AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Tablice wielodzielcze Testy niezależności oraz test jednorodności χ^2

Statystyka

Dr inż. Janusz Majewski
Katedra Informatyki

Tablice 2x2

Przykład: Jako część studiów nad jednostronnością u człowieka badany był związek między prawo- i leworęcznością a dominacją prawo- i lewooczną.

		Dominacja ręki		*Razem*
		Leworęczni	**Praworęczni**	
Dominacja oka	**Lewooczni**	27	110	137
	Prawooczni	27	236	263
Razem		54	346	400

Tablice 2x2

		Dominacja ręki		*Razem*
		Leworęczni	**Praworęczni**	
Dominacja oka	**Lewooczni**	27	110	137
	Prawooczni	27	236	263
Razem		54	346	400

W teście McNemara chodziło o weryfikację hipotezy parametrycznej: czy frakcja leworęcznych jest identyczna jak frakcja lewoocznych? Tutaj chodzi o weryfikację hipotezy nieparametrycznej: czy istnieje związek między leworęcznością a lewoocznością? **Hipoteza zerowa, jak zwykle "neutralna", mówi, że związku nie ma, hipoteza alternatywna twierdzi, że związek ten istnieje.** Omawiany test nazywamy ***testem niezależności chi-kwadrat.***

Tablice 2x2

		Dominacja ręki		*Razem*
		Leworęczni	**Praworęczni**	
Dominacja oka	**Lewooczni**	27	110	137
	Prawooczni	27	236	263
Razem		54	346	400

Jeżeli hipoteza zerowa byłaby prawdziwa, to oczekiwane liczebności w polach tabeli powinny spełniać zależność:

$$E = \frac{suma\ wiersza\ \cdot suma\ kolumny}{liczebność\ całkowita}$$

bo...

$$P\left(\begin{matrix}lewa\\ręka\end{matrix} \wedge \begin{matrix}lewe\\oko\end{matrix}\right) = P\left(\begin{matrix}lewa\\ręka\end{matrix}\right) \cdot P\left(\begin{matrix}lewe\\oko\end{matrix} \middle| \begin{matrix}lewa\\ręka\end{matrix}\right) = P\left(\begin{matrix}lewa\\ręka\end{matrix}\right) \cdot P\left(\begin{matrix}lewe\\oko\end{matrix}\right)$$

gdy nie ma związku, to: **prawdopodobieństwo warunkowe = = prawdopodobieństwo bezwarunkowe**

Tablice 2x2

		Leworęczni	Praworęczni	*Razem*
Lewooczni	**O**	27	110	137
	E	18.5	118.5	
	O-E	**8.5**	-8.5	
Prawooczni	**O**	27	236	263
	E	35.5	227.5	
	O-E	-8.5	8.5	
Razem		54	346	400

Leworęczność "sprzyja" lewooczności: zależność "dodatnia"

Tablice 2x2

		Leworęczni	Praworęczni	*Razem*
Lewooczni	**O** **E** **O-E**	27 18.5 **8.5**	110 118.5 -8.5	137
Prawooczni	**O** **E** **O-E**	27 35.5 -8.5	236 227.5 8.5	263
Razem		54	346	400

$$\chi^2 = \sum_{\substack{po\ wszystkic\ h \\ polac\ h\ tabeli}} \frac{(O-E)^2}{E}$$

$$H_0\ odrzucamy, gdy\ \chi^2 \geq\ {}_{\alpha}\chi^2_{(1)}$$

$$\chi^2 = 6.866, \qquad {}_{0.01}\chi^2_{(1)} = 6.63,$$

$H_0\ odrzucamy!$

Tablice 2x2

Test wolno stosować, gdy wszystkie liczebności oczekiwane są ≥5 oraz N ≥40. Gdy liczebności są w pobliżu dolnej granicy, stosujemy poprawkę na nieciągłość (Yatesa).

$$\chi_c^2 = \sum \frac{\left(|O - E| - \frac{1}{2}\right)^2}{E}$$

Gdy którakolwiek z liczebności oczekiwanych jest mniejsza niż 5 i 20<N<40 lub gdy N<20, możemy zastosować dokładny test Fishera. Nie wolno nam stosować testu χ^2.

Nie polecam testu Fishera – bardzo trudno odrzucić hipotezę zerową przy małej liczebności próby.

Miary siły związku

Przykład: Badano, czy obecność koniczyny białej (*Trifolium repens*) na pastwisku jest związana z obecnością odchodów dżdżownic. Zbadano 80 losowo dobranych powierzchni próbnych o polu 1 stopy kwadratowej każda. Uzyskano wyniki:

		Odchody dżdżownic		*Razem*
		Obecne	**Nieobecne**	
Koniczyna biała	**Obecna**	18	5	23
	Nieobecna	34	23	57
Razem		52	28	80

$$\chi^2 = 2.4952$$

$$_{0.05}\chi^2_{(1)} = 3.84$$

Nie ma podstaw do odrzucenia hipotezy o braku związku.

Miary siły związku

Przypuśćmy, że takie samo badanie wykonano dla 800 powierzchni próbnych i otrzymano wszystkie wyniki 10-krotnie większe.

		Odchody dżdżownic		*Razem*
		Obecne	**Nieobecne**	
Koniczyna biała	**Obecna**	180	50	230
	Nieobecna	340	230	570
Razem		520	280	800

$$\chi^2 = 24.952$$

$$!_{0.01}\,\chi^2_{(1)} = 6.63$$

Hipotezę o braku związku należy odrzucić!

Miary siły związku

χ^2 rośnie proporcjonalnie do N. Naturalnym miernikiem **siły związku** jest

$$\phi^2 = \frac{\chi^2}{N}.$$

Przyjmiemy oznaczenia:

a	b	r_1
c	d	r_2
s_1	s_2	N

Wtedy:

$$\chi^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{r_1 \cdot r_2 \cdot s_1 \cdot s_2}$$

Mamy podstawową miarę siły związku:

$$\phi = \sqrt{\frac{(a \cdot d - b \cdot c)^2}{r_1 \cdot r_2 \cdot s_1 \cdot s_2}} \qquad \phi \in \langle 0,1 \rangle$$

Miary siły związku

oraz miarę Pearsona:

$$r_p = \sqrt{\frac{2 \cdot (a \cdot d - b \cdot c)^2}{r_1 \cdot r_2 \cdot s_1 \cdot s_2 + (a \cdot d - b \cdot c)^2}}$$

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

i miarę Kendalla:

$$Q = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}$$

$$Q \in \langle -1,1 \rangle$$

Miary siły związku

Tablica			Φ	Q	r_p
50	0	50	1	1	1
0	50	50			
50	50	100			
40	0	40	0.82	1 wartość = 1, gdy choć jedno zero	0.89
10	50	60			
50	50	100			
40	10	50	0.60	0.88	0.73
10	40	50			
50	50	100			
20	30	50	0.20	-0.38 znak ujemny (domina-cja pobocznej prze-kątnej)	0.28
30	20	50			
50	50	100			
25	25	50	0	0	0
25	25	50			
50	50	100			

Porównywanie więcej niż dwóch populacji

Dlaczego nie można porównywać parametrów więcej niż dwóch populacji metodą porównywania „każda z każdą”?

Liczba populacji = k

Liczba porównań = $\binom{k}{2}$

Poziom istotności pojedynczego porównania = α

Poziom istotności całego badania (przy nieprawdziwym zresztą założeniu o niezależności wzajemnej poszczególnych porównań)

$$\alpha_{\Sigma} = 1 - (1-\alpha)^{\binom{k}{2}}$$

Porównywanie więcej niż dwóch populacji

Przykład:

$k=4 \quad \alpha=0,05$

$$\binom{k}{2}=\binom{4}{2}=\frac{4\cdot 3}{1\cdot 2}=6$$

$$\alpha_{\Sigma}=1-(1-0,05)^{\binom{4}{2}}=1-(0,95)^{6}=0,265$$

Poziom istotności całego badania statystycznego przy porównywaniu kilku populacji metodą „każda z każdą” jest nie do przyjęcia. Dlatego też porównujemy *k* populacji wykonując jeden „łączny” test z ustalonym poziomem istotności.

Test jednorodności chi-kwadrat dla porównywania kilku frakcji

Tablice wielodzielcze i statystyka chi-kwadrat stosowane są w jeszcze jednym rodzaju analizy: wówczas, gdy chcemy sprawdzić, czy frakcja jakiejś charakterystycznej cechy jest równa w kilku populacjach. Na przykład szpital dziecięcy chce stwierdzić, czy odsetek dzieci – nosicieli i nienosicieli bakterii *Streptococcus pyogenes* – jest taki sam w trzech grupach dzieci: z migdałkami niepowiększonymi, z powiększonymi i z bardzo powiększonymi. W pewnym sensie pytanie o równość frakcji jest pytaniem o to, czy te grupy dzieci są ***jednorodne*** pod względem odsetka nosicielstwa tej bakterii. Z tego powodu testy równości frakcji w kilku populacjach nazywa się również ***testami jednorodności***.

Tablice 2 x k

Porównanie kilku frakcji

Grupa	1	2	...	i	...	k
Sukces	r_1	r_2	...	r_i	...	r_k
Porażka	$n_1 - r_1$	$n_2 - r_2$	...	$n_i - r_i$	...	$n_k - r_k$
Ogółem	n_1	n_2	...	n_i	...	n_k
Frakcja sukcesu	$p_1 = \frac{r_1}{n_1}$	$p_2 = \frac{r_2}{n_2}$	...	$p_i = \frac{r_i}{n_i}$	...	$p_k = \frac{r_k}{n_k}$

} 2 x k

H_0: wszystkie Π_i są równe

H_1: nieprawdą jest, że wszystkie Π_i są równe

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{\sum_{i=1}^{k} \frac{r_i^2}{n_i} - \frac{R^2}{N}}{P(1-P)}$$

gdzie $R = \sum_i r_i, \quad N = \sum_i n_i, \quad P = \frac{R}{N}.$

Tablice 2 x k

Grupa	1	2	...	i	...	k
Sukces	r_1	r_2	...	r_i	...	r_k
Porażka	$n_1 - r_1$	$n_2 - r_2$	...	$n_i - r_i$	...	$n_k - r_k$
Ogółem	n_1	n_2	...	n_i	...	n_k
Frakcja sukcesu	$p_1 = \frac{r_1}{n_1}$	$p_2 = \frac{r_2}{n_2}$	...	$p_i = \frac{r_i}{n_i}$	...	$p_k = \frac{r_k}{n_k}$

} 2 x k

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{\sum_{i=1}^{k} \frac{r_i^2}{n_i} - \frac{R^2}{N}}{P(1-P)}$$

gdzie $R = \sum_i r_i, \quad N = \sum_i n_i, \quad P = \frac{R}{N}.$

Tak obliczone χ^2 porównujemy z wartością krytyczną o (k-1) stopniach swobody. H_0 odrzucamy, gdy:

$$\chi^2 \geq {}_{\alpha}\chi^2_{(k-1)}$$

Tablice 2 x k

Przykład: Liczba dzieci - nosicieli i nienosicieli bakterii *Streptococcus pyogenes* w zależności od wielkości migdałków.

	Migdałki			*Razem*
	Nie powiększone	**Powiększone**	**Bardzo powiększone**	
Nosiciele	19	29	24	72
Nienosiciele	497	560	269	1326
Razem	516	589	293	1398
Frakcja nosicieli	0.0368	0.0492	0.0892	0.0515

$\chi^2 = 7.88$ $\quad {}_{0.05}\chi^2_{(2)} = 5.99$ $\quad \chi^2 > \chi^2_{kryt}$

Hipotezę zerową o równości frakcji nosicielstwa w każdej z grup związanych z wielkością migdałków należy odrzucić.

Obliczona statystyka χ^2 charakteryzuje się dwoma stopniami swobody. Dalej omówiony zostanie test trendu częstości będący jednym z możliwych sposobów wyodrębniania składników sumarycznego χ^2 w celu uzyskania możliwości bardziej precyzyjnego wnioskowania.

Test trendu częstości

Przykład: Badana jest zmienność liczby wystąpień drobnoustrojów na oddziale OIOM w stosunku do liczby wystąpień drobnoustrojów na wszystkich oddziałach zabiegowych pewnego szpitala w latach 1992-95. Zebrane przez pracownię mikrobiologiczną dane przedstawiono w poniższej tabeli.

Oddział	**Rok - x_i**				***Razem***
	1992	**1993**	**1994**	**1995**	
OIOM r_i	1609	1368	1421	1455	**R** 5853
Inne zabiegowe $n_i - r_i$	999	657	624	821	**N-R** 3103
Razem n_i	2608	2025	2045	2276	**N** 8954
$p_i = \frac{r_i}{n_i} = \frac{OIOM}{RAZEM}$	0.617	0.675	0.695	0.639	$P = \frac{R}{N} = 8954$

Test trendu częstości

Oddział	Rok - x_i				*Razem*
	1992	1993	1994	1995	
OIOM r_i	1609	1368	1421	1455	**R** 5853
Inne zabiegowe $n_i - r_i$	999	657	624	821	**N-R** 3103
Razem n_i	2608	2025	2045	2276	**N** 8954
$p_i = \frac{r_i}{n_i} = \frac{OIOM}{RAZEM}$	0.617	0.675	0.695	0.639	$P = \frac{R}{N} = 8954$

Hipoteza zerowa: udziały drobnoustrojów izolowanych na oddziale OIOM w kolejnych latach są takie same.

Hipoteza alternatywna: udziały drobnoustrojów izolowanych na oddziale OIOM w kolejnych latach są różne.

Obliczamy $\chi^2 = 12.145$, $_{0.05}\chi^2_{(3)} = 7.815$

$\chi^2 >$ $_{0.05}\chi^2_{(3)}$ => hipotezę zerową można odrzucić.

Test trendu częstości

Jeśli kategorie klasyfikacji kolumn dadzą się uporządkować i można im przyporządkować pewne liczby x_i, wówczas wskazane może być celowe przeprowadzenie testu trendu. Rozbijamy wtedy całkowite χ^2 na dwa składniki: χ_1^2 i χ_2^2.

$$\chi^2 = \chi_1^2 + \chi_2^2$$

Pierwszy składnik χ_1^2, charakteryzujący się jednym stopniem swobody jest odpowiedzialny za liniowy trend frakcji p_i.

$$\chi_1^2 = \frac{N\left(N\sum_{i=1}^{k} r_i x_i - R\sum_{i=1}^{k} n_i x_i\right)^2}{R(N-R)\left[N\sum_{i=1}^{k} n_i x_i^2 - \left(\sum_{i=1}^{k} n_i x_i\right)^2\right]}$$

Drugi składnik χ_2^2, charakteryzujący się $k - 2$ stopniami swobody

$$\chi_2^2 = \chi^2 - \chi_1^2 ,$$

odpowiedzialny jest za odchylenia od liniowego trendu.

Każdy ze składników testujemy osobno, porównując z wartościami krytycznymi odczytanymi z tablic dla odpowiedniej liczby stopni swobody.

Test trendu częstości

Oddział	Rok - x_i				*Razem*
	1992	1993	1994	1995	
OIOM	1609	1368	1421	1455	**R** 5853
Inne zabiegowe $n_i - r_i$	999	657	624	821	**N-R** 3103
Razem n_i	2608	2025	2045	2276	**N** 8954
$p_i = \frac{r_i}{n_i} = \frac{OIOM}{RAZEM}$	0.617	0.675	0.695	0.639	$P = \frac{R}{N} = 8954$

Przykład c.d.: Ponieważ klasyfikacja czasowa jest w naturalny sposób uporządkowana, zaś obliczone p_i wskazują na możliwość istnienia wzrastającego trendu udziałów drobnoustrojów na oddziale OIOM w stosunku do wszystkich drobnoustrojów wyizolowanych na oddziałach zabiegowych szpitala, przeprowadzamy test trendu, przyjmując np. że $x_1 = 1$ jest odpowiednikiem roku 1992, $x_2 = 2$ odpowiada 1993 rokowi, itd.

Obliczamy:

$$\chi_1^2 = 4.830 \qquad {}_{0.05}\chi_{(1)}^2 = 3.841$$

$$\chi_2^2 = 7.463 \qquad {}_{0.05}\chi_{(2)}^2 = 5.991$$

Mamy:

$$\chi_1^2 > {}_{0.05}\chi_{(1)}^2 \qquad \chi_2^2 > {}_{0.05}\chi_{(2)}^2$$

Test trendu częstości

Obliczamy:

$$\chi_1^2 = 4.830 \qquad {}_{0.05}\chi_{(1)}^2 = 3.841$$

$$\chi_2^2 = 7.463 \qquad {}_{0.05}\chi_{(2)}^2 = 5.991$$

Mamy:

$$\chi_1^2 > {}_{0.05}\chi_{(1)}^2 \qquad \chi_2^2 > {}_{0.05}\chi_{(2)}^2$$

Uzyskaliśmy potwierdzenie istotności liniowego trendu udziałów drobnoustrojów wyizolowanych na oddziale OIOM w kolejnych latach, jak i potwierdzenie istotności odchyleń od tego trendu. Tak więc prawdopodobnie trend w rzeczywistości nie ma charakteru zależności liniowej.

Uwaga: Test trendu wolno przeprowadzać dopiero wówczas, gdy zasadniczy test porównywania frakcji da wynik istotny. Jeśli nie możemy odrzucić hipotezy zerowej o równości porównywanych frakcji - nie należy przeprowadzać testu trendu, gdyż jego ewentualny pozytywny (istotny) wynik nie będzie wiarygodny.

Test niezależności dla tablic *r x c*

r - liczba wierszy tabeli

c - liczba kolumn tabeli

H_0: brak zależności między klasyfikacją "wierszy" a klasyfikacją "kolumn"

H_1: zależność istnieje

Obliczamy wartości oczekiwane dla każdego pola tabeli:

$$E = \frac{suma\ wiersza \cdot suma\ kolumny}{całkowita\ liczba\ obserwacji}$$

i statystykę χ^2

$$\chi^2 = \sum_{po\ wszystkich\ polach\ tabeli} \frac{(O-E)^2}{E}$$

O - liczebności obserwowane

Test niezależności dla tablic *r x c*

$$\chi^2 = \sum_{po\ wszystkic\ h\ polac\ h\ tabeli} \frac{(O-E)^2}{E}$$

O - liczebności obserwowane

χ^2 ma rozkład χ^2 o $(r-1)\cdot(c-1)$ stopniach swobody.

H_0 odrzucamy, gdy

$$\chi^2 \geq {}_{\alpha}\chi^2_{(r-1)\cdot(c-1)}.$$

Z testu można korzystać, gdy tylko nieliczne liczebności oczekiwane są mniejsze niż 5 (jeden wynik na 5 pól tabeli, 2 wyniki na 10 pól, itd.) i w wypadku, gdy żadna z liczebności oczekiwanych nie jest mniejsza od 1! Gdy warunki te nie są spełnione - można zmniejszać wymiarowość agregacji.

Test niezależności dla tablic *r x c*

MIARY SIŁY ZWIĄZKU

- miara Czuprowa T

$$T^2 = \frac{\chi^2}{N\sqrt{(r-1)(c-1)}}$$

Miernik ten przyjmuje wartość maksymalną = 1 tylko dla tablic kwadratowych (r = c).

- miara Cramera V

$$V^2 = \frac{\chi^2}{N \cdot min(r-1, c-1)}$$

- miara Pearsona C

$$C^2 = \frac{\chi^2}{\chi^2 + N}$$

Wartość maksymalna miernika C zależy od wymiarowości tablicy, np. dla tablicy 2 x 2 wynosi $\sqrt{2}/2$. Można go standaryzować, znając tę wartość maksymalną.

Test niezależności dla tablic *r x c*

Przykład: Badano zależność między grupą krwi, a chorobami żołądka. Z badań wyłączono niewielką grupę osób z grupą AB. Uwzględniono raka żołądka i chorobę wrzodową. Do porównań wykorzystano kontrolną grupę ludzi zdrowych.

		Choroba wrzodowa **W**	Rak żołądka **R**	Grupa kontrolna **K**	*Razem*
0	**O**	983	383	2892	4258
	E	872.39	428.91	2956.70	
	O-E	110.61	-45.91	-64.70	
A	**O**	679	416	2625	3720
	E	762.16	374.72	2583.12	
	O-E	-83.16	41.28	41.88	
B	**O**	134	84	570	788
	E	161.44	79.38	547.18	
	O-E	-27.44	4.62	22.82	
Razem		1796	883	6087	8766

Test niezależności dla tablic *r x c*

$\chi^2 = 40.54$ $\quad$ $_{\mathbf{0.001}}\chi^2_{(4)} = 18.47$

$\chi^2 > \chi^2_{kryt}$ => związek między grupą krwi a chorobami żołądka jest wysoce istotny (0.1% błędu)

$\left.\begin{matrix} V = T = 0.0481 \\ C = 0.0678 \end{matrix}\right\}$ => związek jest słaby

Na wysoką wartość χ^2 ma wpływ duża liczebność próby.

Jeżeli w tablicy wielopolowej r x c zaobserwujemy istotną zależność, to można sprawdzić, czy zależność ta utrzymuje się w pewnych fragmentach tablicy i wykryć te "obszary" tablicy, w których zależność ta jest skoncentrowana.

Metodą jest podział tablicy pełnej na kilka tablic "cząstkowych" i obliczenie χ^2 dla tych tablic cząstkowych. Tak wyznaczone "składniki χ^2" powinny po zsumowaniu dać w przybliżeniu wartość "całkowitego χ^2". Dokładność sumowania zależy od sposobu liczenia wartości oczekiwanych.