WNE FiR 2010

STATYSTYKA

Robert Pietrzykowski

Prawa naukowe nie są formułowane na mocy autorytetów ani uzasadniane przez wiarę czy średniowieczną filozofię. Jedynym sądem odwoławczym dla nowej wiedzy jest statystyka

P.C. Mahanalobis

Wykład IV

Badanie zależności cech

Na dziś…

- Sprawy bieżące
- Inne

Zmienne losowe X i Y są niezależne wtedy i tylko wtedy gdy $\forall (x, y) \in R^2$:

$$F_{XY}(x, y) = F_X(x) F_Y(y).$$

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

Współczynnik korelacji

Opis jakościowy zależności

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

	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

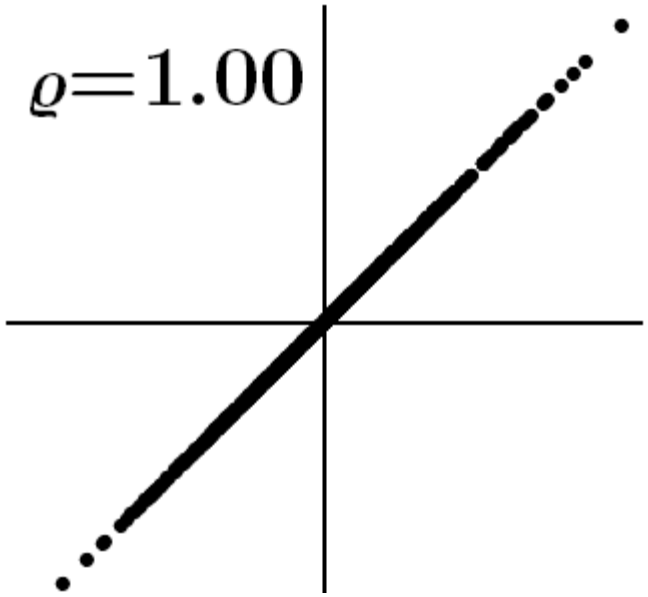
$$P\{X = 1 \ \& \ Y = 2\} \neq P\{X = 1\}P\{Y = 2\}$$

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

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

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

Współczynnik korelacji jest miernikiem zależności między dwiema cechami
Oznaczenie: ϱ



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

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

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

współczynnik korelacji Pearsona

współczynnik korelacji rangowej Spearmana
współczynnik korelacji rangowej Kendalla

test chi–kwadrat niezależności

H_0 : Cechy X oraz Y są niezależne

Współczynnik korelacji Pearsona

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

$$H_0 : \varrho = 0$$

$$H_1 : \varrho \neq 0$$

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

Hipoteza H_0 jest odrzucana jeżeli $|R| \geq r_{(\alpha,n)}$

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

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

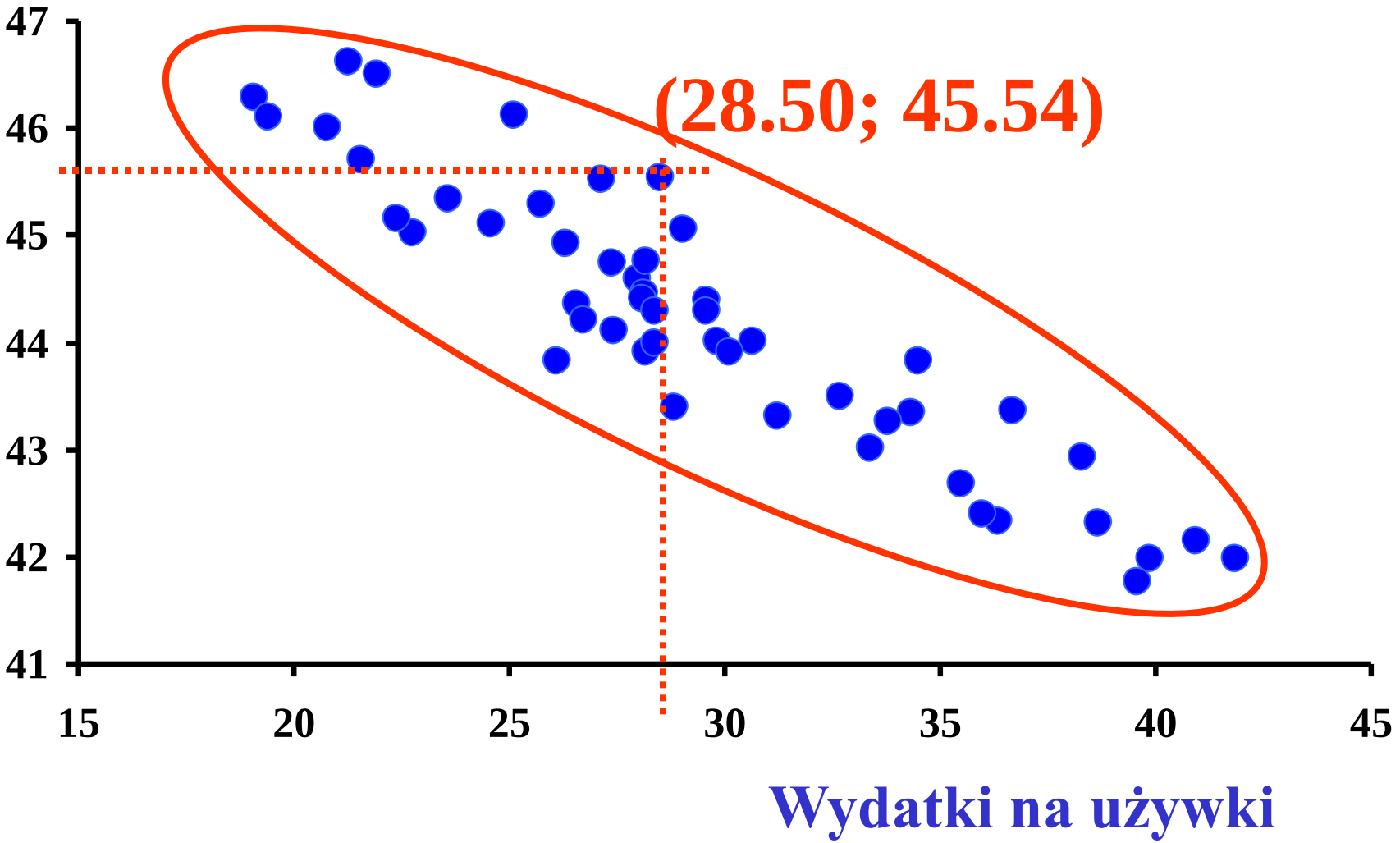
$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

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.

(28.50; 45.54)

Uż	Sp	Uż	Sp	Uż	Sp	Uż	Sp	Uż	Sp
			4.35	28.37	44.00	38.31	42.92	22.78	45.03
			5.71	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

Wydatki na art. spożywcze



Cechy:

X: tygodniowe wydatki na używki

Y: tygodniowe wydatki na artykuły spożywcze

Założenie:

normalność rozkładów badanych cech

Techniki statystyczne:

1. Weryfikacja hipotezy $H_0 : \varrho = 0$

Weryfikacja hipotezy i wnioskowanie:

Wniosek.

Wydatki na używki (X) i wydatki na artykuły spożywcze (Y) są od siebie zależne. Ponieważ współczynnik korelacji jest ujemny, więc zależność ma charakter malejący, tzn. im większe są wydatki na używki, tym mniejsze (średnio) na artykuły spożywcze.

Test współczynnika korelacji 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)

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

Jeżeli $|r_{\text{emp}}| > r_{\alpha/2,n}$, to hipotezę H_0 : *Cechy X oraz Y są niezależne* odrzucamy. Liczby $r_{\alpha,n}$ są stablicowanymi dwustronnymi wartościami krytycznymi rozkładu współczynnika korelacji Spearmana.

Test chi–kwadrat niezależności

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

$$n_{ij}^t = \frac{n_{i.} n_{.j}}{N}, \quad N = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij}$$

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

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

Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; (k-1)(m-1))$,
to hipotezę H_0 odrzucamy

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)	≤ 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)	≥ 2000	71	47	49	50

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

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

Do zobaczenia na następnym wykładzie !!!