Badanie zależności między cechami

Obserwujemy dwie cechy: X oraz Y
Obiekt $\longrightarrow (X, Y)$

$$H_0 : \text{Cechy } X \text{ oraz } Y \text{ są niezależne}$$

Próba: $(X_1, Y_1), \ldots, (X_n, Y_n)$

Cechy X, Y są dowolnego typu:
Test Chi–Kwadrat niezależności

Łączny rozkład cech X, Y jest normalny:
Test współczynnika korelacji Pearsona

Cechy X, Y są typu ciągłego:
Test współczynnika korelacji rangowej Spearmana

Test współczynnika korelacji rangowej Kendalla

Test Chi–Kwadrat niezależności
(poziom istotności α)

Klasy cechy Y	Klasy cechy X			
	1	2	...	m
1	n_{11}	n_{12}	...	n_{1m}
2	n_{21}	n_{22}	...	n_{2m}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
k	n_{k1}	n_{k2}	...	n_{km}

Statystyka testowa

$$\chi^2_{\text{emp}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - n^t_{ij})^2}{n^t_{ij}}$$

$$n^t_{ij} = \frac{n_{i.}n_{.j}}{N},\ N = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}$$

$$n_{i.} = \sum_{j=1}^{m} n_{ij},\ n_{.j} = \sum_{i=1}^{k} n_{ij}$$

Jeżeli $\chi^2_{\text{emp}} > \chi^2(\alpha; (k-1)(m-1))$,
to hipotezę H_0 odrzucamy

Przykład. W celu zbadania istnienia związku między wykształceniem (X) a zarobkami (Y) wylosowano 950 osób. Uzyskano następujące dane

		podstawowe (W_1)	średnie (W_2)	wyższe (W_3)	ponad wyższe (W_4)
(Z_1)	$\leq$500	21	41	93	47
(Z_2)	500−1000	33	37	35	53
(Z_3	1000−1500	45	75	27	43
(Z_4)	1500−2000	30	48	50	55
(Z_5)	$\geq$2000	71	47	49	50

Czy powyższe świadczą o istnieniu zależności między wykształceniem i zarobkami?

Populacja

Cechy X, Y
para cech (*wykształcenie, zarobki*)

Założenia
obie cechy traktowane są jakościowo

Formalizacja
W celu uzyskania odpowiedzi na postawione pytanie formułowana jest hipoteza o wzajemnej niezależności wykształcenia i zarobków

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

Technika statystyczna
Test chi–kwadrat niezależności
poziom istotności $\alpha = 0.05$

Obliczenia
Zbadano łącznie $N = 950$ osób
Liczebności brzegowe:

$$n_{1\cdot} = 21 + 41 + 93 + 47 = 202$$

$$n_{2\cdot} = 158,\ n_{3\cdot} = 190,\ n_{4\cdot} = 183,\ n_{5\cdot} = 217$$

$$n_{\cdot 1} = 21 + 33 + 45 + 30 + 71 = 200$$

$$n_{\cdot 2} = 248,\ n_{\cdot 3} = 254,\ n_{\cdot 4} = 248.$$

	W_1	W_2	W_3	W_4	
Z_1	n_{11}=21	n_{12}=41	n_{13}=93	n_{14}=47	$n_{1.}$=202
Z_2	n_{21}=33	n_{22}=37	n_{23}=35	n_{24}=53	$n_{2.}$=158
Z_3	n_{31}=45	n_{32}=75	n_{33}=27	n_{34}=43	$n_{3.}$=190
Z_4	n_{41}=30	n_{42}=48	n_{43}=50	n_{44}=55	$n_{4.}$=183
Z_5	n_{51}=71	n_{52}=47	n_{53}=49	n_{54}=50	$n_{5.}$=217
	$n_{.1}$=200	$n_{.2}$=248	$n_{.3}$=254	$n_{.4}$=248	N=950

Liczebności teoretyczne:

$$n_{11}^t = \frac{n_{1.} \cdot n_{.1}}{N} = \frac{202 \cdot 200}{950} = 42.5263$$

$$n_{43}^t = \frac{n_{4.} \cdot n_{.3}}{N} = \frac{183 \cdot 254}{950} = 48.9284$$

Wyznaczenie $(n_{ij} - n_{ij}^t)^2/n_{ij}^t$ dla wszystkich dwudziestu kombinacji i, j.

$$\frac{(n_{11} - n_{11}^t)^2}{n_{11}^t} = \frac{(21 - 42.5263)^2}{42.5263} = 10.8964$$

$$\frac{(n_{43} - n_{43}^t)^2}{n_{43}^t} = \frac{(50 - 48.9284)^2}{48.9284} = 0.0235$$

	W_1	W_2	W_3	W_4	
Z_1	$n_{11}^t=$ 42.5263	$n_{12}^t=$ 52.7326	$n_{13}^t=$ 54.0084	$n_{14}^t=$ 52.7326	
Z_2	$n_{21}^t=$ 33.2632	$n_{22}^t=$ 41.2463	$n_{23}^t=$ 42.2442	$n_{24}^t=$ 41.2463	
Z_3	$n_{31}^t=$ 40.0000	$n_{32}^t=$ 49.6000	$n_{33}^t=$ 50.8000	$n_{34}^t=$ 49.6000	
Z_4	$n_{41}^t=$ 38.5263	$n_{42}^t=$ 47.7726	$n_{43}^t=$ 48.9284	$n_{44}^t=$ 47.7726	
Z_5	$n_{51}^t=$ 45.6842	$n_{52}^t=$ 56.6484	$n_{53}^t=$ 58.0189	$n_{54}^t=$ 56.6484	

	W_1	W_2	W_3	W_4	
Z_1	**10.8964**	**2.6104**	**28.1501**	**0.6232**	
Z_2	**0.0021**	**0.4372**	**1.2423**	**3.3494**	
Z_3	**0.6250**	**13.0073**	**11.1504**	**0.8782**	
Z_4	**1.8870**	**0.0011**	**0.0235**	**1.0934**	
Z_5	**14.0287**	**1.6433**	**1.4020**	**0.7803**	

Wartość statystyki testowej

$$\chi^2_{\text{emp}} = 93.8311$$

Wartość krytyczna

$$\chi^2(0.05; 12) = 21.0261$$

Odpowiedź
Hipotezę odrzucamy

Wniosek
Stwierdzamy istnienie zależności między wykształceniem i zarobkami