Post-stratification

сделали обычную рандом-ую, а оценку хотим другую

(2) $\hat{y}_{\substack{post\\strat}} = \sum_{k=1}^{K} p_k \bar{y}_k, \quad \bar{y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} y_{kj}$

$n < N_{\text{ген совокупность}}$

Но $n_k \neq p_k \cdot n$ ← размер выборки

размер k-той страты в выборке

доля k-той страты в ген. совокупности

вытакие: $n_1, n_2, \dots, n_k$

total variance law

т.е. при условии что выборка

$$Var(\hat{Y}_{post}) = E\left(V(\hat{y}_{post} \mid n_1, \dots, n_k)\right) + Var\left(E(\hat{Y}_{post} \mid n_1, \dots, n_k)\right) =$$

$$= E\left(\sum_{k=1}^{K} p_k^2 Var(\bar{y}_k \mid n_k)\right) + Var\left(\sum_{k=1}^{K} p_k \cdot \mu_k\right) = E\left(\sum_{k=1}^{K} p_k^2 \cdot \frac{1}{n_k} \sigma_k^2\right) + \underbrace{Var(\mu)}_{=0} =$$

$$= \sum_{k=1}^{K} p_k^2 \sigma_k^2 \boxed{E\left(\frac{1}{n_k}\right)} \quad (*)$$

n_k – сл. вел, т.к. делаем случ. выборку ⇒ кол-во людей из strata k, которые попадут в выборку, случайное

$$E\left(\frac{1}{n_k}\right) = ?$$

мат.ожидание

дисп. $= n \cdot p_k (1 - p_k)$

n_k - сл. величина Бернулли, мат.ож $= n \cdot p_k$

те выборка размера n, p_k - вер-ть успехов

Расклад. в ряд Тейлора вокруг $1/n \cdot p_k$ $= 0$ $E(n_k) = n \cdot p_k$

$$E\left(\frac{1}{n_k}\right) = E\left(\frac{1}{n \cdot p_k} + \frac{-1}{n^2 p_k^2}(n_k - n \cdot p_k) + \frac{1}{n^3 p_k^3}(n_k - n \cdot p_k)^2 + O\left(\frac{1}{n^2}\right)\right) =$$

$$= \frac{1}{n \cdot p_k} + \frac{1}{n^3 p_k^3} \underbrace{E(n_k - \overbrace{n \cdot p_k}^{E(n_k)})^2}_{Var(n_k)} + O\left(\frac{1}{n^2}\right) = \boxed{\frac{1}{n \cdot p_k} + \frac{\cancel{n \cdot p_k}(1 - p_k)}{n^{\cancel{3}\,2} \cdot p_k^{\cancel{3}\,2}} + O\left(\frac{1}{n^2}\right)}$$

подставляем

потому что

$$E\left(\frac{-n_k}{n^2 p_k^2} + \frac{\cancel{n \cdot p_k}}{\cancel{n^2 p_k^2}}\right) = \frac{-1}{n \cdot p_k} + \frac{1}{n p_k} = 0$$

$$\mathrm{Var}(\hat{Y}_{post}) = \sum_{k=1}^{K} p_k^2 \sigma_k^2 \left(\frac{1}{np_k} + \frac{1-p_k}{n^2 p_k^2} + O\left(\frac{1}{n^2}\right) \right) =$$

$$= \underbrace{\frac{1}{n} \sum_{k=1}^{K} p_k \sigma_k^2}_{\mathrm{Var}(\hat{Y}_{pre\text{-}strat})} + \underbrace{\frac{1}{n^2} \sum_{k=1}^{K} (1-p_k)\sigma_k^2 + O\left(\frac{1}{n^2}\right)}_{A}$$

сравнить

$$\mathrm{Var}(\bar{Y}) = \frac{1}{n} \sum_{k=1}^{K} p_k \sigma_k^2 + \underbrace{\frac{1}{n} \sum_{k=1}^{K} p_k (\mu_k - \mu)}_{B}$$

конечное число

т.к. все μ_k, p_k, K, μ — конечные числа, то $\exists n$ достаточно большое чтобы $A \leq B$

$\frac{1}{n^2}$ убыв быстрее, чем $\frac{1}{n}$ при $n \to \infty$

$\Downarrow$

post str $\downarrow$ дисп. по сравнен. с обычн. случаем на $n \to \infty$

$$\Rightarrow \boxed{\mathrm{Var}(\hat{Y}_{pre}) \leq \mathrm{Var}(\hat{Y}_{post}) \leq \mathrm{Var}(\bar{Y})}$$

$\Rightarrow$ вывод: лучше думать заранее о дизайне эксперимента