Престратификация

Материал написан на основе статьи Xie, H., & Aurisset, J. (2016, August). Improving the sensitivity of online controlled experiments: Case studies at netflix. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 645-654).

1) Стратификация (stratification sampling)

Мы помним, что снижение дисперсии снижает необходимое число наблюдений.

Предположим, что мы выбрали переменную, по которой мы можем разбить нашу выборку на группы (страты). Пусть таких групп K штук, а n_k – численность каждой из них, то есть $\sum_{k=1}^{K} n_k = N$, где N величина генеральной совокупности. Тогда вероятность попасть в одну из групп равняется $p_k = \frac{n_k}{N}$ – доля людей из k-й страты в генеральной совокупности.

Пусть наша зависимая переменная равна Y, причем $\mathbb{E}(Y) = \mu$ и $var(Y) = \sigma^2$. Тогда среднее значение зависимой переменной равно:

- Обычное среднее по всей выборке:
  $\bar{Y} = \dfrac{1}{N} \sum\limits_{k=1}^{K} \sum\limits_{j=1}^{n_k} Y_{kj}$
- Средневзвешенное при стратификации равно сумме средних в каждой страте, взвешенных по вероятностям попасть в каждую из этих страт:
  $\hat{Y}_{\text{strat}} = \sum\limits_{k=1}^{K} \underbrace{\dfrac{n_k}{N}}_{p_k} \bar{Y}_k = \sum\limits_{k=1}^{K} p_k \bar{Y}_k$, где $\bar{Y}_k = \dfrac{1}{n_k} \sum\limits_{j=1}^{n_k} Y_{kj}$ среднее значение метрики внутри k-й страты

Распишем среднее при стратификации подробнее
$\underbrace{\sum\limits_{k=1}^{K} p_k \bar{Y}_k}_{\hat{Y}_{\text{strat}}} = \sum\limits_{k=1}^{K} \dfrac{n_k}{N} \dfrac{1}{n_k} \sum\limits_{j=1}^{n_k} Y_{kj} = \underbrace{\dfrac{1}{N} \sum\limits_{k=1}^{K} \sum\limits_{j=1}^{n_k} Y_{kj}}_{\bar{Y}}$ обе средние равны

Найдем среднее и дисперсию зависимой переменной при стратификации.

- Среднее
  $\mathbb{E}\left(\hat{Y}_{\text{strat}}\right) = \mathbb{E}\left[\sum\limits_{k=1}^{K} p_k \bar{Y}_k\right] = \sum\limits_{k=1}^{K} p_k \mathbb{E}\left(\bar{Y}_k\right) = \sum\limits_{k=1}^{K} p_k \mu_k = \boxed{\mu}$
- Дисперсия
  $var\left(\hat{Y}_{\text{strat}}\right) = \text{var}\left[\sum\limits_{k=1}^{K} p_k \bar{Y}_k\right] = \sum\limits_{k=1}^{K} p_k^2 var\left(\bar{Y}_k\right) = \sum\limits_{k=1}^{K} \dfrac{n_k^2}{N^2} \dfrac{1}{n_k^2} n_k \sigma_k^2 = \dfrac{1}{N} \sum\limits_{k=1}^{K} \dfrac{n_k}{N} \sigma_k^2 = \boxed{\dfrac{1}{N} \sum\limits_{k=1}^{K} p_k \sigma_k^2}$ **(1)**

2) Простая рандомизация (simple random sampling)

Аналогично найдем среднее и дисперсию зависимой переменной при классическом эксперименте.

- Среднее
  $\mathbb{E}(\bar{Y}) = \mathbb{E}\left[\dfrac{1}{N} \sum\limits_{k=1}^{K} \sum\limits_{j=1}^{n_k} Y_{kj}\right] = \dfrac{1}{N} \sum\limits_{k=1}^{K} \sum\limits_{j=1}^{n_k} \mathbb{E}\left(Y_{kj}\right) = \dfrac{1}{N} \sum\limits_{k=1}^{K} \sum\limits_{j=1}^{n_k} \mu = \dfrac{1}{N} \sum\limits_{k=1}^{K} n_k \mu = \dfrac{1}{N} N \mu = \boxed{\mu}$
- Дисперсия
  $\text{var}(\bar{Y}) = var\left[\dfrac{1}{N} \sum\limits_{k=1}^{K} \sum\limits_{j=1}^{n_k} Y_{kj}\right] = \dfrac{1}{N^2} \sum\limits_{k=1}^{K} \sum\limits_{j=1}^{n_k} var\left(Y_{kj}\right) = \dfrac{1}{N^2} \sum\limits_{k=1}^{K} n_k \sigma^2 = \dfrac{N\sigma^2}{N^2} = \boxed{\dfrac{\sigma^2}{N}}$ **(2)**

Дисперсия зависимой переменной при рандомизированном эксперименте может быть представлена в виде суммы внутригрупповой и межгрупповой дисперсии.

Для этого нам понадобится total variance law (см. доказательство на последней странице):
$var(Y) = var[\mathbb{E}(Y \mid X)] + \mathbb{E}[var(Y \mid X)]$

Пусть Z – номер страты от 1 до K, тогда:

$$var(Y) = \mathbb{E}(var(Y \mid Z)) + var(\mathbb{E}(Y \mid Z)) =$$

$$= \{I(Z = k) \text{ индикаторная переменная, равная 1, если } Z = k, \text{ и равная нулю иначе}\} =$$

$$= \mathbb{E}\left[\sum_{k=1}^{K} \sigma_k^2 I(Z = k)\right] + var\left[\sum_{k=1}^{K} \mu_k I(Z = k)\right] =$$

$$= \sum_{k=1}^{K} \sigma_k^2 \mathbb{E}[I(Z = k)] + \mathbb{E}\left[\sum_{k=1}^{K} \mu_k I(Z = k)\right]^2 - \left[\mathbb{E}\left[\sum_{k=1}^{K} \mu_k I(Z = k)\right]\right]^2 =$$

$$= \sum_{k=1}^{K} \sigma_k^2 p_k + \sum_{k=1}^{K} \mu_k^2 p_k - \mu^2 = \{*\} = \boxed{\sum_{k=1}^{K} \sigma_k^2 p_k + \sum_{k=1}^{K} p_k(\mu_k - \mu)^2} \quad \textbf{(3)}$$

$$(*) : \sum_{k=1}^{K} p_k(\mu_k - \mu)^2 = \sum_{k=1}^{K} p_k \left(\mu_k^2 - 2\mu\mu_k + \mu^2\right) =$$
$$= \sum_{k=1}^{K} p_k \mu_k^2 - 2\mu \sum_{k=1}^{K} \mu_k p_k + \mu^2 \sum_{k=1}^{K} p_k = \sum_{k=1}^{K} \mu_k^2 p_k - 2\mu^2 + \mu^2 = \sum_{k=1}^{K} \mu_k^2 p_k - \mu^2$$

Из **(2)** и **(3)** следует:

$$var(\bar{Y}) = \frac{\sigma^2}{N}$$

$$var(Y) = \sum_{k=1}^{K} \sigma_k^2 p_k + \sum_{k=1}^{K} p_k \left(\mu_k - \mu\right)^2$$

$$var(\bar{Y}) = \underbrace{\frac{1}{N} \sum_{k=1}^{K} p_k \sigma_k^2}_{\text{внутригрупповая дисперсия}} + \underbrace{\frac{1}{N} \sum_{k=1}^{K} p_k \left(\mu_k - \mu\right)^2}_{\text{межгрупповая дисперсия}} \quad \textbf{(4)}$$

Из **(1)** и **(4)** следует:

$$var\left(\hat{Y}_{\text{strat}}\right) = \frac{1}{N} \sum_{k=1}^{K} p_k \sigma_k^2$$

$$var(\bar{Y}) = \underbrace{\frac{1}{N} \sum_{k=1}^{K} p_k \sigma_k^2}_{\text{внутригрупповая дисперсия}} + \underbrace{\frac{1}{N} \sum_{k=1}^{K} p_k \left(\mu_k - \mu\right)^2}_{\text{межгрупповая дисперсия}}$$

Дисперсия среднего значения зависимой переменной всегда больше, чем дисперсия среднего значения зависимой переменной в случае стратификации, поскольку межгрупповая дисперсия всегда неотрицательная. Таким образом, $var(\bar{Y}) \geqslant var\left(\hat{Y}_{\text{strat}}\right)$

Условные обозначения

- Y – зависимая переменная, причем
  - $\mathbb{E}(Y) = \mu$
  - $var(Y) = \sigma^2$
- Выборка разбита на k страт по показателю X, то есть
  - μ_k – среднее в страте
  - σ_k^2 – дисперсия в страте
  - n_k – численность в страты, то есть $\sum_{k=1}^{K} n_k = N$
  - $p_k = \frac{n_k}{N}$ – доля людей из k-й страты в генеральной совокупности
  - Y_{kj} – метрика j-го человека из k-й страты
- Тогда среднее значение зависимой переменной
  - $\bar{Y} = \dfrac{1}{N}\sum\limits_{k=1}^{K}\sum\limits_{j=1}^{n_k} Y_{kj}$ обычное среднее
  - $\hat{Y}_{\text{strat}} = \sum\limits_{k=1}^{K} p_k \bar{Y}_k$ среднее при стратификации
    - $\bar{Y}_k = \dfrac{1}{n_k}\sum\limits_{j=1}^{n_k} Y_{kj}$ среднее значение метрики внутри k-й страты
  - $\underbrace{\sum\limits_{k=1}^{K} p_k \bar{Y}_k}_{\hat{Y}_{\text{strat}}} = \sum\limits_{k=1}^{K} \dfrac{n_k}{N}\dfrac{1}{n_k}\sum\limits_{j=1}^{n_k} Y_{kj} = \underbrace{\dfrac{1}{N}\sum\limits_{k=1}^{K}\sum\limits_{j=1}^{n_k} Y_{kj}}_{\bar{Y}}$ обе средние равны

Total variance law

Доказать:
$var(Y) = var[\mathbb{E}(Y \mid X)] + \mathbb{E}[var(Y \mid X)]$

Доказательство:

$$var(Y) = \mathbb{E}\left(Y^2\right) - (E(Y))^2 =$$
$$= \{\text{закон повтороного мат. ожидания } \mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X))\} =$$
$$= \mathbb{E}\left(\mathbb{E}\left(Y^2 \mid X\right)\right) - [\mathbb{E}(\mathbb{E}(Y \mid X))]^2 =$$
$$= \left\{\text{добавим и вычтем } \mathbb{E}\left[(\mathbb{E}(Y \mid X))^2 \mid X\right]\right\} =$$
$$= \underbrace{\mathbb{E}\left[\mathbb{E}\left(Y^2 \mid X\right) - (\mathbb{E}(Y \mid X))^2 \mid X\right]}_{\mathbb{E}[var(Y|X)]} + \underbrace{\mathbb{E}\left[(\mathbb{E}(Y \mid X))^2 \mid X\right] - \mathbb{E}[\mathbb{E}(Y \mid X)]\mathbb{E}[\mathbb{E}(Y \mid X)]}_{var[\mathbb{E}(Y|X)]} =$$
$$= \mathbb{E}[\text{var}(Y \mid X)] + \text{var}[\mathbb{E}(Y \mid X)]$$