TABELE WIELODZIELCZE

W wielu badaniach gromadzimy dane będące liczebnościami. Przykładowo możemy klasyfikować chore zwierzęta w badanej próbie do różnych kategorii pod względem wieku, płci czy skali natężenia choroby. Przedstawiane do tej pory metody statystyczne stają się bezużyteczne dla danych tego typu, zwanych danymi jakościowymi. Testy i techniki statystyczne prezentowane na tych ćwiczeniach należą do najbardziej przydatnych technik analizy danych jakościowych. Techniki te umożliwiają również dokonania oceny zależności pomiędzy zmiennymi tego typu.

Pierwszym krokiem w analizach, o których tu mowa jest przedstawienie zebranych danych indywidualnych w postaci **tablicy wielodzielczej.** Wymaga to zliczenia jednostek w odpowiednich komórkach tabeli z danymi. Zliczanie to bez użycia komputera jest żmudne i męczące zwłaszcza dla dużej ilości przypadków. Tablice wielodzielcze stanowią bowiem podstawę do obliczania pozostałych statystyk określających siłę związku.

Tablica wielodzielcza przedstawia rozkład obserwacji ze względu na kilka cech jednocześnie. Dla dwu zmiennych tablica wielodzielcza pokazuje łączny rozkład obu cech. Liczebności w ostatnim wierszu i w ostatniej kolumnie nazywamy empirycznymi brzegowymi rozkładami, odpowiednio cechy Y i cechy X.

Przykładowo chcemy ocenić czy miejsce przybywania psa (miasto, wieś) wpływ wykonanie obowiązujących szczepień. Wykorzystamy w tym celu zebrane dane w naszej przykładowej bazie 40 psów.

Zliczając otrzymane dane dla miejsca przebywania i szczepienia otrzymamy następującą tablicę wielodzielczą:

	Szczepienie Tak	**Szczepienie Nie**	**Razem**
Miasto	25	3	28
Wieś	7	5	12
Razem	32	8	40

W tabeli zacieniowano rozkłady brzegowe. Z tabeli widać wyraźną przewagę psów z miasta. Z kolei cztery razy więcej jest psów szczepionych niż nieszczepionych. Informacje byłyby bogatsze po dołączenie danych procentowych. Stosuje się procenty liczone względem ostatniej kolumny (względem miejsca przebywania), względem ostatniego wiersza (względem szczepienia) oraz względem całkowitej liczby zwierząt.

Następny etap analizy statystycznej tak zebranych danych, to próba weryfikacji hipotezy mówiącej, że dwie jakościowe cechy w populacji są niezależne. Najczęściej stosowanym „narzędziem" jest test χ^2. Został on opracowany przez Karla Pearsona w 1900 r. i jest metodą, dzięki której można się upewnić, czy dane zawarte w tablicy wielodzielczej dostarczają wystarczającego dowodu na związek tych dwóch zmiennych. Test χ^2 polega na porównaniu częstości zaobserwowanych z częstościami oczekiwanymi przy założeniu hipotezy zerowej (o braku związku pomiędzy tymi dwiema zmiennymi).

Interesuje nas weryfikacja hipotezy zerowej:

$\mathbf{H_0}$ **:** cechy X i Y są niezależne

Wobec hipotezy alternatywnej:

$\mathbf{H_1}$ **:** cechy X i Y są zależne

Do weryfikacji hipotezy stosujemy statystykę:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

gdzie E - oczekiwana częstość komórki oraz O - obserwowana częstość komórki. Przy założeniu hipotezy zerowej opisywana statystyka ma rozkład χ^2 o s = (k - 1)(p - 1) stopniach swobody. Częstości oczekiwane obliczamy wykorzystując częstości marginalne (z tablicy wielodzielczej) według następującego wzoru:

$$\text{E (częstość oczekiwana)} = \frac{(\text{suma rzędu})(\text{suma kolumny})}{(\text{suma całkowita})}$$

Dla tabel dwudzielczych 2x2 postaci

a	b
c	d

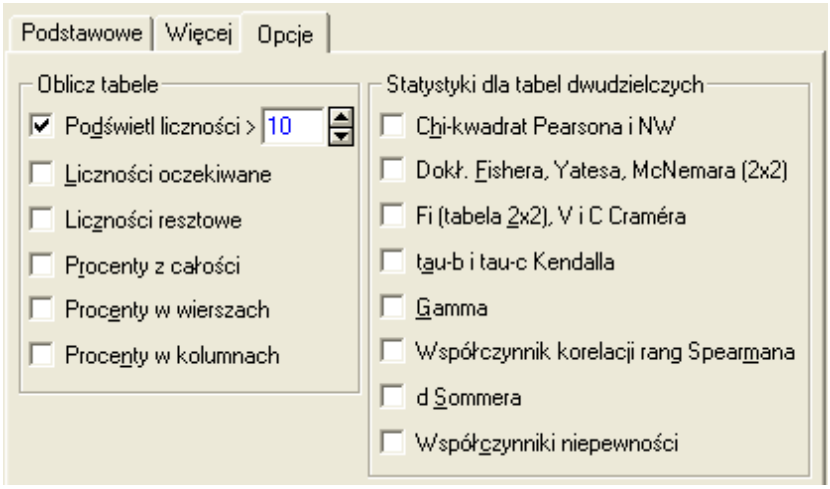
wartość statystyki χ^2 wyznaczamy według prostszego, praktycznego wzoru:

$$\chi^2 = \frac{(ad-bc)^2 N}{(a+b)(c+d)(a+c)(b+d)}$$

Przykładowo dla naszych danych chcemy ocenić, czy szczepienie jest powiązane z miejscem przebywania zwierzęcia. Wyniki obliczeń wartości oczekiwanych przedstawiono w nawiasach obok wartości obserwowanych.

	Szczepienie Tak	**Szczepienie Nie**	**Razem**
Miasto	25 (22,4)	3 (5,6)	28
Wieś	7 (9,6)	5 (2,4)	12
Razem	32	8	40

Oczywiście nie przeprowadzamy weryfikacją „na piechotę”. W praktyce posługujemy się komputerem. W programie *STATISTICA* do analizy tablic wielodzielczych służy opcja **Tabele wielodzielcze** w module **Podstawowe statystyki i tabele**. W module tym możemy wybrać dwie grupy statystycznych analiz dotyczących tablic zbiorczych oraz tablic wielokrotnych odpowiedzi. Możemy utworzyć tabele wielodzielcze i zbiorcze oraz obliczyć różne statystyki związane z takimi tabelami. W module tym możemy analizować tabele dowolnych rozmiarów niekoniecznie 2x2 jak w poprzednim module. Możemy też wybrać jakie podsumowania i jakie statystyki chcemy policzyć. Kartę z możliwymi opcjami pokazuje poniższy rysunek.



Rys. 1 Arkusz wyboru opcji dla testu χ^2

Znajdująca się tam opcja **Chi-kwadrat Pearsona i NW oraz FI, V i C Cramera** umożliwia obliczenie statystyki χ^2 oraz innych statystyk z nią związanych dla tablic wielodzielczych.

Dla danych z naszego przykładu otrzymujemy następujący arkusze wyników.

Podsumowująca tabela dwudzielcza: częst
Liczność oznacz. komórek > 10

Przebywanie	Szczepienie 2 Tak	Szczepienie 2 Nie	Wiersz Razem
Miasto	25	3	28
%kolumny	78,13%	37,50%	
%wiersza	89,29%	10,71%	
%ogółu	62,50%	7,50%	70,00%
Wieś	7	5	12
%kolumny	21,88%	62,50%	
%wiersza	58,33%	41,67%	
%ogółu	17,50%	12,50%	30,00%
Ogół	32	8	40
%ogółu	80,00%	20,00%	100,00%

Statystyka: Przebywanie(2) :

statystyka	Chi-kwadr.	df	p
Chi^2 Pearsona	5,029762	df=1	p=,02492
Chi^2 NW	4,663568	df=1	p=,03081
Chi^2 Yatesa	3,281250	df=1	p=,07008
dokł. Fishera, 1-stronny			p=,03857
2-stronny			p=,03857
Chi^2 McNemara (A/D)	12,03333	df=1	p=,00052
(B/C)	,9000000	df=1	p=,34278
Fi dla tabel 2 x 2	,3546041		
Korel. tetrachoryczne	,5738961		
Wsp. kontyngencji	,3342134		

Rys. 1 Arkusze wyników dla testu χ^2

W pierwszych oknie (po lewej) mamy tabelę z danymi wraz sumami brzegowymi oraz procenty wszystkich wartości wyliczane w stosunku do całkowitej liczebności grupy oraz procenty z kolumn i wierszy. Drugie okno (po prawej) to wartości statystyki χ^2 oraz jej modyfikacje (związane z liczebnością próby) wraz z poziomami istotności. Przykładowo, gdy ogólna liczebność próby jest mała (N<40) i którakolwiek z liczebności oczekiwanych jest < 5 to stosujemy dokładny test Fishera.
Arkusz wyników podaje również wartość współczynnika Φ - Yula (omówiony poniżej) oceniający siłę powiązania pomiędzy dwoma zmiennymi w tabeli 2x2. Jak widzimy powiązanie pomiędzy miejscem pobytu a szczepieniem jest istotne (p = 0,02492), ale słabe (Φ = 0,354). Mamy tym samym podstawy wnioskować, że nieszczepione regularnie psy częściej występują na wsi niż w mieście.
Zauważmy, że bardzo duże wartości χ^2 oznaczają dużą różnicę pomiędzy częstościami obserwowanymi a oczekiwanymi i jest to dowód istnienia zależności. Przeciwnie mała wartość χ^2 (zwłaszcza bliska 0) nie daje dowodu na istnienie korelacji.

UWAGI !

- Dla tabeli 2x2 statystyka χ^2jest często modyfikowana w celu utworzenia bardziej odpowiedniego testu. W większości statystycznych programów komputerowych mamy możliwości obliczenia tych poprawek. Najbardziej popularna to poprawka Yatesa. Stosujemy ją, jeżeli 20 < N <40 i którakolwiek z liczebności oczekiwanych jest mniejsza od 5.
- Statystyka χ^2 sprawdza czy dwie zmienne są ze sobą powiązane. Jednakże oprócz sprawdzenia czy pomiędzy zmiennymi zachodzi związek, interesuje nas jak silne jest to powiązanie. W praktyce najczęściej korzystamy z następujących miar:

    1. Współczynnik Φ - Yula. Jest on miarą korelacji pomiędzy zmiennymi jakościowymi w tabeli 2x2. Przyjmuje on wartości od 0 (brak powiązania) do 1 (doskonałe powiązanie pomiędzy zmiennymi)

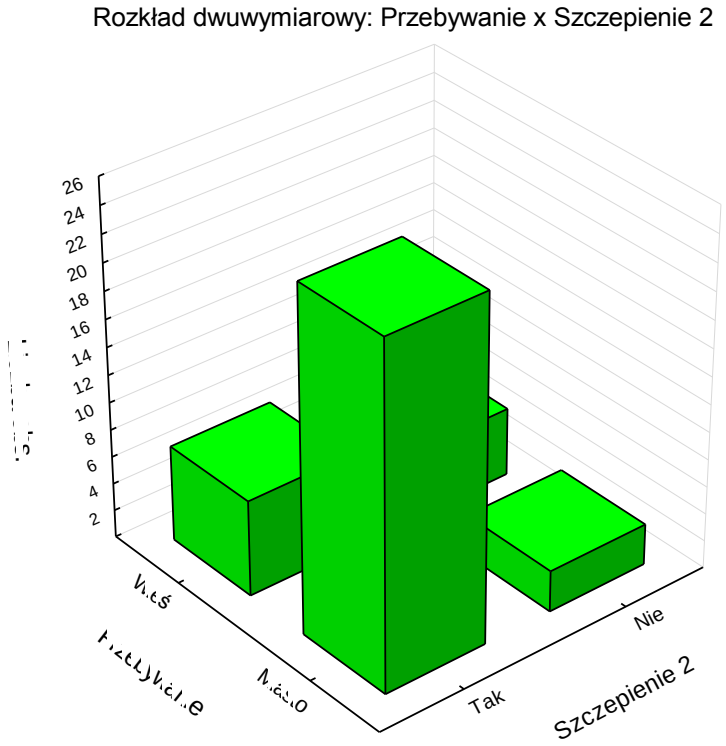
    2. Współczynnik V – Cramera. Przyjmuje on wartości od 0 (brak relacji między zmiennymi) do 1

Interpretacja wszystkich tych współczynników jest taka sama:

- jeżeli przyjmują one wartość zero to cechy X i Y są niezależne

- im bliższe jedności są wartości tych współczynników tym silniejsze jest powiązanie pomiędzy analizowanymi cechami X i Y.

Program udostępnia nam również interpretacje graficzne analizowanych problemów. Przykładowy wykres dla danych opisujących powiązanie między sczepieniem a miejscem przebywania pokazuje poniższy rysunek.



Rys. 3 Trójwymiarowy wykres częstości

Porównanie dwóch wskaźników struktury (proporcji)

Badając dwie populacje ze względu na cechę jakościową, musimy często sprawdzać hipotezę, że wskaźniki struktury (procenty) są w obu populacjach takie same. Podany niżej test pozwala na zweryfikowanie tej hipotezy w oparciu o wyniki dwu dużych prób. W zależności od postaci hipotezy alternatywnej, rozpatrujemy obszar krytyczny dwustronny albo też jednostronny.

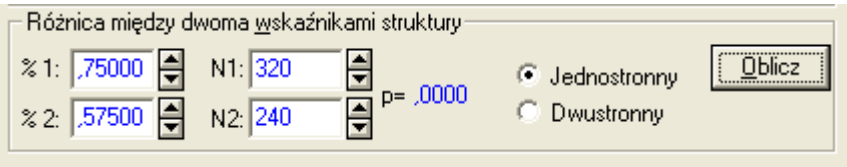
Przykład 20

Badano wpływ nowej szczepionki przeciwko parwowirozie. W tym celu 320 wylosowanym psom zaaplikowano szczepionkę i u 240 chorych stwierdzono po ustalonym okresie leczenia powrót do normy. Natomiast w drugiej 200 osobowej grupie chorych psów, gdzie leczono tradycyjną szczepionką, stwierdzono powrót do normy u 115 psów. Zweryfikujmy hipotezę o większym procencie wyzdrowień w grupie psów leczonych nową szczepionką. Oto kolejne kroki postępowania.

1. Z menu **Statystyka** wybieramy opcję **Statystyki podstawowe i tabele**. Następnie w otwierającym się oknie wybieramy opcję **Inne testy istotności**.
2. W polu **Różnica między dwoma wskaźnikami struktury:**
   a. wprowadzamy konkretne dane z próby (tak jak na rysunku 4),

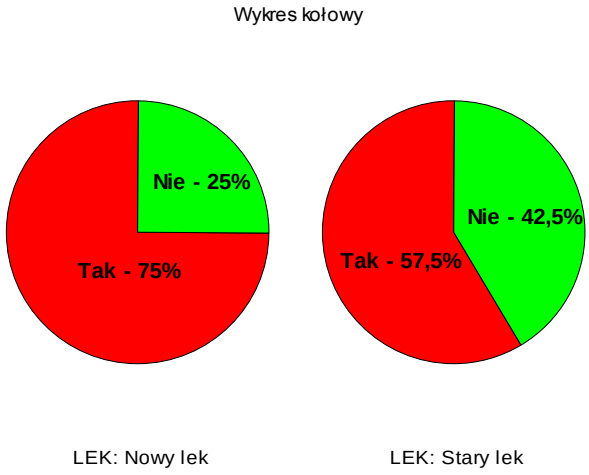
b. wybieramy opcję **Jednostronne** lub **Dwustronne** (w naszym przypadku jednostronne, bo hipoteza alternatywna ma postać H_1: $p_1 > p_2$),

c. naciskamy przycisk **Oblicz.** Otrzymany arkusz wyników pokazuje poniższy rysunek.



Rysunek 4 Okno z wynikami testu dla dwóch wskaźników struktury.

3. Możemy dołączyć interpretacje graficzną w postaci wykresu kołowego przedstawionego na rysunku 66. Otrzymamy go wybierając z menu **Wykresy** opcję **Wykresy skategoryzowane** a następnie **Wykresy kołowe**. W otwierającym się oknie **Skategoryzowane wykresy kołowe** określamy zmienne i interesujące nas opcje. Wszystkie dostępne tu opcje są doskonale opisane w rozbudowanej **Pomocy** pakietu *STATISTICA*. Wykres otrzymamy po kliknięciu na przycisku OK.



Rys. 5 Interpretacja graficzna dla porównania proporcji