Statystyka Opisowa z Demografią oraz Biostatystyka

Analiza Współzależności

Aleksander Denisiuk

denisjuk@euh-e.edu.pl

Elbląska Uczelnia Humanistyczno-Ekonomiczna

ul. Lotnicza 2

82-300 Elbląg

Analiza Współzależności

Najnowsza wersja tego dokumentu dostępna jest pod adresem

http://denisjuk.euh-e.edu.pl/

Współzależność statystyczna

- związki między zjawiskami
- duża liczba obserwacji
- wykrywanie związków
- analiza statystyczna poprzedzana jest analizą logiczną związków
- pary cech, każda jednostka ststystyczna scharakteryzowana jednocześnie dwoma cechami

Metody analizy współzależności

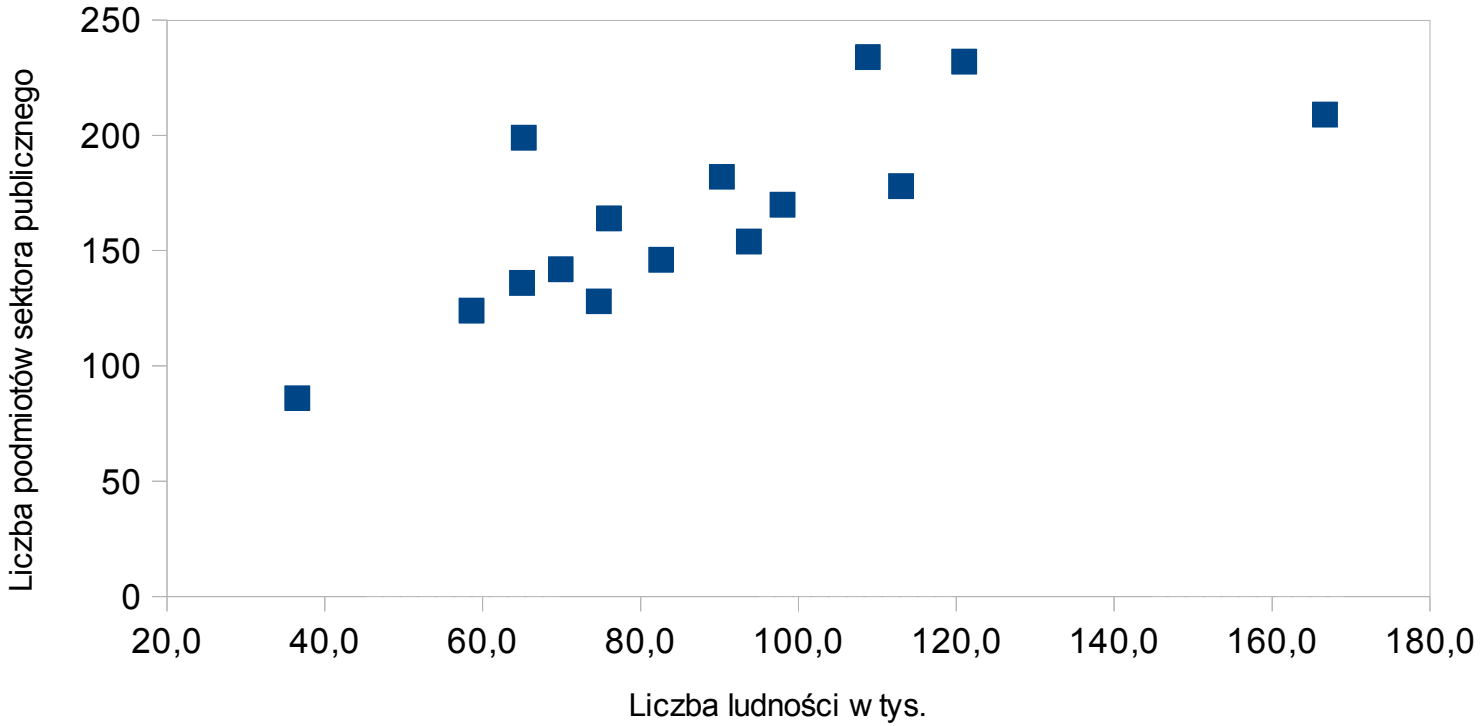
- *korelacja* — wzajemne oddziaływanie
- *regresja* — wpływ jednej cechy (przyczyny) na drugą (skutek)
- forma opisu
  - tabelaryczna (czeregi lub tablice)
  - graficzna (diagram korelacji)
  - parametryczna (liczbowa)
- zależy od rodzaju cech (ilościowe, jakościowe)

Ocena rodzaju związku

- *korelacja dodatnia*
- *korelacja ujemna*
- poszczególne obserwacje mogą odbiegać od stwierdzonej prawidłowości

Przykład (województwo Pomorskie)

powiat	liczba ludności	liczba podmiotów sektora publicznego
bytowski	76,0	164
chojnicki	90,3	182
człuchowski	58,6	124
kartuski	98,0	170
kościerski	65,0	136
kwidzyński	82,6	146
lęborski	65,2	199
malborski	108,8	234
nowodworski	36,5	86
gdański	74,7	128
pucki	69,9	142
słupski	93,7	154
starogardski	121,0	232
tczewski	113,4	178
wejherowski	166,7	209



- związek korelacyjny dodatni
- zbliżony do liniowego
- umiarkowana siła współzależności

Miary korelacji

Miara	Cechy	Związek	Prezentacja
współczynnik korelacji linowej Pearsona	obie ilościowe	liniowy	szeregi, tablice
współczynnik korelacji rang Spearmana	obie ilościowe lub jakościowe w skali porządkowej	liniowy	szeregi
współczynnik zbieżności T Czuprowa	obie ilościowe lub jakościowa i ilościowa	nie ma znaczenia	tablice

Współczynnik korelacji liniowej Pearsona

- pokazuje siłę i kierunek współzależnośi
- $$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n\cdot s_x\cdot s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} =$$
$$\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\left(n\cdot\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right)\cdot\left(n\cdot\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\right)}}$$

Determinacja i indeterminacja

- r^2 — *współczynnik determinacji*
  - pokazuje jaka część zmienności cechy jest wyjaśnona ukształtowaniem się drugiej cechy
- $\phi^2 = 1 - r^2$ — *współczynnik indeterminacji*
  - pokazuje jaka część zmienności cechy nie może być wyjaśnona ukształtowaniem się drugiej cechy

Współczynnik Pearsona. Przykład

- ceny działek budowlanych w zależności od odległości od centrum miasta

odległość (km)	0	1	2	3	4	5	6
cena (zł/m^2)	1000	900	500	500	270	300	100

- $r = -0{,}959$
- $r^2 = 0{,}920$

Współczynnik korelacji rang Spearmana

- wartości cech zastępujemy rangami (od 1 do n)
  - jeżeli kilka cech mają te same rangi, zastępujemy liczbą średnią
- zgodność uporządkować cech
- przyjmuje wartości w przedziale $[-1, 1]$
- znak $+$ oznacza uporządkowanie zgodne
- znak $-$ oznacza uporządkowanie przeciwne
- im bliższy do ± 1, tym związek jest silniejszy
- $r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, gdzie d_i jest różnica między rangami

Współczynnik Spearmana. Przykład

- 1 tura wyborów 1995 roku

kandydat	odsetek głosów	
	miast	wsi
A. Kwaśniewski	31,4	33,4
L. Wałęsa	33,4	33,4
J. Kuroń	13,3	6,3
J. Olszewski	7,0	7,2
W. Pawlak	0,5	9,5
T. Zieliński	4,6	2,5
H. Gronkiewicz-Waltz	3,5	2,7
J. Korwin-Mikke	4,3	1,4

- $r_s = 0{,}542$

Współczynnik zbieżności T Czuprowa

- dowolne dwie cechy, przedstawione w postaci tablicy (zazwyczaj jakościowe)
- przyjmuje wartości w przedziale $[0, 1]$
- im bliżej 1, tym silniejsza jest współzależność
- tablica korelacyjna
  - k wierszy, dotyczących wariantów cechy X
  - l kolumn, dotyczących wariantów cechy Y
  - n_{ij} jest liczbą jednostek, posiadających i-ty wariant cechy X i j-ty wariant cechy Y

Obliczenie współczynnika T Czuprowa

1. dla każdego pola oblicza się liczebności teoretyczne $\hat{n}_{ij} = \frac{n_i \cdot n_j}{n}$, gdzie n_i jest sumą wiersza, n_j — sumą kolumny
2. statystyka chi-kwadrat $\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_i \sum_j \frac{n_{ij}^2}{\hat{n}_{ij}} - n$
3. współczynnik zbieżności T Czuprowa $T = \sqrt{\frac{\chi^2}{n\sqrt{(k-1)(l-1)}}}$,

   gdzie k jest ilością wierszy, l — kolumn

Współczynnik T Czuprowa. Przykład

- czy opinia o pracy służby zdrowia zależy do płci?

opinia	M	K
negatywna	120	240
pozytywna	280	360

- $T = 0{,}10$ (nie zależy)

Model regresji liniowej

- funkcja regresji opisuje wpływ, jaki wywiera *przyczyna* na *skutku*: $\hat{y}_i = f(x_i)$
- funkcja regresji jest funkcją matematyczną określonego typu, która jest przybliżeniem faktycznej zależnośći
- postać funkcji ustala się na podstawie zaobserowanych wartośći (x_i, y_i)
  - zaobserwowane wartości będą się odchylały od funkcji po wpływem cech nie uwzględnionych w badaniu oraz na skutek czynników losowych: $y_i = \hat{y}_i + e_i$
  - e_i nazywają się *resztami*

Model regresji liniowej, cd

- funkcja regresji może przybrać postać liniową lub krzywoliniową
- linowa postać oznacza, że jednakowym przyrostom zmiennej niezależnej odpowiadają jednakowe co do siły i kierunku zmiany zmiennej zależnej
- regresja krzywoliniowa (kwadratowa, sześcienna, wykładnicza, etc) oznacza, że jednakowym przyrostom zmiennej niezależnej odpowiadają różne co do siły i kierunku zmiany zmiennej zależnej
- postać określa się na podstawie wykresu korelacyjnego oraz względów merytorycznych

Metoda najmniejszych kwadratów

- parametry odpowiedzniej funkcji regresja określa się za pomocą metody najmniejszych kwadratów:
  - suma kwadratów odchyleń odległości zaobserowanych wartości y_i od wartości teoretycznych $\hat{y}_i$ jest najmniejsza
  - $\sum_i (y_i - \hat{y}_i)^2 \to \min$

Obliczanie parametrów regresji liniowej

- $\hat{y} = a + bx$
- b jest *współczynnikiem regresji*
- wartości a i b obliczane na posdtawie wyników obserwacji:

  $x_1, \quad x_2, \quad x_3, \quad \ldots, \quad x_n$

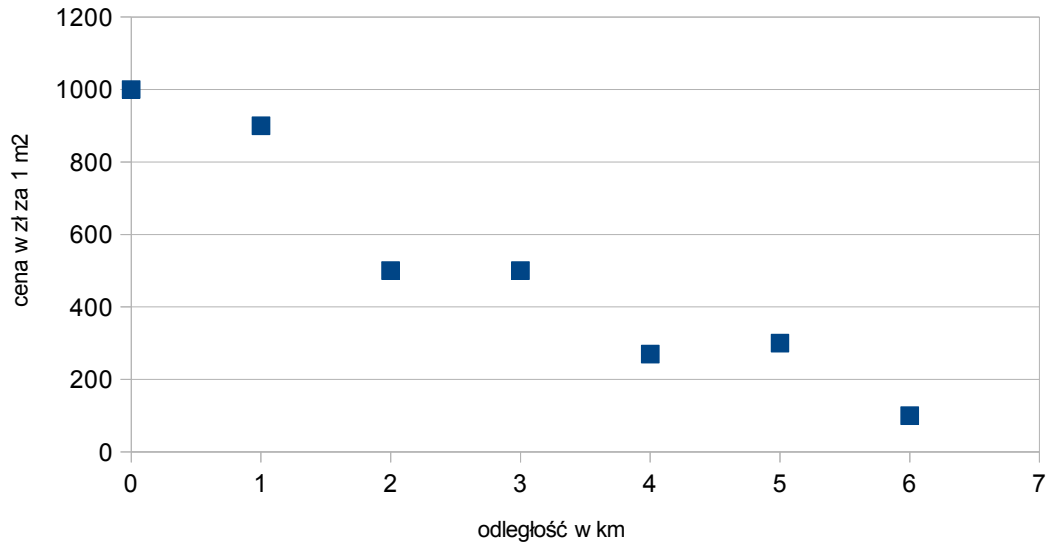
  $y_1, \quad y_2, \quad y_3, \quad \ldots, \quad y_n$

- $b = \frac{n\sum x_i y_i - \sum x_i \sum y_j}{n\sum x_i^2 - (\sum x_i)^2}$
- $a = \frac{\sum y_i - b\sum x_i}{n}$

Regresja liniowa. Przykład

- ceny działek budowlanych w zależności od odległości od centrum miasta

odległość (km)	0	1	2	3	4	5	6
cena (zł/m^2)	1000	900	500	500	270	300	100



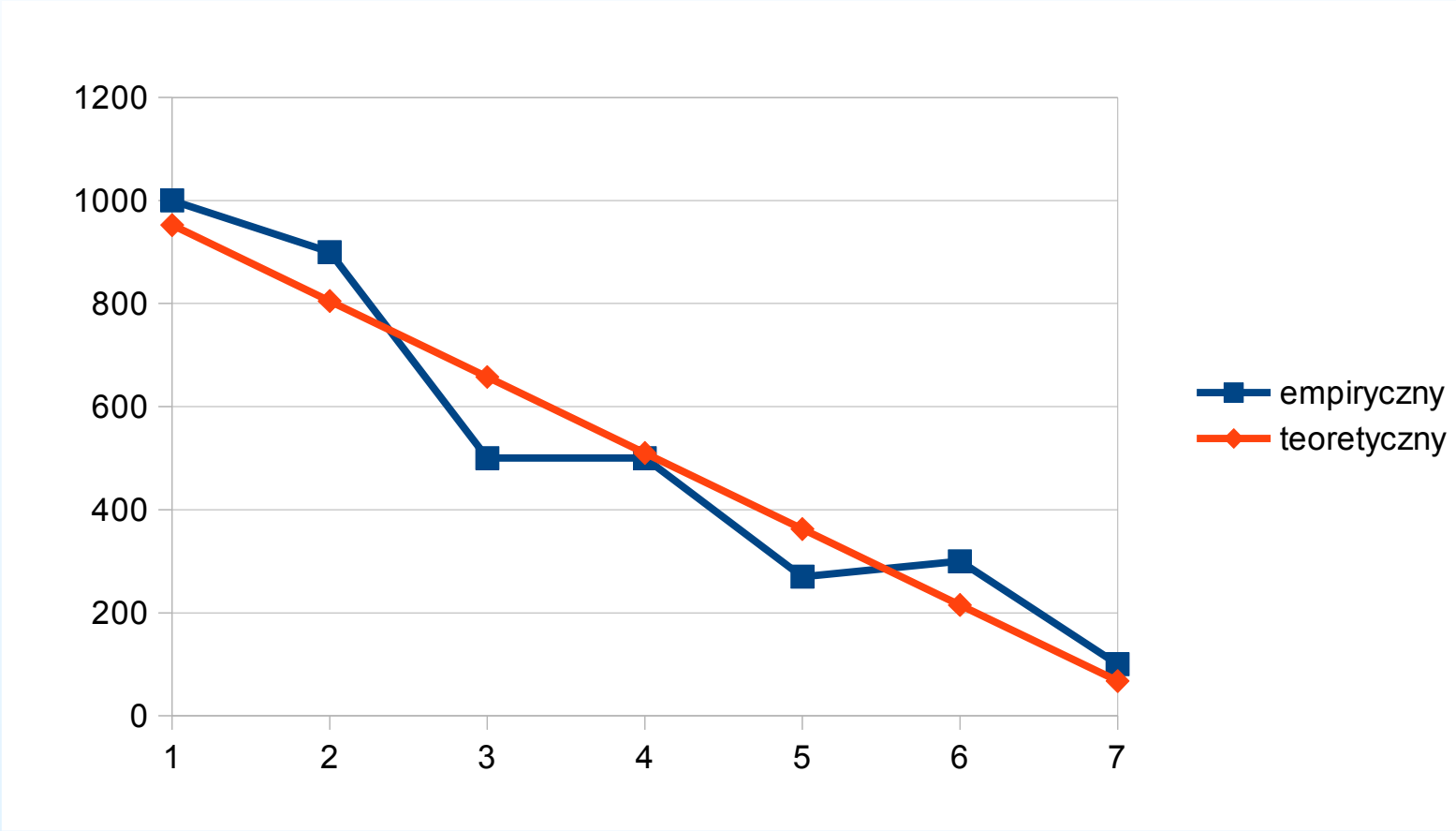
- $\hat{y}_i = 952{,}5 - 147{,}5x_i$

Regresja liniowa

- jeżeli wcześniej zostały obliczone współczynnik korelacji oraz średnie i odchylenia stadardowe, to współczynniki regresji można obliczyć ze wzorów:
- $b = r\frac{s_y}{s_x}$
- $a = \bar{y} - b\bar{x}$

Ocena dopasowania regresji

- sporządzenie wykresu



Wariancja resztowa

- reszty: $e_i = y_i - \hat{y}_i$
- wariancja resztowa: $Se^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$
- odchylenie standardowe składnika resztowego:
  $Se = \sqrt{Se^2} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$
  - pokazuje rozrzut rzeczywistych reszt w stosunku do funkcji regresji

Współczynnik zmienności przypadkowej

- $Ve = \frac{Se}{\bar{y}} \cdot 100$
  - pokazuje natężenie wahań przypadkowych

Oszacowanie warości cech zależnych

- przypusza się, że zmiany w poziomie cechy zależnej są wynikiem
  - oddziaływania zmiennej niezależnej
  - działania czynników losowych
- dobroć dopasowywania regresji do danych rzeczywistych pokazuje
  - *współczynnik deternimacji:* jaka część cechy zależnej jest wyjaśniona kształtowaniem się cechy niezależnej
  - *współczynnik zbieżności (indeternimacji):* jaka część cechy zależnej jest wywołana działaniem czynników losowych

Współczynnik zbieżnośći

- $\varphi^2 = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$
  - przyjmuje wartości od 0 do 1
  - informuje, jaka część zmienności cechy zależnej Y nie jest wyjaśniona zmianami cechy zależnej X
  - im bliższy 0, tym funkcja jest lepiej dopasowana do danych empirycznych

Współczynnik determinacji

- $R^2 = 1 - \varphi^2$
  - indeks korelacji $R = \sqrt{R^2}$
  - w przypdaku regresji liniowej jest równy współczynnikowi korelacji Pearsona $R = \sqrt{1 - \varphi^2} = r$

Przykład z działkami

- $Se = 102{,}98$ zł/m^2
- $\varphi^2 = 0{,}08$ (8%)
- $R^2 = 0,92$
- $r = \pm\sqrt{R} = -0,959$