STATYSTYKA OPISOWA

Dr Alina Gleska

Instytut Matematyki WE PP

28 września 2018

Jednostka zbiorowości - para (X, Y). Przy badaniu korelacji nie ma znaczenia, którą cechę przyjmiemy za X, a którą za Y. Rozróżnienie to jest istotne w analizie regresji.

Badanie związków między cechami ma sens tylko wtedy, gdy występuje między nimi więź przyczynowo-skutkowa, dlatego analiza korelacji powinna być zawsze poprzedzona analizą merytoryczną problemu.

Celem analizy korelacji jest stwierdzenie:

- czy między badanymi cechami występuje współzależność,
- jaki jest kształt zależności (liniowa, nieliniowa),
- jaka jest jej siła,
- jaki jest jej kierunek.

Ponieważ od oceny kształtu zależności zależy wybór właściwej miary korelacji, najpierw ocenia się kształt zależności, a dopiero potem stosując odpowiednie miary, określa się siłę i kierunek tej zależności.

Wyróżniamy dwa rodzaje współzależności:

- funkcyjną - konkretnej wartości jednej zmiennej odpowiada tylko jedna, ściśle określona wartość drugiej zmiennej;
- statystyczną (stochastyczną) - konkretnej wartości jednej zmiennej odpowiada wiele różnych wartości drugiej zmiennej.

Szczególnym przypadkiem zależności statystycznej jest współzależność korelacyjna, kiedy określonym wartościom jednej zmiennej odpowiadają ściśle określone, lecz różne średnie wartości drugiej zmiennej.

Prezentacja graficzna danych:

- szereg korelacyjny – dwuwymiarowy szereg szczegółowy
- diagram korelacyjny – przedstawienie punktów (x_i, y_i) w układzie współrzędnych
- tablica korelacyjna – tabela dla danych pogrupowanych

Prezentacja graficzna danych:

- szereg korelacyjny – dwuwymiarowy szereg szczegółowy
- diagram korelacyjny – przedstawienie punktów (x_i, y_i) w układzie współrzędnych
- tablica korelacyjna – tabela dla danych pogrupowanych

Prezentacja graficzna danych:

- szereg korelacyjny – dwuwymiarowy szereg szczegółowy
- diagram korelacyjny – przedstawienie punktów (x_i, y_i) w układzie współrzędnych
- tablica korelacyjna – tabela dla danych pogrupowanych

Diagram korelacyjny pozwala ocenić, czy między badanymi cechami istnieje zależność (korelacja), o jakiej sile, kierunku i kształcie:

- siła zależności – oceniamy na podstawie rozrzutu punktów na wykresie; niewielki rozrzut wskazuje na dużą siłę związku, im punkty bardziej rozrzucone, tym związek jest słabszy;
- kierunek zależności – korelacja dodatnia lub ujemna (tylko dla cech wyrażonych w skali co najmniej porządkowej)
- kształt zależności – postać funkcji matematycznej opisującej związek między badanymi cechami; korelacja liniowa lub krzywoliniowa.

Diagram korelacyjny pozwala ocenić, czy między badanymi cechami istnieje zależność (korelacja), o jakiej sile, kierunku i kształcie:

- siła zależności – oceniamy na podstawie rozrzutu punktów na wykresie; niewielki rozrzut wskazuje na dużą siłę związku, im punkty bardziej rozrzucone, tym związek jest słabszy;
- kierunek zależności – korelacja dodatnia lub ujemna (tylko dla cech wyrażonych w skali co najmniej porządkowej)
- kształt zależności – postać funkcji matematycznej opisującej związek między badanymi cechami; korelacja liniowa lub krzywoliniowa.

Diagram korelacyjny pozwala ocenić, czy między badanymi cechami istnieje zależność (korelacja), o jakiej sile, kierunku i kształcie:

- siła zależności – oceniamy na podstawie rozrzutu punktów na wykresie; niewielki rozrzut wskazuje na dużą siłę związku, im punkty bardziej rozrzucone, tym związek jest słabszy;
- kierunek zależności – korelacja dodatnia lub ujemna (tylko dla cech wyrażonych w skali co najmniej porządkowej)
- kształt zależności – postać funkcji matematycznej opisującej związek między badanymi cechami; korelacja liniowa lub krzywoliniowa.

Diagram korelacyjny pozwala ocenić, czy między badanymi cechami istnieje zależność (korelacja), o jakiej sile, kierunku i kształcie:

- siła zależności – oceniamy na podstawie rozrzutu punktów na wykresie; niewielki rozrzut wskazuje na dużą siłę związku, im punkty bardziej rozrzucone, tym związek jest słabszy;
- kierunek zależności – korelacja dodatnia lub ujemna (tylko dla cech wyrażonych w skali co najmniej porządkowej)
- kształt zależności – postać funkcji matematycznej opisującej związek między badanymi cechami; korelacja liniowa lub krzywoliniowa.

Tablica korelacyjna

Dla cech skokowych lub jakościowych

Cecha X	Cecha Y				Ogółem
	y_1	y_2	$\dots$	y_ℓ	
x_1	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Ogółem	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet\ell}$	n

Tablica korelacyjna

Dla cech ciągłych

Cecha X	Cecha Y				Ogółem
	(y_0, y_1)	(y_1, y_2)	$\dots$	$(y_{\ell-1}, y_\ell)$	
(x_0, x_1)	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
(x_1, x_2)	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
(x_{k-1}, x_k)	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Ogółem	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet \ell}$	n

Rozkład brzegowy cechy X

Rozkład brzegowy to rozkład jednej cechy niezależnie od tego, jaką wartość przyjmuje druga cecha.

Cecha X	Cecha Y				liczebność
	(y_0, y_1)	(y_1, y_2)	$\dots$	$(y_{\ell-1}, y_\ell)$	
(x_0, x_1)	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
(x_1, x_2)	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
(x_{k-1}, x_k)	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Ogółem	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet \ell}$	n

Rozkład brzegowy cechy Y

Cecha X	Cecha Y				Ogółem
	(y_0, y_1)	(y_1, y_2)	$\dots$	$(y_{\ell-1}, y_\ell)$	
(x_0, x_1)	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
(x_1, x_2)	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
(x_{k-1}, x_k)	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
liczebność	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet \ell}$	n

Rozkład warunkowy

Rozkład warunkowy to rozkład jednej cechy pod warunkiem, że druga cecha przyjmuje określoną wartość.
Rozkład X pod warunkiem $y \in (y_1, y_2)$

Cecha X	Cecha Y				Ogółem
	(y_0, y_1)	(y_1, y_2)	$\dots$	$(y_{\ell-1}, y_\ell)$	
(x_0, x_1)	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
(x_1, x_2)	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
(x_{k-1}, x_k)	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Ogółem	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet\ell}$	n

Rozkład warunkowy

Rozkład Y pod warunkiem $x \in (x_1, x_2)$

Cecha X	Cecha Y				Ogółem
	(y_0, y_1)	(y_1, y_2)	$\dots$	$(y_{\ell-1}, y_\ell)$	
(x_0, x_1)	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
(x_1, x_2)	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
(x_{k-1}, x_k)	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Ogółem	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet \ell}$	n

Współczynnik korelacji liniowej Pearsona

Współczynnik korelacji liniowej Pearsona stosujemy wtedy, gdy związek badanych cech jest liniowy. Wzór:

$$r = \frac{cov(X, Y)}{S_X \cdot S_Y},$$

gdzie $cov(X, Y)$ oznacza kowariancję cech X i Y, a S_X - odch. stand. zmiennej X, S_Y - odch. stand. zmiennej Y.

Współczynnik korelacji liniowej Pearsona

Własności:

- współczynnik r jest miarą symetryczną, tzn. jego wartość nie zależy od tego, czy badamy nim zależność cechy X od cechy Y, czy odwrotnie,
- $r \in [-1, 1]$,
- znak r, który jest zgodny ze znakiem kowariancji, informuje o kierunku korelacji:
  - (i) $r > 0$ – korelacja dodatnia (wzrost wartości jednej cechy pociąga za sobą wzrost średnich wartości drugiej cechy),
  - (ii) $r < 0$ – korelacja ujemna (wzrost wartości jednej cechy pociąga za sobą spadek średnich wartości drugiej cechy),
- siła korelacji:

Współczynnik korelacji liniowej Pearsona

- siła korelacji:
  - (i) $|r| \in [0, 0.2)$ – bardzo słaby związek liniowy (praktycznie brak związku),
  - (ii) $|r| \in [0.2, 0.4)$ – słaby zw. lin.,
  - (iii) $|r| \in [0.4, 0.6)$ – umiarkowany zw. lin.,
  - (iv) $|r| \in [0.6, 0.8)$ – silny zw. lin.,
  - (v) $|r| \in [0.8, 1)$ – bardzo silny zw. lin.,
  - (vi) $|r| = 1$ – zależność funkcyjna (na wykresie punkty empiryczne układają się idealnie na linii prostej).

Wyznaczanie współczynnika korelacji liniowej Pearsona

Wyznaczanie współczynnika korelacji liniowej Pearsona dla DANYCH INDYWIDUALNYCH (W POSTACI SZEREGU KORELACYJNEGO)

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

Uwaga: kowariancja jest wielkością mianowaną (np. lata $\cdot$ cm), więc z powodu jej nienaturalnych jednostek miary nie interpretuje się jej. Jej znaczenie polega na tym, że jej znak informuje nas o kierunku korelacji.

Wyznaczanie współczynnika korelacji liniowej Pearsona

Odchylenia standardowe S_X i S_Y obliczamy zgodnie ze wzorami:

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}$$

oraz

$$S_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \overline{y}^2}.$$

Wyznaczanie współczynnika korelacji liniowej Pearsona

Wyznaczanie współczynnika korelacji liniowej Pearsona dla DANYCH POGRUPOWANYCH

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{k}(x_i^0 - \overline{x})(y_j^0 - \overline{y})n_{ij} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{k}x_i^0 y_j^0 n_{ij} - \overline{xy},$$

gdzie x_i^0 – środek i-tego przedziału cechy X (w przypadku, gdy szereg dla cechy X jest punktowy w miejsce x_i^0 wstawiamy x_i oznaczający i-ty wariant cechy X), podobnie dla y_j^0, zaś n_{ij} to liczebność w kratce tablicy leżącej na przecięciu i-tego wiersza i j-tej kolumny.

Wyznaczanie współczynnika korelacji liniowej Pearsona

Odchylenia standardowe S_X i S_Y obliczamy zgodnie ze wzorami:

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{r}(x_i^0 - \overline{x})^2 n_{i\bullet}} = \sqrt{\frac{1}{n}\sum_{i=1}^{r}(x_i^0)^2 n_{i\bullet} - \overline{x}^2},$$

gdzie $n_{i\bullet}$ – suma liczebności w wierszach (liczebność brzegowa w rozkładzie cechy X)
oraz

$$S_Y = \sqrt{\frac{1}{n}\sum_{j=1}^{k}(y_j^0 - \overline{y})^2 n_{\bullet j}} = \sqrt{\frac{1}{n}\sum_{j=1}^{k}(y_j^0)^2 n_{\bullet j} - \overline{y}^2},$$

gdzie $n_{\bullet j}$ – suma liczebności w kolumnach (liczebność brzegowa w rozkładzie cechy Y).

Współczynnik korelacji rang Spearmana

Współczynnik korelacji rang Spearmana, zwany też współczynnikiem korelacji kolejnościowej, stosujemy do oceny kierunku i siły korelacji w przypadku, gdy:

- cechy są niemierzalne, ale istnieje możliwość uporządkowania wariantów cechy (czyli cechy te są wyrażone w skali porządkowej),
- cechy są mierzalne, przy czym liczba wariantów przyjmowanych przez te cechy musi być skończona.

Współczynnik korelacji rang Spearmana

Najpierw poszczególnym wariantom obu cech nadaje się rangi, czyli numery od 1 do n, które pozwalają uporządkować ciąg obserwacji (rosnąco lub malejąco). Takie tworzenie rankingu cech nazywamy rangowaniem.
Uwaga 1: sposób rangowania musi być jednakowy dla obu cech (tj. dla obu cech w kolejności rosnącej lub dla obu cech w kolejności malejącej).
Uwaga 2: w przypadku, gdy dana wartość (dany wariant) występuje wielokrotnie, wówczas wartościom tym nadajemy tę samą rangę równą średniej arytmetycznej kolejnych numerów pozycji, na których stoją te jednakowe wartości (są to tzw. rangi wiązane).

Współczynnik korelacji rang Spearmana

Cecha X	Cecha Y	Ranga X	Ranga Y
x_1	y_1	1	1
x_2	y_2	2	3
x_3	y_3	3,5	4
x_4	y_4	3,5	2
x_5	y_5	5	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Współczynnik korelacji rang Spearmana

Wzór:

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2,$$

gdzie n to liczba obserwacji jednej z cech, d_i – różnica między rangami, które są przypisane i-tej obserwacji pierwszej i drugiej cechy.

Współczynnik korelacji rang Spearmana

WŁASNOŚCI:

- współczynnik korelacji rang Spearmana r_s jest miarą symetryczną,
- korelacja rang jest niewrażliwa na obserwacje odstające (które mogą zaburzyć wartość współczynnika korelacji liniowej Pearsona),
- $r_s \in [-1, 1]$,

Współczynnik korelacji rang Spearmana

- znak informuje o kierunku współzależności między badanymi cechami:
  - (i) $r_s > 0$ – korelacja dodatnia (występuje zgodność rang, czyli wyższym rangom jednej cechy odpowiadają na ogół wyższe rangi drugiej cechy),
  - (ii) $r_s = 1$ – idealna zgodność rang,
  - (iii) $r_s = 0$ – brak korelacji,
  - (iv) $r_s < 0$ – korelacja ujemna (występuje niezgodność rang, czyli wyższym rangom jednej cechy odpowiadają na ogół niższe rangi drugiej cechy),
  - (v) $r_s = -1$ – idealna niezgodność rang,

Współczynnik korelacji rang Spearmana

- siła korelacji
  - (i) $|r_s| \in [0, 0.2)$ – bardzo słaba współzależność (praktycznie brak związku),
  - (ii) $|r_s| \in [0.2, 0.4)$ – słaba współzależność,
  - (iii) $|r_s| \in [0.4, 0.6)$ – umiarkowana współzależność,
  - (iv) $|r_s| \in [0.6, 0.8)$ – silna współzależność,
  - (v) $|r_s| \in [0.8, 1)$ – bardzo silna współzależność.

PRZYKŁAD.

Analiza korelacji - cechy jakościowe

Jeżeli co najmniej jedna z cech jest cechą jakościową typu nominalnego, to nie ma możliwości nadania rang. Wówczas korzystamy z tzw. współczynników kontyngencji (inaczej: zbieżności korelacyjnej). Należą do nich:

- współczynnik φ Yule'a,
- współczynnik T Czuprowa,
- współczynnik V Cramera,
- współczynnik kontyngencji P Pearsona,
- współczynnik Q Kendalla (tylko do pomiaru cech dychotomicznych).

Analiza korelacji - cechy jakościowe

Wszystkie współczynniki kontyngencji oparte są na tzw. statystyce χ^2 (chi-kwadrat):

$$\varphi = \sqrt{\frac{\chi^2}{n}}, \quad T = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(k-1)}}},$$

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, k-1)}}, \quad P = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

Analiza korelacji - cechy jakościowe

Własności współczynników φ, T, V, P:

- są symetryczne,
- są zawsze nieujemne,
- nie informują o kierunku korelacji,
- $\varphi, T, V, P \in [0,1]$ (na ogół; może się zdarzyć, że niektóre z nich przyjmą wartość większą niż 1),
- informują o sile korelacji:
  - (i) $0-0.2$ – bardzo słaba współzależność (praktycznie brak związku),
  - (ii) $0.2-0.4$ – słaba współzależność,
  - (iii) $0.4-0.6$ – umiarkowana współzależność,
  - (iv) $0.6-0.8$ – silna współzależność,
  - (v) $0.8-1$ – bardzo silna współzależność.

Statystyka χ^2

Statystykę χ^2 obliczamy według wzoru:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - n_{ij}^t)^2}{n_{ij}^t},$$

gdzie n_{ij} – liczebności w poszczególnych kratkach tablicy kontyngencji, tzw. liczebności empiryczne (zaobserwowane w próbie), n_{ij}^t – liczebności teoretyczne (są to takie liczebności, jakie powinny wystąpić w poszczególnych kratkach tablicy, gdyby między badanymi cechami nie istniała zależność).

Statystyka χ^2

Liczebności teoretyczne obliczamy według wzoru:

$$n_{ij}^t = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n},$$

gdzie $n_{i\bullet}$ – suma liczebności z wiersza, $n_{\bullet j}$ – suma liczebności z kolumny, n – liczebność zbiorowości.

Statystyka χ^2

Istotą obliczania charakterystyki χ^2 jest porównanie liczebności zaobserwowanych n_{ij} z liczebnościami teoretycznymi n_{ij}^t, czyli takimi, których należałoby oczekiwać, gdyby cechy były niezależne. Jeżeli okaże się, że liczebności zaobserwowane są bardzo bliskie liczebnościom teoretycznym wówczas można przypuszczać, że badane cechy są niezależne. Natomiast im bardziej liczebności zaobserwowane różnią się od teoretycznych, tym większa jest wartość statystyki χ^2 i związek badanych cech jest silniejszy.
PRZYKŁAD

Statystyka χ^2

Dla cech dychotomicznych tablica kontyngencji jest tablicą czteropolową.

Cecha X	Cecha Y	
	y_1	y_2
x_1	a	b
x_2	c	d

Po dość żmudnych obliczeniach otrzymujemy wzór na statystykę χ^2:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)},$$

który możemy stosować, jeśli wszystkie $a, b, c, d \geq 5$.

Statystyka χ^2

Jeżeli warunek ten nie jest spełniony, tzn. gdy którakolwiek z wartości $a, b, c, d < 5$, wówczas do powyższego wzoru wprowadzamy tzw. poprawkę Yatesa:

$$\chi^2 = \frac{n(|ad - bc| - 0,5n)^2}{(a+b)(a+c)(b+d)(c+d)}.$$

Statystyka χ^2

W przypadku tablic czteropolowych możemy obliczyć WSPÓŁCZYNNIK Q KENDALLA:

$$Q = \frac{ad - bc}{ad + bc}.$$

Współczynnik Q Kendalla może przyjmować wartości ujemne (ale nie oznacza to, że korelacja jest ujemna; żaden ze współczynników kontyngencji nie wskazuje kierunku korelacji).

Statystyka χ^2

Własności współczynnika Q Kendalla:

- $Q \in [-1, 1]$,
- jest symetryczny,
- nie informuje o kierunku korelacji,
- siła korelacji:
  - $|Q| \in [0, 0.2)$ – bardzo słaba współzależność (praktycznie brak związku),
  - $|Q| \in [0.2, 0.4)$ – słaba współzależność,
  - $|Q| \in [0.4, 0.6)$ – umiarkowana współzależność,
  - $|Q| \in [0.6, 0.8)$ – silna współzależność,
  - $|Q| \in [0.8, 1]$ – bardzo silna współzależność.

PRZYKŁAD.