Test niezależności χ^2

Maciej Beręsewicz

1 Test niezależności χ^2

Punktem wyjścia w teście niezależności χ^2 jest dwuwymiarowa tablica kontyngencji $\mathbf{N} = [n_{ij}]$, gdzie n_{ij} oznaczają liczebności empiryczne w i–tym wierszu oraz j–tej kolumnie ($i = 1, \ldots, I$, $j = 1, \ldots, J$).

Tabela 1: Tablica kontyngencji dla dwóch zmiennych X i Y

Kategorie zmiennej X	**Kategorie zmiennej Y**				**Suma wierszy**
	Y_1	Y_2	$\cdots$	Y_J	
X_1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1.}$
X_2	n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I.}$
Suma kolumn	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.J}$	n

Źródło: Opracowanie własne.

Powyższa tabela przedstawia tablicę kontyngencji dla dwóch zmiennych X i Y, z których pierwsza ma I, a druga J wariantów.

W praktyce najczęściej wykorzystywanym narzędziem pozwalającym wykryć zależność między zmiennymi X i Y jest test χ^2 niezależności. *Hipoteza zerowa – H_0 zakłada brak związku między zmiennymi*, czyli zmienne X i Y są niezależne, wobec *hipotezy alternatywnej* – H_1 , że zmienne X i Y nie są niezależne. Sprawdzianem hipotezy zerowej jest statystyka:

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

gdzie n_{ij} oraz $\hat{n}_{ij}$ oznaczają odpowiednio empiryczne i teoretyczne liczebności tabeli kontyngencji. Liczebności teoretyczne $\hat{n}_{ij}$ wyznaczamy ze wzoru:

$$\hat{n}_{ij} = \frac{n_{i.} n_{.j}}{n} \tag{2}$$

gdzie n oznacza liczebność próby, a $n_{i.}$, $n_{.j}$ liczebności brzegowe. Statystyka (1) przy założeniu prawdziwości hipotezy H_0 ma graniczny rozkład χ^2 z $(I-1)(J-1)$ stopniami swobody. Obszar krytyczny w teście χ^2 niezależności buduje się prawostronnie:

$$P(\chi^2 \geqslant \chi^2_{\alpha}) = \alpha \tag{3}$$

W przypadku tabeli 2x2

Tabela 2: Tabela 2x2

Cecha X	**Cecha Y**		**Razem**
	Y_1	Y_2	
X_1	A	B	A+B
X_2	C	D	C+D
Razem	A+C	B+D	N=A+B+C+D

Źródło: Opracowanie własne.

można skorzystać z następującego wzoru:

$$\chi^2 = \frac{n * (ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \tag{4}$$

2 Test niezależności χ^2 w SAS

Zadanie 1. Stosując test niezależności χ^2 sprawdź czy pomiędzy wykształceniem, a tym czy w ciągu ostatnich 3 lat osoba brała udział w szkoleniach istnieje zależność. Na podstawie 141 losowo zapytanych pracowników pewnej firmie uzyskano następujące wyniki. Przyjmij poziom istotności $\alpha = 0,05$.

Tabela 3: Tabla kontyngencji zadania 1

Wykształcenie	**Udział w szkoleniach**	**Brak udziału w szkoleniach**
Zawodowe	18	17
Średnie	34	19
Wyższe	38	15

Wczytanie danych

```
data zad1;
input  wykszt $ szkolenie $ n;
datalines;
zawodowe tak 18
zawodowe nie 17
średnie tak 34
średnie nie 19
wyższe tak 38
wyższe nie 15
;
run;
```

Aby rozwiązać zadanie można wykorzystać do tego procedurę **PROC FREQ**. Kod znajduje się poniżej.

```
proc freq data=zad1;
table wykszt*szkolenie /chisq;
weight n;
run;
```

Na podstawie poniższej tabeli należy stwierdzić, że nie mamy podstaw do odrzucenia H_0 ponieważ p–*value* (*prob*) jest większe od założonego poziomu istotności $\alpha = 0,05$. W związku z tym nie ma zależności między wykształceniem a udziałem w szkoleniach.

Statistic	DF	Value	Prob
Chi-Square	2	3.7550	0.1530
Likelihood Ratio Chi-Square	2	3.7240	0.1554
Mantel-Haenszel Chi-Square	1	0.6492	0.4204
Phi Coefficient		0.1632	
Contingency Coefficient		0.1611	
Cramer's V		0.1632	