Matematyka z elementami statystyki

Łukasz Dawidowski

Instytut Matematyki, Uniwersytet Śląski

Korelacja

Zależność funkcyjna – wraz ze wzrostem jednej zmiennej następuje ściśle określona zmiana druiej zmiennej.

X – zmienne niezależna (objaśniająca)
Y – zmienne zależna (objaśniana)

Korelacja

Zależność stochastyczna – wraz ze zmianą jednej zmiennej następuje zmiana rozkładu drugiej zmiennej.

Zależność statystyczna (korelacyjna) – określonym wartościom jednej zmiennej odpowiadają ściśle określone śrenie wartości drugiej zmiennej.

Korelacja

Współczynnik korelacji – miernik siły zależności między badanymi zmiennymi

- zmiany jednokierunkowe – wartości obu zmiennych na ogół rosną lub maleją – korelacja dodatnia
- zmiany różnokierunkowe – wzrostowi jednego szeregu odpowiada spadek wartości drugiego – korelacja ujenma

Tablica korelacyjna

	y_1	y_2	$\ldots$	y_j	$\ldots$	y_r	suma
x_1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1r}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2r}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{ir}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\ldots$	n_{kj}	$\ldots$	n_{kr}	$n_{k\cdot}$
suma	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot j}$	$\ldots$	$n_{\cdot r}$	n

Tablica korelacyjna

gdzie:

$$n_{i\cdot} = \sum_{j=1}^{r} n_{ij}$$

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

$$n = \sum_{i=1}^{k} \sum_{j=1}^{r} n_{ij}$$

Zamiast częstości absolutnych (bezwzględnych) n_{ij} można w tablicy korelacyjnej podawać częstości względne $\frac{n_{ij}}{n}$.

Korelacja

Rozkład brzegowy – prezentuje strukturę wartości jednej zmiennej bez względu na kształtowanie się wartości drugiej zmiennej.

np.

x_1	x_2	$\cdots$	x_i	$\cdots$	x_k
$n_{1\cdot}$	$n_{2\cdot}$	$\cdots$	$n_{i\cdot}$	$\cdots$	$n_{k\cdot}$

Korelacja

Rozkład warunkowy – prezentuje strukturę wartości jednej zmiennej pod warunkiem, że druga osiągnęła określoną wartość.

Korelacja

Średnie arytmetyczne rozkładów brzegowych:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_{i\cdot}$$

$$\overline{y} = \frac{1}{n} \sum_{j=1}^{r} y_j n_{\cdot j}$$

Korelacja

Średnie arytmetyczne rozkładów warunkowych:

$$\overline{x}_j = \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} x_i n_{ij}$$

$$\overline{y}_i = \frac{1}{n_{i\cdot}} \sum_{j=1}^{r} y_j n_{ij}$$

Korelacja

Wariancje rozkładów brzegowych:

$$s^2(x) = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_{i\cdot} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_{i\cdot} - \overline{x}^2$$

$$s^2(y) = \frac{1}{n}\sum_{j=1}^{r}(y_j - \overline{y})^2 n_{\cdot j} = \frac{1}{n}\sum_{j=1}^{r} y_j^2 n_{\cdot j} - \overline{y}^2$$

Korelacja

Wariancje rozkładów warunkowych:

$$s_j^2(x) = \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} (x_i - \overline{x}_j)^2 n_{ij} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} x_i^2 n_{ij} - \overline{x}_j^2$$

$$s_i^2(y) = \frac{1}{n_{i\cdot}} \sum_{j=1}^{r} (y_j - \overline{y}_i)^2 n_{ij} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{r} y_j^2 n_{ij} - \overline{y}_i^2$$

Korelacja

Cecha X jest stochastycznie niezależna od cechy Y, jeżeli:

$$\overline{x}_1 = \overline{x}_2 = \ldots = \overline{x}_r$$

oraz

$$s_1^2(x) = s_2^2(x) = \ldots = s_r^2(x)$$

Korelacja

Cecha Y jest stochastycznie niezależna od cechy X, jeżeli:

$$\overline{y}_1 = \overline{y}_2 = \ldots = \overline{y}_k$$

oraz

$$s_1^2(y) = s_2^2(y) = \ldots = s_k^2(y)$$

Korelacja

Niezależność korelacyjna – tylko pierwsze warunki.

Uwaga:
Jeżeli zmienne są niezależne stochastycznie, to są również niezależne korelacyjnie.
Ale nie jest na odwrót!

Test niezależności chi–kwadrat

Można go stosować zarówno do cech mierzalnych, jak i niemierzalnych.

Zakładamy, że przedmiotem badania jest populacja generalna scharakteryzowana za pomocą dwóch cech jakościowych. Z populacji tej bierzemy n elementów, a wynik tej próby zapisujemy w tablicy korelacyjnej (zwanej też tablicą niezależności)

Test niezależności chi–kwadrat

Weryfikujemy następującą hipotezę zerową:

$$H_0 : \text{ cechy } X \text{ i } Y \text{ są niezależne}$$

wobec hipotezy alternatywnej:

$$H_2 : \text{ cechy } X \text{ i } Y \text{ nie są niezależne}$$

Test niezależności chi–kwadrat

Liczebności teoretyczne:

$$\hat{n}_{ij} = \frac{\text{suma licz. } i\text{–tego wiesza} \times \text{s. licz. } j\text{–tej kolumny}}{\text{liczebność próby}}$$

$$\hat{n}_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

Test niezależności chi–kwadrat

Do weryfikacji hipotezy H_0 o niezależnoci stochastycznej zmiennych wykorzystujemy statystykę:

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

Statystyka ta ma, przy założeniu prawdziwości hipotezy H_0, dla dużych prób asymptotyczny rozkład χ^2 z $(k-1)(r-1)$ stopniami swobody.

Test niezależności chi–kwadrat

Obszar krytyczny (prawostronny) w rozważanym teście stanowi nierówność:

$$\chi^2 \geqslant \chi^2_\alpha$$

gdzie χ^2_α jest wartością krytyczną odczytywaną z tablicy rozkładu χ^2 dla ustalonego z góry poziomu istotności α i dla $(k-1)(r-1)$ stopni swobody tak, aby zachodziła relacja

$$P(\chi^2 \geqslant \chi^2_\alpha) = \alpha$$

Test niezależności chi–kwadrat

- gdy $\chi^2 \geqslant \chi^2_\alpha$, to hipotezę o niezależności cech X i Y odrzucamy
- gdy $\chi^2 < \chi^2_\alpha$, to nie mamy podstaw do odrzucenia hipotezy H_0

Test niezależności chi–kwadrat

Gdy liczba stopni swobody przekracza 30, to w celu weryfikacji hipotezy H_0 o stochastycznej niezależności zmiennych X i Y wykorzystujemy test:

$$Z = \sqrt{2\chi^2} - \sqrt{2(k-1)(r-1)-1}$$

który ma rozkład normalny $N(0,1)$.

Test niezależności chi–kwadrat

Tablica 2×2 (tablica czteropolowa)

	1	2	$n_{i\cdot}$
1	a	b	$a+b$
2	c	d	$c+d$
$n_{\cdot j}$	$a+c$	$b+d$	n

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

Test niezależności chi–kwadrat

Jeżeli w tablicy czteropolowej przynajmniej jedna liczebność jest mniejsza od 10 wprowadzamy poprawkę Yatesa na nieciągłość:

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

Opisowe miary korelacji dwóch zmiennych

- współczynnik zbieżności Czuprowa,
- wskaźniki (stosunki) korelacyjne Pearsona,
- współczynnik korelacji liniowej Pearsona,
- współczynnik korelacji rangowej Spearmana.

Współczynnik zbieżności Czuprowa

$$T_{XY} = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(k-1)}}}$$

$T_{XY} = T_{YX}$, $T_{XY} \in \langle 0, 1 \rangle$.

Gdy $T_{XY} = 0$, to zmienne są stochastycznie niezależne.

Gdy $T_{XY} = 1$, to obserwujemy zależność funkcyjną zmiennych.

Im bliżej zera, tym zależność między zmiennymi jest słabsza.

Nie wskazuje kierunku korelacji.

Współczynniki korelacyjne Pearsona

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_{i\cdot}$$

$$\overline{y} = \frac{1}{n} \sum_{j=1}^{r} y_j n_{\cdot j}$$

$$\overline{x}_j = \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} x_i n_{ij}$$

$$\overline{y}_i = \frac{1}{n_{i\cdot}} \sum_{j=1}^{r} y_j n_{ij}$$

Współczynniki korelacyjne Pearsona

$$s^2(x) = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_{i\cdot}$$

$$s^2(y) = \frac{1}{n}\sum_{j=1}^{r}(y_j - \overline{y})^2 n_{\cdot j}$$

$$s^2(\overline{x}_j) = \frac{1}{n}\sum_{j=1}^{r}(\overline{x}_j - \overline{x})^2 n_{\cdot j}$$

$$s^2(\overline{y}_i) = \frac{1}{n}\sum_{i=1}^{k}(\overline{y}_i - \overline{y})^2 n_{i\cdot}$$

Współczynniki korelacyjne Pearsona

$$e_{XY} = \sqrt{\frac{s^2(\overline{x}_j)}{s^2(x)}} = \frac{s(\overline{x}_j)}{s(x)}$$

$$e_{YX} = \sqrt{\frac{s^2(\overline{y}_i)}{s^2(y)}} = \frac{s(\overline{y}_i)}{s(y)}$$

$e_{XY} \neq e_{YX}$ (niesymetryczny), $e_{XY}, e_{YX} \in \langle 0, 1 \rangle$.

Współczynniki korelacyjne Pearsona

Są równe zero, gdy cechy są nieskorelowane, zaś równe jeden, kiedy między cechami występuje zależność funkcyjne.
Im wartość jest bliższa jeden, to zależność korelacyjna między cechami jest silniejsza.
Nie wskazują kierunku korelacji.
Ważne jest, aby ustalić, która cecha jest niezależna, a która zależna.

Współczynnik korelacji liniowej Pearsona

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

$cov(X,Y)$ – kowariancja
$cov(X,Y) \in \langle -s(x)s(y), s(x)s(y) \rangle$

- $cov(X,Y) = 0$ – brak zależności korelacyjnej
- $cov(X,Y) > 0$ – dodatnia zależność korelacyjna
- $cov(X,Y) < 0$ – ujemna zależność jkorelacyjna

Współczynnik korelacji liniowej Pearsona

$$r_{XY} = r_{YX} = \frac{cov(X, Y)}{s(x)s(y)}$$

$$r_{XY} \in \langle -1, 1 \rangle$$

- $r_{XY} = 0$ – brak zależności korelacyjnej
- $r_{XY} > 0$ – dodatnia zależność korelacyjna
- $r_{XY} < 0$ – ujemna zależność jkorelacyjna

Współczynnik korelacji liniowej Pearsona

Współczynnik determinacji (określoności)

$$r_{XY}^2$$

informuje o tym jaka część zmian zmiennej objaśnianej (skutek) jest wyjaśniana przez zmiany wartości zmiennej objaśniającej (przyczyna).

Współczynnik korelacji rang Spearmana

- służy do opisu siły korelacji cech mających charakter jakościowy i istnieje możliwość uporządkowania obserwacji empirycznych w określonej kolejności
- może być użyty do badania cech ilościowych w przypadku niewielkiej liczby obserwacji

Rangowanie – uporządkowanym rosnąco lub malejąco wartościom zmiennych nadajemy numery kolejnych liczb naturalnych. W przypadku, gdy występują jednakowe wartości zmiennych, przyporządkowujemy im średnią arytmetyczną obliczoną z ich kolejnych numerów.

Współczynnik korelacji rang Spearmana

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

gdzie d_i oznaczają różnice między rangami odpowiadających sobie wartości cechy x i cechy y.

$r_s \in \langle -1, 1 \rangle$
Interpretacja jak dla współczynnika korelacji liniowej Pearsona.

Cechy niemierzalne

Tablica asocjacji

	+	−	suma
+	a	b	$a+b$
−	c	d	$c+d$
suma	$a+c$	$b+d$	n

+ oznacza, że dana cecha występuje

− oznacza, że danej cechy nie zaobserwowano (nie występuje)

Tablica asocjacji może mieć więcej wierszy i kolumn (decyzja należy do badającego)

Współczynnik φ – Yule'a

służy do badania siły związku dwóch cech jakościowych, z których każda ma dwa warianty

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

gdzie χ^2 jest wartością statystyki dla testu chi–kwadrat

Współczynnik φ – Yule'a

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

$\varphi \in \langle -1, 1 \rangle$

$\varphi = 0$ – zmienne są niezależne

Im bliżej 1 lub -1 tym zależność jest większa

φ nie informuje o kierunku zależności

Współczynnik Cole'a

$$\varphi_{kor} = \begin{cases} \dfrac{ad-bc}{n \cdot min(b,c)+(ad-bc)}, & \varphi \geqslant 0 \\ \\ \dfrac{ad-bc}{n \cdot min(a,d)-(ad-bc)}, & \varphi < 0. \end{cases}$$

Współczynnik kontyngencji C Pearsona

Może być stosowany do tablic wielodzielnych dowolnej wielkości (co najmniej cztery pola) i dowolnej formy (kwadratowe i prostokątne).

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{\varphi^2}{\varphi^2 + n}}$$

$C \in \langle 0, 1 \rangle$
$C = 0$ – cechy są niezależne
$C = 1$ – cechy są zależne

Funkcja regresji liniowej MNK

Funkcja regresji liniowej zmiennej Y względem zmiennej X w populacji generalnej jest dana za pomocą wzoru

$$\hat{y}_i = a_1 x_i + a_0,$$

gdzie

$$a_1 = \frac{cov(X, Y)}{S_X}$$

oraz

$$a_0 = \bar{y} - a_1 \bar{x}.$$