

METODY STATYSTYCZNE W BIOLOGII

1. Wykład wstępny
2. Populacje i próby danych
3. Testowanie hipotez i estymacja parametrów
4. Planowanie eksperymentów biologicznych
5. Najczęściej wykorzystywane testy statystyczne
6. Podsumowanie materiału, wspólna analiza przykładów, dyskusja
7. Regresja liniowa
8. Regresja nieliniowa
9. Określenie jakości dopasowania równania regresji liniowej i nieliniowej
10. Korelacja
11. Elementy statystycznego modelowania danych - EDA
12. Porównywanie modeli
- 13. Analiza wariancji**
14. Analiza kowariancji
15. Podsumowanie materiału, wspólna analiza przykładów, dyskusja

Analiza wariancji

1. Model jednoczynnikowy
2. Model dwuczynnikowy
3. Model hierarchiczny

Analiza Wariancji – założenia

1. Zmienna zależna (y) pochodzi z rozkładu normalnego
2. Podobne wariancje wewnątrz każdej klasy
3. Poszczególne obserwacje są niezależne od siebie

Model Jednoczynnikowy

Model Jednoczynnikowy – próba danych

próba danych

1. Zawartość azotu w trzcinie (% suchej masy)
2. 3 lokalizacje (A, B, C), pomiar w 1996 r.
3. Flowermere, hrabstwo Cambridge

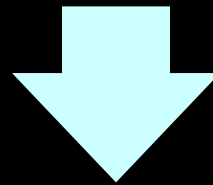
A	B	C
3.06	3.41	2.92
2.60	3.23	2.88
2.55	3.93	3.25
2.42	3.74	2.64
2.35	3.18	3.28



Model Jednoczynnikowy – próba danych

Obserwowana zmienność cechy

3.06	3.41	2.92
2.60	3.23	2.88
2.55	3.93	3.25
2.42	3.74	2.64
2.35	3.18	3.28



Różne lokalizacje

3.06	3.41	2.92
2.60	3.23	2.88
2.55	3.93	3.25
2.42	3.74	2.64
2.35	3.18	3.28

Model Jednoczynnikowy – model

MODEL ANALIZY WARIANCJI:

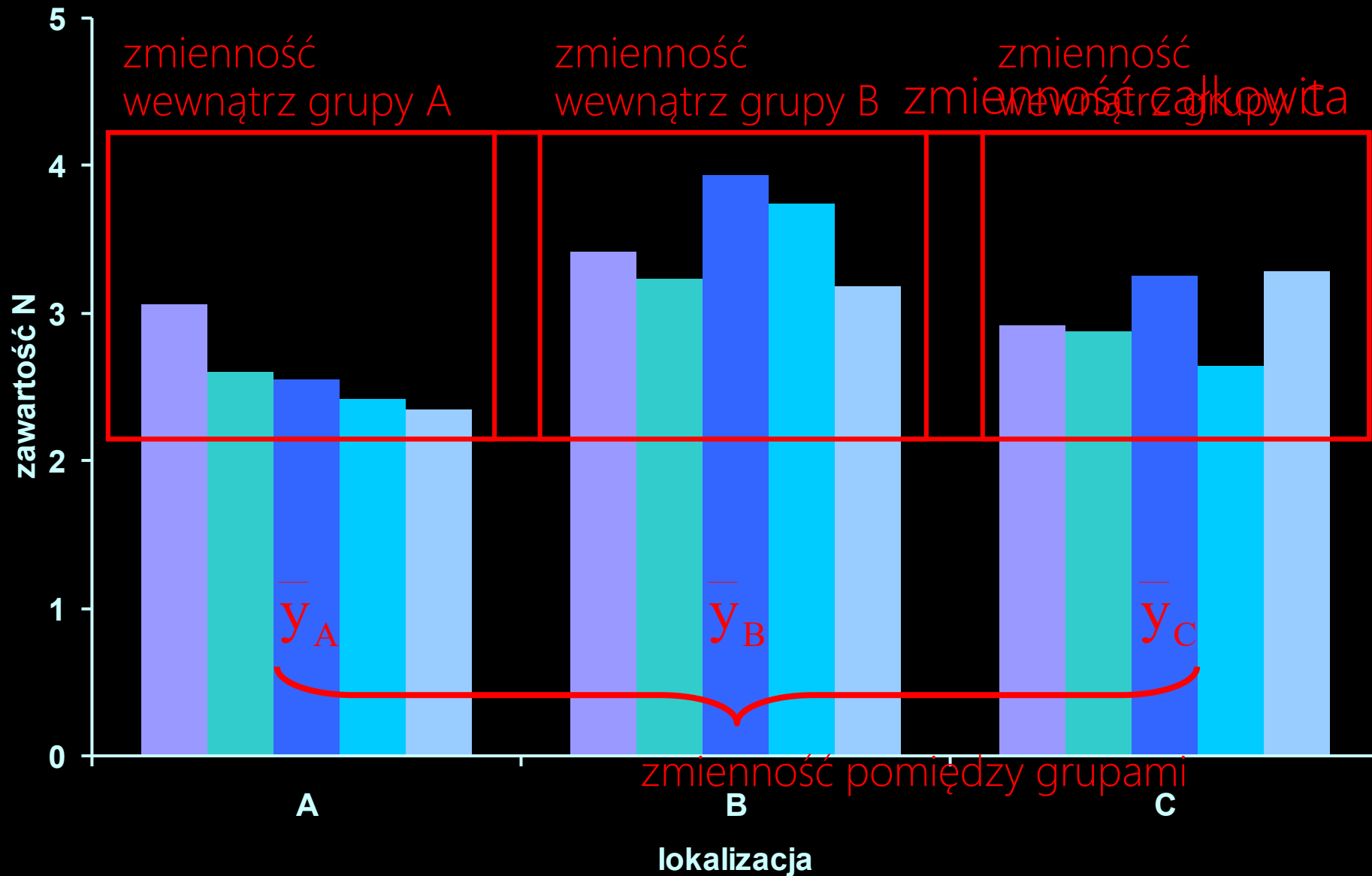
$$\text{Zawartość } N = \mu + \text{lokalizacja} + e$$

Efekty wspólne dla wszystkich obserwacji
Nie wpływają na zmienność

Efekty niezmierzone w doświadczeniu
Wpływają na zmienność

Analizowany czynnik
Testujemy czy wpływa na zmienność ???

Model Jednoczynnikowy – model



Model Jednoczynnikowy – model

Schemat Jednoczynnikowej Analizy Wariancji

$$\sigma_{\text{azot}}^2 = \sigma_{\text{lokalizacja}}^2 + \sigma_e^2$$

źródło zmienności	suma kwadratów	stopnie swobody	średni kwadrat
Pomiędzy gr. (lokalizacja)	$\sum_{i=1}^{N_g} n_i (\bar{y}_i - \bar{y})^2$	$N_g - 1$	$\frac{\sum_{i=1}^{N_g} n_i (\bar{y}_i - \bar{y})^2}{N_g - 1}$
Wewnątrz gr. błąd	$\sum_{i=1}^{N_g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$N - N_g$	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{N - N_g}$
Całkowita	$\sum_{i=1}^{N_g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$	$N - 1$	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2}{N - 1}$

Model Jednoczynnikowy – testowanie hipotez

1. Określenie hipotez H_0 i H_1

H_0 : lokalizacje nie wpływają na zawartość azotu

H_1 : lokalizacje wpływają na zawartość azotu

$$H_0: \sigma_{lok}^2 \leq \sigma_e^2 \quad \mu_1 = \mu_2 = \mu_3$$

$$H_1: \sigma_{lok}^2 > \sigma_e^2 \quad \mu_1 \neq \mu_2 \neq \mu_3$$

2. Ustalenie poziomu istotności

$$\alpha_{MAX} = 0.05$$

3. Wybór testu

test F

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


średnia zmienność wartości y
spowodowana różnymi lokalizacjami

średnia zmienność wartości y nie wyjaśniona
przez różne lokalizacje = średni błąd

3. Wybór i obliczenie wartości testu statystycznego

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

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

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

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

lokalizacje wpływają na zawartość azotu w suchej masie trzciny

Model Dwuczynnikowy

Model Dwuczynnikowy – próba danych

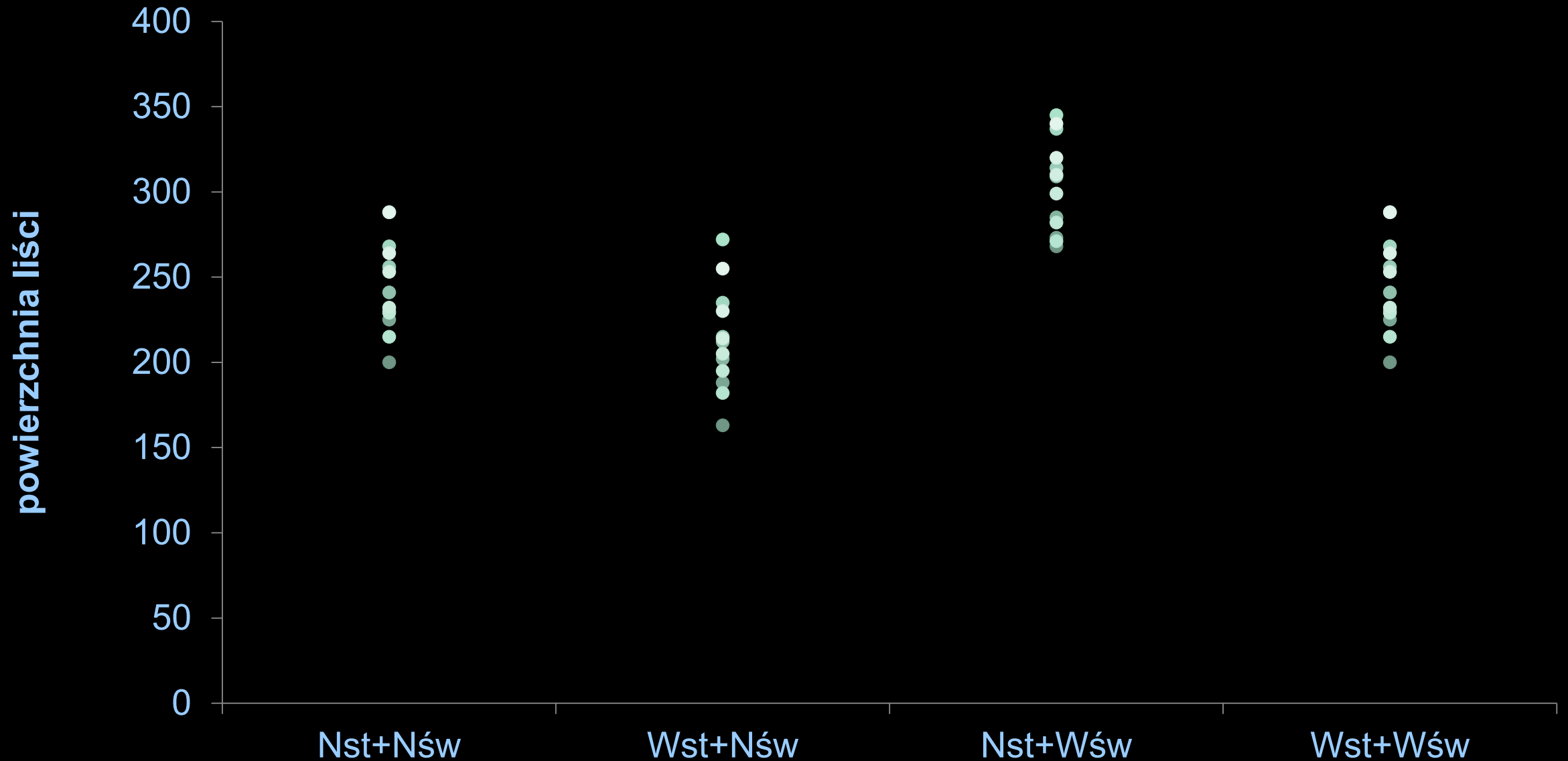
próba danych

1. Wzrost soi - powierzchnia liści
2. Stres mechaniczny
3. Nasłonecznienie



	niski poziom stresu	wysoki poziom stresu
słabe nasłonecznienie	200, 225, 230, ... , 264, 288	163, 188, 202, ... , 230, 255
dobre nasłonecznienie	268, 273, 285, ... , 320, 349	200, 225, 230, ... , 264, 288

Model Dwuczynnikowy – próba danych



Model Dwuczynnikowy – model

źródło zmienności	suma kwadratów	stopnie swobody	średni kwadrat
Pomiędzy gr. (stres)	$\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2$	$N_{st} - 1$	$\frac{\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2}{N_{st} - 1}$
Pomiędzy gr. (słońce)	$\sum_{j=1}^{N_{św}} n_j (\bar{y}_j - \bar{y})^2$	$N_{św} - 1$	$\frac{\sum_{j=1}^{N_{św}} n_j (\bar{y}_j - \bar{y})^2}{N_{św} - 1}$
Interakcja	$\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} n_{ij} [(\bar{y}_{ij} - \bar{y}) - (...)]^2$		
Wewnątrz gr. błąd			
Całkowita			

Model Dwuczynnikowy – model

źródło zmienności	suma kwadratów	stopnie swobody	średni kwadrat
Pomiędzy gr. (stres)	$\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2$	$N_{st} - 1$	$\frac{\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2}{N_{st} - 1}$
Pomiędzy gr. (słońce)	$\sum_{j=1}^{N_{św}} n_j (\bar{y}_j - \bar{y})^2$	$N_{św} - 1$	$\frac{\sum_{j=1}^{N_{św}} n_j (\bar{y}_j - \bar{y})^2}{N_{św} - 1}$
Interakcja	$\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} n_{ij} [(\bar{y}_{ij} - \bar{y}) - (\bar{y}_i - \bar{y}) - (\bar{y}_j - \bar{y})]^2$		
Wewnątrz gr. błąd			
Całkowita			

Model Dwuczynnikowy – model

źródło zmienności	suma kwadratów	stopnie swobody	średni kwadrat
Pomiędzy gr. (stres)	$\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2$	$N_{st} - 1$	$\frac{\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2}{N_{st} - 1}$
Pomiędzy gr. (słońce)	$\sum_{j=1}^{N_{św}} n_j (\bar{y}_j - \bar{y})^2$	$N_{św} - 1$	$\frac{\sum_{j=1}^{N_{św}} n_j (\bar{y}_j - \bar{y})^2}{N_{św} - 1}$
Interakcja	$\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} n_{ij} (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$	$(N_{st} - 1)(N_{św} - 1)$	$\frac{\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} n_{ij} (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2}{N - N_{st} N_{św}}$
Wewnątrz gr. błąd	$\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$	$N - N_{st} N_{św}$	$\frac{\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{N - N_{st} N_{św}}$
Całkowita	$\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y})^2$	$N - 1$	$\frac{\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y})^2}{N - 1}$

Model Dwuczynnikowy – testowanie hipotez

1. Określenie hipotez H_0 i H_1

H_0 : stres mechaniczny nie wpływa na powierzchnię liści

H_1 : stres mechaniczny wpływa na powierzchnię liści

$$H_0: \sigma_{stres}^2 \leq \sigma_e^2 \quad \mu_1 = \mu_2$$

$$H_1: \sigma_{stres}^2 > \sigma_e^2 \quad \mu_1 \neq \mu_2$$

2. Ustalenie poziomu istotności

$$\alpha_{MAX} = 0.05$$

3. Wybór i obliczenie wartości testu statystycznego

Model Dwuczynnikowy – testowanie hipotez

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

Pomiędzy gr.
(stres)

$$\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2$$

$$N_{st} - 1$$

$$\frac{\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2}{N_{st} - 1}$$

Pomiędzy gr.
(słońce)

Interakcja

Wewnątrz gr.
błąd

$$\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$$

$$N - N_{st} N_{św}$$

$$\frac{\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{N - N_{st} N_{św}}$$

Całkowita

F

3. Wybór i obliczenie wartości testu statystycznego

$$F = \frac{\sum_{i=1}^{N_{st}} n_i (\bar{y}_i - \bar{y})^2}{\frac{\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{\acute{s}w}} \sum_{k=1}^{N_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{N - N_{st} N_{\acute{s}w}}} = \frac{27187}{\frac{36527}{52 - 4}} = 35.73$$

4. Określenie rozkładu testu: $\sim F_{N_{st}-1, N - N_{st} N_{\acute{s}w}}$

5. Obliczenie wartości α_t : $\alpha_t = 0.000000027$

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

stres mechaniczny wpływa na powierzchnię liści

Model Dwuczynnikowy – testowanie hipotez

Poziom nastonecznienia: $H_0 : \sigma_{słońce}^2 \leq \sigma_e^2$ $H_1 : \sigma_{słońce}^2 > \sigma_e^2$

Pomiędzy gr.
(stres)

Pomiędzy gr.
(słońce)

Interakcja

Wewnątrz gr.
błąd

Całkowita

$$\sum_{j=1}^{N_{św}} n_j (\bar{y}_j - \bar{y})^2$$

$$N_{św} - 1$$

$$\frac{\sum_{j=1}^{N_{św}} n_j (\bar{y}_j - \bar{y})^2}{N_{św} - 1}$$

F

$$\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$$

$$N - N_{st} N_{św}$$

$$\frac{\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{św}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{N - N_{st} N_{św}}$$

Model Dwuczynnikowy – testowanie hipotez

Interakcja: $H_0 : \sigma_{stres*slo\ncce}^2 \leq \sigma_e^2 \quad H_1 : \sigma_{stres*slo\ncce}^2 > \sigma_e^2$

Pomiędzy gr.
(stres)

Pomiędzy gr.
(słońce)

Interakcja

Wewnątrz gr.
błąd

Całkowita

$$\begin{array}{ccc}
 \sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{s\acute{w}}} n_{ij} (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2 & (N_{st} - 1)(N_{s\acute{w}} - 1) & \frac{\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{s\acute{w}}} \sum_{k=1}^{n_{ij}} n_{ij} (\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2}{N - N_{st}N_{s\acute{w}}} \\
 \sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{s\acute{w}}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2 & N - N_{st}N_{s\acute{w}} & \frac{\sum_{i=1}^{N_{st}} \sum_{j=1}^{N_{s\acute{w}}} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{N - N_{st}N_{s\acute{w}}}
 \end{array}$$

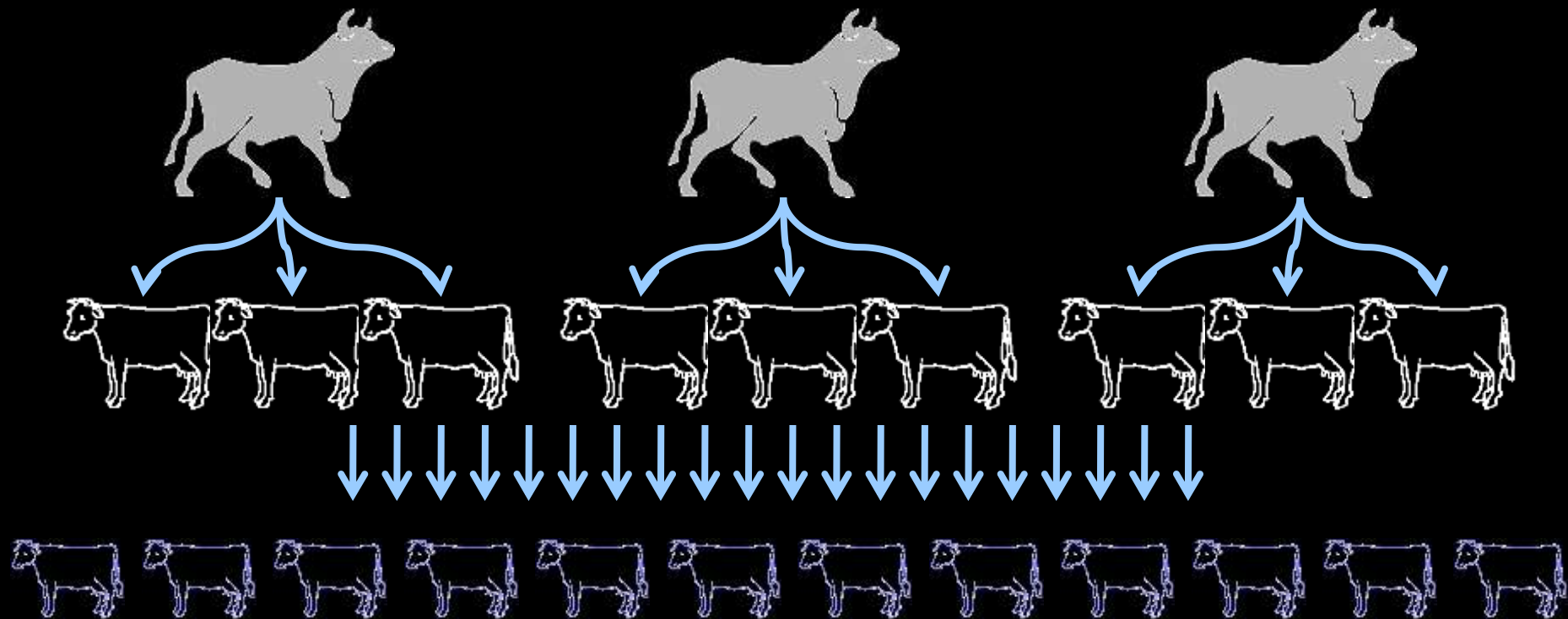
F

Model Hierarchiczny

Model Hierarchiczny – próba danych

próba danych

1. Liczba komórek somatycznych w mleku (SCS) – zdrowotność wymienia
2. Ojciec (genotyp)
3. Matka (genotyp)



Model Hierarchiczny – model

źródło zmienności	suma kwadratów	stopnie swobody	średni kwadrat
Pomiędzy gr. (buhaj)	$\sum_{i=1}^{N_b} n_i (\bar{y}_i - \bar{y})^2$	$N_b - 1$	$\frac{\sum_{i=1}^{N_b} n_i (\bar{y}_i - \bar{y})^2}{N_b - 1}$
Pomiędzy krowami wewnątrz buhajów	$\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} n_{ij} (\bar{y}_{ij} - \bar{y}_i)^2$	$N_b (N_k - 1)$	$\frac{\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} n_{ij} (\bar{y}_{ij} - \bar{y}_i)^2}{N_b (N_k - 1)}$
Wewnątrz gr. błąd	$\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$	$N - N_b N_k$	$\frac{\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{N - N_b N_k}$
Całkowita	$\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y})^2$	$N - 1$	$\frac{\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y})^2}{N - 1}$

Model Hierarchiczny – testowanie hipotez

Wpływ buhaja: $H_0 : \sigma_{buhaj}^2 \leq \sigma_e^2$ $H_1 : \sigma_{buhaj}^2 > \sigma_e^2$

Pomiędzy gr.
(buhaj)

$$\sum_{i=1}^{N_b} n_i (\bar{y}_i - \bar{y})^2$$

$N_b - 1$

$$\frac{\sum_{i=1}^{N_b} n_i (\bar{y}_i - \bar{y})^2}{N_b - 1}$$

Pomiędzy krowami
wewnątrz buhajów

F

Wewnątrz gr.
błąd

$$\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$$

$N - N_b N_k$

$$\frac{\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{N - N_b N_k}$$

Całkowita

1. Określenie hipotez H_0 i H_1

H_0 : genotyp krowy nie wpływa na liczbę komórek somatycznych w mleku córek

H_1 : genotyp krowy wpływa na liczbę komórek somatycznych w mleku córek

$$H_0: \sigma_{krowa(buhaj)}^2 \leq \sigma_e^2 \quad \mu_{1a} = \mu_{1b} = \mu_{1c} \wedge \mu_{2d} = \mu_{2e} = \mu_{2f}$$

$$H_1: \sigma_{krowa(buhaj)}^2 > \sigma_e^2 \quad \mu_{1a} \neq \mu_{1b} \neq \mu_{1c} \vee \mu_{2d} \neq \mu_{2e} \neq \mu_{2f}$$

2. Ustalenie poziomu istotności

$$\alpha_{MAX} = 0.05$$

3. Wybór i obliczenie wartości testu statystycznego

Model Hierarchiczny – testowanie hipotez

$$H_0: \sigma_{krowa(buhaj)}^2 \leq \sigma_e^2 \quad H_1: \sigma_{krowa(buhaj)}^2 > \sigma_e^2$$

Pomiędzy gr.
(buhaj)

Pomiędzy krowami
wewnątrz buhajów

Wewnątrz gr.
błąd

Całkowita

$$\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} n_{ij} (\bar{y}_{ij} - \bar{y}_i)^2$$

$$\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2$$

F

$$N_b(N_k - 1)$$

$$N - N_b N_k$$

$$\frac{\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} n_{ij} (\bar{y}_{ij} - \bar{y}_j)^2}{N_b(N_k - 1)}$$

$$\frac{\sum_{i=1}^{N_b} \sum_{j=1}^{N_k} \sum_{k=1}^{n_{ij}} (y_{ijk} - \bar{y}_{ij})^2}{N - N_b N_k}$$

3. Wybór i obliczenie wartości testu statystycznego

$$F = \frac{\sum_{i=1}^3 \sum_{j=1}^3 3(\bar{y}_{ij} - \bar{y}_j)^2}{3(3-1)} = \frac{14114}{15391} = 2.75$$
$$\frac{\sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 (y_{ijk} - \bar{y}_{ij})^2}{27 - 3 \cdot 3}$$

4. Określenie rozkładu testu: $\sim F_{N_b(N_k-1), N-N_bN_k}$

5. Obliczenie wartości α_t : $\alpha_t = 0.04$

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

genotyp krowy wpływa na liczbę komórek somatycznych w mleku córek

Analiza wariancji

1. Model jednoczynnikowy
2. Model dwuczynnikowy
3. Model hierarchiczny