Uporabna statistika

Gregor Dolinar

Fakulteta za elektrotehniko
Univerza v Ljubljani

8. januar 2013

Analiza variance - ANOVA

Zanima nas vpliv različnih vrednosti nekega parametra na izid poskusa. Izide poskusa pri določeni vrednosti parametra zberemo v isti razred.

razred					vsota	povprečje
1	y_{11}	y_{12}	$\dots$	y_{1n}	$y_{1.}$	$\overline{y}_{1.}$
2	y_{21}	y_{22}	$\dots$	y_{2n}	$y_{2.}$	$\overline{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\dots$	y_{an}	$y_{a.}$	$\overline{y}_{a.}$
					$y_{..}$	$\overline{y}_{..}$

Linearni statistični model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

razredi $i = 1, 2, \ldots, a$,
število meritev v vsakem razredu $j = 1, 2, \ldots, n$,
μ pričakovana vrednost,
τ_i vpliv posameznih razredov (privzamemo $\sum_{i=1}^{a} \tau_i = 0$),
ϵ_{ij} odstopanja ij-tega poskusa od i-tega razreda.

Oziroma

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

$\mu_i = \mu + \tau_i$, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, n$, kjer je μ_i povprečje posameznega razreda.
Privzamemo ϵ_{ij} normalno porazdeljene slučajne spremenljivke, matematično upanje 0, varianca σ^2.

Pozor!

- določeni razredi (naš primer, ne moremo posplošiti na druge razrede)
- naključno izbrani razredi.

Ker so τ_i definirani kot odstopanja od μ, je

$$\sum_{i=1}^{a} \tau_i = 0.$$

Najprej preverimo, ali sprememba vrednosti neodvisne spremenljivke vpliva na rezultat. Če ne vpliva, so vse vrednosti τ_i enake.
Postavimo domnevo.

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$$

$$H_1 : \tau_i \neq 0 \ \text{ za vsaj eno vrednost } \ i$$

Označimo (SS summed squares):

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{..})^2$$

celotna vsota kvadratov odstopanj

$$SS_A = n \sum_{i=1}^{a} (\overline{y}_{i.} - \overline{y}_{..})^2$$

vsota kvadratov med razredi

$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{i.})^2$$

nepojasnjena odstopanja

$$SS_T = SS_A + SS_E$$

$$E(SS_A) = (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2$$

$$E(SS_E) = a(n-1)\sigma^2$$

$$F_0 = \frac{SS_A/(a-1)}{SS_E/(a(n-1))} = \frac{MS_A}{MS_E}$$

F porazdelitev z $a-1$ in $a(n-1)$ prostostnimi stopnjami.
Hipotezo H_0 zavrnemo, če je $f_0 > f_{\alpha,a-1,a(n-1)}$

Fisherjev test LSD (least significant difference)

Vemo, da eden izmed faktorjev vpliva. Kateri?
Za vsak par preverimo hipotezo

$$H_0 : \mu_i = \mu_j$$

s pomočjo t-testa

$$t_0 = \frac{\overline{y}_{i.} - \overline{y}_{j.}}{\sqrt{\frac{2MS_E}{n}}}.$$

Par bomo definirali kot bistveno različen, če je

$$|\overline{y}_{i.} - \overline{y}_{j.}| > \mathrm{LSD},$$

kjer je

$$\mathrm{LSD} = t_{\alpha/2,a(n-1)}\sqrt{\frac{2MS_E}{n}}.$$

Preverjanje modela (normalnost ϵ_{ij})

Statistični model:

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij},$$

razredi $i = 1, 2, \ldots, a$,
bloki $j = 1, 2, \ldots, b$,

razred					vsota	povprečje
1	y_{11}	y_{12}	$\ldots$	y_{1b}	$y_{1.}$	$\overline{y}_{1.}$
2	y_{21}	y_{22}	$\ldots$	y_{2b}	$y_{2.}$	$\overline{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\ldots$	y_{ab}	$y_{a.}$	$\overline{y}_{a.}$
					$y_{..}$	$\overline{y}_{..}$

Imamo b blokov.

Neparametrični testi

Doslej smo pri večini testov predpostavili, da so vzorci naključno izbrani iz populacije, ki ima neko znano slučajno porazdelitev (običajno je bila to normalna porazdelitev).
V veliko primerih je ta predpostavka smiselna, v nekaterih primerih pa populacijo očitno ni normalno porazdeljena in tudi ne vemo, kako bi lahko bila porazdeljena. Kaj storimo v tem primeru?
Pomagamo si z neparamteričnimi testi, pri katerih za populacijo predpostavimo samo, da je zvezno porazdeljena slučajna spremenljivka.

Test predznakov

Test predznakov uporabljamo za preverjanje hipoteze o vrednosti mediane (polovica podatkov ima manjšo vrednost, polovica večjo vrednost).
Za mediano $\tilde{\mu}$ slučajne spremenljivke X velja

$$P[X \leq \tilde{\mu}] = \frac{1}{2}, \quad P[X \geq \tilde{\mu}] = \frac{1}{2}.$$

Če je porazdelitev slučajne spremenljivke simetrična, je mediana enaka povprečni vrednosti (npr. pri normalni porazdelitvi).
Za simetrične porazdelitve lahko torej s testom predznakov preverjamo hipoteze o povprečni vrednosti slučajne spremenljivke.

Naj bo $\tilde{\mu}_0$ izbrana vrednost. Preverjamo domnevo

$$H_0 : \tilde{\mu} = \tilde{\mu}_0$$

$$H_1 : \tilde{\mu} \neq \tilde{\mu}_0$$

Naj bo $X_1, \ldots, X_n$ slučajni vzorec. Oglejmo si razlike

$$X_i - \tilde{\mu}_0, \quad i = 1, \ldots, n.$$

Če je ničelna hipoteza $H_0 : \tilde{\mu} = \tilde{\mu}_0$ pravilna, potem je enaka verjetnost, da je $X_i - \tilde{\mu}_0$ pozitivna ali negativna. V tem primeru je število pozitivnih in negativnih predznakov približno enako.

Testna statistika R^+ je število pozitivnih predznakov. Ničelno hipotezo zavrnemo, če je delež pozitivnih predznakov r^+, ki smo jih izračunali na podlagi opazovanja, značilno različen od $\frac{1}{2}$. P vrednost izračunamo s pomočjo binomske porazdelitve za $p = \frac{1}{2}$. Hipotezo zavrnemo, če je delež pozitivnih predznakov značilno različen od $\frac{1}{2}$, torej značilno manjši ali značilno večji od $\frac{1}{2}$. Če je $r^+ < \frac{n}{2}$, je P vrednost

$$P = 2P[R^+ \leq r^+, p = \frac{1}{2}].$$

Če je $r^+ > \frac{n}{2}$, je P vrednost

$$P = 2P[R^+ \geq r^+, p = \frac{1}{2}].$$

Primer

Testiramo pri $\alpha = 0.05$

$$H_0 : \tilde{\mu} = \tilde{\mu}_0$$

$$H_1 : \tilde{\mu} \neq \tilde{\mu}_0$$

Opravimo 20 meritev in dobimo $r^+ = 15$.
Izračunamo P vrednost

$$P = 2P[R^+ \geq 15, p = \frac{1}{2}] = 2\sum_{r=15}^{20} \binom{20}{r} 0.5^r 0.5^{20-r}$$

$$= 2 \cdot 0.0207 = 0.0414 < 0.05.$$

Imamo tudi tabele za kritične vrednosti ($\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$).
Za $\alpha = 0.05$ je pri $n = 20$ kritična vrednost 5. (zavrnemo, če $\min\{r^+, n - r^+\} \leq 5$).
Za $\alpha = 0.01$ je pri $n = 20$ kritična vrednost 3.

Opomba

Kaj če razlika enaka 0?
Pri zvezni porazdelitvi je verjetnost enaka nič. Praktično: tako vrednost izločimo in delamo z $n - 1$ podatki.

Za $p = \frac{1}{2}$ in $n \geq 10$ je binomska porazdelitev dobro aproksimirana s standardizirano normalno (povprečje $n \cdot p$, varianca $p \cdot (1 - p) \cdot n$):

$$Z_0 = \frac{R^+ - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}}.$$

Ničelno hipotezo zavrnemo, če $|z_0| > z_{\alpha/2}$.

Primer

$n = 20$, $r^+ = 15$, $\alpha = 0.05$, torej zavrnemo, če $|z_0| > z_{0.025} = 1.96$. ($z_0 = 2.24$)

Če imamo enostranski test

$$H_0 : \tilde{\mu} = \tilde{\mu}_0$$

$$H_1 : \tilde{\mu} > \tilde{\mu}_0,$$

je P vrednost

$$P = P[R^+ \geq r^+, p = \frac{1}{2}].$$

Normalna aproksimacija

$$z_0 > z_\alpha.$$

Test predznakov za vzorec parov

Naj bo (X_{1j}, X_{2j}), $j = 1, \ldots, n$, vzorec parov. Definiramo

$$D_j = X_{1j} - X_{2j}, \quad j = 1, \ldots, n.$$

Preverjamo, če imata vzorca parov enako mediano, torej $\tilde{\mu}_1 = \tilde{\mu}_2$. To pomeni, da preverjamo hipotezo, da je $\tilde{\mu}_D = 0$. Torej delamo test predznakov za d_j.

Opomba

S tem testom preverjamo, če imata vzorca parov isto mediano, in ne, če imata dva vzorca isto mediano.

1	2	3	4	5	6	7
0.9	2.1	2.9	4.1	4.9	6.1	6.9
$+$	$-$	$+$	$-$	$+$	$-$	$+$

1	2	3	4	5	6	7
6.9	0.9	2.1	2.9	4.1	4.9	6.1
$-$	$+$	$+$	$+$	$+$	$+$	$+$

Wilcoxonov test

Pogoj, da je porazdelitev zvezna in simetrična.
Če je porazdelitev simetrična, je mediana enaka povprečni vrednosti.
Torej preverjamo domneve o povprečni vrednosti porazdelitve.
Običajni test predznakov upošteva samo predznak razlike opazovanj od mediane, ne pa velikost razlike!

Preverjamo hipotezo

$$H_0 : \tilde{\mu} = \tilde{\mu}_0.$$

Naj bo $X_1, \ldots, X_n$ slučajni vzorec.
Oglejmo si razlike

$$X_i - \tilde{\mu}_0, \quad i = 1, \ldots, n.$$

Absolutne vrednosti razlik uredimo po velikosti $|X_i - \tilde{\mu}_0|$ od najmanjše do največje.
Zaporedni številki dodamo ustrezni predznak.

Označimo z W^+ vsoto pozitivnih števil, ki označujejo zaporedno mesto, z W^- pa vsoto ustreznih negativnih števil.
Definiramo $W = \min\{W^+, W^-\}$.
V preglednici preberemo kritične vrednosti, pri katerih zavrnemo hipotezo pri dani vrednosti α.

Primer

$H_0 : \tilde{\mu} = 4.5$, $H_1 : \tilde{\mu} \neq 4.5$.

X_i	1.1	2.2	3.3	4.4	5.5	6.6	7.7
$X_i - \mu_0$	-3.4	-2.3	-1.2	-0.1	1	1.1	2.2

-0.1	1	1.1	-1.2	2.2	-2.3	-3.4
-1	$+2$	$+3$	-4	$+5$	-6	-7

$W^+ = 10$, $W^- = 18$, $W = 10$

Kritična vrednost za $n = 7$ je 3. Ker je $10 > 3$, ne moremo zavrniti hipoteze, da je 4.5 povprečna vrednost.

Bolj kot sta števili W^+ in W^- blizu, večja je verjetnost, da je μ_0 povprečna vrednost.

Če je $n > 20$, je W^+ normalno porazdeljena slučajna spremenljivka s povprečno vrednostjo

$$\mu_{W^+} = \frac{n(n+1)}{4}$$

in varianco

$$\sigma^2_{W^+} = \frac{n(n+1)(2n+1)}{24}.$$

To lahko izpeljemo iz naslednjih dveh enakosti:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \quad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$