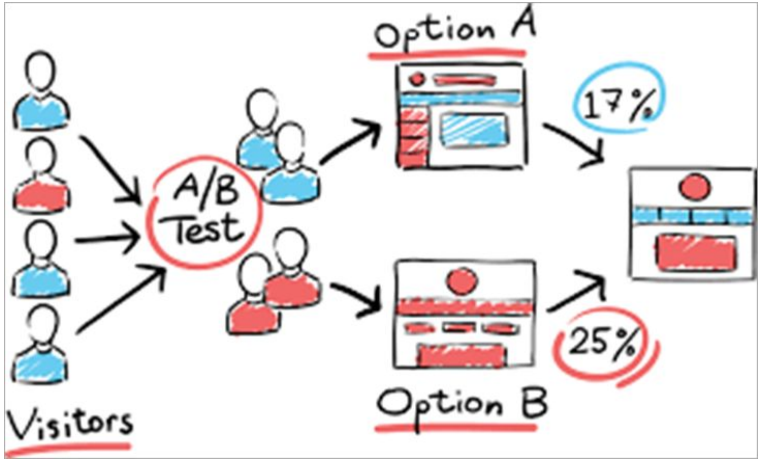
План лекции

- Частотный подход к проведению экспериментов
- Немного о Байесовских алгоритмах в А/Б
- Последовательный анализ: метод Вальда
- Последовательный анализ: always valid p-values

Вспомним базу частотных АВ

- Формулируем гипотезу
- Разбиваем пользователей на группы
- Для сравнения групп выбираем подходящую метрику
- Сравнение проводим с помощью заранее выбранного правила
- Вносим изменения в экспериментальную группу
- Какое-то время пользователи из экспериментальной группы видят нововведения
- По окончании срока теста подводим итоги

$\left.\right\}$ Дизайн



Вспомним базу частотных АВ

Вывод делаем, основываясь на расчете p-value

P-value — вероятность получить значение статистики критерия равное наблюдаемому или более экстремальное при условии справедливости нулевой гипотезы (об отсутствии разницы в сравниваемых группах)

Необходимо дождаться накопления определенного размера выборки + знать показатели дисперсии метрики*

*в случае ряда критериев - дисперсии генеральной совокупности

При этом важно помнить:

- Критерий не изменяется до конца теста
- Подглядывание не разрешено - рискуем увеличить ошибку 1 рода
- Заранее остановить эксперимент не можем
- Нельзя добавить группу в уже запущенный тест
- Дизайн определяет основные параметры эксперимента

Но на этапе дизайна оказывается, что …

при желаемых параметрах тест будет идти слишком долго

Что можно сделать?

- Пожертвовать мощностью и корректностью
- Задать минимальный порог эффекта повыше
- Поискать другие метрики
- Попробовать как-то ускорить тест

За счет чего можно ускорить тест:

- Прокси-метрики
- Сокращение дисперсии: логарифмирование метрики, работа с выбросами, CUPED, стратификация и т.д.
- Выбрать другой подход: Последовательное тестирование или Байесовские методы

План лекции

- Частотный подход к проведению экспериментов
- Немного о Байесовских алгоритмах в А/Б
- Последовательный анализ: метод Вальда
- Последовательный анализ: always valid p-values

Теорема Байеса

P(A) – априорная вероятность гипотезы A, первоначальный уровень доверия предположению A

P(A | B) – апостериорная вероятность гипотезы A при наступлении события B

P(B | A) / P(B) – как событие B помогает изменить уровень доверия к предположению A

Probability of B occurring given evidence A has already occurred

Probability of A occurring

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$$

Probability of A occurring given evidence B has already occurred

Probability of B occurring

Теорема Байеса

Апостериорная вероятность – условная вероятность случайного события при условии того, что известны апостериорные данные (полученные после опыта)

Априорная вероятность – вероятность, присвоенная событию при отсутствии знания, поддерживающего его наступление

Probability of B occurring given evidence A has already occurred

Probability of A occurring

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability of A occurring given evidence B has already occurred

Probability of B occurring

Теорема Байеса: пример

Тест на болезнь «зеленуху» имеет вероятность ошибки 0.1 (как позитивной, так и негативной), зеленухой болеет 10% населения. Какая вероятность того, что человек болен зеленухой, если у него позитивный результат теста?

$$P(\text{болен} \mid +) = \frac{P(+ \mid \text{болен})P(\text{болен})}{P(+ \mid \text{болен})P(\text{болен}) + P(+ \mid \text{здоров})P(\text{здоров})}$$

Получаем, что искомая вероятность: (0.9*0.1)/(0.9*0.1 + 0.1*0.9) = 0.5

https://alexanderdyakonov.wordpress.com/2015/10/12/%d1%84%d0%be%d1%80%d0%bc%d1%83%d0%bb%d0%b0-%d0%b1%d0%b0%d0%b9%d0%b5%d1%81%d0%b0/

Байесовские методы

В байесовской статистике неизвестные величины рассматриваются как некоторые распределения, а не как точечные оценки, в отличие от частотного подхода

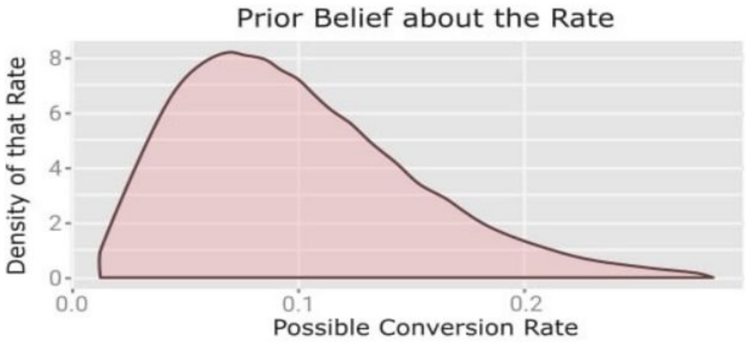
Байесовские методы для А/Б

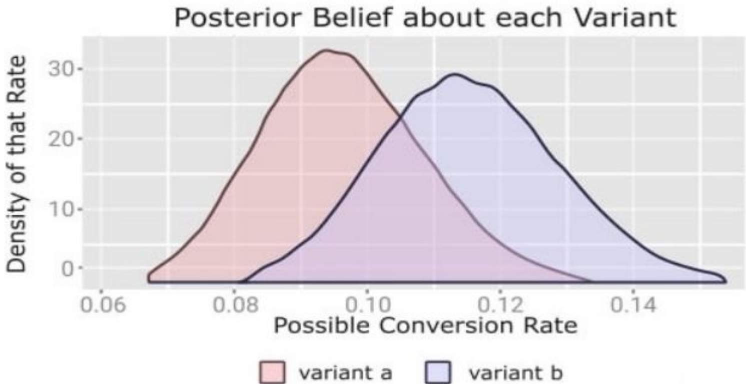
1. Полагаем некоторую априорную вероятность для наших групп
2. Выбираем критерий принятия решения (функция потерь/вероятность, что одна группа лучше другой и т.д)
3. Запускаем тест и собираем данные
4. Получив новые данные, формируем апостериорную вероятность
5. Вычисляем значение критерия принятия решения
6. Решаем, продолжать ли тест

[Байесовский подход к АБ тестированию / Хабр (habr.com)]

Байесовские методы для А/Б

По мере набора данных обновляем наши вероятности и в итоге сходимся к итоговому результату:





Байесовские методы для А/Б

- Беспроблемное введение новых вариантов в тест
- Гибкое реагирование на предпочтения пользователей и переключение на более хорошие и успешные варианты
- Нет потребности в дизайне и долгой подготовке эксперимента

Подробно про алгоритмы Байесовских бандитов для А/Б - здесь
https://habr.com/ru/companies/ods/articles/325416/

План лекции

- Частотный подход к проведению экспериментов
- Немного о Байесовских алгоритмах в А/Б
- Последовательный анализ: метод Вальда
- Последовательный анализ: always valid p-values

Классический вариант Вальда

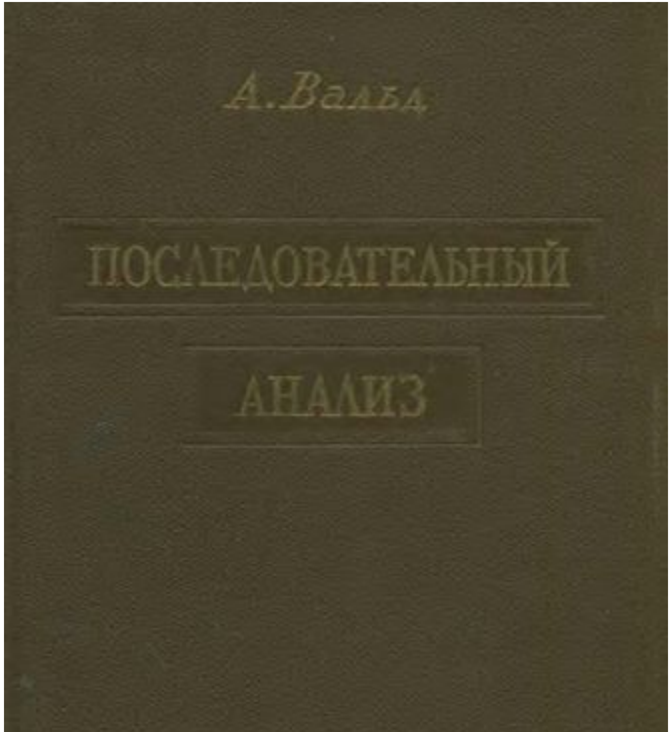
- Разработки начались в 1939 году Статистической исследовательской группой Колумбийского университета (А. Вальд, Д. Вулфовиц, А. Уоллис) для нужд военной промышленности США
- Впервые идею последовательного теста отношения вероятностей высказали экономисты М. Фридман и А. Уоллис: “*it might pay to use a test which would not be as efficient as the classical tests if a sample of exactly N were to be taken, but which would more than offset this disadvantage by providing a good chance of terminating early when used sequentially.*”

Классический вариант Вальда

Метод был рассекречен и напечатан в статье Вальда 1945 года (на русском книга была издана в 1960 году)

«Открыв наук зелёный том,
я долго плакал, а потом
его закрыл и бросил в реку.
~~*Науки вредны человеку.*~~
~~*Науки тянут нас в беду -*~~
~~*возьмёмтесь лучше за еду*~~*!»*

Д. Хармс (1933)



Классический вариант Вальда: описание

2 гипотезы:

- H_0: μ = 0 (разность средних равна 0)
- и альтернативная H_1: μ = 0,1 (разность средних отличается от нуля на фиксированное значение)

Для каждой можем построить функцию правдоподобия* (допустим, по нормальному распределению) с известной дисперсией (выборочной) и параметрами «мю» из гипотез:

$$\mathcal{L}(\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$L(\mathbf{x} \mid \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

(Она же в прологарифмированном виде)

*** Функция правдоподобия:** *Пусть плотность распределения генеральной совокупности p(x, θ) в точке x зависит от параметра θ, и у нас имеется выборка x1, x2, . . . , xn.*

Совместная плотность выборки, которая равна произведению плотностей в силу независимости наблюдений, и есть функция правдоподобия:

L(θ) = p(x1, θ) · . . . · p(xn, θ)

Классический вариант Вальда: описание

Для дальнейшего сравнения будем рассматривать отношение функций правдоподобия для обеих гипотез:

$$\Lambda(X) = \frac{\mathcal{L}(0.1, \sigma^2)}{\mathcal{L}(0, \sigma^2)} = \frac{e^{-\sum_{i=1}^{n} \frac{(X_i - 0.1)^2}{2\sigma^2}}}{e^{-\sum_{i=1}^{n} \frac{(X_i)^2}{2\sigma^2}}}$$

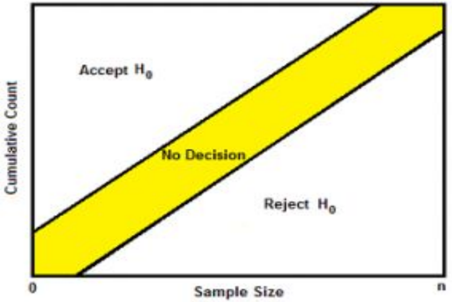
$$\log(\Lambda(X)) = \left(\sum_{i=1}^{n} \frac{(X_i)^2}{2\sigma^2}\right) - \left(\sum_{i=1}^{n} \frac{(X_i - 0.1)^2}{2\sigma^2}\right)$$

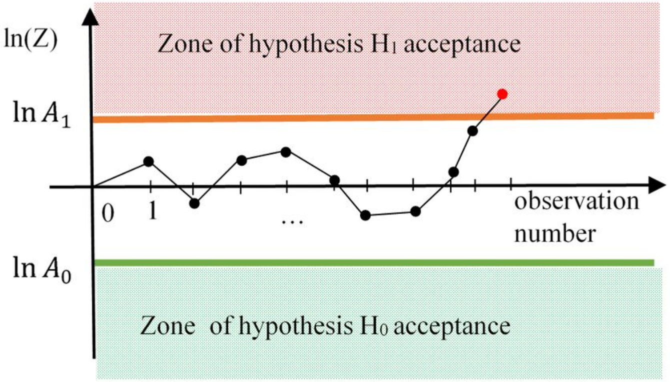
Классический вариант Вальда: описание

Логарифм этого отношения будет сравниваться со значениями a и b, являющимися границами коридора, в котором будет колебаться наш параметр отношения, и зависящими от желаемых ошибок I и II рода. При пересечении одной из границ тест можно считать завершенным и принимать/отвергать одну из гипотез.

- Если «лямбда» >= b – принимаем H_1, если «лямбда» <= a – принимаем H_0, если a < «лямбда» < b – продолжаем тест (на графике, соответственно, зеленая линия – a, красная – b)

Вальд и Вулфовиц (в совместной работе 1948 года) доказали, что тест с этими границами является наиболее мощным последовательным тестом отношения вероятностей





$$a \approx \log \frac{\beta}{1-\alpha}$$

$$b \approx \log \frac{1-\beta}{\alpha}$$

План лекции

- Частотный подход к проведению экспериментов
- Немного о Байесовских алгоритмах в А/Б
- Последовательный анализ: метод Вальда
- Последовательный анализ: always valid p-values

Always valid p-values -1

Так же есть 2 гипотезы:

- H_0: μ = 0 (разность средних равна 0)
- H_1: μ != 0 (разность средних отличается от нуля, значение не фиксируем)

Так же используются функции правдоподобия:

Но теперь берется иное отношение, учитывающее смешанное распределение для H_1 (с учетом наблюдаемого выборочного среднего sn на момент размера выборки = n отношение правдоподобия θ к θ0 принимает вид (fθ(sn)/fθ0 (sn))^n):

$$\Lambda_n^H(s_n) = \int_\Theta \left(\frac{f_\theta(s_n)}{f_{\theta_0}(s_n)} \right)^n dH(\theta)$$

В виде уравнения со всеми параметрами «лямбда» считается так:

$$\tilde{\Lambda}_n^{H,\theta_0} = \sqrt{\frac{V_n}{V_n + n\tau^2}} \exp\left\{ \frac{n^2\tau^2(\bar{Y}_n - \bar{X}_n - \theta_0)^2}{2V_n(V_n + n\tau^2)} \right\}$$

n – накопленный размер выборки, Ȳ□ и X̄□ - средние для групп B и A, v□ - сумма выборочных дисперсий групп. Параметр «тау» отвечает за дисперсию смешанного распределения и может вычисляться по-разному.

Always valid p-values -2

Параметр «тау» отвечает за дисперсию смешанного распределения и может вычисляться по-разному:

Подход из статьи (2019 Johari, Pekelis, Walsh), где полагается, что мы пользуемся нормальным распределением, а параметр b вычисляем, исходя из изначальной дисперсии, умноженной на коэффициент, корректирующий ожидаемое усечение выборки (M). Эффективность повышается за счет взвешивания в сторону больших эффектов, когда доступно мало данных, и меньших эффектов, когда имеется достаточно данных.

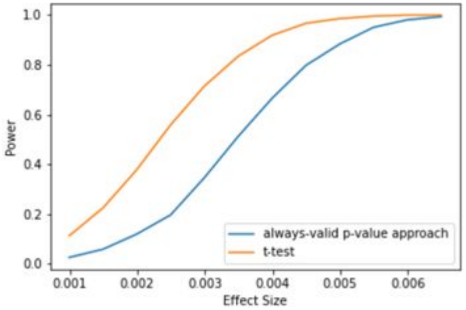
$$\gamma^{2*} = \tau^2 \frac{\Phi(-b)}{\frac{1}{b}\phi(b) - \Phi(-b)} \qquad b = \left(\frac{2\log\alpha^{-1}}{M\tau^2}\right)^{1/2},$$

- Подход из другой статьи (2019, Zhenyu Zhao et. al):
- где z - статистика Z, α - уровень значимости, а v_ctrl и v_trt - выборочные дисперсии контрольной и тестовой групп соответственно.

$$\tau = \delta^2 = z^2_{1-\alpha/2} \frac{v_{ctrl} + v_{trt}}{n}$$

Always valid p-values -3

- **Откуда получаем p-value:** 1/лямбда = наблюдаемое p-value (если оно больше 1, то полученное значение заменяется на 1)
- **Как дела с мощностью метода:** Wish Tackles Peeking with Always Valid p-values | by Qike (Max) Li | Towards Data Science (авторы проверили мощность метода)



- Проверка корректности: 1000 итераций, перемешивание групп, симуляция проведения теста последовательным методом + оценка t-тестом для сравнения

Уровень α	Количество итераций	Ошибки на t-тесте	Ошибки на always valid p-values	Ошибки на t-тесте последовательно (с подглядыванием)
0,05	1000	55	38*	134

*из 38 случаев ложного прокраса 18 приходятся на первый день теста, а 27 - на 1+2 дни теста

Плюсы и минусы метода:

Позволяет «легально» подглядывать в тест

Можно как раньше остановить, так и подержать подольше

Прост в интерпретации – известный p-value

Хорошо совместим с поправками на МПГ

Действительно контролирует ошибку в пределах заданного уровня α, при этом сокращая время проведения теста и во многих случаях почти не жертвуя мощностью

Хуже работает на выборках сильно отличающегося размера (если разбивка не 50/50, выборочная дисперсия может сильно «шуметь»)

Сильно зависит от выбора параметра "тау"

При специфике продукта может сокращать время теста не максимально (например, если нужен учет сезонности)

В первые дни проведения выше вероятность ошибки (но всё равно в пределах уровня α)

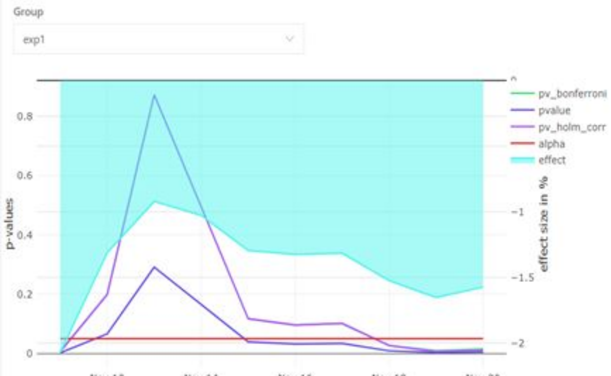
Требует выполнения ряда условий на данные, надо дорабатывать метод для других распределений

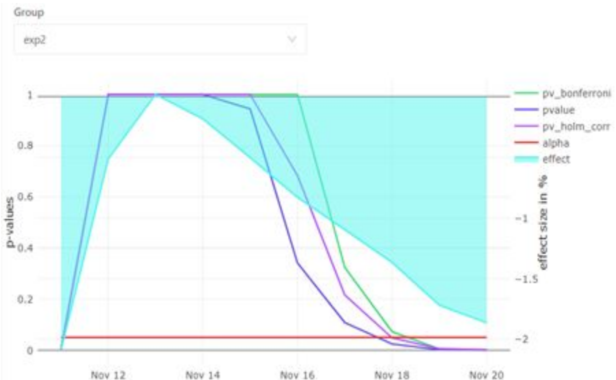
Always valid p-values в AB-тестах в Кионе

- Тест запущен на 3 недели
- Экспериментальных групп 3, контрольная 1

Видим «прокрас» метрики в 1-й день на 1-2 графиках, **как так?** Несмотря на малый размер выборки в первый день теста, эффект был достаточно велик для того, чтобы метрика «прокрасилась». Однако мы наблюдаем не менее 7-ми дней в силу специфики продукта.

В дальнейшем наблюдаем прокрасы на 8-9й день, где эффект ниже, но из-за роста размера выборки чувствительность корректируется.

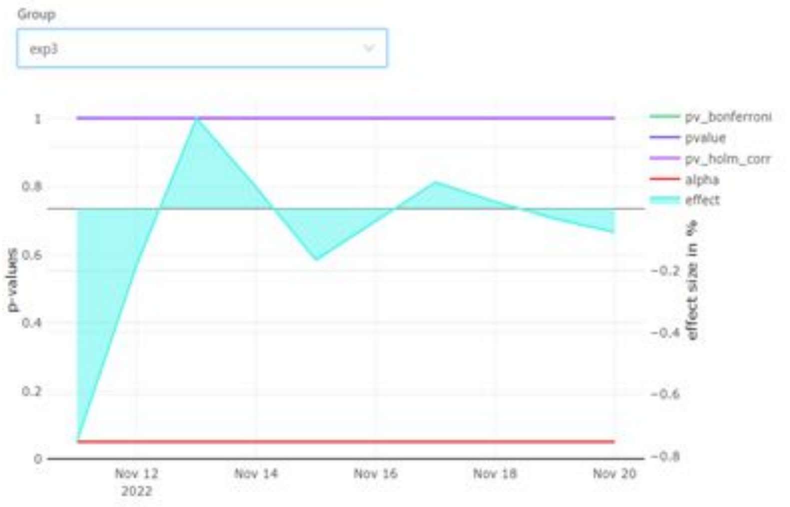




Always valid p-values в AB-тестах в Кионе

Если эффекты совсем незначительны (3 группа), «прокрас» не наступает

Благодаря совместимости с поправками на МПГ можем наблюдать сразу несколько разных вариантов коррекции p-value (для случаев, когда у нас всего 1 экспериментальная группа, они все совпадают)



Полезные ссылки

1. Статья “Always Valid Inference: Continuous Monitoring of A/B Tests” Johari, Pekelis, Walsh (2019) 1512.04922.pdf (arxiv.org)
2. Статья “Safely and Quickly Deploying New Features with a Staged Rollout Framework Using Sequential Test and Adaptive Experimental Design “ Zhenyu Zhao et. al (2019) 1905.10493.pdf (arxiv.org)
3. “Последовательный анализ» А. Вальд (здесь можно скачать Вальд А. Последовательный анализ (studmed.ru))
4. Статья с общим объяснением метода и проверкой мощности Wish Tackles Peeking with Always Valid p-values | by Qike (Max) Li | Towards Data Science
5. Хороший обзор различных техник последовательного тестирования от Spotify
6. Байесовские многорукие бандиты против A/B тестов
7. Байесовский подход к A/B тестированию
8. И снова я вещаю про последовательный анализ в Кионе :)

Спасибо за внимание!