Badanie zależności między cechami

Obiekt $\longrightarrow (X, Y)$

$$H_0 : \text{Cechy } X \text{ oraz } Y \text{ są niezależne}$$

Próba: $(X_1, Y_1), \ldots, (X_n, Y_n)$

test chi–kwadrat niezależności

współczynnik korelacji Pearsona

współczynnik korelacji rangowej Spearmana
współczynnik korelacji rangowej Kendalla

Test chi–kwadrat niezależności

Klasy cechy Y	Klasy cechy X 1	2	$\ldots$	m
1	n_{11}	n_{12}	$\ldots$	n_{1m}
2	n_{21}	n_{22}	$\ldots$	n_{2m}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
k	n_{k1}	n_{k2}	$\ldots$	n_{km}

Statystyka testowa

$$\chi^2_{\text{emp}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n^t_{ij})^2}{n^t_{ij}}$$

$$n^t_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}, \ N = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}$$

$$n_{i\cdot} = \sum_{j=1}^{m} n_{ij}, \ n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$$

$$P_{H_0}\{\chi^2_{\text{emp}} \le y\} \approx \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^y x^{\nu/2-1} e^{-x/2} dx$$

$$\nu = (k-1)(m-1)$$

Przykład. W celu zbadania istnienia związku między wykształceniem (X) a zarobkami (Y) wylosowano 950 osób. Uzyskano następujące dane

		podstawowe (W_1)	średnie (W_2)	wyższe (W_3)	ponad wyższe (W_4)
(Z_1)	$\leq$500	21	41	93	47
(Z_2)	500−1000	33	37	35	53
(Z_3)	1000−1500	45	75	27	43
(Z_4)	1500−2000	30	48	50	55
(Z_5)	$\geq$2000	71	47	49	50

Czy powyższe dane świadczą o istnieniu zależności między wykształceniem i zarobkami?

Zbadano łącznie $N = 950$ osób
Liczebności brzegowe:

$$n_{1\cdot} = 21 + 41 + 93 + 47 = 202$$

$$n_{2\cdot} = 158,\ n_{3\cdot} = 190,\ n_{4\cdot} = 183,\ n_{5\cdot} = 217$$

$$n_{\cdot 1} = 21 + 33 + 45 + 30 + 71 = 200$$

$$n_{\cdot 2} = 248,\ n_{\cdot 3} = 254,\ n_{\cdot 4} = 248.$$

	W_1	W_2	W_3	W_4	
Z_1	n_{11}=21	n_{12}=41	n_{13}=93	n_{14}=47	$n_{1.}$=202
Z_2	n_{21}=33	n_{22}=37	n_{23}=35	n_{24}=53	$n_{2.}$=158
Z_3	n_{31}=45	n_{32}=75	n_{33}=27	n_{34}=43	$n_{3.}$=190
Z_4	n_{41}=30	n_{42}=48	n_{43}=50	n_{44}=55	$n_{4.}$=183
Z_5	n_{51}=71	n_{52}=47	n_{53}=49	n_{54}=50	$n_{5.}$=217
	$n_{.1}$=200	$n_{.2}$=248	$n_{.3}$=254	$n_{.4}$=248	N=950

Liczebności teoretyczne:

$$n_{11}^t = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N} = \frac{202 \cdot 200}{950} = 42.5263$$

$$n_{43}^t = \frac{n_{4\cdot} \cdot n_{\cdot 3}}{N} = \frac{183 \cdot 254}{950} = 48.9284$$

Wyznaczenie $(n_{ij} - n_{ij}^t)^2/n_{ij}^t$ dla wszystkich dwudziestu kombinacji i, j.

$$\frac{(n_{11} - n_{11}^t)^2}{n_{11}^t} = \frac{(21 - 42.5263)^2}{42.5263} = 10.8964$$

$$\frac{(n_{43} - n_{43}^t)^2}{n_{43}^t} = \frac{(50 - 48.9284)^2}{48.9284} = 0.0235$$

Wartość statystyki testowej

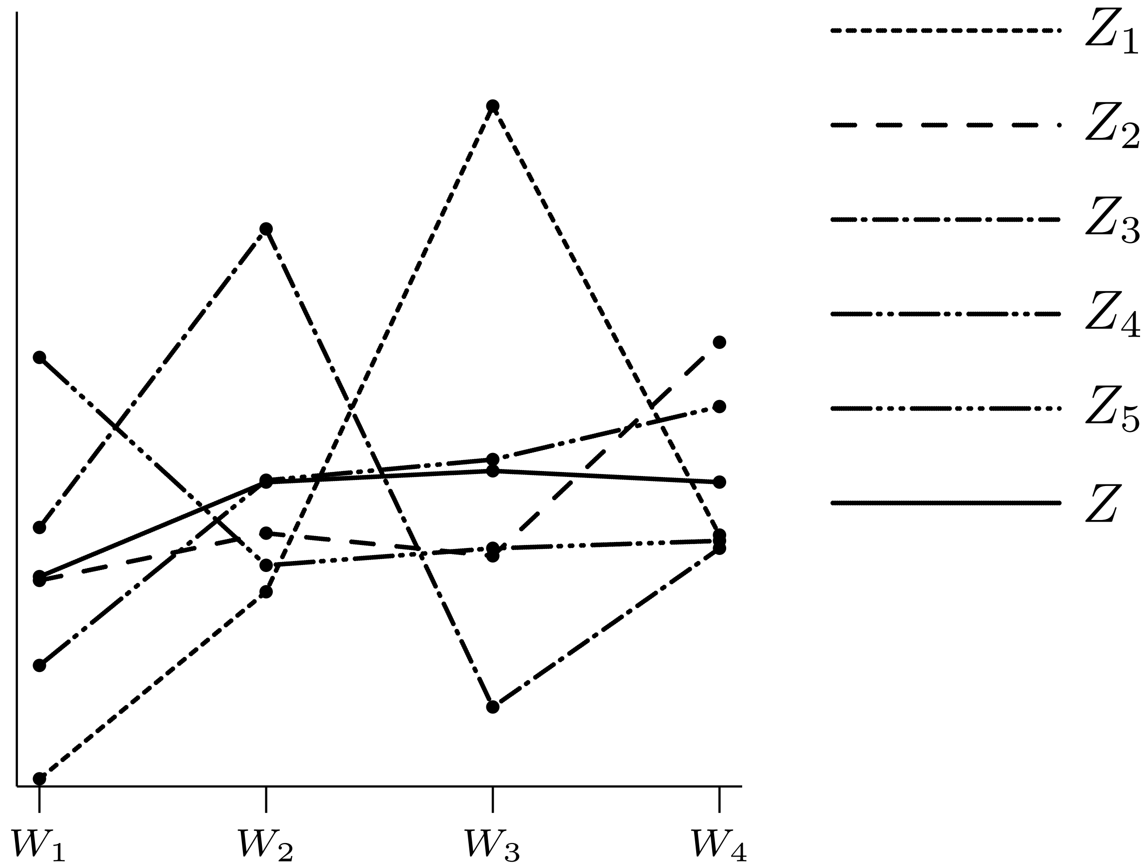
$$\chi^2_{\mathrm{emp}} = 93.8311$$

$$p \approx 8.91264 \times 10^{-15}$$

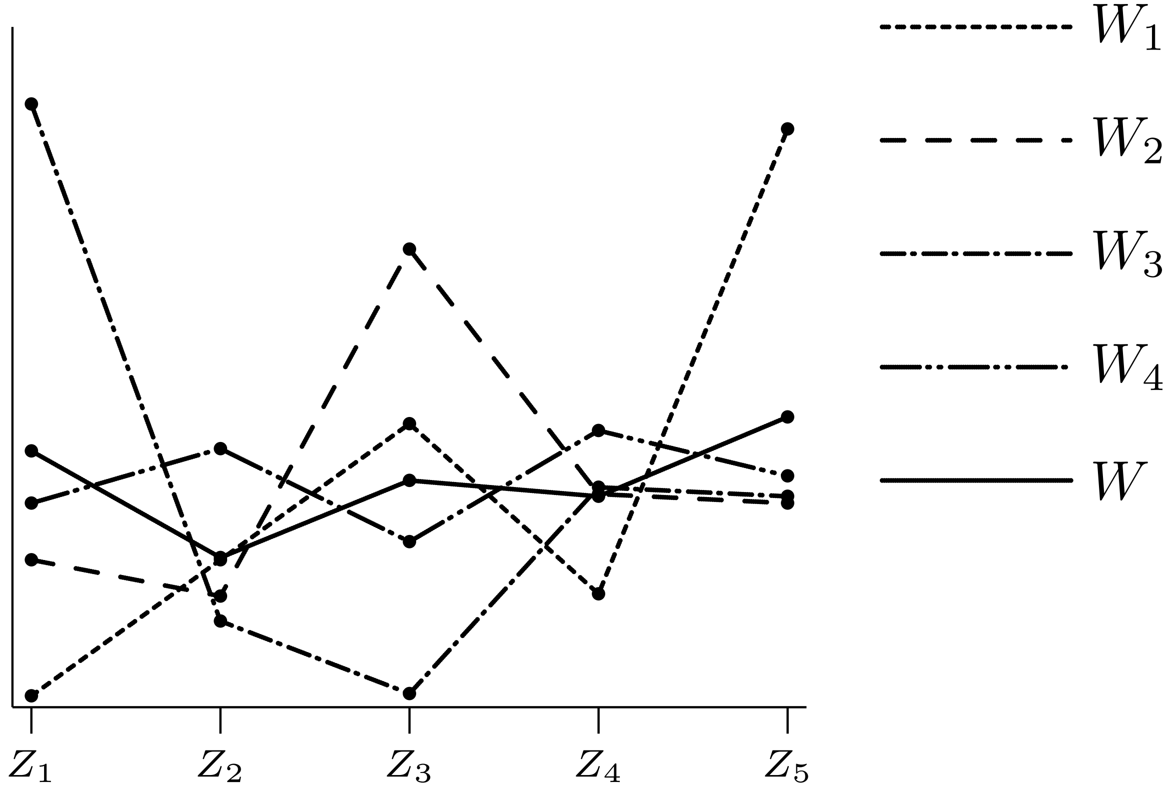
	W_1	W_2	W_3	W_4	
Z_1	$n_{11}^t=$ 42.5263	$n_{12}^t=$ 52.7326	$n_{13}^t=$ 54.0084	$n_{14}^t=$ 52.7326	
Z_2	$n_{21}^t=$ 33.2632	$n_{22}^t=$ 41.2463	$n_{23}^t=$ 42.2442	$n_{24}^t=$ 41.2463	
Z_3	$n_{31}^t=$ 40.0000	$n_{32}^t=$ 49.6000	$n_{33}^t=$ 50.8000	$n_{34}^t=$ 49.6000	
Z_4	$n_{41}^t=$ 38.5263	$n_{42}^t=$ 47.7726	$n_{43}^t=$ 48.9284	$n_{44}^t=$ 47.7726	
Z_5	$n_{51}^t=$ 45.6842	$n_{52}^t=$ 56.6484	$n_{53}^t=$ 58.0189	$n_{54}^t=$ 56.6484	

	W_1	W_2	W_3	W_4	
Z_1	**10.8964**	**2.6104**	**28.1501**	**0.6232**	
Z_2	**0.0021**	**0.4372**	**1.2423**	**3.3494**	
Z_3	**0.6250**	**13.0073**	**11.1504**	**0.8782**	
Z_4	**1.8870**	**0.0011**	**0.0235**	**1.0934**	
Z_5	**14.0287**	**1.6433**	**1.4020**	**0.7803**	

	W_1	W_2	W_3	W_4
(Z_1)	0.104	0.203	0.460	0.233
(Z_2)	0.209	0.234	0.222	0.335
(Z_3)	0.237	0.395	0.142	0.226
(Z_4)	0.164	0.262	0.273	0.301
(Z_5)	0.327	0.217	0.226	0.230
(Z)	0.211	0.261	0.267	0.261



	Z_1	Z_2	Z_3	Z_4	Z_5
(W_1)	0.105	0.165	0.225	0.15	0.355
(W_2)	0.165	0.149	0.302	0.194	0.190
(W_3)	0.366	0.138	0.106	0.197	0.193
(W_4)	0.190	0.214	0.173	0.222	0.202
(W)	0.213	0.166	0.200	0.193	0.228



Tablice kontygencji 2×2

	X_1	X_2	
Y_1	m_1	m_2	m
Y_2	$n_1 - m_1$	$n_2 - m_2$	$n - m$
	n_1	n_2	n

n, m, n_1 — stałe; m_1 — losowe

$$P\{m_1|n,m,n_1\} = \frac{\binom{m}{m_1}\binom{n-m}{n_1-m_1}}{\binom{n}{n_1}}$$

p–value

jeżeli $m_1 < \frac{m}{n}n_1$, to

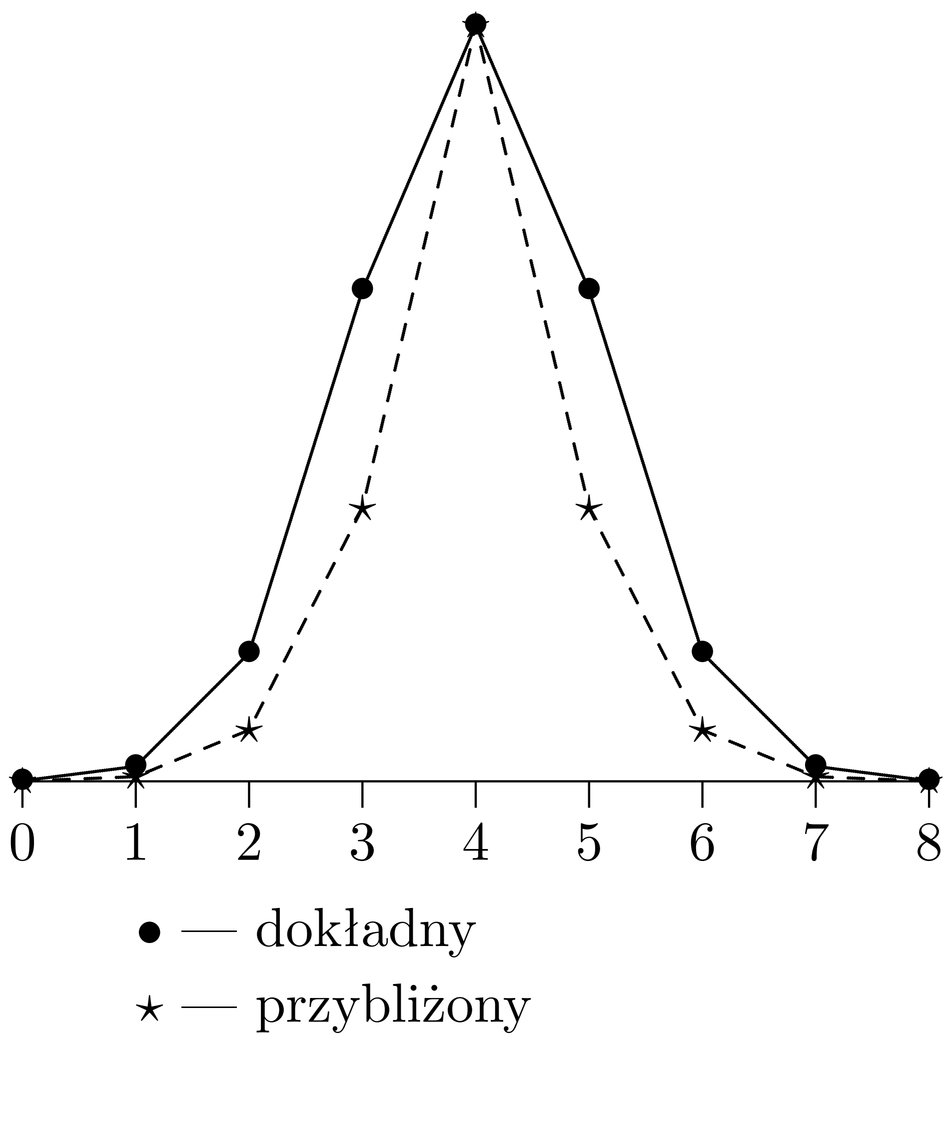
$$\sum_{k=0}^{m_1} P\{k|n,m,n_1\}$$

jeżeli $m_1 > \frac{m}{n}n_1$, to

$$\sum_{k=m_1}^{m \wedge n_1} P\{k|n,m,n_1\}$$

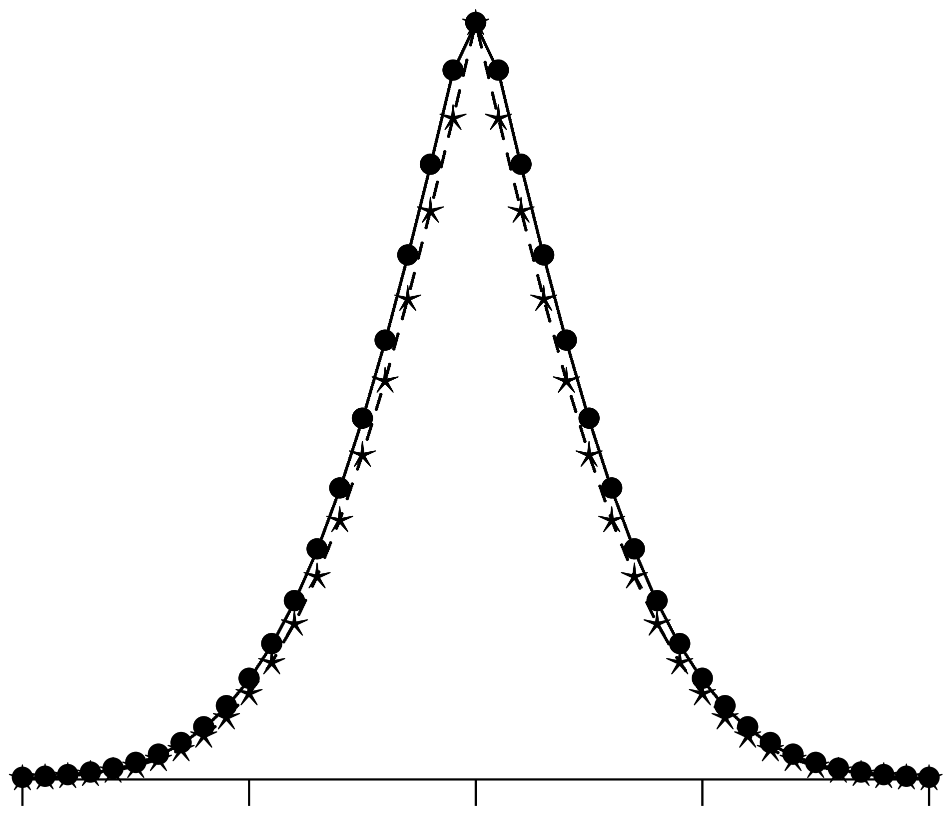
	X_1	X_2	
Y_1	M_1	$8 - M_1$	8
Y_2	$10 - M_1$	$2 + M_1$	12
	10	10	20

$$EM_1 = 4$$



	X_1	X_2	
Y_1	M_1	$800 - M_1$	800
Y_2	$500 - M_1$	$M_1 - 300$	200
	500	500	1000

$$EM_1 = 400$$



● — dokładny

⋆ — przybliżony

Przykład. Szczepionka pomaga?

Wymyślono nową szczepionkę przeciwko cholerze. Wybrano 86 osób spośród których 55 podano szczepionkę. Z badanej grupy osób 41 nie zachorowało, w tym były 32 osoby zaszczepione. Dokładniejsza analiza danych pozwoliła uzyskać informacje o warunkach bytowych badanych osób. Okazało się, że 51 osób żyło w dobrych warunkach. Spośród nich nie chorowało 45, w tym było 30 osób zaszczepionych. Czy szczepionka pomaga uniknąć ataku cholery?

⋆ ⋆ ⋆

S — podano szczepionkę;
s — nie podano szczepionki;

C — zachorowanie;
c — nie zachorowanie;

D — dobre warunki;
s — złe warunki;

	C	c	
S	23	32	55
s	22	9	31
	45	41	86

$$\frac{32 \cdot 22 - 23 \cdot 9}{32 \cdot 22 + 23 \cdot 9} = 0.54$$

		C	c	
D	S	15	30	45
	s	2	4	6
d	S	8	2	10
	s	20	5	25
		45	41	86

$$\frac{15 \cdot 4 - 2 \cdot 30}{15 \cdot 4 + 2 \cdot 30} = 0$$

$$\frac{8 \cdot 5 - 2 \cdot 20}{8 \cdot 5 + 2 \cdot 20} = 0$$

. . . .

Współczynnik korelacji

$$\varrho = \frac{E(X - EX)(Y - EY)}{DX \cdot DY}$$

Własności współczynnika korelacji

1. Współczynnik korelacji jest liczbą niemianowaną

2. $\varrho \in \langle -1, 1 \rangle$

3. Jeżeli $\varrho > 0$, to większym wartościom jednej cechy odpowiadają (średnio) większe wartości drugiej cechy. Zależność dodatnia (rosnąca, stymulująca).

4. Jeżeli $\varrho < 0$, to większym wartościom jednej cechy odpowiadają (średnio) mniejsze wartości drugiej cechy. Zależność ujemna (malejąca, limitująca).

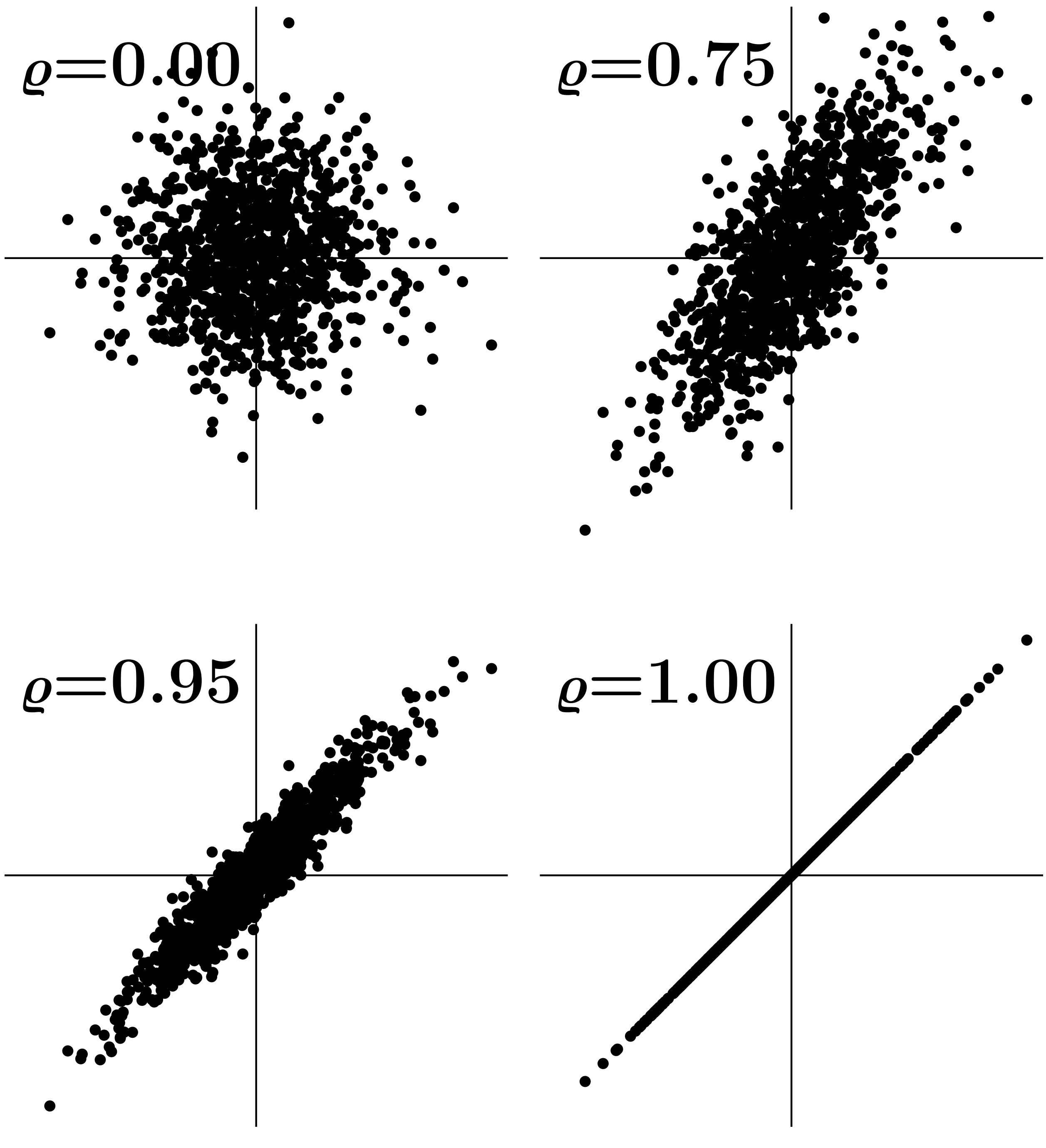
5. Jeżeli $\varrho = 0$, to bez względu na wartości przyjmowane przez jedną z cech, średnie wartości drugiej cechy są takie same. Cechy nieskorelowane.

6. Jeżeli $\varrho = \pm 1$, to istnieją takie liczby rzeczywiste a oraz b, że $Y = aX + b$. Jeżeli $\varrho = 1$, to $a > 0$. Jeżeli $\varrho = -1$, to $a < 0$.

Współczynnik korelacji jest miernikiem **liniowej** zależności między cechami X oraz Y.

Im $|\varrho|$ jest bliższe 1, tym bardziej „liniowa" jest zależność między cechami.

7. Jeżeli (X, Y) ma dwuwymiarowy rozkład normalny, to $\varrho = 0$ jest równoważne **niezależności** cech X, Y.


ϱ=0.00
ϱ=0.75
ϱ=0.95
ϱ=1.00

	1	2	3	4	5	6
1	0	0	1/12	1/12	0	0
2	0	1/12	0	0	1/12	0
3	1/12	0	0	0	0	1/12
4	1/12	0	0	0	0	1/12
5	0	1/12	0	0	1/12	0
6	0	0	1/12	1/12	0	0

$$EX = 3.5 \qquad EY = 3.5$$

$$DX = 1.7078 \qquad DY = 1.7078$$

$$E(X - EX)(Y - EY) = 0$$

$$P\{X = 1 \ \& \ Y = 2\} \neq P\{X = 1\}P\{Y = 2\}$$

Niech $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ będzie próbą

Współczynnik korelacji z próby (próbkowy)

$$R = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var} X}\sqrt{\mathrm{var} Y}}$$

Współczynnik korelacji Pearsona

Suma iloczynów odchyleń

$$\mathrm{cov}(X, Y) = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Kowariancja z próby

$$\mathrm{C}(X, Y) = \frac{\mathrm{cov}(X, Y)}{n - 1}$$

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \right)$$

Jeżeli $\varrho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = 0$, to

$$\frac{R}{\sqrt{1 - R^2}} \sqrt{n-2} \sim t_{n-2}$$

p–value:
jeżeli R_{emp} jest wartością próbkowego współczynnika korelacji oraz $\varrho = 0$, to

$$p = P \left\{ t_{n-2} > \frac{|R_{\mathrm{emp}}|}{\sqrt{1 - R_{\mathrm{emp}}^2}} \sqrt{n-2} \right\}$$

Ilościowy opis zależności

Ilościowy opis zależności Y od X:

$$E(Y|X = x) = f(x)$$

Funkcja f nosi nazwę **funkcji regresji**

Rozkład warunkowy Y dla ustalonego $X = x$:

$$Y|X = x \sim N\left(E(Y|X = x), D^2(Y|X = x)\right)$$

$$E(Y|X = x) = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x - \mu_X)$$

$$D^2(Y|X = x) = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$$

$$E(Y|X=x)=\left(\frac{\sigma_{XY}}{\sigma_X^2}\right)\cdot x+\left(\mu_Y-\frac{\sigma_{XY}}{\sigma_X^2}\mu_X\right)$$

Czyli $f(x)=ax+b$

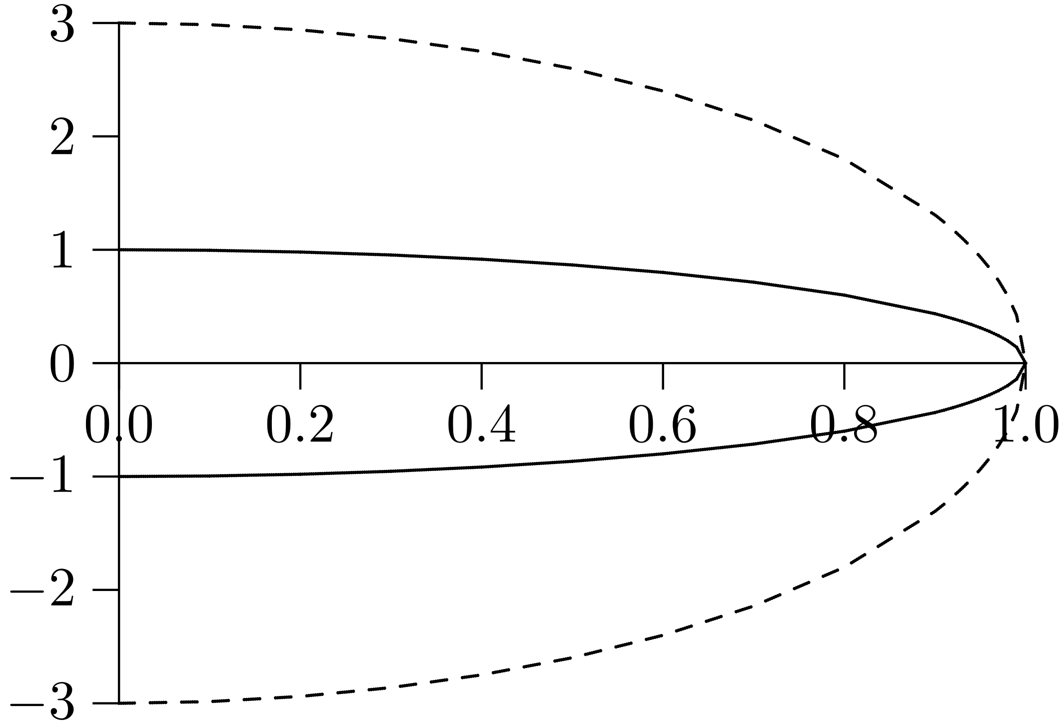
$$a=\frac{\sigma_{XY}}{\sigma_X^2}\qquad b=\mu_Y-a\mu_X$$

Zróżnicowanie Y wokół funkcji regresji

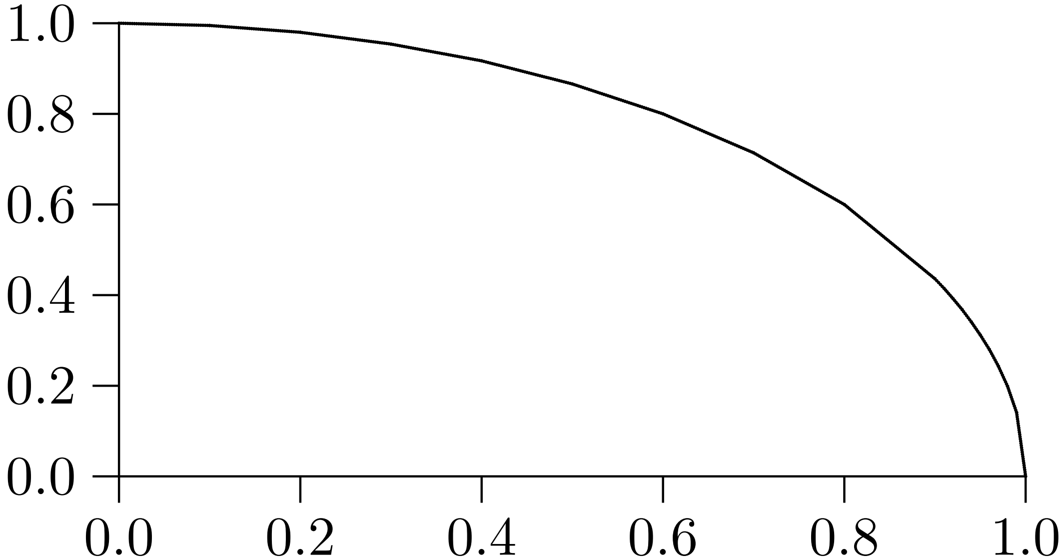
$$D^2(Y|X=x)=\sigma_Y^2(1-\varrho^2)$$

Jeżeli $q\in(0,1)$, to

$$P\left\{|Y-E(Y|X=x)|<u_{1-\frac{q}{2}}\sigma_Y\sqrt{1-\varrho^2}\right\}=q$$



3
2
1
0
−1
−2
−3
0.0
0.2
0.4
0.6
0.8
1.0



1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Ocena parametrów funkcji regresji

Próba $(X_1, Y_1), \ldots, (X_n, Y_n)$

Niezbędne rachunki

$$\bar{X},\ \mathrm{var}X,\ \bar{Y},\ \mathrm{var}Y,\ \mathrm{cov}(X,Y)$$

$$\widehat{a} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}X} \qquad \widehat{b} = \bar{Y} - \widehat{a}\bar{X}$$

Resztowa suma kwadratów

$$\mathrm{var}R = \mathrm{var}Y(1 - R^2)$$

Wariancja resztowa

$$S^2 = \frac{\mathrm{var}R}{n-2}$$

Przykład. W pewnej rodzinie obserwowano tygodniowe wydatki na używki (Uż) i artykuły spożywcze (Sp). Na podstawie poniższych danych zbadać istnienie zależności. Jeżeli taka zależność istnieje, to opisać ją ilościowo.

Uż	Sp	Uż	Sp	Uż	Sp	Uż	Sp	Uż	Sp
28.50	45.54	26.55	44.35	28.37	44.00	38.31	42.92	22.78	45.03
23.61	45.33	21.55	45.71	28.15	44.46	21.94	46.50	25.76	45.29
31.22	43.31	20.77	46.01	36.71	43.36	32.69	43.50	22.39	45.16
36.38	42.33	25.11	46.12	29.57	44.39	34.51	43.82	28.19	44.76
35.99	42.40	26.13	43.82	29.07	45.05	39.59	41.77	29.84	44.01
38.67	42.31	19.41	46.10	27.43	44.11	29.58	44.29	30.14	43.91
19.08	46.28	27.16	45.52	39.86	41.98	27.38	44.74	28.39	44.29
28.83	43.39	27.98	44.59	34.33	43.34	33.38	43.01	40.97	42.14
35.48	42.68	30.67	44.01	41.88	41.98	28.09	44.40	21.29	46.61
24.57	45.10	28.17	43.91	26.73	44.20	33.79	43.26	26.32	44.92

X: tygodniowe wydatki na używki
Y: tygodniowe wydatki na artykuły spożywcze

Współczynnik korelacji

$$R = -0.9255$$

$$p - \text{value} \approx 7.11268 \times 10^{-22}$$

Funkcja regresji

$$E(Y|X = x) = ax + b$$

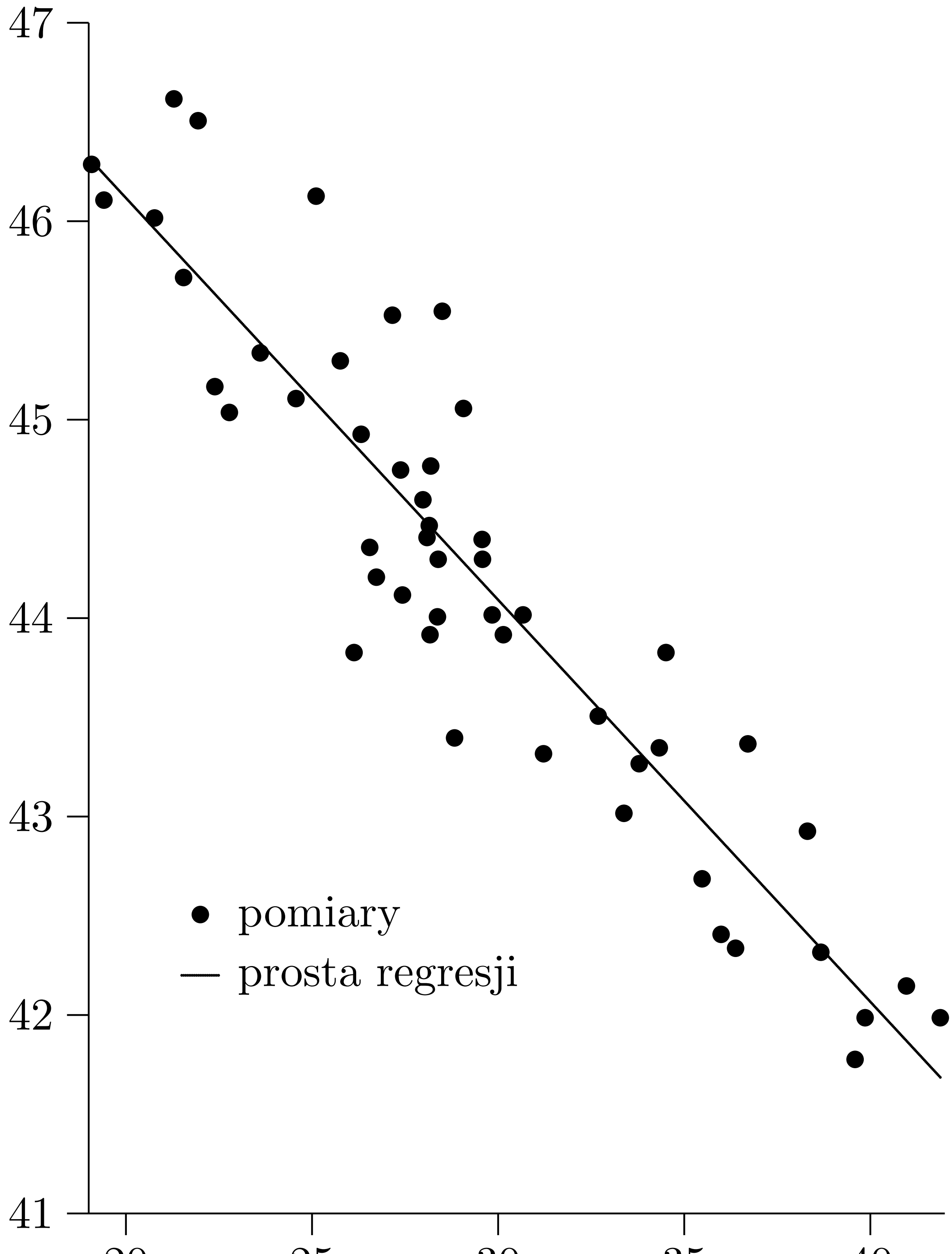
$$\widehat{a} = -0.2025 \qquad \widehat{b} = 50.1680$$

Zależność między średnimi wydatkami na spożywcze a wydatkami na używki opisana jest wzorem

$$\text{średni } y = 50.1680 - 0.2025x$$

Wariancja resztowa

$$s^2 = 0.2374$$


47
46
45
44
43
42
41
20
25
30
35
40
pomiary
prosta regresji

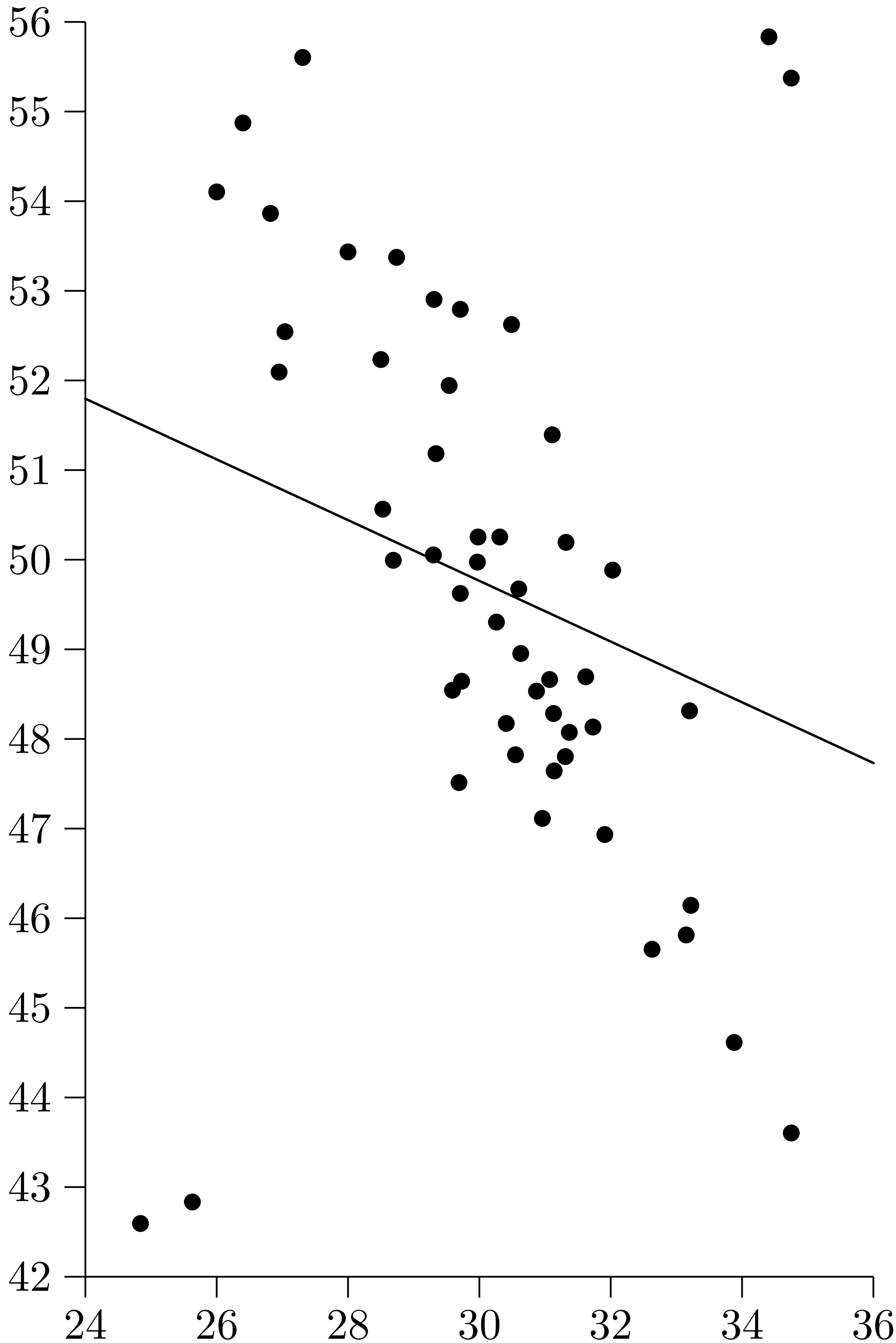
Przykład. podobny

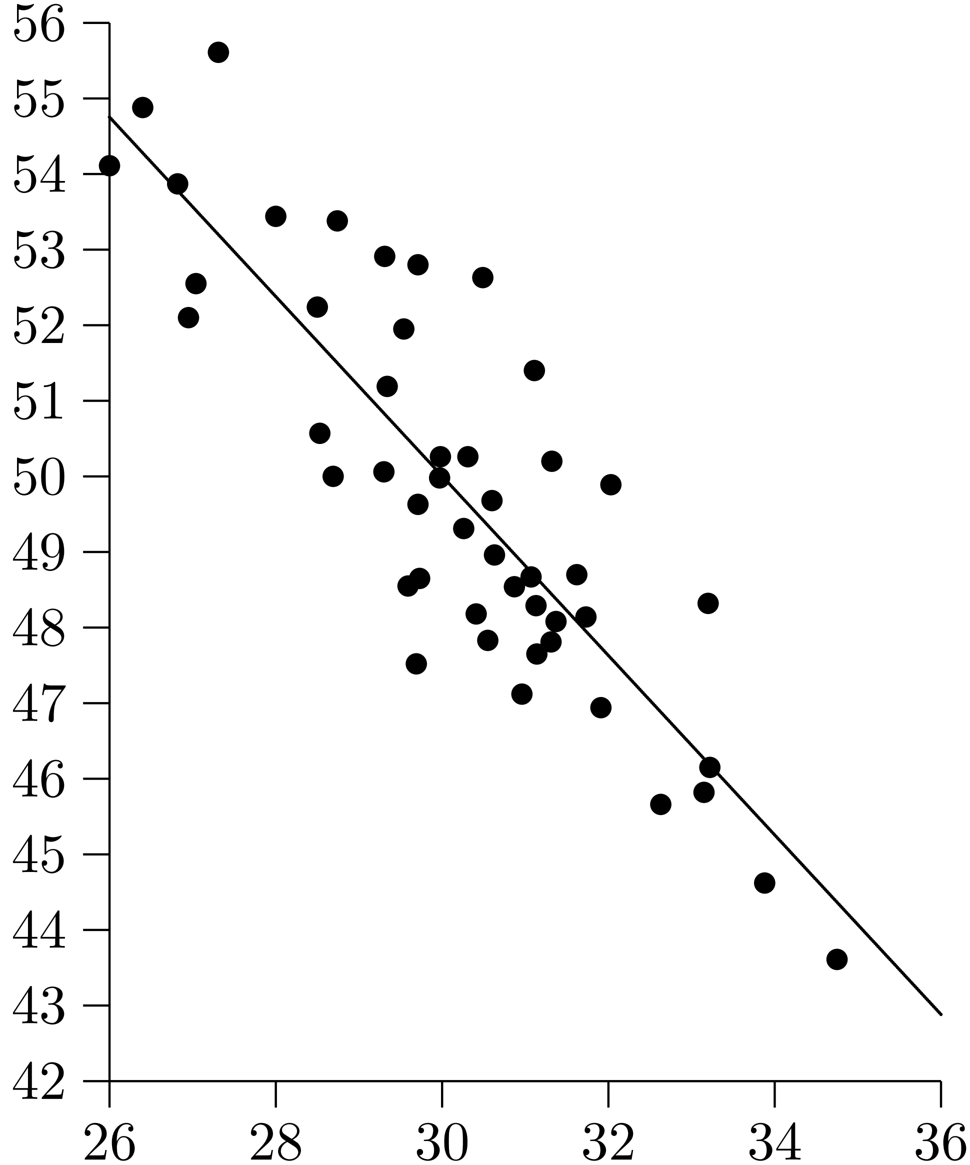
Nr	Uż	Sp	Nr	Uż	Sp	Nr	Uż	Sp	Nr	Uż	Sp
1	29.59	48.53	14	26.95	52.08	27	30.49	52.61	40	30.63	48.94
2	30.96	47.10	15	34.41	55.82	28	27.31	55.59	41	27.04	52.53
3	29.97	49.96	16	29.34	51.17	29	29.71	52.78	42	28.50	52.22
4	29.69	47.50	17	32.63	45.64	30	30.87	48.52	43	26.00	54.09
5	31.91	46.92	18	31.37	48.06	31	24.84	42.58	44	29.98	50.24
6	29.54	51.93	19	30.31	50.24	32	25.63	42.82	45	31.73	48.12
7	30.60	49.66	20	34.75	43.59	33	28.53	50.55	46	26.82	53.85
8	31.11	51.38	21	28.00	53.42	34	31.31	47.79	47	31.07	48.65
9	31.13	48.27	22	33.88	44.60	35	31.14	47.63	48	29.71	49.61
10	31.62	48.68	23	32.03	49.87	36	30.41	48.16	49	31.32	50.18
11	30.55	47.81	24	26.40	54.86	37	33.22	46.13	50	33.15	45.80
12	29.73	48.63	25	28.69	49.98	38	33.20	48.30	51	34.75	55.36
13	30.26	49.29	26	29.30	50.04	39	28.74	53.36	52	29.31	52.89

$$R = -0.2447 \qquad p - \text{value} \approx 0.0867$$

$$y = -0.3387x + 59.925$$

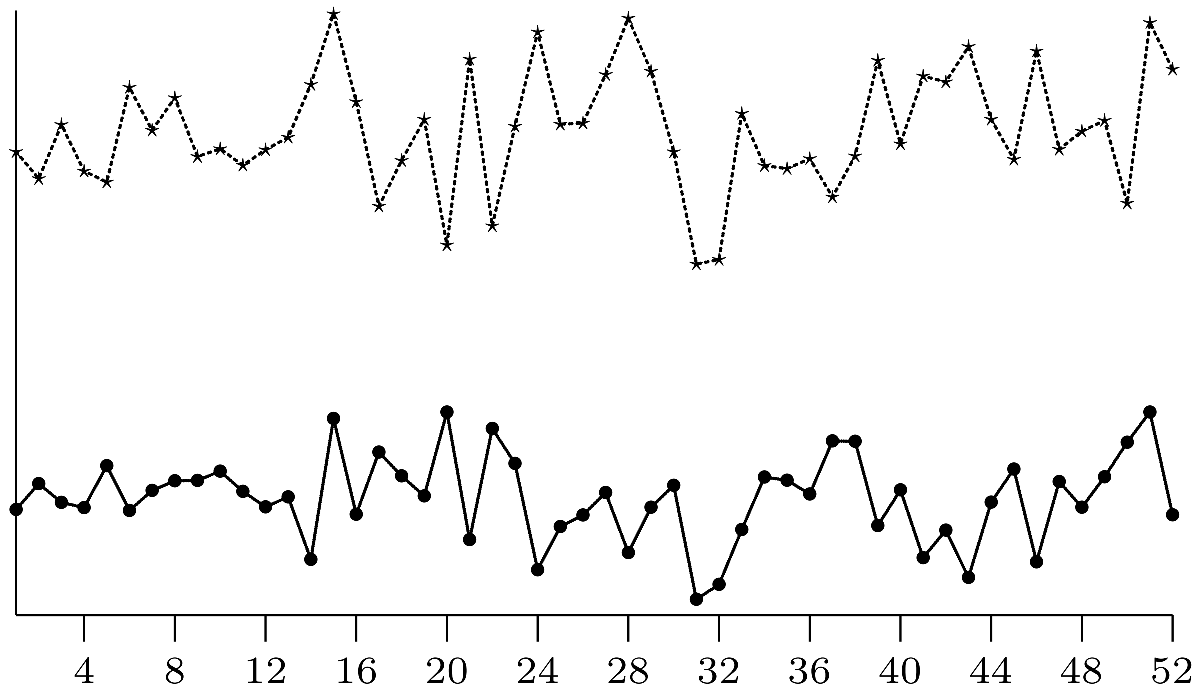
$$\bar{y} = 49.70$$

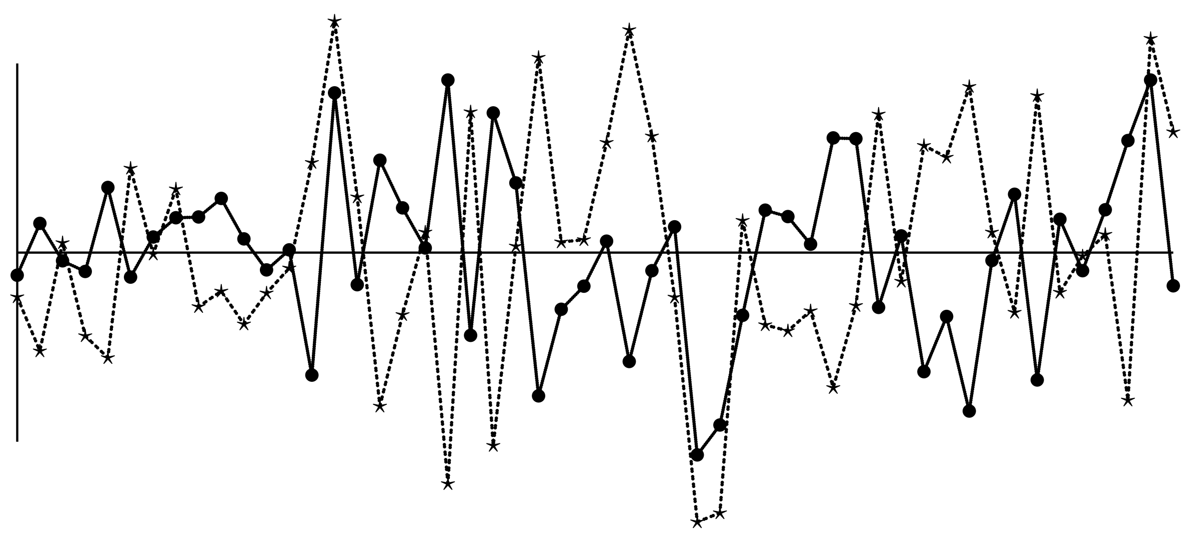

56
55
54
53
52
51
50
49
48
47
46
45
44
43
42
24
26
28
30
32
34
36



$R = -0.8565 \qquad p - \text{value} \approx 2.17 \times 10^{-15}$

$$y = -1.1871x + 85.618$$


4
8
12
16
20
24
28
32
36
40
44
48
52



(X, Y) ma dwuwymiarowy rozkład ciągły

Współczynnik korelacji rangowej Spearmana

Współczynnik korelacji rangowej Kendalla

Rangi

Próba:	1.1	1.2	0.8	0.9	1.5	1.3	1.0	0.7	0.6	1.6
Rangi:	6	7	3	4	9	8	5	2	1	10

Współczynnik korelacji rangowej Spearmana

Obserwacje: $(X_i, Y_i), i = 1, \ldots, n$

Obserwacjom X_i nadajemy rangę R_i

Obserwacjom Y_i nadajemy rangę Q_i

Otrzymujemy pary liczb naturalnych (R_i, Q_i)

$$R = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$

Dla dużych n zmienna losowa R ma w przybliżeniu rozkład

$$N\left(0, \sqrt{\frac{1}{n-1}}\right)$$

Współczynnik korelacji rangowej Kendalla

Obserwacje: $(X_i, Y_i), i = 1, \ldots, n$.

Pary porządkujemy według wzrastających wartości X-ów:

$$(X_{1:n}, Y_1^*), \ldots, (X_{n:n}, Y_n^*).$$

Niech s_i będzie liczbą tych par $(X_{j:n}, Y_i^*), j > i$, w których $Y_j^* > Y_i^*$.

$$T = \frac{4\sum_{i=1}^n s_i}{n(n-1)} - 1$$

Dla dużych n zmienna losowa T ma w przybliżeniu rozkład

$$N\left(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)$$

Przykład.
X — wyniki pierwszego testu inteligencji
Y — wyniki drugiego testu inteligencji

Czy X oraz Y są niezależne?

Obserwacje:

$$\begin{array}{cccc}
(502, 564) & (678, 787) & (727, 851) & (724, 767) \\
(576, 722) & (527, 585) & (705, 739) & (737, 865) \\
(999, 901) & (955, 922) & (529, 444) & (603, 492) \\
(825, 951) & (504, 616) & (646, 635) & (663, 574) \\
(930, 789) & (714, 768) & (858, 809) & (582, 573)
\end{array}$$

Współczynnik korelacji rangowej Spearmana

Współczynnik korelacji rangowej Kendalla

X	Y	R_i	Q_i	$(R_i - Q_i)^2$
502	564	1	3	4
678	787	10	13	9
727	851	14	16	4
724	767	13	11	4
930	789	18	14	16
576	722	5	9	16
527	585	3	6	9
705	739	11	10	1
737	865	15	17	4
714	768	12	12	0
999	901	20	18	4
955	922	19	19	0
529	444	4	1	9
603	492	7	2	25
858	809	17	15	4
825	951	16	20	16
504	616	2	7	25
646	635	8	8	0
663	574	9	5	16
582	573	6	4	4
	$\sum_{i=1}^{20} (R_i - Q_i)^2 =$			170

$$R = 1 - \frac{6}{20(400-1)} 170 = 0.8722$$

$$p - \text{value} \approx 0.000144$$

			i																			
i	$X_{(i)}$	Y_i^*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	502	564																				
2	504	616	1																			
3	527	585	1	0																		
4	529	444	0	0	0																	
5	576	722	1	1	1	1																
6	582	573	1	0	0	1	0															
7	603	492	0	0	0	1	0	0														
8	646	635	1	1	1	1	0	1	1													
9	663	574	1	0	0	1	0	1	1	0												
10	678	787	1	1	1	1	1	1	1	1	1											
11	705	739	1	1	1	1	1	1	1	1	1	0										
12	714	768	1	1	1	1	1	1	1	1	1	0	1									
13	724	767	1	1	1	1	1	1	1	1	1	0	1	0								
14	727	851	1	1	1	1	1	1	1	1	1	1	1	1	1							
15	737	865	1	1	1	1	1	1	1	1	1	1	1	1	1	1						
16	825	951	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1					
17	858	809	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0				
18	930	789	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0			
19	955	922	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1		
20	999	901	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	
		s_i:	17	13	13	16	11	13	13	11	11	7	9	7	7	4	3	0	2	2	0	0

$$T = \frac{4 \cdot 159}{20(20-1)} - 1 = 0.6736$$

$$p - \text{value} \approx 0.000033$$

. . . .

Test serii

$$AA\ BBBBBB\ AAAAA\ B$$

Ilość serii: 4

Jeżeli $P(A) = p$, to $(n_1 + n_2 = n)$

$$P\{\text{ilość serii } = r\} =$$

$$\begin{cases} p^n + (1-p)^n, & r = 1 \\ 2\displaystyle\sum_{n_1=k}^{n-k} \binom{n_1-1}{k-1}\binom{n_2-1}{k-1} p^{n_1}(1-p)^{n_2}, & r = 2k \\ \displaystyle\sum_{n_1=k}^{n-k} \left(\binom{n_1-1}{k-1}\binom{n_2-1}{k} + \right. & \\ \qquad \left. \binom{n_1-1}{k}\binom{n_2-1}{k-1} \right) p^{n_1}(1-p)^{n_2}, & r = 2k+1 \end{cases}$$

Niech $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ będzie próbą

Niech $p \in (0, 1)$

Niech $\hat{Y}_p$ będzie p–tym kwantylem próbkowym

Pary porządkujemy według wzrastających wartości X-ów:

$$(X_{1:n}, Y_1^*), \ldots, (X_{n:n}, Y_n^*).$$

Z próby $Y_1^*, \ldots, Y_n^*$ tworzymy ciąg liter A oraz B:
jeżeli $Y_i^* > \hat{Y}_p$, to piszemy A
jeżeli $Y_i^* < \hat{Y}_p$, to piszemy B

Niech R_p będzie liczbą serii
p–value obliczamy z podanego wzoru

W zastosowaniach wyznacza się $R_{0.5}$

Przykład.

X — wyniki pierwszego testu inteligencji

Y — wyniki drugiego testu inteligencji

Czy X oraz Y są niezależne?

$X_{(i)}$	Y_i^*	
502	564	A
504	616	A
527	585	A
529	444	A
576	722	A
582	573	A
603	492	A
646	635	A
663	574	A
678	787	B
705	739	A
714	768	B
724	767	B
727	851	B
737	865	B
825	951	B
858	809	B
930	789	B
955	922	B
999	901	B

Mediana: 753

$R_{0.5} = 4$

$n = 20$

$P\{R_{0.5} = 4\} = 0.00185$

. . . .