TABLICA KONTYNGENCJI

X_i / Y_j	Y_1	Y_2	...	Y_k	$n_{i\bullet}$
X_1	n_{11}	n_{12}		n_{1k}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	...	n_{2k}	$n_{2\bullet}$
⋮					
X_r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet k}$	n

Rozkłady brzegowe:

zmiennej X, bez względu na to, jakie wartości przyjmuje zmienna Y - pierwsza i ostatnia kolumna (łącznie);

zmiennej Y, bez względu na to, jakie wartości przyjmuje zmienna X - pierwszy i ostatni wiersz (łącznie);

Rozkłady warunkowe:

Struktura wartości jednej zmiennej, pod warunkiem, że druga ze zmiennych przyjmuje określoną wartość

rozkład warunkowy zmiennej $X|Y=y_j$ - tzn. pierwsza i j-ta kolumna (łącznie)

rozkład warunkowy zmiennej $Y|X=x_i$ - tzn. pierwszy i ostatni wiersz (łącznie),

STATYSTYKA CHI-KWADRAT

X_i / Y_j	Y_1	Y_2	...	Y_k	$n_{i\bullet}$	$p_{i\bullet}$
X_1	n_{11}	n_{12}		n_{1k}	$n_{1\bullet}$	$p_{1\bullet}$
X_2	n_{21}	n_{22}	...	n_{2k}	$n_{2\bullet}$	$p_{2\bullet}$
⋮						
X_r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r\bullet}$	$p_{r\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet k}$	n	1
$p_{\bullet j}$	$p_{\bullet 1}$	$p_{\bullet 2}$		$p_{\bullet k}$	1	

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

Liczebności teoretyczne oblicza się według wzoru:

$$\hat{n}_{ij} = n \cdot p_{ij}$$

gdzie:

$$p_{ij} = p_{i\bullet} p_{\bullet j} \qquad p_{i\bullet} = \frac{n_{i\bullet}}{n} \qquad p_{\bullet j} = \frac{n_{\bullet j}}{n}$$

TABLICA CZTEROPOLOWA

Dla tablicy o wymiarach 2 x 2 postaci:

$X \;/\; Y$	Y_1	Y_2	$n_{i.}$
X_1	a	b	$a+b$
X_2	c	d	$c+d$
$n_{.j}$	$a+c$	$b+d$	n

statystykę χ^2 wyznaczyć można z następującego wzoru :

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

W przypadku liczebności poszczególnych komórek $n_{ij}<=8$ stosujemy wzór uwzględniający tzw. poprawkę Yatesa:

$$\chi_Y^2 = \frac{n(|ad - bc| - \frac{1}{2}n)^2}{(a + b)(c + d)(a + c)(b + d)}$$

Współczynnik zbieżności Czuprowa

miernik współzależności oparty na statystyce χ^2

χ^2 - przyjmuje wartości z przedziału: $[0; n\sqrt{(r-1)(k-1)}]$;

stąd wartość :

χ^2=0 występuje gdy wszystkie liczebności teoretyczne są równe empirycznym, tzn. $n_{ij} - \hat{n}_{ij} = 0$ **dla każdego *ij*, występuje wówczas stochastyczna niezależność badanych zmiennych;**

$\chi^2 = n\sqrt{(r-1)(k-1)}$ oznacza występowanie zależności funkcyjnej.

Standaryzacja wielkości χ^2 prowadzi do otrzymania kwadratu współczynnika zbieżności Czuprowa:

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

$T \in \langle 0;1 \rangle$

$T = 0$ stochastyczna niezależność

$T = 1$ zależność funkcyjna

Własności współczynnika zbieżności:

- dotyczy cech mierzalnych i niemierzalnych

- symetryczność

- nie wskazuje kierunku zależności między zmiennymi.

- T^2 * 100% - współczynnik determinacji

ZWIĄZEK CECH NIEMIERZALNYCH

- ustalenie skojarzeń, asocjacji, kontyngencji;

- obie, lub przynajmniej jedna z cech mają charakter jakościowy;

Współczynnik Yule'a

- każda z cech ma dwa warianty

- zależność

$$\varphi^2 \quad od \quad \chi^2$$

$$\chi^2 = n\varphi^2 \quad \text{lub} \quad \varphi^2 = \frac{\chi^2}{n} \qquad \varphi = \sqrt{\frac{\chi^2}{n}}$$

wzór bezpośredni:

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

teoretycznie: $\varphi \in \langle -1;1 \rangle$ **,**

$\varphi = 0$ **- występuje niezależność zmiennych,**

$\varphi = 1$ lub $\varphi = -1$ **zachodzi tylko wtedy, gdy: *a=d=0* lub *b=c=0*,**

w innych przypadkach φ **nie osiąga wartości krańcowych nawet przy bardzo silnych związkach**

Dlatego konkretną wartość odnosi się je do φ_{min} **i** φ_{max}**.**

Ponieważ ustalenie wartości φ_{min} i φ_{max} jest kłopotliwe, oblicza się wartość skorygowanego współczynnika według wzoru Cole'a :

$$\varphi_{kor} = \frac{ad - bc}{n \cdot \min(b,c) + (ad - bc)} \quad dla \quad \varphi \geq 0$$

$$\varphi_{kor} = \frac{ad - bc}{n \cdot \min(a,d) - (ad - bc)} \quad dla \quad \varphi < 0$$

- *min (b,c)* oznacza, że w obliczeniach należy uwzględnić mniejszą z liczb *b* i *c*;

- znak współczynnika φ nie informuje o kierunku zależności między zmiennymi, gdyż zależy on od sposobu uporządkowania wariantów cech w tablicy.

Współczynnik V - Cramera

$$V \in \langle 0,1 \rangle$$

$$V = \sqrt{\frac{\chi^2}{n \cdot \min[(r-1),(k-1)]}} = \sqrt{\frac{\varphi^2}{\min[(r-1),(k-1)]}}$$

Współczynnik kontyngencji C - Pearsona

- może być stosowany przy tablicach wielodzielnych (prostokątnych bądź kwadratowych)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{\varphi^2}{1 + \varphi^2}}$$

$$C \in \langle 0;1 \rangle$$

Kres górny współczynnika *C* zależy od liczby wierszy i kolumn, stąd wartość *C* należy odnosić do wartości maksymalnej dla danej tablicy C_{max}:

- dla tablicy kwadratowej, gdzie *k* - oznacza liczbę kolumn:

$$C_{\max} = \sqrt{\frac{k-1}{k}}$$

- dla tablicy prostokątnej:

$$C_{\max} = \frac{\sqrt{\frac{k-1}{k}} + \sqrt{\frac{r-1}{r}}}{2}$$

Współczynnik skorygowany oblicza się według wzoru:

$$C_{kor} = \frac{C}{C_{\max}}$$

Znając wartość *C* można wyznaczyć wartość φ :

$$\varphi = \frac{C}{\sqrt{1 - C^2}}$$

Przykład 1:

Określ, czy pomiędzy stosunkiem studentów do transmisji sportowych, a ich czynnym uczestnictwem w uprawianiu sportu istnieje współzależność. Jeśli tak, proszę określić jej siłę.

H_0: $P\{X=x_i, Y=y_j\} = P\{X=x_i\}\, P\{Y=y_j\}$

H_1: $P\{X=x_i, Y=y_j\} \neq P\{X=x_i\}\, P\{Y=y_j\}$

Uprawianie sportu (Y)	Opinie o transmisjach TV (X) atrakcyjne	nieatrakcyjne	$n_{i\bullet}$
nie uprawiają	65	17	82
uprawiają	24	14	38
$n_{\bullet j}$	89	31	120

$$\chi^2 = \frac{120\cdot(65\cdot 14 - 17\cdot 24)^2}{89\cdot 31\cdot 82\cdot 38} =$$

$$= \frac{120\cdot(910 - 408)^2}{8597044} = \frac{30240480}{8597044} = 3.52$$

α=0.1 , $\chi^2_{0.1;1}$=2.706 $\chi^2 > \chi^2_{0.1;1}$ stąd:
H_0 należy odrzucić.

Współczynnik zbieżności Czuprowa T:

$$T^2 = \frac{\chi^2}{n\sqrt{(r-1)(k-1)}} = \frac{3.52}{120\cdot 1} = 0.029$$

$$T = 0.17$$

Współczynnik Yule'a:

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{3.52}{120}} = 0.17$$

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} = \frac{910 - 408}{\sqrt{82 \cdot 38 \cdot 89 \cdot 31}} =$$

$$= \frac{502}{2932.072} = 0.17$$

$$dla \quad \varphi \geq 0$$

$$\varphi_{kor} = \frac{ad - bc}{n \cdot \min(b,c) + (ad - bc)} = \frac{910 - 408}{120 \cdot 17 + 502} =$$

$$= \frac{502}{2040 + 502} = \frac{502}{2542} = 0.197 \approx 0.2$$

Współczynnik V - Cramera:

$$V = \sqrt{\frac{\varphi^2}{\min[(r-1),(k-1)]}} = \sqrt{\varphi^2} = 0.2 \quad (0.17)$$

Współczynnik kontyngencji C - Pearsona:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{3.52}{3.52 + 120}} = \sqrt{0.028} = 0.1688 \approx 0.17$$

$$C_{max} = \sqrt{\frac{2-1}{2}} = \sqrt{\frac{1}{2}} = 0.707$$

$$C_{kor} = \frac{0.1688}{0.707} = 0.2387 \approx 0.24$$

Przykład 2:

Relację pomiędzy poziomem wykształcenia i czasem poszukiwania pracy przez osoby bezrobotne w trzecim kwartale 2001 r. przedstawiono w tabeli

Bezrobocie	Poziom wykształcenia		
	Podstawowe	Średni	Wyższy
Krótko-okresowe	264	205	56
Średnio-okresowe	676	373	38
Długo-okresowe	710	328	23

Źródło: Aktywność Ekonomiczna ludności Polski III kwartał 2000, GUS, Warszawa

Formułujemy hipotezę zerową i alternatywną:

$$H_0: \quad P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}$$
$$H_1: \quad P\{X = x, Y = y\} \neq P\{X = x\}P\{Y = y\}$$

Bezrobocie	Poziom wykształcenia			$n_{i\bullet}$	$p_{i\bullet}$
	Podstawowe	Średni	Wyższy		
Krótko-okresowe	0,122 **264** 325	0,067 **205** 178	0,008 **56** 21	**525**	0,197
Średnio-okresowe	0,250 **676** 667	0,138 **373** 368	0,016 **28** 43	**1077**	0,405
Długo-okresowe	0,247 **710** 658	0,135 **328** 360	0,016 **23** 43	**1061**	0,398
$n_{\bullet j}$	**1650**	**906**	**107**	**2663**	
$p_{\bullet j}$	0,620	0,340	0,04		1

Wartość statystyki χ^2 można obliczyć w formie tabelarycznej:

n_{ij}	np_{ij}	$(n_{ij} - np_{ij})^2 / np_{ij}$
264	325	11,45
676	667	0,12
710	658	4,11
205	178	4,10
373	368	0,07
328	360	2,84
56	21	58,3
28	43	5,23
23	43	9,3
		95,52

Tak więc wartość statystyki $\chi^2 = 95{,}52$.

chi-kwadrat= 95,52

wymiar tablicy kontyngencji 3 x 3

Współczynnik Czuprowa

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

T^2= 0,0179 T= 0,1339

Współczynnik V Cramera

$$V \in \langle 0,1 \rangle$$

$$V = \sqrt{\frac{\chi^2}{n \cdot \min[(r-1),(k-1)]}} = \sqrt{\frac{\varphi^2}{\min[(r-1),(k-1)]}}$$

V= 0,2678

Współczynnik kontyngencji Pearsona

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}}$$

$$C \in \langle 0;1 \rangle$$

C= 0,1861

$$C_{max} = \sqrt{\frac{k-1}{k}}$$ 0,816

$$C_{kor} = \frac{C}{C_{max}}$$

C_{kor}= 0,2279

X_i / Y_j	Y_1	Y_2	...	Y_k	$n_{i\bullet}$
X_1	n_{11}	n_{12}		n_{1k}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	...	n_{2k}	$n_{2\bullet}$
⋮					
X_r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet k}$	n

CHARAKTERYSTYKI ROZKŁADÓW BRZEGOWYCH I WARUNKOWYCH

- Średnie arytmetyczne rozkładów brzegowych

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} x_i n_{i\bullet} \qquad \bar{y} = \frac{1}{n}\sum_{j=1}^{k} y_j n_{\bullet j}$$

- Średnie arytmetyczne rozkładów warunkowych

$$\bar{x}_j = \frac{1}{n_{\bullet j}}\sum_{i=1}^{r} x_i n_{ij} \qquad \bar{y}_i = \frac{1}{n_{i\bullet}}\sum_{j=1}^{k} y_j n_{ij}$$

- Wariancje rozkładów brzegowych

$$S(x)^2 = \frac{\sum_{i=1}^{r} (x_i - \bar{x})^2 n_{i\bullet}}{n} \qquad S(y)^2 = \frac{\sum_{j=1}^{k} (y_j - \bar{y})^2 n_{\bullet j}}{n}$$

- Wariancje rozkładów warunkowych

$$S_j(x)^2 = \frac{\sum_{i=1}^{r} (x_i - \bar{x}_j)^2 n_{ij}}{n_{\bullet j}} \qquad S_i(y)^2 = \frac{\sum_{j=1}^{k} (y_j - \bar{y}_i)^2 n_{ij}}{n_{i\bullet}}$$

NIEZALEŻNOŚĆ ZMIENNYCH

- **Zależność stochastyczna dwóch zmiennych losowych polega na tym, że zmiana jednej z nich zmienia rozkład prawdopodobieństwa drugiej zmiennej**
  (warunkowy rozkład jednej zmiennej losowej zależy od wartości, jaką przyjmuje druga)
- Cecha ***X*** jest stochastycznie niezależna od cechy ***Y*** gdy:

$$\bar{x}_1 = \bar{x}_2 = \ldots \bar{x}_k \qquad \text{oraz} \qquad s_1^2(x) = s_2^2(x) = \ldots = s_k^2(x)$$

- Cecha ***Y*** jest stochastycznie niezależna od cechy ***X*** gdy:

$$\bar{y}_1 = \bar{y}_2 = \ldots \bar{y}_r \qquad \text{oraz} \qquad s_1^2(y) = s_2^2(y) = \ldots = s_r^2(y)$$

- **Zależność korelacyjna dwóch zmiennych losowych polega na tym, że zmiana jednej z nich pociąga za sobą zmianę średnich wartości drugiej cechy**
- Cecha ***X*** jest korelacyjnie niezależna od cechy ***Y*** gdy:

$$\bar{x}_1 = \bar{x}_2 = \ldots \bar{x}_k$$

- Cecha ***Y*** jest korelacyjnie niezależna od cechy ***X*** gdy:

$$\bar{y}_1 = \bar{y}_2 = \ldots \bar{y}_r$$

- **Zależność funkcyjna dwóch zmiennych losowych polega na tym, że zmiana wartości jednej z nich pociąga za sobą ściśle określoną zmianę wartości drugiej cechy**

WSKAŹNIKI (STOSUNKI KORELACYJNE)

Równość wariancyjna

$$S^2(y) = \overline{S_i^2(y)} + S^2(\bar{y}_i) \qquad \text{oraz} \quad S^2(x) = \overline{S_j^2(x)} + S^2(\bar{x}_j)$$

$S^2(x) \quad S^2(y) \rightarrow$ miary ogólnego zróżnicowania cech X i Y

$$S^2(y) = \frac{1}{n}\sum_i (y_i - \bar{y})^2 n_{i\cdot} \qquad S^2(x) = \frac{1}{n}\sum_j (x_j - \bar{x})^2 n_{\cdot j}$$

Dla zależności *Y(X)*

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

Stosunek korelacyjny zmiennej Y względem X

$$e_{y(x)} = \sqrt{1 - \frac{\overline{S_i^2(y)}}{S^2(y)}} = \sqrt{\frac{S^2(\bar{y}_i)}{S^2(y)}}$$

Dla zależności *X(Y)*

$$\overline{S^2(x_j)} = \frac{1}{n}\sum_{j=1}^{r} S_j^2(x) n_{.j} \qquad S^2(\bar{x}_j) = \frac{1}{n}\sum_{j=1}^{r} (\bar{x}_j - \bar{x})^2 n_{.j}$$

Stosunek korelacyjny zmiennej X względem Y

$$e_{x(y)} = \sqrt{1 - \frac{\overline{S_j^2(x)}}{S^2(x)}} = \sqrt{\frac{S^2(\bar{x}_j)}{S^2(x)}}$$

Własności:

- wskaźnik korelacyjny nie jest symetryczny, tzn. $e_{y(x)} \neq e_{x(y)}$, poza sytuacją niezależności cech lub związku funkcyjnego,
- przyjmuje wartości <0;1>,
- nie wskazuje kierunku korelacji dwóch cech,
- w przypadku stochastycznej niezależności dwóch cech $e_{y(x)} = 0$, natomiast w przypadku związku funkcyjnego $e_{y(x)} = 1$
- przynajmniej cecha zależna musi być mierzalna,
- może być stosowany zarówno w przypadku związków korelacyjnych liniowych jak i nieliniowych.

WSPÓŁCZYNNIK KORELACJI RANG SPEARMANA

- uszeregowanie badanych jednostek według kryterium porządkującego, niezależnie ze względu na badane cechy

- nadanie rang wszystkim badanym jednostkom, tzn. numerów miejsc zajmowanych przez badane jednostki w ciągu uporządkowanych ze względu na badane cechy

- obliczenie różnic pomiędzy rangami przyporządkowanymi poszczególnym badanym jednostkom w obu ciągach

- obliczenie współczynnika korelacji rang:

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

gdzie:
R_{Xi}, R_{Yi} – rangi nadane *i*-tej badanej jednostce w poszczególnych, uporządkowanych ciągach:

$d_i = R_{Xi} - R_{Yi}$

Własności:

- umożliwia ocenę zarówno siły jak i kierunku zależności pomiędzy cechami niemierzalnymi
- przyjmuje wartości z przedziału <-1,1>
- jest miarą symetryczną

WSPÓŁCZYNNIK KORELACJI LINIOWEJ PEARSONA

$$r(yx) = \frac{\mathrm{cov}(yx)}{S(y)S(x)}$$

gdzie:

$$\mathrm{cov}(yx) = \frac{1}{n}\sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y})n_{ij}$$

Własności:

- jest symetryczny, tzn. $r(yx) = r(xy)$,
- przyjmuje wartości z przedziału <-1; 1>,
- charakteryzuje zarówno kierunek jak i siłę zależności dwóch cech
- ma zastosowanie wyłącznie gdy związek dwóch cech ma charakter liniowy,
- może być wyznaczany wyłącznie w przypadku cech mierzalnych.