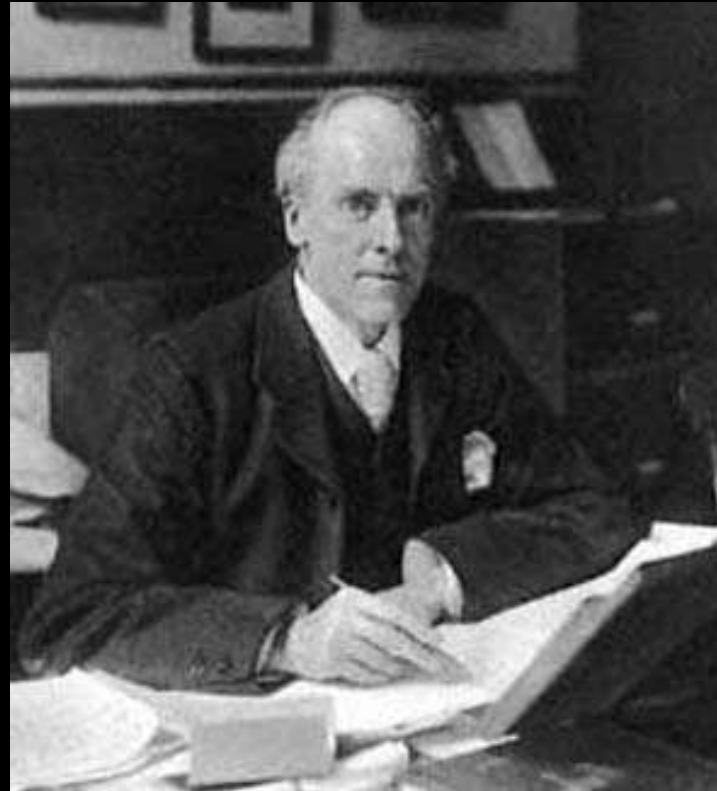


METODY STATYSTYCZNE W BIOLOGII

1. Wykład wstępny
2. Populacje i próby danych
3. Testowanie hipotez i estymacja parametrów
4. Planowanie eksperymentów biologicznych
5. Najczęściej wykorzystywane testy statystyczne I
6. Najczęściej wykorzystywane testy statystyczne II
7. Regresja liniowa
8. Regresja nieliniowa
9. Określenie jakości dopasowania równania regresji liniowej i nieliniowej
- 10. Korelacja**
11. Elementy statystycznego modelowania danych
12. Porównywanie modeli
13. Analiza wariancji
14. Analiza kowariancji
15. Podsumowanie materiału, wspólna analiza przykładów, dyskusja

1. Korelacja liniowa Pearsona
 - obliczanie
 - testowanie
 - korelacja wielokrotna
2. Korelacja rangowa Spearmana
 - obliczanie
 - testowanie
3. Inne miary *korelacji*

Współczynnik korelacji Pearsona



Współczynnik korelacji Pearsona - definicja

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

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

odchylenia od średniej:

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

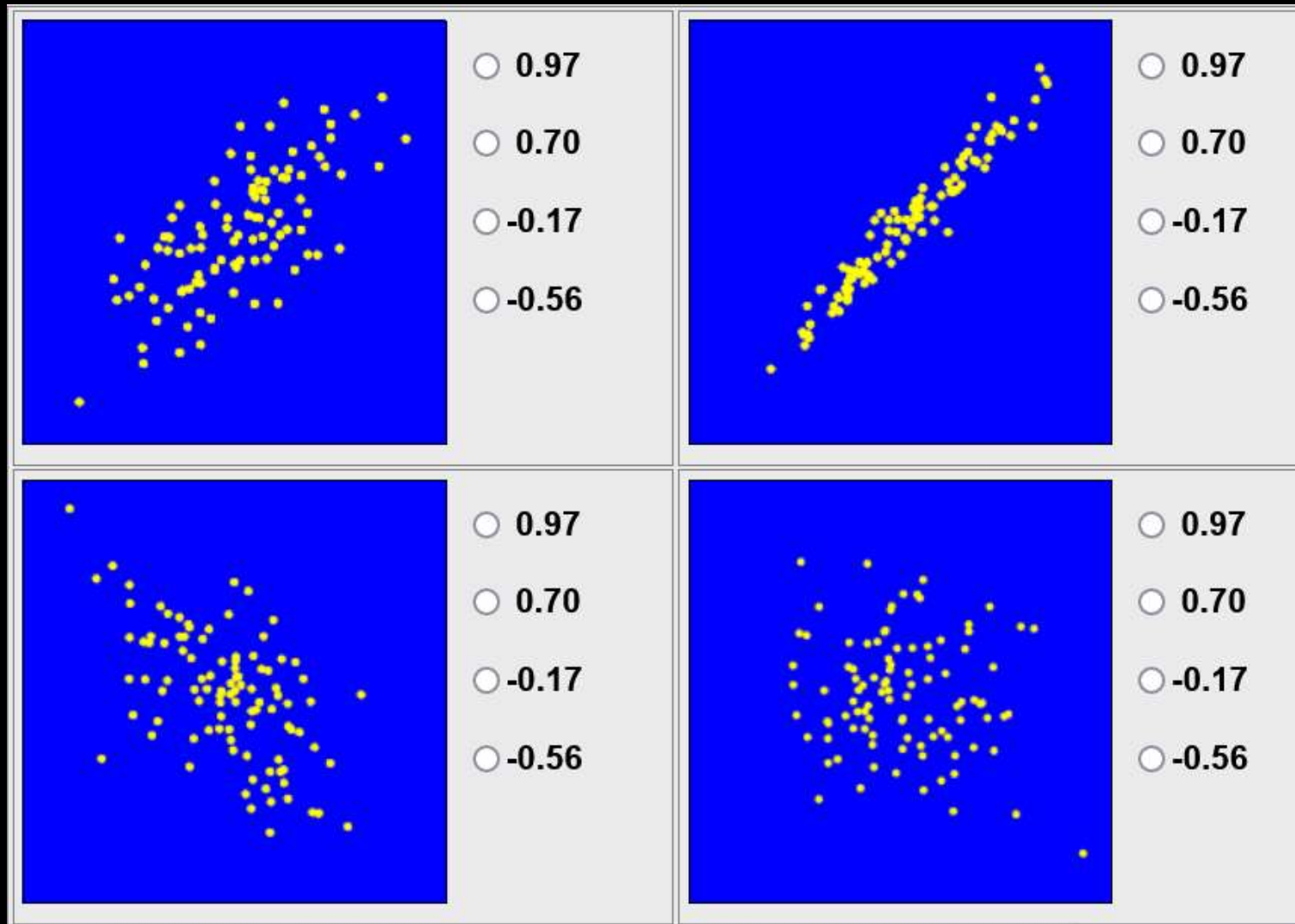
Współczynnik korelacji Pearsona - definicja

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

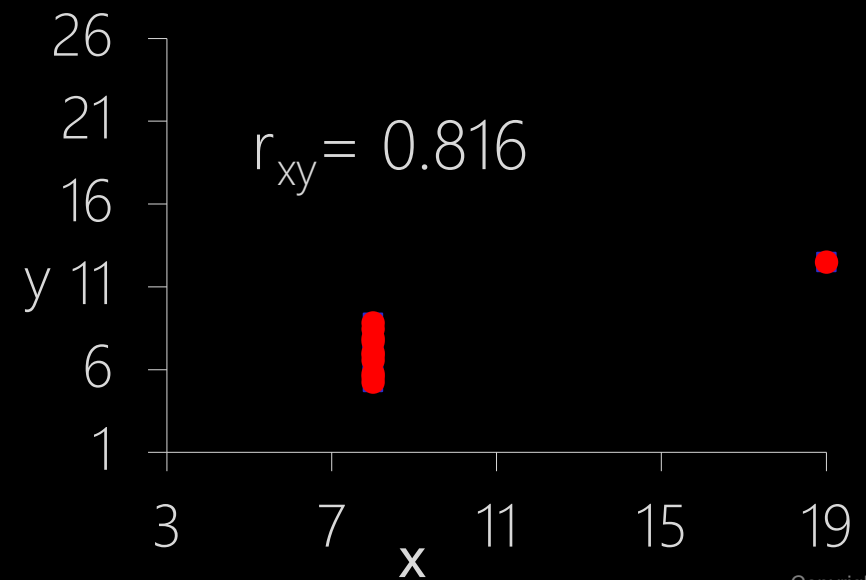
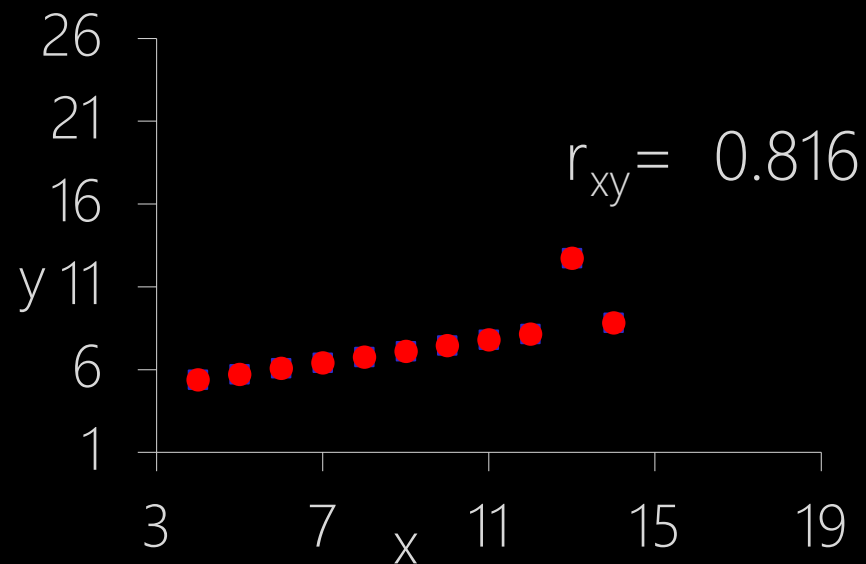
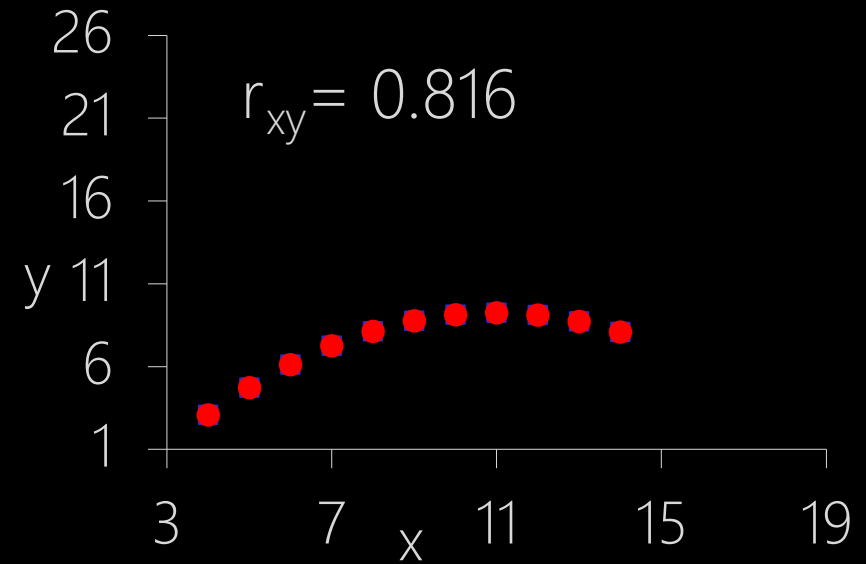
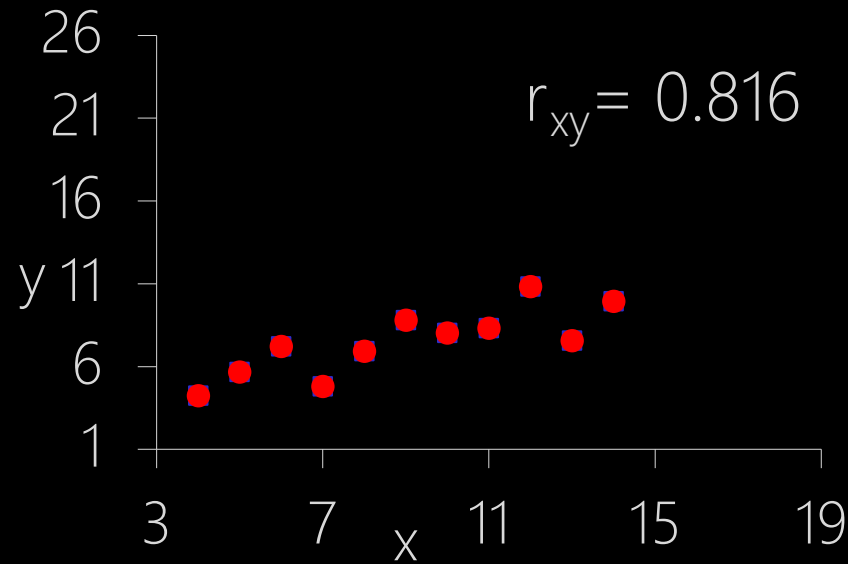
1. Miara zależności pomiędzy 2ma zmiennymi (x,y)
2. Założenia:
 - ciągłe wartości zmiennych
 - normalny rozkład zmiennych
 - mierzy zależność **liniową**
3. Wartości [-1, 1]

Współczynnik korelacji Pearsona - przykłady

Correlation Game
istics.net/Correlations/



Współczynnik korelacji Pearsona - przykłady



Współczynnik korelacji Pearsona - przykłady

próba danych

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



$$r_{mt} = 0.94$$

Współczynnik korelacji Pearsona - testowanie

1. Hipotezy

- H_0 : brak korelacji między masą ciała, a zaw. tłuszczu
- H_1 : istnieje korelacja między masą ciała, a zaw. tłuszczu
- $H_0: r_{mt} = 0$ $H_1: r_{mt} \neq 0$

2. Założone maksymalne prawdopodobieństwo błędu $\alpha_{MAX} = 0.01$

3. Test:

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

4. Prawdopodobieństwo błędu dla $t=7.47$ wynosi $\alpha_T=0.00007$

5. $\alpha_{MAX} > \alpha_T$

6. $\rightarrow H_1$

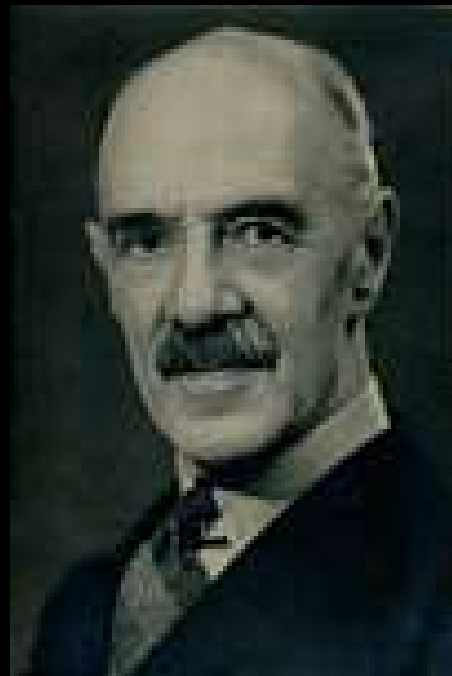
7. Występuje pozytywna korelacja między masą ciała, a zawartością tłuszczu

Współczynnik korelacji wielokrotnej - definicja

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

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

Współczynnik korelacji Spearmana

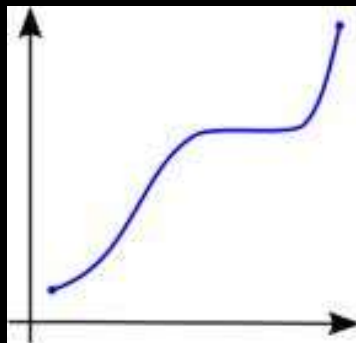
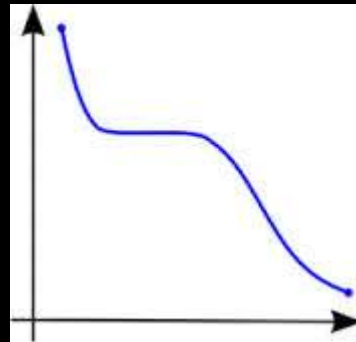


Współczynnik korelacji Spearmana - definicja

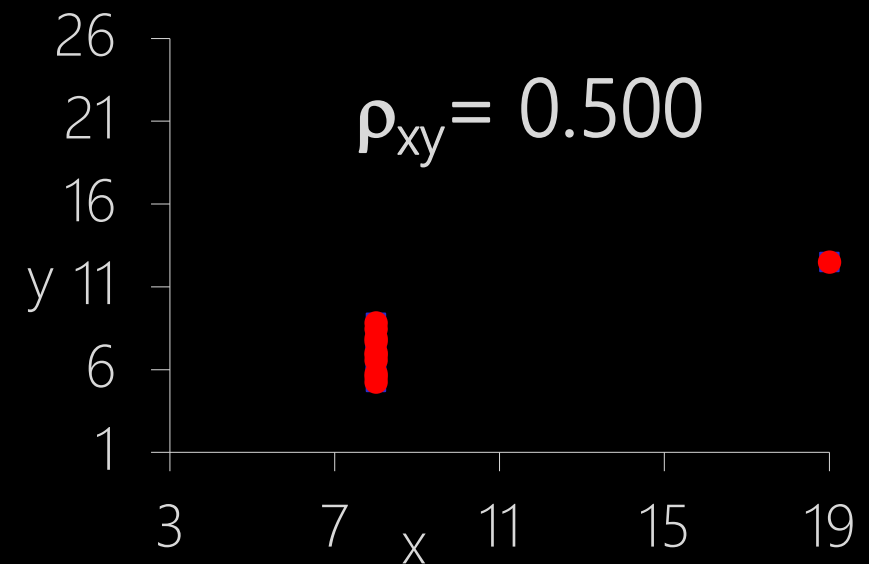
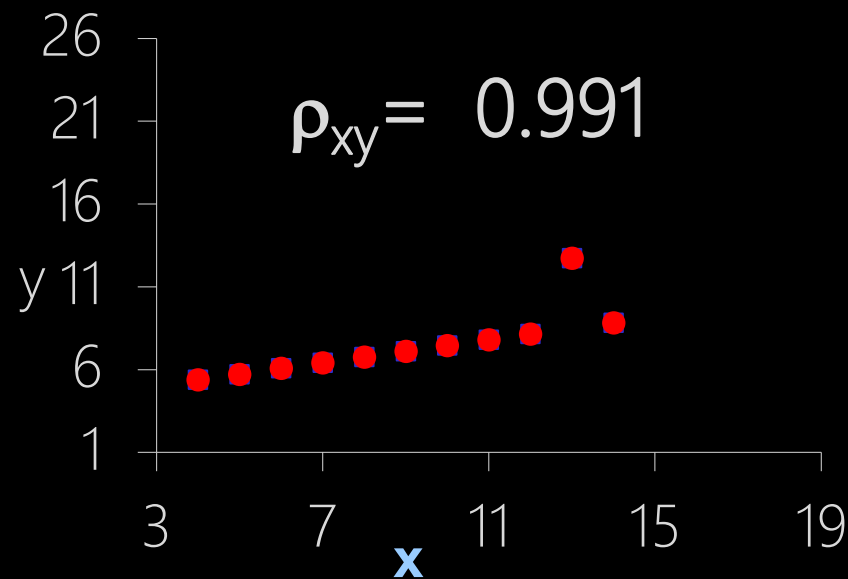
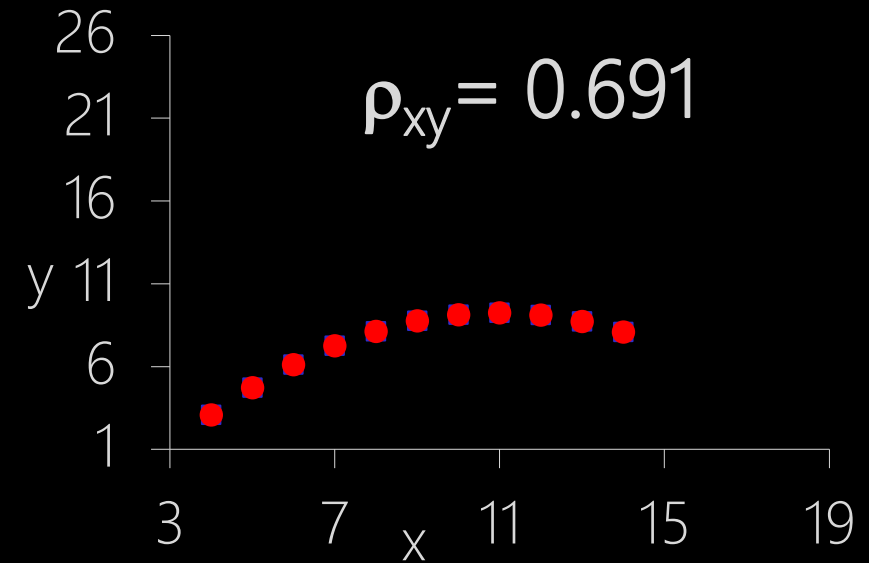
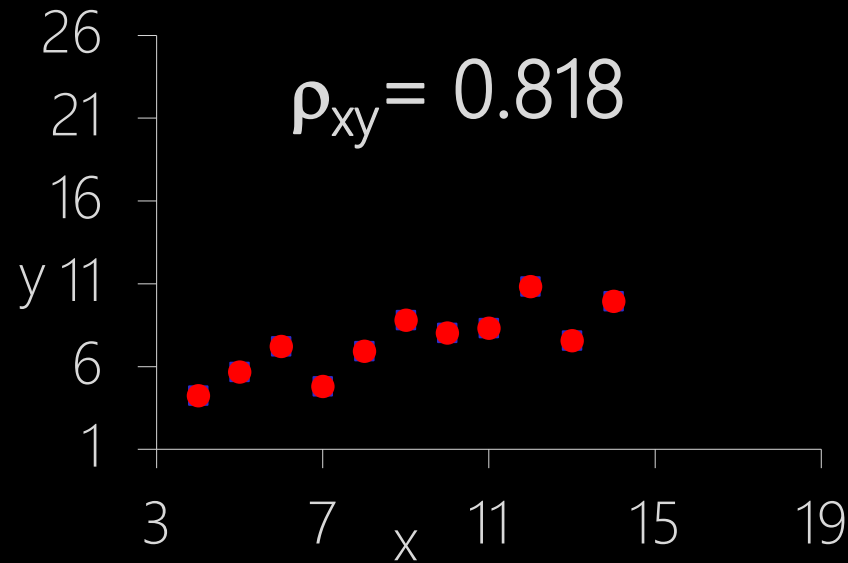
$$\rho_{xy} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{N(N^2 - 1)}$$

różnica w
rankingu
zmiennych x i y

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



Współczynnik korelacji Spearmana - przykłady



Współczynnik korelacji Spearmana- przykłady

próba danych

1. 18 samców *Fregata magnificens*
2. Powiązanie objętości worka z częstotliwością wydawanego dźwięku

objętość [cm ³]	częstotliwość [Hz]
1760	529
2040	566
2440	473
2550	461
2730	465
2740	532
3010	484
3080	527
3370	488
3740	485



$\rho_{ct} = -0.76$

Współczynnik korelacji Spearmana - testowanie

1. Hipotezy

- H_0 : brak korelacji między objętością, a częstotliwością
- H_1 : istnieje korelacja między objętością, a częstotliwością
- $H_0: \rho_{ct} = 0$ $H_1: \rho_{ct} \neq 0$

2. Założone maksymalne prawdopodob. błędu $\alpha_{MAX} = 0.01$

3. Test:

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

4. Prawdopodobieństwo błędu dla $t=-4.68$ wynosi $\alpha_T=0.00019$

5. $\alpha_{MAX} > \alpha_T$

6. $\rightarrow H_1$

7. Występuje ujemna korelacja między objętością worka, a częstotliwością dźwięku

Współczynnik korelacji Pearsona vs. Spearmana

x	y	x	y
1	3	1	1
2	4	2	2
3	5	3	3
4	6	4	4
5	7	5	5
6	8	6	6
7	9	7	7
8	10	8	8
9	11	9	9
10	12	10	10

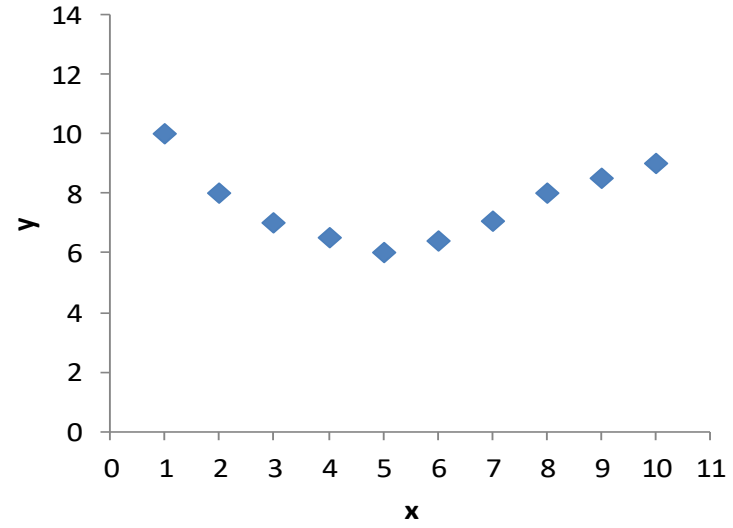
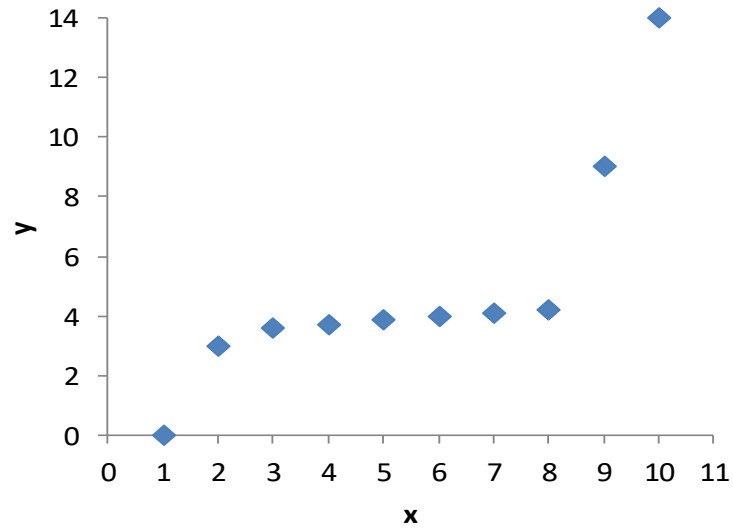
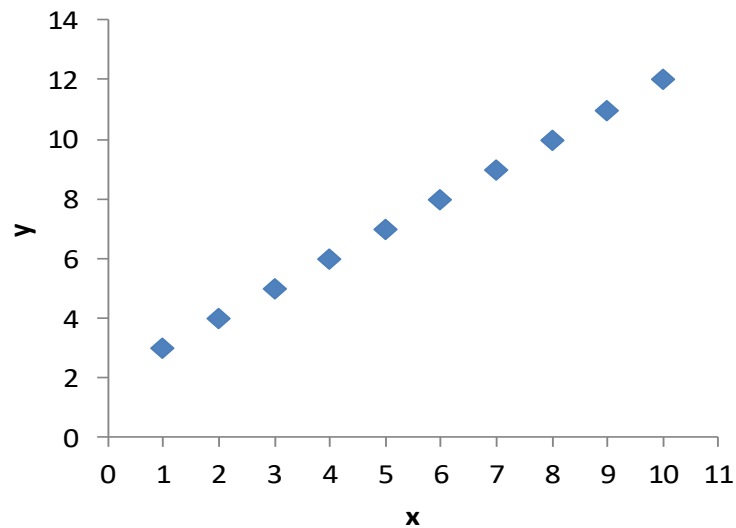
P S

x	y	x	y
1	0	1	1
2	3	2	2
3	3.6	3	3
4	3.7	4	4
5	3.9	5	5
6	4	6	6
7	4.1	7	7
8	4.2	8	8
9	9	9	9
10	14	10	10

P S

x	y	x	y
1	10	1	10
2	8	2	6
3	7	3	4
4	6.5	4	3
5	6	5	1
6	6.4	6	2
7	7.1	7	5
8	8	8	6
9	8.5	9	8
10	9	10	9

P S



Inne miary korelacji

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

Hoeffding's dependence coefficient

$$H = \frac{(n-2)(n-3)D_1 + D_2 - (n-2)2D_3}{n(n-1)(n-2)(n-3)(n-4)}$$

$$D_1 = \sum_{i=1}^n (Q_i - 1)(Q_i - 2)$$

$$D_2 = \sum_{i=1}^n (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$$

$$D_3 = \sum_{i=1}^n (R_i - 2)(S_i - 2)(Q_i - 1)$$

R_i ranga X_i

S_i ranga Y_i

Q_i liczba obserwacji dla których pozostałe X i $Y < X_i$ i Y_i



Arkusz programu
Microsoft Excel

1. Korelacja liniowa Pearsona

- obliczanie
- testowanie
- korelacja wielokrotna

2. Korelacja rangowa Spearmana

- obliczanie
- testowanie

3. Inne miary *korelacji*