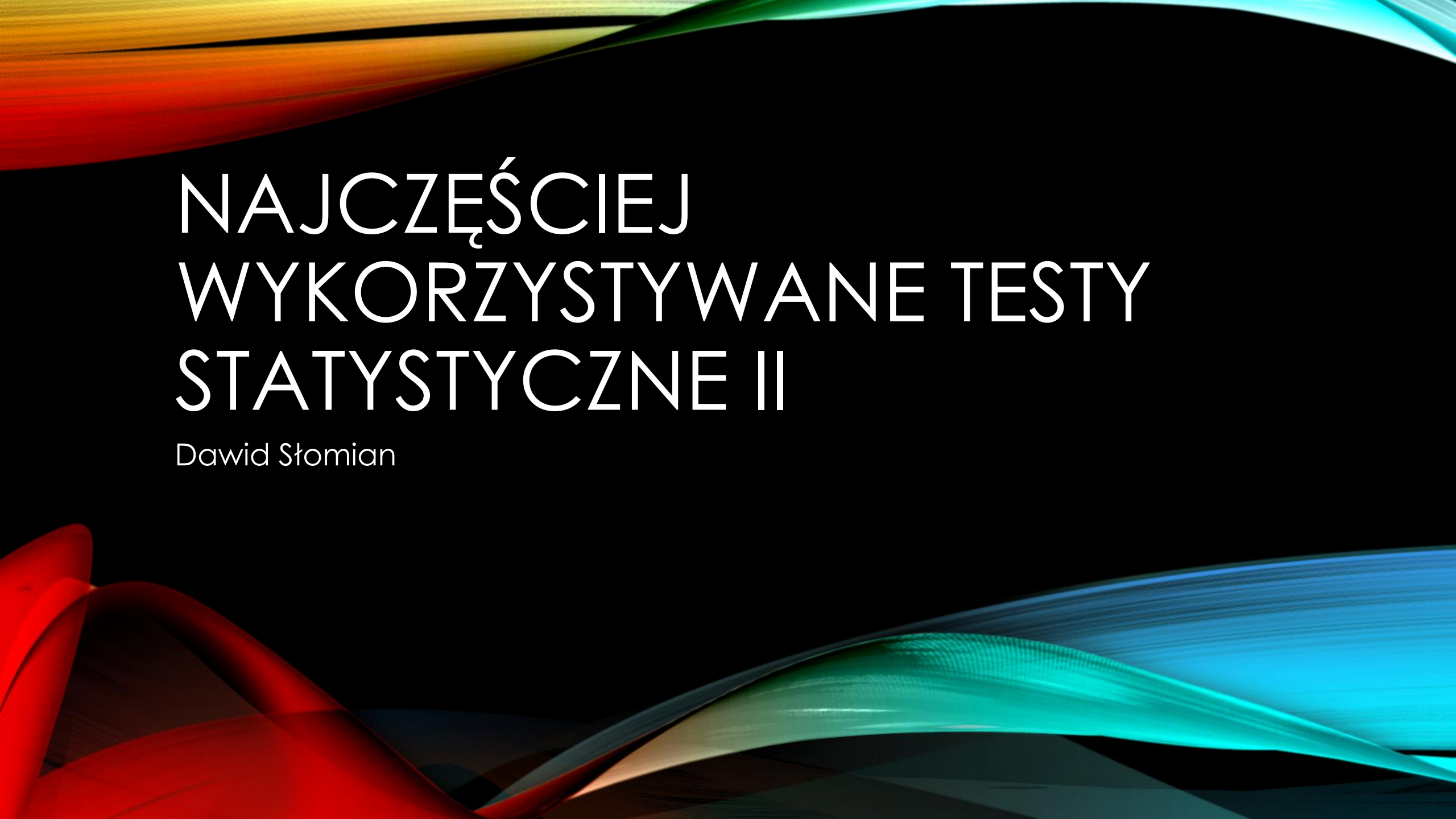


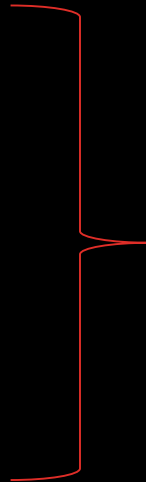


NAJCZĘŚCIEJ WYKORZYSTYWANE TESTY STATYSTYCZNE II

Dawid Słomian

WSTĘP

- Test F
- Test Kruskala-Wallisa
- Test chi-kwadrat

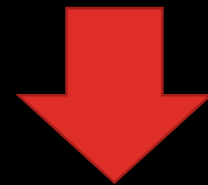
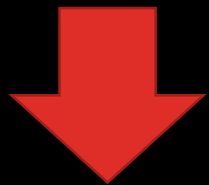


- Zakres stosowania
- Definicja
- Przykład

WSTĘP

Hipotezy

Próba danych



Test

TEST F

- Porównanie zmienności
- Dane ciągłe
- Rozkład normalny
- Warianty testu:

- Analiza wariancji
- Regresja

one-way ANOVA $H_0 : \sigma_A^2 \leq \sigma_e^2$

TEST F – ANALIZA WARIANCJI

- Zawartość azotu w trzcinie (% suchej masy)
- 3 lokalizacje (A, B, C), pomiar w 1996 r.
- Flowermere, hrabstwo Cambridge

A

3.06

2.60

2.55

2.42

2.35

B

3.41

3.23

3.93

3.74

3.18

C

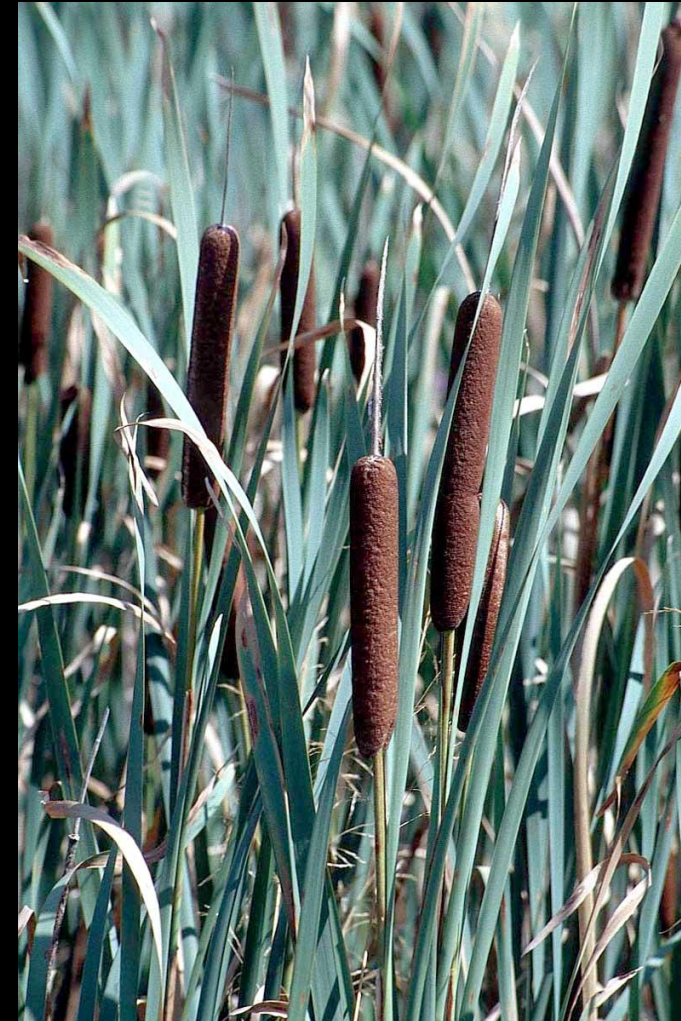
2.92

2.88

3.25

2.64

3.28



TEST F – ANALIZA WARIANCJI

1. Określenie hipotez H_0 i H_1

H_0 : lokalizacje nie mają wpływu na zawartość azotu

H_1 : lokalizacje wpływają na zawartość azotu

$$H_0: \sigma_{lok}^2 \leq \sigma_e^2$$

$$H_1: \sigma_{lok}^2 > \sigma_e^2$$

2. Ustalenie poziomu istotności

$$\alpha_{MAX} = 0.05$$

3. Wybór i obliczenie wartości testu statystycznego:

$$F = \frac{\frac{\sum_{i=1}^{N_g} n_i (\bar{y}_i - \bar{y})^2}{N_g - 1}}{\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{N - N_g}}$$

TEST F – ANALIZA WARIANCJI

3. Wybór i obliczenie wartości testu statystycznego

$$F = \frac{\frac{\sum_{i=1}^{N_g} n_i (\bar{y}_i - \bar{y})^2}{N_g - 1}}{\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{N - N_g}} = \frac{\frac{2.04}{3 - 1}}{\frac{1.03}{15 - 3}} = 11.95$$

4. Określenie rozkładu testu: $\sim F_{N_g - 1, N - N_g}$

Excel: 

5. Obliczenie wartości α_t : $\alpha_t = 0.00139$

R: `aov(surface~localisation,data=mays)`

6. Decyzja: $\alpha_t < \alpha_{\max}$

~~H_0~~ H_1

lokalizacje wpływają na zawartość azotu w suchej masie trzciny

TEST F – ZAKRES STOSOWANIA

1. Porównanie zmienności
2. Dane ciągłe
3. Rozkład normalny
4. Warianty testu:
 - Analiza wariancji
 - Regresja *regression* $H_0: \beta = 0$

TEST F – REGRESJA

MASA
CIAŁA

ZAW.
TŁUSZCZU

89	28
88	27
66	24
59	23
93	29
73	25
82	29
77	25
100	30
67	23

1. Masa ciała [kg] i grubość tkanki tłuszczowej [mm]
2. 10 osób dorosłych, różnej płci, tej samej rasy



TEST F - REGRESJA

1. Określenie hipotez H_0 i H_1

H_0 : grubość tkanki tłuszczowej nie zależy od masy ciała

H_1 : grubość tkanki tłuszczowej zależy od masy ciała

$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$

2. Ustalenie poziomu istotności

$$\alpha_{MAX} = 0.05$$

3. Wybór i obliczenie wartości testu statystycznego:

$$F = \frac{\frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{N_\beta - 1}}{\frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N - N_\beta}}$$

TEST F - REGRESJA

3. Wybór i obliczenie wartości testu statystycznego

$$F = \frac{\frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{N_\beta - 1}}{\frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{N - N_\beta}} = \frac{\frac{54.32}{2 - 1}}{\frac{7.78}{10 - 2}} = 55.8$$

4. Określenie rozkładu testu: $\sim F_{N_\beta - 1, N - N_\beta}$

Excel: 

5. Obliczenie wartości α_t : $\alpha_t = 0.000071$

R: `lm(tluszcz~masa,data=human_weight)`

6. Decyzja: $\alpha_t < \alpha_{\max}$ ~~H_0~~ H_1

grubość tkanki tłuszczowej zależy od masy ciała

TEST KRUSKALA-WALLISA

- Porównanie zmienności
- Dane ciągłe lub porządkowe (rangi)
- Dane nie pochodzą z rozkładu normalnego
- Analiza wariancji

TEST KRUSKALA-WALLISA

- Wzrost dorosłych kobiet w USA
- 3 przedziały wiekowe

20-29	30-39	40-49
161.925	164.465	173.990
173.355	171.450	175.260
158.115	173.355	167.640
170.815	175.260	166.370
179.705	164.465	168.910



TEST KRUSKALA-WALLISA

1. Określenie hipotez H_0 i H_1

H_0 : wzrost kobiet jest jednakowy w każdym p. wiekowym

H_1 : wzrost kobiet różni się w zależności od p. wiekowego

$$H_0: \sigma_{wiek}^2 \leq \sigma_e^2 \quad H_1: \sigma_{wiek}^2 > \sigma_e^2$$

2. Ustalenie poziomu istotności

$$\alpha_{MAX} = 0.05$$

3. Wybór i obliczenie wartości testu statystycznego:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{N_A} n_i (\bar{R}_i - \bar{R})^2 \sim \chi_{N_A-1}^2$$

N liczba obserwacji

N_A liczba grup

$\bar{R}_i$ śr. ranking obserwacji
w grupie i

$\bar{R}$ śr. ranking wszystkich obs.

TEST KRUSKALA-WALLISA

3. Wybór i obliczenie wartości testu statystycznego

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{N_A} n_i (\bar{R}_i - \bar{R})^2 = \frac{12}{15(15+1)} 5(7.2 - 8)^2 + 5(8 - 8)^2 + 5(8 - 8)^2 = 6.45$$

4. Określenie rozkładu testu: $\sim \chi^2_{3-1}$

5. Obliczenie wartości α_t : $\alpha_t = 0.88$

Excel: przykład 

6. Decyzja: $\alpha_t > \alpha_{\max}$ H_0 ~~H_1~~

wzrost dorosłych kobiet różni się w poszczególnych przedziałach wiekowych

TEST CHI-KWADRAT

- Testowanie liczebności poszczególnych kategorii
- Dane w formie liczebności → liczba obserwacji w danej kategorii
~~%, średnia, prawdopodobieństwo~~
- Dane podzielone na kategorie – nominalne, porządkowe ciągłe
- Nie należy stosować testu χ^2 gdy oczekiwana liczebność dla kategorii jest mała (<5)
- Warianty testu:
- Klasyfikacja jednoczynnikowa One-way classification
- Klasyfikacja dwuczynnikowa

TEST CHI-KWADRAT – KLASYFIKACJA JEDNOCZYNNIKOWA

KOLOR	CZĘSTOŚĆ
Biały	Żółty
Żółty	8
Żółty	Czerwony
Czerwony	5
Żółty	Biały
Żółty	4
Żółty	
Żółty	SUMA
Czerwony	17
Biały	
Czerwony	
Żółty	
Czerwony	
Czerwony	
Biały	
Żółty	
Biały	

- Klasyfikacja danych wg pojedynczego kryterium
- Kolor kwiatów krokusa



TEST CHI-KWADRAT – KLASYFIKACJA JEDNOCZYNNIKOWA

1. Określenie hipotez H_0 i H_1

H_0 : liczebności wystąpienia kolorów są jednakowe

H_1 : liczebności wystąpienia kolorów są różne

H_0 : $n_B = n_Z = n_C$

H_1 : $n_B \neq n_Z \neq n_C$

2. Ustalenie poziomu istotności

$$\alpha_{MAX} = 0.05$$

3. Wybór i obliczenie wartości testu statystycznego

$$\chi^2 = \sum_{i=1}^{N_K} \frac{(n_{obs}^i - n_{exp}^i)^2}{n_{exp}^i} = \frac{(n_{obs}^{bi} - n_{exp}^{bi})^2}{n_{exp}^{bi}} + \frac{(n_{obs}^{zó} - n_{exp}^{zó})^2}{n_{exp}^{zó}} + \frac{(n_{obs}^{cz} - n_{exp}^{cz})^2}{n_{exp}^{cz}}$$

TEST CHI-KWADRAT – KLASYFIKACJA JEDNOCZYNNIKOWA

3. Wybór i obliczenie wartości testu statystycznego

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{N_K} \frac{(n_{obs} - n_{exp})^2}{n_{exp}} = \frac{(n_{obs}^{bi} - n_{exp}^{bi})^2}{n_{exp}^{bi}} + \frac{(n_{obs}^{zó} - n_{exp}^{zó})^2}{n_{exp}^{zó}} + \frac{(n_{obs}^{cz} - n_{exp}^{cz})^2}{n_{exp}^{cz}} \\ &= \frac{(4 - 5.7)^2}{5.7} + \frac{(8 - 5.7)^2}{5.7} + \frac{(5 - 5.7)^2}{5.7} = 1.53\end{aligned}$$

4. Określenie rozkładu testu: $\sim \chi_{K-1=2}^2$

5. Obliczenie wartości α_t : $\alpha_t = 0.47$

6. Decyzja: $\alpha_t > \alpha_{max}$ $H_0 \not\equiv H_1$ Excel: przykład 

liczebności wystąpienia kolorów są jednakowe

TEST CHI-KWADRAT

- Testowanie liczebności poszczególnych kategorii
- Dane w formie liczebności → liczba obserwacji w danej kategorii
~~%, średnia, prawdopodobieństwo~~
- Dane podzielone na kategorie – nominalne, porządkowe ciągłe
- Nie należy stosować testu χ^2 gdy oczekiwana liczebność dla kategorii jest mała (<5)
- Warianty testu:
- Klasyfikacja jednoczynnikowa One-way classification
- Klasyfikacja dwuczynnikowa

TEST CHI-KWADRAT – KLASYFIKACJA DWUCZYNNIKOWA

1. Klasyfikacja danych wg dwu kryteriów
2. Liczebność słoni w Parku Narodowym Mikumi, Tanzania



		kategoria			
		samotny samiec	grupa samców	grupa rodzinna	grupa rodzinna + samiec
Pora	sucha	43	4	196	7
	deszczowa	92	17	195	8

TEST CHI-KWADRAT – KLASYFIKACJA DWUCZYNNIKOWA

1. Określenie hipotez H_0 i H_1

H_0 : pora roku nie wpływa na zmiany liczebności słoń w poszczególnych kategoriach (oba kryteria są niezależne)

H_1 : pora roku wpływa na zmiany liczebności słoń w poszczególnych kategoriach (oba kryteria są zależne)

H_0 : $n_{Di} = n_{Si}$ dla $i=1...4$ H_1 : $n_{Di} \neq n_{Si}$

2. Ustalenie poziomu istotności

$$\alpha_{MAX} = 0.05$$

3. Wybór i obliczenie wartości testu statystycznego

$$\chi^2 = \sum_{i=1}^{N_K} \sum_{j=1}^{N_M} \frac{(n_{obs}^{ij} - n_{exp}^{ij})^2}{n_{exp}^{ij}} = \frac{(n_{obs}^{D-1} - n_{exp}^{D-1})^2}{n_{exp}^{D-1}} + \dots + \frac{(n_{obs}^{S-4} - n_{exp}^{S-4})^2}{n_{exp}^{S-4}}$$

TEST CHI-KWADRAT – KLASYFIKACJA DWUCZYNNIKOWA

$$\chi^2 = \sum_{i=1}^{N_K} \sum_{j=1}^{N_M} \frac{(n_{obs}^{ij} - n_{exp}^{ij})^2}{n_{exp}^{ij}}$$

liczebność klasy oczekiwana dla H_0

$$n_{exp}^{ij} = \frac{\sum_{i=1}^{N_K} n_{obs}^{i \cdot}}{N} \cdot \frac{\sum_{j=1}^{N_M} n_{obs}^{\cdot j}}{N} \cdot N = \frac{\sum_{i=1}^{N_K} n_{obs}^{i \cdot} \sum_{j=1}^{N_M} n_{obs}^{\cdot j}}{N}$$

	samotny samiec	grupa samców	grupa rodzinna	grupa rodzinną + samiec	suma
sucha	43	4	196	7	250
deszczowa	92	17	195	8	312
suma	135	21	391	15	562

	samotny samiec	grupa samców	grupa rodzinną	grupa rodzinną + samiec
sucha	60.5	9.34	173.93	6.67
deszczowa	74.95	11.66	217.07	8.33

TEST CHI-KWADRAT – KLASYFIKACJA DWUCZYNNIKOWA

3. Wybór i obliczenie wartości testu statystycznego

$$\chi^2 = \sum_{i=1}^{N_K} \sum_{j=1}^{N_M} \frac{(n_{obs}^{ij} - n_{exp}^{ij})^2}{n_{exp}^{ij}} =$$
$$\frac{(43 - 60.05)^2}{60.05} + \frac{(92 - 74.95)^2}{74.95} + \frac{(4 - 9.34)^2}{9.34} + \frac{(17 - 11.66)^2}{11.66}$$
$$+ \frac{(196 - 173.93)^2}{173.93} + \frac{(195 - 217.07)^2}{217.07} + \frac{(7 - 6.67)^2}{6.67} + \frac{(8 - 8.33)^2}{8.33} =$$
$$19.30$$

4. Określenie rozkładu testu: $\sim \chi_{(K-1)(M-1)=3}^2$

5. Obliczenie wartości α_t : $\alpha_t = 0.0002$ Excel: przykład



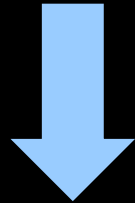
TEST CHI-KWADRAT – KLASYFIKACJA DWUCZYNNIKOWA

6. Decyzja: $\alpha_t < \alpha_{\max}$ ~~H_0~~ H_1

liczebność poszczególnych grup słońi różni się w zależności od pory roku

WYBÓR TESTU

porównanie zmienności

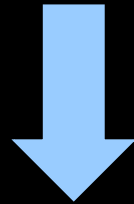


rozkład normalny

test F

WYBÓR TESTU

porównanie zmienności



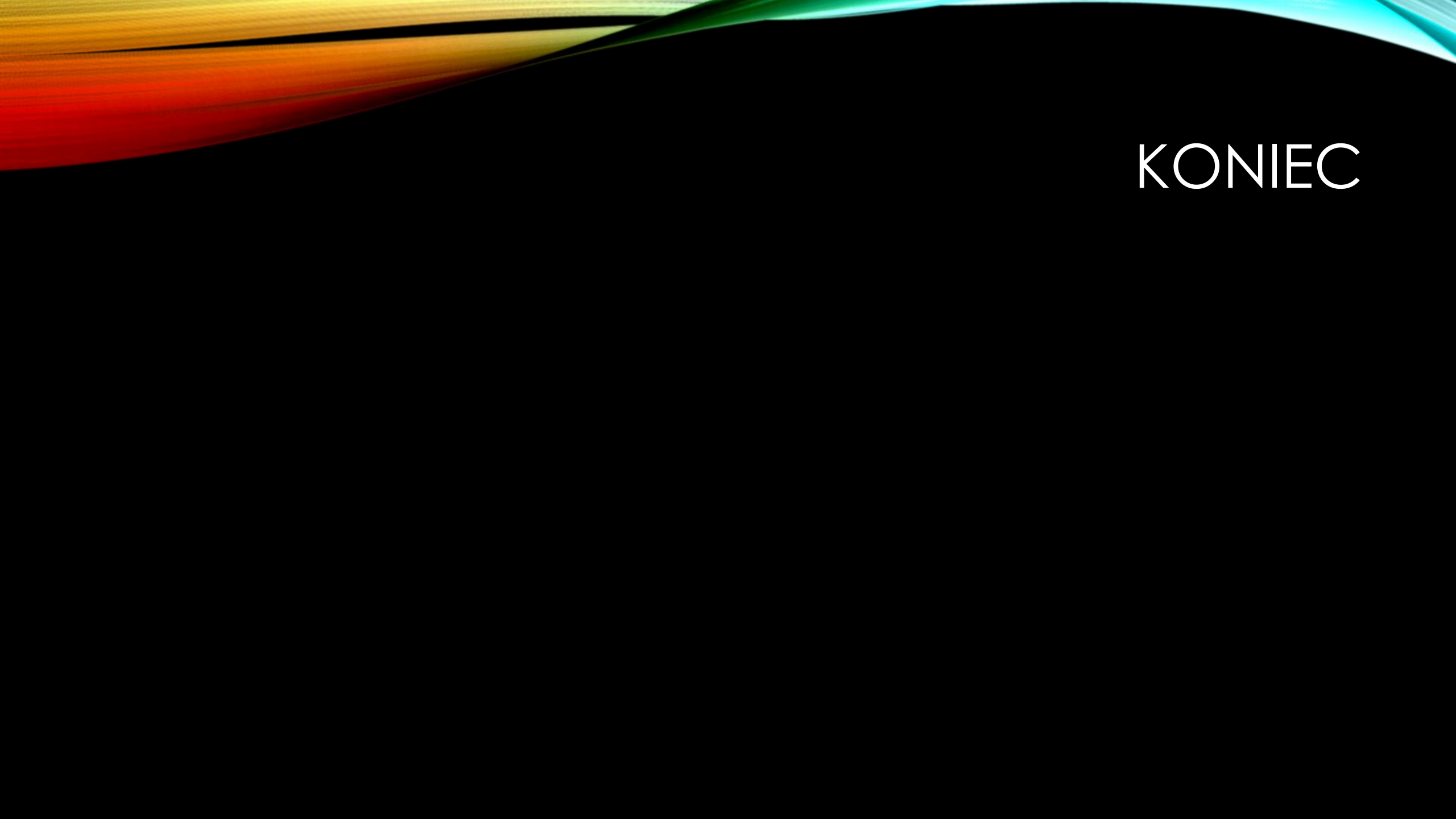
rozkład inny niż normalny

test Kruskala-Wallisa

WYBÓR TESTU

porównanie liczebności

test χ^2



KONIEC