



> Конспект > 6 урок > Линеаризация. Многопараметрический дельта метод

> Оглавление

> Оглавление

> Ratio-метрики

> Пользовательские метрики vs метрики отношения

> Пользовательские метрики

> Ratio метрики

> Проблема

> Наивный способ

> Зависимость данных

> User average

> Среднее по пользователям

> Направленность

> Бутстреп

> Дельта-метод

> Информация Фишера

> Вычисление информации Фишера

> Информация Фишера

> Дельта метод. MLE подход

> Распределение Бернулли

> MLE оценка для распределения Бернулли

> Нормальное распределение

> Параметры нормального распределения

> Вычисление se

> Применение дельта-метода к ratio-метрикам

> Математическое ожидание

> Дисперсия

> Многопараметрический дельта-метод

КУПЛЕНО НА
SKLADCHIK.COM

- > Информационная матрица Фишера
 - > MLE оценка параметров
 - > Логарифмическое правдоподобие
 - > Матрица Фишера
- > Многопараметрический дельта-метод
- > Применение многопараметрического дельта-метода
- > Линеаризация
 - > Направленность
 - > Уровень значимости
 - > Оценка значимости
 - > Проблема зависимости данных
 - > Алгоритм применения линеаризации
- > Если данных очень много. Метод бакетов
 - > Проблемы производительности и хранения данных
 - > Метод бакетов
 - > Бакетная метрика
 - > Дисперсия среднего
 - > Нормальное распределение
 - > Преимущества метода бакетов
 - > Суточные продажи в магазине
 - > Зависимость в данных ломает A/A-тест
 - > Независимые покупки
 - > Зависимые покупки
 - > Бакеты помогают бороться с зависимостью в данных
- > Дополнительные материалы

> Ratio-метрики

> 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:

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

Недостатки такого подхода:

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

> Зависимость данных

Представим, что по России собраны данные о распределении цвета глаз:

- 50% — серый;
- 20% — синий и голубой;
- 25% — карий;
- 5% — чёрный и зелёный.



Попробуем угадать, какого цвета глаза у мужчин на фото. Это разные люди, независимые между собой.



Не угадали по ММП



Рассмотрим ещё один пример: по ММП голосуем за серый цвет



Мы не угадали (цвет зелёный), но пришла информация, что эти люди — близнецы, тогда меняется апостериорная вероятность, по которой выбираем зелёный цвет



Теперь верно!

Если есть зависимые данные, то получение новой части данных не несёт в себе столько же информации, сколько было бы в независимых данных. Будет казаться, что данных много, но они не будут нести в себе большого количества новой информации.

> User average

> Среднее по пользователям

Предполагаем, что хотя бы внутри одного пользователя данные зависимы, но между пользователями — нет.

Предложение: 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:

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

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

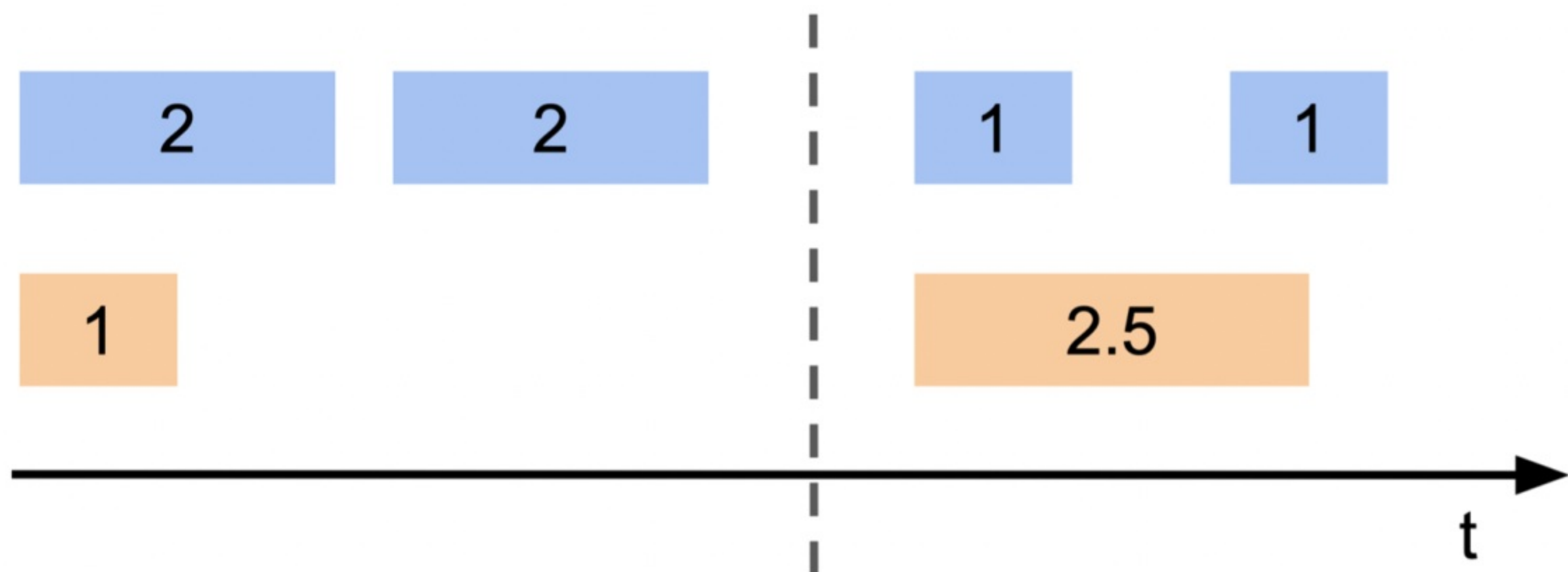
у них разный вклад.

- Не сохраняется направленность. При увеличении метрики M_A , метрика M_A^{avg} может уменьшаться, и наоборот.

> Направленность

Пример несохранения направленности метрик.

Даны длины сессий двух пользователей, вычислим изменения метрик.



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

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

> Бутстреп

Как действовать в случае ratio-метрик:

1. Генерируем случайные выборки пользователей из пилотной и контрольной групп (A_{bs}, B_{bs}).
2. Для каждой пары сэмплов считаем разность их ratio метрик:
$$\Delta = \mathfrak{R}_{B_{bs}} - \mathfrak{R}_{A_{bs}} = \sum_{u \in B_{bs}} X(u) / \sum_{u \in B_{bs}} Y(u) - \sum_{u \in A_{bs}} X(u) / \sum_{u \in A_{bs}} Y(u)$$
3. Повторяем первые два шага, чтобы набрать достаточно данных.
4. По полученному множеству дельт оцениваем значимость отличия от нуля разности метрик.

Получаем корректную оценку.

Недостатки такого подхода:

КУПЛЕНО НА
SKLADCHIK.COM

- Много времени уходит на вычисления;
- Нет пользовательской метрики.

> Дельта-метод

> 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

$$I_n(\theta) = E_\theta \left(\frac{\delta L(\theta, X_1, \dots, X_n)^2}{\delta \theta} \right), L = \sum_{i=1}^n \ln f(\theta, X_1, \dots, X_n),$$

где $L(\theta, X_1, \dots, X_n)$ — лог. функция правдоподобия, а E_θ — мат. ожидание 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

$$I_n(\theta) = E_\theta \left(\frac{\delta L(\theta, X)^2}{\delta \theta} \right) = -E_\theta \left(\frac{\delta^2 L(\theta, X)}{\delta \theta^2} \right)$$

При этом должно выполняться условие регулярности модели:

$$E_\theta \left(\frac{\delta L(\theta, X_1, \dots, X_n)}{\delta \theta} \right) = 0$$

Фишеровское количество информации в одном наблюдении:

$$I_i(\theta) = E_\theta \left(\frac{\delta \ln f(\theta, X_i)}{\delta \theta} \right)^2$$

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

> Информация Фишера

Теорема

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

$$se \approx \sqrt{1/I_n(\theta_*)} \text{ и } \frac{\hat{\theta}_n - \theta_*}{se} \rightsquigarrow N(0, 1)$$

> Дельта метод. MLE подход

- Пусть $\tau = g(\theta)$, где g гладкая функция.

- MLE оценка для τ имеет вид $\hat{\tau}_n = g(\hat{\theta}_n)$
- Как распределена $\hat{\tau}_n$?

Теорема

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

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

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

Доказательство

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

$$\sqrt{n}(\hat{\tau}_n - \tau_*) \approx \sqrt{n}(\hat{\theta}_n - \theta_*)g'(\theta_*),$$

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

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

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

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

> Распределение Бернулли

> MLE оценка для распределения Бернулли

- Пусть $X_1, \dots, X_n \sim \text{Bernoulli}(p)$

--

- Оценим $\psi = g(p) = \log \frac{p}{1-p}$
- Логарифмическое правдоподобие равно:

$$l(p) = \sum_{i=1}^n X_i \cdot \log(p) + (n - \sum_{i=1}^n X_i) \cdot \log(1-p)$$

Найдя производную и приравняв к нулю, мы получим MLE оценку:

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

Найдём информацию по Фишеру:

$$\log f(X; p) = X \log(p) + (1 - X) \log(1 - p)$$

$$\frac{\delta \log f(X; p)}{\delta p} = \frac{X}{p(1-p)} - \frac{1}{1-p}$$

$$\frac{\delta^2 \log f(X; p)}{\delta p^2} = -\frac{(1-2p)X + p^2}{p^2(1-p)^2}$$

Учитывая $EX = p$, получаем:

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

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

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

> Нормальное распределение

> Параметры нормального распределения

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

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

Взяв производную и приравняв её к нулю, получаем:

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

> Вычисление se

Найдём информацию по Фишеру:

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

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

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

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

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

> 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:

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

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 $\theta = (EX, EY) = (\mu_x, \mu_y)$.

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 $E f(X, Y) \approx \frac{\mu_x}{\mu_y}$

> Дисперсия

Дисперсия по определению:

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

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

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

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

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

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

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

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

$$f'^2_x(\theta) = \frac{1}{\mu_y^2}$$

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

$$f'^2_y(\theta) = \frac{\mu_x^2}{\mu_y^4}$$

Подставив значения производных, получим итоговую формулу:

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

Теперь можем вычислять t-статистику и p-value для ratio-метрик.

Недостаток такого подхода: нет пользовательской метрики.

> Многопараметрический дельта-метод

> Информационная матрица Фишера

> MLE оценка параметров

В многомерном случае мы можем ставить задачу отыскания наилучшим образом набора параметров, 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:

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

> Матрица Фишера

Введём определение:

$$H_{jj} = \frac{\delta^2 l_n}{\delta \theta_j^2}, H_{jk} = \frac{\delta^2 l_n}{\delta \theta_j \delta \theta_k}$$

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 назовём precision матрицей:

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

> Многопараметрический дельта-метод

Теорема

При соответствующих условиях регулярности

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

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

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

где $\hat{se}_j^2 = J_n(j, j)$ — 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

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

Теорема

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

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

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

$$\hat{J}_n = J_n(\text{hat}\theta_n), \hat{\nabla} g = \nabla g(\text{hat}\theta)$$

> Применение многопараметрического дельта-метода

Пример

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

Информационная матрица Фишера:

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

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

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

> Линеаризация

Ratio-метрика:

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

Переходим в **другое признаковое пространство**.

Линеаризованная **пользовательская** метрика (одно число для одного пользователя):

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

Линеаризованная метрика (усреднение по пользовательским метрикам):

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

> Направленность

Теорема

Пусть X и Y — пользовательские метрики, причём Y положительная. $\mathfrak{R}$ — ratio метрика, $\mathcal{L}$ — линеаризованная метрика.

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

Тогда разница метрик между группами связана следующим уравнением:

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

Доказательство

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

$$\Delta(\mathcal{L}_{\kappa(0)}) = \Delta(X) - k(0)\Delta(Y) = (X_B - X_A) - (X_A/Y_A)(Y_B - Y_A) = X_B - X_A Y_B / Y_A = Y_B (X_B / Y_B - X_A / Y_A) = Y_B \Delta(\mathfrak{R})$$

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

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

$$\Delta(\mathcal{L}_\eta) = \Delta(X) - ((1 - \eta)\mathfrak{R}_A + \eta\mathfrak{R}_B)\Delta(Y) = (1 - \eta)\Delta(\mathcal{L}_{\mathfrak{R}_A}) + \eta\Delta(\mathcal{L}_{\mathfrak{R}_B}) = (1 - \eta)Y_B \Delta(\mathfrak{R}) + \eta Y_A \Delta(\mathfrak{R})$$

Получаем систему:

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

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

Следствие

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

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

> Уровень значимости

> Оценка значимости

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

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

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

$$L = X - \kappa Y$$

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

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

> Проблема зависимости данных

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

- Значения $\mathcal{L}_A$ и $\mathcal{L}_B$ не независимы;
- 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 $\Delta(\mathcal{L})$.

Теорема

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

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

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

$D(\mathfrak{R}) = \Delta(\mathfrak{R}) / \sqrt{\delta(\mathfrak{R}_A) + \delta(\mathfrak{R}_B)}$ — симптотически стандартная нормальная статистика, полученная дельта-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:

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

где $\gamma = (Y_A^2/Y_B^2 - 1)\delta(\mathfrak{R}_A) + \beta$ и

$$\beta = \frac{\Delta(\mathfrak{R})((\mathfrak{R}_A + \mathfrak{R}_B)\sigma_B^2(Y) - 2\text{cov}_B(X, Y))}{|B|Y_B^2}$$

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

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

при достаточно малых относительных изменениях $|\Delta X/X_B| < \epsilon_1(c)$ и $|\Delta Y/Y_B| < \epsilon_2(c)$, где константы $C_1(c)$, $C_2(c)$, $\epsilon_1(c)$, $\epsilon_2(c)$ зависят только от границы c

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

Суть

Асимптотически критерий Стьюдента для линеаризованной метрики работает корректно.

На практике относительные изменения метрик $\Delta X/X_B$ и $\Delta Y/Y_B$ невелики и составляют не более нескольких процентов, в этом случае отличие между $T(\mathcal{L})$ и $D(\mathfrak{R})$ того же порядка. На практике уровни значимости $\Delta(\mathcal{L})$, 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 A/B-теста, в котором в качестве основной метрики выступает метрика-отношение

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

1. Для каждого пользователя, участвовавшего в A/B-тесте, **вычислить линеаризованную метрику** по формуле:

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

- 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.

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

Сравнение методов

> Если данных очень много. Метод бакетов

> Проблемы производительности и хранения данных

Если мы логируем всё, то упираемся в проблемы:

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

Существует объективный предел масштабирования.

Бакет — новая экспериментальная единица.

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

Необходимо сокращать объёмы. Мы можем объединять пользователей по их *UID*.



$\text{hash}(\text{Uid}) \% B$



> **Метод бакетов**

> **Бакетная метрика**

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

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

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

1 млн ➔ 200

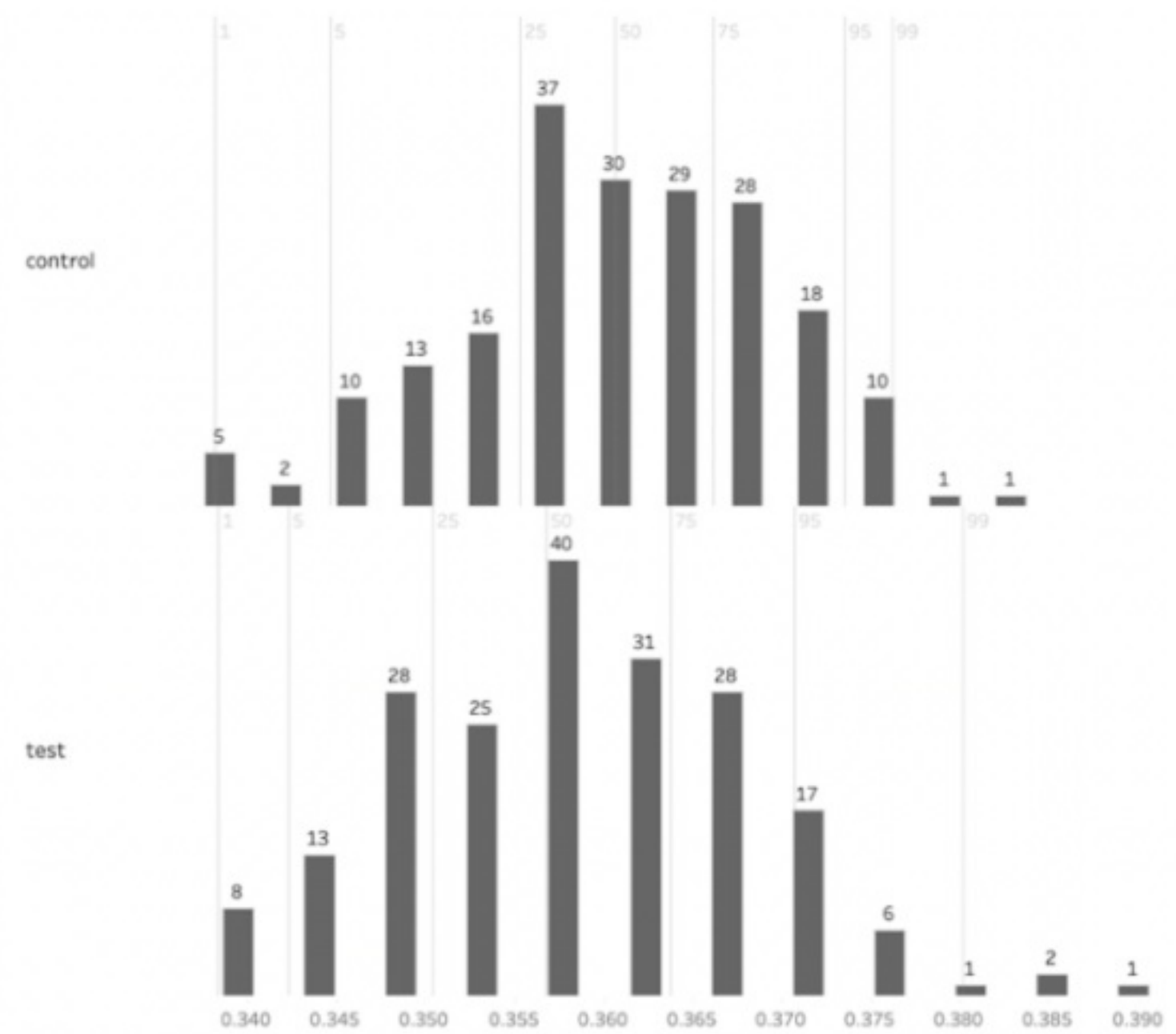
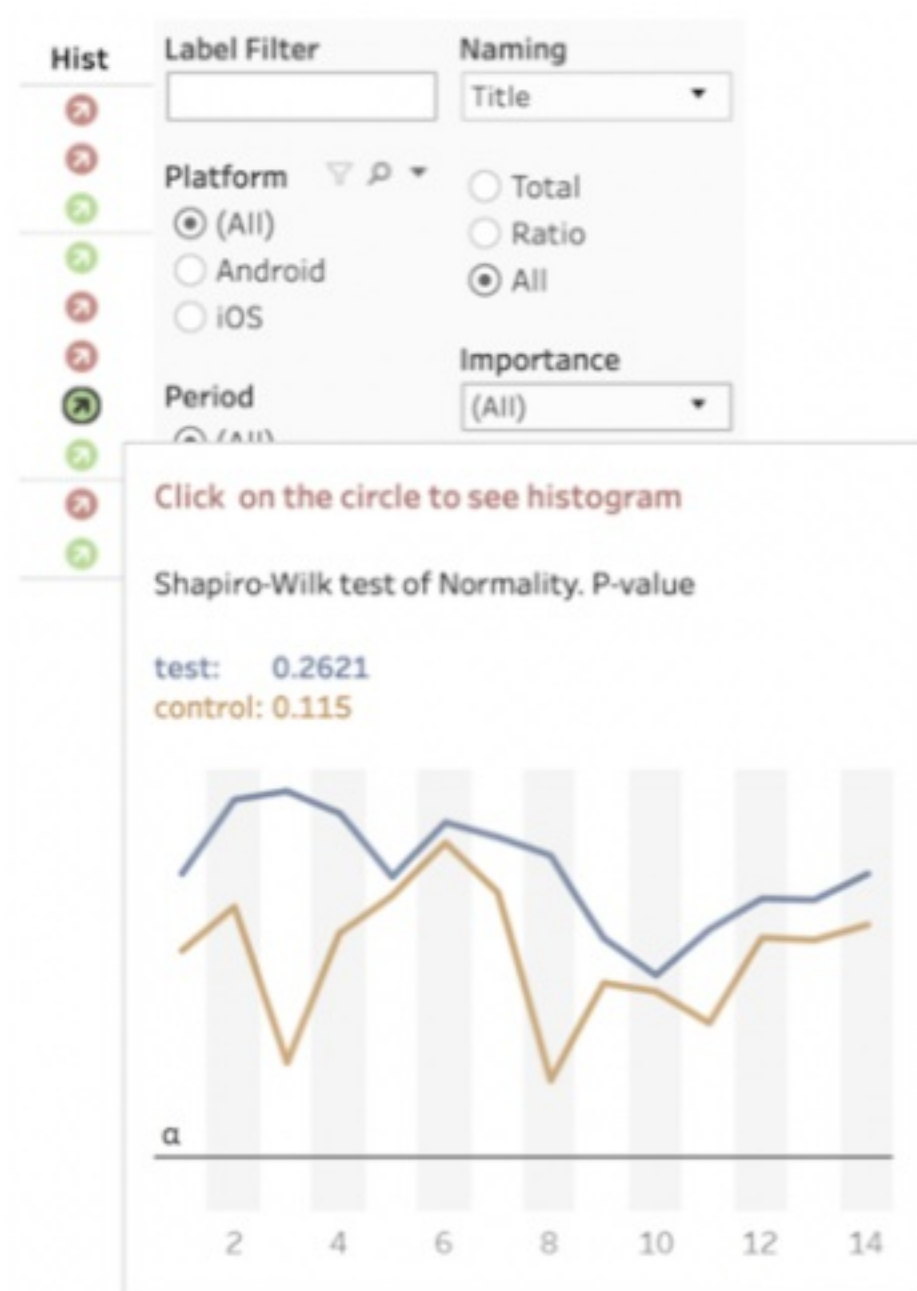


Плотность распределения метрики после бакетного преобразования всегда становится схожа с нормальным.

Вместе с этим мы **сокращаем размер выборки** до фиксированных размеров.

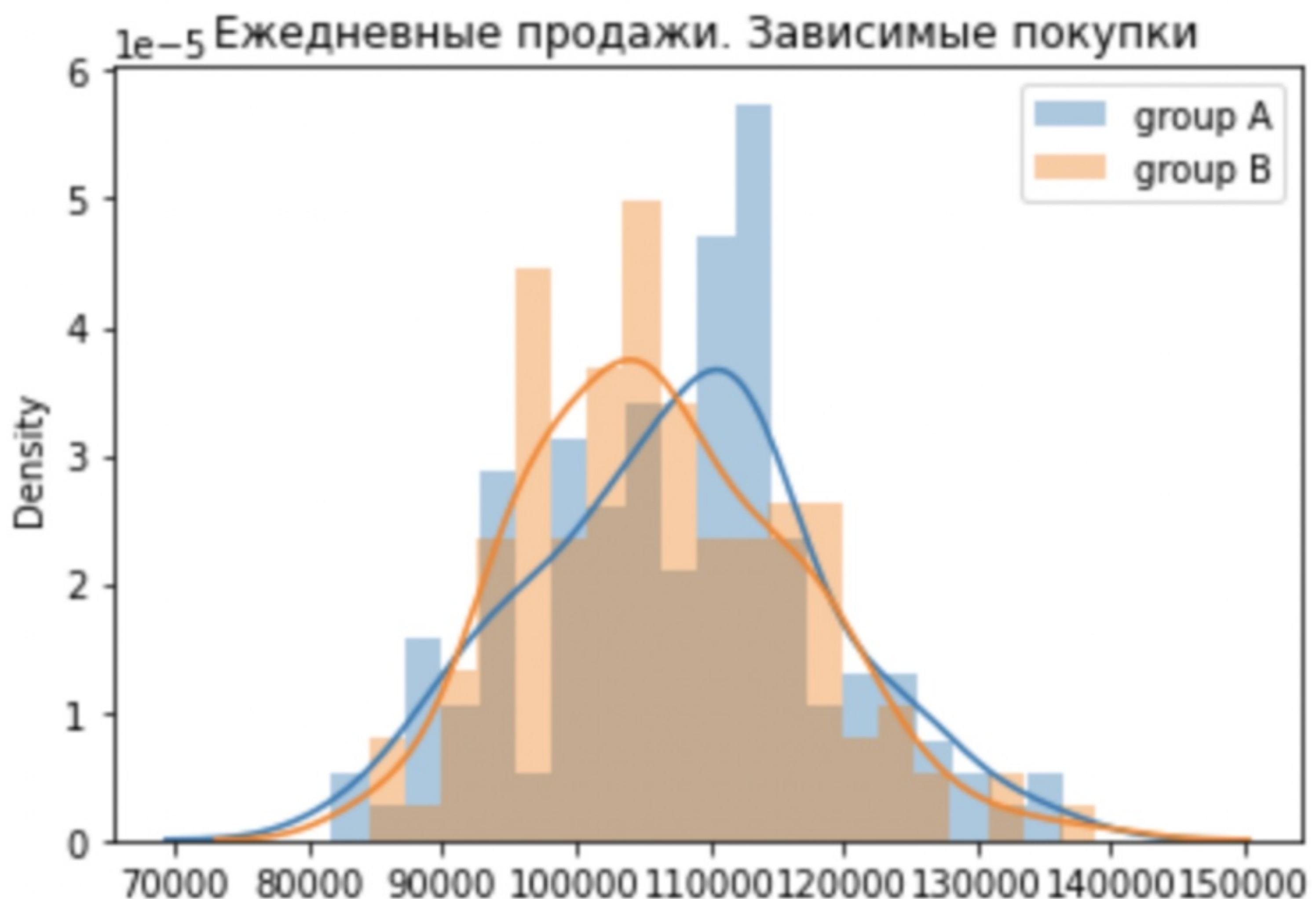
> Преимущества метода бакетов

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



> Суточные продажи 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.



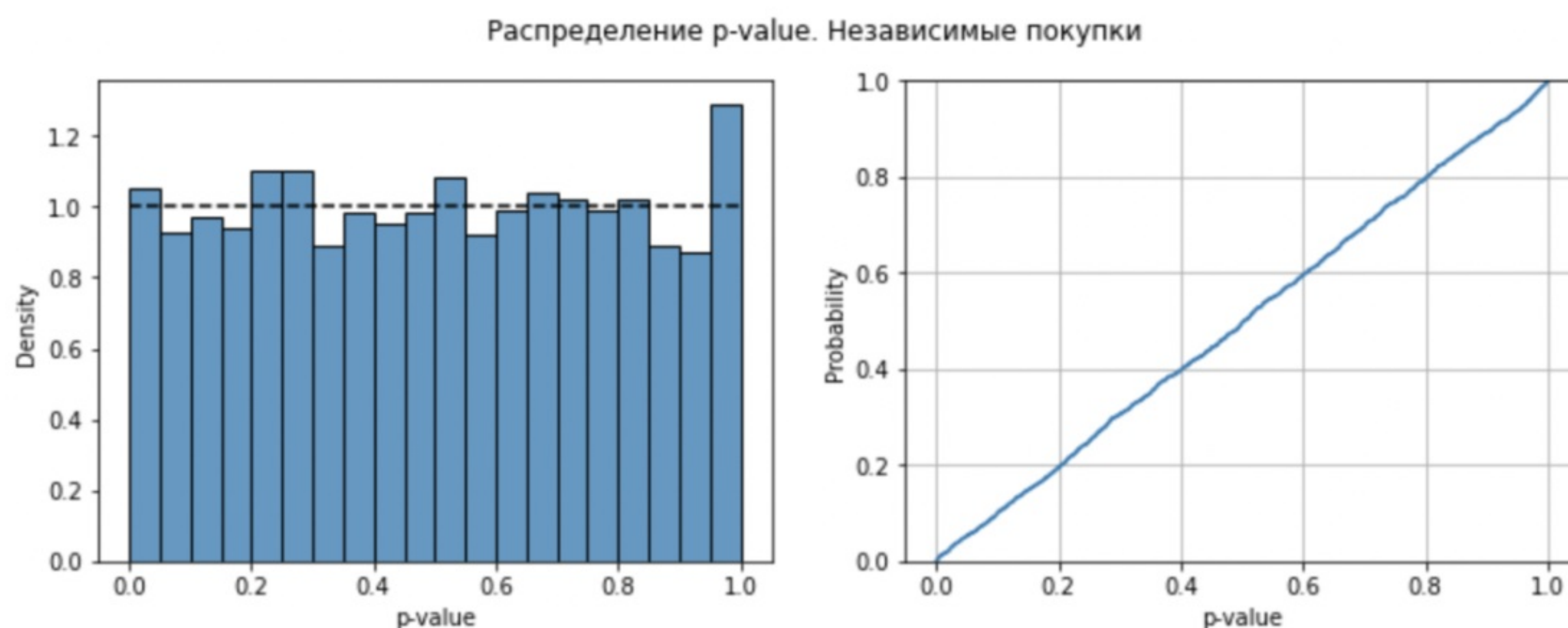
Но для проведения A/B-теста такие данные непригодны. Привычки покупателей делают данные зависимыми:

- Кто-то не ходит в магазин два дня подряд;
- Другие делают покупки только по средам.

> Зависимость в данных ломает A/A-тест

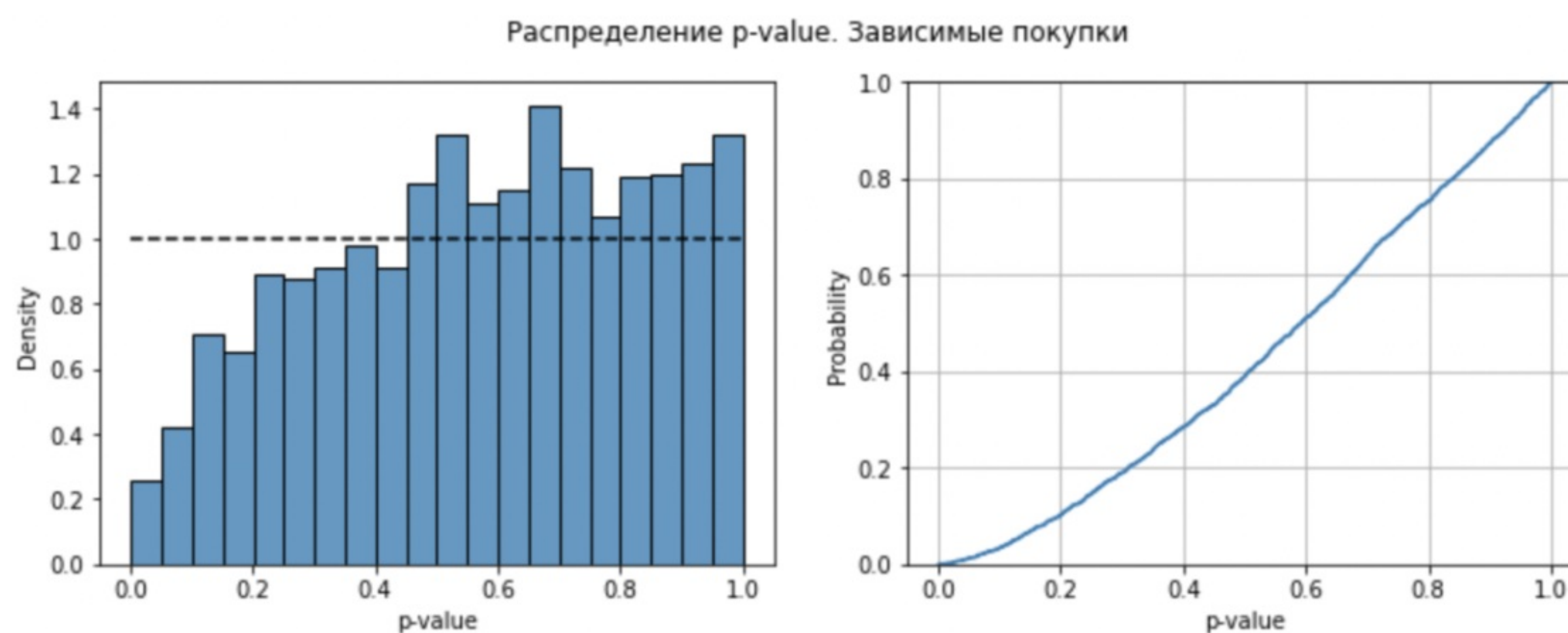
> Независимые покупки

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



> Зависимые покупки

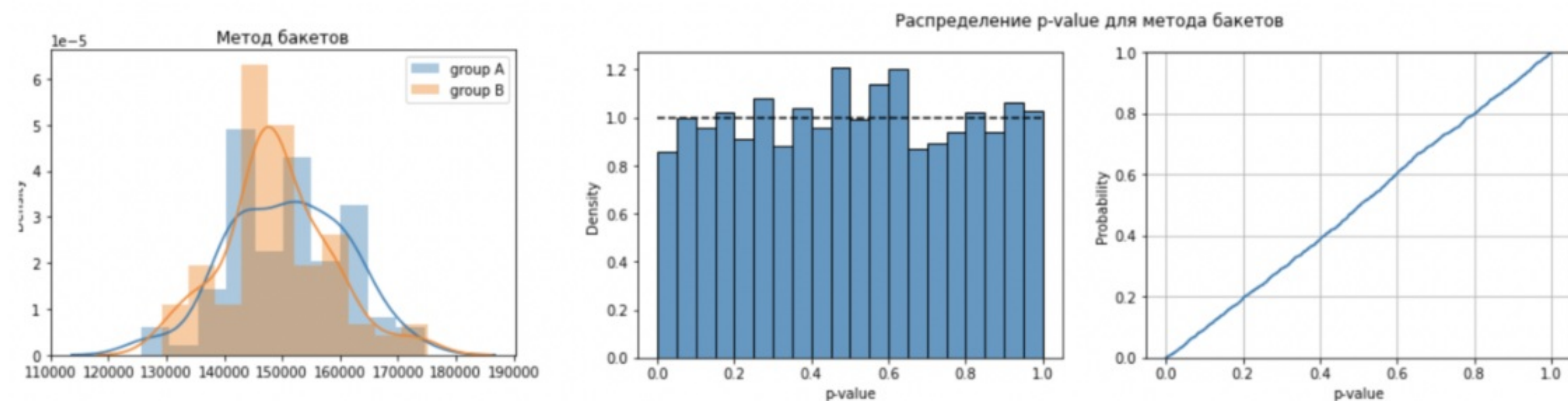
Если же конкретный покупатель никогда не ходит в магазин два раза подряд или у него есть любимые дни для посещения магазина, то A/A-тест разваливается.



> Бакеты помогают бороться с зависимостью в данных

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

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



> **Дополнительные материалы**

1. [Как устроено A/B-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](#)