KORELACJA I REGRESJA

Populacja jest opisana dwoma lub większą liczbą cech. Czy istnieją związki pomiędzy tymi cechami, w tym:

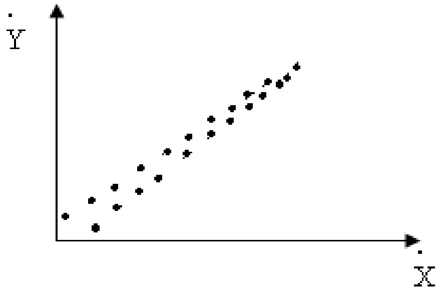
1. **Czy dwie zmienne są ze sobą skorelowane, czy też są niezależne,**
2. **Jaka jest siła zależności pomiędzy cechami,**
3. **Czy liczbowo (wzorem) można wyrazić zależność pomiędzy cechami.**

A) Badanie siły zależności pomiędzy cechami.

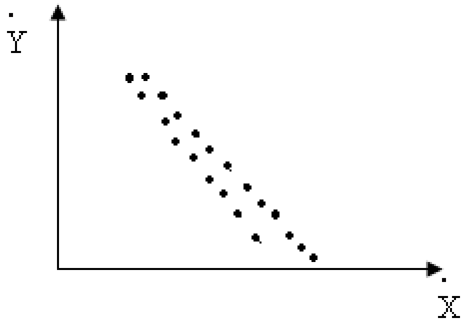
W populacji badano dwie cech *X* i *Y*. Otrzymano wyniki:

$$X: \quad x_1, x_2, \dots, x_n$$
$$Y: \quad y_1, y_2, \dots, y_n$$

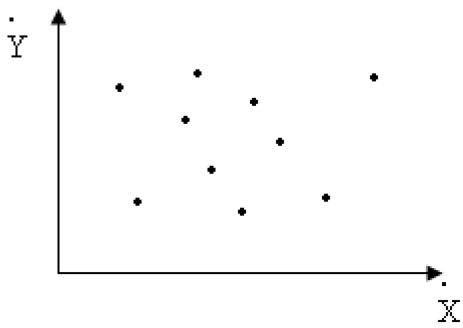
Pary liczb (x_i, y_i) można przedstawić na płaszczyźnie.



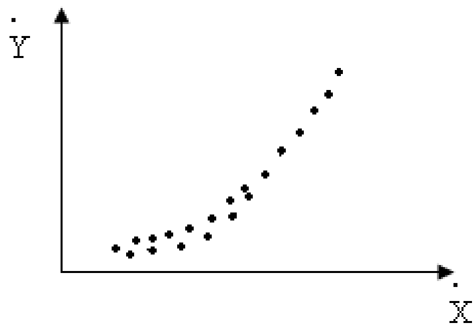
Korelacja liniowa, dodatnia, silna.



Korelacja liniowa, ujemna, silna



Brak korelacji



Korelacja krzywoliniowa, dodatnia

Kowariancja **pomiędzy cechami *X* i *Y***

- dla ciągu statystycznego $c_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

- dla szeregu statystycznego $c_{xy} = \frac{1}{n}\sum_{i=1}^{k}(x_i^s - \bar{x}) \bullet n_i (y_i - \bar{y}) \bullet m_i$

Współczynnik korelacji (Pearsona): $r_{XY} = \frac{c_{XY}}{s_X s_Y}$.

Określa on stopień (siłę zależności pomiędzy cechami. Przyjmuje się, że

$\lvert r_{XY} \rvert$	**zależność liniowa**
<0.2	**brak zależności,**
0.2 - 0.4	**zależność niska**
0.4 - 0.7	**zależność umiarkowana,**
0.7 - 0.9	**zależność znacząca**
> 0.9	**zależność bardzo silna**

Współczynnik korelacji jest określonym wskaźnikiem, a nie pomiarem na skali liniowej o jednakowych jednostkach. Oznacza to, że zależność r_{xy} = 0,90 nie jest dwukrotnie większa od r_{xy} = 0,45.

Przykład

wiek kobiety (x)	liczba dzieci (y)	(x-sr x)*(y-sr y)	x-sr x	y-sr y	Średnia X	Średnia Y	Odchylenie X	Odchylenie Y
55	5	64,5207	19,1818	3,36364	35,81818	1,636364	11,86145922	1,431637795
21	1	9,42975	-14,818	-0,6364				
35	2	-0,2975	-0,8182	0,36364				
58	2	8,06612	22,1818	0,36364				
28	1	4,97521	-7,8182	-0,6364				
30	2	-2,1157	-5,8182	0,36364				
32	3	-5,2066	-3,8182	1,36364				
20	0	25,8843	-15,818	-1,6364				
35	0	1,33884	-0,8182	-1,6364				
46	0	-16,661	10,1818	-1,6364				
34	2	-0,6612	-1,8182	0,36364				
	Kowariancja	8,11572						
	Korelacja	0,477921						

B) Regresja liniowa.

Po stwierdzeniu dużej zależności pomiędzy cechami można wyznaczyć liniową zależność pomiędzy nimi. To jest liniową funkcję regresji:

a) zmiennej *Y* względem zmiennej niezależnej *X*

$$\hat{y} = a_y x + b_y .$$

Współczynniki prostej wyznacza się zazwyczaj metodą najmniejszych kwadratów. Wówczas:

$$a_y = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_X^2} = \frac{c_{XY}}{s_X^2}$$

$$b_y = \bar{y} - a_y \bar{x}.$$

Parametry a_y, b_y nazywamy <u>*współczynnikami regresji*</u>.

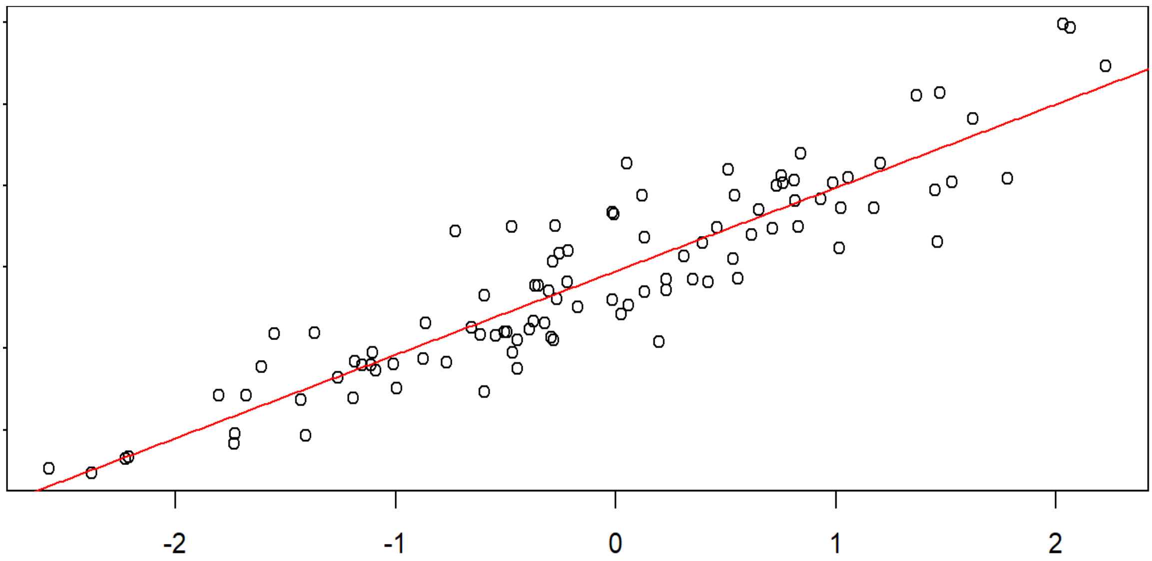
b) zmiennej *X* względem zmiennej niezależnej *Y*

$$\hat{x} = a_x y + b_x .$$

Współczynniki prostej:

$$a_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{s_Y^2} = \frac{c_{XY}}{s_Y^2}$$

$$b_x = \bar{x} - a_x \bar{y}.$$



Współczynnik korelacji kolejnościowej (rang) Spearmana

Współczynnik ten służy do opisu siły korelacji dwóch cech, szczególnie w przypadku niewielkiej liczby obserwacji.

Współczynnik rang Spearmana:

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

gdzie:

d_i – różnice między rangami odpowiadających sobie wartości cechy x_i i cechy y_i (i=1, 2, ..., n).

Współczynnik rang przyjmuje wartości z przedziału

$$-1 \le r_s \le +1,$$

a jego interpretacja jest identyczna jak współczynnika korelacji Pearsona.

Nadawanie cech

Uporządkowanym wartościom nadajemycech (czynność tęnazywamy*rangowaniem*). Może się ono odbywać od najmniejszej do największej wartości cechy, lub odwrotnie. Przy czym, sposób rangowania musi być jednakowy dla obydwu zmiennych.

W przypadku, gdy występują jednakowe wartości, wówczas przyporządkowujemy im średnią arytmetyczną obliczoną z ich kolejnych numerów.

Uwaga.

Jednakowe rangi wartości świadczą o istnieniu dodatniej korelacji między zmiennymi. Natomiast przeciwstawna numeracja sugeruje istnienie korelacji ujemnej.

Przykład6.

Na podstawie przeprowadzonej kontroli,kierownik i kontroler wydali opinię (w punktach) o każdym z pracowników.

Pracownik	A	B	C	D	E	F	G	H	I	J	K
Kierownik	41	27	35	33	25	47	38	53	43	35	36
Kontroler	38	24	34	29	27	47	43	52	39	31	29

Ustalić zależność między opiniami kierownika i kontrolera.

Rangi ocen											
Pracownik	A	B	C	D	E	F	G	H	I	J	K
Kierownik	4	10	7,5	9	11	2	5	1	3	7,5	6
Kontroler	5	11	6	8,5	10	2	3	1	4	7	8,5
różnice rang	-1	-1	1,5	0,5	1	0	2	0	-1	0,5	-2,5
kwadraty	1	1	2,25	0,25	1	0	4	0	1	0,25	6,25

Stąd:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 17}{11(121-1)} = 0,92$$

Otrzymany wynik wskazuje, że współzależność opinii jest bardzo silna. Oceniający kierowali się podobnymi kryteriami.

Przykład c.d. Obliczyć współczynnik korelacji Pearsona.