Zadania ze statystyki cz.5 I rok socjologii – miary związków między zmiennymi jakościowymi

Zadanie 1
Zdaniem wielu komentatorów, kobiety częściej niż mężczyźni głosują na partię rządzącą. Wyniki badań przedstawia tabela poniżej. Należy sprawdzić czy istnieje związek między płcią a głosowaniem w ostatnich wyborach na PO lub PIS.

	kobieta	mężczyzna
PO	350	160
PIS	280	150

Zadanie 2
Czy kolor oczu ma znaczenie na wybór partnera? Ten problem został zbadany a wyniki rozkładu cech (kolor oczu badanego i kolor oczu partnera) prezentuje poniższa tabela. Należy ustalić, czy istnieje związek między kolorem oczu osób w związkach oraz siłę tego związku, wykorzystując współczynniki oparte na chi-kwadrat.

		badany		
		zielone	niebieskie	brązowe
partner	zielone	15	12	22
	niebieskie	12	28	10
	brązowe	29	14	11

Zadanie 3
Dla danych z tabeli z zadania 2 należy ustalić jaka będzie wielkość redukcji błędu przewidywania, jeśli będziemy chcieli przewidzieć wybór partnera ze względu na kolor oczu, wykorzystując współczynnik Lambda.

Zadanie 4
W pewnym badaniu podjęto się analizy problemu relacji między krajem zamieszkania emigrantów z Polski a sektorem zatrudnienia. Należy ustalić, w oparciu o współczynnik Tau Goodmana-Kruskala, czy silnej sektor zatrudnienia determinuje wybór kraju emigracji, czy też kraj emigracji silniej wpływa na wybór sektora pracy.

	Niemcy	Wielka Bryt.	Holandia
budownictwo	13	14	39
usługi	17	48	11
nauka i eduk	47	13	8
opieka zdrow.	7	16	13

Zadanie 5
Czy poziom zadowolenia z życia ma znaczenie dla uczestniczenia w różnego rodzaju aktywności obywatelskiej? Proszę zbadać, wykorzystując współczynniki Gamma i d Somersa, czy istnieje związek między poziomem zadowolenia a poziomem aktywności.

		poziom zadowolenia z życia		
		wysoki	średni	niski
poziom aktywności obywatelskiej	wysoki	20	30	10
	średni	30	25	15
	niski	20	30	40

Zadanie 6
Czy istnieje związek między miejscem zamieszkania a wysokością dochodów? W tabeli poniżej przedstawiono procentowy udział mieszkańców w rozkładzie dochodów. Czy miejsce zamieszania wpływa na dochody? Jak silna jest to zależność?

	wieś	miasto
0-2000	45	35
2001 - 4000	25	45
4001 - 6000	20	10
powyżej 6000	10	10

PODSUMOWANIE WYKŁADU

Test χ^2 chi-kwadrat i mierniki oparte na χ^2 chi-kwadrat

- Test χ^2 (chi – kwadrat), inaczej nazywany testem niezależności, bada odstępstwa rozkładu empirycznego zmiennych z rozkładem teoretycznym (oczekiwanym), zakładającym pełną niezależność zmiennych.
- Idea χ^2 opiera się na poszukiwaniu takiego rozkładu liczebności , który wskazywałby na niezależność rozkładów dwóch zmiennych względem siebie (rozkład oczekiwany).

$$\chi^2 = \sum \frac{(n_e - n_t)^2}{n_t}$$

n_e – *liczebności empiryczne rozkładów warunkowych*
n_t – *liczebności teoretyczne rozkładów warunkowych*

- Rozkład empiryczny dwóch zmiennych to łączny rozkład współzmienności dwóch zmiennych, pochodzący z pomiarów
- Rozkład teoretyczny (oczekiwany) to rozkład warunkowy pochodzący z analizy rozkładów brzegowych zmiennych.

$$n_{tij} = \frac{n_{i.} * n_{.j}}{n}$$

$n_{i.} - \ liczebność\ brzegowa\ i - tej\ kategorii\ pierwszej\ zmiennej\ (np. w\ kolumnie)$
$n_{.j} - liczebność\ brzegowa\ j - tej\ kategorii\ drugiej\ zmiennej\ (np. w\ wierszu)$

Przykład:

Oto mamy rozkład empiryczny, z rozkładem warunkowym a, b, c, i d:

	i1	i2	
j1	a	b	a+b
j2	c	d	c+d
	a+c	b+d	n

Tworzymy dla tego rozkładu rozkład teoretyczny (oczekiwany)

$$n_{ta} = \frac{(a+b)*(a+c)}{n} \qquad n_{tb} = \frac{(b+a)*(b+d)}{n}$$
$$n_{tc} = \frac{(c+a)*(c+d)}{n} \qquad n_{td} = \frac{(d+c)*(b+d)}{n}$$

$$n = a + b + c + d$$

Następnie porównujemy oba rozkłady, sumujemy ostatnią kolumnę i otrzymujemy χ^2 chi-kwadrat

				$\frac{(n_e - n_t)^2}{n_t}$
a	ta	a – ta	(a – ta)²	$\frac{(a-ta)^2}{ta}$
b	tb	b – tb	(b – tb)²	$\frac{(b-tb)^2}{tb}$
c	tc	c – tc	(c – tc)²	$\frac{(c-tc)^2}{tc}$
d	td	d - td	(d – td)²	$\frac{(d-td)^2}{td}$

				$\chi^2 = \sum \frac{(n_e - n_t)^2}{n_t}$

- Związek między zmiennymi występuje wtedy, gdy wartość testu χ^2 (chi – kwadrat) jest większa od zera.
- Brak związku (pełna niezależność dwóch zmiennych) między zmiennymi wyznaczona jest wartością zero testu χ^2 (chi – kwadrat).
- Wadą testu χ^2 jest uzależnienie wysokości wyniku od wielkości badanej grupy (od liczebności rozkładów brzegowych i warunkowych). Im większa grupa, a tym samym większa liczebność rozkładów brzegowych i warunkowych, tym większa wartość testu χ^2 dla tego samego związku między zmiennymi.

Mierniki oparta na chi – kwadrat

współczynnik $\boldsymbol{\Phi_Y}$ *Yula*

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

współczynnik $\mathbf{V_C}$ *Cramera*

$$V_C = \sqrt{\frac{\chi^2}{(m-1)n}}$$

współczynnik $\mathbf{C_P}$ *kontyngencji (Pearsona)*

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

Współczynniki oparte na koncepcji PRE (proporcjonalnej redukcji błędu przewidywania

Współczynnik λ Guttmana (Lambda)

- Jeżeli będziemy chcieli przewidzieć, jak zmienna niezależna X wpływa na zmienną Y (będziemy chcieli określić wielkość błędu przewidywania wartości zmiennej Y na podstawie wartości zmiennej X), to współczynnik obliczymy:

$$\lambda_{(Y/X)} = \frac{\sum max_x n_{ij} - max_x n_{.j}}{n - max_x n_{.j}}$$

$max_x n_{ij}$ – liczebność dominującego wariantu cechy X w i– tej kategorii cechy Y

$max_x n_{.j}$ – liczebność dominującego wariantu brzegowego cechy X

n_{ij} – warunkowa czestość i– teao wariantu zmiennej zależnej Y w j –tej kategorii zmiennej niezależnej X

$n_{.j}$ – częstość brzegowa zmiennej niezależnej X

- Miernik λ Guttmana pokazuje wielkość redukowanego błędu popełnianego przy przewidywaniu jednej zmiennej (Y) na podstawie drugiej zmiennej (X).
- Wartość miernika λ Guttmana zawiera się w przedziale od 0 do 1

Współczynnik **τ Goodmana-Kruskala** (Tau)

- Współczynnik **τ Goodmana-Kruskala** (Tau) mówi w jakim stopniu znajomość rozkładu cechy X pozwala zredukować oczekiwaną liczbę błędów przy przewidywaniu rozkładu cechy Y.

$$\tau_{(Y/X)} = \frac{n\sum\sum\frac{n_{ij}^{\ 2}}{n_{i.}} - \sum n_{.j}^{\ 2}}{n^2 - \sum n_{.j}^{\ 2}}$$

n_{ij} – warunkowa czestość i– teao wariantu zmiennej zależnej Y w j –tej kategorii zmiennej niezależnej X

$n_{.j}$ – częstość brzegowa zmiennej niezależnej X

$ni.$ –częstość brzegowa zmiennej zależnej Y

Współczynniki dla zmiennych mierzonych na skali porządkowej.

- Współczynnik **γ** (Gamma) Goodmana-Kruskala

$$\gamma = \frac{P - Q}{P + Q}$$

- Miara **d** Somersa

$$d = \frac{2P - 2Q}{B_{(W)}}$$

- Współczynnik **τb** (Tau-be) Kendalla

$$\tau_b = \frac{2P - 2Q}{\sqrt{B_K * B_W}}$$

Gdzie dla każdego z tych mierników:

$$P = \sum_{i=1}^{w}\sum_{j=1}^{k} n_{ij}c_{ij} \qquad Q = \sum_{i=1}^{w}\sum_{j=1}^{k} n_{ij}d_{ij}$$

$$c_{ij} = \sum_{h<i}^{w}\sum_{l<j}^{k} n_{hl} + \sum_{h>i}^{w}\sum_{l>j}^{k} n_{hl} \qquad d_{ij} = \sum_{h<i}^{w}\sum_{l>j}^{k} n_{hl} + \sum_{h>i}^{w}\sum_{l<j}^{k} n_{hl}$$

$$B_{(w)} = n^2 - \sum n_{i.}^2 \qquad B_{(k)} = n^2 - \sum n_{.j}^2$$

$B_{(w/k)}$ – *brak uporządkowania ogólnego w wierszach* (B_w) *lub w kolumnach* (B_k)

$n_{.j}$ – *częstość brzegowa zmiennej niezależnej X*

$ni.$ – *częstość brzegowa zmiennej zależnej Y*

- Mierniki te opierają się na sposobie uporządkowania danych w tabeli kontyngencji:
  - P – uporządkowanych zgodnie z kierunkiem kategorii;
  - Q – uporządkowanych przeciwnie do kierunku uporządkowania kategorii
  - B – braku uporządkowania rozkładów brzegowych (kolumnowych lub wierszowych).
- Informują o stopniu uporządkowania danych w tabeli kontyngencji w kierunku zgodnym z rozkładem kategorii zmiennych (+) lub przeciwnym (-).

Przykład:

Pary uporządkowane zgodnie z kierunkiem uporządkowania kategorii cechy to a i d, pary uporządkowane w kierunku przeciwnym do kierunku uporządkowania cechy to pary b i c.

	wysoki	niski
wysoki	**a**	**b**
niski	**c**	**d**

W takim razie:

$$P = a \cdot d$$

$$Q = b \cdot c$$

Inny przykład. Rozważając wszystkie pary iloczynów o zgodnym i przeciwnym kierunku uporządkowania, tabelę większą niż 2x2, będziemy traktować jak serię takich tabel (2x2)

	wysoki	średni	niski
wysoki	a	b	c
średni	d	e	f
niski	g	h	i

Tym samym:

$$P = a(e + f + h + i) + b(f + i) + d\,(h + i) + ei$$

$$Q = c(d + e + g + h) + b(d + g) + f(g + h) + eg$$