

STATYSTYKA

Korelacja



Pojęcie korelacji

Korelacja (współzależność cech) określa wzajemne powiązania pomiędzy wybranymi zmiennymi.

Charakteryzując korelację dwóch cech podajemy dwa czynniki: kierunek oraz siłę.

Rodzaje korelacji

Ze względu na sposób analizy oraz charakter analizowanych zmiennych wyróżniamy:

- **korelację prostą** – badającą związek zachodzący pomiędzy dwoma cechami lub zjawiskami (r_{xy} , r_{12}),
- **korelację cząstkową** – informującą o związku dwóch cech z wyłączeniem trzeciej zmiennej ($r_{xy.z}$, $r_{12.H}$),
- **korelację wieloraką** – informującą o związku jednej cechy z kilkoma ujętymi łącznie ($r_{x.yz}$, $r_{1.2H}$).

Interpretacja wyników korelacji

Wyrazem liczbowym korelacji jest **współczynnik korelacji** (r lub R), zawierający się w przedziale [-1; 1].

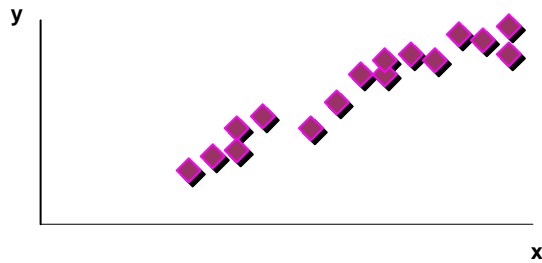
- **korelacja dodatnia** (wartość współczynnika korelacji **od 0 do 1**) – informuje, że wzrostowi wartości jednej cechy towarzyszy wzrost średnich wartości drugiej cechy,
- **korelacja ujemna** (wartość współczynnika korelacji **od -1 do 0**) - informuje, że wzrostowi wartości jednej cechy towarzyszy spadek średnich wartości drugiej cechy.

Siła związków korelacyjnych

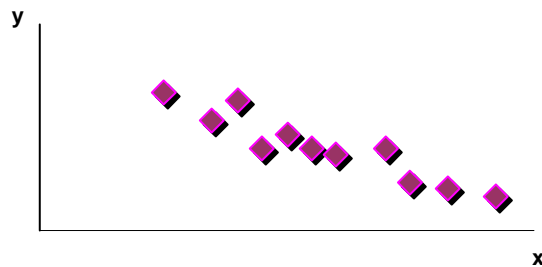
- poniżej 0,2** - **korelacja słaba** (praktycznie brak związku)
- 0,2 – 0,4** - **korelacja niska** (zależność wyraźna)
- 0,4 – 0,6** - **korelacja umiarkowana** (zależność istotna)
- 0,6 – 0,8** - **korelacja wysoka** (zależność znaczna)
- 0,8 – 0,9** - **korelacja bardzo wysoka** (zależność bardzo duża)
- 0,9 – 1,0** - **zależność praktycznie pełna**

NAJWAŻNIEJSZA JEST ISTOTNOŚĆ KORELACJI

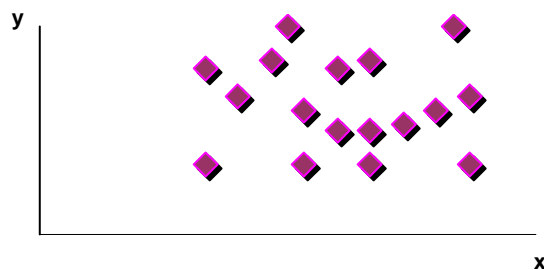
Korelacyjne wykresy rozrzutu



zależność liniowa dodatnia
($r > 0$)



zależność liniowa ujemna ($r < 0$)



brak zależności ($r = 0$)

Współczynnik korelacji Pearsona

Współczynnik ten wykorzystywany jest do badania związków prostoliniowych badanych zmiennych, w których zwiększenie wartości jednej z cech powoduje proporcjonalne zmiany średnich wartości drugiej cechy (wzrost lub spadek).

Współczynnik ten obliczamy na podstawie wzoru:

$$r_{xy} = \frac{\text{cov}(x, y)}{Sd_x \cdot Sd_y} \quad \text{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Przykład Nr 1

Badaniu poddano długość kończyny dolnej (x_i) oraz moc (y_i) u siedmiu uczniów IV klasy szkoły podstawowej. Na podstawie poniższych danych oszacować współzależność obu analizowanych cech.

x_i	83,1	88,2	87,3	90,4	80,6	87,1	85,3
y_i	41	45	42	52	52	46	47

W pierwszej kolejności należy wyliczyć:

\bar{X}_{x_i} ; \bar{y}_{y_i} ; Sd_x ; Sd_y oraz wartość kowariancji

tabela pomocnicza

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
83,1	41	-2,9	-5,4	15,66	8,41	29,16
88,2	45	2,2	-1,4	-3,08	4,84	1,96
87,3	42	1,3	-4,4	-5,72	1,69	19,36
90,4	52	4,4	5,6	24,64	19,36	31,36
80,6	52	-5,4	5,6	-30,24	29,16	31,36
87,1	46	1,1	-0,4	-0,44	1,21	0,16
85,3	47	-0,7	0,6	-0,42	0,49	0,36
$\Sigma=602,0$	$\Sigma=325,0$			$\Sigma=-1,24$	$\Sigma=65,16$	$\Sigma=113,72$

$$\bar{x} = \frac{602}{7} = 86; \bar{y} = \frac{325}{7} = 46,4$$

$$Sd_x = \sqrt{\frac{\sum (x_i - \bar{x}_x)^2}{n}} = \sqrt{\frac{65,16}{7}} = 3,05$$

$$Sd_y = \sqrt{\frac{\sum (y_i - \bar{x}_y)^2}{n}} = \sqrt{\frac{113,72}{7}} = 4,03$$

$$\text{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{-1,24}{7} = -0,18$$

$$r_{xy} = \frac{\text{cov}(x, y)}{Sd_x \cdot Sd_y} = \frac{-0,18}{3,05 \cdot 4,03} = -0,015$$

Interpretacja: wyliczony współczynnik korelacji wskazuje na **brak związku** długości kończyny z mocą objętych badaniem uczniów.

Wskaźnik determinacji liniowej

Na podstawie wyliczonego współczynnika korelacji obliczyć można tzw. **wskaźnik determinacji liniowej**, informujący o procencie wyjaśnionej liniowo zmienności zmiennej zależnej przez zmienną niezależną. Wskaźnik ten oblicza się na podstawie wzoru:

$$WD = r_{xy}^2 \cdot 100\%$$

Zadanie Nr 1

Sportowców poddano badaniom szybkości reakcji na bodziec wzrokowy (y_i) oraz badaniom wzroku (x_i). Oszacować współzależność obu analizowanych cech.

x_i	3,5	3,4	2,1	5,4	1,1	5,1	6,9	4,0	4,5	2,5
y_i	1,6	2,9	1,5	3,5	0,6	2,5	7,1	3,5	2,1	2,6

Tabela pomocnicza do zad. 1

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
3,5	1,6	-0,35	-1,19	0,42	0,12	1,42
3,4	2,9	-0,45	0,11	-0,05	0,20	0,01
2,1	1,5	-1,75	-1,29	2,26	3,06	1,66
5,4	3,5	1,55	0,71	1,10	2,40	0,50
1,1	0,6	-2,75	-2,19	6,02	7,56	4,80
5,1	2,5	1,25	-0,29	-0,36	1,56	0,08
6,9	7,1	3,05	4,31	13,15	9,30	18,58
4	3,5	0,15	0,71	0,11	0,02	0,50
4,5	2,1	0,65	-0,69	-0,45	0,42	0,48
2,5	2,6	-1,35	-0,19	0,26	1,82	0,04
38,5	27,9			22,445	26,485	28,069

Interpretacja

Występuje bardzo wysoka dodatnia korelacja pomiędzy analizowanymi cechami. Wzrostowi jakości wzroku towarzyszy wzrost szybkości reakcji.

$$r=0,8232 \in < 0,7;0,9)$$

$$WD=67,77\%$$

Współczynnik R Spearmana

- Współczynnik korelacji rang Spearmana wykorzystywany jest do opisu siły korelacji dwóch cech, w przypadku gdy:
 - **cechy mają charakter jakościowy,** pozwalający na uporządkowanie ze względu na siłę tej cechy,
 - **cechy mają charakter ilościowy, ale ich liczebność jest niewielka.**



kobieta



mężczyzna

Przykład pomiaru nominalnego: **płeć**



niskie



średnie



wysokie

Przykład pomiaru porządkowego: **stopień zaangażowania w trening sportowy**



0

25 m

50 m

75 m

100 m

125 m

Przykład pomiaru ilorazowego: **odległość w skokach narciarskich [m]**

$$r_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

d_i^2 - różnica pomiędzy rangami odpowiadających sobie wartości cech x_i i y_i

Ranga jest to liczba odpowiadająca miejscu w uporządkowaniu każdej z cech. Jeśli w badanej zbiorowości jest więcej jednostek z identycznym natężeniem badanej cechy, to jednostkom tym przypisuje się identyczne rangi, licząc średnią arytmetyczną z rang przynależnych tym samym jednostkom.

Współczynnik korelacji rang przyjmuje wartości z przedziału $[-1; 1]$. Interpretacja jest podobna do współczynnika korelacji liniowej Pearsona.

Przykład Nr 2

Na podstawie opinii o zdrowiu 10 pacjentów wydanych przez dwóch lekarzy chcemy ustalić współzależność między tymi opiniami, które zostały wyrażone w punktach.

Nr Pacjenta	1	2	3	4	5	6	7	8	9	10
Punkty od I lekarza	42	27	36	33	24	47	39	52	43	37
Punkty od II lekarza	39	24	35	29	26	47	44	51	39	32

Nr Pacjenta	1	2	3	4	5	6	7	8	9	10
Rangi od I lekarza	7	2	4	3	1	9	6	10	8	5
Rangi od II lekarza	6,5	1	5	3	2	9	8	10	6,5	4

Tabela pomocnicza

x_i	y_i	Ranga x	Ranga y	d_i	d_i^2
42	39	7	6,5	0,5	0,25
27	24	2	1	1	1
36	35	4	5	-1	1
33	29	3	3	0	0
24	26	1	2	-1	1
47	47	9	9	0	0
39	44	6	8	-2	4
52	51	10	10	0	0
43	39	8	6,5	1,5	2,25
37	32	5	4	1	1
Σ					10,5

$$r_s = 1 - \frac{6 \cdot 10,5}{10(100 - 1)} = 1 - \frac{63}{990} = 0,936$$