Analiza variance

Enakost povprečij $r \geq 2$ normalno porazdeljenih populacij s podobnimi variancami lahko preverimo s testom, ki se imenuje analiza variance. Pri tem testu ničelna hipoteza trdi, da so povprečja vseh populacij med seboj enaka, alternativna pa, da se vsaj eno povprečje razlikuje od drugih:

$$\begin{aligned} &\mathrm{H}_0 : m_1 = m_2 = \cdots = m_r \,, \\ &\mathrm{H}_1 : m_i \neq m_j \,, \quad \text{za vsaj en par } (i,j)\,. \end{aligned} \tag{1}$$

Za preverjanje ničelne hipoteze potrebujemo iz vsake od r populacij en vzorec z n_i elementi, skupaj torej r vzorcev z $n = \sum_{i=1}^{r} n_i$ elementi.

Test temelji na primerjavi odstopanja izmerjene spremenljivke med vzorci z odstopanjem spremenljivke znotraj vzorcev. Odstopanje med vzorci izrazimo s povprečnim kvadratičnim odstopanjem med vzorci S_1^2, odstopanje znotraj vzorcev pa s povprečnim kvadratičnim odstopanjem znotraj vzorcev S_2^2. V kolikor so odstopanja med vzorci značilno večja kot odstopanja znotraj vzorcev, H_0 zavrnemo. Uporabimo testno statistiko:

$$F = \frac{S_1^2}{S_2^2}\,. \tag{2}$$

Testna statistika F je porazdeljena s porazdelitvijo F, ki ima $(r-1, n-r)$ prostostnih stopenj. Interval zavračanja je $(f_{r-1,n-r;\alpha}, \infty)$. Količine, potrebne za izračun testne statistike, vpisujemo v tabelo:

Odstopanje	Vsota kvadratičnih odstopanj	Število prostostnih stopenj	Povprečje kvadratičnih odstopanj	Statistika
Med skupinami	q_1	$r-1$	$S_1^2 = \dfrac{q_1}{r-1}$	$F = \dfrac{S_1^2}{S_2^2}$
Znotraj skupin	q_2	$n-r$	$S_2^2 = \dfrac{q_2}{n-r}$	
Celotno	q	$n-1$		

Pri izračunu si pomagamo z naslednjimi formulami:

$$q = \left(\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}^2\right) - nm^2\,, \qquad m = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}\,, \tag{3}$$

$$q_1 = \left(\sum_{i=1}^{r} n_i m_i^2\right) - nm^2\,, \qquad m_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}\,, \tag{4}$$

$$q_2 = q - q_1\,, \qquad n = \sum_{i=1}^{r} n_i\,. \tag{5}$$

Linearna regresija

S *korelacijskim koeficientom* r opišemo primernost uporabe linearne regresije za ocenjevanje medsebojne odvisnosti naključnih spremenljivk X in Y:

$$r = \frac{\mathrm{Cov}\left[X, Y\right]}{\sqrt{\mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]}}\,. \tag{6}$$

Zaloga vrednosti za r je $[-1, 1]$. Vrednosti $|r| \approx 1$ kažejo na izrazito linearno odvisnost med X in Y, medtem ko vrednosti $|r| < 0.5$ kažejo, da je regresijska premica neprimerna za upodobitev povezanosti X in Y, saj sta spremenljivki bodisi neodvisni ali pa je njuna odvisnost nelinearna. Linearno regresijo je smiselno uporabljati, če je $r \geq 0.75$. Za vzorec meritev $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ izračunamo korelacijski koeficient po enačbi:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}\,. \tag{7}$$

Pri linearni regresiji vzamemo za matematični model odvisnosti linearno funkcijo:

$$Y = aX + b\,, \tag{8}$$

cilj pa je določiti koeficienta a in b tako, da bo vsota kvadratov napak najmanjša. Rešitev je:

$$a = \frac{\mathrm{Cov}\left[X, Y\right]}{\mathrm{Var}\left[X\right]}\,, \qquad\qquad b = \mathrm{E}\left[Y\right] - a\mathrm{E}\left[X\right]\,. \tag{9}$$

Cenilki $\hat{a}$ in $\hat{b}$ za koeficienta a in b določimo na podlagi vzorca meritev:

$$\hat{a} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\,, \qquad\qquad \hat{b} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i\,. \tag{10}$$

Linearni model lahko uporabimo tudi, kadar zveza med spremenljivkama ni linearna. V takih primerih si lahko pomagamo z logaritmiranjem zveze, v kolikor jo s tem linearizirамo, ali pa uvedemo nove spremenljivke:

$$\begin{aligned} Y &= b\,\mathrm{e}^{aX} \quad &\rightarrow \quad &\ln Y = \ln b + aX\,, \\ Y &= aX^2 + b \quad &\rightarrow \quad &Y = aZ + b\,, \quad \text{kjer je } X^2 = Z\,. \end{aligned} \tag{11}$$