Priloga - Statistični izračuni

Luka Štrlekar

Izračuni za številske (intervalne) spremnljivke

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 povprečji 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.

Uteženi podatki

Utežujemo s poststratifikacijskimi utežmi, uteži so normalizirane (povprečje uteži je 1 in vsota uteži je enaka velikosti vzorca $\sum_{i=1}^{n} w_i = n$).

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

$$\hat{\mu}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

Oglejmo si še oceno vzorčne variance. Ker ima utežena aritmetična sredina slučajni spremenljivki v števcu in v imenovalcu, moramo uporabiti izraz za varianco razmernostne cenilke. Če je koeficient variacije uteži manjši od 0.2, je približna cenilka variance za uteženo aritmetično sredino enaka (pri čemer je spremenljivka u_i definirana kot $u_i = w_i x_i$ ter je n velikost vzorca) (Kalton & Vehovar, 2001, str. 95):

$$var(\hat{\mu}_w) \approx \frac{var(u)n + \hat{\mu}_w^2 var(w)n - 2\hat{\mu}_w cov(u, w)n}{(\sum_{i=1}^{n} w_i)^2} \tag{3}$$

Ker so uteži normalizirane ($\sum_{i=1}^{n} w_i = n$), se enačba (3) poenostavi v:

$$var(\hat{\mu}_w) \approx \frac{var(u) + \hat{\mu}_w^2 var(w) - 2\hat{\mu}_w cov(u,w)}{n} \tag{4}$$

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 neuteženе ocene (Kalton & Vehovar, 2001, str. 111).

Podobno kot pri neuteženih podatkih ocenimo testno statistiko:

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

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

$$\nu_w \approx \frac{\left(var(\hat{\mu}_{w1}) + var(\hat{\mu}_{w2})\right)^2}{\left(var(\hat{\mu}_{w1})\right)^2 / (n_1 - 1) + \left(var(\hat{\mu}_{w2})\right)^2 / (n_2 - 1)}$$

Izračuni za opisne (nominalne) spremnljivke

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 (skupini) ($H_0 : p_1 = p_2$).

Ker sta vzorca neodvisna, velja $Var(p_1 - p_2) = \frac{p_1(1-p_1)}{n_1 - 1} + \frac{p_2(1-p_2)}{n_2 - 1}$. Uporabili smo nepristransko cenilko za elementarno varianco binomsko porazdeljene spremenljivke, ki je $Var = \frac{n}{n-1} p(1-p)$, iz tega potem izhaja, da je vzorčna varianca $Var(p) = \frac{np(1-p)/n-1}{n} = \frac{p(1-p)}{n-1}$.

Normalna porazdelitev začne precej dobro aproksimirati binomsko, ko velja pravilo čez palec $np > 5$ in $n(1-p) > 5$ (pri bolj ekstremnih deležih je potreben večji vzorec). Na vzorcu ocenimo testno statistiko:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1}}} \tag{5}$$

ki je ob izpolnjenih pogojih (dovolj velik vzorec) porazdeljena približno po standardni normalni (z) porazdelitvi. Ta izraz vrne identično vrednost kot enačba (1).

Formalno gledano ta testna statistika ni porazdeljena kot t porazdelitev (ocenimo le en parameter, ne dva), vendar bomo pri poročanju p vrednosti in IZ uporabili t porazdelitev, ker bomo tako bolj konzervativni in tako upamo, da bo napaka prve vrste res $\leq \alpha$ (predvsem pri manjših vzorcih, ko ima t porazdelitev širše repe kot normalna in bo dala višje p vrednosti in širše IZ).

Uteženi podatki

Postopamo isto kot v poglavju za številske spremenljivke.

Simulacije (izračun vzorčne variance)

V izdelavi...