Metody probabilistyczne i statystyka - wykład trzynasty[1]
Testowania hipotez: testy istotności część II

dr Jarosław Kotowicz

Instytut Informatyki Uniwersytet w Białymstoku

wersja z roku ak. 2020/21

Spis treści

1. Dwuwymiarowy rozkład empiryczny
2. Test niezależności chi-kwadrat
3. Miary zależności cech oparte na statystyce chi-kwadrat
   - Współczynnik zbieżności V Cramera
   - Inne miary zależności cech jakościowych
4. Badanie zależności dwóch cech
   - Kowariancja rozkładu empirycznego – przypomnienie
   - Współczynnik korelacji liniowej Pearsona (cechy mierzalne)
   - Współczynnik korelacji rang Spearmana (skala porządkowa)
   - Empiryczne krzywe regresji
   - Stosunki korelacyjne

Uwagi

Przyjmujemy, że populacja generalna jest badana jednocześnie pod względem dwóch cech X i Y. Odpowiada to sytuacji w rachunku prawdopodobieństwa, że na zbiorze zdarzeń elementarnych została określona dwuwymiarowa zmienna losowa (X, Y).

Niech tak jak poprzednio cecha X przyjmuje k wartości tj. $x_1, \ldots, x_k$, zaś cecha Y przyjmuje l wartości $y_1, \ldots, y_l$. Dla każdej cechy (oddzielnie) można określić jednowymiarowy rozkład empiryczny tzn. uporządkować w postaci szeregu rozdzielczego punktowego.

W celu określenia łącznego rozkładu obu cech należy ustalić, ile jednostek zbiorowości przyjmuje możliwe pary wartości (x_i, y_j). Tę ilość oznaczamy n_{ij}, gdzie $i \in \overline{1, k}$; $j \in \overline{1, l}$.

Liczebności empirycznego rozkładu dwuwymiarowego

Definicja 1

Empiryczny dwuwymiarowy rozkład cechy (X, Y) *(empiryczny łączny rozkład cech* X, Y*) określają liczebności* n_{ij}*, gdzie* $i \in \overline{1,k}$*;* $j \in \overline{1,l}$*, odpowiadające parom wartości* (x_i, y_j)*.*

Uwaga 1

Przyjmujemy

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

Liczebności rozkładów brzegowych rozkładu empirycznego

Definicja 2

Rozkład brzegowy cechy X wyznaczają liczebności $n_{i\cdot} = \sum_{j=1}^{l} n_{ij}$, *dla* $i \in \overline{1,k}$.

Definicja 3

Rozkład brzegowy cechy Y wyznaczają liczebności $n_{\cdot j} = \sum_{i=1}^{k} n_{ij}$, *dla* $j \in \overline{1,l}$.

Tablica korelacyjna rozkładu empirycznego

		Y				
		y_1	y_2	$\dots$	y_l	$\sum$
X	x_1	n_{11}	n_{12}	$\dots$	n_{1l}	$n_{1\cdot}$
	x_2	n_{21}	n_{22}	$\dots$	n_{2l}	$n_{2\cdot}$
	$\vdots$	$\dots$				$\vdots$
	x_k	n_{k1}	n_{k2}	$\dots$	n_{kl}	$n_{k\cdot}$
	$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot l}$	n

Częstotliwości względne rozkładu empirycznego

Ostatni wiersz określa rozkład brzegowy cechy Y, zaś ostatnia kolumna określa rozkład brzegowy cechy X.

Można określić również częstotliwości względne rozkładu łącznego

$$\omega_{ij} = \frac{n_{ij}}{n},$$

jak i częstotliwości względne rozkładów brzegowych

$$\omega_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \text{oraz} \quad \omega_{\cdot j} = \frac{n_{\cdot j}}{n}.$$

Rozkłady warunkowe rozkładu empirycznego

W rozkładzie empirycznym dwuwymiarowym można określić rozkłady warunkowe tj. rozkłady jednej zmiennej przy ustalonej drugiej zmiennej.

W tablicy korelacyjnej warunkowe rozkłady zmiennej X są to kolejne kolumn tablicy. Natomiast warunkowe rozkłady zmiennej Y to są kolejne wiersz tablicy korelacyjnej.

	y_1	y_2	$\ldots$	y_l
x_i	n_{i1}	n_{i2}	$\ldots$	n_{il}

Tabela: Rozkłady warunkowe zmiennej Y

Parametry rozkładów brzegowych rozkładu empirycznego. I

Rozkłady brzegowe cechy X:

1. Średnia

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

2. Wariancja

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

Rozkłady brzegowe cechy Y:

1. Średnia

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

Parametry rozkładów brzegowych rozkładu empirycznego. II

2. Wariancja

$$\tilde{s}_Y^2 := \frac{1}{n-1}\sum_{j=1}^{l}(y_j - \overline{y})^2 \cdot n_{\cdot j}.$$

Uwaga 2

Można też określić s_X^2 *i* s_Y^2, *gdzie* $s_X^2 := \frac{n-1}{n}\tilde{s}_X^2$ *i* $s_Y^2 := \frac{n-1}{n}\tilde{s}_Y^2$.

Parametry rozkładów warunkowych rozkładu empirycznego. I

Rozkłady warunkowy cechy X dla $j \in \overline{1, l}$:

1. Średnia
$$\overline{x_j} := \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} x_i \cdot n_{ij}.$$
2. Wariancja
$$\tilde{s}^2_{j,X} := \frac{1}{n_{\cdot j} - 1} \sum_{i=1}^{k} (x_i - \overline{x})^2 \cdot n_{ij}.$$

Rozkłady warunkowy cechy Y dla $i \in \overline{1, k}$:

1. Średnia
$$\overline{y_i} := \frac{1}{n_{i \cdot}} \sum_{j=1}^{l} y_j \cdot n_{ij}.$$

Parametry rozkładów warunkowych rozkładu empirycznego. II

2. Wariancja

$$\tilde{s}^2_{i,Y} := \frac{1}{n_{i\cdot} - 1} \sum_{j=1}^{l} (y_j - \overline{y})^2 \cdot n_{ij}.$$

Uwaga 3

Podobnie, jak dla rozkładów brzegowych, tak i dla rozkładów warunkowych można też określić $s^2_{j,X}$ *i* $s^2_{i,Y}$, *gdzie* $s^2_{j,X} := \frac{n-1}{n}\tilde{s}^2_{j,X}$ *i* $s^2_{i,Y} := \frac{n-1}{n}\tilde{s}^2_{i,Y}$.

Uwagi

1. Będziemy rozważać dwuwymiarową zmienną losową (X, Y).
2. Dla (X, Y) będącej zmienną losową dyskretną, to będziemy zakładać, że zmienna losowa X przyjmuje k wartości tj. $x_1, \ldots, x_k$, zaś zmienna Y przyjmuje l wartości $y_1, \ldots, y_l$.
3. Będziemy też zakładać, że liczba obserwacji w próbie wynosi n.

Spis treści

1. Dwuwymiarowy rozkład empiryczny
2. Test niezależności chi-kwadrat
3. Miary zależności cech oparte na statystyce chi-kwadrat
   - Współczynnik zbieżności V Cramera
   - Inne miary zależności cech jakościowych
4. Badanie zależności dwóch cech
   - Kowariancja rozkładu empirycznego – przypomnienie
   - Współczynnik korelacji liniowej Pearsona (cechy mierzalne)
   - Współczynnik korelacji rang Spearmana (skala porządkowa)
   - Empiryczne krzywe regresji
   - Stosunki korelacyjne

Test niezależności – omówienie. I

Będziemy sprawdzać niezależność zmiennych losowych. Stawiamy hipotezę zerową i alternatywną:

$$H_0: \quad p_{ij} = p_{i\cdot} \cdot p_{\cdot j} \text{ dla wszystkich } i \in \overline{1,k}, j \in \overline{1,l},$$
$$H_1: \quad p_{ij} \neq p_{i\cdot} \cdot p_{\cdot j} \text{ dla niektórych } i \in \overline{1,k}, j \in \overline{1,l}.$$

Gdyby prawdopodobieństwa $p_{i\cdot}$ i $p_{\cdot j}$ były znane, to można byłoby wyznaczyć p_{ij} (przy założeniu prawdziwości hipotezy zerowej), a następnie obliczyć oczekiwana liczność rozkładu dwuwymiarowego $\widehat{n_{ij}} = n \cdot p_{ij}$.

W celu podjęcia decyzji odnośnie hipotezy zerowej należałoby porównać liczebność rozkładu rzeczywistego n_{ij} z liczebnością rozkładu hipotetycznego $\widehat{n_{ij}}$ za pomocą testy χ^2.

Jednak sformułowana hipoteza zerowa nie precyzuje wartości nieznanych $k + l$ prawdopodobieństw rozkładów brzegowych.

Test niezależności – omówienie. II

Z zależności

$$\sum_{i=1}^{k} p_{i\cdot} = \sum_{j=1}^{l} p_{\cdot j} = 1$$

można wyznaczyć dwie nieznane wartości. Zostają wtedy $k + l - 2$ nieznane parametry.

Przy stosowaniu testu χ^2 nieznane prawdopodobieństwa brzegowe należy szacować za pomocą częstotliwości względnych

$$\widehat{p_{i\cdot}} = \frac{n_{i\cdot}}{n} \text{ dla } i \in \overline{1, k-1}, \tag{1}$$

$$\widehat{p_{\cdot j}} = \frac{n_{\cdot j}}{n} \text{ dla } j \in \overline{1, l-1}. \tag{2}$$

Test niezależności – omówienie. III

Oczekiwane liczebności w tablicy korelacyjnej przy założeniu prawdziwości hipotezy zerowej wynoszą

$$\widehat{n_{ij}} = n \cdot \widehat{p_{i\cdot}} \cdot \widehat{p_{\cdot j}} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}.$$

Test zgodności wykorzystuje statystykę

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

gdzie

- próba musi być duża: $\widehat{n_{ij}} \geqslant 5$;
- liczba stopni swobody wynosi: $kl - (k + l - 2) - 1 = (k - 1)(l - 1)$.

Test niezależności – omówienie. IV

Obszar krytyczny, dla poziomu istotności α, opisuje zależność

$$P(\{\chi^2 \geqslant \chi^2_{\alpha,(k-1)(l-1)}\}) = \alpha.$$

Jeżeli wartości obliczona jest nie mniejsza niż wartość hipotetyczna, to hipotezę zerową odrzucamy.

Spis treści

1. Dwuwymiarowy rozkład empiryczny
2. Test niezależności chi-kwadrat
3. Miary zależności cech oparte na statystyce chi-kwadrat
   - Współczynnik zbieżności V Cramera
   - Inne miary zależności cech jakościowych
4. Badanie zależności dwóch cech
   - Kowariancja rozkładu empirycznego – przypomnienie
   - Współczynnik korelacji liniowej Pearsona (cechy mierzalne)
   - Współczynnik korelacji rang Spearmana (skala porządkowa)
   - Empiryczne krzywe regresji
   - Stosunki korelacyjne

Motywacja

Uwaga 4

1. *Do pomiaru siły związku (niezależności lub zależności) wykorzystuje się statystykę chi-kwadrat*

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

   która przyjmuje wartości z przedziału $[0, n(\min\{k, l\} - 1)]$. *Przyjęcie wartości zero przez statystykę oznacza niezależność cech, a przyjęcie wartości maksymalnej oznacza zależność funkcyjną.*

2. *Odrzucenie hipotezy zerowej (o niezależności cech) sugeruje istnienie związku między tymi zmiennymi.*

Współczynnika kontyngencji V Cramera. I

Definicja 4

Współczynnik kontyngencji V Cramera nazywamy unormowaną miarę sił zależności cech wyrażoną za pomocą wzoru

$$V = \sqrt{\frac{\chi^2}{n(\min\{k, l\} - 1)}}. \tag{3}$$

Własności współczynnika V Cramera.

1. Wartości współczynnika V Cramera przyjmują wartości z odcinka $[0, 1]$.
2. Wartości współczynnika V Cramera równą 0 oznacza niezależność cech, a wartość równa 1 zależność funkcyjną.
3. Przy wyznaczaniu wartości współczynnika V Cramera nie jest ważne, którą z cech traktujemy jako zależną, a którą jako niezależną.

Współczynnika kontyngencji V Cramera. II

4. Współczynnik V Cramera stosowany jest w przypadku cech nominalnych[2].
5. Jest miarą współzależności krzywoliniowej.

[2] Bywa też stosowany w przypadku cech mierzalnych

Współczynnik φ Yule'a. I

Definicja 5

Współczynnikiem φ Yule'a jest postaci

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

Współczynnik φ Yule'a ma następujące własności:

1. jeżeli $k = 2$ i l jest dowolne, to $\varphi \in [0, 1]$,
2. jeżeli $k > 2$ i l jest dowolne, to $\varphi \in [0, +\infty[$,
3. współczynnik φ Yule'a może zawyżać wyniki, jeżeli liczba wierszy nie równa jest 2,
4. mierzy współzależność krzywoliniową.

Współczynnik φ Yule'a. II

Uwaga 5

1. *Współczynnikiem φ Yule'a służy do badania zależności cech niemierzalnych (skala nominalna).*
2. *Współczynnik określony w definicji 5 nazwany jest też w literaturze współczynnikiem φ i jako jego autora podaje się K. Pearsona*

Współczynnik kontyngencji T Czuprowa. I

Definicja 6

Współczynnik kontyngencji T Czuprowa jest postaci

$$T := \sqrt{\frac{\chi^2}{n\sqrt{(k-1)(l-1)}}}.$$

Własności współczynnika kontyngencji T Czuprowa:

1. jeżeli $k = l$, to $T \in [0, 1]$,
2. jeżeli $k \neq l$, to T może być znacznie mniejsze niż 1,
3. może on zaniżać wyniki, gdy tablica nie jest symetryczna,
4. nie można przy jego pomocy wskazać kierunku współzależności.

Współczynnik kontyngencji T Czuprowa. II

Uwaga 6

Współczynnik zbieżność T Czuprowa służy do badania zależności cech niemierzalnych (nominalnych).

Współczynnik kontyngencji C Pearsona. I

Definicja 7

Współczynnik kontyngencji C Pearsona jest postaci

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

Własności współczynnika kontyngencji C Pearsona:

1. może przyjmować (teoretycznie) wartości od 0 do 1,
2. kres górny wartości współczynnika faktycznie zależy jednak od wymiarów tablicy; im większa tablica, tym wyższa maksymalna wartość, którą może osiągnąć C,
3. może on zaniżać wyniki, gdy tablica jest mała (mniejsza niż 10×10),
4. mierzy współzależność krzywoliniową.

Spis treści

1. Dwuwymiarowy rozkład empiryczny
2. Test niezależności chi-kwadrat
3. Miary zależności cech oparte na statystyce chi-kwadrat
   - Współczynnik zbieżności V Cramera
   - Inne miary zależności cech jakościowych
4. Badanie zależności dwóch cech
   - Kowariancja rozkładu empirycznego – przypomnienie
   - Współczynnik korelacji liniowej Pearsona (cechy mierzalne)
   - Współczynnik korelacji rang Spearmana (skala porządkowa)
   - Empiryczne krzywe regresji
   - Stosunki korelacyjne

Współczynnika korelacji zmiennych losowych

Dla zmiennych losowych, całkowalnych z kwadratem i o niezerowej wariancji, współczynnik korelacji definiujemy równością

$$\rho(X, Y) := \frac{\operatorname{cov}(X, Y)}{\sqrt{\mathbb{D}^2(X)\mathbb{D}^2(Y)}}. \tag{4}$$

Uwaga 7

Z nierówności Schwarza wynika, że przyjmuje on wartości z przedziału $[-1, 1]$.
Ponadto zeruje się dla zmiennych niezależnych.

Kowariancja rozkładu empirycznego

Definicja 8

Kowariancją dwuwymiarowego rozkładu empirycznego, oznaczaną c_{xy}, nazywamy nieobciążony estymator kowariancji w populacji wyrażający się wzorem

$$\tilde{c}_{xy} := \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{l} (x_i - \overline{x})(y_j - \overline{y}) \cdot n_{ij}. \tag{5}$$

Uwaga 8

W przypadku danych indywidualnych (x_i, y_i) (dla $i \in \overline{1, n}$) kowariancją dwuwymiarowego rozkładu empirycznego wyraża się wzorem

$$\tilde{c}_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \tag{6}$$

Uwaga

1. Mamy następująca zależność
$$-\tilde{s}_X \tilde{s}_Y \leqslant \tilde{c}_{xy} \leqslant s_X s_Y, \tag{7}$$
gdzie s_X oraz s_Y są odchyleniami standardowymi zmiennych X i Y.
2. Definiując kowariancję wzorem
$$c_{xy} := \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{l} (x_i - \overline{x})(y_j - \overline{y}) \cdot n_{ij},$$
otrzymujemy
$$-s_X s_Y \leqslant c_{xy} \leqslant s_X s_Y.$$

Własności kowariancji rozkładu empirycznego

Kowariancję wykorzystujemy do oceny stopnia współzależności zmiennych.

1. Jeżeli kowariancja jest ujemna, to na ogół niskim wartościom jednej cechy odpowiadają wysokie wartości drugiej i na odwrót.
2. Jeżeli kowariancja jest dodatnia, to niskim (wysokim) wartościom jednej cechy odpowiadają niskie (wysokie) wartości drugiej.
3. Jeżeli kowariancja jest bliska zeru, to przy różnych wartościach jednej cechy poziom wartości drugiej pozostaje (w przybliżeniu) ten sam.

Współczynnik korelacji liniowej Pearsona

Definicja 9

Współczynnikiem korelacji liniowej Pearsona w rozkładzie empirycznym nazywamy wielkość

$$r_{xy} := \frac{c_{xy}}{s_X s_Y}, \tag{8}$$

gdzie c_{xy} jest kowariancją rozkładu empirycznego, s_X, s_Y są odchyleniami standardowymi w rozkładach brzegowych.

Uwaga

1. W definicji współczynnika korelacji liniowej Pearsona mogliśmy użyć $\tilde{c}_{xy}$, $\tilde{s}_X$ oraz $\tilde{s}_Y$.
2. Współczynnik korelacji liniowej Pearsona może być rozpatrywany jako estymator współczynnika korelacji ρ w populacji generalnej albo jako parametr rozkładu empirycznego w skończonej zbiorowości.
3. Służy jednocześnie zarówno do określenia siły zależności, jak też do wskazania jej kierunku

Własności współczynnika korelacji liniowej Pearsona

1. $r_{xy} \in [-1, 1]$;
2. $r_{xy} = r_{yx}$;
3. $r_{xy} = 0$, gdy cechy są liniowo nieskorelowane (nieskorelowane albo skorelowane nieliniowo);
4. $|r_{xy}| = 1$ wtedy i tylko wtedy, gdy związek między cechami jest funkcją liniową.

Uwaga 9

Należy podkreślić, że warunek zerowania się współczynnika korelacji może oznaczać bark korelacji, jak i korelację nieliniową.

Testowanie hipotezy o liniowej nieskorelowaności cech. I

Załóżmy, że dwuwymiarowy rozkład zmiennych losowa X i Y w populacji generalnej jest normalny. Na podstawie n-elementowej próby z tej populacji należy sprawdzić, że zmienne w populacji generalnej są liniowo nieskorelowane tzn. współczynnik korelacji ρ w populacji generalnej jest równy zeru.

Testujemy hipotezę

$$\begin{aligned} H_0 &: \quad \rho = 0, \\ H_1 &: \quad \rho \neq 0. \end{aligned}$$

Jeżeli prawdziwa jest hipoteza zerowa, to statystyka

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{n-2},$$

Testowanie hipotezy o liniowej nieskorelowaności cech. II

gdzie r_{xy} jest współczynnikiem korelacji z próby, ma rozkład t-Studenta o $n-2$ stopniach swobody.

Obszar krytyczny, dla poziomu istotności α, zadaje równość

$$P(\{|t| \geqslant t_\alpha\}) = \alpha.$$

Motywacja. I

Problem

Jak mierzyć zależność cech niemierzalnych?

Propozycja rozwiązania.

Można też wykorzystać opisany dokładanie przez Charlesa Spearmana współczynnik korelacji rang.

Uwaga 10

W rachunku prawdopodobieństwa mamy w takim wypadku doczynienie ze zmienna losową, która zdarzeniom przypisuje arbitralna wartość.

Gdy badane cechy niemierzalne mają charakter porządkowy, możliwe jest nadanie wariantom tych cech rang tzn. umownych liczbowych wartości (np. numerów miejsc w ciągu). Badanie

Motywacja. II

zależności między cechami niemierzalnymi może polegać wtedy na badaniu korelacji między rangami przyporządkowanymi wariantom tych cech tzn. na badaniu stopnia odpowiedniości między rangami.

Estymator nieuwzględniający rang wiązanych

Definicja 10

Niech a_i będzie rangą przyporządkowaną i-tej obserwacji z pierwszego ciągu oraz b_i będzie rangą przyporządkowaną i-tej obserwacji z drugiego ciągu. Jeżeli w zbiorze danych nie ma obserwacji powiązanych tzn. nie istnieje podzbiór obserwacji, których nie można uporządkować, to wtedy współczynnik korelacji rang Spearmana określa równość

$$r_d := 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{9}$$

gdzie $d_i = a_i - b_i$.

Własności współczynnika korelacji rang Spearmana. I

1. $r_d \in [-1, 1]$.
2. Jeżeli $r_d = 1$, to występuje idealna zgodność rang.
3. Jeżeli $r_d = -1$, to występuje maksymalna niezgodność rang (najwyższej randze w jednym ciągu odpowiada najniższa ranga w drugim).
4. Jeżeli $r_d = 0$, to rangi w obu ciągach są niezależne (losowe kojarzenie rang w obu ciągach).

Nieoczekiwane własności współczynnika korelacji rang Spearmana

1. nie jest wówczas prawdą, iż $r_d(X, -Y) = -r_d(X, Y)$,
2. nie jest wtedy zgodny z pierwotną definicją korelacji rang Spearmana jako zwykłego współczynnika korelacji liczonego dla rang,

Własności współczynnika korelacji rang Spearmana. II

3. dla zmiennych dyskretnych, minimalną wartością jego granicy, przy rozmiarze próby dążącym do nieskończoności, jest $\frac{2}{(\min(m_X, m_Y))^2} - 1$, gdzie m_X to liczba różnych wartości przyjmowanych przez zmienną X, a m_Y liczba różnych wartości zmiennej Y.
4. Wynika stąd, że estymator ten jest dla zmiennych dyskretnych niezgodny i asymptotycznie obciążony.

Ogólna postać estymatora oparta na różnicy rang. I

Sposób prezentacji i materiał przedstawiony tutaj został zaczerpnięty z [2] i [1].

W oryginalnym ujęciu Spearmana, jego korelacja rang jest współczynnikiem korelacji Pearsona liczonym dla rang zmiennych.

$$r_S = \rho(RX, RY),$$

gdzie

ρ	-	klasyczny współczynnik korelacji,
RX	-	rangi zmiennej X w próbie,
RY	-	rangi zmiennej Y w próbie.

Ogólna postać estymatora oparta na różnicy rang. II

Ten sam estymator można też zapisać w innej, równoważnej wersji

$$r_S = \frac{\frac{1}{6}(n^3 - n) - \left(\sum_{i=1}^{n} d_i^2\right) - T_X - T_Y}{\sqrt{\frac{1}{6}(n^3 - n) - 2T_X)\frac{1}{6}(n^3 - n) - 2T_Y)}},$$

gdzie

$$\begin{aligned} d_i &= Rx_i - Ry_i, \\ T_X &= \frac{1}{12}\sum_j (t_j^3 - t_j), \\ T_Y &= \frac{1}{12}\sum_k (u_k^3 - u_k), \end{aligned}$$

Ogólna postać estymatora oparta na różnicy rang. III

natomiast t_j jest liczbą obserwacji w próbie posiadających tę samą j-tą wartość rangi zmiennej X, u_k liczbą obserwacji w próbie posiadających tę samą k-tą wartość rangi zmiennej Y, a sumowanie przebiega po wszystkich wartościach rang – wystarczy zsumować rangi wiązane, bo dla pozostałych $t_j^3 - t_j = 1^3 - 1 = 0$ (analogicznie u_k), gdy w danej zmiennej nie ma rang wiązanych, T_X lub T_Y jest równe zeru.

Właściwości estymatora

1. Współczynnik jest unormowany tzn. $|r_S(X, Y)| \leqslant 1$.
2. Im bardziej wartości oddalone są od zera, tym większa siła związku między zmiennymi.
3. Gdy każda zmienna jest ściśle rosnącą funkcją drugiej, to występuje idealna zgodność rang i ich korelacja przyjmuje wartość 1. W szczególności wartość ta jest przyjmowana, gdy zmienna jest korelowana sama ze sobą tzn. $r_S(X, X) = 1$.

Ogólna postać estymatora oparta na różnicy rang. IV

4. Gdy każda zmienna jest ściśle malejącą funkcją drugiej zmiennej, występuje maksymalna niezgodność rang i ich korelacja przyjmuje wartość -1. W szczególności wartość ta jest przyjmowana, gdy zmienna X korelowana jest z $-X$ to

$$r_S(X, -X) = -1. \tag{10}$$

5. Dla niezależnych zmiennych losowych wartością oczekiwaną estymatorów jest 0, a rozkład każdego z nich nie zależy od rozkładu zmiennych przed rangowaniem;
6. Zachodzi symetria ze względu na zamianę zmiennych: $r_S(X, Y) = r_S(Y, X)$.
7. Zachodzi symetria ze względu na zmianę znaku zmiennej tzn.

$$r_S(X, Y) = -r_S(X, -Y) = -r_S(-X, Y). \tag{11}$$

W przypadku wystąpienia rang wiązanych część z tych właściwości nie jest spełniona dla niektórych estymatorów [zob. 2]. Dla estymatora określonego wzorem (9) nie są prawdziwe własności (10) i (11).

Testowanie hipotezy o współczynniku korelacji rang Spearmana. I

Załóżmy, że mamy dwie zmienne losowa X i Y w populacji generalnej. Na podstawie n-elementowej próby z tej populacji należy sprawdzić, że rangi są nieskorelowane tzn. współczynnik korelacji rang Spearmana jest równy zeru.

Testujemy hipotezę

$$\begin{aligned} H_0: &\quad r_S = 0, \\ H_1: &\quad r_S \neq 0. \end{aligned}$$

Jeżeli prawdziwa jest hipoteza zerowa i próba jest mała, to statystyka

$$t = \frac{r_S}{\sqrt{1 - r_S^2}} \sqrt{n-2}$$

Testowanie hipotezy o współczynniku korelacji rang Spearmana. II

ma rozkład t-Studenta o $n-2$ stopniach swobody. Obszar krytyczny, dla poziomu istotności α zadawany jest tutaj standardowo.

Jeżeli natomiast próba jest duża i hipoteza zerowa jest prawdziwa, to r_S ma rozkład normalny $\mathcal{N}(0, \frac{1}{\sqrt{n-1}})$ i należy rozważyć statystykę

$$u = r_S\sqrt{n-1},$$

która ma rozkład normalny standardowy.

Motywacja

Do oceny współzależności zmiennych może być użyta analiza rozkładów warunkowych zmiennych określonych w tablicy korelacyjnej.

Porównanie średnich warunkowych

1. Jeżeli $\overline{x_1} = \ldots = \overline{x_l} = \overline{x}$, to zmienna Y nie wpływa na zmienną X.
2. Jeżeli $\overline{y_1} = \ldots = \overline{y_k} = \overline{y}$, to zmienna X nie wpływa na zmienną Y.

Jeśli cechy są skorelowane, to średnie warunkowe zmiennej uznanej za zależną będą różne. Zależność jest tym silniejsza, im mocniej różne wartości przyjmowane przez cechę niezależną różnicują średni poziom wartości cechy zależnej.

Uwaga 11

Średnie warunkowe cechy zależnej możemy traktować jako funkcje wartości cechy niezależnej (funkcje regresji I rodzaju).

Warunkowe wartości oczekiwane – zmienne dyskretne. I

Definicja 11

Warunkową wartością oczekiwaną zmiennej X pod warunkiem $\{Y = y_j\}$, oznaczaną $\mathbb{E}(X|Y = y_j)$, jest liczba wyrażona wzorem

$$\sum_{i=1}^{k} x_i P(X = x_i|Y = y_j).$$

Warunkową wartością oczekiwaną zmiennej Y pod warunkiem $\{X = x_i\}$, oznaczaną $\mathbb{E}(Y|X = x_i)$, jest liczba wyrażona wzorem

$$\sum_{j=1}^{l} y_j P(Y = y_j|X = x_i).$$

Warunkowe wartości oczekiwane – zmienne dyskretne. II

Twierdzenie 1

Mamy

$$\mathbb{E}(X|Y = y_j) = \sum_{i=1}^{k} x_i p_{i|j}.$$

oraz

$$\mathbb{E}(Y|X = x_i) = \sum_{j=1}^{l} y_j p_{j|i}.$$

Warunkowe wartości oczekiwane – zmienne ciągłe. I

Definicja 12

Warunkową wartością oczekiwaną zmiennej X pod warunkiem $\{Y = y\}$, oznaczaną $\mathbb{E}(X|Y = y)$, jest liczba wyrażona wzorem

$$\int_{-\infty}^{+\infty} xf(x|y)dx.$$

Warunkową wartością oczekiwaną zmiennej Y pod warunkiem $\{X = x\}$, oznaczaną $\mathbb{E}(Y|X = x)$, jest liczba wyrażona wzorem

$$\int_{-\infty}^{+\infty} yf(y|x)dy.$$

Warunkowe wartości oczekiwane – zmienne ciągłe. II

Uwaga 12

Przypomnijmy definicję gęstości warunkowej:
Niech dwuwymiarowa zmienna losowa (X, Y) ma rozkład ciągły z gęstością f tzn. f jest funkcją dwóch zmiennych x i y. Gęstością rozkładu warunkowego X pod warunkiem $Y = y$ nazywamy funkcję określoną dla $x \in \mathbb{R}$ wzorem

$$f(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)} & gdy\ f_Y(y) > 0 \\ 0 & w\ p.p. \end{cases}$$

gdzie $f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx$ jest gęstością rozkładu brzegowego Y.
Analogicznie definiujemy gęstością rozkładu warunkowego Y pod warunkiem $X = x$.

Funkcje regresji I rodzaju. I

Definicja 13

Funkcją regresji I rodzaju zmiennej X względem zmiennej Y nazywamy warunkową wartość oczekiwaną zmiennej X jako funkcję wartości zmiennej Y wyrażoną wzorem

$$m_1(y) := \mathbb{E}(X|Y = y_j) \text{ dla zmiennej } Y \text{ dyskretnej}, \tag{12}$$

$$m_1(y) := \mathbb{E}(X|Y = y) \text{ dla zmiennej } Y \text{ ciągłej}. \tag{13}$$

Definicja 14

Funkcją regresji I rodzaju zmiennej Y względem zmiennej X nazywamy warunkową wartość oczekiwaną zmiennej Y jako funkcję wartości zmiennej X wyrażoną wzorem

$$m_2(x) := \mathbb{E}(Y|X = x_i) \text{ dla zmiennej } X \text{ dyskretnej}, \tag{14}$$

$$m_2(x) := \mathbb{E}(Y|X = x) \text{ dla zmiennej } X \text{ ciągłej}. \tag{15}$$

Funkcje regresji I rodzaju. II

Uwaga 13

1. *Empiryczna krzywa regresji cechy X względem cechy Y jest to łamana łącząca punkty $(\overline{x_j}, y_j)$ dla $j = 1, \ldots, l$.*
2. *Empiryczna krzywa regresji cechy Y względem cechy X jest to łamana łącząca punkty $(x_i, \overline{y_i})$ dla $i = 1, \ldots, k$.*

Stosunki korelacyjne zmiennych

Jeżeli regresja zmiennych jest nieliniowa, to do pomiaru siły zależności cech wykorzystuje się tzw. stosunki (wskaźniki) korelacyjne wykorzystujące funkcję regresji.

Wykorzystując w tym celu zależności

$$\begin{aligned}\mathbb{E}((Y-\mathbb{E}(Y))^2) &= \mathbb{E}((m_2(x)-\mathbb{E}(Y))^2)+\mathbb{E}((Y-m_2(x))^2),\\ \mathbb{E}((X-\mathbb{E}(X))^2) &= \mathbb{E}((m_1(y)-\mathbb{E}(X))^2)+\mathbb{E}((X-m_1(y))^2).\end{aligned}$$

mamy

$$\begin{aligned}\eta_{xy} &:= \sqrt{\frac{\mathbb{E}(m_1(y)-\mathbb{E}(X))^2}{\mathbb{D}^2(X)}}=\sqrt{1-\frac{\mathbb{E}(X-m_1(y))^2}{\mathbb{D}^2(X)}},\\ \eta_{yx} &:= \sqrt{\frac{\mathbb{E}(m_2(x)-\mathbb{E}(Y))^2}{\mathbb{D}^2(Y)}}=\sqrt{1-\frac{\mathbb{E}(Y-m_2(x))^2}{\mathbb{D}^2(Y)}}.\end{aligned}$$

Motywacja

W celu stwierdzenia, czy istnieje zależność między cechami i określenia jej siły dla zmiennej uznanej za zależną można przeprowadzić analizę wariancji.

Dla zmiennej X.

Dokonujemy podziału całkowitej sumy kwadratów odchyleń od średniej na sumę kwadratów międzygrupową i wewnątrz grupową tj.

$$\sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot n_{i\cdot} = \sum_{j=1}^{l}(\overline{x_j} - \overline{x})^2 \cdot n_{\cdot j} + \sum_{j=1}^{l}\sum_{i=1}^{k}(x_i - \overline{x_j})^2 \cdot n_{ij}, \tag{16}$$

gdzie $\overline{x}$ średnia w brzegowym rozkładzie cechy X, $\overline{x_j}$ średnie w warunkowych rozkładach cechy X.

Pierwszy składnik sumy po prawej stronie jest to zróżnicowanie wyjaśniające regresje, a drugi to zróżnicowanie nie wyjaśniane regresją.

Dla zmiennej Y.

$$\sum_{j=1}^{l}(y_j - \overline{y})^2 \cdot n_{\cdot j} = \sum_{i=1}^{k}(\overline{y_i} - \overline{y})^2 \cdot n_{i\cdot} + \sum_{i=1}^{k}\sum_{j=1}^{l}(y_j - \overline{y_i})^2 \cdot n_{ij}. \tag{17}$$

Uwaga 14

Zróżnicowanie nie wyjaśniane regresją reprezentuje rozrzut indywidualnych wartości zmiennej uznawanej za zależną wokół empirycznej linii regresji.

Definicja stosunków korelacyjnych. I

Definicja 15

Wskaźnikiem (stosunkiem) korelacyjnym zmiennej zależnej X względem zmienne Y nazywamy liczbę e_{xy} taką, że

$$e_{xy}^2 := \frac{\sum_{j=1}^{l} (\overline{x_j} - \overline{x})^2 \cdot n_{.j}}{\sum_{i=1}^{k} (x_i - \overline{x})^2 \cdot n_{i.}} = 1 - \frac{\sum_{j=1}^{l} \sum_{i=1}^{k} (x_i - \overline{x}_j)^2 n_{ij}}{\sum_{i=1}^{k} (x_i - \overline{x})^2 \cdot n_{i.}}. \tag{18}$$

Definicja stosunków korelacyjnych. II

Definicja 16

Wskaźnikiem (stosunkiem) korelacyjnym zmiennej zależnej Y względem zmienne X nazywamy liczbę e_{yx} taką, że

$$e_{yx}^2 := \frac{\sum_{i=1}^{k} (\overline{y_i} - \overline{y})^2 \cdot n_{i\cdot}}{\sum_{j=1}^{l} (y_j - \overline{y})^2 \cdot n_{\cdot j}} = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} (y_j - \overline{y}_i)^2 n_{ij}}{\sum_{j=1}^{l} (y_j - \overline{y})^2 \cdot n_{\cdot j}}. \tag{19}$$

Uwaga 15

1. *Ponieważ $e_{xy}^2 \in [0,1]$ i $e_{yx}^2 \in [0,1]$, więc $e_{xy}, e_{yx} \in [0,1]$.*
2. *Stosunki korelacyjne e_{xy} i e_{yx} nie muszą być równe.*

Testowanie hipotez o niezależności jednej zmiennej od drugiej. I

Będziemy weryfikować hipotezę, że zmienna losowa X w populacji generalnej jest niezależna od zmiennej Y.

Hipotezę traktuje się, jako przypuszczenie, że stosunek korelacyjny zmiennej X względem zmiennej Y w populacji generalnej równe są zero.

Mamy więc

$$\begin{aligned} H_0: &\quad \eta_{xy} = 0 \\ H_1: &\quad \eta_{xy} > 0. \end{aligned}$$

W celu weryfikacji tej hipotezy stosowana jest statystyka

$$F = \frac{\frac{e_{xy}^2}{l-1}}{\frac{1-e_{xy}^2}{n-l}},$$

Testowanie hipotez o niezależności jednej zmiennej od drugiej. II

gdzie n jest liczebnością próby, l liczba warunkowych rozkładów zmiennej X, e_{xy} stosunek korelacyjny z próby.

Jeżeli hipoteza zerowa jest prawdziwa, to statystyka F ma rozkład Fishera-Snedecora o $l-1$ stopniach swobody licznika i $n-l$ stopniach swobody mianownika.

Dla poziomu istotności α obszar krytyczny określony jest przez zależność

$$P(\{F \geqslant F_\alpha\}) = \alpha.$$

Testowanie hipotez o niezależności jednej zmiennej od drugiej. III

Uwaga 16

W przypadku weryfikowani hipotezy, że zmienna losowa Y w populacji generalnej jest niezależna od zmiennej X mamy następujące hipotezy (zerowa i alternatywna):

$$\begin{aligned} H_0: &\quad \eta_{yx} = 0, \\ H_1: &\quad \eta_{yx} > 0. \end{aligned}$$

A w celu weryfikacji tej hipotezy stosowana jest statystyka

$$F = \frac{\frac{e_{yx}^2}{k-1}}{\frac{1-e_{yx}^2}{n-k}},$$

gdzie n jest liczebnością próby, k liczba warunkowych rozkładów zmiennej Y, e_{yx} stosunek korelacyjny z próby.

Miara siły korelacji nieliniowej

W przypadku nieliniowej zależności wartość współczynnika korelacji jest niższa niż powinno to wynikać z siły związku między cechami.

Miarą siły korelacji nieliniowej są stosunki korelacyjne:

$$\widehat{m}_{xy} = e_{xy}^2 - r^2 \tag{20}$$

$$\widehat{m}_{yx} = e_{yx}^2 - r^2, \tag{21}$$

gdzie $\widehat{m}_{xy}, \widehat{m}_{yx} \in [0, 1]$.

Im bliższe jedności są wartości tych wskaźników, tym bardziej związki między cechami odchylają się od zależności liniowej.

Bibliografia

[1] *Spearman's rank correlation coefficient*. Wikipedia. 2020. URL: https://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient (term. wiz. 15. 03. 2020).

[2] *Współczynnik korelacji rang Spearmana*. Wikipedia. 2020. URL: https://pl.wikipedia.org/wiki/Wsp%C3%B3%C5%82czynnik_korelacji_rang_Spearmana (term. wiz. 15. 03. 2020).