Izračun variance razlike povprečji vzorca in podvzorca

Luka Štrlekar

Imamo vzorec n enot, kjer del vzorca pripada neki skupini (podvzorec n_p enot). Želimo preveriti statistično značilnost razlike povprečji med celotnim vzorcem ($\bar{x}_s$) in podvzorcem ($\bar{x}_p$).

Neuteženi podatki

Zapišimo najprej povprečje celotnega vzorca $\bar{x}_s$ v obliki linearne kombinacije povprečji obeh podvzorcev ($w \in (0,1)$ je delež vsebovanosti prvega podvzorca v celotnem vzorcu):

$$\bar{x}_s = w\bar{x}_{p1} + (1-w)\bar{x}_{p2}$$

Razliko med povprečjem celotnega vzorca in podvzorca lahko izrazimo kot:

$$\begin{aligned} \bar{x}_{p1} - \bar{x}_s &= \bar{x}_{p1} - [w\bar{x}_{p1} + (1-w)\bar{x}_{p2}] \\ &= \bar{x}_{p1} - w\bar{x}_{p1} - \bar{x}_{p2} + w\bar{x}_{p2} \\ &= (1-w)(\bar{x}_{p1} - \bar{x}_{p2}) \end{aligned} \tag{1}$$

$$\begin{aligned} \bar{x}_s - \bar{x}_{p1} &= w\bar{x}_{p1} + (1-w)\bar{x}_{p2} - \bar{x}_{p1} \\ &= w\bar{x}_{p1} + \bar{x}_{p2} - w\bar{x}_{p2} - \bar{x}_{p1} \\ &= (1-w)(\bar{x}_{p2} - \bar{x}_{p1}) \end{aligned} \tag{2}$$

, oziroma kot:

$$\begin{aligned} \bar{x}_{p2} - \bar{x}_s &= \bar{x}_{p2} - [w\bar{x}_{p1} + (1-w)\bar{x}_{p2}] \\ &= \bar{x}_{p2} - w\bar{x}_{p1} - \bar{x}_{p2} + w\bar{x}_{p2} \\ &= w(\bar{x}_{p2} - \bar{x}_{p1}) \end{aligned} \tag{3}$$

$$\begin{aligned} \bar{x}_s - \bar{x}_{p2} &= w\bar{x}_{p1} + (1-w)\bar{x}_{p2} - \bar{x}_{p2} \\ &= w\bar{x}_{p1} + \bar{x}_{p2} - w\bar{x}_{p2} - \bar{x}_{p2} \\ &= w(\bar{x}_{p1} - \bar{x}_{p2}) \end{aligned} \tag{4}$$

Razliko vzorca in podvzorca, ki nista neodvisna zaradi prekrivanja, smo prevedli na razliko dveh neodvisnih podvzorcev, pomnoženih s konstanto prekrivanja (w). Izračun variance razlike je zaradi neodvisnosti sedaj:

$$\begin{aligned} Var(\bar{x}_{p1} - \bar{x}_s) &= Var[(1-w)(\bar{x}_{p1} - \bar{x}_{p2})] \\ &= (1-w)^2[Var(\bar{x}_{p1}) + Var(\bar{x}_{p2})] = Var(\bar{x}_s - \bar{x}_{p1}) \end{aligned} \tag{5}$$

, oziroma:

$$\begin{aligned} Var(\bar{x}_{p2} - \bar{x}_s) &= Var[w(\bar{x}_{p2} - \bar{x}_{p1})] \\ &= w^2[Var(\bar{x}_{p1}) + Var(\bar{x}_{p2})] = Var(\bar{x}_s - \bar{x}_{p2}) \end{aligned} \tag{6}$$

, kjer sta $Var(\bar{x}_{p1})$ in $Var(\bar{x}_{p2})$ vzorčni varianci podvzorcev:

$$Var(\bar{x}_{p1}) = \frac{s_{p1}^2}{n_{p1}} \qquad Var(\bar{x}_{p2}) = \frac{s_{p2}^2}{n_{p2}}$$

Na koncu ocenimo običajno testno statistiko Z, kjer preverjamo ničelno domnevo $H_0 : \bar{x}_s = \bar{x}_p$.

$$Z = \frac{\bar{x}_{p1} - \bar{x}_s}{\sqrt{(1-w)^2(s_{p1}^2/n_{p1} + s_{p2}^2/n_{p2})}} \qquad \text{oziroma} \qquad Z = \frac{\bar{x}_{p2} - \bar{x}_s}{\sqrt{w^2(s_{p1}^2/n_{p1} + s_{p2}^2/n_{p2})}}$$

Testna statistika je porazdeljena približno po standardizirani normalni (z) porazdelitvi (v teoriji se sicer porazdeljuje po približno t-porazdelitvi, vendar pa je izpeljava izračuna stopinj prostosti prezapletena za praktične potrebe, saj imamo opravka z večjimi vzorci ($n > 100$), kjer pa razlika med t ali z porazdelitvijo postane neznatna).

Simulacije

Preverimo teoretične izračune še s simulacijami. Pri vsaki iteraciji simulacije generiramo 1000 enot iz normalne porazdelitve ($N \sim (3, 1^2)$) in jih razdelimo v dva podvzorca skladno z w. Za vsak w izvedemo 10^4 simulacij.

Primer $\bar{x}_{p1} - \bar{x}_s$:

w	Izračunana vrednost	Ocenjena vrednost (simulacije)
0.1	0.00900	0.00891
0.2	0.00400	0.00396
0.3	0.00233	0.00235
0.4	0.00150	0.00149
0.5	0.00100	0.00100
0.6	0.00067	0.00066
0.7	0.00043	0.00043
0.8	0.00025	0.00025
0.9	0.00011	0.00011

Primer $\bar{x}_{p2} - \bar{x}_s$:

w	Izračunana vrednost	Ocenjena vrednost (simulacije)
0.1	0.00900	0.00891
0.2	0.00400	0.00396
0.3	0.00233	0.00235
0.4	0.00150	0.00149
0.5	0.00100	0.00100
0.6	0.00067	0.00066
0.7	0.00043	0.00043

w	Izračunana vrednost	Ocenjena vrednost (simulacije)
0.8	0.00025	0.00025
0.9	0.00011	0.00011

Uteženi podatki

Uteževanje vzorca in podvzorca posebej

Če poststratifikacijsko utežimo posebej celoten vzorec in posebej podvzorec, ki je cilj primerjave, potem ne moremo več primerjave vzorca in podvzorca prevesti na primerjavo dveh neodvisnih podvzorcev.

Zapišimo uteženo povprečje ($\bar{x}_u = \frac{\sum w_i x_i}{\sum w_i}$) celotnega vzorca v obliki linearne kombinacije povprečji obeh podvzorcev ($W \in (0,1)$ je delež vsebovanosti prvega podvzorca v celotnem vzorcu), pri čemer to velja zgolj, če so uteži normirane ($\sum w_i = n$ in $\bar{w} = 1$):

$$\bar{x}_{u_s} = W\bar{x}_{u_{sp1}} + (1-W)\bar{x}_{u_{sp2}}$$

Vendar, če je potem podvzorec utežen posebej, bodo uteži podvzorca drugačne, zato ne velja, da je povprečje podvzorca v celotnem vzorcu enako povprečju podvzorca ($\bar{x}_{u_{sp1}} \neq \bar{x}_{u_{p1}}$ oziroma $\bar{x}_{u_{sp2}} \neq \bar{x}_{u_{p2}}$).

Zapišimo razliko med povprečjem celotnega vzorca in podvzorca 1:

$$\bar{x}_{u_s} - \bar{x}_{u_{p1}} = W\bar{x}_{u_{sp1}} + (1-W)\bar{x}_{u_{sp2}} - \bar{x}_{u_{p1}} \tag{7}$$

Varianca razlike pa je potem:

$$\begin{aligned} Var(\bar{x}_{u_s} - \bar{x}_{u_{p1}}) &= Var[W\bar{x}_{u_{sp1}} + (1-W)\bar{x}_{u_{sp2}} - \bar{x}_{u_{p1}}] \\ &= W^2 Var(\bar{x}_{u_{sp1}}) + (1-W)^2 Var(\bar{x}_{u_{sp2}}) + Var(\bar{x}_{u_{p1}}) \\ &+ 2\ W(1-W)Cov(\bar{x}_{u_{sp1}}, \bar{x}_{u_{sp2}}) - 2\ WCov(\bar{x}_{u_{sp1}}, \bar{x}_{u_{p1}}) - 2\ (1-W)Cov(\bar{x}_{u_{sp2}}, \bar{x}_{u_{p1}}) \\ &= W^2 Var(\bar{x}_{u_{sp1}}) + (1-W)^2 Var(\bar{x}_{u_{sp2}}) + Var(\bar{x}_{u_{p1}}) - 2\ WCov(\bar{x}_{u_{sp1}}, \bar{x}_{u_{p1}}) \end{aligned} \tag{8}$$

Členi kovarianc med $\bar{x}_{u_{sp1}}$ in $\bar{x}_{u_{sp2}}$ ter med $\bar{x}_{u_{sp2}}$ in $\bar{x}_{u_{p1}}$ so 0, ostane kovarianca med $\bar{x}_{u_{sp1}}$ in $\bar{x}_{u_{p1}}$, ki ne more biti 0, saj se gre za isti podvzorec, le da so uteži drugačne.

Zapišimo v splošnem kovarianco med $\bar{x}_{u_{sp}}$ in $\bar{x}_{u_p}$. Ponovno je treba poudariti, da to velja zgolj v primeru normiranih uteži.

$$\begin{aligned} Cov(\bar{x}_{u_{sp}}, \bar{x}_{u_p}) &= Cov(\frac{\sum_{i=1}^{n_p} w_{sp_i} x_i}{\sum_{i=1}^{n_p} w_{sp_i}}, \frac{\sum_{j=1}^{n_p} w_{p_j} x_j}{\sum_{j=1}^{n_p} w_{p_j}}) \\ &= Cov(\frac{\sum_{i=1}^{n_p} w_{sp_i} x_i}{n_p}, \frac{\sum_{j=1}^{n_p} w_{p_j} x_j}{n_p}) \\ &= \frac{1}{n_p^2} \sum_{i=1}^{n_p} \sum_{j=1}^{n_p} Cov(w_{sp_i} x_i, w_{p_j} x_j) \\ &= \frac{1}{n_p^2} \sum_{i=1}^{n_p} Cov(w_{sp_i} x_i, w_{p_i} x_i) \quad \text{ker sta } w_{sp_i} x_i \text{ in } w_{p_j} x_j \text{ neodvisna za } i \neq j \\ &= \frac{1}{n_p} Cov(w_{sp} x, w_p x) \end{aligned} \tag{9}$$

Rezultat je smiseln, saj je v primeru neuteženih podatkov ($w_i = 1$) rezultat enak kot smo že pokazali, saj v splošnem velja $\frac{1}{n_p} Cov(X, X) = \frac{1}{n_p} Var(X)$.

Ker imamo utežene podatke, za oceno vzorčne variance uporabimo recimo cenilko pridobljeno preko Taylorjeve linearizacije:

$$Var(\bar{x}_u) \approx \frac{n}{(n-1)(\sum w_i)^2} \sum w_i^2 (x_i - \bar{x}_u)^2$$

Uteževanje vzorca

Če utežimo zgolj vzorec in pri primerjavi podvzorca obdržimo uteži, smo na istem kot če imamo neutežene podatke, le da je izračun standardne napake drugačen.