Autor: Dariusz Piwczyński 2011-12-01 14:07

Analiza danych jakościowych – testy oparte o statystykę χ^2.

Cechy jakościowe są to cechy, których jednoznaczne i oczywiste scharakteryzowanie za pomocą liczb jest niemożliwe lub bardzo utrudnione.

Zanim rozpoczniemy analizę statystyczną konieczne jest ustalenie skali, w jakiej wyrażana jest nasza cecha!

Skale pomiaru (przypomnienie):

- **nominalna** – porządek właściwie dowolny

  Forma spędzania wolnego czasu

- **nominalna dychotomiczna, PŁEĆ, „CHORY, ZDROWY”**

- **porządkowa** (dokładniejsza), można przypisać rangi interwałowa (możemy przyporządkować poszczególnym obiektom wartość mierzoną w ściśle określonych jednostkach)

Przykład problemów:

1. Tablice liczebności

Wykształcenie	Liczność	Procent	Liczność Skumulowana	Procent Skumulowany
Podstawowe	13	5,10	13	5,10
Zawodowe	111	43,53	124	48,63
Średnie	91	35,69	215	84,31
Pomaturalne	5	1,96	220	86,27
Wyższe	35	13,73	255	100,00

2. Asocjacje, czyli badanie zależności między cechami

Czy humor Szefa związany jest z pogodą???

Skala nominalna

	☺	😐	☹	
☼	40	20	15	75
🌥	41	80	29	150
🌧	10	17	48	75
	91	117	92	300

Konkluzja: CHYBA JEST!!!

Asocjacja między dwiema zmiennymi nominalnymi istnieje, jeżeli rozkład jednej zmiennej ulega zmianie, gdy zmienia się poziom drugiej cechy (lub wartość) ulega zmianie.

	☺	😐	☹	
☼	24	26	25	75
🌥	50	49	51	150
🌧	25	25	25	75
	89	90	91	300

Konkluzja: Szef jest jak skała!!!

Asocjacje nie występują, gdy rozkład pierwszej zmiennej nie zależy od rozkładu drugiej zmiennej!

Weryfikacja H_0 oraz dobór testu

1. Tabele liczebności

H_0: Rozkład badanej cechy jest zgodny z teoretycznym
H_1: Rozkład badanej cechy nie jest zgodny z teoretycznym

Test X^2 – test zgodności (Liczebności rzeczywiste są zgodne z oczekiwanymi!)

$$X^2_{emp.} = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o – wartość otrzymana;
f_e – wartość oczekiwana

- Obliczone $X^2_{emp.}$ porównujemy z $X^2_{tab.}$ odczytanym z tabel statystycznych dla określonej liczby stopni swobody (df) i poziomu istotności.
- Jeżeli otrzymane X^2_{emp} jest większe lub równe wartości odczytanej z tabel to istnieją podstawy do odrzucenie hipotezy zerowej.
- Oznacza to, że analizowany rozkład nie jest zgodny z rozkładem teoretycznym.

df – liczba stopni swobody (liczba grup – 1). Przy 3 grupach df = 2, ponieważ różnice między f_o i f_e mogą się swobodnie kształtować tylko w dwóch grupach, trzecia grupa zdeterminowana jest przez sumę obserwacji dla wszystkich grup).

Założenia i ograniczenia testu X^2

- wartość oczekiwana w każdej klasie nie powinna być mniejsza niż 5. Przy większej liczbie klas i tak niskiej wartości oczekiwanej, pewne klasy można ze sobą połączyć. Dzięki temu istnieje szansa na zwiększenie tej wartości.
- Przy dwóch grupach, df = 1, stosujemy tzw. poprawkę Yates'a na nieciągłość. Wynika to z tego, iż rozkład X^2 jest rozkładem ciągłym, zaś frekwencje przyjmują liczby naturalne. Przy małych liczebnościach może to spowodować odrzucenie hipotezy zerowej z większym prawdopodobieństwem aniżeli założony poziom istotności.

$$X^2_{emp.} = \sum \frac{(|f_o - f|_e - 0{,}5)^2}{f_e}$$

Czy struktura płci studiujących osób na Wydziale Hodowli i Biologii Zwierząt jest zgodna z teoretycznym rozkładem płci w populacji ludzkiej, tj. 1:1?

Tabela liczności: plec (ankieta2009.sta)

Klasa	Liczba	Skumulow. Liczba	Procent	Skumulow. Procent
kobieta	226	226	71.74603	71.7460
mezczyzn	89	315	28.25397	100.0000

$$X^2_{emp.} = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(226 - 157{,}5)^2}{157{,}5} + \frac{(89 - 157{,}5)^2}{157{,}5} = 59{,}89$$

$X^2_{(df=1;\alpha=0{,}05)} = 3{,}841$

Obliczona wartość statystyki $X^2_{emp.}$ jest większa niż wartość odczytana z tabel (wartość krytyczna), co pozwala odrzucić hipotezę zerową i stwierdzić, że rozkład płci studiujących osób nie jest zgodny z rozkładem 1:1.

Przykład (EXCEL): (Wprowadzenie do statystyki dla przyrodników, Adam Łomnicki):

Czy istnieje związek między wiekiem ślimaków a skłonnością do przebywania w określonym siedlisku?

f_o (otrzymane)	odkryty grunt (G)	Roślinność zielona (R)	pnie drzew (D)	Σwieku
Młode (M)	52	43	17	112
Dorosłe (D)	108	15	74	197
Σsiedlisk	160	58	91	**309**

$$X^2_{emp.} = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o – wartość otrzymana;
f_e – wartość oczekiwana

Co oznaczają wartości rzeczywiste i oczekiwane?

Wartość oczekiwaną (f_e) obliczamy z proporcji:

$$\frac{f_e}{\text{suma kolumny, w której znajduje się } f_e} = \frac{\text{suma wiersza, w którym } f_e \text{ występuje}}{\text{suma wszystkich obserwacji}}$$

W naszym zadaniu wartość oczekiwaną liczby ślimaków młodych na odkrytym gruncie obliczamy następująco:

$$\frac{f_e}{\mathbf{160}} = \frac{\mathbf{112}}{309}$$

zatem $f_e = (160 \times 112) / 309 = \mathbf{57{,}99}$

Oczekiwana liczebność osobników na gruncie odkrytym wynosi 57,99. W sposób analogiczny obliczamy pozostałe wartości oczekiwane.

$X^2_{emp} = 49{,}163$

Wartość krytyczną statystyki X^2 odczytujemy z tabel statystycznych w zależności od ustalonego poziomu istotności i liczby stopni swobody.

Liczba stopni swobody obliczana jest następująco: $df = (k - 1) \times (w - 1)$, gdzie:
k – liczba kolumn (3 – odkryty grunt, roślinność zielona, pnie drzew)
w – liczba wierszy (2 – młode, dorosłe). W naszym przykładzie $df = 2$ $((3 - 1) \times (2 - 1))$
$X^2_{0,05} = 5{,}991$.
Ze względu na fakt, iż obliczona przez nas wartość statystyki X^2_{emp} jest większa niż wartość krytyczna odrzucamy hipotezę zerową.

2. Tablice kontyngencji (badanie zależności między cechami jakościowymi)

H_0: Między badanymi cechami nie ma zależności
H_1: Istnieje zależność między badanymi cechami

W celu weryfikacji tych hipotez wykorzystujemy na ogół test **X^2, ale nie zawsze...**

Liczebności	**Rodzaj testu:**
N > 40 i wszystkie liczebności oczekiwane > 10	**X^2**
N > 40 i którakolwiek liczebność oczekiwana < 5	**test X^2 z poprawką Yatesa**
20 < N =< 40 i wszystkie liczebności oczekiwane> 5	**test X^2 z poprawką Yatesa**
20 < N =< 40 i którakolwiek liczebność oczekiwana< 5	**dokładny test Fishera**
N =< 20 i którakolwiek liczebność oczekiwana < 5	**dokładny test Fishera**

- X^2 z poprawką Yatesa – współczynnik X^2 z poprawką Yatesa ze względu na niską liczebność w podgrupie. Powyższa poprawka powoduje bardziej ostrożną ocenę.
- Test Mantel-Haenszel przeznaczony jest do badania zależności między cechami wyrażonymi skalami porządkowymi.
  H_0: Brak porządkowej zależności między zmiennymi wierszy i kolumn
  H_1: Istnieje porządkowa zależność między zmiennymi wierszy i kolumn
- Współczynnik Fi (ϕ) = pierwiastek kwadratowy (X^2/n). Stosowany w odniesieniu do tabel 2x2. Przyjmuje wartości od 0 do 1. 0 - brak zależności, 1 – całkowita zależność. Stosowany w odniesieniu do zmiennych jakościowych.
- Statystyka V Cramera jest miarą siły zależności między badanymi cechami. Jej wartość zawiera się w przedziale od -1 do 1 w przypadku tabel dwudzielczych, zaś dla tabel większych przyjmuje wartości od 0 do 1.

0 **1**

brak zależności **pełna zależność**

Wyniki z SAS, Przykład (SAS EG) – ankieta.xls:

Sprawdź czy istnieje zależność między płcią a udzielonymi w ankiecie odpowiedziami (kwiek, papierosy, miejsce).

H_0: nie istnieje zależność między płcią osób a treścią udzielanych odpowiedzi
H_1: istnieje zależność między płcią osób a treścią udzielanych odpowiedzi

Tabela plec wg papierosy				
		papierosy		
		nie	tak	Razem
plec				
kobieta	Liczebność	193	53	246
	Proc. wier.	78.46	21.54	
	Proc. kol.	69.93	72.60	
mezczyzn	Liczebność	83	20	103
	Proc. wier.	80.58	19.42	
	Proc. kol.	30.07	27.40	
Razem	Liczebność	276	73	349
Liczebność braków danych = 3				

Statystyki dla tabeli plec wg papierosy

Statystyka	St. sw.	Wartość	Prawd.
Chi-kwadrat	1	0.1986	0.6558
Chi-kw. ilorazu wiarygodn.	1	0.2007	0.6541
Poprawka uciągl. chi-kwadrat	1	0.0908	0.7631
Chi-kwadrat Mantela-Haenszela	1	0.1980	0.6563
Współczynnik FI		-0.0239	
Współczynnik wielodzielczości		0.0238	
V Cramera		-0.0239	

Prawdopodobieństwo, jakie wynika z przeprowadzonego testu X^2 (0,6558) nie pozwala odrzucić hipotezy zerowej i stwierdzić, że istnieje zależność istotna między płcią osób a odpowiedzią na pytanie: czy palisz papierosy? Zestawione w tabeli wyniki pozwala stwierdzić, że wśród kobiet udział palących wyniósł 21,546% i był około 2 j. p. niższy niż wśród mężczyzn (19,42%).

STATISTICA

Podsumowująca tabela dwudzielcza: częstości obserwowane (ankieta2009)
Liczność oznacz. komórek > 10

plec	slodycze tak	slodycze nie	Wiersz Razem
kobieta	210	16	226
%kolumny	73.68%	53.33%	
%wiersza	92.92%	7.08%	
mezczyzn	75	14	89
%kolumny	26.32%	46.67%	
%wiersza	84.27%	15.73%	
Ogół	285	30	315

Statystyka: plec(2) x slodycze(2) (ankieta2009)

statystyka	Chi-kwadr.	df	p
Chi^2 Pearsona	5.545548	df=1	p=.01853
Chi^2 NW	5.095818	df=1	p=.02399
Chi^2 Yatesa	4.587049	df=1	p=.03222
dokł. Fishera, 1-stronny			p=.01886
2-stronny			p=.03087
Chi^2 McNemara (A/D)	169.7545	df=1	p=0.0000
(B/C)	36.96703	df=1	p=.00000

Dokładny test Fishera

(wartości empiryczne)

1	**3**	**4**
0	**5**	**5**
1	**8**	**9**

Możliwe wartości

0	**4**	**4**
1	**4**	**5**
1	**8**	**9**

W ramach testu Fishera obliczane jest prawdopodobieństwo otrzymania danego rozkładu z tablicy. Rozpatrywane są wszelkie możliwe kombinacje liczebności komórek w oparciu o liczebności brzegowe. Prawdopodobieństwo związane z dokładnym testem Fishera wykazuje tendencje do przyjmowania wyższych wartości, aniżeli asymptotyczny test X^2, ponieważ jest testem bardziej konserwatywnym.

Czy istnieje zależność wśród osób chorych na trądzik między płcią a ubytkami naskórka?

Tabela plec wg ubytki_n			
plec	**ubytki_n**		
Liczebność Procent Proc. wier. Proc. kol.	**0**	**1**	**Razem**
kobieta	68 68.00 95.77 74.73	3 3.00 4.23 33.33	71 71.00
mezczyzna	23 23.00 79.31 25.27	6 6.00 20.69 66.67	29 29.00
Razem	91 91.00	9 9.00	100 100.00

Statystyka	**St. sw.**	**Wartość**	**Prawd.**
Chi-kwadrat	1	6.8149	0.0090
Chi-kw. ilorazu wiarygodn.	1	6.0824	0.0137
Poprawka uciągl. chi-kwadrat	1	4.9529	0.0260

Statystyka	St. sw.	Wartość	Prawd.
Chi-kwadrat Mantela-Haenszela	1	6.7467	0.0094
Współczynnik FI		0.2611	
Współczynnik kontyngencji		0.2526	
V Cramera		0.2611	
OSTRZEŻENIE: 25% komórek ma oczekiwane liczby wyst. mniejsze niż 5. Chi-kwadrat może nie być właściwym testem.			

Pojawiło się ostrzeżenie ze strony programu SAS o niewystarczających liczebnościach oczekiwanych. Decyzję o odrzuceniu hipotezy zerowej odczytujemy zatem w oparciu o test z poprawką Yates'a (**Poprawka uciągl. chi-kwadrat)**, a w skrajnych przypadkach o dokładny test Fishera.

Dokładny test Fishera	
Komórka (1,1) liczebność (F)	68
Lewostronne pr. <= F	0.9978
Prawostronne pr. >= F	0.0165
Tabela prawdopodobieństwa (P)	0.0143
Dwustronne pr. <= P	0.0165

Test McNemary

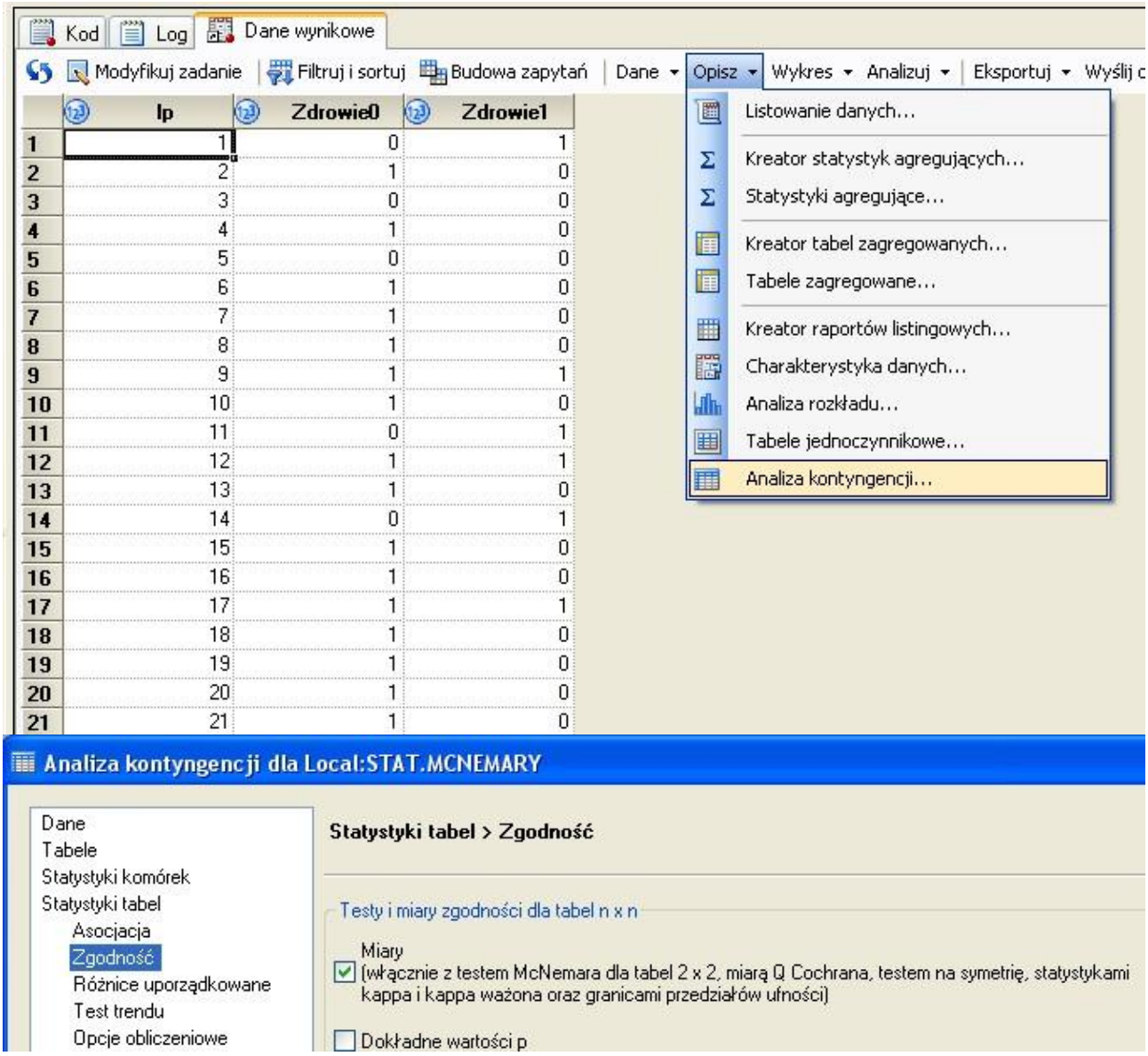
$H_0 : p_1 = p_2$
$H_1 : p_1 \neq p_2$

p_1 – proporcja pacjentów, którzy wykazują objawy choroby przed leczeniem
p_2 – proporcja pacjentów, którzy wykazują objawy choroby po leczeniem

$$\chi^2 = \frac{(A-D)^2}{A+D}$$

A – Liczba osób wśród których doszło do zmiany wyniku z dodatniego na ujemny
D – Liczba osób wśród których doszło do zmiany wyniku z ujemnego na dodatni

Wykorzystywany do analizy tabel 2x2, gdy doświadczenie zorganizowane jest w układzie zależnym. Przykładem niech będzie wpływ prowadzonej kuracji na stan zdrowia pacjentów pacjentów. 1 – pacjent zdrowy; 0 – pacjent chory.



	lp	Zdrowie0	Zdrowie1
1	1	0	1
2	2	1	0
3	3	0	0
4	4	1	0
5	5	0	0
6	6	1	0
7	7	1	0
8	8	1	0
9	9	1	1
10	10	1	0
11	11	0	1
12	12	1	1
13	13	1	0
14	14	0	1
15	15	1	0
16	16	1	0
17	17	1	1
18	18	1	0
19	19	1	0
20	20	1	0
21	21	1	0

Wyniki

0 – pacjent bez objawów choroby; 1 – pacjent z objawami choroby

Rezultaty analizy kontyngencji

Procedura FREQ

Tabela Zdrowie0 wg Zdrowie1				
		Zdrowie1		
		0	1	Razem
Zdrowie0				
0	Liczebność	14	7	21
	Proc. wier.	66.67	33.33	
1	Liczebność	37	12	49
	Proc. wier.	75.51	24.49	
Razem	Liczebność	51	19	70

Statystyki dla tabeli Zdrowie0 od Zdrowie1

Test McNemara	
Statystyka (S)	20.4545
DF	1
Pr. > S	<.0001

Współczynnik kappa zwykły	
Kappa	-0.0628
Std. błąd asympt.	0.0858
Dolna granica prz. ufn. 95%	-0.2310
Górna granica prz. ufn. 95%	0.1054

Wielkość próby = 70

Przeprowadzony test zgodności McNemara pozwala wnioskować, że przeprowadzona kuracja wpłynęła statystycznie na stan zdrowia pacjentów.