Politechnika Gdańska

Mariusz Kaszubowski

Katedra Statystyki

Wydział Zarządzania i Ekonomii

Politechnika Gdańska

Korelacja

Ćwiczenia nr 6 i 7

Statystyka opisowa

Podstawowe pojęcia

- korelacja (dodatnia, ujemna)
- współczynnik korelacji liniowej Pearsona
- współczynnik determinacji
- współczynnik korelacji rang (Spearmana, Kendalla)

Współczynniki korelacji

Współczynnik korelacji	szereg szczegółowy	szereg w tablicy korelacyjnej
Kowariancja	$\operatorname{cov}(xy)=\frac{1}{n}\cdot\sum(x_i-\bar{x})\cdot(y_i-\bar{y})$ $\operatorname{cov}(xy)=\frac{1}{n}\cdot\sum x_i\cdot y_i-\bar{x}\cdot\bar{y}$	$\operatorname{cov}(xy)=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i-\bar{x})\cdot(y_j-\bar{y})\cdot n_{ij}$ $\operatorname{cov}(xy)=\frac{1}{n}\cdot\sum_{i=1}^{k}\sum_{j=1}^{l}x_i\cdot y_j\cdot n_{ij}-\bar{x}\cdot\bar{y}$
Współczynnik korelacji liniowej Pearsona	$r_{xy}=\frac{\operatorname{cov}(xy)}{s_x\cdot s_y}=r_{yx}\qquad r_{xy}\in\langle-1;1\rangle$	$r_{xy}=\frac{\operatorname{cov}(xy)}{s_x\cdot s_y}=r_{yx}\qquad r_{xy}\in\langle-1;1\rangle$
Współczynnik determinacji	$R^2=r_{xy}^2\qquad R^2\in\langle 0;1\rangle$	
Współczynnik korelacji rang Spearmana	$R_s=1-\frac{6\cdot\sum_{i=1}^{n}d_i^2}{n^3-n}\qquad d_i-$ *różnice rang*	
Współczynnik korelacji rang Kendalla	$\tau=\frac{2S}{n(n-1)}\qquad S-$ *suma not*	

Siła i kierunek korelacji

$r_{xy} > 0$	zależność dodatnia (wprost proporcjonalna)
$r_{xy} = 0$	brak zależności
$r_{xy} < 0$	zależność ujemna (odwrotnie proporcjonalna)
$r_{xy} = 1$ lub $r_{xy} = -1$	związek funkcyjny
$\left\|r_{xy}\right\| < 0,2$	brak związku liniowego
$0,2 \le \left\|r_{xy}\right\| < 0,4$	zależna liniowa lecz słaba
$0,4 \le \left\|r_{xy}\right\| < 0,7$	zależność umiarkowana
$0,7 \le \left\|r_{xy}\right\| < 0,9$	zależność znacząca
$\left\|r_{xy}\right\| \ge 0,9$	bardzo silna zależność

Zadanie 1

Na podstawie rocznych danych dotyczących populacji bocianów X oraz ilości urodzeń żywych Y w gminie Z ustalić czy między zmiennymi X i Y istnieje (z punktu widzenia statystycznego) zależność korelacyjna. Jeśli tak, to określić jej siłę i kierunek.
Do obliczeń wykorzystaj współczynnik korelacji liniowej Pearsona, współczynnik rang Spearmana oraz rang Kendalla.

X	136	132	141	144	152	148	158	163	154	155
Y	12	4	7	11	8	5	14	12	9	7

Rozwiązanie (Pearson)

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$
1	136	12					
2	132	4					
3	141	7					
4	144	11					
5	152	8					
6	148	5					
7	158	14					
8	163	12					
9	154	9					
10	155	7					
$\sum$			X	X			

i	x_i	y_i	$x_i-\overline{x}$	$y_i-\overline{y}$	$(x_i-\overline{x})^2$	$(y_i-\overline{y})^2$	$(x_i-\overline{x})\cdot(y_i-\overline{y})$
1	136	12	-12,3	3,1	151,29	9,61	-38,13
2	132	4	-16,3	-4,9	265,69	24,01	79,87
3	141	7	-7,3	-1,9	53,29	3,61	13,87
4	144	11	-4,3	2,1	18,49	4,41	-9,03
5	152	8	3,7	-0,9	13,69	0,81	-3,33
6	148	5	-0,3	-3,9	0,09	15,21	1,17
7	158	14	9,7	5,1	94,09	26,01	49,47
8	163	12	14,7	3,1	216,09	9,61	45,57
9	154	9	5,7	0,1	32,49	0,01	0,57
10	155	7	6,7	-1,9	44,89	3,61	-12,73
$\sum$	1483	89	X	X	890,1	96,9	127,3

$$\operatorname{cov}(xy)=\frac{1}{n}\cdot\sum(x_i-\overline{x})\cdot(y_i-\overline{y})=12,73$$

$$s_x=9,43$$

$$s_y=3,11$$

$$r_{xy}=\frac{\operatorname{cov}(xy)}{s_x\cdot s_y}=0,43$$

$$R^2=r_{xy}^2=0,19$$

Rozwiązanie (Spearman)

i	x_i	y_i	rangi X	rangi Y	d_i	d_i^2
1	136	12				
2	132	4				
3	141	7				
4	144	11				
5	152	8				
6	148	5				
7	158	14				
8	163	12				
9	154	9				
10	155	7				
$\sum$			x	x		

i	x_i	y_i	rangi X	rangi Y	d_i	d_i^2
1	136	12	9	2,5	6,5	42,25
2	132	4	10	10	0	**0**
3	141	7	8	7,5	0,5	0,25
4	144	11	7	4	3	9
5	152	8	5	6	-1	1
6	148	5	6	9	-3	9
7	158	14	2	1	1	1
8	163	12	1	2,5	-1,5	2,25
9	154	9	4	5	-1	1
10	155	7	3	7,5	-4,5	20,25
$\sum$	1483	89	X	X	0	86

$$R_{xy} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n} = 0,48$$

Rozwiązanie (Kendall)

i	x_i	y_i	rangi X	rangi Y									
			1										
			2										
			3										
			4										
			5										
			6										
			7										
			8										
			9										
			10										
suma rang													

i	x_i	y_i	rangi X	rangi Y									
8	**163**	12	1	2,5									
7	**158**	14	2	1	-1								
10	**155**	7	3	7,5	1	1							
9	**154**	9	4	5	1	1	-1						
5	**152**	8	5	6	1	1	-1	1					
6	**148**	5	6	9	1	1	1	1	1				
4	**144**	11	7	4	1	1	-1	-1	-1	-1			
3	**141**	7	8	7,5	1	1	0	1	1	-1	1		
1	**136**	12	9	2,5	0	1	-1	-1	-1	-1	-1	-1	
2	**132**	4	10	10	1	1	1	1	1	1	1	1	1
suma rang					6	8	-2	2	1	-2	1	0	1

$$\tau = \frac{2S}{n(n-1)} = 0,33$$

Zadanie 2

W pewnej miejscowości zbadano wiek par heteroseksualnych przystępujących do związku małżeńskiego. Dane przedstawione w poniższej tabeli zbiorczej. Policz współczynnik korelacji liniowej Pearsona i zinterpretuj go.

Małżeństwa w 2008 roku						
Grupa wiekowa męża	**Grupa wiekowa żony**					
	y_i	15 – 25	25 – 35	35 – 45	45 – 100	n_i
x_j	środki	20	30	40	72,5	
18 – 24	21	7	5	1	0	13
25 – 35	30	12	7	2	1	22
35 – 45	40	2	6	4	0	12
45 – 100	72,5	1	2	2	3	8
n_j		22	20	9	4	55

Rozwiązanie

j	x_j	n_j	$x_j \cdot n_j$	$x_j - \overline{x}$	$(x_j - \overline{x})^2$	$(x_j - \overline{x})^2 \cdot n_j$
1	21	13				
2	30	22				
3	40	12				
4	72,5	8				
$\sum$	X	55		X	X	

i	y_i	n_i	$y_i \cdot n_i$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$(y_i - \overline{y})^2 \cdot n_i$
1	20	22				
2	30	20				
3	40	9				
4	72,5	4				
$\sum$	X	55		X	X	

Rozwiązanie

j	x_j	n_j	$x_j \cdot n_j$	$x_j - \overline{x}$	$(x_j - \overline{x})^2$	$(x_j - \overline{x})^2 \cdot n_j$
1	21	13	273	-15,24	232,15	3017,91
2	30	22	660	-6,24	38,89	855,63
3	40	12	480	3,76	14,16	169,98
4	72,5	8	580	36,26	1315,05	10520,41
$\sum$	X	55	1993	X	X	14563,93

i	y_i	n_i	$y_i \cdot n_i$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$(y_i - \overline{y})^2 \cdot n_i$
1	20	22	440	-10,73	115,07	2531,64
2	30	20	600	-0,73	0,53	10,58
3	40	9	360	9,27	85,98	773,85
4	72,5	4	290	41,77	1744,96	6979,84
$\sum$	X	55	1690	X	X	10295,91

Rozwiązanie

$$A = \begin{bmatrix} 7 & 5 & 1 & 0 \\ 12 & 7 & 2 & 1 \\ 2 & 6 & 4 & 0 \\ 1 & 2 & 2 & 3 \end{bmatrix} \qquad B = \begin{bmatrix} -15{,}24 \\ -6{,}24 \\ 3{,}76 \\ 36{,}26 \end{bmatrix} \qquad C = \begin{bmatrix} -10{,}73 \\ -0{,}73 \\ 9{,}27 \\ 41{,}77 \end{bmatrix}$$

$$\sum_{ij} \left(x_j - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) \cdot n_{ij} = \det\left(B^T \cdot A \cdot C\right) = 6334,296$$

$$\operatorname{cov}(xy) = \frac{6334,296}{55} = 115,169$$

$$s_x = 16,27$$

$$s_y = 13,68$$

$$r_{xy} = \frac{\operatorname{cov}(xy)}{s_x \cdot s_y} = 0,52$$

Zadanie 3

Na podstawie danych dotyczących wydajności pracy Y i stażu pracy X dla 10 robotników pewnej firmy ustalić czy między zmiennymi X i Y istnieje zależność korelacyjna. Jeśli tak, to określić jej kierunek.

Do obliczeń wykorzystaj współczynnik korelacji liniowej Pearsona, współczynnik rang Spearmana oraz rang Kendalla.

X	1	5	10	8	9	1	2	4	5	6
Y	120	115	132	123	128	102	106	109	112	110

Zadanie 4

Policzyć macierz korelacji pomiędzy następującymi zmiennymi i podaj interpretację obliczonych współczynników.

rok	rozwody	zgony	małżeństwa	urodzenia
1999	42020	382170	204656	384379
2000	42770	368782	208366	380476
2001	45308	363963	192295	370247
2002	45414	360170	191978	355526
2003	48632	365945	195495	352785
2004	56332	363522	191890	357884
2005	67578	368285	206974	366095
2006	71912	369686	226257	376035