Biomeetria

7(8). praktikum – dispersioonanalüüs (ANOVA)

Kasutame jälle andmestikku fishcatch.dat:

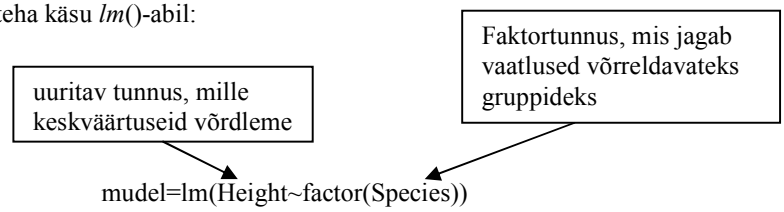
andmed=read.table("http://www.ms.ut.ee/mart/biomeetria2012/fishcatch.dat", header=TRUE)

Dispersioonanalüüs

Soovime teada, kas kõigi kalaliikide kõrguste keskväärtused (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 $\text{Height}_{\text{särg}}$ = ... = E $\text{Height}_{\text{ahven}}$
H_1: leiduvad vähemalt kaks kalaliiki, mille kõrguste keskväärtused pole võrdsed.

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



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

```
> mudel

Call:
lm(formula = Height ~ factor(Species))

Coefficients:
     (Intercept)  factor(Species)2  factor(Species)3  factor(Species)4
         39.5257          -10.3257          -12.7907           -0.2166
factor(Species)5  factor(Species)6  factor(Species)7
        -22.6400          -23.6845          -13.2686
```

Latikate (*Species*=1) keskmine kõrgus on 39,5 (mõõdetuna kui % nende pikkusest)

Ahvenate (Species=7) keskmine kõrgus on 13,3 võrra väiksem kui latikate oma

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

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

```
> 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
```

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

```
> 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 ahvenate 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~relevel(factor(Species),ref="2")))

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)                          29.2000     0.6568  44.455  < 2e-16 ***
relevel(factor(Species), ref = "2")1  10.3257     0.7109  14.524  < 2e-16 ***
relevel(factor(Species), ref = "2")3  -2.4650     0.7489  -3.291  0.00124 **
relevel(factor(Species), ref = "2")4  10.1091     0.8166  12.380  < 2e-16 ***
relevel(factor(Species), ref = "2")5 -12.3143     0.7851 -15.685  < 2e-16 ***
relevel(factor(Species), ref = "2")6 -13.3588     0.7640 -17.485  < 2e-16 ***
relevel(factor(Species), ref = "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 ühiku võrra väiksema kõrgusega

Veidi mugavamat (ja mõistlikumat) võimalust kõikki gruppe omavahel võrrelda pakub lisamoodulit multcomp.

Lisame lisamooduli oma arvutisse (käsku pole vaja anda, kui multcomp-lisamoodul on juba sinu arvutisse installeeritud):

```
install.packages("multcomp")
```

Võtame lisamooduli kasutusse:

```
library("multcomp")
```

Multcomp-lisamoodul võimaldab teha kõigi gruppide vahelisi võrduseid automaatselt.

Hindame mudeli:

```
m1=lm(Height~factor(Species))
```

Ütleme, et tahame mudelis m1 võrrelda kõiki tunnuse factor(Species) tasemeid omavahel:

```
a=glht(m1, linfct=mcp("factor(Species)"="Tukey"))
```

Lihtne tulemuste tabel: kas gruppide keskmised on statistiliselt oliselt erinevad:

```
summary(a)
```

Usaldusintervallid gruppide keskväärtuste erinevustele:

```
confint(a)
```

Joonistame usaldusintervallid graafikule:

```
plot(confint(a))
```

Tasub tähele panna, et saadud paarikaupa võrdluste p-väärtused (glht funktsiooni kasutades summary(a)-käsuga saadud) on mõnevõrra erinevad nendest, mida saame summary(m1) käsuga. Vaata näiteks, milline on 1. ja 4. kalaliigi keskväärtuste võrdsuse kontrollimisel saadud p-väärtus:

summary(m1)-käsu abil: .....

summary(a)-käsu abil: .....

Millest tuleneb erinevus?

Küsimus on mitmese testimise korrektsioonis.

Teeme praegu 21 erinevate kalaliikide keskväärtuste võrdlust. Kui kasutame olulisuse nivood 0,05, siis lubame (halvimal võimalikul juhul) iga testi puhul I-liiki vea tegemise tõenäosust 0,05. Kui grupid on üsna juhuslikult valitud, mingit erinevust uuritava tunnuse keskväärtuste vahel pole, siis ootaksime 21 testi peale keskmiselt 21*0,05=1,05 valesti vastu võetud alternatiivset hüpoteesi (st 21 testist kipub üks näitama statistiliselt olulist erinevust keskväärtuste vahel ka siis, kui tegelikult kõigi 7 kalaliigi keskväärtused oleksid täpselt samasugused). Käsud

```
a=glht(m1, linfct=mcp("factor(Species)"="Tukey"))
summary(a)
```

aga tagavad, et kõigi tehtavate võrdluste peale kokku (praegu teeme 21 erinevat keskväärtuste võrdlemist) ei tekiks I-liiki viga suurema tõenäosusega kui 0,05.

Kui soovime saada summary(m1) käsuga võrreldavaid tulemusi, siis saaksime neid paluda käsuga **(ÄRA KASUTA):**

```
summary(a, test=adjusted("none"))
```

Eelduste kontroll

1. Kas dispersioonmudeli jäägid on normaaljaotusega?

```
plot(mudel,2)
```

Normaaljaotuse eeldust saab kontrollida ka formaalse statistilise testi abil, näiteks Shapiro-Wilk'i testi abil. Shapiro-Wilk'i testi saab R'is kasutada shapiro.test-käsu abil:

```
shapiro.test(resid(mudel))
shapiro.test(stdres(mudel))
```

Märkus: enamik jääkide normaaljaotust kontrollivaid teste pole matemaatilises mõttes täiesti korrektsed, kuid annavad siiski enamasti õigele p-väärtusele väga lähedase tulemuse.

2. Kas uuritava tunnuse hajuvus kõigis gruppides (kõigi kalaliikide korral) on ligikaudu samasuur? Vaatame (standardiseeritud) jääkide hajuvust kalaliigiti:

```
tapply(stdres(mudel), Species, sd)
```

Kõigis gruppides (kõigi kalaliikide korral) peaks standardiseritud jääkide standardhälve olema ligikaudu 1. Väga suurte erinevuste korral (kui standardhälbe hinnangud erinevad teineteisest mitmeid kordi) peaksime kahtlema jääkide hajuvuse võrdsuse eelduses.

Uuritava tunnuse hajuvust grupiti ehk kalaliigiti saab võrrelda ka graafiliselt:

```
boxplot(stdres(mudel)~Species)
```

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

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

        Bartlett test of homogeneity of variances

data:  stdres(mudel) by factor(Species)
Bartlett's K-squared = 10.8114, df = 6, p-value = 0.09438
```

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!

Alternatiivina võib proovida ka Fligner-Killeeni testi hajuvuste võrdlemiseks. Fligner-Killeeni test on nn mitteprameetriline test – ta ei eelda, et uuritav tunnus on normaaljaotusega.

```
> fligner.test(stdres(mudel)~factor(Species))

        Fligner-Killeen test of homogeneity of variances

data:  stdres(mudel) by factor(Species)
Fligner-Killeen:med chi-squared = 13.5669, df = 6, p-value = 0.03487
```

Näeme, et Fligner-Killeeni test suudab vastu võtta alternatiivse hüpoteesi: uuritava tunnuse hajuvused on kalaliigiti erinevad.

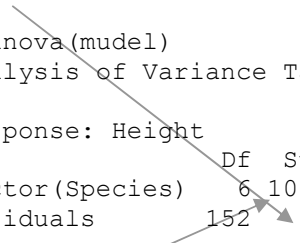
Antud olukorras kahtleksin prognoosiintervallides (võimalik, et mõne kalaliigi jaoks on nad liiga pisikesed, teise jaoks aga jälle liiga laiad). Küll aga teeksin näo, nagu antud eeldusega oleks kõik korras, kui vaatan näiteks p-väärtuseid: tavaliselt on p-väärtuste arvutustes märkimisväärse vea saavutamiseks vajalik väga mitmekordne dispersioonide erinevus (kui standardhälvete hinnangud erineksid kalaliigiti näiteks kümnekordselt, siis muretseksin ka leitud p-väärtuste pärast).

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
```

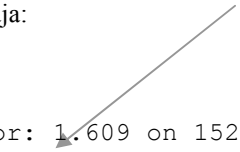
Kui kasutame ka kala liiki kala kõrguse prognoosimisel, siis kahaneb prognoosivigade ruutude summa 393,5-ks – võid seda väidet kontrollida ka käsuga *sum(resid(mudel)**2)*, aga anova-käsk annab ka vastuse:



```
> 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, teisisõnu, 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:



```
> summary(mudel)
[...]
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
```

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 ekspressioonitasemega (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 kirjapandud geenide 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)
```

Ülesanne
Uuri, kas *Escherichia coli* puhul on keskmine ekspressioonitase erinev erinevate alguskoodonite puhul. Kasuta selleks ühefaktorilist dispersioonanalüüsi:

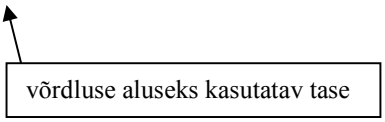
```
model=lm(lcai~factor(first_codon))
summary(model)
```

Milline on sinu otsus? Proovi ka käsku:

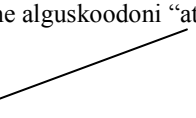
```
drop1(model, test="F")
```

R kasutab vaikimisi alati esimest faktori taset võrdlustasemena:

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



Võrdluse aluseks valitavat taset (referentstaset) saab muuta. Valime alguskoodoni "atg" (kõige sagedamini esinev alguskoodon) võrdlustasemeks:




```
> model=lm(lcai~relevel(factor(first_codon),ref="atg"))
> summary(model)

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 ***
relevel(factor(first_codon), ref = "atg")aat -0.017910   0.285967   -0.063    0.950
relevel(factor(first_codon), ref = "atg")att  0.134050   0.285967    0.469    0.639
relevel(factor(first_codon), ref = "atg")ctg -0.505684   0.285967   -1.768    0.077 .
relevel(factor(first_codon), ref = "atg")gtg -0.094878   0.012507   -7.586 4.03e-14 ***
relevel(factor(first_codon), ref = "atg")ttg -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"!

Muuseas, kui palju ekspressioonitaseme varieeruvusest on võimalik ära kirjeldada geeni alguskoodonit teades?

Tee kõigi alguskoodonite jaoks ekspressioonitasemete keskväärtuste võrdlus (täpsemalt, tunnuse lcai keskväärtuste võrdlus). Milliste alguskoodoni-paaride korral saame tõestada ekspressioonitasemete erinevust? Kasuta selleks funktsiooni glht!