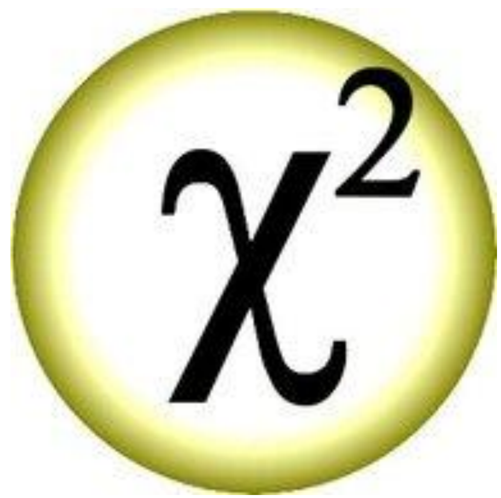
Badanie zależności między zmiennymi (test niezależności chi-kwadrat)



dr Jolanta Mirek

Uwagi wstępne

Szanowni Państwo,

Na ostatnim wykładzie była „mowa” między innymi o tabulacji.

To, że stworzymy tabelę dwudzielczą, nie oznacza jeszcze, że pomiędzy zmiennymi istnieje istotna statystycznie zależność.

Do zbadania zależności pomiędzy zmiennymi, pomiędzy którymi taka logiczna zależność naszym zdaniem może zachodzić (na mapie zmiennych już określiliśmy sobie takie zależności), służy między innymi test niezależności χ2 (chi-kwadrat). Jest to jeden z częściej w tym celu wykorzystywanych testów.

To wydaje się takie trochę nielogiczne, test niezależności do badania zależności, ale może dzięki temu będzie to łatwiej zapamiętać tym, którzy będą się uczyć ☺

Test niezależności χ^2 (chi-kwadrat)

Pozwala na **wykrycie związków** pomiędzy dwiema zmiennymi.

Pozwala on **z określonym prawdopodobieństwem** zweryfikować hipotezę o **niewystępowaniu współzależności** pomiędzy badanymi cechami.

Na wstępie konstruuje się dwudzielczą **tabelę kontyngencji** przedstawiającą **rozkład** poszczególnych odpowiedzi w ramach wyróżnionych cech.

Na jej podstawie ustala się **teoretyczny rozkład liczebności**, który wystąpiłby, gdyby **nie zachodziła żadna współzależność**.

Test niezależności χ^2 (chi-kwadrat) – tabela kontyngencji

Rozkład ocen respondentów

Ocena / Płeć	Optymi-styczna	Pesymi-styczna	Razem
Mężczyźni	n_{11}	n_{12}	$N_{1\cdot}$
Kobiety	n_{21}	n_{22}	$N_{2\cdot}$
Razem	$N_{\cdot 1}$	$N_{\cdot 2}$	N

n_{ij} **–** liczebności respondentów sklasyfikowanych ze względu na określone cechy (**i** – numer wiersza, **j** – numer kolumny)

$N_{i\cdot}$- suma **i**-tego wiersza

$N_{\cdot j}$- suma **j**-tej kolumny

N - suma całości (= liczba respondentów)

Test niezależności χ^2 (chi-kwadrat) – formułowanie hipotez

Hipoteza zerowa:

H_0: Nie istnieje istotna statystycznie zależność pomiędzy zmiennymi (w tym przypadku pomiędzy płcią a oceną przyszłej sytuacji rynkowej).

lub:

H_0: Płeć nie wpływa w istotny statystycznie sposób na **ocenę** przyszłej sytuacji rynkowej.

Hipoteza alternatywna:

H_1: Istnieje istotna statystycznie zależność pomiędzy zmiennymi (w tym przypadku pomiędzy płcią a oceną przyszłej sytuacji rynkowej).

lub:

H_1: Płeć wpływa w istotny statystycznie sposób na **ocenę** przyszłej sytuacji rynkowej.

Test niezależności χ^2 (chi-kwadrat)

Rozkład ocen respondentów

Ocena / Płeć	Optymi-styczna	Pesymi-styczna	Razem
Mężczyźni	n_{11}	n_{12}	$N_{1}.$
Kobiety	n_{21}	n_{22}	$N_{2}.$
Razem	$N._{1}$	$N._{2}$	N

Rozkład ocen respondentów (wartości obserwowane, empiryczne - w lewym górnym rogu))

Ocena / Płeć	Optymi-styczna	Pesymi-styczna	Razem
Mężczyźni	31	9	40
Kobiety	19	21	40
Razem	50	30	80

Test niezależności χ^2 (chi-kwadrat)

Rozkład ocen respondentów

Ocena / Płeć	Optymi-styczna	Pesymi-styczna	Razem
Mężczyźni	n_{11}	n_{12}	$N_{1\cdot}$
Kobiety	n_{21}	n_{22}	$N_{2\cdot}$
Razem	$N_{\cdot 1}$	$N_{\cdot 2}$	N

W celu zbadania ewentualnego związku pomiędzy badanymi cechami należy dla każdej wartości $\mathbf{n_{ij}}$ obliczyć **wartość teoretyczną (liczność oczekiwaną)**:

$$\hat{n}_{ij} = \frac{N_{i\cdot} \bullet N_{\cdot j}}{N}$$

Test niezależności χ^2 (chi-kwadrat)

Rozkład ocen respondentów

Ocena / Płeć	Optymi-styczna	Pesymi-styczna	Razem
Mężczyźni	31	9	40
Kobiety	19	21	40
Razem	50	30	80

$$\hat{n}_{ij} = \frac{N_{i\cdot} \bullet N_{\cdot j}}{N}$$

Musimy policzyć następujące wartości teoretyczne (liczności oczekiwane):

$$\hat{n}_{11}, \hat{n}_{12}, \hat{n}_{21}, \hat{n}_{22}$$

Test niezależności χ^2 (chi-kwadrat)

$$\hat{n}_{11} = \frac{40 \bullet 50}{80} = 25$$

$$\hat{n}_{12} = \frac{40 \bullet 30}{80} = 15$$

$$\hat{n}_{21} = \frac{40 \bullet 50}{80} = 25$$

$$\hat{n}_{22} = \frac{40 \bullet 30}{80} = 15$$

Test niezależności χ^2 (chi-kwadrat)

Rozkład ocen respondentów

(z wartościami teoretycznymi (oczekiwanymi) w dolnych prawych rogach)

Ocena / Płeć	Optymi-styczna	Pesymi-styczna	Razem
Mężczyźni	31 / 25	9 / 15	40
Kobiety	19 / 25	21 / 15	40
Razem	50	30	80

Postać testu χ^2:

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

Proszę się nie przestraszyć tym wzorem.

Jedynymi symbolami, które należy wyjaśnić, są:

r – liczba wierszy

s – liczba kolumn

Nasza tabela ma wymiary 2x2 (2 wiersze i 2 kolumny)

Dzięki tabeli, w której mamy wartości empiryczne (obserwowane – będące wynikiem badań) i już policzone wartości teoretyczne (oczekiwane) możemy łatwo policzyć wartość testu niezależności $\chi 2$ (chi-kwadrat) za pomocą szeregu statystycznego.

Test niezależności χ^2 (chi-kwadrat)

Ocena / Płeć	Optymi-styczna	Pesymi-styczna	Razem
Mężczyźni	31 / 25	9 / 15	40
Kobiety	19 / 25	21 / 15	40
Razem	50	30	80

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

Różnica pomiędzy wartościami z pierwszej kolumny a wartościami z drugiej kolumny

Wartości empiryczne	Wartości teoretyczne	Różnice (1-2)	Kwadraty różnic	Ilorazy (4:2)
----------------	----------------	-----------	------------	$\sum$

Ilorazy wartości z czwartej kolumny i z drugiej kolumny

Test niezależności χ^2 (chi-kwadrat)

Wartości empiryczne	Wartości teoretyczne	Różnice (1-2)	Kwadraty różnic	Ilorazy (4:2)
31	25	6	36	1,44
9	15	-6	36	2,4
19	25	-6	36	1,44
21	15	6	36	2,4
				7,68

Wartość testu niezależności χ2

Test niezależności χ^2 (chi-kwadrat)

Otrzymaną wartość χ^2

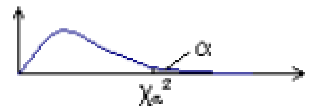
- porównujemy z **wartością krytyczną χ^2_α** z tablic dla

- **k = (r - 1)(s - 1) stopni swobody**
(r – liczba wierszy, s – liczba kolumn)

- na **poziomie istotności α** = 0,05 (czyli z prawdopodobieństwem 95%) – oznacza to możliwość popełnienia pięcioprocentowego błędu – w naukach społecznych zazwyczaj taki się przyjmuje, choć można założyć mniejszy

Wybrane tablice statystyczne

Wartości krytyczne z rozkładu chi-kwadrat.



k	α = 0,995	0,99	0,975	0,95	0,9	0,1	0,05	0,025	0,01	0,005
1	0,000	0,000	0,001	0,004	0,016	2,706	$\boxed{3,841}$	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
35	17,192	18,509	20,569	22,465	24,797	46,059	49,802	53,203	57,342	60,275
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
45	24,311	25,901	28,366	30,612	33,350	57,505	61,656	65,410	69,957	73,166

$k = (r - 1)(s - 1)$

$k = (2 - 1)(2 - 1) = 1$

$\chi^2_\alpha = 3,84 < 7,68$

Test niezależności χ^2 (chi-kwadrat) – interpretacja wyniku

Jeśli $\chi^2 \geq \chi^2_\alpha$ to **odrzucamy hipotezę zerową** o istnieniu niezależności cech, czyli istnieje **istotna statystycznie zależność** pomiędzy badanymi cechami.

Jeśli $\chi^2 < \chi^2_\alpha$ to **przyjmujemy hipotezę zerową** o istnieniu niezależności cech, czyli **nie istnieje istotna statystycznie zależność** pomiędzy badanymi cechami.

Test niezależności χ^2 (chi-kwadrat) – interpretacja wyniku

$\chi^2_\alpha = 3,84 < 7,68$

Jeśli $\chi^2 \geq \chi^2_\alpha$ to **odrzucamy hipotezę zerową** o istnieniu niezależności cech, czyli istnieje **istotna statystycznie zależność** pomiędzy badanymi cechami.

A zatem z prawdopodobieństwem 95% przyjmujemy, że płeć w **sposób istotny statystycznie** wpływa **na ocenę** przyszłej sytuacji rynkowej.

Test niezależności χ^2 (chi-kwadrat) – interpretacja wyniku

Test niezależności chi-kwadrat stosuje się w przypadku **dwóch zmiennych**.

Należy dodać, że **test niezależności chi-kwadrat** pozwala jedynie stwierdzić, **czy** istnieje statystycznie istotna zależność między zmiennymi.

Do **oceny siły tego związku** służą bowiem **inne statystyki**.

Test niezależności χ^2 (chi-kwadrat) – założenia testu χ^2

Test ten jak każdy test statystyczny ma swoje ograniczenia:

Test $\boldsymbol{\chi^2}$ nie powinien być stosowany, gdy:

- **liczba obserwacji** ogółem jest **mniejsza niż 20**;
- **liczba obserwacji** mieści się **między 20 – 40**, a **najmniejsza liczebność oczekiwana** (nie obserwowana!) w jakimkolwiek polu tabeli jest **mniejsza niż 5**;
- w przypadku **tabeli wielopolowej**, jeżeli **więcej niż 1/5 (20%)** pól zawiera **liczebności oczekiwane poniżej 5** lub jakakolwiek z liczebności oczekiwanych jest **mniejsza niż 1.**

Test niezależności χ^2 (chi-kwadrat) – założenia testu χ^2

Przy względnie **małych liczebnościach**, tzn. ogólnej liczbie obserwacji poniżej 100 oraz jeśli przynajmniej w jednym polu znajduje się **liczba mniejsza niż 10**, powinno się stosować **tzw. poprawkę Yatesa**.

Postać testu χ^2 z poprawką Yatesa:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left\{\left(n_{ij} - \hat{n}_{ij}\right) - 0{,}5\right\}^2}{\hat{n}_{ij}}$$

Proszę zwrócić uwagę na:

- sposób formułowania hipotez
- sposób obliczania wartości teoretycznych
- sposób obliczania stopni swobody
- interpretację wyniku

Test niezależności χ^2 (chi-kwadrat)

Procedura, którą zaprezentowano jest **tradycyjnym sposobem postępowania**, bez wykorzystania statystycznych programów komputerowych.

Jest ona przydatna w sytuacji, kiedy nie potrafimy obsługiwać takich programów.

Test niezależności χ^2 (chi-kwadrat)

W przypadku, kiedy korzystamy z programów statystycznych, wartość **χ^2** ma **znaczenie informacyjne**.

Decyzję o odrzuceniu (lub nie odrzucaniu) hipotezy zerowej podejmujemy na podstawie wielkości **prawdopodobieństwa testowego p**.

Jeżeli **$p < 0{,}05$**, to **odrzucamy hipotezę zerową** o istnieniu niezależności cech, czyli istnieje **istotna statystycznie zależność** pomiędzy badanymi cechami.

Jeżeli **$p \geq 0{,}05$**, to to **przyjmujemy hipotezę zerową** o istnieniu niezależności cech, czyli **nie istnieje istotna statystycznie zależność** pomiędzy badanymi cechami.