Testy χ^2

Na podstawie materiałów W. Kotłowskiego[1]

[1] Poprawne treści są autorstwa WK, niepoprawne – mojego (SW)

Testy χ^2

- Dotyczą zmiennej/zmiennych **dyskretnych**, ze skończoną liczbą możliwych wartości:
  - Płeć, kolor, uporządkowane kategorie, narodowość, wynik rzutu kostką
  - Także zdykretyzowane wartości zmiennych ciągłych (ostatni przykład)
- Nie testują jednego parametru rozkładu, ale **cały rozkład prawdopodobieństwa**.
- Tutaj poznamy dwie wersje: test rozkładu **jednej** zmiennej oraz test rozkładu **dwóch zmiennych**.

Test dla jednej zmiennej

Dyskretna zmienna X przyjmująca jedną z wartości $\{x_1, \ldots, x_k\}$, $P(X = x_i) = p_i$.

Układ hipotez:

H_0 :	Zmienna X ma rozkład P $(\forall_{i=1\ldots k}\ p_i = p_{i,0})$
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

$$e_i = p_{i,0} \cdot n$$

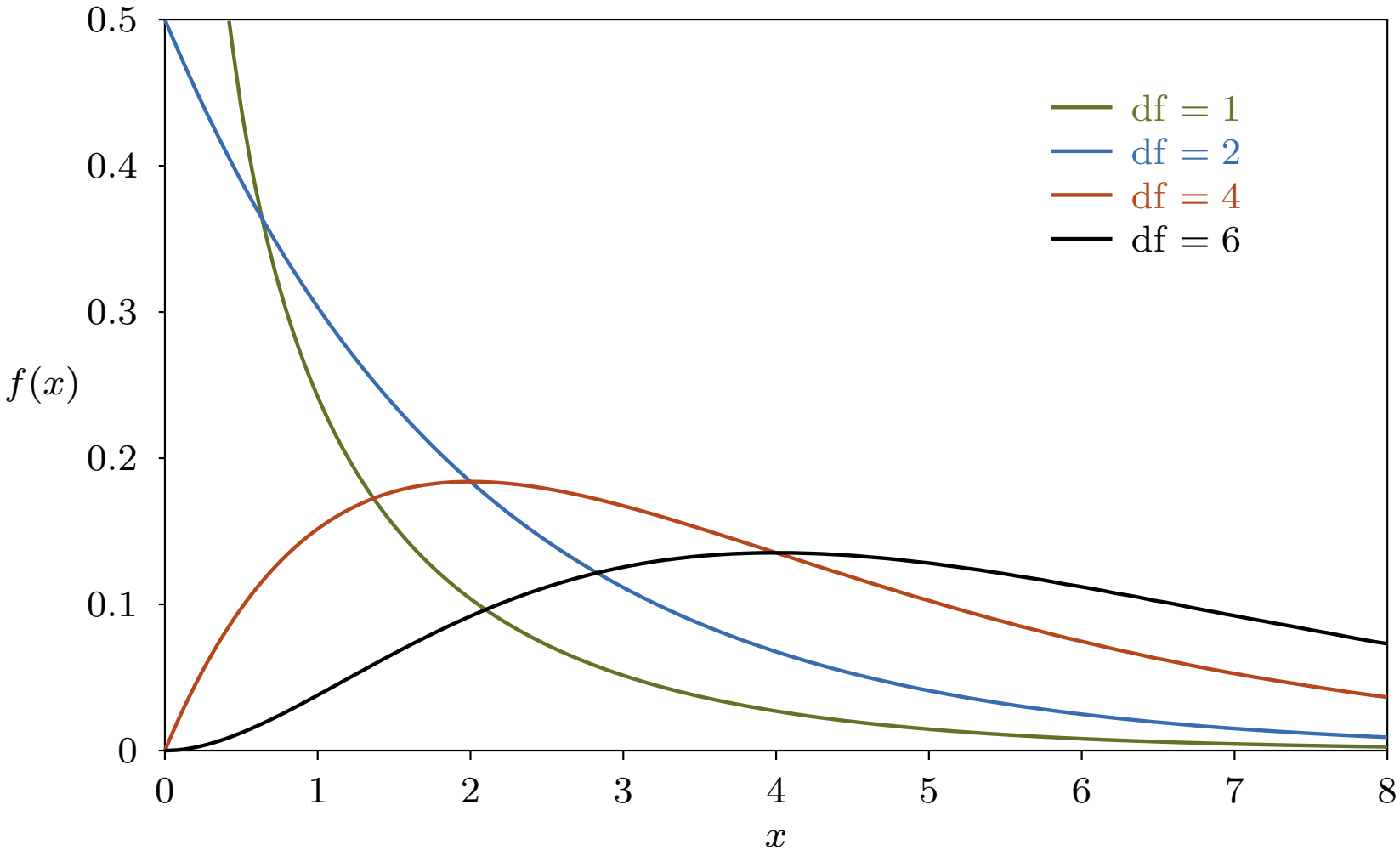
Test jednej zmiennej

Statystyka testowa:

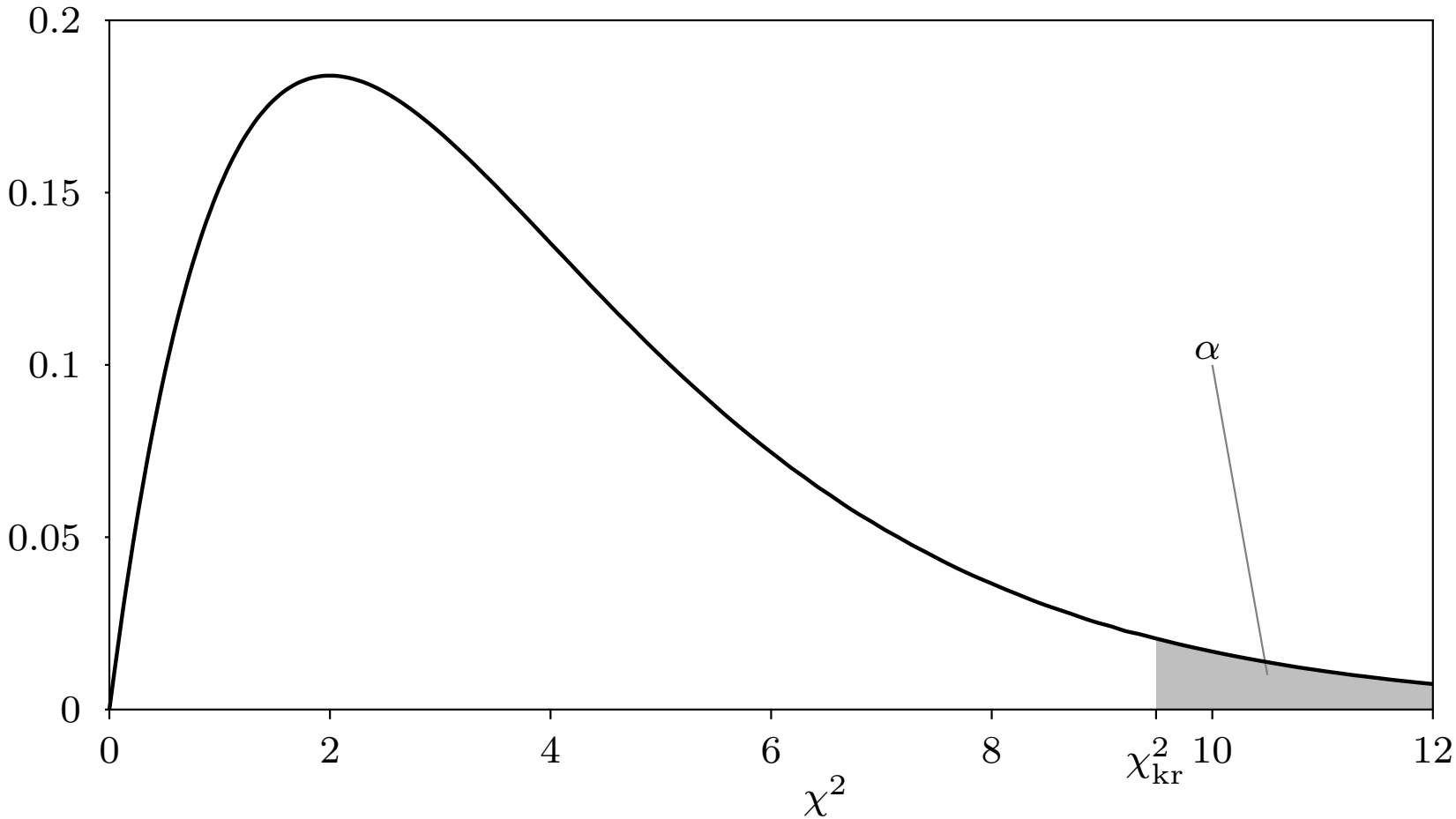
$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(k-1)$$

Jeśli $\chi^2 > \chi^2_{\text{kr}}$, odrzucamy H_0 („**test prawostronny**").

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	$\cdots$	y_k	Σ
x_1	$o_{1,1}$	$o_{1,2}$	$\cdots$	$o_{1,k}$	W_1
x_2	$o_{2,1}$	$o_{2,2}$	$\cdots$	$o_{2,k}$	W_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_w	$o_{w,1}$	$o_{w,2}$	$\cdots$	$o_{w,k}$	W_w
Σ	K_1	K_2	$\cdots$	K_k	n

Tabela w. oczekiwanych

	y_1	y_2	$\cdots$	y_k	Σ
x_1	$e_{1,1}$	$e_{1,2}$	$\cdots$	$e_{1,k}$	W_1
x_2	$e_{2,1}$	$e_{2,2}$	$\cdots$	$e_{2,k}$	W_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_w	$e_{w,1}$	$e_{w,2}$	$\cdots$	$e_{w,k}$	W_w
Σ	K_1	K_2	$\cdots$	K_k	n

Wartości oczekiwane: $e_{ij} = \frac{W_i \, K_j}{n}$ $\left(\frac{\text{suma wiersza} \times \text{suma kolumny}}{\text{podsumowanie tabeli}}\right)$

Test dla dwóch zmiennych

Statystyka testowa:

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

Jeśli $\chi^2 > \chi^2_{\text{kr}}$, odrzucamy H_0 (takie samo postępowanie, jak dla jednej zmiennej – różnica w wyznaczaniu liczby stopni swobody).

Wartości oczekiwane

Skąd wzór $e_{ij} = \frac{W_i \, K_j}{n}$?

Wartości oczekiwane

Skąd wzór $e_{ij} = \frac{W_i \, K_j}{n}$?

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 $e_{ij} = \frac{W_i \, K_j}{n}$?

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:

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

Pomnożenie przez n to właśnie ten wzór.

Miary siły związku

- Statystyka χ^2 nie pozwala na pomiar siły związku, a jedynie na stwierdzenie jego obecności
- Konieczne są **znormalizowane** miary siły związku – współczynniki kontyngencji
  - współczynnik Φ Yule'a: $\Phi = \sqrt{\frac{\chi^2}{n}}$ (dla macierzy 2×2)
  - współczynnik V Crammer'a: $V = \sqrt{\frac{\chi^2}{n \cdot min(w-1, k-1)}}$
  - współczynnik C Pearson'a: $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$