

Pakiety statystyczne - Wykład 9

Tomasz Suchocki

Uniwersytet Przyrodniczy we Wrocławiu
Katedra Genetyki



Analiza wariancji

1. Rys historyczny
2. Podstawy teoretyczne i przykłady zastosowania
3. ANOVA w pakiecie R



1. ANalysis Of VAriance → ANOVA

- ▶ metoda statystyki matematycznej
- ▶ badanie obserwacji zależnych od jednego lub wielu czynników
- ▶ porównywanie wielu średnich



2. Sir Ronald Fisher (1890 – 1962)

- ▶ statystyk i genetyk
- ▶ instytut badań rolniczych w Rothamsted koło Harpenden
- ▶ 1918 - *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*
- ▶ pierwsze zastosowania zostały opublikowane w 1921 roku
- ▶ szeroko stosowana od 1925 roku po ukazaniu się książki *Statistical Methods for Research Workers*





Analiza wariancji

- ▶ jednoczynnikowa
- ▶ wieloczynnikowa
- ▶ interakcje?!



Jak to działa?

1. Hipoteza zerowa

- ▶ Nie ma różnic pomiędzy grupami! To tylko błąd!

2. Hipoteza alternatywna

- ▶ Istnieją różnice pomiędzy grupami, które nie są spowodowane przez błędy!

3. Statystyka F

- ▶ Oblicz statystykę F dla odpowiedniego modelu.



Jednoczynnikowa analiza wariancji:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- ▶ $j = 1, \dots, n_i$ oraz $i = 1, \dots, k$
- ▶ α_i efekt czynnika i
- ▶ ϵ_{ij} niezależne zmienne o identycznych rozkładach normalnych $\mathcal{N}(0, \sigma^2)$



Model można zapisać w innej postaci:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

▶ $\mu_i = \mu + \alpha_i$



Model można zapisać w innej postaci:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- ▶ $\mu_i = \mu + \alpha_i$
- ▶ zatem $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$



Model można zapisać w innej postaci:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- ▶ $\mu_i = \mu + \alpha_i$
- ▶ zatem $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$

Testujemy hipotezy:

- ▶ $H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu$
- ▶ $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$



Model można zapisać w innej postaci:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- ▶ $\mu_i = \mu + \alpha_i$
- ▶ zatem $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$

Testujemy hipotezy:

- ▶ $H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu$
- ▶ $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$

A jaką postać ma hipoteza alternatywna?



Hipoteza alternatywna:

$$\blacktriangleright H_1 : \exists_{i,j} \quad \mu_i \neq \mu_j$$



Hipoteza alternatywna:

▶ $H_1 : \exists_{i,j} \quad \mu_i \neq \mu_j$

Ale które średnie się różnią?

- ▶ testy post hoc - porównują wszystkie pary średnich



Statystyka F:

$$F = \frac{\sum_{i=1}^k n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2} \cdot \frac{n - k}{k - 1}$$

gdzie

- ▶ $\bar{Y}_{..} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$
- ▶ $\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$
- ▶ statystyka F ma rozkład F Snedecora z $k - 1$ i $n - k$ stopniami swobody



Testy post hoc:

- ▶ Test HSD Tukeya
- ▶ Test Studenta-Newmana-Keulsa
- ▶ Test LSD Fishera
- ▶ Test Scheffego



Przykład - ANOVA jednoczynnikowa

- ▶ Obecność witaminy D₃ wpływa na aktywność genu CD14, co powoduje uruchomieniem procesu wykorzystywanego w terapii przeciwnowotworowej u chorych na białaczkę.
- ▶ Podanie witaminy D₃ w odpowiednio wysokim stężeniu powoduje między innymi odwapnienie kości.
- ▶ Chcemy sprawdzić, czy istnieją analogi witaminy D₃ o podobnym wpływie na ekspresję genu CD14 ale bez skutków ubocznych.



Przykład - ANOVA jednoczynnikowa

Zbiór danych, dotyczy 66 linii komórkowych pobranych od dzieci chorych na ostrą białaczkę szpikową. Każda z linii komórkowych została poddana działaniu kilku substancji, by zobaczyć, jak dana substancja działa na aktywność genu CD14. Liczony był % komórek z obecnym receptorem CD14. Kolejne kolumny zawierają:

- ▶ zmienną mutacja - czy w danej linii komórkowej występuje mutacja genu związanego z białaczką (cztery grupy: mutacja w genie $CBF\alpha$, mutacja w FTL-3, mutacja w innym genie lub brak mutacji);
- ▶ pomiar % obecności CD14 w komórkach hodowlanych bez ingerencji innej substancji;
- ▶ pomiar % obecności CD14 w komórkach hodowlanych dla określonego molowego stężenia witaminy D₃ lub jej analogu 1906 lub 2191.

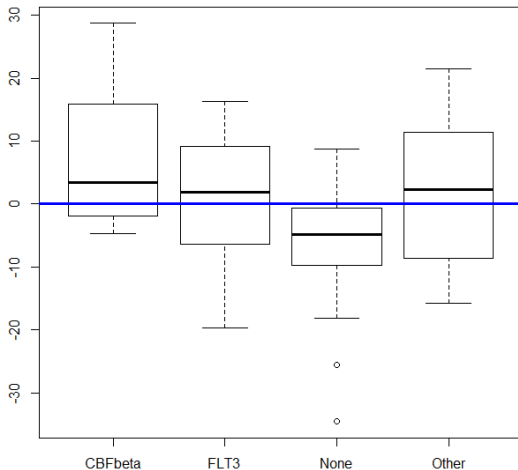


Przykład - ANOVA jednoczynnikowa

```
> require(PBImisc)
Loading required package: PBImisc
Loading required package: MASS
Warning message:
package 'PBImisc' was built under R version 2.15.1
> data(AML)
> AML$change.2191=AML$CD14.2191-AML$CD14.control
> attach(AML)
> by(change.2191,Mutation,mean)
Mutation: CBFbeta
[1] 7.814
-----
Mutation: FLT3
[1] 0.8271429
-----
Mutation: None
[1] -6.680556
-----
Mutation: Other
[1] 1.753158
```



Przykład - ANOVA jednoczynnikowa





Przykład - ANOVA jednoczynnikowa

Hipoteza zerowa:

- ▶ średni poziom receptora CD14 po podaniu analogu 2191 nie zależy od mutacji
- ▶ $H_0 : \forall_{i,j} \quad \mu_i = \mu_j$



Przykład - ANOVA jednoczynnikowa

Hipoteza zerowa:

- ▶ średni poziom receptora CD14 po podaniu analogu 2191 nie zależy od mutacji
- ▶ $H_0 : \forall_{i,j} \quad \mu_i = \mu_j$

Hipoteza alternatywna:

- ▶ średni poziom receptora CD14 po podaniu analogu 2191 jest różny przynajmniej dla dwóch mutacji
- ▶ $H_1 : \exists_{i,j} \quad \mu_i \neq \mu_j$



Przykład - ANOVA jednoczynnikowa

Wyniki:

```
> summary(aov(CD14.2191~Mutation))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Mutation	3	9404	3134.7	6.136	0.00101 **
Residuals	62	31673	510.8		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Przykład - ANOVA jednoczynnikowa

Pomiędzy którymi mutacjami są istotne różnice:

```
> pairwise.t.test(CD14.2191,Mutation)
```

```
Pairwise comparisons using t tests with pooled SD
```

```
data: CD14.2191 and Mutation
```

	CBFbeta	FLT3	None
FLT3	0.1201	-	-
None	0.0060	0.5145	-
Other	0.9598	0.1087	0.0031

```
P value adjustment method: holm
```



Przykład - ANOVA jednoczynnikowa

Diagnostyka modelu:

```
> shapiro.test(lm(CD14.2191~Mutation)$residuals)
```

```
Shapiro-Wilk normality test
```

```
data:  lm(CD14.2191 ~ Mutation)$residuals  
W = 0.9804, p-value = 0.3787
```



Przykład - ANOVA jednoczynnikowa

Diagnostyka modelu cd.:

```
> bartlett.test(CD14.2191~Mutation)
```

```
Bartlett test of homogeneity of variances
```

```
data: CD14.2191 by Mutation
```

```
Bartlett's K-squared = 1.2159, df = 3, p-value = 0.7492
```



Wieloczynnikowa analiza wariancji:

$$Y_{ijm} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijm}$$

- ▶ $j = 1, \dots, r$, $i = 1, \dots, k$ oraz $m = 1, \dots, n_{i,j}$
- ▶ α_i efekt czynnika i
- ▶ β_j efekt czynnika j
- ▶ γ_{ij} interakcja pomiędzy czynnikami α_i i β_j
- ▶ ϵ_{ijm} niezależne zmienne o identycznych rozkładach normalnych $\mathcal{N}(0, \sigma^2)$



Hipotezy zerowe

- ▶ $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$
- ▶ $H_0 : \beta_1 = \beta_2 = \dots = \beta_r = 0$
- ▶ $H_0 : \forall_i \forall_j \gamma_{i,j} = 0$

Hipotezy alternatywne

- ▶ $H_A : \exists_i \alpha_i \neq 0$
- ▶ $H_A : \exists_j \beta_j \neq 0$
- ▶ $H_A : \exists_i \exists_j \gamma_{i,j} \neq 0$



Przykład - ANOVA wieloczynnikowa

Zbiór danych dotyczy 98 osób. Dla każdej z nich mamy informację o genotypach dwóch markerów typu SNP. Dokładna struktura kolumn prezentuje się następująco:

- ▶ genotyp SNP w genie kodującym białko kompleksu $Nf\kappa\beta$
- ▶ genotyp SNP w genie CD28
- ▶ genotyp genu kodującego jeden z interferonów IFN (tutaj mamy informację, czy pacjent jest hetero-, czy homozygotą)
- ▶ natężenie objawów maniakałnych
- ▶ natężenie objawów zniekształcenia postrzegania rzeczywistości
- ▶ natężenie objawów depresyjnych
- ▶ natężenie objawów negatywnych (niewidzenie rzeczy istniejących)
- ▶ natężenie łącznych objawów



Przykład - ANOVA wieloczynnikowa

```
> data(schizophrenia)
> summary(schizophrenia)
```

NfkB	CD28	IFN	Dikeos.manic	Dikeos.reality.distortion
AA:37	TC+CC:33	AT :53	Min. : 0.000	Min. :10.00
AG:43	TT :65	TT+AA:45	1st Qu.: 4.000	1st Qu.:24.00
GG:18			Median : 6.000	Median :26.50
			Mean : 5.847	Mean :24.91
			3rd Qu.: 8.000	3rd Qu.:28.00
			Max. :13.000	Max. :31.00

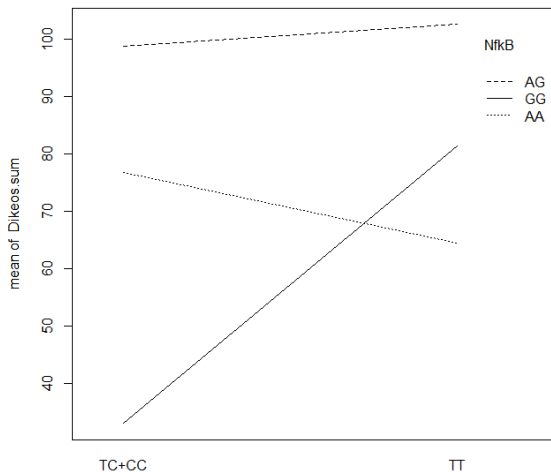
Dikeos.depression	Dikeos.disorganization	Dikeos.negative	Dikeos.sum
Min. : 2.00	Min. : 2.00	Min. :0.000	Min. : 22.00
1st Qu.: 9.00	1st Qu.:26.00	1st Qu.:2.000	1st Qu.: 72.00
Median :12.50	Median :41.00	Median :2.000	Median : 82.50
Mean :12.29	Mean :39.19	Mean :2.633	Mean : 84.87
3rd Qu.:16.00	3rd Qu.:52.50	3rd Qu.:4.000	3rd Qu.:102.75
Max. :19.00	Max. :75.00	Max. :8.000	Max. :128.00

```
> |
```



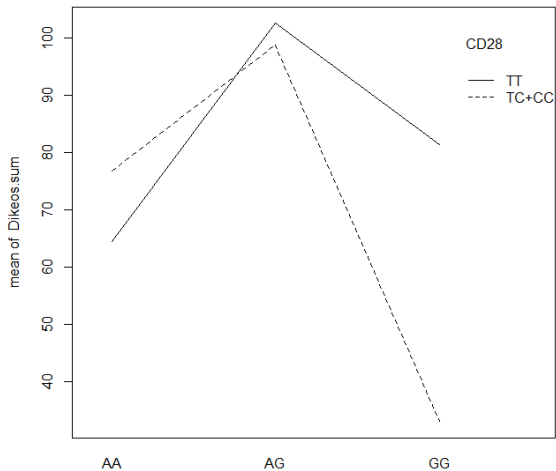
Przykład - ANOVA wieloczynnikowa

interaction.plot



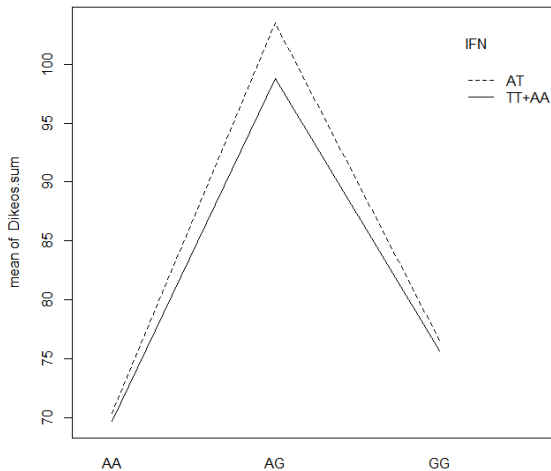


Przykład - ANOVA wieloczynnikowa



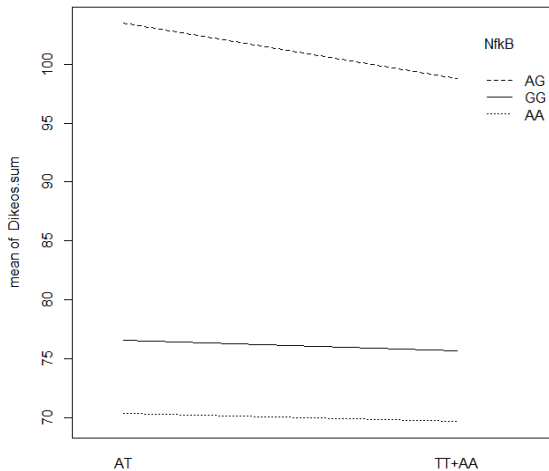


Przykład - ANOVA wieloczynnikowa





Przykład - ANOVA wieloczynnikowa





Przykład - ANOVA wieloczynnikowa

Model z interakcją, czy bez?

```
> m1=lm(Dikeos.sum~NfkB+CD28)
> m2=lm(Dikeos.sum~NfkB*CD28)
> anova(m1,m2)
```

Analysis of Variance Table

Model 1: Dikeos.sum ~ NfkB + CD28

Model 2: Dikeos.sum ~ NfkB * CD28

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	94	20715				
2	92	15040	2	5675	17.357	4.021e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Przykład - ANOVA wieloczynnikowa

Wyniki:

```
> summary(aov(Dikeos.sum~NfkB*CD28))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
NfkB	2	21119	10560	64.593	< 2e-16 ***
CD28	1	5	5	0.028	0.868
NfkB:CD28	2	5675	2838	17.357	4.02e-07 ***
Residuals	92	15040	163		

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
< |
```

Przykład - ANOVA wieloczynnikowa

Testy post hoc:

```
> TukeyHSD(aov(Dikeos.sum~NfkB*CD28))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Dikeos.sum ~ NfkB * CD28)

$NfkB
      diff      lwr      upr      p adj
AG-AA 31.221245 24.391131 38.05136 0.0000000
GG-AA  5.918919 -2.834188 14.67203 0.2462074
GG-AG -25.302326 -33.853237 -16.75141 0.0000000

$CD28
      diff      lwr      upr      p adj
TT-TC+CC 0.4391011 -4.988792 5.866994 0.8727066

$`NfkB:CD28`
      diff      lwr      upr      p adj
AG:TC+CC-AA:TC+CC 22.008403  8.577121 35.439685 0.0000991
GG:TC+CC-AA:TC+CC -43.705882 -71.526223 -15.885542 0.0002126
AA:TT-AA:TC+CC -12.255882 -24.532720  0.020955 0.0506551
AG:TT-AA:TC+CC  25.845842 14.477934 37.213750 0.0000000
GG:TT-AA:TC+CC  4.669118 -8.293646 17.631882 0.9000700
GG:TC+CC-AG:TC+CC -65.714286 -93.846640 -37.581932 0.0000000
AA:TT-AG:TC+CC -34.264286 -47.232660 -21.295911 0.0000000
AG:TT-AG:TC+CC  3.837438 -8.274016 15.948893 0.9398394
GG:TT-AG:TC+CC -17.339286 -30.958803 -3.719768 0.0047113
AA:TT-GG:TC+CC  31.450000  3.850168 59.049832 0.0159841
AG:TT-GG:TC+CC  69.551724 42.344017 96.759432 0.0000000
GG:TT-GG:TC+CC  48.375000 20.463295 76.286705 0.0000330
AG:TT-AA:TT  38.101724 27.284667 48.918782 0.0000000
GG:TT-AA:TT  16.925000  4.442506 29.407494 0.0020898
GG:TT-AG:TT -21.176724 -32.766429  -9.587020 0.0000107
```



Przykład - ANOVA jednoczynnikowa

Diagnostyka modelu:

```
> shapiro.test(lm(Dikeos.sum~NfkB*CD28)$residuals)
```

```
Shapiro-Wilk normality test
```

```
data:  lm(Dikeos.sum ~ NfkB * CD28)$residuals  
W = 0.9901, p-value = 0.6881
```



Dziękuję za uwagę

