Vzorčenje po skupinicah

Predpostavljajte, da je populacija velikosti N razdeljena na $K = N/M$ delov velikosti M. Vzorec velikosti km izberemo na naslednji način:

- Najprej izberemo k od K delov z enostavnim slučajnim vzorčenjem **s ponavljanjem**.
- Na drugem koraku izberemo v vsakem od izbranih k delov še enostavni slučajni vzorec velikosti m **s ponavljanjem**.
- Povprečje spremenljivke za celotno populacijo ocenimo tako, da izračunamo povprečje $\bar{y}$ vrednosti spremenljivke za vse enote izbrane v vzorec.

Izračunati želimo standardno napako ocene povprečja pri takem vzorčenju.

Označite z μ_i povprečje v i-tem delu populacije za $i = 1, 2, \ldots, K$. Naj bo

$$\sigma_u^2 = \frac{1}{K} \sum_{i=1}^{K} (\mu_i - \mu)^2 ,$$

kjer je $\mu = \sum_{i=1}^{K} \mu_i/K$. Označite

$$\sigma_w^2 = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{M} (y_{ij} - \mu_i)^2 ,$$

kjer z y_{ij} označimo vrednost spremenljivke za j-to enoto v i-tem delu populacije.

a. Naj bo najprej $k = 1$. Prepričajte se, da lahko zapišemo oceno povprečja v tem primeru kot

$$\bar{y} = \sum_{i=1}^{K} I_i Y_i ,$$

kjer je

$$I_i = \begin{cases} 1 & \text{če smo izbrali del } i. \\ 0 & \text{sicer} \end{cases}$$

in $\mathrm{var}(Y_i) = \sigma_i^2/m$. Poleg tega so I_i neodvisni od Y_i. Vsaj v mislih si lahko predstavljamo, da smo vzorce po posameznih delih populacije izbrali "na zalogo" in potem neodvisno izbrali enega od teh vnaprej izbranih vzorcev. Pri tem je σ_i^2 varianca spremenljivke v i-tem delu populacije. Izračunajte $\mathrm{var}(\bar{y})$.

b. Če neodvisno ponavljamo izbiro dela populacije in nato izbiro vzorca znotraj izbranega dela, dobimo neodvisne ocene $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$, povprečje pa ocenimo z

$$\bar{y} = \frac{1}{k} \sum_{i=1}^{k} \bar{y}_k \,.$$

Pokažite, da je

$$\mathrm{var}(\bar{y}) = \frac{\sigma_u^2}{k} + \frac{\sigma_w^2}{km} \,.$$

Utemeljite, da je to tudi odgovor na vprašanje zastavljeno na začetku. Zgornji izraz je varianca ocene povprečja, če vzorčimo **s ponavljanjem**.

c. Predpostavka, da vzorčimo s ponavljanjem, je nekoliko nerealistična. Naj bo spet $k = 1$ in predpostavimo, da vzorčimo znotraj izbranega dela brez ponavljanja. Izračunajte varianco

$$\bar{y} = \sum_{i=1}^{K} I_i Y_i \,,$$

kjer je

$$I_i = \begin{cases} 1 & \text{če smo izbrali del } i. \\ 0 & \text{sicer} \end{cases}$$

Spet si lahko mislite, da vzorce v posameznih delih izberemo "na zalogo".

d. Predpostavite zdaj, da tudi k kosov izberete brez ponavljanja. Ocena povprečja bo v tem primeru enaka

$$\bar{y} = \frac{1}{k} \sum_{i=1}^{K} I_i Y_i \,,$$

kjer je

$$I_i = \begin{cases} 1 & \text{če smo izbrali del } i. \\ 0 & \text{sicer} \end{cases}$$

Spremenljivke $I_1, \ldots, I_K$ so neodvisne od $Y_1, \ldots, Y_K$. Pri računanju kovarianc $\mathrm{cov}(I_i, I_j)$ uporabite razmislek, ki smo ga uporabili pri izračunu variance hiper-geometrijske porazdelitve. Izračunajte $\mathrm{var}(\bar{y})$.

e. Ali mislite, da je vzorčna porazdelitev približno normalna? Zakaj? Podajte teoretično utemeljitev. Prepričajte se še s kratko simulacijo.