Testy χ^2

Statystyka i analiza danych 2019

Jurek Błaszczyński,
na podstawie slajdów Wojtka Kotłowskiego

25 maja 2019

Testy χ^2

- Dotyczą zmiennej/zmiennych **dyskretnych**, ze skończoną liczbą możliwych wartości:
  - uporządkowane kategorie, płeć, kolor, narodowość, wynik rzutu kostką, itp.
- Nie dotyczą jednego parametru rozkładu, ale **całego rozkładu prawdopodobieństwa**.
- Będziemy zajmować się dwoma rodzajami (wersjami) testów: test rozkładu **jednej** zmiennej oraz test rozkładu **dwóch zmiennych**.

Test dla jednej zmiennej

Dyskretna zmienna X przyjmująca jedną z wartości $\{x_1, \ldots, x_k\}$.

- **Układ hipotez**:

H_0 :	Zmienna X ma rozkład P
H_1 :	Zmienna X ma rozkład różny od P

- **Tabela wartości obserwowanych (*observed*)**:

x_1	x_2	x_3	$\ldots$	x_k	Σ
O_1	O_2	O_3	$\ldots$	O_k	n

- **Tabela wartości oczekiwanych (*expected*) z H_0**:

x_1	x_2	x_3	$\ldots$	x_k	Σ
E_1	E_2	E_3	$\ldots$	E_k	n

$$E_i = P(X = x_i) \cdot n$$

- **Statystyka testowa**:

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

Jeśli $\chi^2 > \chi^2_{\mathrm{kr}}$, odrzucamy H_0.

Test dla jednej zmiennej – przykład

X – wynik rzutu kostką. Testujemy, czy kostka jest uczciwa

Test dla jednej zmiennej – przykład

X – wynik rzutu kostką. Testujemy, czy kostka jest uczciwa

- **Układ hipotez**:

H_0 :	Zmienna X ma rozkład jednostajny na $\{1, 2, 3, 4, 5, 6\}$
H_1 :	Zmienna X nie ma rozkładu jednostajnego

Test dla jednej zmiennej – przykład

X – wynik rzutu kostką. Testujemy, czy kostka jest uczciwa

- **Układ hipotez**:

H_0 :	Zmienna X ma rozkład jednostajny na $\{1, 2, 3, 4, 5, 6\}$
H_1 :	Zmienna X nie ma rozkładu jednostajnego

- **Tabela wartości obserwowanych** przy $n = 30$ rzutach:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
4	6	3	6	8	3	30

Test dla jednej zmiennej – przykład

X – wynik rzutu kostką. Testujemy, czy kostka jest uczciwa

- **Układ hipotez**:

H_0 :	Zmienna X ma rozkład jednostajny na $\{1, 2, 3, 4, 5, 6\}$
H_1 :	Zmienna X nie ma rozkładu jednostajnego

- **Tabela wartości obserwowanych** przy $n = 30$ rzutach:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
4	6	3	6	8	3	30

- **Tabela wartości oczekiwanych z** H_0:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
						30

Test dla jednej zmiennej – przykład

X – wynik rzutu kostką. Testujemy, czy kostka jest uczciwa

- **Układ hipotez**:

H_0 :	Zmienna X ma rozkład jednostajny na $\{1, 2, 3, 4, 5, 6\}$
H_1 :	Zmienna X nie ma rozkładu jednostajnego

- **Tabela wartości obserwowanych** przy $n = 30$ rzutach:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
4	6	3	6	8	3	30

- **Tabela wartości oczekiwanych z** H_0:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
5	5	5	5	5	5	30

Test dla jednej zmiennej – przykład

X – wynik rzutu kostką. Testujemy, czy kostka jest uczciwa

- **Układ hipotez**:

H_0 : Zmienna X ma rozkład jednostajny na $\{1, 2, 3, 4, 5, 6\}$

H_1 : Zmienna X nie ma rozkładu jednostajnego

- **Tabela wartości obserwowanych** przy $n = 30$ rzutach:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
4	6	3	6	8	3	30

- **Tabela wartości oczekiwanych z** H_0:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
5	5	5	5	5	5	30

- **Statystyka testowa**:

Test dla jednej zmiennej – przykład

X – wynik rzutu kostką. Testujemy, czy kostka jest uczciwa

- **Układ hipotez**:

H_0 : Zmienna X ma rozkład jednostajny na $\{1, 2, 3, 4, 5, 6\}$

H_1 : Zmienna X nie ma rozkładu jednostajnego

- **Tabela wartości obserwowanych** przy $n = 30$ rzutach:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
4	6	3	6	8	3	30

- **Tabela wartości oczekiwanych z** H_0:

$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
5	5	5	5	5	5	30

- **Statystyka testowa**:

$$\chi^2 = \frac{(4-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(3-5)^2}{5}$$
$$= \frac{1}{5} + \frac{1}{5} + \frac{4}{5} + \frac{1}{5} + \frac{9}{5} + \frac{4}{5} = \frac{20}{5} = 4$$

Test dla jednej zmiennej – przykład

X – wynik rzutu kostką. Testujemy, czy kostka jest uczciwa

- **Układ hipotez**:

H_0 :	Zmienna X ma rozkład jednostajny na $\{1, 2, 3, 4, 5, 6\}$
H_1 :	Zmienna X nie ma rozkładu jednostajnego

- **Tabela wartości obserwowanych** przy $n = 30$ rzutach:

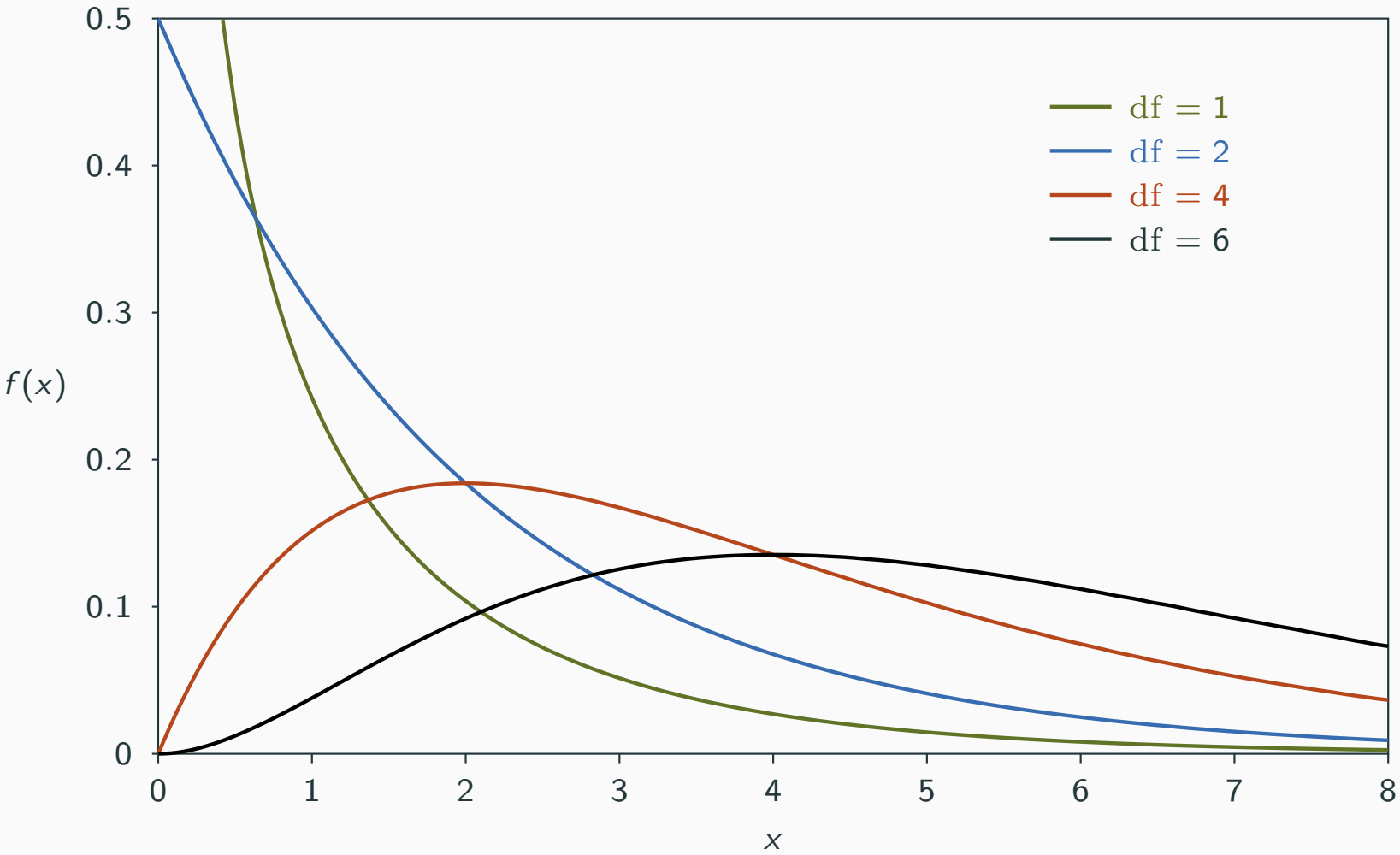
$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
4	6	3	6	8	3	30

- **Tabela wartości oczekiwanych z** H_0:

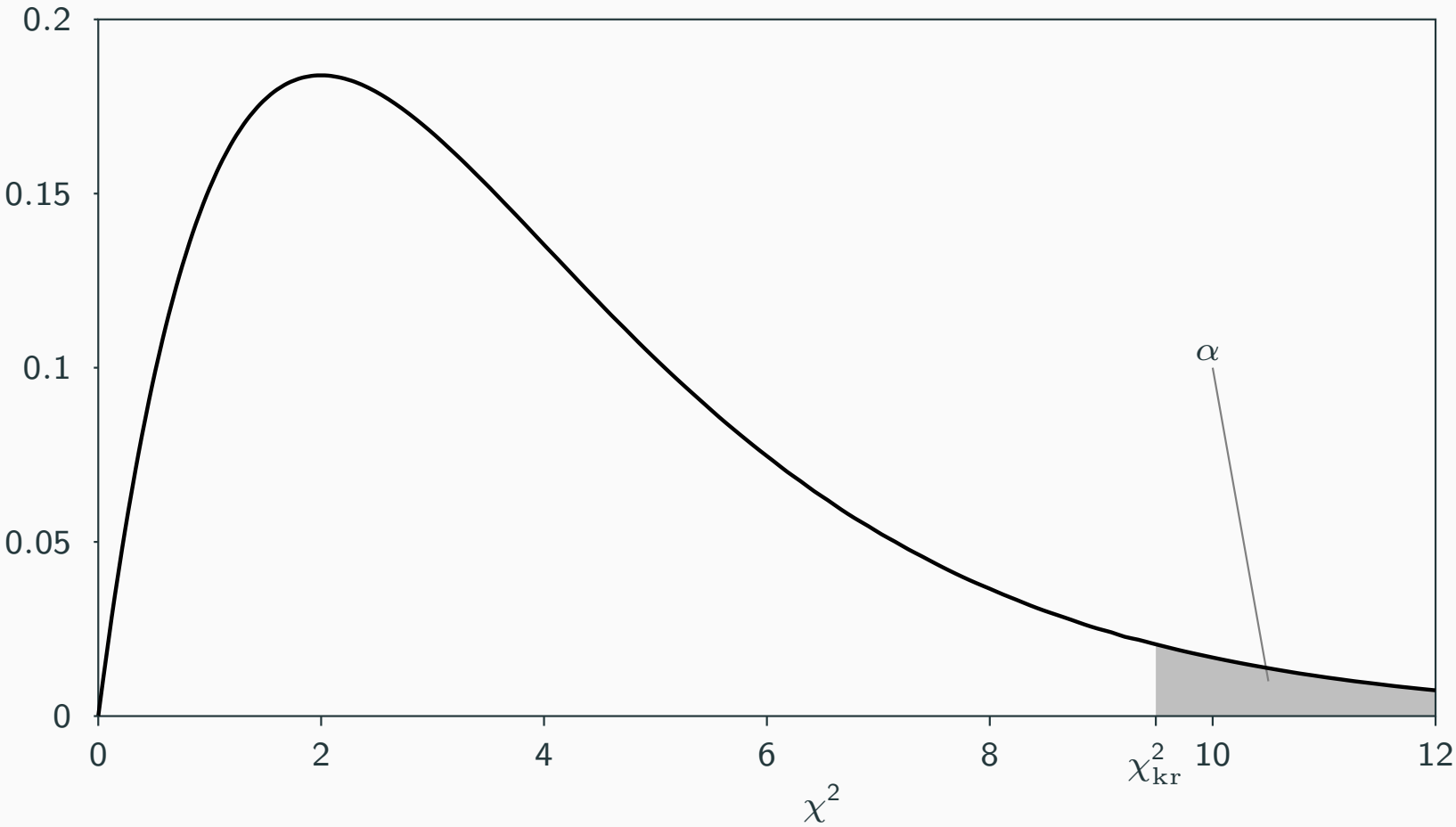
$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	Σ
5	5	5	5	5	5	30

- **Statystyka testowa**:
  - Wartość statystyki: $\chi^2 = 4$
  - Stopnie swobody: $k - 1 = 5$
  - Dla $\alpha = 0.01$, $\chi^2_{\text{kr}} = 15.08$ (z tablic)
  - **Wniosek**: $\chi^2 < \chi^2_{\text{kr}}$, więc brak podstaw do odrzucenia H_0.

Rozkład $\chi^2(k)$



Rozkład $\chi^2(4)$



Obszar krytyczny zawsze z prawej strony: $C_{\mathrm{kr}} = (\chi^2_{\mathrm{kr}}, \infty)$.

Test dla dwóch zmiennych $X \in \{x_1, \ldots, x_w\}$ i $Y \in \{y_1, \ldots, y_k\}$

Układ hipotez:

H_0 :	Zmienne X i Y są **niezależne**
H_1 :	Zmienne X i Y są **zależne**

Tabela w. obserwowanych

	y_1	y_2	$\ldots$	y_k	Σ
x_1	$O_{1,1}$	$O_{1,2}$	$\ldots$	$O_{1,k}$	R_1
x_2	$O_{2,1}$	$O_{2,2}$	$\ldots$	$O_{2,k}$	R_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_w	$O_{w,1}$	$O_{w,2}$	$\ldots$	$O_{w,k}$	R_w
Σ	C_1	C_2	$\ldots$	C_k	n

Tabela w. oczekiwanych

	y_1	y_2	$\ldots$	y_k	Σ
x_1	$E_{1,1}$	$E_{1,2}$	$\ldots$	$E_{1,k}$	R_1
x_2	$E_{2,1}$	$E_{2,2}$	$\ldots$	$E_{2,k}$	R_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_w	$E_{w,1}$	$E_{w,2}$	$\ldots$	$E_{w,k}$	R_w
Σ	C_1	C_2	$\ldots$	C_k	n

Wartości oczekiwane: $E_{ij} = \frac{R_i \, C_j}{n}$ $\left(\frac{\text{suma wiersza} \times \text{suma kolumny}}{\text{podsumowanie tabeli}}\right)$

Statystyka testowa:

$$\chi^2 = \sum_{i=1}^{w} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((w-1)\cdot(k-1))$$

Wartości oczekiwane

Skąd wzór $E_{ij} = \frac{R_i \; C_j}{n}$?

Wartości oczekiwane

Skąd wzór $E_{ij} = \frac{R_i \; C_j}{n}$?

Spodziewamy się wystąpienia:

$$n \cdot P(X = x_i, Y = y_j)$$

obserwacji dla których $X = x_i$ i $Y = y_j$.

Wartości oczekiwane

Skąd wzór $E_{ij} = \frac{R_i \; C_j}{n}$?

Spodziewamy się wystąpienia:

$$n \cdot P(X = x_i, Y = y_j)$$

obserwacji dla których $X = x_i$ i $Y = y_j$.

Przy założeniu H_0 zmienne są **niezależne**, a więc:

$$\begin{aligned} P(X = x_i, Y = y_j) &= P(X = x_i) \cdot P(Y = y_j) \\ &= \frac{R_i}{n} \cdot \frac{C_j}{n} \end{aligned}$$

Pomnożenie przez n to właśnie ten wzór.

Test dla dwóch zmiennych – przykład

W USA przeprowadzono sondaż opinii na 1000 losowo wybranych osób. Sprawdź, czy istnieje zależność między płcią odpytanych osób a ich preferencjami politycznymi.

	republican	democrat	independent	Σ
male	200	150	50	400
female	250	300	50	600
Σ	450	450	100	1000

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

H_0 :	Brak zależności między płcią a pref. wyborczymi
H_1 :	Istnieje zależność

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

H_0 :	Brak zależności między płcią a pref. wyborczymi
H_1 :	Istnieje zależność

Tabela w. obserwowanych

	rep	dem	ind	Σ
M	200	150	50	400
F	250	300	50	600
Σ	450	450	100	1000

Tabela w. oczekiwanych

	rep	dem	ind	Σ
M				400
F				600
Σ	450	450	100	1000

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

H_0 :	Brak zależności między płcią a pref. wyborczymi
H_1 :	Istnieje zależność

Tabela w. obserwowanych

	rep	dem	ind	Σ
M	200	150	50	400
F	250	300	50	600
Σ	450	450	100	1000

Tabela w. oczekiwanych

	rep	dem	ind	Σ
M	180			400
F				600
Σ	450	450	100	1000

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

H_0 :	Brak zależności między płcią a pref. wyborczymi
H_1 :	Istnieje zależność

Tabela w. obserwowanych

	rep	dem	ind	Σ
M	200	150	50	400
F	250	300	50	600
Σ	450	450	100	1000

Tabela w. oczekiwanych

	rep	dem	ind	Σ
M	180	180		400
F				600
Σ	450	450	100	1000

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

H_0 :	Brak zależności między płcią a pref. wyborczymi
H_1 :	Istnieje zależność

Tabela w. obserwowanych

	rep	dem	ind	Σ
M	200	150	50	400
F	250	300	50	600
Σ	450	450	100	1000

Tabela w. oczekiwanych

	rep	dem	ind	Σ
M	180	180	40	400
F				600
Σ	450	450	100	1000

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

H_0 :	Brak zależności między płcią a pref. wyborczymi
H_1 :	Istnieje zależność

Tabela w. obserwowanych

	rep	dem	ind	Σ
M	200	150	50	400
F	250	300	50	600
Σ	450	450	100	1000

Tabela w. oczekiwanych

	rep	dem	ind	Σ
M	180	180	40	400
F	270	270	60	600
Σ	450	450	100	1000

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

H_0 :	Brak zależności między płcią a pref. wyborczymi
H_1 :	Istnieje zależność

Tabela w. obserwowanych

	rep	dem	ind	Σ
M	200	150	50	400
F	250	300	50	600
Σ	450	450	100	1000

Tabela w. oczekiwanych

	rep	dem	ind	Σ
M	180	180	40	400
F	270	270	60	600
Σ	450	450	100	1000

Statystyka testowa:

$$\chi^2 = \frac{(200-180)^2}{180} + \frac{(150-180)^2}{180} + \frac{(50-40)^2}{40} + \frac{(250-270)^2}{270} + \frac{(300-270)^2}{270} + \frac{(50-60)^2}{60} = \frac{20}{9} + 5 + \frac{5}{2} + \frac{40}{27} + \frac{10}{3} + \frac{5}{3} = 16.2$$

Test dla dwóch zmiennych – przykład

X – płeć, Y – preferencje wyborcze.

Układ hipotez:

H_0 :	Brak zależności między płcią a pref. wyborczymi
H_1 :	Istnieje zależność

Tabela w. obserwowanych

	rep	dem	ind	Σ
M	200	150	50	400
F	250	300	50	600
Σ	450	450	100	1000

Tabela w. oczekiwanych

	rep	dem	ind	Σ
M	180	180	40	400
F	270	270	60	600
Σ	450	450	100	1000

Statystyka testowa:

- Wartość statystyki: $\chi^2 = 16.2$
- Stopnie swobody: $(w-1)(k-1) = 1 \cdot 2 = 2$
- Dla $\alpha = 0.01$, $\chi^2_{\text{kr}} = 9.21$ (z tablic)
- **Wniosek**: $\chi^2 > \chi^2_{\text{kr}}$, więc odrzucamy H_0.