



> Конспект > 4 урок > Стратификация

> Оглавление

> Оглавление

> Стратификация

> Как проводить стратификацию?

> Популяционное среднее

> Точечные оценки

> Реализация

> Стратифицированное семплирование

> Условное математическое ожидание

> Полные математическое ожидание и дисперсия

> Закон полного математического ожидания

> Понижение дисперсии

> Межгрупповая и внутригрупповая дисперсия

> Дисперсия случайного семплирования

> Дисперсия стратифицированного семплирования

> Понижение дисперсии

> Преимущества стратифицированного семплирования

> Постстратификация

> Проблемы случайного разбиения

> Дисперсия при постстратификации

> Сравнение методов семплирования

> Дисперсии для различных методов

> Соотношения дисперсий

> Оценка пилота

> Резюме

> Материалы для самостоятельного изучения

КУПЛЕНО НА
SKLADCHIK.COM

> Стратификация

Ситуация: хотим изменить рекламу встроенных покупок, чтобы увеличить их продажи.

Мы уже определили метрики, размер пилотной и контрольной групп. Теперь хотим отнести к ним пользователей. Как это сделать?

Первое, что приходит на ум — случайно распределяем пользователей по группам и начинаем пилот.

Однако пользователи могут отличаться по различным характеристикам: полу, возрасту и т.д.

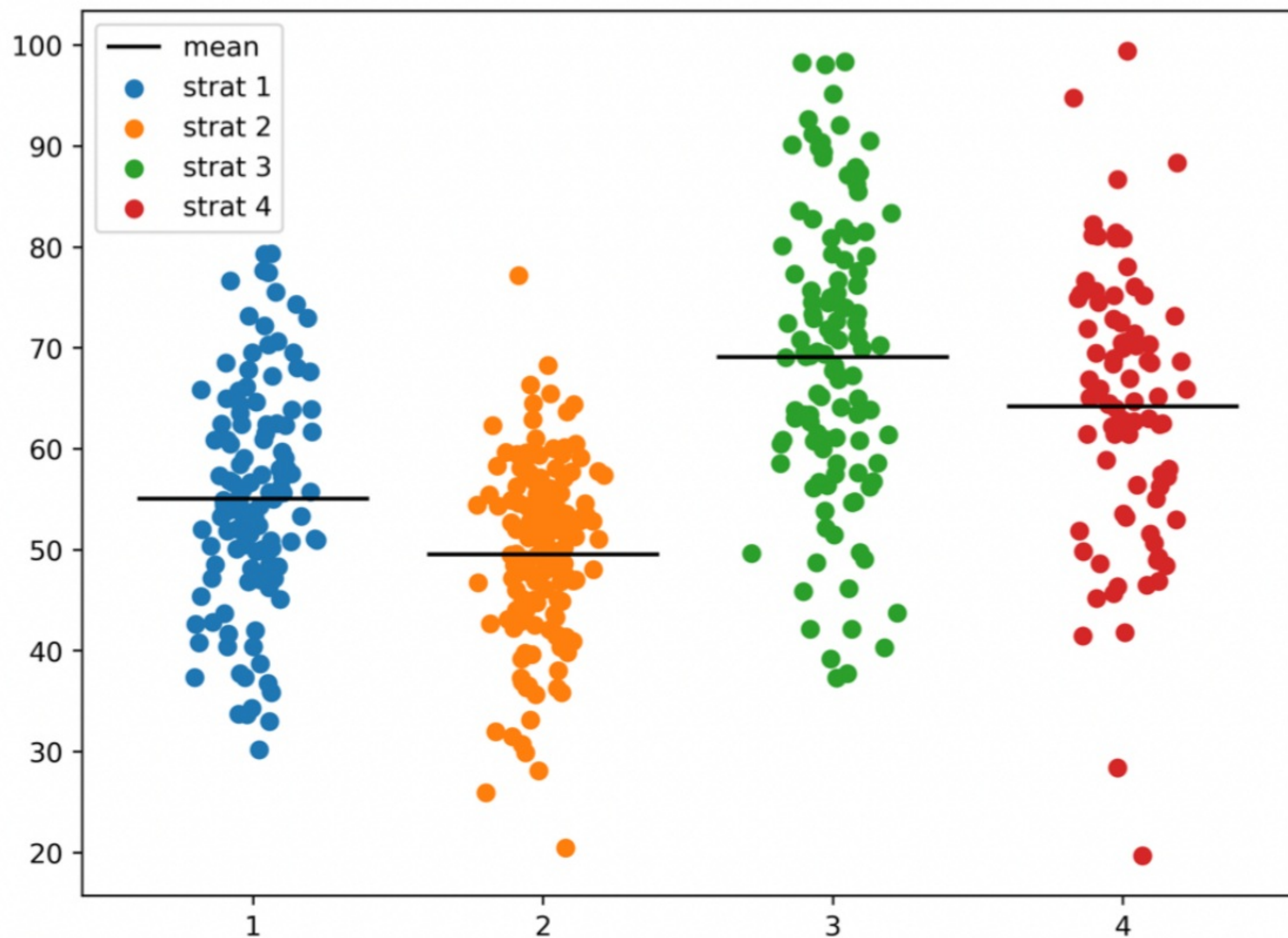
> Как проводить стратификацию?

Начнём с того, что такое 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 могут быть пол, возраст или иные характеристики пользователя. Для международных онлайн-платформ хорошим признаком будет страна проживания пользователя.

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



Замечание: распределения целевой метрики в различных стратах должны отличаться, иначе стратификация не имеет смысла.

> Популяционное среднее

Нам необходимо оценить популяционное среднее бизнес метрики Y .

Введём **обозначения** для лекции:

$\mu = EY$ — популяционное среднее.

$\sigma^2 = VY$ — 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 ГС, где Y_{kj} — метрика для j -го пользователя k -й страты.

> Точечные оценки

Для популяционного среднего можно рассмотреть две несмещённые точечные оценки:

1. Простое среднее

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

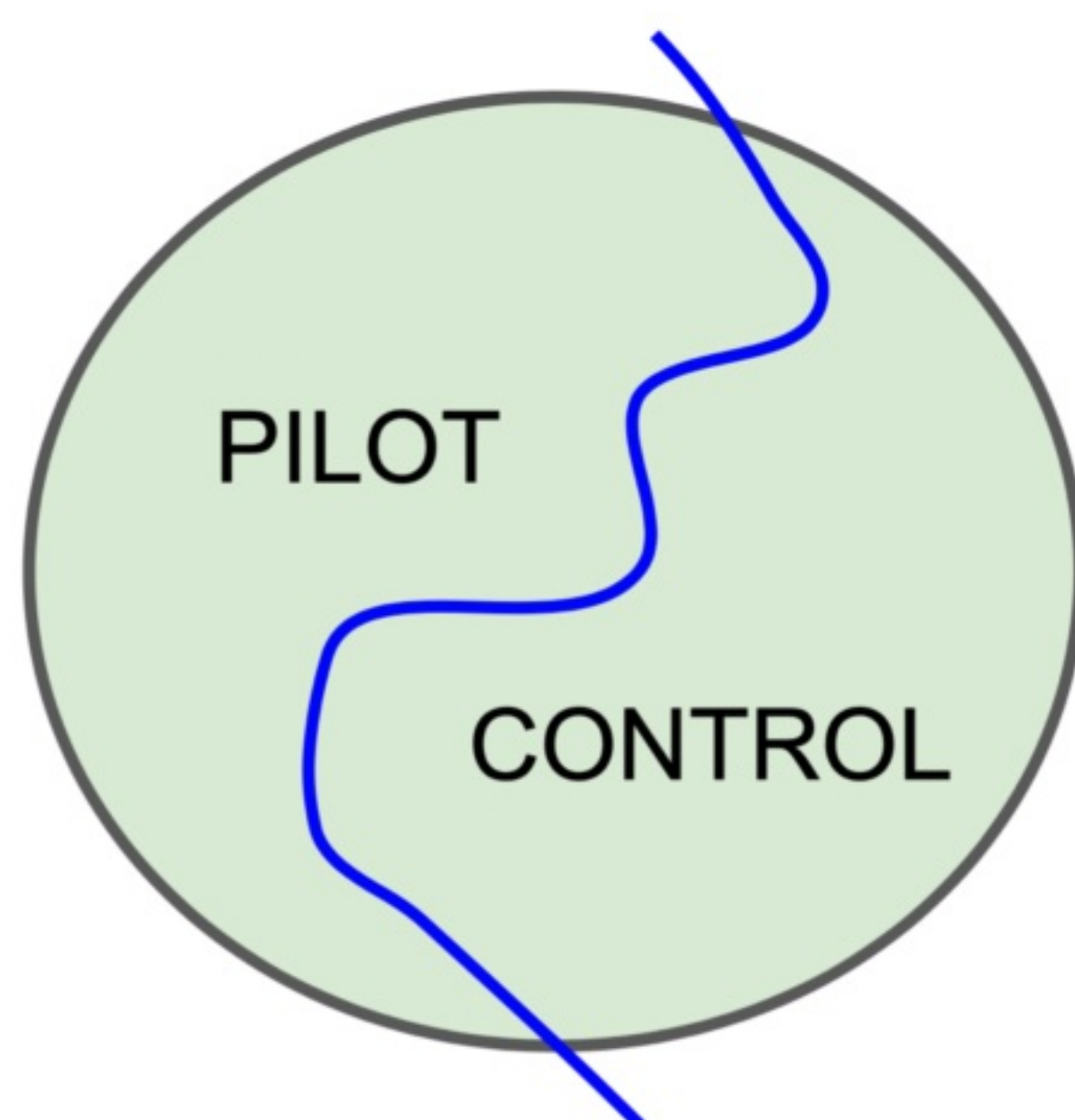
2. Взвешенное среднее (стратифицированное среднее)

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

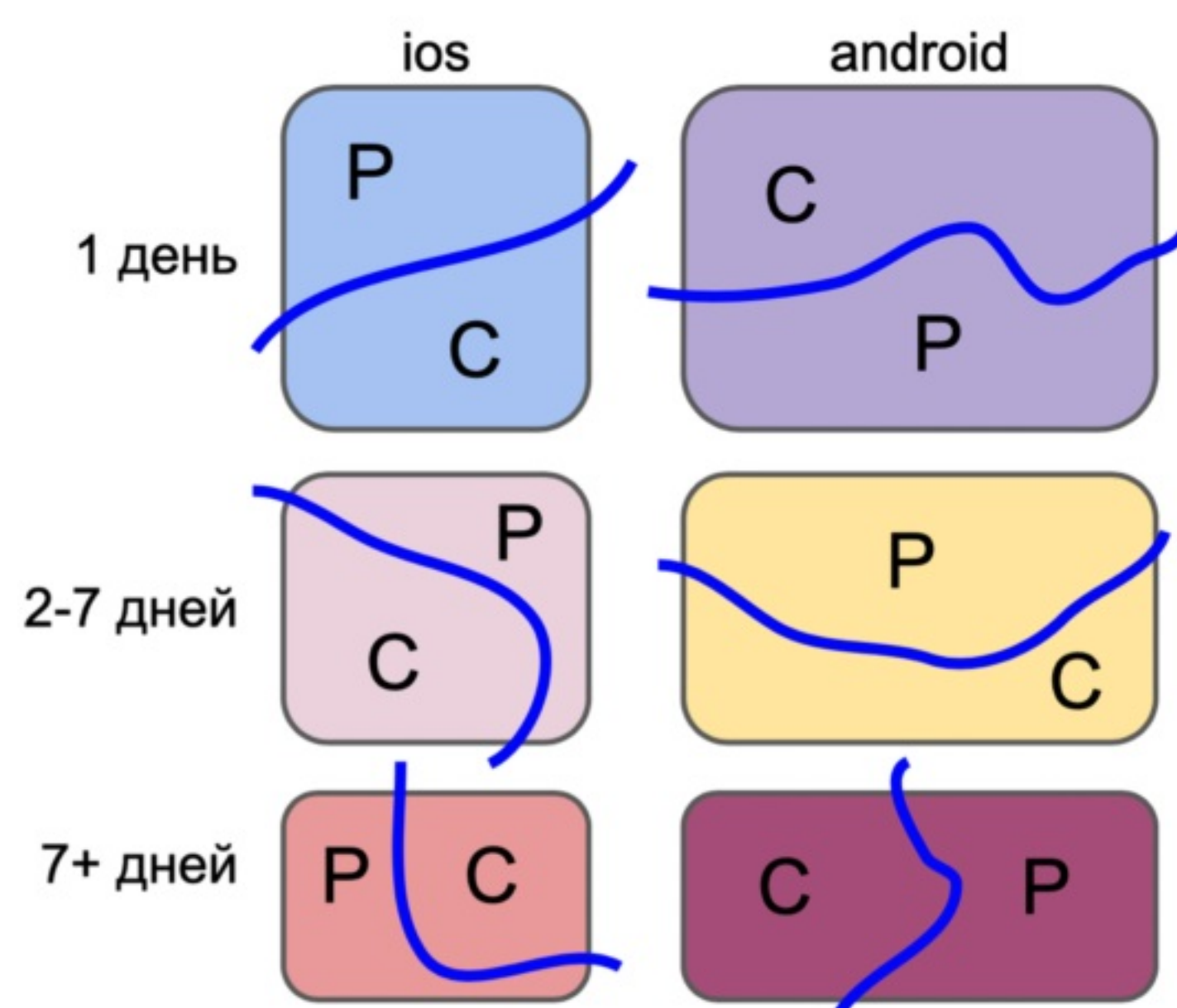
> Реализация

При стратифицированном разбиении необходимо контролировать, чтобы в каждой группе на протяжении всего эксперимента сохранялся баланс между контрольной и пилотной группами. В случайном разбиении этого может не происходить.

Случайное разбиение



Стратифицированное разбиение



> Стратифицированное семплирование

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 к доле каждой из страт. Среднее и дисперсию для этого способа обозначим E_{srs} и V_{srs}
2. **Стратифицированное семплирование** — частота каждой страты должна быть такой же, как и в генеральной совокупности. Обозначения статистик E_{strat} и V_{strat}

Покажем, что в условиях стратифицированного семплирования две приведённые точечные оценки совпадают:

$$\begin{aligned}\hat{Y}_{strat} &= \sum_{k=1}^K w_k \bar{Y}_k = \sum_{k=1}^K w_k \frac{1}{n_k} \sum_{j=1}^{n_k} j = 1Ykj = \\ &= \sum_{k=1}^K \frac{n_k}{n} \frac{1}{n_k} \sum_{j=1}^{n_k} j = 1Ykj = \frac{1}{n} \sum_{k=1}^K \sum_{j=1}^{n_k} j = 1Ykj = \bar{Y}\end{aligned}$$

Покажем, что эти точечные оценки — несмещённые оценки математического ожидания:

Случайное семплирование:

$$\begin{aligned}E_{srs}(\bar{Y}) &= E_{srs}\left(\frac{1}{n} \sum_{k=1}^K \sum_{j=1}^{n_k} j = 1Ykj\right) = \frac{1}{n} \sum_{k=1}^K \sum_{j=1}^{n_k} j = 1E_{srs}(Y_{kj}) = \\ &= \frac{1}{n} \sum_{k=1}^K \sum_{j=1}^{n_k} \mu = \mu\end{aligned}$$

Стратифицированное семплирование:

$$E_{strat}(\hat{Y}_{strat}) = \sum_{k=1}^K w_k E_{strat}(\bar{Y}_k) = \sum_{k=1}^K w_k \mu_k = \mu$$

> **Условное математическое ожидание**

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

$$\int_A X dP = \int_A E(X|\mathcal{G}) dP$$

Пусть X и Y — случайные величины. $EX < \infty$. Тогда условным математическим ожиданием случайной величины X относительно случайной величины Y назовём:

$$E(X|Y) = E(X|\sigma(Y)), \sigma(Y) = (Y^{-1}(B), B \in \mathcal{B})$$

> Полные математическое ожидание и дисперсия

> Закон полного математического ожидания

Найдём $E(V(X|Y))$ и $V(E(X|Y))$ для X и Y :

$$E(V(X|Y)) = E[E[X^2|Y] - (E[X|Y])^2] = E[E[X^2|Y]] - E[(E[X|Y])^2] = E[X^2] - E[(E[X|Y])^2]$$

$$V(E[X|Y]) = E[(E[X|Y])^2] - (E[E[X|Y]])^2 = E[(E[X|Y])^2] - (E[X])^2$$

Из полученных равенств:

$$V(X) = E[X^2] - (E[X])^2 = E[V(X|Y)] + V(E[X|Y])$$

> Понижение дисперсии

> Межгрупповая и внутригрупповая дисперсия

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

$$V_{srs}(Y) = E_{srs}(V_{srs}(Y|Z)) + V_{srs}(E_{srs}(Y|Z)) = E_{srs} \sum_{k=1}^K \sigma_k^2 l(Z=k) + V_{srs} \sum_{k=1}^K \mu_k l(Z=k) = \sum_{k=1}^K \sigma_k^2 E_{srs}(l(Z=k)) + E_{srs} \left(\sum_{k=1}^K \mu_k l(Z=k) \right)^2 -$$

$$(E_{srs} \sum_{k=1}^K \mu_k l(Z = k))^2 = \sum_{k=1}^K \sigma_k^2 w_k + \sum_{k=1}^K \mu_k^2 w_k - \mu^2 = \sum_{k=1}^K \sigma_k^2 w_k + \sum_{k=1}^K w_k (\mu_k - \mu)^2$$

> Дисперсия случайного семплирования

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

> Дисперсия стратифицированного семплирования

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

> Понижение дисперсии

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

> Преимущества стратифицированного семплирования

- Стратифицированное среднее даёт **несмещённую оценку** популяционного среднего: $E_{strat}(\hat{Y}_{strat}) = \mu$
- У этой оценки **дисперсия ниже**, чем при случайном семплировании:

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

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

Но что делать, если семплирование не было произведено заранее?

> Постстратификация

> Проблемы случайного разбиения

Ситуация: имеем 200 000 пользователей для проведения пилота, среди них 100 пользователей старше 35 лет и пользуются iOS.

Какова вероятность, что отличие будет более чем в полтора раза, т.е. в одной из групп будет менее 40 пользователей старше 35 лет с iOS?

Количество пользователей в пилотной группе: $N_{pilot} \sim B(n = 100, p = 0.5)$

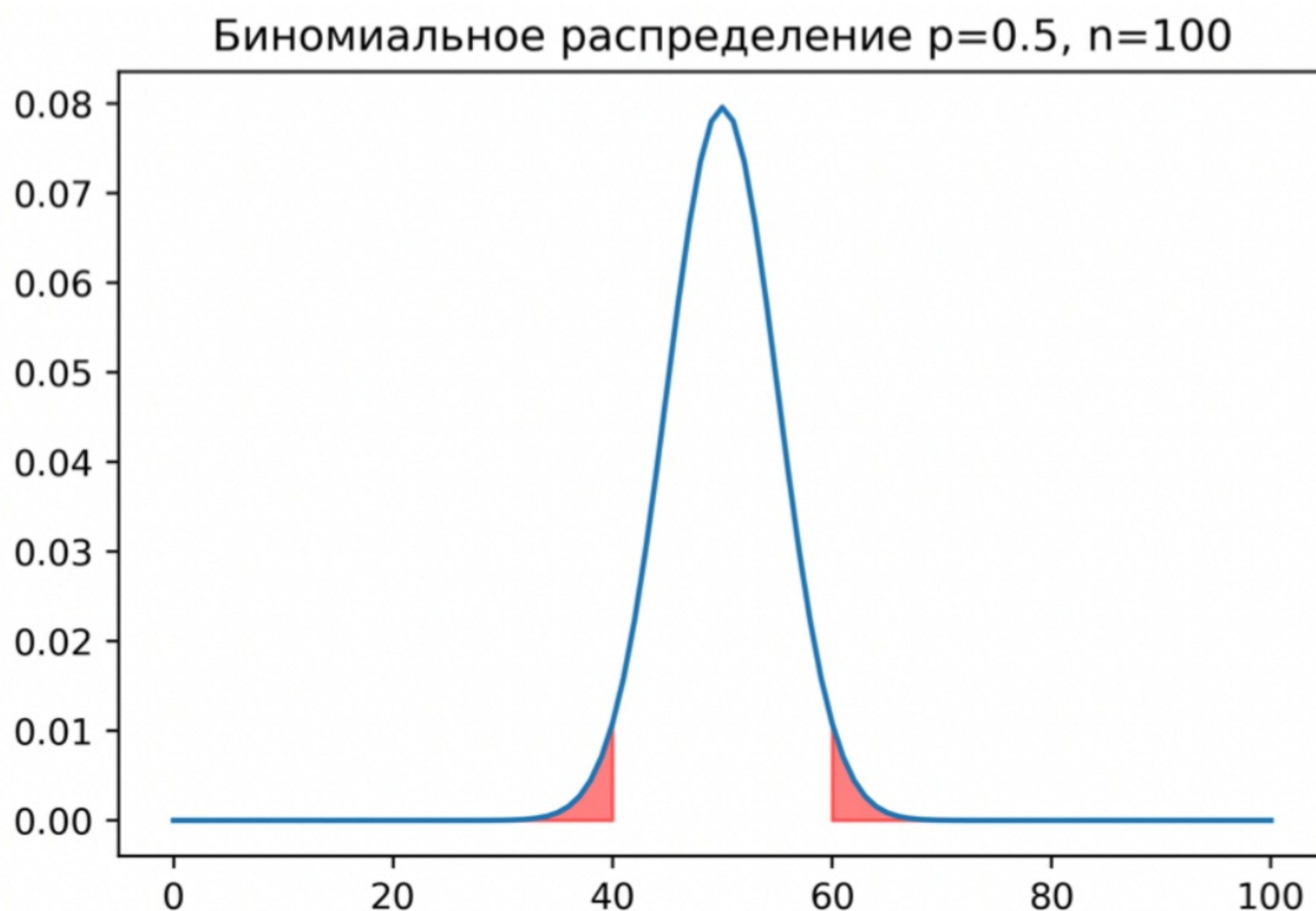
$$P(\{N_{pilot} < 40\} \cup \{N_{control} < 40\}) \approx 0.06$$

Вероятность перекоса более чем в полтора раза при различных N :

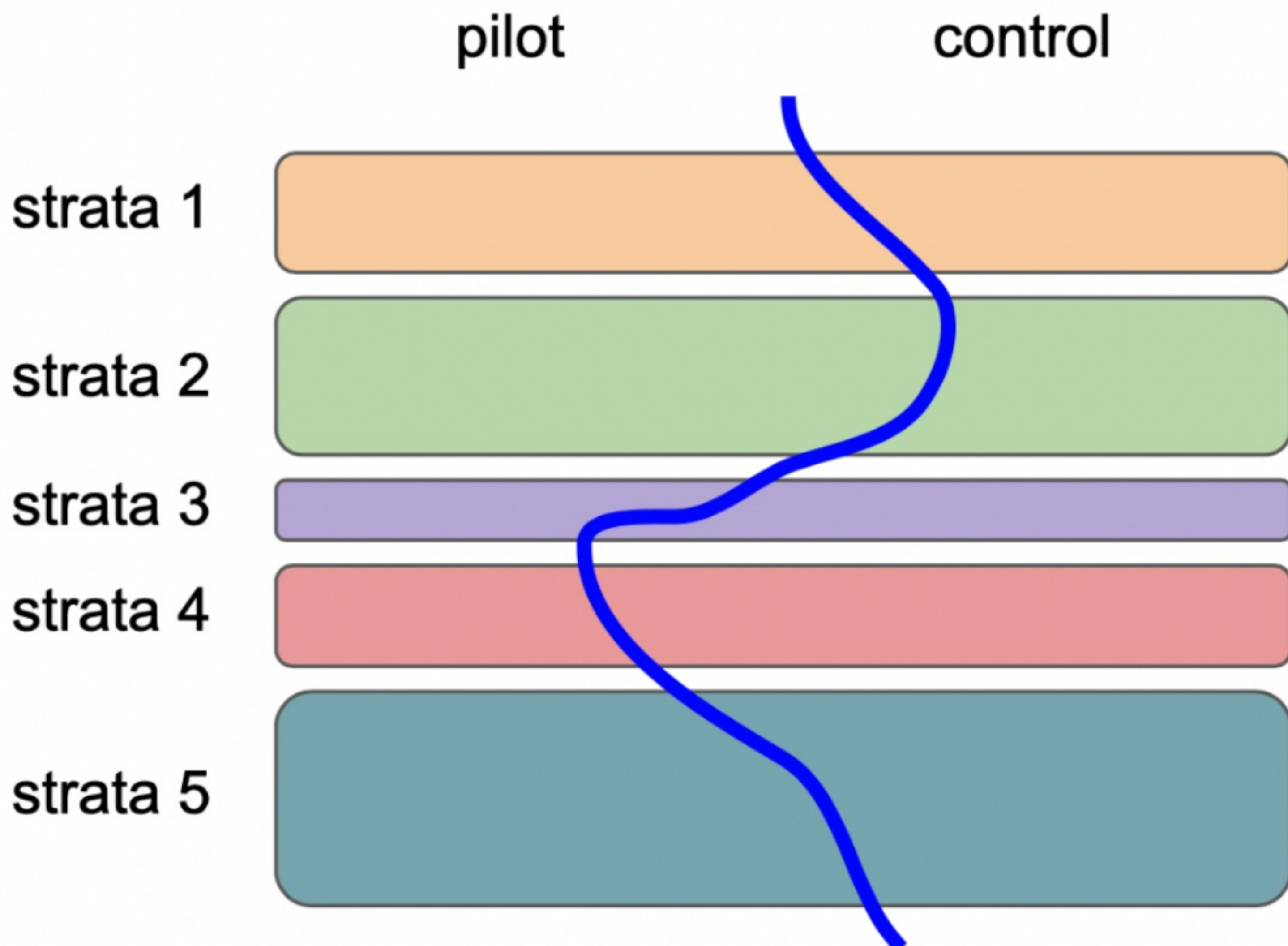
$$N = 500, q \approx 0.001\%$$

$$N = 100, q \approx 6\%$$

$$N = 20, q \approx 50\%$$



Что делать, если провели эксперимент без стратификации?



То есть в различных стратах группы pilot и control представлены в различных долях

Мы также можем **заменить** случайное среднее на стратифицированное среднее:

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

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:

$$V_{srs}(\hat{Y}_{strat}) = E_{srs}(V_{srs}(\hat{Y}_{strat} | n_1, \dots, n_K)) +$$

$$V_{srs}(E_{srs}(\hat{Y}_{strat} | n_1, \dots, n_K)) = E_{srs}\left(\sum_{k=1}^K w_k^2 V_{srs}(\bar{Y}_k | n_k)\right) +$$

$$V_{srs}\left(\sum_{k=1}^K w_k \mu_k\right) = E_{srs}\left(\sum_{k=1}^K w_k^2 \frac{1}{n_k} \sigma_k^2\right) + V_{srs}(\mu) = \sum_{k=1}^K w_k^2 \sigma_k^2 E_{srs}\left(\frac{1}{n_k}\right)$$

n_k — биномиальная случайная величина.

Дисперсия биномиальной случайной величины $V(n_k) = nw_k(1-w_k)$

Применим разложение Тейлора для функции $\frac{1}{n_k}$ в точке $\frac{1}{nw_k}$:

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

Получаем:

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

> Сравнение методов семплирования

> Дисперсии для различных методов

$$V_{srs}(\bar{Y}) = \frac{1}{n} \sum_{k=1}^K w_k \sigma_k^2 + \frac{1}{n} \sum_{k=1}^K w_k (\mu_k - \mu)^2$$

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

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

> Соотношения дисперсий

$$V_{strat}(\hat{Y}_{strat}) = V_{srs}(\hat{Y}_{strat}) + O\left(\frac{1}{n^2}\right) = V_{srs}(\bar{Y}) + O\left(\frac{1}{n}\right)$$

$$V_{strat}(\hat{Y}_{strat}) \leq V_{srs}(\hat{Y}_{strat}) \leq V_{srs}(\bar{Y})$$

> Оценка пилота

С помощью **бутстрэпа** оцениваем распределение разности **средних стратифицированных** $\hat{Y}_{strat}$:

- Сэмплируем данные пилотной и контрольной групп;
- Считаем разность **стратифицированных средних**: $\hat{Y}_{strat}^{bs} - \hat{X}_{strat}^{bs}$
- Строим доверительный интервал;
- Проверяем, входит ли ноль в доверительный интервал.

С помощью **теста Стьюдента**:

- Считаем стратифицированные средние: $\hat{Y}_{strat}$
- Считаем оценку дисперсий: $\sigma_Y^2 = V(\hat{Y}_{strat}) \approx \frac{1}{n} \sum_{k=1}^K w_k \sigma_k^2$
- Считаем t-статистику и p-value:

$$t = \frac{\hat{Y}_{strat} - \hat{X}_{strat}}{\sqrt{\frac{\sigma_Y^2 + \sigma_X^2}{n}}}$$

> Резюме

1. Познакомились с методом стратификационного семплирования.
2. Убедились, что этот метод значительно снижает дисперсию и его полезно использовать в A/B-экспериментах.
3. Если забыли провести стратификацию, то на выручку может прийти постстратификация.

> Материалы для самостоятельного изучения

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