Семинар 7

вторник, 14 марта 2023 г. 21:17

Формула размера выборки

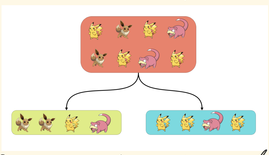
$$n = \frac{[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)]^2 (\sigma_1^2 + \sigma_2^2)}{MDE^2}$$

Стратификация

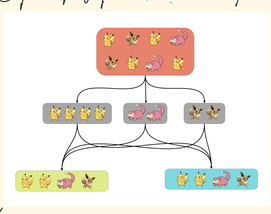
Стратификация — Страт. семплирование + стратифиц. оценка

Пост-стратификация — случайная выборка с стратифицированной оценкой на основе знания известных вероятностей попадания в группу

Случайное семплирование



Стратифицированное семплирование



Обозначения

- $M = E(Y)$ — популяционное среднее
- $\sigma^2 = D(Y)$
- M_k, σ_k^2 — среднее и дисперсия для k-й страты
- W_k — доля k-й страты в популяции
- n_k — число пользователей k-й страты
- $n = \sum_{k=1}^{K} n_k$ — общий размер группы
- $Y_{11}, \dots, Y_{1n_1}, \dots, Y_{Kn_K}$ — выборка из ГС, где Y_{kj} — метрика для j-го пользователя k-й страты

Простое среднее

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

Стратифицированное среднее

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

	Обычное среднее $\bar{Y} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} Y_{kj}$	Стратифицированное среднее $\bar{Y}_{strat} = \sum_{k=1}^{K} w_k \bar{Y}_k$
Случайное семплирование (страта 1 → группа A, страта 2 → группа B)	Классический подход Без стратификации	**Постстратификация**
Стратифицированное семплирование (страта 1 → группа A, страта 2 → группа B)	Не контролирует вероятность ошибки I рода	**Стратификация**

Свойства

1. $\bar{Y}_{strat} = \bar{Y}$ (оценки равны)
2. $E(\bar{Y}_{strat}) = E(\bar{Y}) = E(\bar{Y}_{post}) = M$ (↘ post-стратификация)

Дисперсии

1. $D(\bar{Y}) = \frac{1}{n}\sum_{k=1}^{K} W_k \sigma_k^2 + \frac{1}{n}\sum_{k=1}^{K} W_k (M_k - M)^2$
   — внутригрупповая дисперсия + межгрупповая дисперсия
2. $D(\bar{Y}_{strat}) = \frac{1}{n}\sum_{k=1}^{K} W_k \sigma_k^2$

$$D(\bar{Y}_{post\text{-}st}) = \frac{1}{n}\sum_{k=1}^{K} W_k \sigma_k^2 + \frac{1}{n^2}\sum_{k=1}^{K}(1-W_k)\sigma_k^2 + O\left(\frac{1}{n^2}\right)$$

$$D(\bar{Y}_{strat}) = D(\bar{Y}_{post\text{-}strat}) + O\left(\frac{1}{n^2}\right) = D(\bar{Y}) + O\left(\frac{1}{n^2}\right)$$

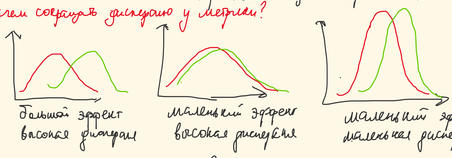
$$D(\bar{Y}_{strat}) \le D(\bar{Y}_{post\text{-}strat}) \le D(\bar{Y})$$

CUPED

Controlled-experiment **U**sing **P**re-**E**xperimental **D**ata

Зачем сокращать дисперсию у метрики?

- A
- B



Большой эффект, высокая дисперсия — маленький эффект, высокая дисперсия — маленький эффект, маленькая дисперсия

Как сокращать дисперсию?

Можно использовать дополнительную информацию (историю)

1) Стратификация
2) Использование ковариат

$\hat{\Theta}$ — несмещённая оценка на Θ

$$E(\hat{\Theta}) = \Theta$$

$$\hat{\Theta}_{cuped} = \hat{\Theta} - \alpha\hat{\Theta}_a, \quad E(\hat{\Theta}_a) = 0$$ — для несмещённости, $\hat{\Theta}_a$ — ковариата

$$D(\hat{\Theta}_{cuped}) = D(\hat{\Theta}) + \alpha^2 D(\hat{\Theta}_a) - 2\alpha\,cov(\hat{\Theta}, \hat{\Theta}_a)$$

— оценки должны коррелировать

Оптимизируем по α, чтобы $D(\hat{\Theta}_{cuped}) \to \min$

$$\frac{\partial}{\partial\alpha} = 2\alpha D(\hat{\Theta}_a) - 2\,cov(\hat{\Theta}, \hat{\Theta}_a) = 0$$

$$\alpha = \frac{cov(\hat{\Theta}, \hat{\Theta}_a)}{D(\hat{\Theta}_a)}$$

$$\hat{\Theta}_{cuped} = \hat{\Theta} - \frac{cov(\hat{\Theta}, \hat{\Theta}_a)}{D(\hat{\Theta}_a)} \cdot \hat{\Theta}_a$$

$$D(\hat{\Theta}_{cuped}) \le D(\hat{\Theta})$$

На практике, чтобы $E(\hat{\Theta}_a) = 0$ делают так:

$\hat{\Theta}_{cuped} = \hat{\Theta} - \alpha\cdot(\hat{\Theta}_a - \bar{\hat{\Theta}_a})$ — среднее $\hat{\Theta}_a$

! Среднее по ковариате ($\bar{\hat{\Theta}_a}$) и α считаются по всем данным (тестовая группа + контрольная & те, которые вне из групп)