Analiza zależności zmiennych jakościowych. Testy chi - kwadrat.

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska

Plan wykładu

1. Analiza danych jakościowych
   - Zmienne jakościowe (nominalne i inne …)
   - Tablice wielodzielcze
2. Analiza tablic wielodzielczych – zależność zmiennych
3. Test zgodności chi - kwadrat
   - Prosta hipoteza o zgodności rozkładów
   - Statystyka chi - kwadrat, …
4. Test niezależności dwóch zmiennych
   - Analiza tablic wielodzielczych z wykorzystaniem testu chi-kwadrat
   - Poprawki w testach
5. Jak to się liczy w różnych programach?
6. Miary siły związku

Analiza danych jakościowych

- Dotychczas → omawiano zmienne ilościowe i procedury służące do ich statystycznych analizy.
- W praktyce często stosujemy także **zmienne jakościowe**!
- Przykłady:
  - Płeć, status małżeński, zawód, kolor, marka towarowa, rodzaj lub profil firmy,…
  - upodobanie do marki handlowej, gatunku towarów, partii politycznej, umowna skala natężenia choroby,…
- Często występują w analizie danych pochodzących z różnych ankiet czy badania opinii ludzi, …, lecz także inne zastosowania.

Dotychczas w zakresie tego przedmiotu:

- Wykłady 1 i 2 → właściwe metody statystyki opisowej.

Potrzeba metod wnioskowania statystycznego dla takich danych!

Typy danych jakościowych

- Skale pomiarowe – nominalne i porządkowe.
- Rzeczywiste zmienne jakościowe:
  - Dane (zmienne) nominalne,
  - Zmienne o wartościach uporządkowanych.

Uwaga!

- Zapis liczbowy tych zmiennych → kodowanie!

- Ponadto, zmienne ilościowe mogą być zdyskretyzowane!
  - np. wielkość miasta (wg. liczby mieszkańców), wielkość dochodów,
- Ogólnie mówimy o zmiennych skategoryzowanych!
- Dalsze rozważania → zmienne nominalne.

Analiza pary zmiennych jakościowych

- Ocena zależności między zmiennymi jakościowymi:
- Przedstawienie danych jednostkowych w postaci tablicy wielodzielczej.
  - Inne nazwy → tablica kontyngencji (ang.); dla pary zmiennych → tablica dwudzielcza.

Założenia:

- Zmienna *X* ma *k* kategorii kodowanych jako $x_1, x_2, \ldots, x_k$
- Zmienna *Y* ma l kategorii kodowanych jako $y_1, y_2, \ldots, y_l$
- Próba liczy *n* par (*x*,*y*)
- n_{ij} jest liczbą wystąpienia w próbie par obserwacji (x_i, y_j)

$\sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = n$

Tablica wielodzielcza

- Macierz o *k* wierszach i l kolumnach z elementami n_{ij} na przecięciu *i*-tego wiersza i *j*-tej kolumny, *i* = 1,2, …, *k* oraz *j* = 1,2, …, l.

	y_1	y_2	…	y_j	…	y_l
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1l}
x_2	n_{21}	n_{21}	…	n_{2j}	…	n_{2l}
…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{il}
…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kl}

- W analizie zależności często oblicza się rozkłady brzegowe.

Tablice wielodzielcze – przykłady

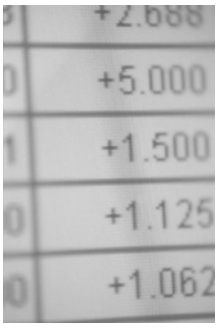
- Dane ankietowe nt. oceny wpływu używek (papieros, kawa, alkohol) oraz płeć na pewną chorobę, …, 90 osób.

Lp.	Kawa	Papierosy	Alkohol	Płeć
1	Nigdy	Dużo	Niewiele	M
2	Niewiele	Nigdy	Nigdy	K
3				
4				
5				
…				
87				
88				
89				
90	Średnio	Średnio	Nigdy	K

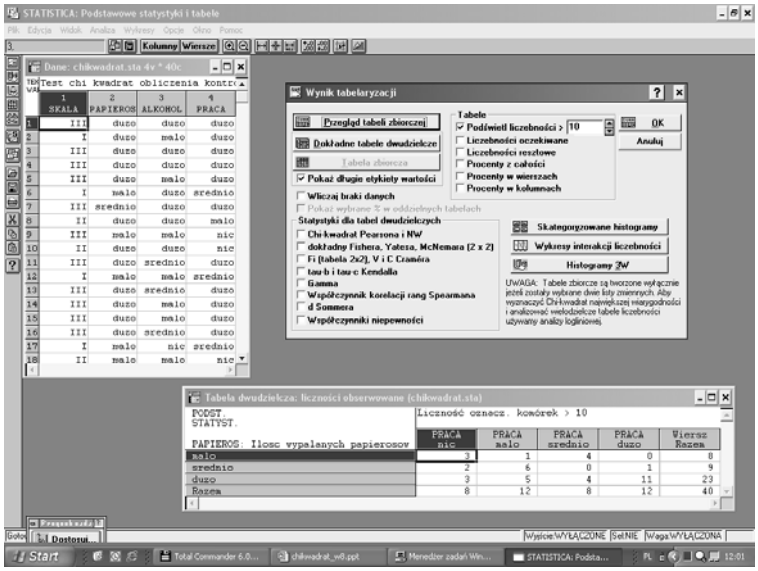
Płeć	Papieros Nigdy	Papieros Niewiele	Papieros Średnio	Papieros Dużo	Suma
Kobieta	**11**	**8**	**6**	**5**	30
Mężczyzna	**4**	**4**	**28**	**24**	60
Suma	15	12	34	29	90

Jak tworzyć tablice automatycznie

- Czy mamy wsparcie ze strony oprogramowania?

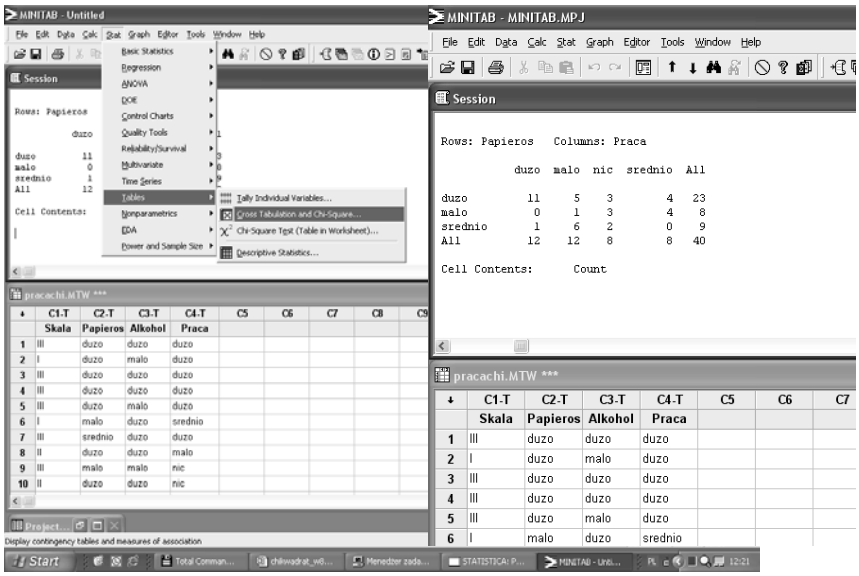


Statistica – dobrze wspiera! (www.statsoft.pl)



www.statsoft.com

Mini-tab -> wspiera



Ale co w Excel-u?

Poczekaj i sprawdź na laboratorium!

Analiza współzależności zmiennych w tablicach …

- Przykładowa tablica zestawiająca osoby wg. wykształcenia i miejsca zamieszkania:

Wykształcenie	**Miejsce zamieszkania**	
	Miasto	Wieś
Podstawowe lub zawod.	100	195
Średnie lub wyższe	350	55

- Czy patrząc na zawartość tablicy możesz wyciągnąć pewne wnioski?

Inne przykłady tabel wielodzieczych

	A	¬A
B	49	2
¬B	1	48

	A	¬A
B	23	27
¬B	26	24

- Co możesz wywnioskować z analizy zawartości powyższych tabel?
- Narzędziem analizy rozkładów wartości jest test wykorzystujący statystykę o rozkładzie χ^2

Schemat wykonywania testu chi-kwadrat

- Dotyczy badania zgodności obserwowanego rozkładu z rozkładem zadanym, a także niezależności zmiennych.
- Sformułuj przypuszczenie co do populacji przez określenie odpowiednich hipotez.
- Oblicz częstości występowania pewnych zdarzeń spodziewanych przy założeniu prawdziwości H_0 $\rightarrow$ tzw. liczności oczekiwane w różnych klasach (kategoriach).
- Zanotuj zaobserwowane liczności punktów pomiarowych w poszczególnych klasach.
- Zbadaj różnicę między wartościami obserwowanymi i oczekiwanymi $\rightarrow$ wzór na statystykę χ^2
- Podejmij decyzję (poziom istotności, l. st. swobody).

Statystyka chi-kwadrat

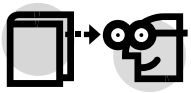
$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

gdzie

- O_i – obserwowana liczność w klasie (kategorii) i.
- E_i – hipotetyczna liczność w klasie (kategorii) i ; oczekiwana przy założeniu prawdziwości H_0.

Statystyka na rozkład χ^2 z odpowiednia liczbą stopni swobody.

Więcej informacji o rozkładzie χ^2 w literaturze!



Testy zgodności rozkładu

- Jest to procedura statystyczna pozwalająca na ustalenie czy dostępne dane potwierdzają założenie o określonym rozkładzie w populacji interesującej nas zmiennej.
- Innymi słowami: czy rozkład zmiennej jakościowej jest zgodny z pewnym rozkładem zadanym.
- Przykład prostego testu zgodności dla przypadku rozkładu wielomianowego
  - mamy $k > 2$ kategorii / klas i prawdopodobieństwo, że punkt / obserwacja należy do i-tej kategorii jest równa p_i.

Prosty test zgodności rozkładu

- Niech zmienna jakościowa *X* ma *k* możliwych wartości (kategorii) $x_1, x_2, \ldots, x_k$ i niech prawdopodobieństwo wystąpienia wartości x_i wynosi p_i dla *i* = 1, 2, …, *k*.
- Zakładamy, że wartości p_i w populacji są nieznane.
- Ponadto niech będzie dany pewien ustalony rozkład prawdopodobieństwa $\{p_1^0, p_2^0, \ldots, p_k^0\}$.
- Rozważamy problem testowania hipotezy o zgodności rozkładu $\{p_1, p_2, \ldots, p_k\}$ z zadanym rozkładem $\{p_1^0, p_2^0, \ldots, p_k^0\}$
- Hipotezy:

  $H_0 : p_i = p_i^0$ dla *i* = 1, 2, …, *k*

  H1: hipoteza zerowa jest fałszywa.

Przykład → preferencja co do koloru.

- Producent zegarków przed wprowadzeniem nowego modelu chce sprawdzić, czy ludzie mają specjalne preferencje co do koloru paska do zegarka, lub czy też wszystkie 4 rozpatrywane kolory są tak samo lubiane.
- Wybrano losowo próbę 80 osób planujących zakup zegarka; każdej z nich pokazano model zegarka z 4 wersjami kolorystycznymi i poproszono o wybór jednej.

Piaskowy	Brązowy	Kasztanowy	Czarny	Suma
12	40	8	20	

- H0: wszystkie kolory pasków do zegarka są jednakowo preferowane, tj. prawdopodobieństwa wyboru są

  $p_1 = p_2 = p_3 = p_4 = 0{,}25$

- H1: nie wszystkie kolory są tak samo preferowane.

Przykład preferencji co do koloru …

Kilka pytań:

- Jak obliczyć liczności oczekiwane E_i ?
- Jaka jest wartość statystyki testowej χ^2 ?
- Jaka jest liczba stopni swobody dla obliczenia wartości krytycznych rozkładu χ^2 ?

Oczekiwana liczność w *i*-tej klasie / kategorii:

$$E_i = n \cdot p_i$$

Liczba stopni swobody *k* - 1

Założenia wykonywania testu chi - kwadrat

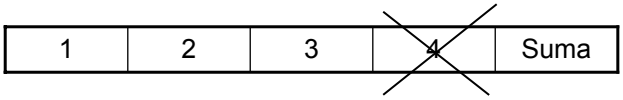
Przy jakich założeniach rozkład statystyki testu jest dobrze przybliżony przez rozkład χ^2 ?

- Im większe *n*, tym przybliżenie jest lepsze.
- Ponadto oczekiwana liczność klas nie może być zbyt mała!

- Rozkład chi – kwadrat można stosować, gdy oczekiwana liczność w każdej klasie jest równa przynajmniej 5.
  - Co robić jeśli dla jednej lub kilku klas, oczekiwana liczba elementów jest < 5?
  - Połączyć klasy tak aby otrzymać większą liczność.
- Więcej → Aczel A. Statystyka w zarządzaniu, str. 751.

Określania liczby stopni swobody

- Powróćmy do przykładu kolorystycznego



- Sumaryczna liczność – pozwala nie znać jednej jakiejkolwiek liczności klasy;
  czyli redukuje liczbę stopni swobody o 1!

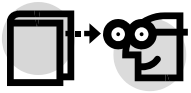
Ogólna zasada (Aczel, str. 753):

- Jeśli wykorzystuje się dane do estymacji parametrów rozkładu prawdopodobieństwa określonego przez hipotezę zerową, to dla każdego parametru estymowanego na podstawie tych danych traci się dodatkowy stopień swobody.

Zadanie domowe – genetyczne!

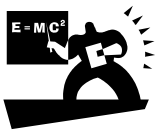
- Zajrzyj do książki Koronacki, Mielniczuk: Statystyka, str. 367 $\rightarrow$ przykład 6.3.
- Przykład dotyczy analizy doświadczeń **Gregora Mendla** z dziedziny genetyki na przykładzie grochu o określonym genotypie.
- Test zgodności można wykorzystać do sprawdzenia, czy rzeczywiste wyniki doświadczenia z grochem (potomkowie grochu zielonego o genotypie *aa*) potwierdziły tezę Mendla o dziedziczeniu!

- Nie czekaj, sprawdź w książce o co chodzi!
- Ponadto, rozdziały 6.2 – 6.3 zawierają opisy innych bardziej złożonych testów.



Gdzie jesteśmy w trakcie wykładu?

1. Analiza danych jakościowych
   - Zmienne jakościowe (nominalne i inne …)
   - Tablice wielodzielcze
2. Analiza tablic wielodzielczych – zależność zmiennych
3. Test zgodności chi - kwadrat
   - Prosta hipoteza o zgodności rozkładów
   - Statystyka chi - kwadrat, …
4. **Test niezależności dwóch zmiennych**
   - Analiza tablic wielodzielczych z wykorzystaniem testu chi-kwadrat
   - Poprawki w testach
5. Jak to się liczy w różnych programach?
6. Miary siły związku



Analiza dwóch zmiennych losowych – testowanie niezależności

Założenia

- Rozważamy populację opisaną przez parę jakościowych zmiennych losowych.
- Dysponujemy *n*-elementową próbą, gdzie każda obserwacja musi należeć do jednej z *kl* możliwych kombinacji kategorii pierwszej i drugiej zmiennej.
- Otrzymane informacje prezentujemy w postaci tablicy dwudzielczej.
- Niech $p(i,j)$ oznacza prawdopodobieństwo zaobserwowania w jednym doświadczeniu *i*-tej kategorii *X* oraz *j*-tej kategorii *Y*.

Przykład:

- Zbadajmy czy choroba wieńcowa współwystępuje z podwyższonym ciśnieniem tętniczym dla grupy osób po 50tce (za Stanisz, str. 227).

Choroba	Ciśnienie podwyższone		Suma
	Nie	Tak	
Nie	37	17	54
Tak	8	38	46
Suma	45	55	100

Test niezależności chi-kwadrat

Schemat postępowania (Karl Pearson 1900r)

- Hipotezy:
  H_0: Zmienne *X* i *Y* są wzajemnie niezależne,
  H_1: Zmienne *X* i *Y* są zależne.
- Statystyka testowa

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  gdzie *E* oczekiwana i *O* obserwowana liczności komórki
- Przy założeniu hipotezy zerowej statystka ma asymptotyczny rozkład $\chi 2$ o $df = (k - 1)\cdot(l - 1)$ stopniach swobody.
- Dla założonego poziomu istotności α podjęcie decyzji.

Jak obliczać wartość oczekiwaną?

- Liczności oczekiwana obliczamy wykorzystując rozkłady brzegowe

$$E = \frac{(\text{suma rzędu}) \cdot (\text{suma kolumny})}{(\text{suma})}$$

- Skąd się to wzięło?
- Zasada niezależności prawdopodobieństwa

$$P(i \cap j) = P(i) \cdot P(j)$$

- więc

$$E_{ij} = n \cdot p(i) \cdot p(j) = n \cdot (R_i / n) \cdot (C_j / n) = R_i \cdot C_j / n$$

Policzmy przykład

Wartości obserwowane

Choroba	Ciśnienie podwyższone		Suma
	Nie	Tak	
Nie	37	17	54
Tak	8	38	46
Suma	45	55	100

Wartości oczekiwane?

Poziom istotności α = 0,001

Przykład (ciśnieniowo-sercowy)

- Wartości oczekiwane:

$$E_{11} = \frac{45 \cdot 54}{100} = 24{,}3 \qquad E_{12} = \frac{55 \cdot 54}{100} = 29{,}7 \qquad E_{21} = \frac{45 \cdot 46}{100} = 20{,}7 \qquad E_{22} = \frac{55 \cdot 46}{100} = 25{,}3$$

Choroba	Ciśnienie podwyższone		Suma
	Nie	Tak	
Nie	**24,3**	**29,7**	54
Tak	**20,7**	**25,3**	46
Suma	45	55	100

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(37-24{,}3)^2}{24{,}3} + \frac{(17-29{,}7)^2}{29{,}7} + \frac{(8-20{,}7)^2}{20{,}7} + \frac{(38-25{,}3)^2}{25{,}3} = 26{,}54$$

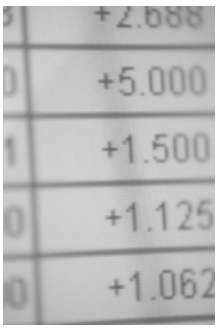
$$\chi^2_{(\alpha,df)} = 10{,}83 \qquad \chi^2_{oblicz} > \chi^2_{kryt}$$

Zadania – trzeba nabyć wprawy!

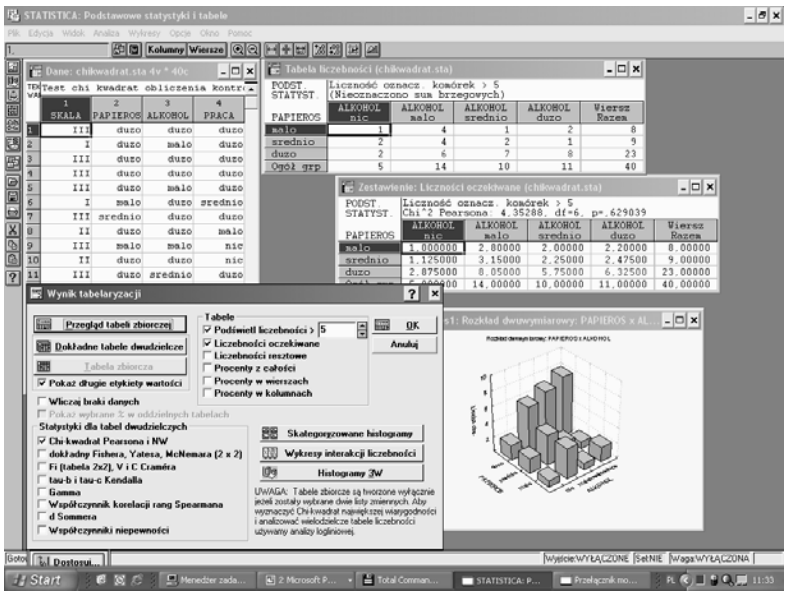
- A teraz trochę popiszemy i policzymy!

Jak obliczać automatycznie test chi – kwadrat?

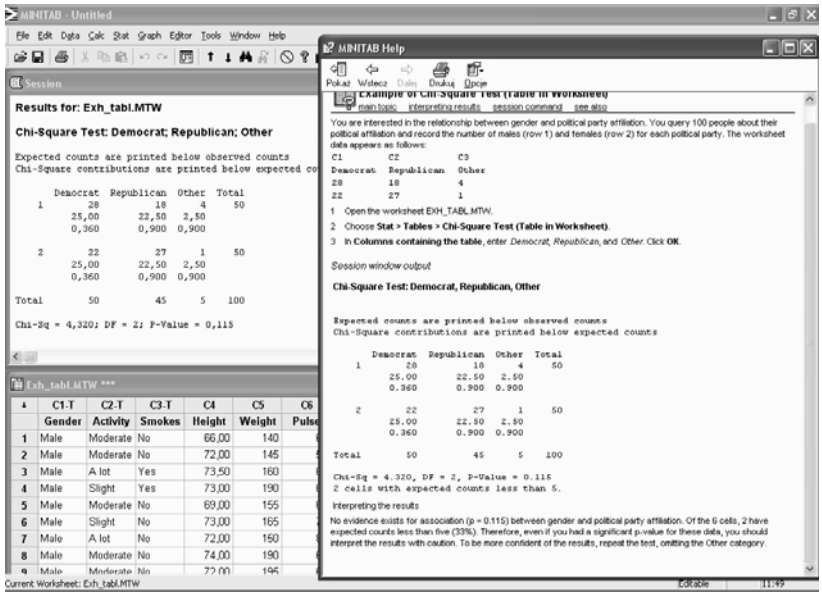
- Czy mamy wsparcie ze strony oprogramowania?



Statsoft : opcja Tablice Wielodzielcze - Chi-kwadrat

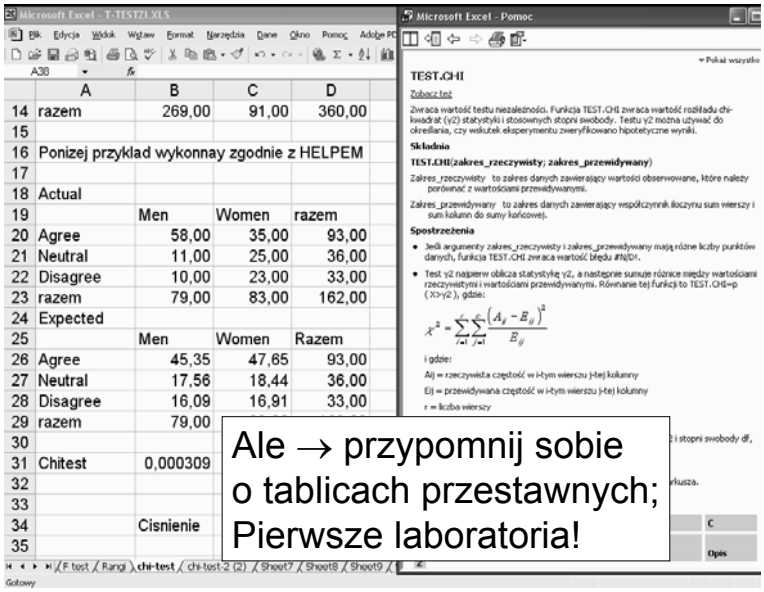


Minitab – opcja Stat:Tables:Chi-squaretest



Excel – co zrobić?

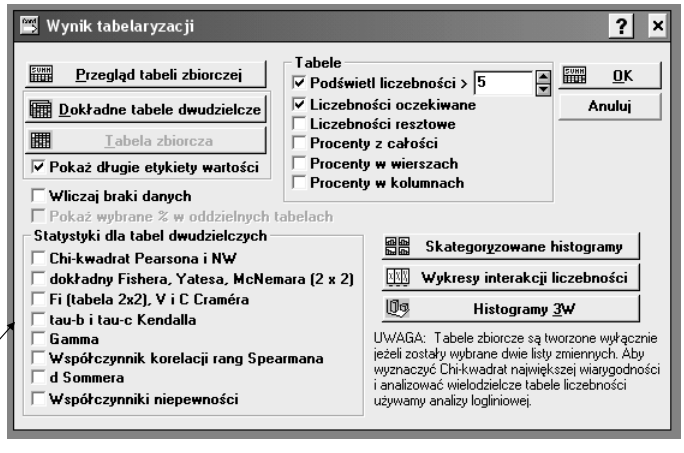
No i nie jest łatwo!



	A	B	C	D
14	razem	269,00	91,00	360,00
15				
16	Ponizej przyklad wykonnay zgodnie z HELPEM			
17				
18	Actual			
19		Men	Women	razem
20	Agree	58,00	35,00	93,00
21	Neutral	11,00	25,00	36,00
22	Disagree	10,00	23,00	33,00
23	razem	79,00	83,00	162,00
24	Expected			
25		Men	Women	Razem
26	Agree	45,35	47,65	93,00
27	Neutral	17,56	18,44	36,00
28	Disagree	16,09	16,91	33,00
29	razem	79,00		
30				
31	Chitest	0,000309		
32				
33				
34		Cisnienie		
35				

Ale → przypomnij sobie o tablicach przestawnych; Pierwsze laboratoria!

Wybór testów niezależności w Statistica



Co to za wersje testów?

StatSoft POLSKA

Poprawki w teście dla tabel 2×2

- Dla tablic typu

a	b
c	d

- Można stosować prostszy wzór

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

- Poprawka Yatesa (gdy n < 40 i którakolwiek z liczności oczekiwanych < 5)

$$\chi^2 = \frac{(|ad-bc|-n/2)^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)}$$

Kilka uwag o innych wersjach testu chi-kwadrat

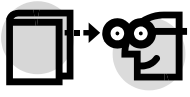
- χ^2 największej wiarygodności (taki sam jak test Pearsona, lecz innych sposób obliczeń oparty na teorii największej wiarygodności (Wyniki obu testów są zbliżone).

Tabele o rozmiarach 2×2

- χ^2 z poprawką Yatesa (liczności oczekiwane małe).
- Dokładny test Fishera – stosowany gdy całkowita liczność obserwacji jest mała lub jeśli małe są liczności oczekiwane).

Powiązanym pojęciem jest test McNemary (gdy liczności reprezentują zmienne zależne).

Więcej w książkach!



Miary siły związku

- Sama wartość statystyki χ^2 pozwala na sprawdzenie tylko czy występuje współzależność! Nie pozwala na pomiar siły tego związku, bo …:)
- Potrzebne są inne miary siły związku! Pożądana normalizacja wartości!

Przykłady miar

- Współczynnik Φ -Yula

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

Miara siły związku w tabeli 2×2; wartość znormalizowana → od 0 (brak związku) do 1 (całkowite powiązanie).

Miary siły związku – cd.

- Współczynnik V – Creamera:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1; l-1)}}$$

gdzie *k* i *l* są wymiarami tablicy wielodzielcznej.

Wartości znormalizowane od 0 (brak związku) do 1.

- Współczynnik kontyngencji Pearsona:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

gdy zmienne niezależne C = 0, lecz max(C) < 1 i zależy od liczby wierszy i kolumn.

Miary siły związku - podsumowanie

Interpretacja wartości współczynników:

- jeśli wartość współczynnika jest równa zero, to cechy *X* i *Y* są niezależne,
- im wartość bliższa jedynki, tym silniejsze jest powiązanie między *X* i *Y*.

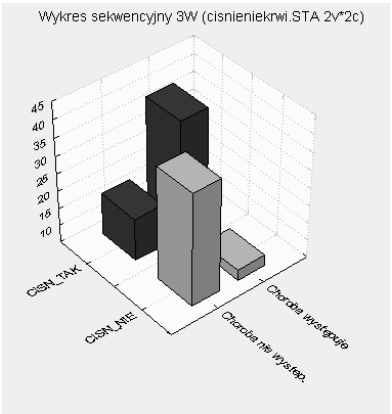
Zastosowania:

- porównywanie między sobą siły zależności różnych par zmiennych,
- Możliwości redukcji liczby atrybutów w tablicach danych.
- Systemy odkrywania wiedzy i eksploracji danych (ang. *Knowledge Discovery in Databases*)
  - 49ner Jan Żytkow

Miary siły związku - Przykład (sercowy):

Badanie czy choroba wieńcowa jest współzależna z podwyższonym ciśnieniem tętniczym → wg. testu TAK!

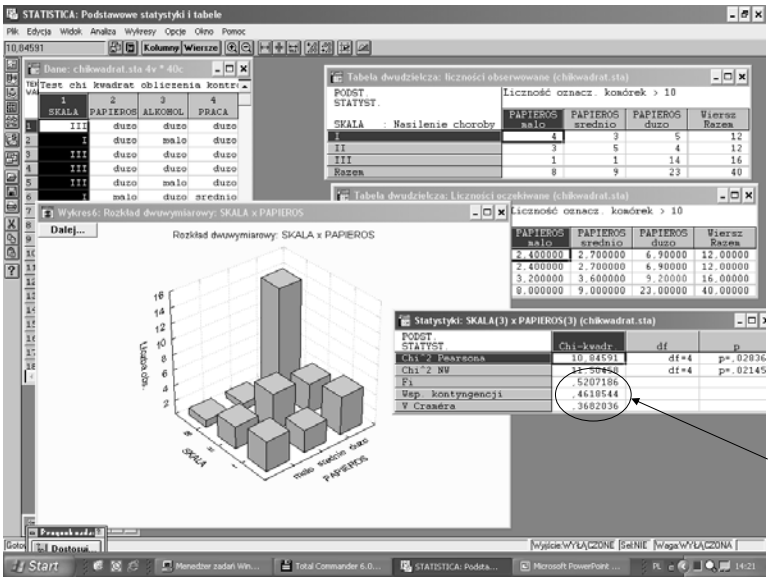
Choroba	Ciśnienie podwyższone		Suma
	Nie	Tak	
Nie	37	17	54
Tak	8	38	46
Suma	45	55	100



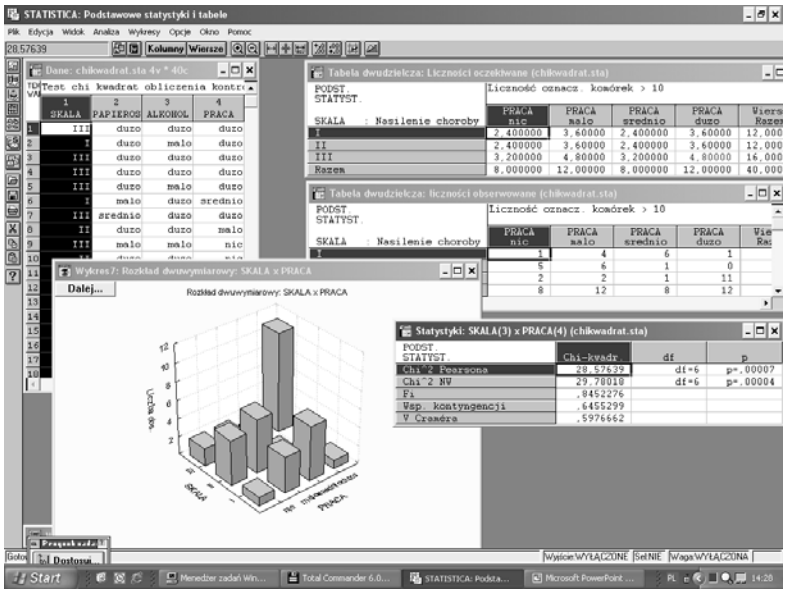
Miary siły związku:

- $\Phi = V = 0{,}51$
- $C = 0{,}46$

Statistica → przykład obliczania miary siły związku para zmiennych *Papierosy* × *Skala choroby*



Statistica → przykład obliczania miary siły związku para zmiennych *Praca* × *Skala choroby*



Dyskusja nt. stosowalności testu chi-kwadrat

- Uwagi o poprawkach testów → minimalne liczności (całkowite i oczekiwane).
- Pamiętaj, że test chi-kwadrat jest skonstruowany dla zmiennych nominalnych!
  - Zmienne porządkowe i liczbowe (dyskretyzowane) → nie wykorzystuje się informacji o porządku wartości;
  - Zalecana ostrożność i wnikliwość analizy (patrz dalej).
- Inne trudności:
  - Dobór właściwego poziomu istotności (zwłaszcza dla danych o dużych rozmiarach); zbyt wysokie wartości mogą prowadzić do identyfikacji losowych efektów, zbyt rygorystyczne prowadzą do pominięcia niektórych regularności zwłaszcza krytyczne w zdaniach eksploracji danych → *data mining*.
  - tzw. paradoks Simpsona / błąd pominięcia ukrytej zmiennej (zaraz będzie przykład)
  - …

Wątpliwości w analizie danych porządkowych

- „Złośliwy” przykład analizy dwóch zmiennych, których wartości są uporządkowane → odpowiednio jako 1, 2, 3, 4 oraz A, B, C, D.

	A	B	C	D
1	9	11	26	14
2	11	14	28	33
3	12	14	33	38
4	3	17	31	39

- Test chi daje *p*-wartość 0,22 , więc nie można odrzucić hipotezy o niezależności.
- Lecz przyjrzyjmy się dokładniej układowi liczności w tabeli, np. (4,A) i (1,D),…, Czy istnieje monotoniczna regularność „małe z małymi, duże z dużymi”?
- Poszukaj innych metod do uwzględnienia porządku.

Uwagi o paradoksie Simpsona

Paradoks Simpsona: pominięcie w analizie zmiennej uwikłanej może zmienić – nawet diametralnie – otrzymane związki między dwiema innymi zmiennymi jakościowymi.

Koronacki, Mielniczuk, Statystyka, str. 392.

Inne sformułowanie:

- „There is a positive dependence between attributes A and B in two complementary sets of data C and $\neg$C, but the dependence becomes negative or vanishes, when we add all data together"

  Jan Żytkow, Automation of discovery in databases: combining AI, statistics and theory of knowledge.

Przykład – paradoks Simpsona

- Pewien Wydział Informatyki i Elektroniki rekrutuje studentów na oba kierunki (przed egzaminem kandydat podaje na jaki kierunek zdaje); Oto wyniki egzaminu.

	Kobiety	Mężczyźni
Osoby odrzucone	102	111
Osoby przyjęte	56	92

- Zauważmy, że procent przyjętych mężczyzn jest wyższy od procentu przyjętych kobiet:
  - przyjęto 100(92/203) = 45% kandydatów mężczyzn
  - oraz tylko 100(56/158) = 35% kandydatek.

Tablica wyników egzaminu – test chi-kwadrat

- Czy można podejrzewać Wydział o dyskryminację kobiet?

Tablica 2 x (chikwadrat.sta)

STATYST. NIEPAR.	Kolumna1	Kolumna2	Wiersz Razem
Liczności, wiersz	102	111	213
Procent całości	28,255%	30,748%	59,003%
Liczności, wiersz	56	92	148
Procent całości	15,512%	25,485%	40,997%
Razem w kol.	158	203	361
Procent całości	43,767%	56,233%	
Chi-kwadrat(df=1)	3,58	p= ,0584	
V-kwadrat(df=1)	3,57	p= ,0587	
Chi-kwadrat skoryg. Yatesa	3,19	p= ,0743	
Fi-kwadrat	,00993		
dokł. p Fishera, jednostr.		p= ,0369	
dwustr.		p= ,0669	
Chi-kwadrat McNemary(A/D	,42	p= ,5182	
Chi-kwadrat(B/C)	17,46	p= ,0000	

Przykład – paradoks Simpsona – więcej szczegółów

- W analizie nie uwzględniliśmy trzeciej zmiennej – kierunku studiów. Jej uwzględnienie → dwie tablice dwudzielcze:

Elektronika		
	Kobiety	Mężczyźni
Osoby odrzucone	11	71
Osoby przyjęte	12	73

Informatyka		
	Kobiety	Mężczyźni
Osoby odrzucone	91	40
Osoby przyjęte	44	19

- Przeanalizujmy procent przyjętych mężczyzn i kobiet:
  - Na elektronikę przyjęto 52% kandydatek oraz 51% kandydatów.
  - Na informatykę przyjęto 32% kandydatek i 32% kandydatów.

Przykład cd. – testy chi-kwadrat.

Elekronika

Tablica 2 x (chikwadrat.sta)

STATYST. NIEPAR.	Kolumna1	Kolumna2	Wiersz Razem
Liczności, wiersz	11	71	82
Procent całości	6,587%	42,515%	49,102%
Liczności, wiersz	12	73	85
Procent całości	7,186%	43,713%	50,898%
Razem w kol.	23	144	167
Procent całości	13,772%	86,228%	
Chi-kwadrat(df=1)	,02	p= ,8951	
V-kwadrat(df=1)	,02	p= ,8955	
Chi-kwadrat skoryg. Yatesa	,01	p= ,9261	
Fi-kwadrat	,00010		
dokł. p Fishera, jednostr.		p= ,5373	
dwustr.		p=1,0000	
Chi-kwadrat McNemary(A/D	44,30	p= ,0000	
Chi-kwadrat(B/C)	40,53	p= ,0000	

Informatyka

(chikwadrat.sta)

Dalej...	Kolumna1	Kolumna2	Wiersz Razem
Liczności, wiersz	91	40	131
Procent całości	46,907%	20,619%	67,526%
Liczności, wiersz	44	19	63
Procent całości	22,680%	9,794%	32,474%
Razem w kol.	135	59	194
Procent całości	69,588%	30,412%	
Chi-kwadrat(df=1)	,00	p= ,9575	
V-kwadrat(df=1)	,00	p= ,9576	
Chi-kwadrat skoryg. Yatesa	,01	p= ,9097	
Fi-kwadrat	,00001		
dokł. p Fishera, jednostr.		p= ,5481	
dwustr.		p=1,0000	
Chi-kwadrat McNemary(A/D	45,83	p= ,0000	
Chi-kwadrat(B/C)	,11	p= ,7434	

Podsumowanie Paradoksu Simpsona

- Związek między dwiema zmiennymi, ujawniany dla każdej kategorii trzeciej zmiennej oddzielnie, może zostać diametralnie zmieniony przez zagregowanie danych, polegające na zsumowaniu wyników dla różnych kategorii trzeciej zmiennej.

Przykład egzaminów

- Fałszywe wrażenie dyskryminacji kobiet, gdyż nie dostrzeżono faktu znacznie trudniejszego wstępu na kierunek informatyka, przy jednoczenie większej popularności tego kierunku wśród kobiet.

Podsumujmy

- Reszta ciekawych dyskusji i przykładów $\rightarrow$ w zalecanych książkach.
- Naa..aaprawde warto zajrzeć!

- Studiowanie to także (i przede wszystkim) samodzielne poszukiwania!

Literatura

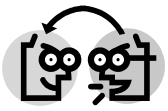
- Statystyka dla studentów kierunków technicznych i przyrodniczych, Koronacki Jacek, Mielniczuk Jan, WNT, 2001.
- Statystyka w zarządzaniu, A.Aczel, PWN, 2000.
- Po prostu statystyka, Frances Clegg, WSiP, 1994.
- Przystępny kurs statystyki, Stanisz A., 1997.
- I wiele innych …

W przykładach wykorzystano także oprogramowanie:

- Statistica 5.0 (© Statsoft Inc.),
- Minitab rel. 14 (© Minitab Inc.),
- Microsoft Excel (© Microsoft).

Czego się dowiedzieliśmy podczas tego wykładu?

1. Analiza danych jakościowych
   - Zmienne jakościowe (nominalne i inne …)
   - Tablice wielodzielcze
2. Analiza tablic wielodzielczych – zależność zmiennych
3. Test zgodności chi - kwadrat
   - Prosta hipoteza o zgodności rozkładów
   - Statystyka chi - kwadrat, …
4. Test niezależności dwóch zmiennych
   - Analiza tablic wielodzielczych z wykorzystaniem testu chi-kwadrat
   - Poprawki w testach
5. Jak to się liczy w różnych programach?
6. Miary siły związku
7. Dyskusja problemów w stosowaniu testów chi-kwadrat



Dziękuję za uwagę

Czytaj także podręczniki
oraz sam eksperymentuj
z danymi!