

# Biomeetria

## 8. praktikum - Dispersioonanalüüs

Kasutame jälle andmestikku fishcatch.dat:

```
andmed=read.table("http://www.ms.ut.ee/mart/biomeetria2007/fishcatch.dat", header=TRUE)
```

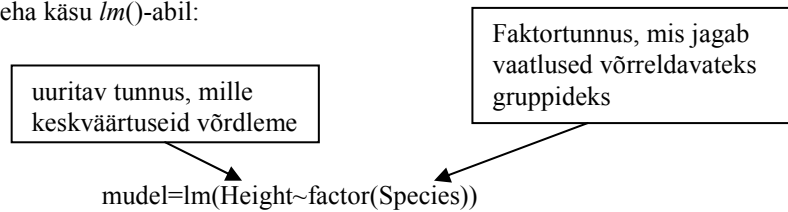
### Dispersioonanalüüs

Soovime teada, kas kõigi kalaliikide kõrguste keskvaärtused (tunnus *Height*, mõõdetud kui % kala pikkusest) on võrdsed või mitte, ehk tahame kontrollida järgmiseid hüpoteese:

$H_0: E \text{Height}_{\text{latikas}} = E \text{Height}_{\text{siig}} = E \text{Height}_{\text{särg}} = \dots = E \text{Height}_{\text{ahven}}$

$H_1$ : leiduvad vähemalt kaks kalaliiki, mille kõrguste keskvaärtused pole võrdsed.

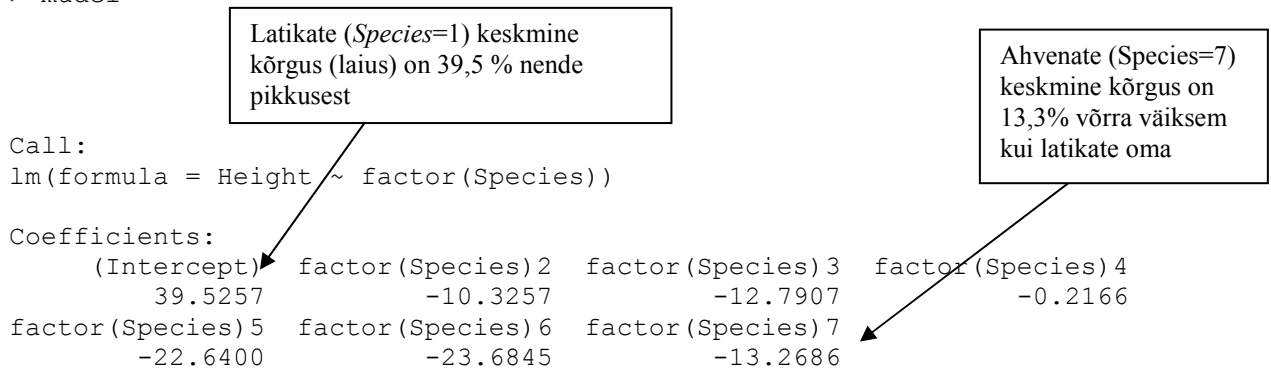
Dispersioonanalüüsi saab teha käsu *lm()*-abil:



Tulemuste vaatamine

1. Vaikimisi trükitakse välja kalaliikide keskmiste kõrguste erinevused latikate keskmisest kõrgusest:

```
> mudel
```



2. Näeme, et valimite keskmised erinevad teineteisest. Kas populatsioonide keskväärtused ka erinevad üksteisest?

```
> drop1(mudel, test="F")
Single term deletions
```

Model:

```
Height ~ factor(Species)
```

	Df	Sum of Sq	RSS	AIC	F value	Pr (F)
<none>			393.5	158.1		
factor(Species)	6	10494.1	10887.6	674.0	675.64	< 2.2e-16 ***

---

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

Olulisustõenäosus on väiksem kui 0,05, järelikult leidub vähemalt kaks kalaliiki, mille kõrguste keskväärtused pole võrdsed

3. Dispersioonanalüüsi mudelit võib uurida ka *summary()*-käsu abil, mille tulemusena trükitakse välja erinevused võrdlustasemega (latikate kaaluga):

```
> summary(mudel)
```

Call:

```
lm(formula = Height ~ factor(Species))
```

Residuals:

Min	1Q	Median	3Q	Max
-4.95714	-1.13500	0.01429	1.08260	4.97429

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	39.5257	0.2720	145.34	<2e-16 ***
factor(Species)2	-10.3257	0.7109	-14.52	<2e-16 ***
factor(Species)3	-12.7907	0.4510	-28.36	<2e-16 ***
factor(Species)4	-0.2166	0.5561	-0.39	0.697
factor(Species)5	-22.6400	0.5088	-44.50	<2e-16 ***
factor(Species)6	-23.6845	0.4756	-49.80	<2e-16 ***
factor(Species)7	-13.2686	0.3467	-38.27	<2e-16 ***

---

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

Residual standard error: 1.609 on 152 degrees of freedom

Multiple R-Squared: 0.9639, Adjusted R-squared: 0.9624

F-statistic: 675.6 on 6 and 152 DF, p-value: < 2.2e-16

Ahvenad on meie valimis keskmiselt latikatest 13,3% võrra väiksema kõrgusega

Näitab kui täpselt me oleme hinnanud latikate ja siigade erinevust – keskväärtuste erinevusele antud hinnangu standardhälve.

T-testi abil kontrollitakse, kas latikate ja ahvenate kõrguste keskväärtused on teineteisest erinevad. Antud juhul võib keskväärtuste erinevuse tõestatuks lugeda.

Võimalik on muuta ka võrdlustaset ehk gruppi, kellega teisi võrreldakse. Näiteks soovime võrrelda teiste kalade kõrguseid siigade (Species=2) kõrgustega:

```
> summary(lm(Height~C(factor(Species),base= 2)))
```

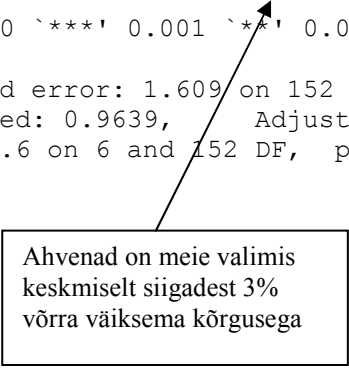
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	29.2000	0.6568	44.455	< 2e-16	***
C(factor(Species), base = 2)1	10.3257	0.7109	14.524	< 2e-16	***
C(factor(Species), base = 2)3	-2.4650	0.7489	-3.291	0.00124	**
C(factor(Species), base = 2)4	10.1091	0.8166	12.380	< 2e-16	***
C(factor(Species), base = 2)5	-12.3143	0.7851	-15.685	< 2e-16	***
C(factor(Species), base = 2)6	-13.3588	0.7640	-17.485	< 2e-16	***
C(factor(Species), base = 2)7	-2.9429	0.6911	-4.258	3.6e-05	***

---

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

Residual standard error: 1.609 on 152 degrees of freedom  
Multiple R-Squared: 0.9639, Adjusted R-squared: 0.9624  
F-statistic: 675.6 on 6 and 152 DF, p-value: < 2.2e-16



Ahvenad on meie valimis keskmiselt siigadest 3% võrra väiksema kõrgusega

Miks kasutan lisaparaameetrit *base= 2*? Kas sellepärast, et siigade puhul tunnus Species=2? Ei! Valik *base=2* näitab, et tahan kasutada faktori factor(Species) 2. taset, ükskõik kas selle teise taseme väärtuseks on siis kas „Tamm“ või „87“ või ... . Kuidas näha faktori tasemete järjestust? Üks võimalus faktorite tasemete järjestuse nägemiseks on kasutada *levels*-käsku:

```
> levels(factor(Sex))  
[1] "0" "1"
```

Seega, soovides võrdluse aluseks isaseid kalu (Sex=1) peame disersioonanalüüsi mudelit hindava programmis kasutama faktori Sex teist taset, *base=2*:

```
> lm(Weight~C(factor(Sex),base=2))
```

Call:

```
lm(formula = Weight ~ C(factor(Sex), base = 2))
```

Coefficients:

(Intercept)	C(factor(Sex), base = 2)1
341.6	120.2

## Eelduste kontroll

1. Kas dispersioonmudeli jäägid on normaaljaotusega?

Mudeli jääke (kalade kõrguste erinevust oma grupi ehk oma kalaliigi keskmisest) saab vaadata *resid()*-käsu abil:

```
resid(mudel)
```

Seda, kas tegemist võiks olla normaaljaotusega, võime näiteks kontrollida kas tõenäosuspaberi või Shapiro-Wilki testi abil:

```
qqnorm(resid(mudel))
qqline(resid(mudel))

ajut=resid(mudel)
shapiro.test(ajut)
```

2. Kas uuritava tunnuse hajuvus kõigis gruppides on ligikaudu samasuur?

Seda saab kontrollida näiteks vaadates, milline on uuritava tunnuse hajuvus kalaliigiti:

```
tapply(Height, Species, sd)
või
tapply(resid(mudel), Species, sd)
```

Mis on vahet nende käskude tulemustel, miks?

Uuritava tunnuse hajuvust grupiti ehk kalaliigiti saab võrrelda ka graafiliselt:

```
boxplot(resid(mudel)~Species)
```

On võimalik ka teostada statistilist testi, mis kontrollib, kas uuritava tunnuse hajuvus grupiti on sama:

```
> bartlett.test(resid(mudel)~factor(Species))
```

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

```
data: resid(mudel) by factor(Species)
Bartlett's K-squared = 11.9809, df = 6, p-value = 0.0624
```

Saame tulemuseks, et olulisustõenäosus on suurem 0,05-st (kuigi napilt) ja järelikult võime jääda oletuse juurde, et uuritava tunnuse hajuvus grupiti ei muutu.

NB! Bartlett test eeldab, et uuritav tunnus oleks normaaljaotusega!

### Dispersioonanalüüsi tabel ja arvutuslikud seosed

Kala kõrguse prognoosimisel tehtavate vigade ruutude summa, juhul kui me kala liiki ei tea ja prognoosimisel kasutada ei saa, on

```
> sum((Height-mean(Height))**2)
[1] 10887.56
```

Kui kasutame ka kala liiki kala kõrguse prognoosimisel, siis kahaneb prognoosivigade ruutude summa 393,5-ks:

```
> anova(mudel)
Analysis of Variance Table

Response: Height
          Df Sum Sq Mean Sq F value    Pr(>F)
factor(Species) 6 10494.1  1749.0   675.64 < 2.2e-16 ***
Residuals      152   393.5     2.6
```

Ehk, teisõõnu, oleme tänu liigi lisamisele prognoosivigade ruutude summat vähendanud 10887,56-393,5=10494,1 võrra. Jagatis 10494,1/10887,56=0,9638615 (prognoosi täpsuse suhteline paranemine) on sama mis determinatsioonikordaja.

## Loeme sisse *Escherichia coli* andmestiku

```
mydata=read.table("http://www.ms.ut.ee/BDA/m52orfs.txt", header=TRUE)
head(mydata)
attach(mydata)
```

Andmestiku lühikirjeldus on kättesaadav järgmiselt aadressilt:

<http://www.ms.ut.ee/BDA/datadesc.pdf>

Peamised meid hetkel huvitavad tunnused:

**first\_codon** – geeni esimene koodon

**cai** – Geeni Koodonikohastumusindeks CAI - see on üsna hästi korrelatsioonis antud valgu ekspresioonitasemega (valgu suhtelise esinemissagedusega) rakkudes

**orientation** – geeni suund kromosoomil

### Regression indikaatortunnustega = ANOVA

Proovime korra läbi, et indikaatortunnust kasutav regressioonanalüüs annab sama tulemuse, mis dispersioonanalüüs. Esmalt teeme siiski ühe teisenduse – mõistlik on kasutada logaritmitud cai väärtuseid, sestap transformeerime esmalt selle tunnuse:

```
lcai=log(cai)
```

Seejärel tekitame indikaatortunnuse

ind=0, kui orientation on "<";

ind=1, kui orientation on ">":

```
ind=1*(orientation==">")
```

Keskmine lcai on veidi erinev eripidi kirjutatud geennide jaoks:

```
by(lcai, orientation, mean)
```

Hindame regressioonmudeli kasutades indikaatortunnust:

```
m1=lm(lcai~ind)
```

```
summary(m1)
```

Kas suudad leida mõlema grupi keskmised kasutades vaid summary-käsu poolt antud informatsiooni?

Enamasti ei pea sa indikaatortunnust ise tekitama. Võrdle eelmise käsu poolt tagastatud tulemusi järgmise, dispersioonanalüüsi tulemustega:

```
summary(lm(lcai~factor(orientation)))
```

Antud juhul (2 gruppi) oleks sama vastuseni olnud võimalik jõuda ka t-testi abil:

```
t.test(lcai~orientation, var.equal=TRUE)
```

## Harjutus

Vaata järgmist väljundit (teil pole ligipääsu antud tunnustele/andmetele, seega piirduge praegu vaid esitatud analüüsiväljundiga:

```
> summary(lm(Expression~factor(Treatment)))
```

Call:

```
lm(formula = Expression ~ factor(Treatment))
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-282.241  -63.892   -4.897   64.105  341.040
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      5499.83     13.70  401.37  <2e-16 ***
factor(Treatment)B -1479.14     19.38  -76.33  <2e-16 ***
factor(Treatment)C  -490.43     19.38  -25.31  <2e-16 ***
---
```

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

Residual standard error: 96.89 on 147 degrees of freedom

Multiple R-Squared: 0.9763, Adjusted R-squared: 0.9759

F-statistic: 3023 on 2 and 147 DF, p-value: < 2.2e-16

Kasutades esitatud väljatrukki, vastake järgmisele küsimusele:

Milline on

keskmise ekspressioonitase grupi A jaoks: .....

keskmise ekspressioonitase grupi B jaoks: .....

keskmise ekspressioonitase grupi C jaoks: .....

## Ülesanne

Uuri, kas *Escherichia coli* puhul on keskmise ekspressioonitase erinev erinevate alguskoodonite puhul.

Kasuta selleks ühefaktorilist dispersioonanalüüsi:

```
model=lm(lcai~factor(first_codon))
```

```
summary(model)
```

Milline on sinu otsus? Proovi ka käsku:

```
drop1(model, test="F")
```

R kasutab vaikimisi alati esimest faktori taset võrdlustasemena:

```
> levels(factor(first_codon))
[1] "aat" "atg" "att" "ctg" "gtg" "ttg"
```

võrdluse aluseks kasutatav tase

Võrdluse aluseks valitavat taset (referentstaset) saab muuta. Valime alguskoodoni "atg" (kõige sagedamini esinev alguskoodon) võrdlustasemeks („atg“ –oli faktori first\_codon teine tase):

```
> model=lm(lcai~C(factor(first_codon),base=2))
> summary(model)
```

```
Call:
lm(formula = lcai ~ C(factor(first_codon), base = 2))
```

„atg“ ja „aat“ võrdlus

```
Residuals:
    Min       1Q   Median       3Q      Max
-1.33919 -0.18122 -0.01407  0.16331  1.02472
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.131818	0.004803	-235.651	< 2e-16 ***
C(factor(first_codon), base = 2)1	-0.017910	0.285967	-0.063	0.950
C(factor(first_codon), base = 2)3	0.134050	0.285967	0.469	0.639
C(factor(first_codon), base = 2)4	-0.505684	0.285967	-1.768	0.077 .
C(factor(first_codon), base = 2)5	-0.094878	0.012507	-7.586	4.03e-14 ***
C(factor(first_codon), base = 2)6	-0.122599	0.025533	-4.802	1.63e-06 ***

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

```
Residual standard error: 0.2859 on 4284 degrees of freedom
Multiple R-Squared: 0.01809,    Adjusted R-squared: 0.01694
F-statistic: 15.79 on 5 and 4284 DF,  p-value: 1.960e-15
```

Proovi, mis juhtub, kui kasutad võrdluse aluseks taset "gtg"!

Veel näiteid. Teeme uue grupeeriva tunnuse (näitab, kas geen kodeerib väga aluselist/ aluselist/ happelist väga happelist valku):

```
pIGroup=rep(NA, length(pI))
pIGroup[pI<5.6]="A"
pIGroup[pI>=5.6 & pI<7]="B"
pIGroup[pI>=7 & pI<9]="C"
pIGroup[pI>=9]="D"
```

```
table(pIGroup)
```

```
summary(lm(lcai~factor(pIGroup)))
```

Katsetame Tukey mitmese võrdluse meetodit (antud kujul eeldab ligikaudu samasuuruseid grupe):

```
TukeyHSD(aov(lm(lcai~factor(pIGroup))))
plot(TukeyHSD(aov(lm(lcai~factor(pIGroup)))))
```