



KORELACJA

Dawid Słomian

WSTĘP

- Korelacja liniowa Pearsona
 - obliczanie
 - testowanie
 - korelacja wielokrotna
- Korelacja rangowa Spearmana
 - obliczanie
 - testowanie
- Inne miary korelacji

WSPÓŁCZYNNIK KORELACJI PEARSONA



WSPÓŁCZYNNIK KORELACJI PEARSONA - DEFINICJA

oryginalne		odchylenia	
masa	tłuszcz	masa	tłuszcz
89	28	9.6	1.7
88	27	8.6	0.7
66	24	-13.4	-2.3
59	23	-20.4	-3.3
93	29	13.6	2.7
73	25	-6.4	-1.3
82	29	2.6	2.7
77	25	-2.4	-1.3
100	30	20.6	3.7
67	23	-12.4	-3.3
średnia:			
79.4	26.3	0.000	0.000

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- odchylenia od średniej:

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

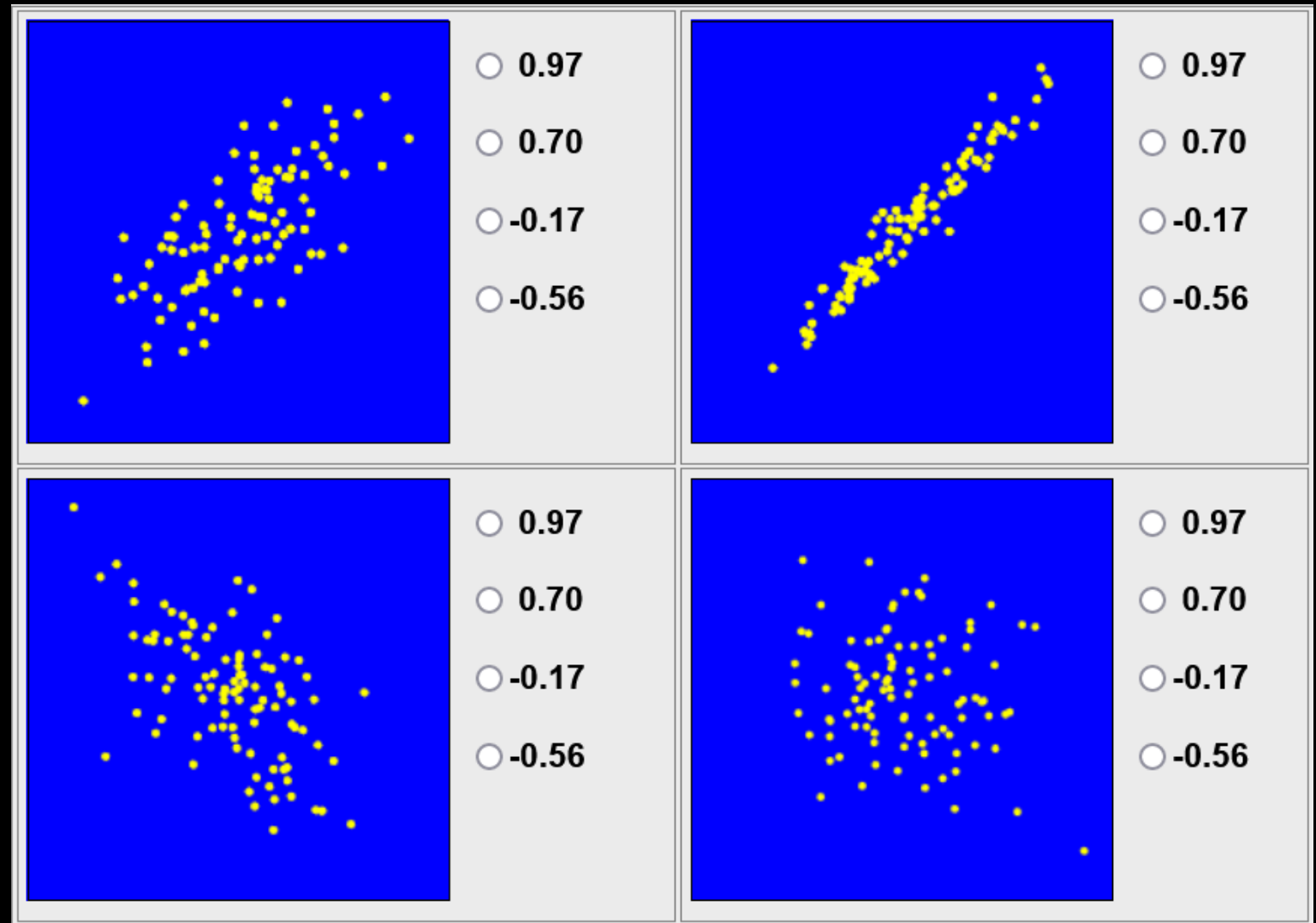
WSPÓŁCZYNNIK KORELACJI PEARSONA - DEFINICJA

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

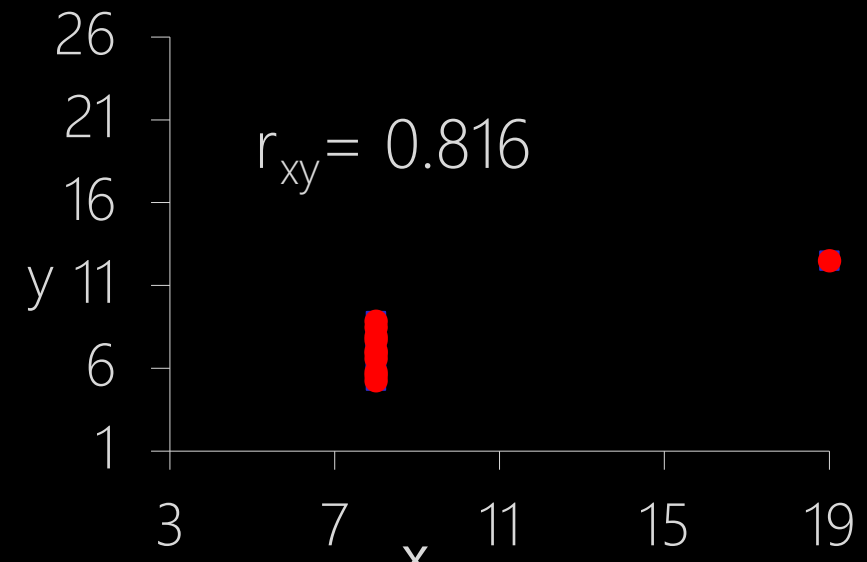
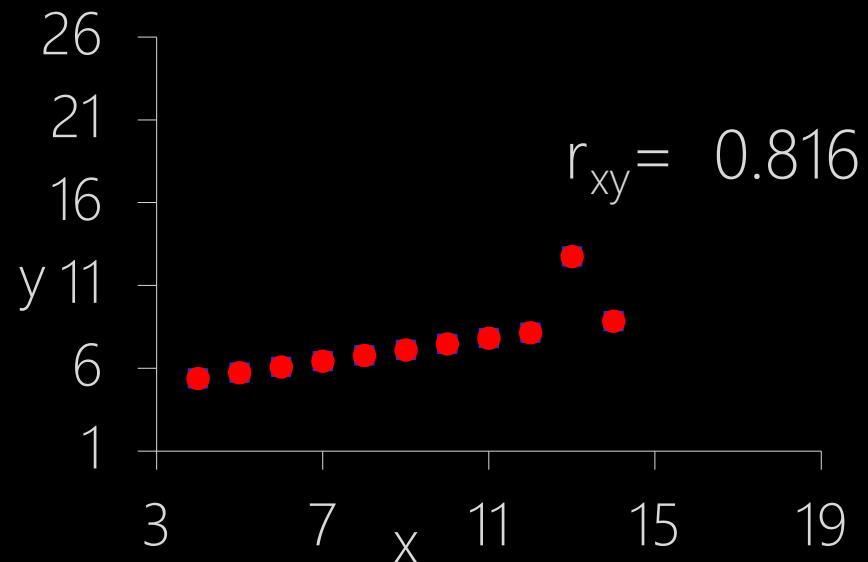
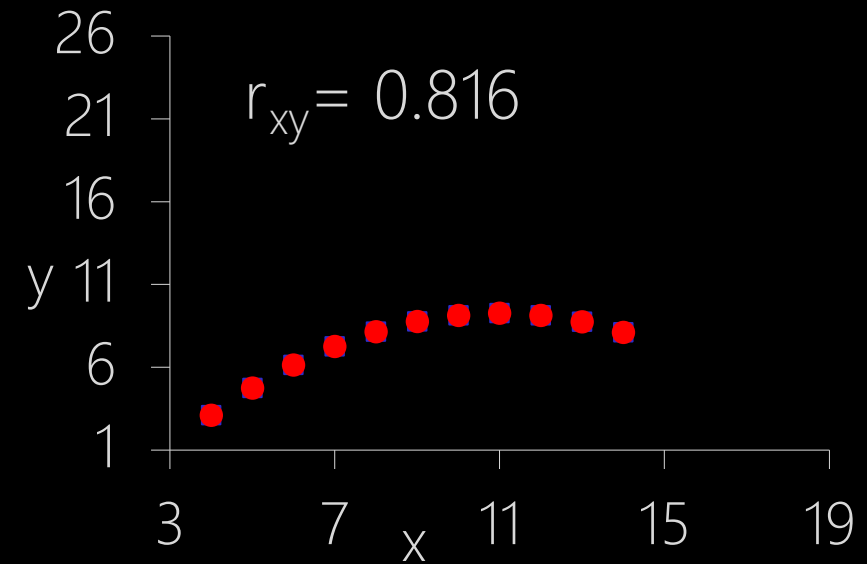
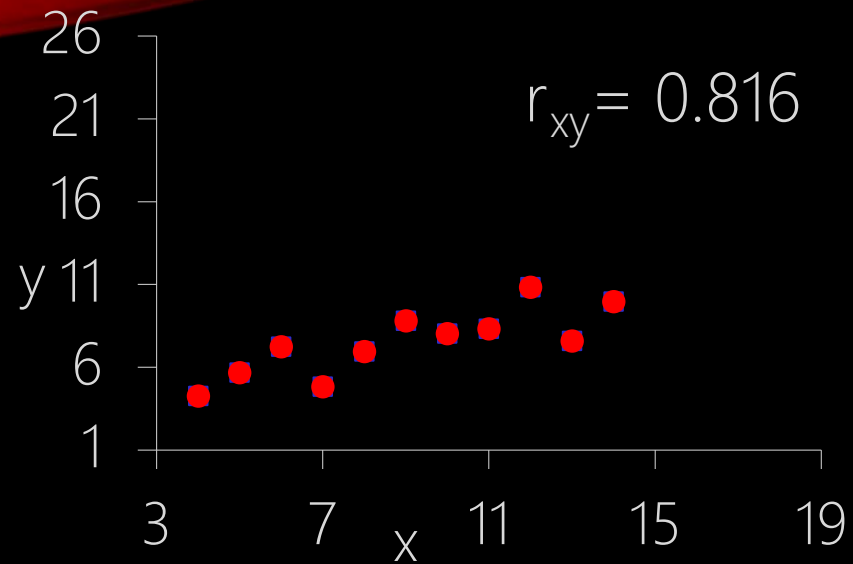
- Miara zależności pomiędzy 2ma zmiennymi (x,y)
- Założenia:
 - ciągłe wartości zmiennych
 - normalny rozkład zmiennych
 - mierzy zależność liniową
- Wartości [-1, 1]

WSPÓŁCZYNNIK KORELACJI PEARSONA - PRZYKŁADY

- Correlation Game
- istics.net/Correlations/



WSPÓŁCZYNNIK KORELACJI PEARSONA - PRZYKŁADY



WSPÓŁCZYNNIK KORELACJI PEARSONA - PRZYKŁADY

MASA CIAŁA	ZAW. TŁUSZCZU
89	28
88	27
66	24
59	23
93	29
73	25
82	29
77	25
100	30
67	23



Współczynnik korelacji
Pearsona dla danych
 $r_{mt} = 0.94$

WSPÓŁCZYNNIK KORELACJI PEARSONA - TESTOWANIE

- Hipotezy
 - H_0 : brak korelacji między masą ciała, a zaw. tłuszczu
 - H_1 : istnieje korelacja między masą ciała, a zaw. tłuszczu
 - $H_0: r_{mt} = 0$ $H_1: r_{mt} \neq 0$
- Założone maksymalne prawdopodobieństwo błędu $\alpha_{MAX} = 0.01$

- Test:

$$t = \frac{r_{tm} \sqrt{N-2}}{\sqrt{1-r_{tm}^2}} \sim t_{N-2}$$

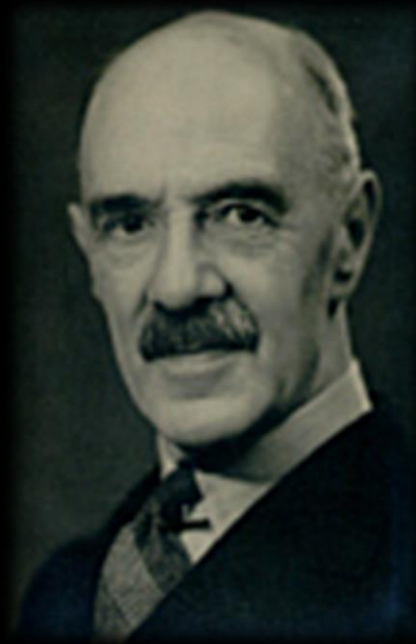
- Prawdopodobieństwo błędu dla $t=7.47$ wynosi $\alpha_T=0.00007$
- $\alpha_{MAX} > \alpha_T$
- $\rightarrow H_1$
- Występuje pozytywna korelacja między masą ciała, a zawartością tłuszczu

WSPÓŁCZYNNIK KORELACJI WIELOKROTNEJ - DEFINICJA

$$R_{xy_n} = \begin{bmatrix} r_{xy_1} & r_{xy_2} & r_{xy_3} \end{bmatrix} \begin{bmatrix} 1 & r_{y_1y_2} & r_{y_1y_3} \\ & 1 & r_{y_2y_3} \\ \text{symm} & & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_{xy_1} \\ r_{xy_2} \\ r_{xy_3} \end{bmatrix}$$

- Miara jak daną zmienną można przewidzieć stosując liniową funkcję innych zmiennych
- Mierzy jedynie siłę lecz nie kierunek zmian
- Wartości [0, 1]

WSPÓŁCZYNNIK KORELACJI SPEARMANA

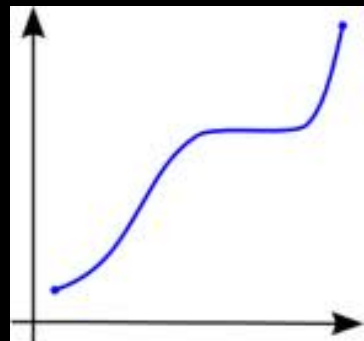
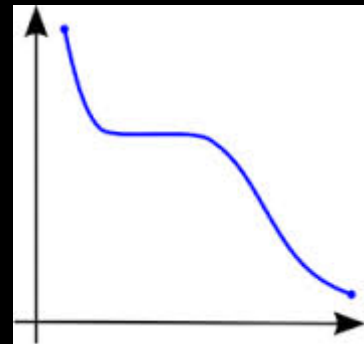


WSPÓŁCZYNNIK KORELACJI SPEARMANA - DEFINICJA

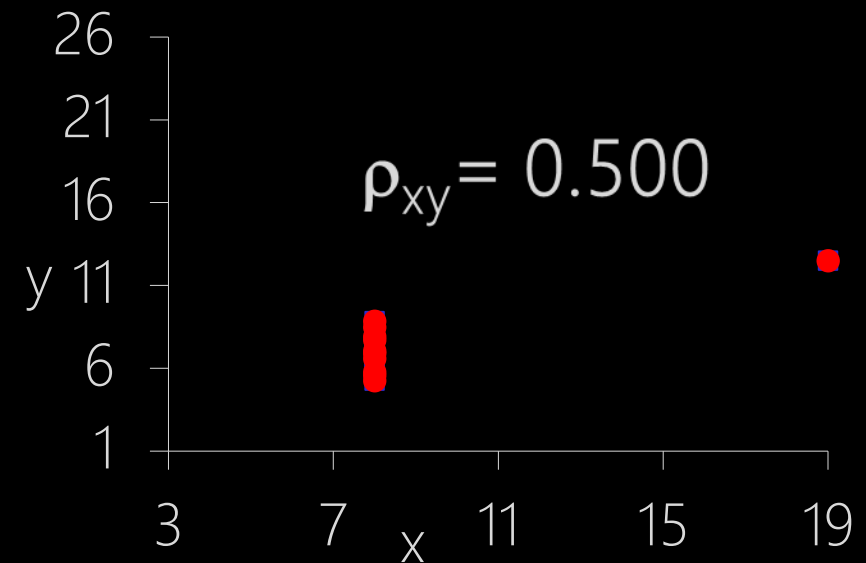
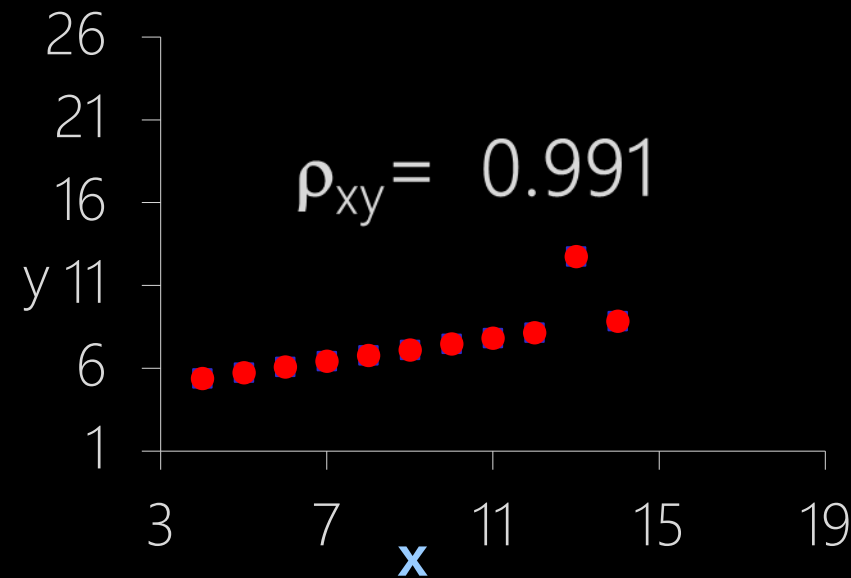
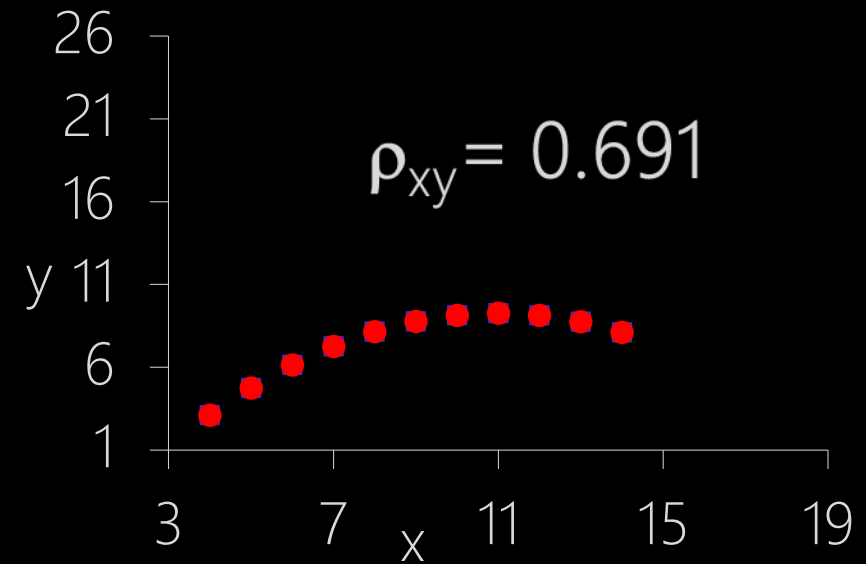
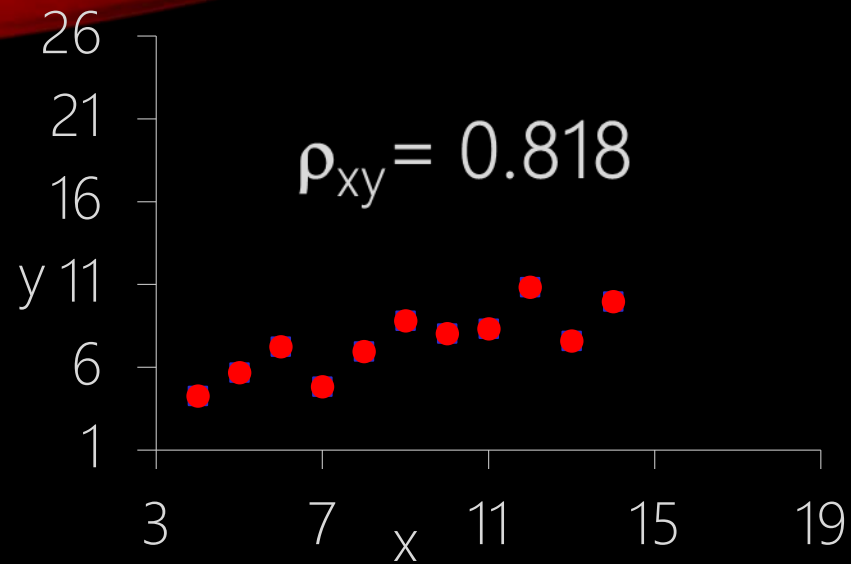
$$\rho_{xy} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{N(N^2 - 1)}$$

różnica w
rankingu
zmiennych x i y

- Miara zależności pomiędzy 2ma zmiennymi (x,y)
- Brak założeń dotyczących rozkładu zmiennych
- Brak założeń dotyczących liniowej zależności pomiędzy zmiennymi
- Mierzy zależność monotoniczną
- Wykorzystuje ranking obserwacji
- Wartości [-1, 1]



WSPÓŁCZYNNIK KORELACJI SPEARMANA - PRZYKŁADY



WSPÓŁCZYNNIK KORELACJI SPEARMANA- PRZYKŁADY

- 18 samców Fregata magnificens
- Powiązanie objętości worka z częstotliwością wydawanego dźwięku

The Condor 106:156-160
© The Cooper Ornithological Society 2004

BIMODAL SIGNALING OF A SEXUALLY SELECTED TRAIT: GULAR POUCH DRUMMING IN THE MAGNIFICENT FRIGATEBIRD

S. BALSBY², TORBEN DABELSTEEN² AND JOSÉ LUIS OSORNO¹
¹Universidad Nacional Autónoma de México, Apartado Postal 70-275,
Ciudad de México, D. F., México
²University of Copenhagen, Tagensvej 16, 2200 Copenhagen N, Denmark



$$\rho_{ct} = -0.76$$

objętość
[cm³]

częstotliwość
[Hz]

1760

529

2040

566

2440

473

2550

461

2730

465

2740

532

3010

484

3080

527

3370

488

3740

485

WSPÓŁCZYNNIK KORELACJI SPEARMANA - TESTOWANIE

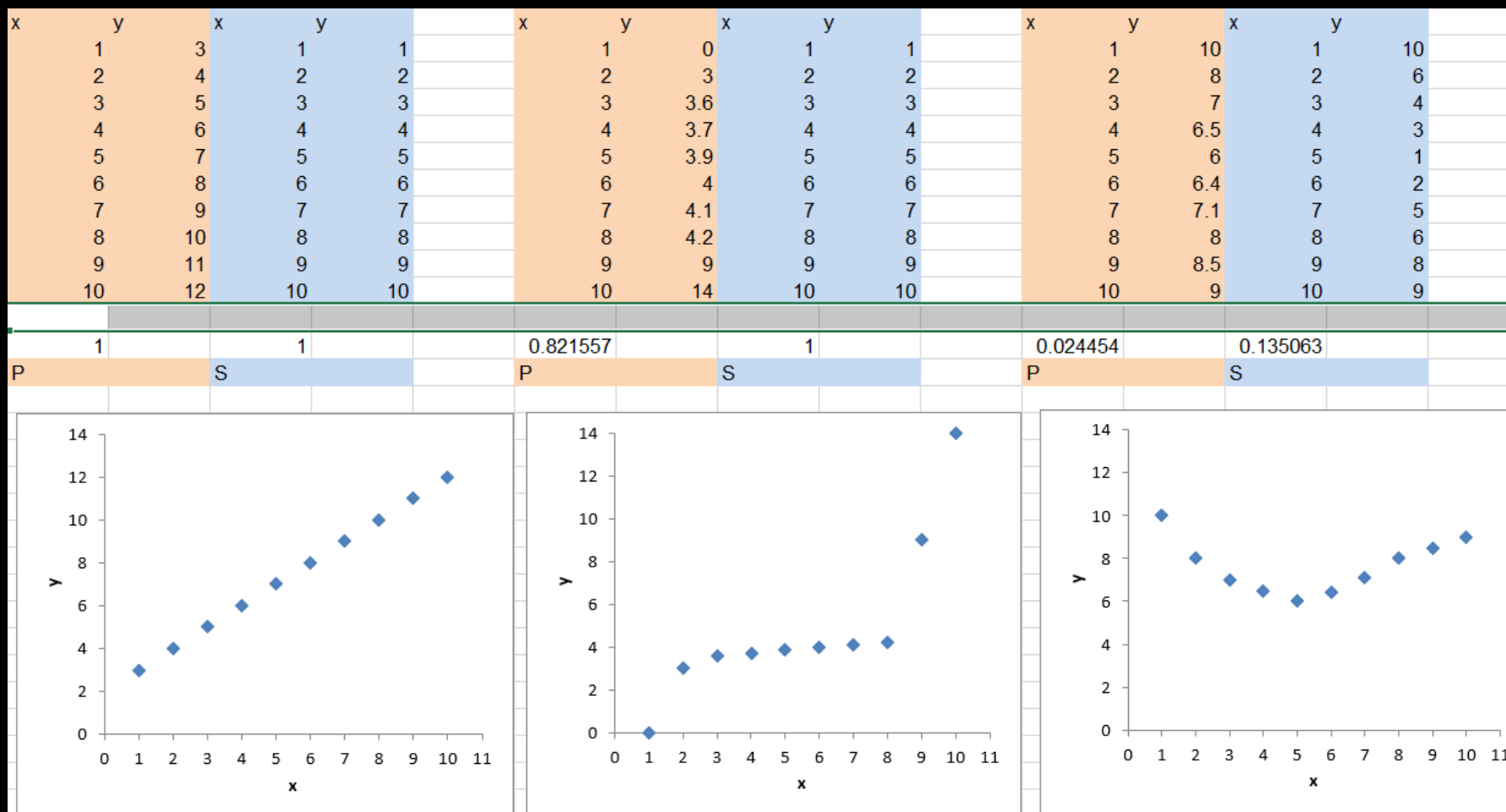
- Hipotezy
 - H0: brak korelacji między objętością, a częstotliwością
 - H1: istnieje korelacja między objętością, a częstotliwością
 - H0: $\text{sigmsact} = 0$ H1: $\text{sigmact} \neq 0$
- Założone maksymalne prawdopodob. błędu $\alpha_{\text{MAX}} = 0.01$

- Test:

$$t = \frac{\rho \sqrt{N-2}}{\sqrt{1-\rho^2}} \sim t_{N-2}$$

- Prawdopodobieństwo błędu dla $t = -4.68$ wynosi $\alpha_T = 0.00019$
- $\alpha_{\text{MAX}} > \alpha_T$
- $\rightarrow H1$
- Występuje ujemna korelacja między objętością worka, a częstotliwością dźwięku

WSPÓŁCZYNNIK KORELACJI PEARSONA VS. SPEARMANA



INNE MIARY

Abbre.	Method	Symbol	Description	References
Pearson	Pearson's	r	Linear, widely-used, no parameter, coeff. $\in [-1, 1]$	Pearson, 1895
Spearman	Spearman's	ρ	Monotonic, rank-based, no parameter, coeff. $\in [-1, 1]$	Spearman, 1904
Kendall	Kendall's	τ	Monotonic, rank-based, no parameter, coeff. $\in [-1, 1]$	Kendall, 1938
Hoeffding	Hoeffding's	D	Non-linear, rank-based, no parameter, coeff. $\in [0, 1]$	Hoeffding, 1948
Blomqvist	Blomqvist's	β	Monotonic, rank-based, no parameter, coeff. $\in [-1, 1]$	Blomqvist, 1950
Goodman	Goodman and Kruskal's	γ	Monotonic, cross classifications, rank-based, no parameter, coeff. $\in [-1, 1]$	Goodman and Kruskal, 1954
WWH	Wang, Waterman, Huang's	wwh	Monotonic, rank-based, no parameter, coeff. $\in [0, +\infty]$	Wang et al., 2014
MI	Mutual information	I	Non-linear, entropy-based, no parameter, coeff. $\in [0, +\infty]$	Shannon, 1948
MIC	Maximum information correlation	mic	Non-linear, entropy-based, 1 parameter, coeff. $\in [0, 1]$	Reshef et al., 2011
Wilks	Wilks'	W	Linear, covariance-based, no parameter, coeff. $\in [0, 1]$	Wilks, 1935
KCCA	Kernel canonical correlation analysis	$kcca$	Non-linear, covariance-based, 1 parameter, coeff. $\in [0, 1]$	Bach and Jordan, 2002
dCor	Distance correlation	$dCor$	Non-linear, covariance-based, 1 parameter, coeff. $\in [0, 1]$	Szekely and Rizzo, 2009
CMMD	copula-based maximum mean discrepancy	$cmmd$	Non-linear, copulas-based, 1 parameter, coeff. $\in [0, 1]$	Poczos et al., 2012
RDC	Randomized dependence coefficient	rdc	Non-linear, copulas-based, 2 parameters, coeff. $\in [0, 1]$	Lopez-Paz et al., 2013

