NIEZALEŻNOŚĆ i ZALEŻNOŚĆ między cechami Test chi-kwadrat, OR, RR

M Zalewska
Zakład Profilaktyki Zagrożeń Środowiskowych i Alergologii

Analiza niezależności zmiennych jakościowych

(test niezależności Chi-kwadrat)

- Rozważmy parę jakościowych zmiennych losowych *X* i *Y*
- Zmienna *X* przyjmuje *k* kategorii
- Zmienna *Y* przyjmuje *l* kategorii
- UWAGA: W ten sposób można również analizować niezależność dla zmiennych ilościowych dyskretnych i ciągłych po ich kategoryzacji

Wyniki *n* elementowej próby zapisujemy w tabeli kontyngencji:

X \ *Y*	**1**	**2**	**…**	***l***
1	n_{11}	n_{12}	…	n_{1l}
2	n_{21}	n_{22}	…	n_{2l}
k	n_{k1}	n_{k2}	…	n_{kl}

Zauważmy, że

- obserwujemy pary zmiennych (X,Y), więc n_{ij} oznacza ilość współwystąpienia elementów cechy X o kategorii i oraz elementów cechy Y o kategorii j
- suma wszystkich elementów w tabeli kontyngencji wynosi n tyle ile wynosi rozmiar próby

Liczebności brzegowe

Dla wierszy:

$n_{i\bullet}$ – liczebności brzegowe *i*-tego wiersza (suma elementów w *i*-tym wierszu)

Dla kolumn:

$n_{\bullet j}$ – liczebności brzegowe *j*-tej kolumny (suma elementów w *j*-tej kolumnie)

Przykład

- Obserwowano 4 metody leczenia oraz stan poprawy zdrowia pacjentów. Informacje zestawiono w tablicy kontyngencji (łącznie leczonych było 400 pacjentów)

Stan poprawy zdrowia (X)	**metoda leczenia (Y)**				$n_{i\bullet}$
	A	**B**	**C**	**D**	
mierny	**30**	**40**	**40**	**20**	**130**
dostateczny	**30**	**40**	**20**	**40**	**130**
dobry	**40**	**20**	**40**	**40**	**140**
$n_{\bullet j}$	**100**	**100**	**100**	**100**	**400**

Pytanie badawcze:

Czy stan poprawy zdrowia pacjentów zależy od metody leczenia?

Możemy wykorzystać :

Test Chi-kwadrat niezależności

Hipoteza zerowa i alternatywna

H_0: brak zależności między cechami X i Y przeciw hipotezie alternatywnej

H_1: cechy są zależne

Statystyka testowa jest postaci:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - n_{i\bullet} n_{\bullet j} / n\right)^2}{n_{i\bullet} n_{\bullet j} / n}$$

Dla dużych prób:

Rozkład statystyki *chi-kwadrat* przy prawdziwej hipotezie zerowej jest zbliżony do rozkładu Chi-kwadrat o

(k-1)(l-1) stopniach swobody

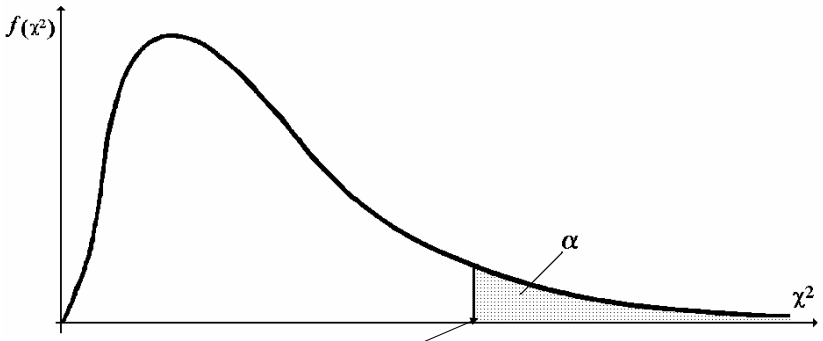
Statystyka testowa

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - n_{i\bullet} n_{\bullet j} / n\right)^2}{n_{i\bullet} n_{\bullet j} / n}$$

$$chi\_kw. = \sum \frac{(liczeb.empiryczne - liczeb.teoretyczne)^2}{liczeb.teoretyczne}$$

Przyjmujemy poziom istotności 0.05

Odrzucamy H_0 gdy statystyka testowa Chi-kw przekracza odpowiedni kwantyl z rozkładu chi-kwadrat



qchisq((1-*alfa*),(*k*-1)*(*l*-1))

Wartości statystyki *Chi-kw* >qchisq(0.95,(*k*-1)*(*l*-1)) świadczą przeciw hipotezie zerowej

Uwaga

- W tablicach rozkładu Chi-kwadrat zazwyczaj są podane prawe ogony a nie kwantyle.

Przykład c.d. Obliczanie liczebności teoretycznych

- wyznaczamy tabelę liczebności teoretycznych $n_{i\bullet}n_{\bullet j}/n$

Stan poprawy zdrowia (X)	metoda leczenia (X)				$n_{i\bullet}$
	A	B	C	D	
mierny	32,5	32,5	32,5	32,5	130
dostatecz ny	32,5	32,5	32,5	32,5	130
dobry	35,0	35,0	35,0	35,0	140
$n_{\bullet j}$	100	100	100	100	400

- Wyznaczamy tabelę wartości $\dfrac{\left(n_{ij} - n_{i\bullet}n_{\bullet j}/n\right)^2}{n_{i\bullet}n_{\bullet j}/n}$

(30-32.5)^2/32.5

Stan poprawy zdrowia (X)	metoda leczenia (X)				Σ
	A	B	C	D	
mierny	0,19	1,73	1,73	4,81	8,46
dostatecz ny	0,19	1,73	4,81	1,73	8,46
dobry	0,71	6,43	0,71	0,71	8,56
$n_{\bullet j}$	1,09	9,89	7,25	7,25	25,48

Wartość statystyki Chi –kwadrat jest sumą elementów tabeli

Obliczamy wartość statystyki testowej Chi-kwadrat

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - n_{i\bullet}n_{\bullet j}/n\right)^2}{n_{i\bullet}n_{\bullet j}/n}$$

= 25,49

Podjęcie decyzji odnośnie hipotezy zerowej

W naszym przykładzie obliczona statystyka Chi-kw =25.49

a kwantyl rzędu 0.95 z rozkładu chi-kwadrat o **(3-1)*(4-1) = 6 stopniach swobody**

wynosi :

wartość krytyczna = 12.59,

Zatem na poziomie istotności α = 0,05 odrzucamy hipotezę zerową o niezależności stanu poprawy zdrowia od zastosowanej metody leczenia.

Test chi-kwadrat niezależności przykłady

- Czy istnieje zależność między wykształceniem (W) i zarobkami (Z)?
- Czy istnieje zależność między rozkładem stężenia białka a rodzajem stosowanej diety?
- Czy jest zależność miedzy leczeniem astmy a wiekiem pacjenta?
- Czy istnieje zależność miedzy wiekiem i objawami astmy?

Przykład

W trzech szpitalach zastosowano nowa metodę leczenia pewnej choroby.

W szpitalu A na n_1=100 leczonych zaobserwowano 80 przypadków poprawy,

w szpitalu B na n_2=50 leczonych - 30 przypadków poprawy, a

w szpitalu C na n_3=80 leczonych - 60.

Czy szansa wyleczenia zależy od szpitala?

Przyjąć poziom istotności równy 0.05.

	1	2	3		
niepopr	20	20	20	60	tabelka wyjściowa
popr	80	30	60	170	
suma	100	50	80	230	obliczamy wartości brzegowe

obserwowane (empiryczne)

obliczamy wartości oczekiwane

60*100/230
26.08696

niepopr	26,08696	13,04348	20,86957
popr	73,91304	36,95652	59,13043

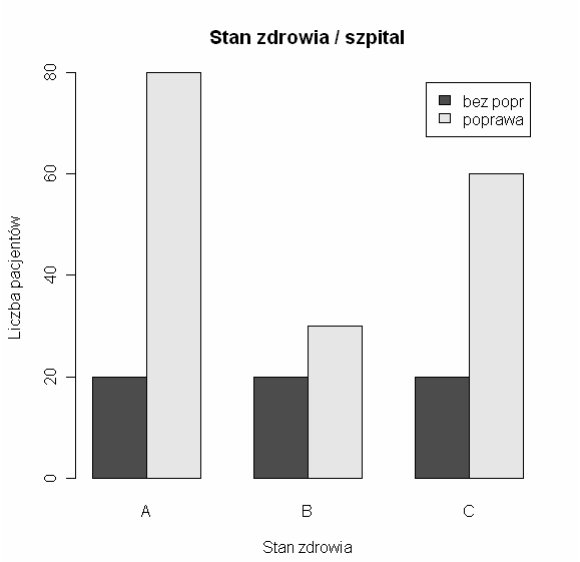
kwadraty reszt jako składniki Chi kwadrat

((20-26.08696)^2)/26.08696
1.420291

niepopr	1,42029	3,710145	0,036232
popr	0,501279	1,309463	0,012788

chi-kwadrat oblicz	6,990196

Suma kwadratów reszt



chi-kwadrat tablic
5.99

obszar krytyczny
(5.99,+niesk)

df= (w-1)*(k-1)=(2-1)*(3-1)=2

alfa=0.05

Ho: wiersze i kolumny niezależne (nie ma zależności miedzy stanem pacjenta a szpitalami)

Decyzja: statystyka obliczona (6.99) wpada w obszar krytyczny (5.99,+niesk) odrzucamy Ho na korzyść H1: wykryto zależność miedzy szpitalami a stanem pacjenta na poziomie istotności 0.05 (szanse wyzdrowienia zależą od szpitala)

Pearson's Chi-squared test

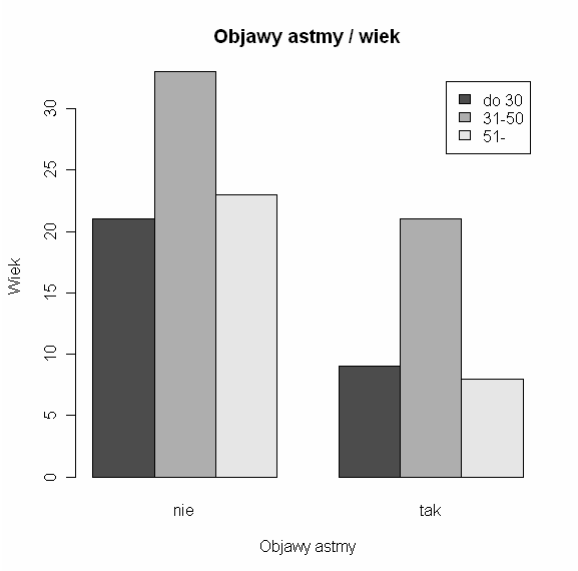
data: rbind(niepopr, popr)
X-squared = 6.9902, df = 2, p-value = 0.03035 p-value <0.05

Decyzja: p-value <0.05 odrzucamy H_0

Przykład

Objawy astmy

Wiek	nie	tak
do 30	21	9
31-50	33	21
51-	23	8



Pearson's Chi-squared test

data: astmalecz1
X-squared = 1.6934, df = 2,
p-value = 0.4288

Typy badań pozwalające oceniać powiązanie zmiennych

- Badania prospektywne kohortowe
  - Ustalone liczności dla grup narażonych i nie narażonych na badany czynnik
  - Obserwacje wystąpienia choroby w obu grupach
- Badania retrospektywne przypadek-kontrola (case-control)
  - Ustalona liczba przypadków i kontroli
  - Określić kto był narażony na czynnik

Schemat kohortowych badań prospektywnych

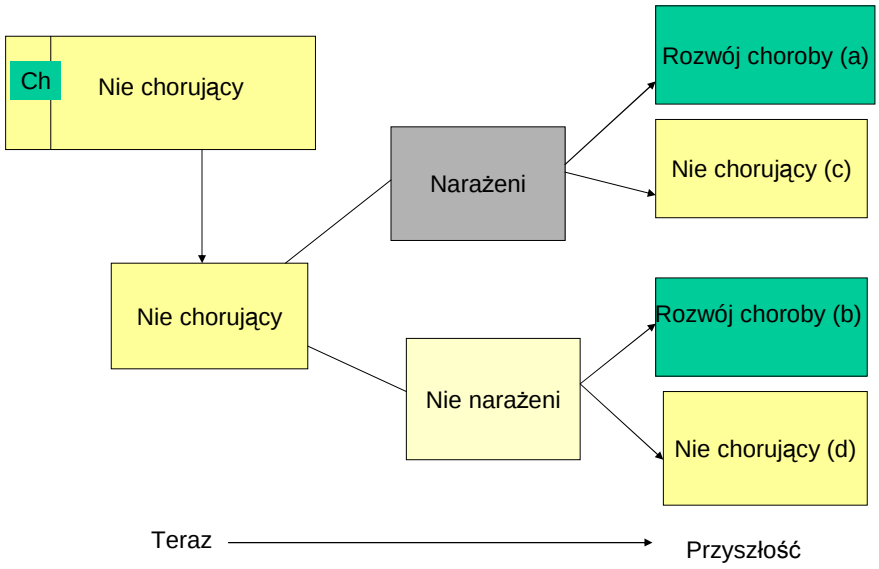


Tabela 2 x 2

	Szczepiony	Nie szczepiony
Polio (+)	82	162
Polio (-)	200663	201067
	200745	201229

2 x 2

X / Y — Badana choroba	Narażeni na czynnik (szczepieni)	Nie narażeni (nie szczepieni)
Choroba - Tak	a	b
Nie	c	d

n=a+b+c+d

Badania prospektywne

- Możemy oszacować ryzyko zachorowania

liczba zachorowań w okresie badania

całkowita liczba w kohorcie

$$ryzyko\ zachorowan\ ia = \frac{a+b}{n}$$

6/10000

Ryzyko względne (RR)

$$RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

Ryzyko w grupie1 (narażonych) / ryzyko w grupie 2 (nie narażonych)

RR

Ogólna interpretacja (RR)

- RR > 1 dodatnia zależność (pozytywny wpływ) czynnika ryzyka na rozwój choroby
- RR = 1 brak związku
- RR < 1 ujemna zależność (negatywny wpływ)

The “grupa referencyjna” w mianowniku

Grupa referencyjna wybierana jako “nie narażeni”

Typ badań: Prospektywne badanie kohortowe

- Związek między antykoncepcją ustną (OC) a chorobami krążenia
- Plan badań:
  - Identyfikacja 23000 (tych co stosują) oraz 23000 (nie stosują)
  - Ustalić czy wystąpiło zachorowanie na choroby krążenia

Wyniki tabela 2 x 2

Choroba	Narażeni (Czynnik ryzyka) OC	Nie stosują
Tak	24	5
Nie	22976	22995
	23000	23000

$p < .001$ (Fisher's Exact Test)

Ryzyko względne RR

- Estymator ryzyka względnego w przykładzie OC/choroby krążenia

$$\hat{RR} = \frac{\hat{p}_{OC}}{\hat{p}_{Nie-OC}} = \frac{24/23000}{5/23000} = 4.8$$

- Interpretacja
  - U stosujących OC blisko 5 razy bardziej prawdopodobne jest wystąpienie chorób układu krążenia niż u nie stosujących

RR

- Uwaga: Można także estymować RR chorób krążenia dla nie stosujących ustnej antykoncepcji w stosunku do stosujących

$$\hat{RR}^* = \frac{\hat{p}_{Nie-OC}}{\hat{p}_{OC}} = \frac{5/23000}{24/23000} = 0.21$$

$$\left(\hat{RR}^* = \frac{1}{\hat{RR}} = \frac{1}{4.8} \right)$$

95% przedział ufności dla RR

$$[\ln RR - 1.96 * \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}; \qquad \ln RR + 1.96 * \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}]$$

- The 95% CI dla ryzyka względnego wystąpienia chorób układu krążenia u stosujących OC w porównaniu do nie stosujących wynosi 1.8–12.6
- Warto zauważyć, że wyznaczony przedział nie zawiera 1

Pamiętać o próbkowym estymatorze vs parametr populacji

- Przypomnienie:

$H_o: p_1 = p_2$ $\qquad$ $H_o: RR = 1$

$H_1: p_1 \neq p_2$ $\qquad$ $H_1: RR \neq 1$

- p = .0004 z testu Chi-kwadrat

RR vs poziom krytyczny (p-value)

- Duży RR nie znaczy że p-value jest małe
- Duży RR może wystąpić jeśli próbka jest mała
- Poziom krytyczny (p-value) zależy od zarówno od wielkości RR jak i rozmiaru próbki.

OR iloraz szans

- Szansa wystąpienia choroby jest zdefiniowana jako

$$\frac{\Pr\ awdopodobienstwo\_wystapienia\_choroby}{\Pr\ awdopodobienstwo\_nie\_wystapienia\_choroby}$$

lub:

$$\frac{\Pr\ awdopodobienstwo\_wystapienia\_choroby}{1-(\Pr\ awdopodobienstwo\_wystapienia\_choroby)}$$

OR

- Dane w tabeli 2 x 2

Choroba	OC Tak	Nie
Tak	24	5
Nie	22976	22995
	23000	23000

Iloraz szans OR

- Szansa wystąpienia choroby w grupie OC

$$\frac{\hat{p}_1}{1-\hat{p}_1} = \frac{24/23000}{22976/23000} = \frac{24}{22976}$$

- Szansa wystąpienia choroby w grupie nie -OC

$$\frac{\hat{p}_2}{1-\hat{p}_2} = \frac{5/23000}{22995/23000} = \frac{5}{22995}$$

OR

$$OR = \frac{24/22976}{5/22995} = \frac{24 \cdot 22995}{5 \cdot 22976} = 4.8$$

OR

- Dla dowolnej tabeli 2 x 2

		Czynnik narażenia: T	Czynnik narażenia: N
Choroba	T	a	b
	N	c	d

$$OR = \frac{ad}{bc}$$

OR

Ma podobną interpretację co RR:

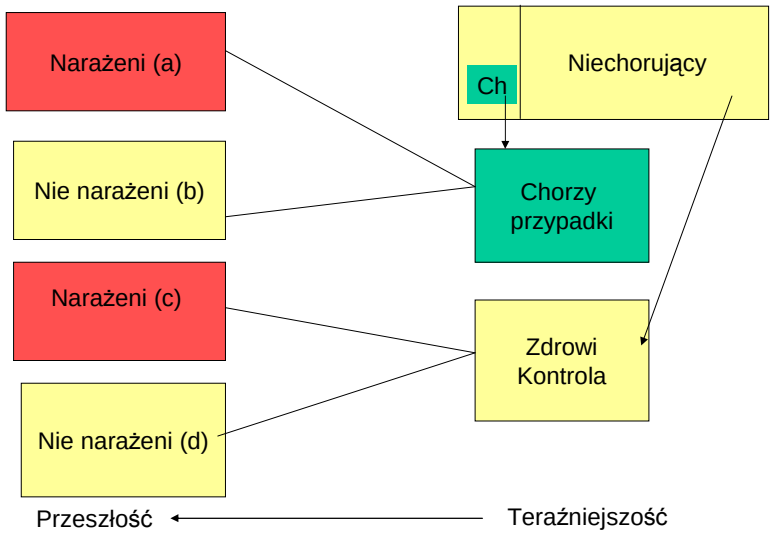
$$\text{Iloraz szans} \begin{cases} > 1 & \text{+ powiazane} \\ = 1 & \text{nie są powiazane} \\ < 1 & \text{- powiazane} \end{cases}$$

H_o: $p_1 = p_2$ H_o: RR = 1 H_o: OR = 1

H_1: $p_1 \neq p_2$ H_1: RR ≠ 1 H_1: OR ≠ 1

W poszukiwaniu związku występowania chorób i czynnika ryzyka 3 modele są równoznaczne

Schemat badań przypadek-kontrola



RR i OR w badaniach przypadek-kontrola

- Nie możemy obliczać RR z badań typu przypadek – kontrola
- Możemy obliczać iloraz szans OR

Przykład badań przypadek-kontrola

- Związek między alkoholem i nowotworem przełyku
  - Grupa 200 przypadków (case) and 775 (kontrolna)
  - Pytamy o spożywanie alkoholu
- Ważne pytanie
  - Czy możemy obliczyć prawdopodobieństwo wystąpienia nowotworu przy spożywaniu więcej niż 80 g alkoholu dziennie na podstawie badań przypadek-kontrola (case-control)?

Wyniki tabela 2 x 2

	Alkohol (g/dzień)		
	> 80	< = 80	
Przypadek	96	104	200
grupa kontr.	109	666	775
	205	770	

OR w badaniach przypadek-kontrola

- W przykładzie alkohol/nowotwór przełyku:

$$\text{Estymator ilorazu szans } (\hat{\text{OR}}) = \frac{96 \times 666}{109 \times 104} = 5.64$$

- Interpretacja
  - U osobników z wysokim spożyciem alkoholu (> 80 gram/dzień) szansa wystąpienie nowotworu przełyku jest ponad pięciokrotnie wyższa niż szansa wystąpienia nowotworu przełyku u osobników z niższym spożyciem alkoholu

OR

- W jaki sposób sprawdzić, czy OR w populacji jest równe 1, czy też nie jest równe 1?
  - Dokładny test Fisher'a
  - χ^2 chi-kwadrat (test przybliżony)
- Obliczamy 95% przedział ufności dla OR w populacji.

Przedział ufności dla OR

95% CI dla ilorazu szans:

$$[\hat{OR} \cdot \exp(-1.96\sqrt{1/a + 1/b + 1/c + 1/d}\ );$$
$$\hat{OR} \cdot \exp(1.96\sqrt{1/a + 1/b + 1/c + 1/d}\ )]$$

95% przedział ufności i poziom krtytczny (p-value)

- 95% CI dla ilorazu szans wystąpienia nowotworu przełyku u osobników spożywających > 80 gramów alkoholu dziennie w porównaniu do spożywających 80 gramów lub mniej wynosi od 4 do 8
- Poziom krytyczny (p-value) dla OR = 1 jest <0.0001

OR

Tabela 2. Liczności obserwowane w badaniu złamań Źródło: Patrie 2006 str 43

		X: leczone HRT	X: nie leczone HRT	łącznie
Y	z złamaniem (chore)	40(a)	1287(b)	1327
	bez zlamania (kontrola)	239(c)	3023(d)	3262
	razem	279	4310	4589

O1=(40/1327)/(1- (40/1327)) # 0.031
O2=(239/3262)/(1- (239/3262)) #0.079
OR=O1/O2 # 0.39
(40/1287)/(239/3023) #0.39
(40*3023)/ (239*1287) #0.39

OR iloraz szans

$$O_i = p_i/(1-p_i)$$

$$OR = \frac{O_1}{O_2}$$

$$[OR * \exp(1.96 * \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}) \quad ; \quad OR * \exp(1.96 * \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}})]$$

Interpretacja

OR=1 brak zależności miedzy czynnikiem (HRT X) a zmienną objaśnianą (złamania Y)

OR>1 szkodliwy wpływ związany z narażeniem na czynnik

OR<1 protekcyjny charakter badanego czynnika względem wystąpienie zmiennej objaśnianej