Priloga - Statistični izračuni za primerjave povprečij in deležev med dvema vzorcema

Luka Štrlekar

Izračuni za številske (intervalne, razmernostne) spremenljivke

Neuteženi podatki

Želimo primerjati povprečja (μ_1 in μ_2) iz dveh neodvisnih vzorcev (skupin) in ugotoviti ali se na populaciji razlikujeta. Ničelna domneva je $H_0 : \mu_1 = \mu_2$, torej, da sta povprečji enaki. Za preverjanje slednje ničelne domneve uporabimo *t-test* za neenake variance (t.i. *Welchev t-test*), ki med drugim predpostavlja normalno porazdelitev povprečij v obeh skupinah oziroma dovolj velik vzorec (v vsaki skupini), da velja centralni limitni izrek (običajno pravilo čez palec $n \geq 30$).

Ker sta vzorca neodvisna, velja $Var(\mu_1 - \mu_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$. Na vzorcu ocenimo testno statistiko:

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \tag{1}$$

ki je približno porazdeljena po t porazdelitvi z ν stopinjami prostosti (df) izračunanimi po naslednji formuli:

$$\nu \approx \frac{\left(\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2\right)^2}{\left(\hat{\sigma}_1^2/n_1\right)^2 / (n_1 - 1) + \left(\hat{\sigma}_2^2/n_2\right)^2 / (n_2 - 1)}$$

Pogosto prakso, ki je npr. implementirana v SPSS, da se najprej preveri enakost varianc (npr. z Levenovim testom) in nato uporabi ustrezen t-test (z enakimi ali neenakimi variancami), se odsvetuje (Zimmerman, 2010). Welchev t-test se lahko privzeto uporabi vsakič, saj je tudi moč tega testa podobna t-testu z enakimi variancami v primeru enakih populacijskih varianc in uravnoteženih vzorcev (Ruxton, 2006).

Uteženi podatki

Vzorčno uteženo aritmetično sredino lahko v splošnem zapišemo kot (w_i so uteži) (Kalton & Vehovar, 2001, str. 93):

$$\mu_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

Oglejmo si še oceno vzorčne variance. Ker ima utežena aritmetična sredina slučajni spremenljivki v števcu in v imenovalcu, je variabilnost odvisna od variabilnosti slučajnih spremenljivk tako v števcu kot v imenovalcu, kot tudi od njihove korelacije. Ker ne obstaja natančna analitična oblika za izračun te variance, se za približno oceno uporabljajo različne metode, predvsem Taylorjeva linearizacija ali metode replikacij kot je bootstrap/jackknife (Kalton & Vehovar, 2001).

Uporaba Taylorjeve linearizacije, kar imenujemo tudi metoda delta, poda naslednjo cenilko za vzorčno varianco utežene aritmetične sredine, pri čemer predpostavljamo, da so uteži fiksne (izhajajo iz neenakih verjetnosti izbora enote v vzorec) (Gatz & Smith, 1995; Lohr, 2019):

$$var(\mu_w) \approx \frac{n}{(n-1)(\sum w_i)^2} \sum w_i^2 (x_i - \mu_w)^2 \tag{3}$$

Zgornjo formulo lahko zapišemo tudi na spodnji način, kjer je bolj razvidno, da v izračun vstopa tudi kovarianca (Kalton & Vehovar, 2001, str. 95, $u_i = w_i x_i$):

$$var(\mu_w) \approx \frac{var(u)n + \mu_w^2 var(w)n - 2\mu_w cov(u,w)n}{(\sum w_i)^2} \tag{4}$$

V praksi sicer prevladuje prepričanje, da je vpliv obravnavane korelacije majhen, saj običajno ni posebnih razlogov za izrazitejšo povezanost med utežmi in vrednostmi spremenljivk, pa tudi vpliv korelacije ni neposreden, ampak je precej kompleksen. Po drugi strani velja opozoriti, da je korelacija med utežmi in spremenljivko - na nivoju celotnega vzorca - predpogoj za nastanek in tudi za zmanjšanje pristranskosti neutežene ocene (Kalton & Vehovar, 2001, str. 111).

Velja pa poudariti, da smo predpostavljali, da so uteži fiksne in se pri ponavljajočem izbiranju vzorca ne spreminjajo, kar pa pri poststratifikacijskem uteževanju ne velja. Vseeno je ta ocena vzorčne variance varna (tj. konzervativna) (Lu in Gelman, 2003; Lumley, 2011).

Podobno kot pri neuteženih podatkih na vzorcu ocenimo testno statistiko:

$$T_w = \frac{\hat{\mu}_{w1} - \hat{\mu}_{w2}}{\sqrt{var(\hat{\mu}_{w1}) + var(\hat{\mu}_{w2})}} \tag{5}$$

Izračuni za kategorialne (nominalne, ordinalne) spremenljivke

Neuteženi podatki

Imamo nominalno spremenljivko s k kategorijami. Delež je pravzaprav povprečje binomsko porazdeljene slučajne spremenljivke (lahko zavzame vrednosti 0 in 1, $\sim \frac{1}{n}Binom(n,p)$). V našem primeru taka spremenljivka zavzame vrednost 1, če je v k-ti kategoriji in 0, če ni (skupaj n vrednosti). Ne želimo preverjati ničelne domneve, da so hkrati deleži vseh kategorij enaki ($H_0 : p_1 = p_2 = \ ... \ = p_k$), temveč želimo preveriti ničelno domnevo, da je delež v k-ti kategoriji v prvem vzorcu (skupini) enak deležu v isti kategoriji v drugem vzorcu ($H_0 : p_1 = p_2$).

Na vzorcu ocenimo običajno testno statistiko (z-test za neodvisna deleža):

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{6}$$

kjer je

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2},$$

ki je ob izpolnjenih pogojih (dovolj velik vzorec v obeh skupinah - pravilo čez palec $n\hat{p} > 5$ in $n(1-\hat{p}) > 5$) porazdeljena približno po standardni normalni (z) porazdelitvi.

Uteženi podatki

Ker v literaturi ne obstaja formula za skupen (*pooled* - $\hat{p}$) utežen delež, ocenimo standardno napako kar po formuli (3) za vsak delež posebej (kot smo že omenili, je delež povprečje dihotomne spremenljivke), saj v primeru neodvisnih vzorcev velja $Var(p_1 - p_2) = Var(p_1) + Var(p_2)$.

Testna statistika za utežen delež je torej:

$$Z_w = \frac{\hat{p}_{w1} - \hat{p}_{w2}}{\sqrt{var(\hat{p}_{w1}) + var(\hat{p}_{w2})}} \tag{7}$$

Viri

Gatz, D. F., & Smith, L. (1995). The standard error of a weighted mean concentration—I. Bootstrapping vs other methods. *Atmospheric Environment, 29*(11), 1185-1193.

Kalton, G. & Vehovar, V. (2001). *Vzorčenje v anketah.* Ljubljana : Fakulteta za družbene vede

Lohr, S.L. (2019). *Sampling: Design and Analysis (2nd ed.).* Chapman and Hall/CRC. https://doi.org/10.1201/9780429296284

Lu, H., & Gelman, A. (2003). A method for estimating design-based sampling variances for surveys with weighting, poststratification, and raking. *Journal of Official Statistics, 19*(2), 133-151.

Lumley, T. (2011). *Complex surveys: a guide to analysis using R.* John Wiley & Sons.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology, 17*(4), 688-690.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*(1), 173-181.