Стратификация

Александр Сахнов
linkedin.com/in/amsakhnov

Staff MLE at Alibaba Group

2 сентября 2021 г.

Оглавление

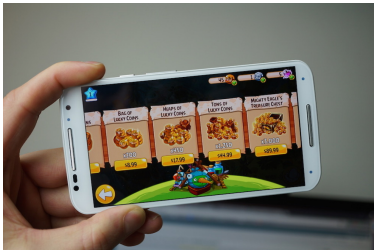
1. Стратификация
2. Точечные оценки популяционного среднего
3. Стратифицированное семплирование
4. Понижение дисперсии
5. Пост стратификация
6. Сравнение методов семплирования

Стратификация

Изменим рекламу встроенных покупок, чтобы увеличить их продажи.

Определили метрики, размер пилотной и контрольной групп.

Случайно распределяем пользователей по группам и начинаем пилот.

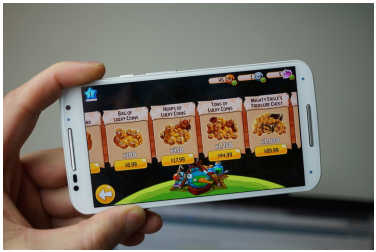


Стратификация

Изменим рекламу встроенных покупок, чтобы увеличить их продажи.

Определили метрики, размер пилотной и контрольной групп.

Случайно распределяем пользователей по группам и начинаем пилот.



Кто наши пользователи?

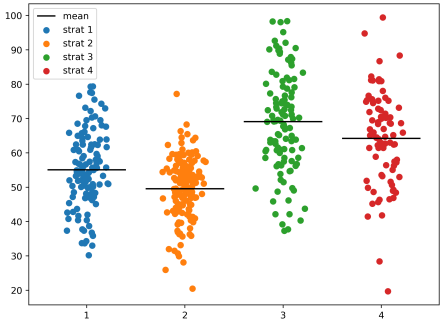
Как проводить стратификацию?

Стратификация

Предположим, что нам удалось найти один или несколько признаков, которые коррелируют с исследуемой *бизнес метрикой* Y. Такие признаки X мы будем называть *ковариатами*. Эти величины должны быть измеримы до эксперимента.

Например, это может быть пол, возраст или иные характеристики пользователя. Для международных онлайн платформ хорошим признаком будет страна проживания пользователя.

Ковариаты используются для того, чтоб разделить всю генеральную совокупность на K непересекающихся подмножеств, называемых *стратами*.



Замечание:

Распределения целевой метрики в различных стратах должно отличаться. Иначе стратификация не имеет смысла.

Популяционное среднее

Нам необходимо оценить популяционное среднее бизнес метрики Y.

Обозначения:

- $\mu = \mathbb{E}Y$ — популяционное среднее;
- $\sigma^2 = \mathbb{V}Y$ — популяционная дисперсия;
- μ_k, σ_k^2 — среднее значение и дисперсия бизнес метрики для k-й страты;
- w_k — доля k-й страты в популяции;
- n_k — число пользователей из k-й страты в рассматриваемой группе;
- $n = \sum_{k=1}^{K} n_k$ — общий размер группы.
- $Y_{11}, \dots Y_{1n_1}, \dots Y_{K1}, \dots Y_{Kn_K}$ — выборка из г.с., где Y_{kj} — метрика для j-го пользователя k-й страты.

Точечные оценки

Для популяционного среднего можно рассмотреть две несмещенные точечные оценки:

1. **Простое среднее**.

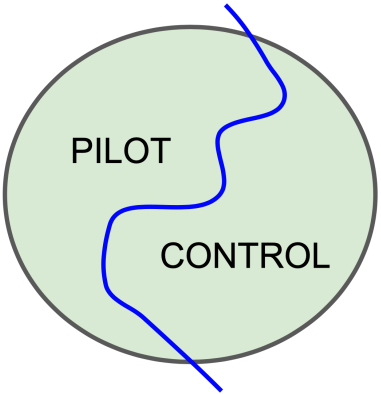
$$\overline{Y} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}.$$

2. **Взвешенное среднее** (стратифицированное среднее).

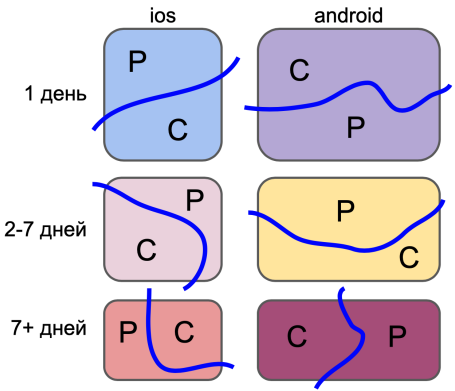
$$\hat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k, \quad \overline{Y}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj}.$$

Реализация

Случайное разбиение



Стратифицированное разбиение



Стратифицированное семплирование

Основная идея метода

Стратифицированное семплирование — метод понижения дисперсии. Для выборки мы должны обеспечить такие же доли каждой страты, что и в генеральной совокупности. Размер страт равен

$$n_k = nw_k.$$

Взвешенное среднее обычно используется для оценки популяционного среднего μ.

Способы семплирования

1. **Случайное семплирование** — мы выбираем элементы без дополнительных требований по доле каждой из страт. Среднее и дисперсию для этого способа обозначим $\mathbb{E}_{srs}$ и $\mathbb{V}_{srs}$.
2. **Стратифицированное семплирование** — частота каждой страты должна быть такой же, как в генеральной совокупности. Обозначения статистик $\mathbb{E}_{strat}$ и $\mathbb{V}_{strat}$.

Свойства оценок популяционного среднего

Совпадение точечных оценок

В условиях стратифицированного семплирования две приведенные точечные оценки совпадают:

$$\hat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k = \sum_{k=1}^{K} w_k \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj} = \sum_{k=1}^{K} \frac{n_k}{n} \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj} = \overline{Y}.$$

Случайное семплирование

$$\mathbb{E}_{srs}(\overline{Y}) = \mathbb{E}_{srs}\left(\frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}\right) =$$

$$= \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \mathbb{E}_{srs}(Y_{kj}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \mu = \mu$$

Стратифицированное семплирование

$$\mathbb{E}_{strat}(\hat{Y}_{strat}) = \sum_{k=1}^{K} w_k \mathbb{E}_{strat}(\overline{Y}_k) =$$

$$= \sum_{k=1}^{K} w_k \mu_k = \mu$$

Условное математическое ожидание

Definition

Условным математическим ожиданием измеримой функции $X : \Omega \to \overline{\mathbb{R}}$ относительно сигма-алгебры $\mathcal{G} \subseteq \mathcal{F}$ называется $\mathcal{G}$-измеримая функция $\mathbb{E}(X|\mathcal{G}) : \Omega \to \overline{\mathbb{R}}$ такая, что для любого $A \in \mathcal{G}$ выполняется равенство

$$\int_A X \mathrm{d}\mathbb{P} = \int_A \mathbb{E}(X|\mathcal{G}) \mathrm{d}\mathbb{P}$$

Definition

Пусть X и Y – случайные величины. $\mathbb{E}X < \infty$. Тогда *условным математическим ожиданием случайной величины* X *относительно случайной величины* Y назовем

$$\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma(Y)), \quad \sigma(Y) = (Y^{-1}(B), B \in \mathrm{B}).$$

Полные математическое ожидание и дисперсия

Закон полного математического ожидания

Полное математическое ожидание сводится к взвешенной сумме по отдельным стратам:

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

Закон полной дисперсии

Найдем $\mathbb{E}[\mathbb{V}(X|Y)]$ и $\mathbb{V}[\mathbb{E}(X|Y)]$ для X и Y:

$$\begin{aligned}\mathbb{E}[\mathbb{V}(X|Y)] &= \mathbb{E}[\mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2] = \mathbb{E}[\mathbb{E}[X^2|Y]] - \mathbb{E}[(\mathbb{E}[X|Y])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X|Y])^2] \\ \mathbb{V}(\mathbb{E}[X|Y]) &= \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[\mathbb{E}[X|Y]])^2 = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[X])^2\end{aligned}$$

Из полученных равенств

$$\mathbb{V}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[\mathbb{V}(X|Y)] + \mathbb{V}(\mathbb{E}[X|Y]).$$

Межгрупповая и внутригрупповая дисперсия

Дисперсия случайного семплирования может быть представлена в виде суммы дисперсии внутри стратифицированной группы, и между стратифицированными группами.

$$
\begin{aligned}
\mathbb{V}_{srs}(Y) =& \mathbb{E}_{srs}(\mathbb{V}_{srs}(Y|Z)) + \mathbb{V}_{srs}(\mathbb{E}_{srs}(Y|Z)) \\
=& \mathbb{E}_{srs}\left(\sum_{k=1}^{K} \sigma_k^2 I(Z=k)\right) + \mathbb{V}_{srs}\left(\sum_{k=1}^{K} \mu_k I(Z=k)\right) \\
=& \sum_{k=1}^{K} \sigma_k^2 \mathbb{E}_{srs}(I(Z=k)) + \mathbb{E}_{srs}\left(\sum_{k=1}^{K} \mu_k I(Z=k)\right)^2 \\
& - \left(\mathbb{E}_{srs}\left(\sum_{k=1}^{K} \mu_k I(Z=k)\right)\right)^2 \\
=& \sum_{k=1}^{K} \sigma_k^2 w_k + \sum_{k=1}^{K} \mu_k^2 w_k - \mu^2 = \sum_{k=1}^{K} \sigma_k^2 w_k + \sum_{k=1}^{K} w_k (\mu_k - \mu)^2
\end{aligned}
$$

Понижение дисперсии

Дисперсия случайного семплирования

$$\mathbb{V}_{srs}(\overline{Y}) = \frac{1}{n}\sigma^2 = \frac{1}{n}\sum_{k=1}^{K} w_k \sigma_k^2 + \frac{1}{n}\sum_{k=1}^{K} w_k (\mu_k - \mu)^2$$

Дисперсия стратифицированного семплирования

$$\mathbb{V}_{strat}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k \sigma_k^2$$

Понижение дисперсии

$$\mathbb{V}_{srs}(\overline{Y}) - \mathbb{V}_{strat}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k (\mu_k - \mu)^2$$

Преимущества стратификационного семплирования

- Стратификационное среднее дает несмещенную оценку популяционного среднего:

$$\mathbb{E}_{strat}(\hat{Y}_{strat}) = \mu$$

- У этой оценки дисперсия ниже, чем при случайном семплировании:

$$\mathbb{V}_{srs}(\overline{Y}) - \mathbb{V}_{strat}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k(\mu_k - \mu)^2$$

- В AB-тестах мы можем получать большую чувствительность за счет сниженной дисперсии.
- Подсчет статистики по стратам и семплирование можно проводить прямо во время эксперимента. Например, выделяя каждого 100го представителя страты и распределяя их между пилотом и контролем.

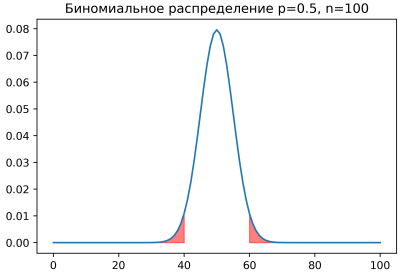
Но что делать, если семплирование не было произведено заранее?

Проблемы случайного разбиения

200‘000 пользователей для проведения пилота.
Среди них **100** пользователей старше 35 лет и с ОС ios.

Какова вероятность, что отличие будет **более чем в полтора раза?**
То есть в одной из групп будет менее 40 пользователей старше 35 лет с ios.

Кол-во пользователей в пилотной группе $N_{pilot} \sim B(n = 100, p = 0.5)$.
$\mathbb{P}(\{N_{pilot} < 40\} \bigcup \{N_{control} < 40\}) \approx 0.06$

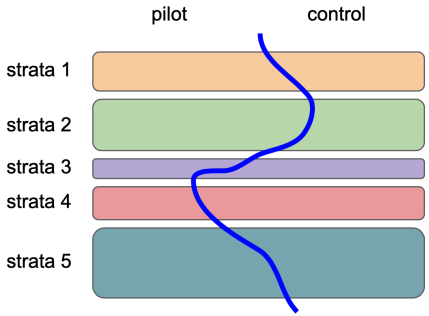


Вероятность перекоса более чем в полтора раза при различных N:

$N = 500$	$q \approx 0.001\%$
$N = 100$	$q \approx 6\%$
$N = 20$	$q \approx 50\%$

Пост стратификация

Что делать, если провели эксперимент без стратификации?



Мы можем так же заменить случайное среднее на стратифицированное среднее

$$\hat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k$$

Это соответствует перевзвешиванию каждой страты в соответствии с долей в генеральной совокупности.

Мы надеемся, что так можно снизить дисперсию.

Дисперсия при постстратификации

Оценим дисперсию стратифицированного среднего при случайном семплировании.

$$
\begin{aligned}
\mathbb{V}_{srs}(\hat{Y}_{strat}) =& \mathbb{E}_{srs}(\mathbb{V}_{srs}(\hat{Y}_{strat}|n_1, \dots, n_K)) + \mathbb{V}_{srs}(\mathbb{E}_{srs}(\hat{Y}_{strat}|n_1, \dots, n_K)) \\
=& \mathbb{E}_{srs}\left(\sum_{k=1}^{K} w_k^2 \mathbb{V}_{srs}(\overline{Y}_k|n_k)\right) + \mathbb{V}_{srs}\left(\sum_{k=1}^{K} w_k \mu_k\right) \\
=& \mathbb{E}_{srs}\left(\sum_{k=1}^{K} w_k^2 \frac{1}{n_k}\sigma_k^2\right) + \mathbb{V}_{srs}(\mu) \\
=& \sum_{k=1}^{K} w_k^2 \sigma_k^2 \mathbb{E}_{srs}\left(\frac{1}{n_k}\right)
\end{aligned}
$$

n_k - биномиальная случайная величина

Дисперсия при постстратификации

Дисперсия биномиальной случайной величины $\mathbb{V}(n_k) = nw_k(1 - w_k)$.
Применим разложение Тейлора для функции $\frac{1}{n_k}$ в точке $\frac{1}{nw_k}$

$$\begin{aligned}
\mathbb{E}_{srs}\left(\frac{1}{n_k}\right) =& \mathbb{E}_{srs}\left(\frac{1}{nw_k} - \frac{1}{n^2 w_k^2}(n_k - nw_k) + \frac{1}{n^3 w_k^3}(n_k - nw_k)^2\right) + O\left(\frac{1}{n^2}\right) \\
=& \frac{1}{nw_k} + \frac{1}{n^3 w_k^3}\mathbb{E}(n_k - nw_k)^2 + O\left(\frac{1}{n^2}\right) \\
=& \frac{1}{nw_k} + \frac{1}{n^3 w_k^3} nw_k(1 - w_k) + O\left(\frac{1}{n^2}\right) \\
=& \frac{1}{nw_k} + \frac{1}{n^2 w_k^2}(1 - w_k) + O\left(\frac{1}{n^2}\right)
\end{aligned}$$

Получаем

$$\mathbb{V}_{srs}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k \sigma_k^2 + \frac{1}{n^2}\sum_{k=1}^{K}(1 - w_k)\sigma_k^2 + O\left(\frac{1}{n^2}\right)$$

Сравнение дисперсий различных методов

Дисперсии для различных методов

$$\mathbb{V}_{srs}(\overline{Y}) = \frac{1}{n}\sum_{k=1}^{K} w_k\sigma_k^2 + \frac{1}{n}\sum_{k=1}^{K} w_k(\mu_k - \mu)^2$$

$$\mathbb{V}_{strat}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k\sigma_k^2$$

$$\mathbb{V}_{srs}(\hat{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} w_k\sigma_k^2 + \frac{1}{n^2}\sum_{k=1}^{K}(1 - w_k)\sigma_k^2 + O\left(\frac{1}{n^2}\right)$$

Соотношения дисперсий

$$\mathbb{V}_{strat}(\hat{Y}_{strat}) = \mathbb{V}_{srs}(\hat{Y}_{strat}) + O\left(\frac{1}{n^2}\right) = \mathbb{V}_{srs}(\overline{Y}) + O\left(\frac{1}{n}\right),$$

$$\mathbb{V}_{strat}(\hat{Y}_{strat}) \leqslant \mathbb{V}_{srs}(\hat{Y}_{strat}) \leqslant \mathbb{V}_{srs}(\overline{Y})$$

Оценка пилота

Бутстреп

Оцениваем распределение разности средних стратифицированных $\hat{Y}_{strat}$.

- семплируем данные пилотной и контрольной групп
- считаем разность стратифицированных средних $\hat{Y}^{bs}_{strat} - \hat{X}^{bs}_{strat}$
- строим доверительный интервал
- проверяем входит ли ноль в доверительный интервал

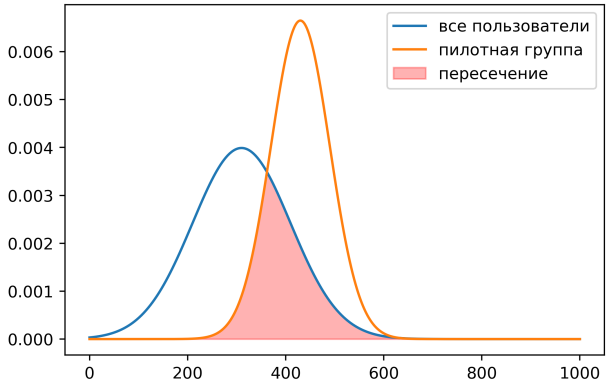
Тест Стьюдента

- считаем стратифицированные средние $\hat{Y}_{strat}$
- считаем оценку дисперсий $\sigma^2_Y = \mathbb{V}(\hat{Y}_{strat}) \approx \frac{1}{n}\sum_{k=1}^{K} w_k \sigma_k^2$
- считаем t-статистику и pvalue

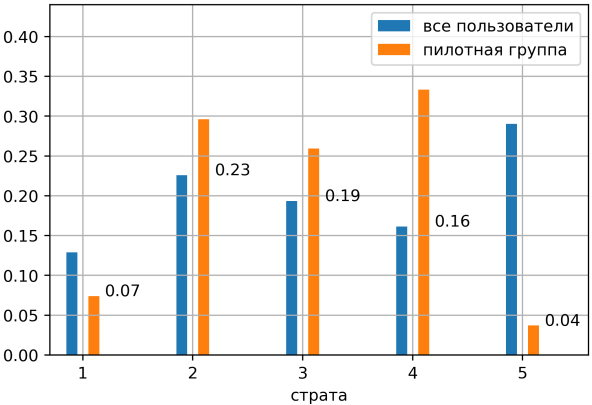
$$t = \frac{\hat{Y}_{strat} - \hat{X}_{strat}}{\sqrt{\frac{\sigma^2_Y + \sigma^2_X}{n}}}$$

Обобщающая способность

Можно ли масштабировать полученные выводы на всех пользователей?



Дискретный случай



- агрегируем признаки, формируем страты
- считаем кол-во пользователей в каждой страте, нормируем
- суммируем минимумы каждой страты

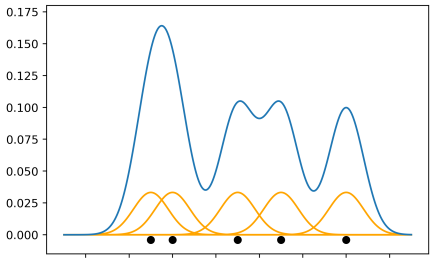
Пересечение 0.69

KDE

Kernel Density Estimation (Ядерная оценка плотности) - непараметрический способ оценки плотности случайной величины.

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

K - ядро, h - ширина ядра. $K(x) \geq 0$, $\int K(x)dx = 1$, $\int xK(x)dx = 0$.



Примеры ядер:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$K(x) = (1 - |x|)\ \mathbb{I}(|x| < 1)$$

$$K(x) = \frac{3}{4}(1 - x^2)\ \mathbb{I}(|x| < 1)$$

Резюме

Сегодня узнали:

- Познакомились с методом стратификационного семплирования
- Убедились, что этот метод значительно снижает дисперсию и его полезно использовать в AB-экспериментах
- Если же мы, забыли провести стратификицию, то на выручку может прийти **пост**стратификация

Ссылки для самостоятельного изучения

1. W. G. Cochran. Sampling Techniques. Wiley, 1977.
2. Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix