

Лекция 6.

Проверка гипотез и A/B тестирование

Март 2019

Краткое содержание

Математический аппарат

- Проверка гипотез

- Статистические критерии

- Интерпретация результата

Примеры критериев

- Параметрические критерии

- Непараметрические критерии

Реалии А/В тестирования

- Кейс 1

- Кейс 2

- Кейс 3

- Кейс 4

- Кейс 5

Самый практичный стат. тест

Проверка гипотез

выборка: $X^n = (X_1, \dots, X_n)$, $X \sim P \in \Omega$

нулевая гипотеза H_0 : $P \in \omega$, $\omega \subset \Omega$

альтернатива H_1 : $P \notin \omega$

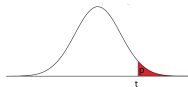
статистика: $T(X^n)$, $T(X^n) \sim F(x)$ при $P \in \omega$, $T(X^n) \approx F(x)$ при $P \notin \omega$



реализация выборки: $x^n = (x_1, \dots, x_n)$

реализация статистики: $t = T(x^n)$

достигаемый уровень значимости: $p(x^n)$ — вероятность при H_0
получить $T(X^n) = t$ или ещё более экстремальное



$p(x^n) = P(T \geq t \mid H_0)$ — обычно это значение называют p-value
Гипотеза отвергается при $p(x^n) \leq \alpha$, α — уровень значимости

Ошибки первого и второго рода

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода (False negative)
H_0 отвергается	Ошибка первого рода (False positive)	H_0 верно отвергнута

Type I error
(false positive)



Type II error
(false negative)



Ошибки первого и второго рода

Задача проверки гипотез несимметрична. Поэтому есть два параметра оценки критерия:

Корректность критерия: $P(p(T) \leq \alpha \mid H_0) \leq \alpha$

Мощность критерия: $\text{pow} = P(p(T) \leq \alpha \mid H_1) \rightarrow \max$

Интерпретация результата

Если величина p достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

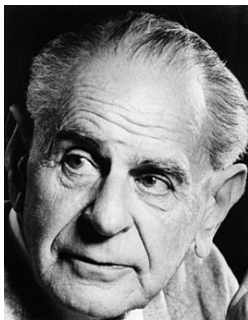
Интерпретация результата

Если величина p достаточно мала, то данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Если величина p не достаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы.

Интерпретация результата

При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы. В философии подобные рассуждения встречаются в критерии научности Карла Поппера.



Отсутствие доказательств \neq доказательство отсутствия.

Уровень значимости

Достижимый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы!

$$p = P(T \geq t \mid H_0) \neq P(H_0 \mid T \geq t)$$



Однако, отсутствие свидетельств = свидетельство
отсутствия.

Мощность

$\text{pow} = P(p(T) \leq \alpha \mid H_1)$ — вероятность отвергнуть H_0 , если верна альтернатива

Мощность критерия зависит от следующих факторов:

- ▶ размер выборки;
- ▶ размер отклонения от нулевой гипотезы;
- ▶ чувствительность статистики критерия;
- ▶ тип альтернативы.

Размер выборки

Обеспечение требуемой мощности: размеры выборки подбирается так, чтобы при размере отклонения от нулевой гипотезы не меньше заданного мощность была не меньше заданного порога.

Руководствуясь этим правилом, оценивается время A/B тестирования. Например, вы хотите показать увеличение конверсии с 0.51 до 0.53.

$$P(\text{pvalue}(T) \leq 0.05 \mid \text{конверсия больше } 0.53) \geq 0.85$$

Статистическая и практическая значимости

Харви Дент. "Либо ты умираешь героем, либо живёшь до тех пор, пока не становишься негодяем."

Статистическая и практическая значимости

Харви Дент. "Либо ты умираешь героем, либо живёшь до тех пор, пока не становишься негодяем."

На практике вероятность отвергнуть нулевую гипотезу зависит не только от того, насколько она отличается от истины, но и от размера выборки.

Статистическая и практическая значимости

- ▶ (Lee et al, 2010): за три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$). Разница в набранном весе составила 150 г.
- ▶ (Ellis, 2010, гл. 2): в 2002 году клинические испытания гормонального препарата Премарин, облегчающего симптомы менопаузы, были досрочно прерваны. Было обнаружено, что его приём ведёт к значимому увеличению риска развития рака груди на 0.08%, риска инсульта на 0.08% и инфаркта на 0.07%.
- ▶ (Kirk, 1996): если при испытании гипотетического лекарства, позволяющего замедлить прогресс ослабления интеллекта больных Альцгеймером, оказывается, что разница в IQ контрольной и тестовой групп составляет 13 пунктов, что статистически незначимо.

Параметрические критерии

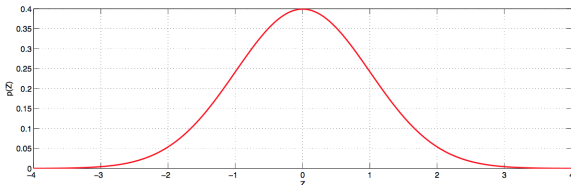


Параметрические критерии проверки гипотез допускают дополнительное знание о распределении выборок, что позволяет составлять более мощные критерии.

К сожалению, реальные данные очень редко распределены как табличные распределения. Но есть ряд популярных случаев, когда это так, их и разберём.

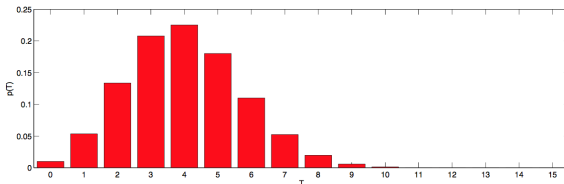
Z-критерий меток для доли

выборка: $X^n = (X_1, \dots, X_n), X \sim \text{Ber}(p)$
нулевая гипотеза: $H_0: p = p_0$
альтернатива: $H_1: p < \neq > p_0$
статистика: $Z_S(X^n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
нулевое распределение: $N(0, 1)$



Биномиальный критерий

выборка: $X^n = (X_1, \dots, X_n), X \sim \text{Ber}(p)$
нулевая гипотеза: $H_0: p = p_0$
альтернатива: $H_1: p < \neq > p_0$
статистика: $T(X^n) = \sum_{i=1}^n X_i$
нулевое распределение: $\text{Bin}(n, p_0)$



Z-критерий разности долей, независимые выборки

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim \text{Ber}(p_1)$
 $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim \text{Ber}(p_2)$
выборки независимы

Исход \ Выборка	Выборка	
	$X_1^{n_1}$	$X_2^{n_2}$
1	a	b
0	c	d
Σ	n_1	n_2

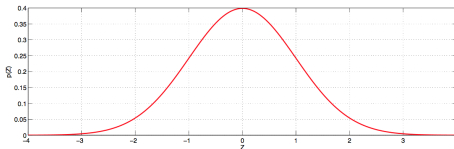
нулевая гипотеза: $H_0: p_1 = p_2$

альтернатива: $H_1: p_1 \neq p_2$

статистика: $Z(X_1^{n_1}, X_2^{n_2}) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$$P = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}, \hat{p}_1 = \frac{a}{n_1}, \hat{p}_2 = \frac{b}{n_2}$$

нулевое распределение: $N(0, 1)$



Z-критерий разности долей, связанные выборки

выборки: $X_1^n = (X_{11}, \dots, X_{1n}), X_1 \sim \text{Ber}(p_1)$
 $X_2^n = (X_{21}, \dots, X_{2n}), X_2 \sim \text{Ber}(p_2)$
выборки связанные

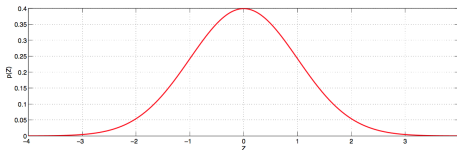
$X_1^n \backslash X_2^n$	1	0
1	e	f
0	g	h

нулевая гипотеза: $H_0: p_1 = p_2$

альтернатива: $H_1: p_1 \neq p_2$

$$\text{статистика: } Z(X_1^n, X_2^n) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{f+g}{n^2} - \frac{(f-g)^2}{n^3}}} = \frac{f-g}{\sqrt{f+g - \frac{(f-g)^2}{n}}}$$

нулевое распределение: $N(0, 1)$



Z-критерий

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$
 $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$

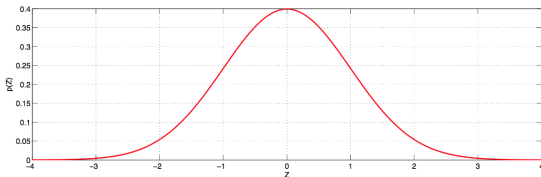
σ_1, σ_2 известны

нулевая гипотеза: $H_0: \mu_1 = \mu_2$

альтернатива: $H_1: \mu_1 < \neq > \mu_2$

статистика: $Z(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

нулевое распределение: $N(0, 1)$



t-критерий Стьюдента

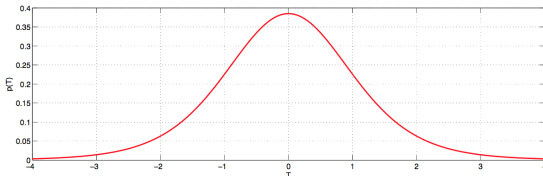
выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$
 $X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$
 σ_1, σ_2 неизвестны

нулевая гипотеза: $H_0: \mu_1 = \mu_2$

альтернатива: $H_1: \mu_1 < \neq > \mu_2$

статистика: $T(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$$

нулевое распределение: $\approx St(\nu)$



Приближение достаточно точно при $n_1 = n_2$ или $[n_1 > n_2] = [\sigma_1 > \sigma_2]$.

Достоины упоминания

1. Доверительные интервалы Вальда, Уилсона — доверительные интервалы для Z -тестов
2. Критерий Харке-Бера, критерий согласия Пирсона — проверка данных на нормальность

Непараметрические критерии



Существует специальный набор критериев, которые можно применять, не зная точного распределения выборки.

Критерий Мана-Уитни

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$

$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$

выборки независимые

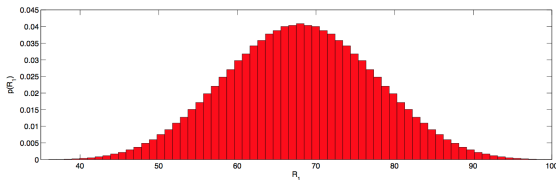
нулевая гипотеза: $H_0: F_{X_1}(x) = F_{X_2}(x)$

альтернатива: $H_1: F_{X_1}(x) = F_{X_2}(x + \Delta), \Delta \neq 0$

статистика: $X_{(1)} \leq \dots \leq X_{(n_1+n_2)}$ — вариационный ряд
объединённой выборки $X = X_1^{n_1} \cup X_2^{n_2}$

$$R_1(X_1^{n_1}, X_2^{n_2}) = \sum_{i=1}^{n_1} \text{rank}(X_{1i})$$

нулевое распределение: табличное



Критерии однородности

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$

$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$

выборки независимые

нулевая гипотеза: $H_0: F_{X_1}(x) = F_{X_2}(x)$

альтернатива: $H_1: H_0$ неверна

Критерий Смирнова

статистика: $D(X_1^{n_1}, X_2^{n_2}) = \sup_{-\infty < x < \infty} |F_{n_1 X_1}(x) - F_{n_2 X_2}(x)|$

Критерий Андерсона (модификация критерия Смирнова-Крамера-фон Мизеса)

статистика:
$$T(X_1^{n_1}, X_2^{n_2}) = \frac{1}{n_1 n_2 (n_1 + n_2)} \left(n_1 \sum_{i=1}^{n_1} (\text{rank}(X_{1i}) - i)^2 + \right. \\ \left. + n_2 \sum_{j=1}^{n_2} (\text{rank}(X_{2j}) - j)^2 \right) - \frac{4n_1 n_2 - 1}{6(n_1 + n_2)}$$

Статистики имеют табличные распределения при H_0 .

Пример АВ теста

Горные велосипеды. Распродажа 60%

В наличии более 1 000 моделей!

bike.ru

Распродажа! Sale! Rebajas! Saldi!

Не пропустите! До 1 числа продаем горные велосипеды по смешным ценам! :)

bike.ru



Офисные кресла

Скидка от 10 кресел -
15%! Скидка в
шоу-рум до 50%!
Доставка –
бесплатно!!!



Офисные кресла

Скидка от 10 кресел -
15%! Скидка в
шоу-рум до 50%!
Доставка –
бесплатно!!!

Практические вопросы

А теперь давайте разберём несколько практических моментов, с которыми каждый, кто проводит АБ тесты рано или поздно сталкивается.

В омут с головой

Вы делаете рекламу аксессуаров к заказу в интернет магазине. Вам требуется проверить наличие и оценить экономический эффект от использования вашей модели.

Прежде всего нужно ответить на следующие вопросы:

1. Что является целевой метрикой?
2. На какое увеличение мы рассчитываем?
3. Как проверять статистическую значимость результата?
4. Как оценить эффект?

Разбиение на тестовые группы

Мы разбили пользователей на две группы. Как лучше это сделать? Как проверить, что полученное разбиение — хорошее?

Разбиение на тестовые группы

Мы разбили пользователей на две группы. Как лучше это сделать? Как проверить, что полученное разбиение — хорошее?

Для разбиения обычно используют хэш от id-шника пользователя с солью. Как вы думаете, зачем соль?

Разбиение на тестовые группы

Для проверки корректности разбиение обычно сравнивают:

- ▶ Статические фичи (пол, возраст и т.п.) распределены одинаково — Критерии однородности и др
- ▶ Исторические фичи (конверсии за какой-то период, покупки и и.д.). Распределения врядли будут прям совпадать, но стоит проверить разные статистики (среднее, медианы, дисперсии)
- ▶ АА-тест. Разбиение всё равно может оказаться плохим, поэтому стоит провести АА тест, в рамках которого убедиться, что в группах нет значимых различий между целевыми метриками.

Кстати, похожий процесс происходит, когда придумывают новую метрику. Ведь нужно убедиться, что она корректно красится в АА тестах.

Сроки теста

Никто не будет запускать тест просто так, нужно заранее оценить на какой срок нужно запускать тест. Как это сделать?

Сроки теста

Никто не будет запускать тест просто так, нужно заранее оценить на какой срок нужно запускать тест. Как это сделать? Часто встречаются следующие две стратегии:

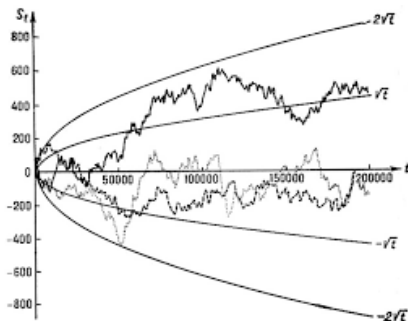
- ▶ Из соображений мощности: нужно взять такое n , чтобы $P(p(X) \leq \alpha | H_1) \geq \beta$
- ▶ Из соображений эффекта: нужно взять такое n , чтобы $p(X) \leq \alpha$ при условии, что есть эффект. Например, чтобы в Z-тесте $pvalue \leq 0.05$ при условии, что $\hat{p}_2 - \hat{p}_1 > 0.01$

Как вы думаете, какая оценка даёт бОльшие числа?

Досрочная остановка теста

Вы распланировали АБ тест. По вашим оценкам(вы использовали биномиальный тест) за 81 день отклонение изменяемой величины на 1 процент является значимым. Вы ежедневно мониторили результаты теста и через 9 дней обнаружили отклонение в 5 процентов, что является значимым для теста длиной 9 дней. Можно в этом случае досрочно завершить АБ тест?

Последовательный анализ



Как с этим жить: Нужно применять Статистический последовательный анализ (Sequential analysis). Он даёт, во-первых, корректные, а во-вторых, более мощные критерии, пользуясь дополнительным знанием о потоковости данных.

Поиск разных эффектов

Допустим, что вы проверяете средний чек, среднее число товаров в чеке, среднее число аксессуаров в чеке. Для каждой из этих величин вы составили свой критерий для проверки гипотезы о наличие эффекта. Каков уровень значимости для такой одновременной проверки гипотез?

Множественная проверка гипотез

Поскольку величина чека, число товаров в нём и число аксессуаров в нём — зависимые величины, то нельзя в точности найти уровень значимости, но можно его оценить:

$$\alpha \leq P(p_1 \leq \alpha \text{ or } p_2 \leq \alpha \text{ or } p_3 \leq \alpha | H_0) \leq \sum_i P(p_i \leq \alpha | H_0) = 3\alpha$$

Причём скорее всего самое первое неравенство строгое. Получается, что из-за того, что мы проверяем несколько гипотез, вероятность ошибки первого рода повышается. Она будет вызвана не особенностью данных, а тем, что мы несколько раз её проверяем.

Множественная проверка гипотез

Ошибка первого рода вызвана не особенностью данных, а тем, что мы несколько раз её проверяем.

Как с этим жить: Нужно применять методы Множественной проверки гипотез (гуглите Multiple comparisons problem).

Самый простой способ — уменьшить α в число гипотез раз, но есть и более сложные подходы. Есть хорошая реализация в Python — `statsmodels.sandbox.stats.multicomp.multipletests`.

Пример реального теста

Допустим, вы сравниваете конверсию пользовательской сессии в приложение в клик по рекламному блоку. Тест длится около месяца. Вы для всех сессий имеет величину 0 или 1 — сконвертировалась ли она в клик. Значимость оценивается с помощью Z-теста на выборках из вышеописанных величин.

Оцените насколько хороша такая схема теста.

Пример реального теста

Допустим, вы сравниваете конверсию пользовательской сессии в приложение в клик по рекламному блоку. Тест длится около месяца. Вы для всех сессий имеет величину 0 или 1 — сконвертировалась ли она в клик. Значимость оценивается с помощью Z-теста на выборках из вышеописанных величин.

Оцените насколько хороша такая схема теста.

Объекты выборки зависимы. Из-за этого вы будете занижать p -value! На практике эта ситуация сплошь и рядом, поэтому давайте разберёмся как решить эту проблему.

Постановка задачи

Данные

Две "выборки" вида $\{(k_i, x_i)\}_{i=1}^n$, где k_i – ключ i -го события, x_i – произвольная статистика события (может несколько чисел). События не независимы, независимость есть только между событиями разных ключей.

Постановка задачи

Данные

Две "выборки" вида $\{(k_i, x_i)\}_{i=1}^n$, где k_i – ключ i -го события, x_i – произвольная статистика события (может несколько чисел). События не независимы, независимость есть только между событиями разных ключей.

Примеры

1. Конверсия сессии в поездку, ключ – id пользователя
2. Стоимость заказа, ключ – id пользователя
3. Конверсия предложения заказа в его принятие, ключ – id водителя
4. Количество заказов в день, ключ – id водителя

Постановка задачи

Данные

Две "выборки" вида $\{(k_i, x_i)\}_{i=1}^n$, где k_i – ключ i -го события, x_i – произвольная статистика события (может несколько чисел). События не независимы, независимость есть только между событиями разных ключей.

Метод обчёта

Функция f , которая для произвольного набора $\{(k_j, x_j)\}_{j=1}^m$ находит значение статистики

Постановка задачи

Данные

Две "выборки" вида $\{(k_i, x_i)\}_{i=1}^n$, где k_i – ключ i -го события, x_i – произвольная статистика события (может несколько чисел). События не независимы, независимость есть только между событиями разных ключей.

Метод обсчёта

Функция f , которая для произвольного набора $\{(k_j, x_j)\}_{j=1}^m$ находит значение статистики

Задача

Ответить на вопрос стат. значимо ли $f(\{(k_i^{\text{test}}, x_i^{\text{test}})\}_{i=1}^{n_{\text{test}}})$ отличается от $f(\{(k_i^{\text{control}}, x_i^{\text{control}})\}_{i=1}^{n_{\text{control}}})$

Возможные решения

Задача

Ответить на вопрос стат. значимо ли $f(\{(k_i^{\text{test}}, x_i^{\text{test}})\}_{i=1}^{n_{\text{test}}})$ отличается от $f(\{(k_i^{\text{control}}, x_i^{\text{control}})\}_{i=1}^{n_{\text{control}}})$

Возможные решения

Задача

Ответить на вопрос стат. значимо ли $f(\{(k_i^{\text{test}}, x_i^{\text{test}})\}_{i=1}^{n_{\text{test}}})$ отличается от $f(\{(k_i^{\text{control}}, x_i^{\text{control}})\}_{i=1}^{n_{\text{control}}})$

Простой случай

Все ключи различные, f – среднее, медиана, гистограмма.

Тогда можно использовать обычные тесты для соответствующей нулевой гипотезы

Бакетное сэмплирование

Описание

Случайно разобьём события на бакеты (от 1 до B) при помощи некоторой хэш-функции от ключа. Для каждого бакета применим функцию f , получив B чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

Бакетное сэмплирование

Описание

Случайно разобьём события на бакеты (от 1 до B) при помощи некоторой хэш-функции от ключа. Для каждого бакета применим функцию f , получив B чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

Основные вопросы

1. Почему в итоге получается выборка?
2. Почему применение теста корректно?
3. Почему мы не потеряем в мощности теста?

Бакетное сэмплирование

Описание

Случайно разобьём события на бакеты (от 1 до B) при помощи некоторой хэш-функции от ключа. Для каждого бакета применим функцию f , получив B чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

Выборочность. Нестрого

Все зависимые события (события одного ключа) попадают в один бакет. То есть все события разных бакетов попарно независимы.

Бакетное сэмплирование

Описание

Случайно разобьём события на бакеты (от 1 до B) при помощи некоторой хэш-функции от ключа. Для каждого бакета применим функцию f , получив B чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

Корректность. Нестрого

Если эффекта нет и данных достаточно много, то $f(\text{bucket})$ будет распределено одинаково в тесте и контроле. Кстати, поэтому требуется, чтобы размер теста и контроля совпадали.

Бакетное сэмплирование

Описание

Случайно разобьём события на бакеты (от 1 до B) при помощи какой-то функции от ключа. Для каждого бакета применим функцию f , получив B чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

Мощность. Нестрого

Если все ключи уникальны, то в результате бакетирования количество объектов выборки уменьшается в $\frac{n}{B}$ раз. Но дисперсия значений выборки тоже уменьшется в $\frac{n}{B}$ раз. Эти два эффекта компенсируют друг друга и дают надежду на мощность теста.

Бакетное сэмплирование на примере

Пусть $k_i = i$, $x_i \sim \text{Bern}(p)$, то есть x_i — независимые конверсии. Метод обчёта — это среднее, а поверх бакетов применяется ttest.

В ttest-е статистика это
$$\frac{\bar{x}_{\text{test}} - \bar{x}_{\text{control}}}{\sqrt{\frac{s^2_{\text{test}}}{n_{\text{test}}} + \frac{s^2_{\text{control}}}{n_{\text{control}}}}}.$$

\bar{x} из-за бакетов не сильно изменится, так как размеры бакетов примерно одинаковые и среднее средних будет почти совпадать с исходным средним.

s^2 из-за усреднения бакетов уменьшится примерно в $\frac{n}{B}$ раз. Но и размер выборки будет не n , а B , то есть знаменатель тоже уменьшится в $\frac{n}{B}$ раз. Получается, что статистика изменится не сильно.

В полусухом остатке

1. Бакетное сэмплирование позволяет быстро сравнивать почти произвольные метрики на выборках больших размеров.
2. Но нужно быть аккуратным с размером групп и с метриками (не везде так хорошо как со средними). Требуется проводить АА тесты.
3. Если у вас небольшая выборка, то лучше использовать классические тесты
4. Практически все метрики, которые используются,
можно представить в виде $\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$ и поэтому бакетный тест
можно эффективно реализовать на MapReduce

В сухом остатке

1. Существует концепция проверки статистических гипотез.
2. Для проверки гипотез применяются различные статистические критерии.
3. Бывают параметрические и непараметрические критерии.
4. На практике машинного обучения это всё применяют для АВ и АА тестирования.
5. Существуют разные подводные камни.