Porównanie średnich w wielu populacjach o rozkładzie normalnym – Analiza wariancji (ANOVA)

Założenia:

$X_i \sim N(m,\sigma^2)$ $\sigma_1 = \sigma_2 = \sigma_3 = ... = \sigma_i$

model analizy wariancji:

$$y_{ij} = m + a_i + e_{ij}$$

gdzie:

y_{ij} – wielkość cechy
m – średnia ogólna (może być oznaczana również literą μ)
a_i – efekt i-tego poziomu czynnika
e_{ij} – błędy losowe, o rozkładzie $N(0, \sigma_e)$

Hipoteza zerowa H_0: $m_1 = m_2 = m_3 = ... = m_i$ (średnie nie różnią się)
Hipoteza alternatywna H_1: $m_i \neq m_i'$ (co najmniej dwie średnie różnią się)
Tabela analizy wariancji:

Źródło zmienności	Stopnie swobody	Sumy kwadratów	Średnie kwadraty	F_{emp}
Czynnik	$k-1$	$\mathrm{var}A$	$S_a^2 = \frac{\mathrm{var}A}{k-1}$	S_a^2/S_e^2
Błąd losowy	$N-k$	$\mathrm{var}E$	$S_e^2 = \frac{varE}{N-k}$	
Ogółem	$N-1$	$\mathrm{var}T$		

$$\mathrm{var}A = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \mathrm{var}E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2,$$

$$\mathrm{var}T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2,$$

$$\mathrm{var}A + \mathrm{var}E = \mathrm{var}T$$

Funkcja testowa $\boldsymbol{F_{emp}}$

Wartość krytyczna $\boldsymbol{F_{\alpha,k-1,n-k}}$

α – poziom istotności (najczęściej przyjmujemy 0,05)

k – liczba poziomów czynnika

n – liczebność prób

Jeżeli $\boldsymbol{F_{emp} > F_{\alpha,k-1,n-k}}$ to $\mathbf{H_0}$ odrzucamy i przyjmujemy $\mathbf{H_1}$

Porównania wielokrotne (szczegółowe)

Grupy jednorodne — podzbiory średnich, które można uznać za takie same (nie różnią się istotnie)

Procedury porównań wielokrotnych — postępowanie statystyczne zmierzające do podzielenia zbioru średnich na grupy jednorodne

Procedury: Tukeya, Scheff´ego, Bonfferroniego, Duncana, Newmana–Kuelsa i inne.

NIR — najmniejsza istotna różnica

$$\text{Jeżeli } |\bar{X}_i - \bar{X}_j| < NIR, \text{ to uznajemy, że } \mu_i = \mu_j.$$

Jeżeli $\left|\overline{X}_i - \overline{X}_j\right| \geq NIR$ to uznajemy, że średnie różnią się (różnica istotna statystycznie)

Procedura Tukeya

$$NIR = t_{\alpha,k,n-k} \cdot S_e \sqrt{\frac{1}{n}}$$

$t_{\alpha,k,n-k}$ – wartość krytyczna studentyzowanego rozstępu
S_e-błąd standardowy różnicy średnich

WSPÓŁCZYNNIK KORELACJI

Współczynnik korelacji liniowej Pearsona (oznaczany najczęściej symbolem - *r*) określa poziom zależności liniowej między zmiennymi losowymi.

$$r = \frac{\mathrm{cov}(X, Y)}{s_x \cdot s_y}$$

gdzie, wartość kowariancji (cov) na podstawie próby liczymy wg następującego wzoru:

$$\mathrm{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

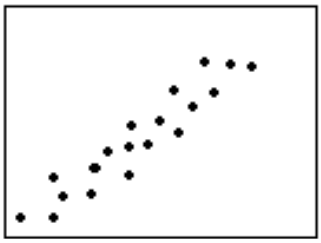
natomiast s_x i s_y są odchyleniami standardowymi dla zmiennej *X* i *Y*

Współczynnik korelacji liniowej dwóch zmiennych jest, zatem ilorazem kowariancji i iloczynu odchyleń standardowych.

Współczynnik korelacji liniowej przyjmuje zawsze wartości w zakresie [-1,1].

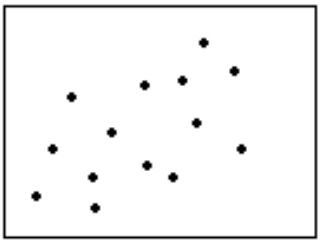
Im większa wartość bezwzględna współczynnika, tym większa jest zależność liniowa między zmiennymi. $r_{xy} = 0$ oznacza brak korelacji, $r_{xy} = 1$ oznacza silną korelację dodatnią, jeżeli jedna zmienna (*X*) rośnie to również rośnie druga zmienna (*Y*), natomiast $r_{xy} = -1$ oznacza korelację ujemną (jeżeli zmienna *X* rośnie, to *Y* maleje i na odwrót).

Stopień korelacji

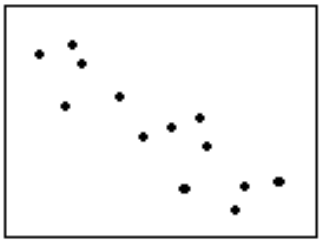


silna dodatnia (r = 0,8)

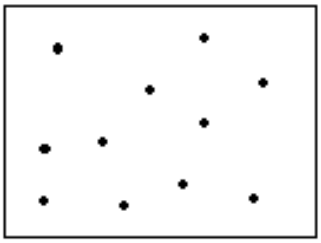
silna ujemna (r = -0,8)



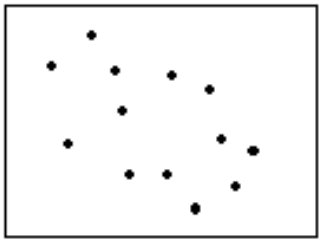
słaba dodatnia (r = 0,3)



umiarkowana ujemna (r = -0,5)



brak korelacji (r = 0,0)



słaba ujemna (r = -0,3)

Testowanie istotności korelacji

Hipoteza zerowa: $H_0:\rho=0$

ρ- wartość współczynnika korelacji dla całej populacji

Jeżeli $|r_{emp}|>r_{\alpha,2,n-2}$ to H_0 odrzucamy.

$r_{\alpha,2,n-2}$ – jest wartością krytyczną współczynnika korelacji prostej Pearsona