Analiza variance (ANOVA)

Analiza variance je ena najpomembnejših kemometrijskih metod. Pri ANOVI gre za analizo oziroma za primerjavo povprečij večjega števila vzorcev (k). Ničelna hipoteza ANOVE je vedno statistična enakost vseh povprečij. Testiramo predpostavko, da so vsi vzorci vzeti iz iste populacije. ANOVA je uporabna predvsem pri vrednotenju laboratorijev za delo z neko predpisano metodo, za primerjanje rezultatov iste metode, ki jih dajejo različni laboratoriji ali en laboratorij pod različnimi pogoji (posledično je torej uporabna za odkrivanje slabosti v laboratorijih), za testiranje ustreznosti kalibracije analitskih inštrumentov (glej poglavje o kalibracijski premici!), za testiranje modelov dobljenih z MLR (ali drugimi modelnimi metodami) s katerimi napovedujemo lastnosti kompleksnih vzorcev in podobno.

ANOVA je osnovana na dveh predpostavkah:

a) vse meritve x_{ij} so normalno porazdeljene in
b) variance v vzorcih so homogene (Bartlettov test)

K sreči je ANOVA zelo robustna metoda, kar pomeni, da daje uporabne rezultate tudi, če podatki malo odstopajo od obeh gornjih predpostavk.

Pri ANOVI vedno najprej sestavimo več skupini vzorcev, ki smo jih dobili z meritvami pri obdelavi ali ovrednotenju nekega analitskega procesa. Ničelna hipoteza H_0 pri ANOVA testu je, da so povprečja v skupinah vzorcev statistično enaka. Z drugimi besedami, da se ob upoštevanem tveganju α, ne razlikujejo med seboj. Ničelno hipotezo pri ANOVI preverjamo z F-testom, to je s primerjavo variance povprečij med skupinami vzorcev, ki smo jih izbrali, in variance povprečij znotraj vzorcev.

$$F = \frac{\sigma^2_{med\ skupinami}}{\sigma^2_{znotraj\ skupin}} < F^{tabele}_{\alpha, ps_1 ps_2} \rightarrow H_0\ sprejeta$$

Obe oznaki ps_1 in ps_2 v zgornji enačbi pomenita število prostostnih stopenj v števcu (ps_1) in v imenovalcu (ps_2)

Večkratni t-test

Če želimo primerjati več povprečnih vrednosti bi lahko za to uporabili večkraten t-test. Vendar pri taki primerjavi uporabimo isto povprečno vrednost večkrat, zato posamezni t-testi niso več med seboj neodvisni. Kot posledica se verjetnost da bo vsaj en test spodletel pri primerjavi povprečnih vrenosti iz iste polulacije poveča. En način da se znebimo tega problema je da prilagodimo tveganje (Bonferronijeva prilagoditev) posameznega testa (α') tako, da bo skupno tveganje (a) take določitve doseglo predpisano vrednost.

$$\alpha' = 1-(1-\alpha)^{\frac{1}{k}}$$

Primer:

Primerjamo 5 povprečij. Za tako primerjavo moramo narediti 10 t-testov. Če želimo, da bo skupno tveganje napovedi α=0.05, potem mora biti izbrano tveganje posameznega testa a':

$$\alpha' = 1-(1-\alpha)^{\frac{1}{k}} = 1-(1-0.05)^{\frac{1}{10}} = 0.005$$

Eno-faktorska ANOVA (one-way ANOVA), medlaboratorijski test

Najenostavnejši primer ANOVE je primerjava povprečij k vzorcev. Vsak j-ti vzorec ima n_j meritev. Testiramo ali so vsa povprečja iz iste populacije. Če gre pri tem za medlaboratorijsko primerjavo k laboratorijev pri kateri je bilo v vsakem laboratoriju narejenih n_j meritev, preverjamo ali dobijo vsi laboratoriji statistično enake rezultate (seveda ob naprej predpisani stopnji tveganja α).

Oznaka x_{ij} pomeni i-to meritev v j-tem laboratoriju, $\overline{x}_j$ je povprečje meritev j-tega laboratorija. Skupno število meritev v vseh laboratorijih $(n_1 + n_2 + \ldots + n_j + \ldots + n_k)$ je enako N. Z $\overline{\overline{x}}$ pa označimo celotno povprečje vseh N meritev.

Izvor variance	Vsota kvadratov napak (SS)	SS/(pr.s t)	F-test
Med k laboratoriji	$SS_{med} = \sum_{j=1}^{k} n_j(\overline{x}_j - \overline{\overline{x}})^2 = \sum_{j=1}^{k} n_j \overline{x}_j - N\overline{\overline{x}}^2$	$SS_{med} /(k-1)$	
Znotraj laboratorija	$SS_{znotraj} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_j)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{k} n_j \overline{x}_j^2$	$SS_{znotraj} /(N=k)$	$F = \frac{SS_{med} /(k-1)}{SS_{znotraj} /(N-k)} \rightarrow F^{tabele}_{\alpha,k-1,N-k}$
Skupna varianca	$SS_{Tot} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \overline{\overline{x}})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - N\overline{\overline{x}}^2$	$SS_{Tot} /(N=1)$	

Ker je $SS_{Tot} = SS_{med\ lab.} + SS_{znotraj\ lab.}$ ni treba računati vseh treh vrednosti, ampak moramo le dve. Najlažje je izračunati SS_{tot}, in $SS_{med\ lab.}$, zato navadno izračunamo $SS_{znotraj\ lab.}$ posredno, iz razlike med prvima dvema:

$$SS_{znotraj\ lab.} = SS_{Tot} - SS_{med\ lab.}$$

V večini primerov je $SS_{znotraj\ lab.}$, ki mu pravimo tudi ostanek ("residual", SS_{res}), manjši kot $SS_{med\ lab.}$ Zato bo količnik F večji od ena (medlaboratorijske razlike so večje kot znotraj laboratorijske). Če temu ni tako, potem razlike med laboratoriji z gotovostjo niso signifikantne. Take postanejo šele, ko iz podatkov izračunani F preseže $F^{tabelaričen}$.

Testiranje enakosti varianc

Ker ANOVA zahteva enakost varianc moramo najprej preveriti ali je ta predpostavka izpoljnjena. Na voljo imamo več testov. Pogledali si bomo tri najpogostejše.

Hartleyev test

Pri njem primerjamo največjo in najmanjšo varianco posameznih serij. Dobljen faktor primerjamo z kritičnimi F vrednostmi.

$$F_{izr} = \frac{s_{\max}^2}{s_{\min}^2}$$

Cochranov test

Pri tem testu primerjamo največjo varianco z vsoto vseh varian. Uporabimo naslednjo zvezo

$$C_{izr} = \frac{s_{\max}^2}{\sum_{j=1}^{k} s_j^2}$$

Osnovna predpostavka pri Cochranovem testu je, da je število elementov v posamezni seriji enako. Pir anjših odstopanjih od te zahteve test še vedno lahko izvajamo. Namesto varianc lahko test izvedemo tudi s kvadrati območij. Kritične vrednosti dobimo iz Cochranovih C tabel. Kritične vrednosti so pri izbrani stopji zaupanja odvisne od števila serij k in števila elementov v posamezni seriji.

Bartlettov test varianc večjega števila vzorcev

Testiranje varianc, če so statisatično enake ali ne, je v bistvu test porazdelitev, iz katerih smo variance izračunali. Ker porazdelitve primerjamo z uporabo χ^2 testa, je logično, da χ^2 tabele uporablja tudi Bartlettov test.

Najprej moramo izračunati skupno varianco (pooled variance) vseh *k* vzorcev, katerih povprečja prizkušamo. Če ima vsak od *k* vzorcev n_j meritev in že določeno ali izračunano varianco s_j^2 , potem je skupna *varianca* enaka:

$$s_{pooled}^2 = \frac{\sum_{j=1}^{k}(n_j - 1)s_j^2}{\sum_{j=1}^{k}(n_j - 1)}$$

Pri Bartlettovem testu moramo vse izračunane variance logaritmirati z naravnim logaritmom in potem izračunati izraz *M*. Število prostostnih stopenj pri testiranju je število vzorcev *k*, oziroma število varianc, ki jih primerjamo, zmanjšano za eno (*k-1*):

$$M = [\sum_{j=1}^{k}(n_j - 1)]\, ln(s_{pooled}^2) - \sum_{j=1}^{k}(n_j - 1) ln(s_j^2) \;<\; \chi^2(\alpha, k-1) \rightarrow H_0 \;sprejeta$$

Podobno kot moramo pred primerjavo dveh povprečij s *t*-testom, preveriti ali imata vzorca primerljivi varianci (to naredimo z *F*-testom; $F = s_1^2/s_2^2$), moramo tudi preden pričnemo primerjati več povprečij z ANOVO, narediti Bartlettov test. Ta nam zagotov, bodisi da so variance, ki jih obravnavamo homogene, bodisi da to niso.

Testiranje ustreznosti modelov z ANOVA testom

V analizni kemiji pogosto pridemo do problema modeliranja, t.j., do iskanja takega matematičnega modela, ki bo dani niz meritev opisal najbolje. Z modelnom, ki je lahko polinomska ali kakšna drugačna analitična funkcija, matrično-vektorski izraz, ali nevronska mreža, želimo napovedati lastnosti neznanih vzorcev. Najpreprostejša uporaba modelov je kalibracijska premica. Modelna funkcija je premica z dvema parametroma: *f(x)= ax + b*. Predno model (v našem primeru premico, lahko pa je model tudi bolj kompliciran) začnemo uporabljati, moramo z ANOVO preveriti ali je, glede na izračunano ali izmerjeno natančnost meritev, ustrezen ali ne.

V skladu z idejo ANOVA testa preverjamo ali sta dve skupini povprečij enaki ali ne. Prva skupina so variance meritev v posameznih točkah (ponovitve ali replikacije meritev) na katerih smo model naredili (razpršenost merskih povprečij), druga skupina pa predstavlja razpršenost ali odstopanje točk s katerimi smo model izdelali, od dejanskih napovedi modela. Z drugimi bedeami: primerjamo variance posamenznih meritev glede na povprečje ponovitev z varianco, ki jo dajo povprečne vrednosti ponovitev glede na napovedi modela,

$$F = \frac{\sigma_{mod\,ela}^2}{\sigma_{meritve}^2} < F(\alpha, ps_1, ps_2) \;\rightarrow\; mod\,el\;ustrezen$$

V *F*- testu sta ps_1 in ps_2 števili prostostnih stopenj modela ($ps_1 = k-p$) in meritev povprečij ($ps_2 = N-k$). Modelna funkcija $f(x_j)$ ima *p* parametrov (v primeru premice sta to *a* in *b*, *p = 2*) in je narejena na osnovi *k* merskih točk x_j, $j = 1 \ldots k$. V primeu kalibracijske premice je x_j *j* -ta koncentracija. Pri vsaki točki x_j je bilo izmerjeno n_j ponovitev y_{ij}. ($i = 1 \ldots n_j$). Vrednost modelne funkcije $f(x_j)$ v točki x_j velikokrat označimo s strešico $\hat{y}_j$.

$$\left.\begin{aligned} \sigma_{mod\,ela}^2 &= \frac{SS_{mod\,ela}}{k-p} = \frac{\sum_{j=1}^{k} n_j[\bar{y}_j - f(x_j)]^2}{k-p} \\ \sigma_{meritev}^2 &= \frac{SS_{meritev}}{N-k} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{N-k} \end{aligned}\right\} \quad F = \frac{\frac{SS_{mod\,ela}}{k-p}}{\frac{SS_{meritev}}{N-k}} = \frac{\frac{\sum_{j=1}^{k} n_j(\bar{y}_j - \hat{y}_j)^2}{k-p}}{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{N-k}} < F_{\alpha,k-p,N-k}^{tabele} \rightarrow$$

Model je ustrezen

Ovrednotenje modela premice (ANOVA)

Na osnovi 20 meritev ($N = 20$), t.j., po 4 ponovitve ($n_1 + n_2 + n_3 + n_4 = 4$) v vsaki od petih ($k = 5$) točk, smo določili model kalibraciske premice z dvema parametroma ($p = 2$): $a = 4.050$ in $b = 9.327$.

			1	2	3	4	5
		x_j	1.00	2.00	3.00	5.00	10.00
Ponovitve meritev y_{ij}		i	10.60	24.80	31.00	52.30	102.40
			8.70	22.20	32.30	51.40	85.90
			12.80	23.80	35.20	59.30	95.20
			9.50	21.80	32.30	54.60	98.40
Povprečje replikacij	A	$\bar{y}_j$	10.40	23.15	32.70	54.40	95.48
Modelne vrednosti	B	$\hat{y}_j = a + bx_j$	13.38	22.71	32.03	50.69	97.32
Račun za SS_{pe}	C	$(\bar{y}_j - \hat{y}_j)^2$	35.47	0.79	1.78	55.15	13.66
Račun za SS_{lof}	D	$\sum_{i=1}^{nj} y_{ij}^2$	442.14	2149.56	4286.62	11874.90	36610.17
	E	$\frac{1}{n_j}(\sum_{i=1}^{nj} y_{ij})^2$	432.64	2143.69	4277.16	11837.44	36461.90
	D-E	$\sum_{i=1}^{n_j} y_{ij}^2 - \frac{1}{n_j}(\sum_{i=1}^{n_j} y_{ij})^2$	9.50	5.87	9.46	37.46	148.27

SSpe = vsota vrstice C = 106.86
SSlof = vsota vrstice D-E = 210.56

Število prostostnih stopenj:
N – k = 15 pri napaki (PE – pure error)
K – p = 3 pri modelu (LOF – lack of fit)

$$F = \frac{\frac{SS_{LOF}}{k-p}}{\frac{SS_{PE}}{N-k}} = \frac{\frac{\sum_{j=1}^{n_j} n_j(\bar{y}_j - \hat{y}_j)^2}{k-p}}{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_n}(y_{ij} - \bar{y}_j)^2}{N-k}} = \frac{\frac{210.56}{3}}{\frac{106.86}{15}} = \mathbf{2.537}$$

F tabelarična = F(0.05, 3, 15) = 3.287