

Линеаризация

АЛЕКСАНДР САХНОВ

linkedin.com/in/amsakhnov

Staff MLE at Alibaba Group

2 сентября 2021 г.

Оглавление

- 1 Ratio метрики
- 2 User average
- 3 Бутстреп
- 4 Дельта метод
- 5 Применение дельта-метода к ratio-метрикам
- 6 Многопараметрический дельта-метод
- 7 Линеаризация
- 8 Если данных очень много. Метод бакетов

Пользовательские метрики vs метрики отношения

Пользовательские метрики

$$M_A = \frac{1}{|A|} \sum_{u \in A} X(u)$$

Примеры:

- средняя выручка с пользователя
- доля пользователей с подпиской

Ratio метрики

$$\mathfrak{R}_A = \frac{\sum_{u \in A} X(u)}{\sum_{u \in A} Y(u)}$$

Примеры:

- средняя длина сессии
- доля просмотренных постов

Пользовательские метрики vs метрики отношения

Пользовательские метрики

$$M_A = \frac{1}{|A|} \sum_{u \in A} X(u)$$

Примеры:

- средняя выручка с пользователя
- доля пользователей с подпиской

Ratio метрики

$$\mathfrak{R}_A = \frac{\sum_{u \in A} X(u)}{\sum_{u \in A} Y(u)}$$

Примеры:

- средняя длина сессии
- доля просмотренных постов

Проблема

Для ratio метрик не работают:

- t-test
- методы повышения чувствительности

Наивный способ

Предложение: перейдём от пользователей к сессиям.

Длины сессий пользователя u : $t_{u1}, t_{u2}, \dots, t_{uN_u}$.

Средняя длина сессий

$$\mathfrak{R}_A = \frac{\sum_{u \in A} \sum_{i=1}^{N_u} t_{ui}}{\sum_{u \in A} N_u}$$

Наивный способ

Предложение: перейдём от пользователей к сессиям.

Длины сессий пользователя u : $t_{u1}, t_{u2}, \dots, t_{uN_u}$.

Средняя длина сессий

$$\mathfrak{R}_A = \frac{\sum_{u \in A} \sum_{i=1}^{N_u} t_{ui}}{\sum_{u \in A} N_u}$$

Недостатки такого подхода:

- Данные не будут независимыми. Сессии одного пользователя связаны, у одних сессии длиннее, у других короче.
- Группы будут разного размера, так как у пользователей может быть разное количество сессий. Это вносит дополнительную дисперсию.

Зависимость данных

Распределение цвета глаз:

50% - серый
25% - карий

20% - синий и голубой
5% - чёрный и зелёный



Зависимость данных

Распределение цвета глаз:

50% - серый
25% - карий

20% - синий и голубой
5% - чёрный и зелёный



Зависимость данных

Распределение цвета глаз:

50% - серый
25% - карий

20% - синий и голубой
5% - чёрный и зелёный



Зависимость данных

Распределение цвета глаз:

50% - серый
25% - карий

20% - синий и голубой
5% - чёрный и зелёный



Зависимость данных

Распределение цвета глаз:

50% - серый
25% - карий

20% - синий и голубой
5% - чёрный и зелёный



Зависимость данных

Распределение цвета глаз:

50% - серый
25% - карий

20% - синий и голубой
5% - чёрный и зелёный



Среднее по пользователям

Предложение: будем использовать средние длины сессий по пользователям.

Средние длины сессий пользователей: $M_u = \frac{\sum_{i=1}^{N_u} t_{ui}}{N_u}$

Среднюю длину сессий будем считать как

$$M_A^{\text{avg}} = \frac{1}{|A|} \sum_{u \in A} M_u$$

Среднее по пользователям

Предложение: будем использовать средние длины сессий по пользователям.

Средние длины сессий пользователей: $M_u = \frac{\sum_{i=1}^{N_u} t_{ui}}{N_u}$

Среднюю длину сессий будем считать как

$$M_A^{\text{avg}} = \frac{1}{|A|} \sum_{u \in A} M_u$$

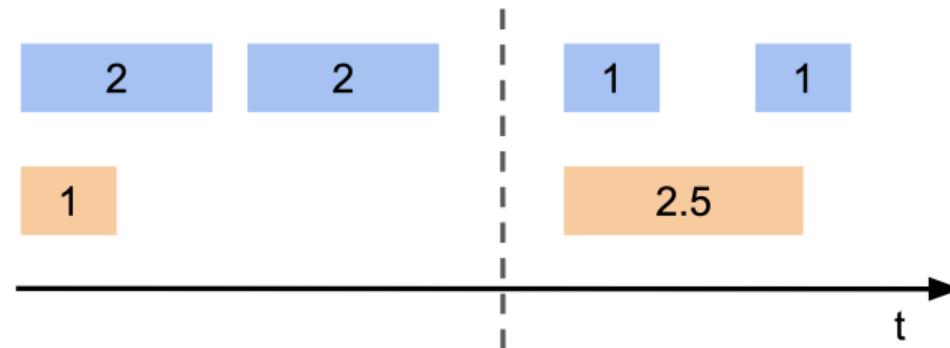
Недостатки такого подхода:

- Теряем часть информации из-за объединения сессий по пользователям. У кого-то мало сессий, у кого-то много, усредняем пользователей с одинаковыми весами, хотя у них разный вклад.
- Не сохраняется направленность. При увеличении метрики M_A , метрика M_A^{avg} может уменьшаться, и наоборот.

Направленность

Пример не сохранения направленности метрик.

Даны длины сессий двух пользователей, вычислим изменения метрик.



$$\Delta \mathfrak{R}_A = \frac{1 + 1 + 2.5}{3} - \frac{2 + 2 + 1}{3} = \frac{-0.5}{3} \approx -0.17$$

$$\Delta M_A^{\text{avg}} = \frac{1 + 2.5}{2} - \frac{2 + 1}{2} = \frac{0.5}{2} = 0.25$$

Бутстреп

1. Генерируем случайные выборки пользователей из пилотной и контрольной групп (A_{bs}, B_{bs});
2. Для каждой пары сэмплов считаем разность их ratio метрик

$$\Delta = \mathfrak{R}_{B_{bs}} - \mathfrak{R}_{A_{bs}} = \frac{\sum_{u \in B_{bs}} X(u)}{\sum_{u \in B_{bs}} Y(u)} - \frac{\sum_{u \in A_{bs}} X(u)}{\sum_{u \in A_{bs}} Y(u)}$$

3. Повторяем первые два шага, чтобы набрать достаточно данных;
4. По полученному множеству дельт оцениваем значимость различия от нуля разности метрик.

Получаем корректную оценку.

Бутстреп

- Генерируем случайные выборки пользователей из пилотной и контрольной групп (A_{bs}, B_{bs});
- Для каждой пары сэмплов считаем разность их ratio метрик

$$\Delta = \mathfrak{R}_{B_{bs}} - \mathfrak{R}_{A_{bs}} = \frac{\sum_{u \in B_{bs}} X(u)}{\sum_{u \in B_{bs}} Y(u)} - \frac{\sum_{u \in A_{bs}} X(u)}{\sum_{u \in A_{bs}} Y(u)}$$

- Повторяем первые два шага, чтобы набрать достаточно данных;
- По полученному множеству дельт оцениваем значимость различия от нуля разности метрик.

Получаем корректную оценку.

Недостатки такого подхода:

- Долгое время вычислений.
- Нет пользовательской метрики.

Информация Фишера

Информация Фишера

Информация Фишера — математическое ожидание квадрата относительной скорости изменения условной плотности вероятности $p(x|\theta)$.

Пусть $f(\theta, X_1, \dots, X_n)$ — плотность распределения для данной статистической модели. Тогда если определена функция

$$I_n(\theta) = \mathbb{E}_\theta \left(\frac{\partial L(\theta, X_1, \dots, X_n)}{\partial \theta} \right)^2, \quad L = \sum_{i=1}^n \ln f(\theta, X_i)$$

где $L(\theta, X_1, \dots, X_n)$ — лог. функция правдоподобия, а \mathbb{E}_θ — матожидание по X при данном θ , то она называется *информацией Фишера* при n независимых испытаниях.

Вычисление информации Фишера

Если $\ln f(x; \theta)$ дважды дифференцируем по θ , то информацию Фишера можно переписать как

$$I_n(\theta) = \mathbb{E}_\theta \left(\frac{\partial L(\theta, X)}{\partial \theta} \right)^2 = -\mathbb{E}_\theta \left(\frac{\partial^2 L(\theta, X)}{\partial \theta^2} \right)$$

при этом должно выполняться условие регулярности модели:

$$\mathbb{E}_\theta \left(\frac{\partial L(\theta, X_1, \dots, X_n)}{\partial \theta} \right) = 0$$

Фишеровское количеством информации в одном наблюдении:

$$I_i(\theta) = \mathbb{E}_\theta \left(\frac{\partial \ln f(\theta, X_i)}{\partial \theta} \right)^2.$$

Для n независимых испытаний $I_n(\theta) = nI(\theta)$.

Информация Фишера

Theorem

Пусть $se = \sqrt{\mathbb{V}(\hat{\theta}_n)}$. При соответствующих условиях регулярности

1. $se \approx \sqrt{1/I_n(\theta_*)}$ и $\frac{\hat{\theta}_n - \theta_*}{se} \rightsquigarrow \mathcal{N}(0, 1)$
2. Пусть $\hat{se} = \sqrt{1/I_n(\hat{\theta}_n)}$. Тогда $\frac{\hat{\theta}_n - \theta_*}{\hat{se}} \rightsquigarrow \mathcal{N}(0, 1)$

Дельта метод. МЛЕ подход

- Пусть $\tau = g(\theta)$, где g гладкая функция.
- МЛЕ оценка для τ имеет вид $\hat{\tau}_n = g(\hat{\theta}_n)$.
- Как распределена $\hat{\tau}_n$?

Theorem

Если $\tau = g(\theta)$ и при этом g дифференцируема, $g'(\theta) \neq 0$, то

$$\frac{\hat{\tau}_n - \tau_*}{\widehat{se}(\hat{\tau}_n)} \rightsquigarrow \mathcal{N}(0, 1),$$

где $\hat{\tau}_n = g(\hat{\theta}_n)$ и $\widehat{se}(\hat{\tau}_n) = |g'(\hat{\theta})| \widehat{se}(\hat{\theta}_n)$.

Дельта метод. МЛЕ подход

- Пусть $\tau = g(\theta)$, где g гладкая функция.
- MLE оценка для τ имеет вид $\hat{\tau}_n = g(\hat{\theta}_n)$.
- Как распределена $\hat{\tau}_n$?

Theorem

Если $\tau = g(\theta)$ и при этом g дифференцируема, $g'(\theta) \neq 0$, то

$$\frac{\hat{\tau}_n - \tau_*}{\widehat{se}(\hat{\tau}_n)} \rightsquigarrow \mathcal{N}(0, 1),$$

где $\hat{\tau}_n = g(\hat{\theta}_n)$ и $\widehat{se}(\hat{\tau}_n) = |g'(\hat{\theta})| \widehat{se}(\hat{\theta}_n)$.

Доказательство:

$$\begin{aligned}\hat{\tau}_n &= g(\hat{\theta}_n) \approx g(\theta_*) + (\hat{\theta}_n - \theta_*)g'(\theta_*) = \\ &= \tau_* + (\hat{\theta}_n - \theta_*)g'(\theta_*), \\ \sqrt{n}(\hat{\tau}_n - \tau_*) &\approx \sqrt{n}(\hat{\theta}_n - \theta_*)g'(\theta_*),\end{aligned}$$

$$\frac{\sqrt{nI(\theta_*)}(\hat{\tau}_n - \tau_*)}{g'(\theta_*)} \approx \sqrt{nI(\theta_*)}(\hat{\theta}_n - \theta_*).$$

$\sqrt{nI(\theta_*)}(\hat{\theta}_n - \theta_*)$ сходится к $\mathcal{N}(0, 1)$. Тогда

$$\frac{\sqrt{nI(\theta_*)}(\hat{\tau}_n - \tau_*)}{g'(\theta_*)} \rightsquigarrow \mathcal{N}(0, 1).$$

$$\hat{\tau}_n \approx \mathcal{N}(\tau_*, se^2(\hat{\tau}_n)), se^2(\hat{\tau}_n) = \frac{(g'(\theta_*))^2}{nI(\theta_*)}.$$

Распределение Бернулли

MLE оценка для распределения Бернулли

- Пусть $X_1, \dots, X_n \sim Bernoulli(p)$.
- Оценим $\psi = g(p) = \log \frac{p}{1-p}$.
- Логарифмическое правдоподобие равно

$$\ell(p) = \sum_{i=1}^n X_i \cdot \log(p) + \left(n - \sum_{i=1}^n X_i \right) \cdot \log(1-p).$$

Найдя производную и приравняв к нулю мы получим MLE оценку:

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}.$$

Найдем информацию по Фишеру

$$\begin{aligned}\log f(X; p) &= X \log(p) + (1-X) \log(1-p), \\ \frac{\partial \log f(X; p)}{\partial p} &= \frac{X}{p(1-p)} - \frac{1}{1-p}, \\ \frac{\partial^2 \log f(X; p)}{\partial p^2} &= -\frac{(1-2p)X + p^2}{p^2(1-p)^2}.\end{aligned}$$

Учитывая $\mathbb{E}X = p$ получаем

$$I(p) = \frac{1}{p(1-p)}.$$

- Так мы получаем $\widehat{se} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$.
- Учитывая $g'(p) = 1/(p-p^2)$, получаем

$$\widehat{se}(\widehat{\psi}_n) = |g'(\widehat{p}_n)| \quad \widehat{se}(\widehat{p}_n) = \frac{1}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}.$$

Нормальное распределение

Параметры нормального распределения

- Пусть $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.
- Пусть нам известно μ , но неизвестна σ .
- Мы хотим оценить $\psi = \log \sigma$.
- Логарифмическое правдоподобие равно

$$\ell(\sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2.$$

Взяв производную и приравняв её к нулю получаем:

$$\hat{\sigma}_n = \sqrt{\frac{\sum_{i=1}^n (X_i - \mu)^2}{n}}.$$

Вычисление se

Найдем информацию по Фишеру

$$\log f(X; \sigma) = -\log \sigma - \frac{(X - \mu)^2}{2\sigma^2},$$

$$\frac{\partial^2 \log f(X; \sigma)}{\partial \sigma^2} = \frac{1}{\sigma^2} - \frac{3(X - \mu)^2}{\sigma^4},$$

$$I(\sigma) = -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}.$$

- Так мы получаем $\widehat{se} = \frac{\hat{\sigma}_n}{\sqrt{2n}}$.
- Пусть $\psi = g(\sigma) = \log \sigma$, тогда $\widehat{\psi}_n = \log \hat{\sigma}_n$.
- Учитывая $g'(\sigma) = 1/\sigma$, получаем

$$\widehat{se}(\widehat{\psi}_n) = \frac{1}{\hat{\sigma}_n} \frac{\hat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}}.$$

Применение дельта-метода к ratio-метрикам

Метрика отношения

$$\mathfrak{R}_A = \frac{\sum_{u \in A} X(u)}{\sum_{u \in A} Y(u)}$$

Статистика теста Стьюдента

$$t = \frac{\mathfrak{R}_B - \mathfrak{R}_A}{\sqrt{\sigma^2(\mathfrak{R}_A) + \sigma^2(\mathfrak{R}_B)}}$$

Проблема

Как оценить дисперсии \mathfrak{R}_A и \mathfrak{R}_B ?

Математическое ожидание

Для любой $f(x, y)$, разложение Тейлора в окрестности точки $\theta = (\theta_x, \theta_y)$

$$f(x, y) = f(\theta) + f'_x(\theta)(x - \theta_x) + f'_y(\theta)(y - \theta_y) + R$$

где R - остаток меньшего порядка чем члены уравнения.

Положим $\theta = (\mathbb{E}X, \mathbb{E}Y) = (\mu_x, \mu_y)$.

Тогда математическое ожидание $f(X, Y)$

$$\begin{aligned}\mathbb{E}(f(X, Y)) &= \mathbb{E}[f(\theta) + f'_x(\theta)(X - \mu_x) + f'_y(\theta)(Y - \mu_y) + R] \\ &\approx \mathbb{E}[f(\theta)] + \mathbb{E}[f'_x(\theta)(X - \mu_x)] + \mathbb{E}[f'_y(\theta)(Y - \mu_y)] \\ &= \mathbb{E}[f(\theta)] + f'_x(\theta)\mathbb{E}[(X - \mu_x)] + f'_y(\theta)\mathbb{E}[(Y - \mu_y)] \\ &= \mathbb{E}[f(\theta)] + 0 + 0 \\ &= f(\mu_x, \mu_y)\end{aligned}$$

Для $f = f(X, Y) = \frac{X}{Y}$ получаем $\mathbb{E}f(X, Y) \approx \frac{\mu_x}{\mu_y}$.

Дисперсия

Дисперсия по определению

$$\mathbb{V}(f(X, Y)) = \mathbb{E} \{ [f(X, Y) - \mathbb{E}(f(X, Y))]^2 \}$$

Воспользуемся полученным приближением $\mathbb{E}(f(X, Y)) \approx f(\theta)$

$$\mathbb{V}(f(X, Y)) \approx \mathbb{E} \{ [f(X, Y) - f(\theta)]^2 \}$$

Применим разложение Тейлора для $f(X, Y)$ в окрестности точки $\theta = (\mu_x, \mu_y)$

$$\begin{aligned}\mathbb{V}(f(X, Y)) &\approx \mathbb{E} \{ [f(\theta) + f'_x(\theta)(X - \mu_x) + f'_y(\theta)(Y - \mu_y) - f(\theta)]^2 \} \\&= \mathbb{E} \{ [f'_x(\theta)(X - \mu_x) + f'_y(\theta)(Y - \mu_y)]^2 \} \\&= \mathbb{E} \{ f'^2_x(\theta)(X - \mu_x)^2 + 2f'_x(\theta)(X - \mu_x)f'_y(\theta)(Y - \mu_y) \\&\quad + f'^2_y(\theta)(Y - \mu_y)^2 \} \\&= f'^2_x(\theta)\mathbb{V}(X) + 2f'_x(\theta)f'_y(\theta)\text{cov}(X, Y) + f'^2_y(\theta)\mathbb{V}(Y)\end{aligned}$$

Дисперсия

Вычислим производные для $f(X, Y) = \frac{X}{Y}$

$$f'_x = \frac{1}{Y} \quad f'_y = -\frac{X}{Y^2}$$

Тогда в точке $\theta = (\mu_x, \mu_y)$ получаем

$$f'^2_x(\theta) = \frac{1}{\mu_y^2} \quad f'_x(\theta)f'_y(\theta) = -\frac{\mu_x}{\mu_y^3} \quad f'^2_y(\theta) = \frac{\mu_x^2}{\mu_y^4}$$

Подставив значения производных, получим итоговую формулу

$$\mathbb{V}\left(\frac{X}{Y}\right) \approx \frac{1}{\mu_y^2} \mathbb{V}(X) - 2\frac{\mu_x}{\mu_y^3} \text{cov}(X, Y) + \frac{\mu_x^2}{\mu_y^4} \mathbb{V}(Y)$$

Теперь можем вычислять t-статистику и pvalue для ratio метрик.

Недостатки такого подхода:

- Нет пользовательской метрики.

Информационная матрица Фишера

MLE оценка параметров

В многомерном случае мы можем ставить задачу о поиске набора параметров наилучшим образом описывающих выборку.

Пусть $\theta = (\theta_1, \dots, \theta_k)$ — набор интересующих нас параметров. А $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ — MLE параметров для имеющейся выборки.

Логарифмическое правдоподобие

Логарифмическое правдоподобие может быть записано в виде

$$\ell_n(\theta) = \sum_{i=1}^n \log f(X_i; \theta).$$

Матрица Фишера

Введем определение:

$$H_{jj} = \frac{\partial^2 \ell_n}{\partial \theta_j^2}, \quad H_{jk} = \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta_k}.$$

Тогда матрица Фишера записывается так:

$$I_n(\theta) = - \begin{pmatrix} \mathbb{E}_{\theta}(H_{11}) & \mathbb{E}_{\theta}(H_{12}) & \cdots & \mathbb{E}_{\theta}(H_{1k}) \\ \mathbb{E}_{\theta}(H_{21}) & \mathbb{E}_{\theta}(H_{22}) & \cdots & \mathbb{E}_{\theta}(H_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}_{\theta}(H_{k1}) & \mathbb{E}_{\theta}(H_{k2}) & \cdots & \mathbb{E}_{\theta}(H_{kk}) \end{pmatrix}$$

Тогда обратную к ней мы назовем precision матрицей:

$$J_n(\theta) = I_n^{-1}(\theta).$$

Многопараметрический дельта-метод

Theorem

При соответствующих условиях регулярности,

$$\hat{\theta} - \theta_* \rightsquigarrow \mathcal{N}(0, J_n).$$

Если $\hat{\theta}_j$ — j -я компонента вектора $\hat{\theta}$

$$\frac{\hat{\theta}_j - \theta_{j,*}}{\hat{s}\hat{e}_j} \rightsquigarrow \mathcal{N}(0, 1),$$

где $\hat{s}\hat{e}_j^2 = J_n(j, j)$ диагональный элемент
матрицы J_n .

$$\text{Cov}(\hat{\theta}_j, \hat{\theta}_k) \approx J_n(j, k).$$

Многопараметрический дельта-метод

Theorem

При соответствующих условиях регулярности,

$$\hat{\theta} - \theta_* \rightsquigarrow \mathcal{N}(0, J_n).$$

Если $\hat{\theta}_j$ — j -я компонента вектора $\hat{\theta}$

$$\frac{\hat{\theta}_j - \theta_{j,*}}{\widehat{se}_j} \rightsquigarrow \mathcal{N}(0, 1),$$

где $\widehat{se}_j^2 = J_n(j, j)$ диагональный элемент матрицы J_n .

$$\text{Cov}(\hat{\theta}_j, \hat{\theta}_k) \approx J_n(j, k).$$

Пусть $\tau = g(\theta_1, \dots, \theta_k)$ — гладкая функция параметров и

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial \theta_1} \\ \vdots \\ \frac{\partial g}{\partial \theta_k} \end{pmatrix}.$$

Theorem

Если $\nabla g(\hat{\theta}) \neq 0$, $\hat{\tau} = g(\hat{\theta})$, то

$$\frac{\hat{\tau} - \tau_*}{\widehat{se}(\hat{\tau})} \rightsquigarrow \mathcal{N}(0, 1),$$

где $\widehat{se}(\hat{\tau}) = \sqrt{(\hat{\nabla}g)^T \hat{J}_n(\hat{\nabla}g)}$,
 $\hat{J}_n = J_n(\hat{\theta}_n)$, $\hat{\nabla}g = \nabla g(\hat{\theta})$.

Применение многопараметрического дельта-метода

Example

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\tau = g(\mu, \sigma) = \sigma/\mu$.

Информационная матрица Фишера

$$I_n(\mu, \sigma) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix},$$

$$J_n = I_n^{-1}(\mu, \sigma) = \frac{1}{n} \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{pmatrix}, \quad \nabla g = \begin{pmatrix} -\frac{\sigma}{\mu^2} \\ \frac{1}{\mu} \end{pmatrix},$$

$$\hat{se}(\hat{\tau}) = \sqrt{(\hat{\nabla}g)^T \hat{J}_n(\hat{\nabla}g)} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{\hat{\mu}^4} + \frac{\hat{\sigma}^2}{2\hat{\mu}^2}}.$$

Линеаризация

Ratio метрика

$$\mathfrak{R}_A = \frac{\sum_{u \in A} X(u)}{\sum_{u \in A} Y(u)}$$

Линеаризованная пользовательская метрика

$$L(u) = X(u) - \kappa Y(u), \quad \forall u \in A$$

Линеаризованная метрика

$$\mathfrak{L}_A = \text{avg}_{u \in A} L(u)$$

- Почему это работает?
- Как выбрать κ ?

Направленность

Theorem

Пусть X и Y — пользовательские метрики, причём Y положительная.

\mathfrak{R} — ratio метрика, \mathfrak{L} - линеаризованная метрика.

Параметр κ определим как $\kappa(\eta) = (1 - \eta)\mathfrak{R}_A + \eta\mathfrak{R}_B$, $\eta \in \mathbb{R}$.

Тогда разности метрик между группами связаны следующим уравнением

$$\Delta(\mathfrak{L}) = ((1 - \eta)Y_B + \eta Y_A) \Delta(\mathfrak{R})$$

Доказательство. При $\eta = 0$ получаем $\kappa = \mathfrak{R}_A$ и $\Delta(\mathfrak{L}_{\kappa(0)}) = Y_B \Delta(\mathfrak{R})$:

$$\begin{aligned} \Delta(\mathfrak{L}_{\kappa(0)}) &= \Delta(X) - \kappa(0)\Delta(Y) = (X_B - X_A) - (X_A/Y_A)(Y_B - Y_A) \\ &= X_B - X_A Y_B / Y_A = Y_B(X_B/Y_B - X_A/Y_A) = Y_B \Delta(\mathfrak{R}) \end{aligned}$$

Аналогично, можно показать $\Delta(\mathfrak{L}_{\kappa(1)}) = Y_A \Delta(\mathfrak{R})$ при $\eta = 1$.

Представим $\mathfrak{L}_{\kappa(\eta)}$ как линейную комбинацию $\mathfrak{L}_{\kappa(0)}$ и $\mathfrak{L}_{\kappa(1)}$

$$\begin{aligned} \Delta(\mathfrak{L}_{\kappa(\eta)}) &= \Delta(X) - ((1 - \eta)\mathfrak{R}_A + \eta\mathfrak{R}_B) \Delta(Y) \\ &= (1 - \eta)\Delta(\mathfrak{L}_{\mathfrak{R}_A}) + \eta\Delta(\mathfrak{L}_{\mathfrak{R}_B}) = (1 - \eta)Y_B \Delta(\mathfrak{R}) + \eta Y_A \Delta(\mathfrak{R}) \quad \square \end{aligned}$$

Направленность

Ratio метрика: $\mathfrak{R}_A = \frac{\sum_{u \in A} X(u)}{\sum_{u \in A} Y(u)}$

Линеаризованная пользовательская метрика: $L(u) = X(u) - \kappa Y(u), \forall u \in A$

Линеаризованная метрика: $\mathfrak{L}_A = \text{avg}_{u \in A} L(u)$

$$\kappa(\eta) = (1 - \eta)\mathfrak{R}_A + \eta\mathfrak{R}_B, \eta \in \mathbb{R}$$

$$\Delta(\mathfrak{L}) = ((1 - \eta)Y_B + \eta Y_A) \Delta(\mathfrak{R})$$

Следствие

Если $\kappa \in [\min\{\mathfrak{R}_A, \mathfrak{R}_B\}, \max\{\mathfrak{R}_A, \mathfrak{R}_B\}]$, то

$$\text{sgn}\Delta(\mathfrak{R}) = \text{sgn}\Delta(\mathfrak{L})$$

Уровень значимости

Оценка значимости

Для оценки значимости изменения линеаризованной метрики нужно построить её распределение при нулевой гипотезе.

$$\Delta(\mathfrak{L}) = \mathfrak{L}_B - \mathfrak{L}_A$$

$$\mathfrak{L} = \text{avg } L$$

$$L = X - \kappa Y$$

\mathfrak{L} — среднее значение линеаризованной метрики по объектам рандомизации, которые считаются независимыми.

Можно применить тест Стьюдента, он будет работать корректно при заранее фиксированном и независящем от наблюдений параметре κ .

Проблема зависимости данных

Если устанавливаем $\kappa = \mathfrak{R}_A$, нарушаются два ключевых условия критерия Стьюдента:

- a. Значения \mathfrak{L}_A и \mathfrak{L}_B не независимые;
- b. Значения внутри множеств $\{L(u) | u \in V\}$, $V = A, B$ не независимые.

Тем не менее, есть ряд теорем, которые показывают, что критерий Стьюдента применим для $\Delta(\mathfrak{L})$.

Уровень значимости

Theorem

Пусть X и Y — пользовательские метрики, причём Y положительная.

\mathfrak{R} — ratio метрика, \mathfrak{L} — линеаризованная метрика с параметром $\kappa = \mathfrak{R}_A$.

Пусть $T(\mathfrak{L})$ — t -статистика для среднего метрики L .

$D(\mathfrak{R}) = \Delta(\mathfrak{R}) / \sqrt{\delta(\mathfrak{R}_A) + \delta(\mathfrak{R}_B)}$ — асимптотически стандартная нормальная статистика, полученная делта методом, где $\delta(\mathfrak{R}_A)$ и $\delta(\mathfrak{R}_B)$ — дисперсии метрик \mathfrak{R}_A и \mathfrak{R}_B , а $\Delta(\mathfrak{R}) = \mathfrak{R}_B - \mathfrak{R}_A$.

1. Тогда верно следующее тождество

$$T(\mathfrak{L}) = D(\mathfrak{R}) \sqrt{1 - \frac{\gamma}{\delta(\mathfrak{R}_A) + \delta(\mathfrak{R}_B) + \gamma}}$$

$$\text{где } \gamma = (Y_A^2/Y_B^2 - 1)\delta(\mathfrak{R}_A) + \beta \text{ и } \beta = \frac{\Delta(\mathfrak{R}) ((\mathfrak{R}_A + \mathfrak{R}_B)\sigma_B^2(Y) - 2\widehat{\text{cov}}_B(X, Y))}{|B|Y_B^2}$$

Уровень значимости

Theorem (продолжение)

2. Если выборочная корреляция ограничена $|\widehat{\text{corr}}_B(X, Y)| < c < 1$, то верно

$$\left| \frac{T(\mathfrak{L})}{D(\mathfrak{R})} - 1 \right| \leq C_1(c) \left| \frac{\Delta X}{X_B} \right| + C_2(c) \left| \frac{\Delta Y}{Y_B} \right|$$

при достаточно малых относительных изменениях $|\Delta X/X_B| < \varepsilon_1(c)$ и $|\Delta Y/Y_B| < \varepsilon_2(c)$; где константы $C_1(c), C_2(c), \varepsilon_1(c), \varepsilon_2(c)$ зависят только от границы c .

3. Если $|\text{corr}(X, Y|B) < c < 1|$ и $\mathbb{E}[X|A] \neq 0, \mathbb{E}[Y|A] \neq 0$, то t -статистика $T(\mathfrak{L})$ асимптотически нормальна при нулевой гипотезе, что $\mathbb{E}[\Delta X] = 0$ и $\mathbb{E}[\Delta Y] = 0$.

Суть

Асимптотически критерий Стьюдента для линеаризованной метрики работает корректно. На практике, относительные изменения метрик $\Delta X/X_B$ и $\Delta Y/Y_B$ невелики, не более нескольких процентов, в этом случае отличие между $T(\mathfrak{L})$ и $D(\mathfrak{R})$ того же порядка. На практике, уровни значимости $\Delta(\mathfrak{L})$, рассчитанные с помощью критерия Стьюдента, согласуются с уровнями значимости $\Delta(\mathfrak{R})$, полученными дельта-методом и бутстрепом.

Алгоритм применения линеаризации

Алгоритм применения линеаризации для оценки результатов АВ теста, в котором в качестве основной метрики выступает метрика-отношение

$$\mathfrak{R}_V = \frac{\sum_{u \in V} X(u)}{\sum_{u \in V} Y(u)}, \quad V = A, B$$

- Для каждого пользователя, участвовавшего в АВ-тесте, вычислить линеаризованную метрику по формуле

$$L(u) = X(u) - \kappa Y(u), \quad \kappa = \mathfrak{R}_A$$

- Опционально, к полученной пользовательской метрике можно применить какие-либо способы повышения чувствительности;
- Вычислить значение t-статистики критерия Стьюдента и соответствующее значение p_{value} ;
- Принять решение согласно оговоренному до начала эксперимента правилу, основываясь на полученном значении p_{value} и требуемом уровне значимости.

Сравнение методов

	легкость вычисления	корректное p-value	наличие пользовательской метрики	направленность
по сессиям	+	-	-	+
среднее по пользователям	+	-	+	-
бутстреп	-	+	-	+
дельта метод	+	+	-	+
линеаризация	+	+	+	+

Если данных слишком много

Проблемы производительности и хранения данных

Если мы логируем всё, то упираемся в проблемы:

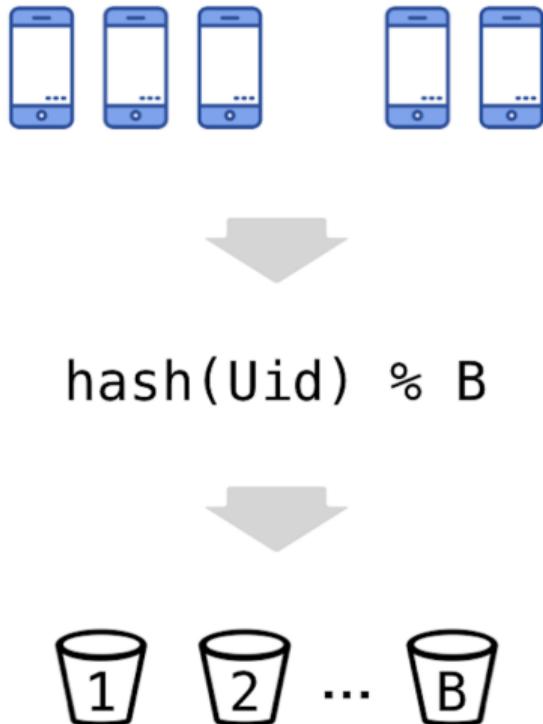
- Нужны очень большие и дорогие хранилища
- Подсчет даже самых простых метрик упирается в чтение с диска

Существует объективный предел масштабирования.

Бакет — новая экспериментальная единица

Рассчет пользовательских метрик в крупной компании ежедневно порождает сотни миллиардов значений.

Необходимо сокращать объемы. Мы можем объединять пользователей по их UID.



Метод бакетов

Бакетная метрика

Выбираем число бакетов B и проводим объединение на основе UID по величине $\text{hash}(UID) \% B$:

$$x_j^b = \sum_{i=1}^{N_j} x_{ij}$$

Для ratio отдельно считаем числитель и знаменатель.



Дисперсия среднего

Легко проверить, что дисперсия среднего не изменяется:

$$\frac{s^2}{N} \simeq \frac{s_b^2}{B}$$

Значение метрики тоже сохраняется.

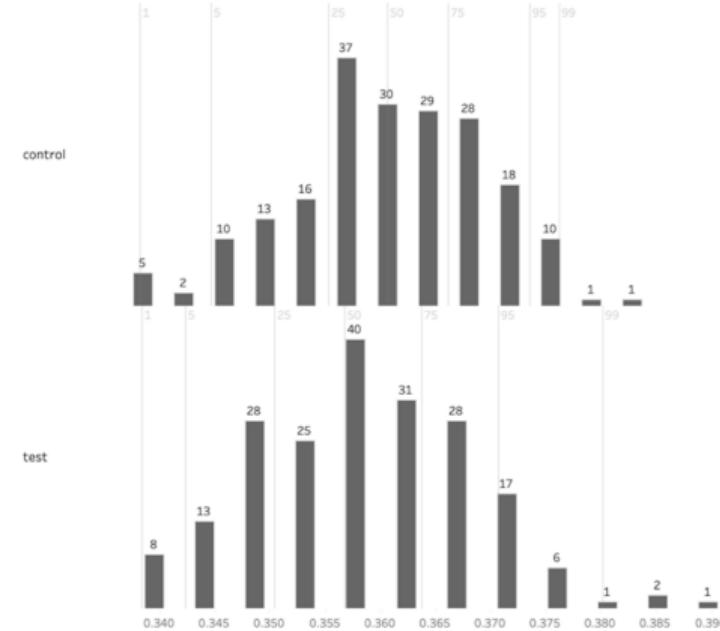
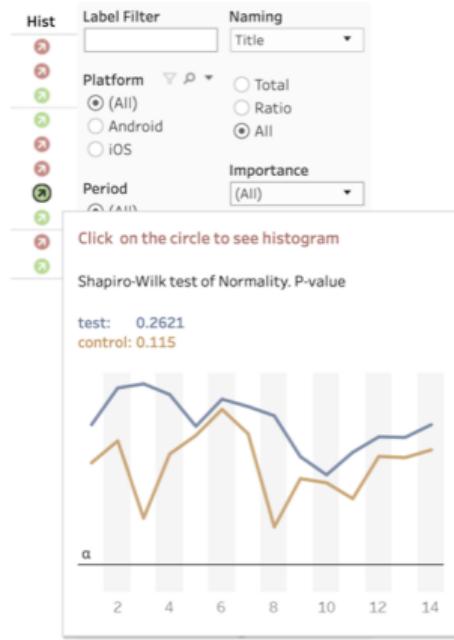
Нормальное распределение

Плотность распределения метрики после бакетного преобразования всегда становится схожа с нормальным.

Вместе с этим мы сокращаем размер выборки до фиксированных размеров.

Преимущества метода бакетов

Мы радикально сокращаем объем метрик, при этом переходим к нормальным распределениям. Значение метрики и дисперсия среднего не меняются.



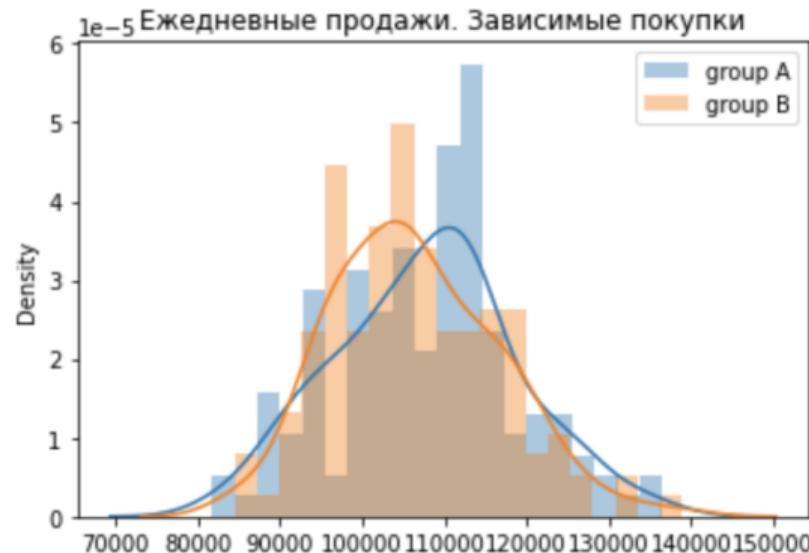
Суточные продажи в магазине

Посуточные данные продаж в магазине

Если мы анализируем покупки в магазине, то самой естественной мыслью будет объединить данные по продажам за каждый день. За несколько дней мы получим несколько значений. Можно даже построить красивое распределение.

Но для проведения АВ-теста такие данные непригодны. Привычки покупателей делают данные зависимыми:

- Кто-то не ходит в магазин два дня подряд
- Другой делает покупки только по средам



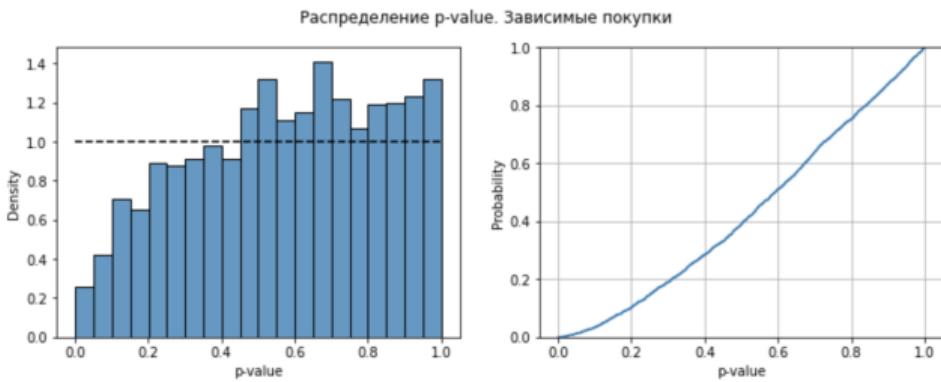
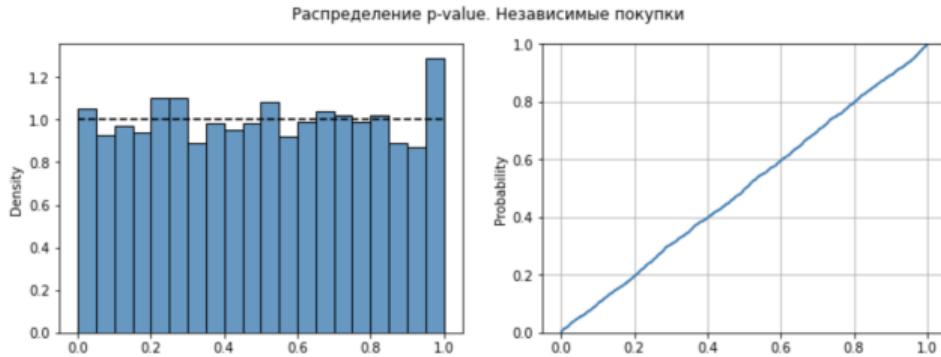
Зависимость в данных ломает АА-тест

Независимые покупки

Если для каждого посетителя магазина покупки не зависят от его поведения в прошлом, то проблем не возникает. Мы можем объединять покупки всех посетителей в посугочную статистику.

Зависимые покупки

Если же конкретный покупатель никогда не ходит в магазин два раза подряд или у него есть любимые дни для посещения магазина, то АА-тест разваливается.

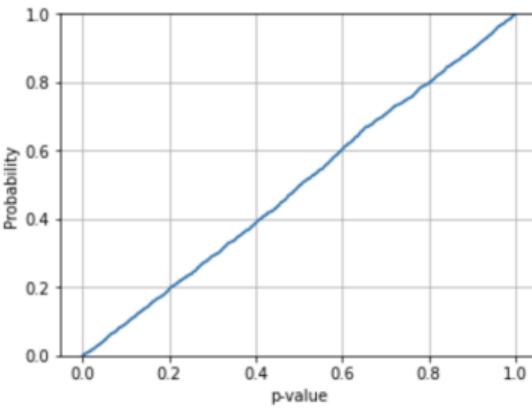
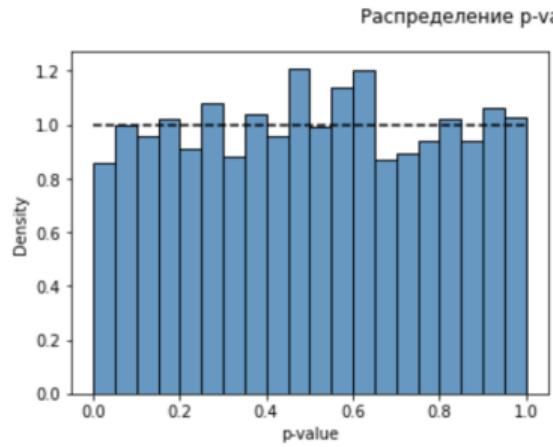
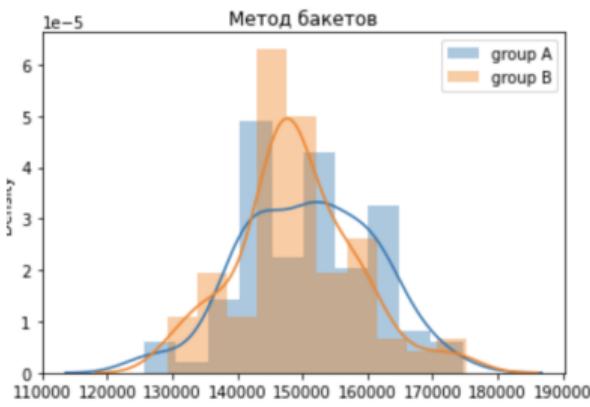


Бакеты помогают бороться с зависимостью в данных

Переход к анализу бакетов

Мы видим, что при агрегации по магазинам происходит протекание зависимости. Эта зависимость связана с паттернами поведения покупателей.

Если каждый покупатель дает только одно число, то зависимости в данных не возникает. Мы можем объединить их в группы на основе хэша и вычислить агрегированные показатели.



Резюме

- 1 Ratio метрики
- 2 User average
- 3 Бутстреп
- 4 Дельта метод
- 5 Применение дельта-метода к ratio-метрикам
- 6 Многопараметрический дельта-метод
- 7 Линеаризация
- 8 Если данных очень много. Метод бакетов

Дополнительные материалы

Ссылки для самостоятельного изучения

1. Как устроено А/В-тестирование в Авито
2. VK Tech. Practitioner's Guide to Statistical Tests
3. Applying the Delta Method in Metric Analytics: A Practical Guide with Novel Ideas
4. Как измерить счастье пользователя
5. Approximations for Mean and Variance of a Ratio