Repetition samt introduktion til forsøgsplanlægning

Statistisk Dataanalyse 2

Anders Tolver

Uge 6, torsdag d. 12/10-2017

Program

Vi repeterer og uddyber forskellige punkter fra undervisningen i kursusugerne 1-5.

Desuden snuser vi lidt til emnet forsøgsplanlægning, som vil være centrum for undervisningen i kursusuge 7.

Vi kommer i hvert tilfælde ind på følgende emner

- Varianshomogenitet og sammenligning af grupper
- Transformation af respons ved varianshomogenitet
- Samme model - forskellige parametriseringer i `R` (Hydrolyse-eksempel)
- Faktordiagrammer og vekselvirkninger
- Styrkeberegninger og introduktion til forsøgsplanlægning

Eksempel: hjertevolumen for raske hunde

For 97 hunde har man målt volumen af venstre forkammer i hjertet (`maxLA`).

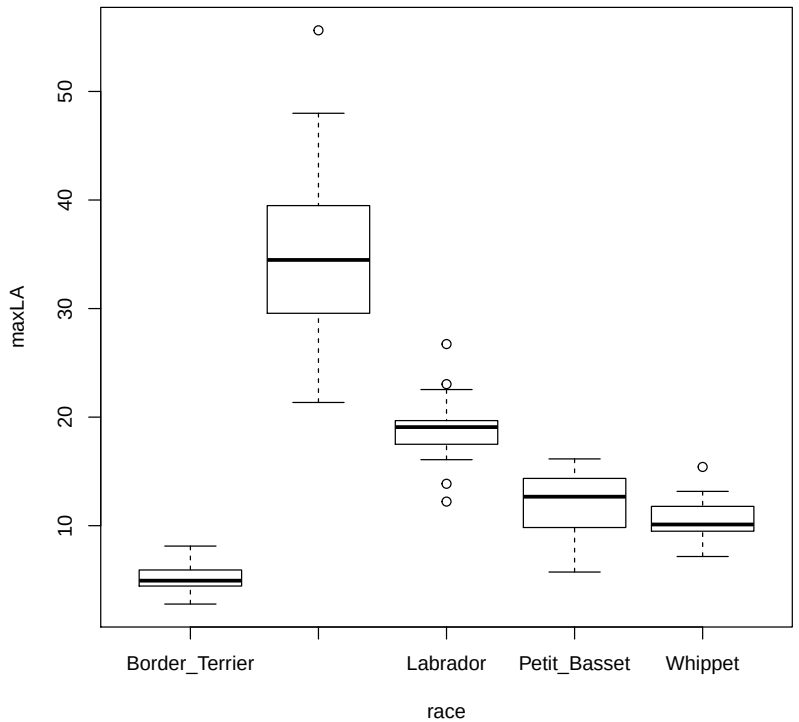
Hundene repræsenterer 5 forskellige racer (`race`).

Data er venligst stillet til rådighed af Miriam Höllmer.

```
##                race  wgt maxLA
## 1 Border_Terrier  9.2  6.07
## 2 Border_Terrier  6.9  4.67
## 3 Border_Terrier  7.7  4.40
## 4 Border_Terrier 11.0  4.48
## 5 Border_Terrier  6.3  2.78
```

- Vi ønsker at teste, om volumenet af venstre forkammer i hjertet (`maxLA`) afhænger af `race`.
- Mere præcist: hvilket racer kan antages at have samme størrelse af venstre forkammer?

Hjertevolumen for hunde - boxplot



- Tilsyneladende *IKKE* varianshomogenitet
- Resultater baseret på ensidet ANOVA ikke pålidelige!

Forskellige tests for varianshomogenitet

Bartlett's test assumes normally distributed residuals and compares variances across groups.

```
bartlett.test(maxLA~race,data=hunde)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  maxLA by race
## Bartlett's K-squared = 78.132, df = 4, p-value = 4.331e-16
```

Levene's test is more robust against departures from normal distribution.

```
library(car)
```

```
leveneTest(maxLA~race,data=hunde)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value    Pr(>F)
## group  4  11.735 9.607e-08 ***
##       92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tests som ikke kræver varianshomogenitet

Der findes situationer hvor...

- ... man ikke kan finde passende transformation som giver varianshomogenitet
- ... den nødvendige transformation vil gøre det svært at *regne* tilbage og fortolke resultaterne på oprindelig skala

Test i R som ikke kræver varianshomogenitet

```
oneway.test(maxLA~race,data=hunde,var.equal=FALSE)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  maxLA and race
## F = 136.62, num df = 4.000, denom df = 40.312, p-value < 2.2e
```

Modelkontrol: transformation

Modelkontrol i praksis (gode råd)

- Kig efter om størrelsen af de standardiserede residualer afhænger af prædikterede værdier og variable i den systematiske del af modellen
  [Ved eksamen: forklar, hvad du kigger efter!]

Er du i tvivl, om det ser pænt nok ud så

- se på residual plot hvor responsvariablen transformeres først [kan du se en synlig forbedring?]

Valg af transformation

- Søg evt. inspiration i box-cox plot
- Vælg helst pæn transformation $y^2, \sqrt{y}, \log(y)$
- Endnu bedre: kan transformation motiveres ud fra en forståelse af den respons der måles på

NB: Overvej samtidig om kovariater på 'højre side' i modellen også bør transformeres.

Hjertevolumen for hunde: model1 - fortolkning?

$$\log(\texttt{maxLA}_i) \quad = \quad \gamma(\texttt{race}_i) + e_i \quad \text{1-sidet ANOVA}$$

```
mod1 <- lm(log(maxLA) ~ race, data = hunde)
summary(mod1)
```

```
##
## Call:
## lm(formula = log(maxLA) ~ race, data = hunde)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71922 -0.09587  0.00793  0.13221  0.49884
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.59549    0.04675   34.13   <2e-16 ***
## raceGrand_Danois   1.94684    0.07033   27.68   <2e-16 ***
## raceLabrador       1.32955    0.07260   18.31   <2e-16 ***
## racePetit_Basset   0.86945    0.06934   12.54   <2e-16 ***
## raceWhippet        0.74124    0.07260   10.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.229 on 92 degrees of freedom
## Multiple R-squared:  0.9007, Adjusted R-squared:  0.8964
## F-statistic: 208.7 on 4 and 92 DF,  p-value: < 2.2e-16
```

Hjertevolumen for hunde: model2 - fortolkning?

$$\log(\texttt{maxLA}_i) = \alpha + \beta \cdot \log(\texttt{wgt}_i) + e_i \quad \text{log-log linear regression}$$

```
mod2 <- lm(log(maxLA) ~ log(wgt), data = hunde)
summary(mod2)
```

```
##
## Call:
## lm(formula = log(maxLA) ~ log(wgt), data = hunde)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55682 -0.13167 -0.00815  0.18163  0.44446
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.11931    0.09684  -1.232    0.221
## log(wgt)     0.89163    0.03172  28.107   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2344 on 95 degrees of freedom
## Multiple R-squared:  0.8927, Adjusted R-squared:  0.8915
## F-statistic:   790 on 1 and 95 DF,  p-value: < 2.2e-16
```

Afskårne roser (fra 12/9-2017): data

```
##     Obs gartner handler kunde flor  tid
## 1     1       0       0     0    1 10.1
## 2     2       0       0     0    2  8.9
## 3     3       0       0     0    3 11.4
## 4     4       1       0     0    1 10.9
## 5     5       1       0     0    2  6.9
## 6     6       1       0     0    3 11.2
## 7     7       0       1     0    1  9.6
## 8     8       0       1     0    2  9.9
## 9     9       0       1     0    3 11.3
## 10   10       0       0     1    1 11.8
```

[... more datalines here ...]

Afskårne roser (fra 12/9-2017): statistisk model

```
library(lme4)


## Loading required package: Matrix

model0 <- lmer(tid ~ handler*kunde + gartner * kunde + (1 | flor), data = roser)
model0


## Linear mixed model fit by REML ['lmerMod']
## Formula: tid ~ handler * kunde + gartner * kunde + (1 | flor)
##    Data: roser
## REML criterion at convergence: 66.3596
## Random effects:
##  Groups   Name        Std.Dev.
##  flor     (Intercept) 1.286
##  Residual             1.019
## Number of obs: 24, groups:  flor, 3
## Fixed Effects:
##     (Intercept)         handler1           kunde1         gartner1
##          9.8333           0.7333           1.3667           0.1333
## handler1:kunde1  kunde1:gartner1
##          2.1333          -0.5667
```

- Opskriv den statistiske model svarende til `model0`
- Hvad er den forventede holdbarhed hvis konserveringsmidlet tilsættes hos kunden og hos gartneren?

Eksempel 4.2: serin-indhold i foderprøver

```
data<-read.table(file="../data/hydrolysis.txt",header=T)
data$hourfac<-factor(data$hour)
data[1:12,]
```

```
##      feed hour serine hourfac
## 1  barley    8   4.47       8
## 2  barley   16   4.34      16
## 3  barley   24   4.22      24
## 4  barley   32   4.10      32
## 5  barley   72   3.48      72
## 6  barley    8   4.46       8
## 7  barley   16   4.30      16
## 8  barley   24   4.19      24
## 9  barley   32   4.08      32
## 10 barley   72   3.53      72
## 11   fish    8   4.23       8
## 12   fish   16   4.09      16
```

```
[  ... more data lines here  ...  ]
```

Forklar hvilke modeller der fittes her

Brug eventuelt hjælpe-appen via linket:
`http://shiny.science.ku.dk/AT/SD2/hydrolyse/`

```
mA<-lm(log(serine)~hour+feed,data)
mB<-lm(log(serine)~hourfac*feed,data)
mC<-lm(log(serine)~hour*feed,data)
mD<-lm(log(serine)~hourfac+feed,data)
mE<-lm(log(serine)~feed+feed:hourfac,data)
mF<-lm(log(serine)~hour+feed+hour:feed,data)
mG<-lm(log(serine)~feed+feed:hour,data)
mH<-lm(log(serine)~feed+feed:hour-1,data)
mI<-lm(log(serine)~feed+hourfac+feed:hourfac,data)
mJ<-lm(log(serine)~hour*feed-1,data)
mK<-lm(log(serine)~feed+hour-1,data)
mL<-lm(log(serine)~feed:hour,data)
mM<-lm(log(serine)~feed:hourfac,data)
mN<-lm(log(serine)~feed,data)
mO<-lm(log(serine)~hour,data)
mP<-lm(log(serine)~hour-1,data)
mQ<-lm(log(serine)~hourfac:feed-1,data)
mR<-lm(log(serine)~feed:hour-1,data)
mS<-lm(log(serine)~feed+feed:hourfac-1,data)
```

Faktorer og faktordiagrammer

Nogle væsentlige overvejelser er følgende ...

- Hvilke faktorer indgår i datasættet (=hovedeffekter)?
- Er der nogle faktorer, som er grovere / finere end hinanden?
- Hvilke vekselvirkninger skal med i modellen / faktordiagrammet?
  - Er der "gentagelser" for (nogle) kombinationer af de to faktorer?
  - Er produktfaktoren i virkeligheden identisk med andre faktorer, som allerede er medtaget (som hovedeffekter)?
- Hvilke faktorer skal indgå i modellen med tilfældig effekt?

Øvelse (fra slides d. 7/9-2017)

Der indgår ofte flere faktorer i et eksperiment...

Vækstforsøg med 16 planter i fire vækstkamre under forskellige gødnings- og lysforhold. Planternes vækst er målt.

	Lys 1				Lys 2			
	Kammer 1		Kammer 2		Kammer 3		Kammer 4	
Gødning	4.70	5.14	4.49	4.42	4.42	4.80	4.81	4.95
Ingen gødning	5.28	4.28	4.50	4.30	4.61	4.68	4.77	5.11

- Hvad er forsøgsenhederne og hvor mange er der?
- Hvad er de relevante faktorer, og hvad er deres niveauer?
- Er faktorerne balancerede? Angiv n_F for de balancerede faktorer.
- Tegn et faktordiagram og opskriv en statistisk model.

Introduktion til forsøgsplanlægning

I forbindelse med dit speciale skal du udføre et forsøg.

- 2 behandlinger skal sammenlignes: `kontrol` og `behandling`
- Det påtænkes at benytte n forsøgsenheder for hver behandling
- Det samlede ressourceforbrug ved udførelsen af forsøget afhænger af det totale antal forsøgsenheder $2n$

Spørgsmål: Hvor mange forsøgsenheder n skal der være i hver gruppe?

Svaret på spørgsmålet afhænger af

- vores ressourcer (maksimalt antal forsøgsenheder $2n$)
- hvor står en forskel mellem behandlingsgrupperne ønsker vi at kunne opdage?
- hvor sikre ønsker vi at være på at opdage/afsløre en eventuel forskel?
- hvor stor er variationen inden for forsøgsenheder som får den samme behandling?
- hvilken statistisk metode (test og signifikansniveau) skal benyttes ved analysen?

Introduktion til forsøgsplanlægning

Lad os lægge en statistisk model ned over forsøget

- kontrolgruppe: $X_1, \ldots, X_{n_1}$ uafh. $\sim N(\mu_X, \sigma^2)$
- behandlingsgruppe: $Y_1, \ldots, Y_{n_2}$ uafh. $\sim N(\mu_Y, \sigma^2)$
- hypotese, $H_0 : \mu_X = \mu_Y$
- teststørrelse,

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s \cdot \sqrt{1/n_1 + 1/n_2}},$$

  hvor $\hat{\mu}_X, \hat{\mu}_Y$ og s er estimater for μ_X, μ_Y og σ.

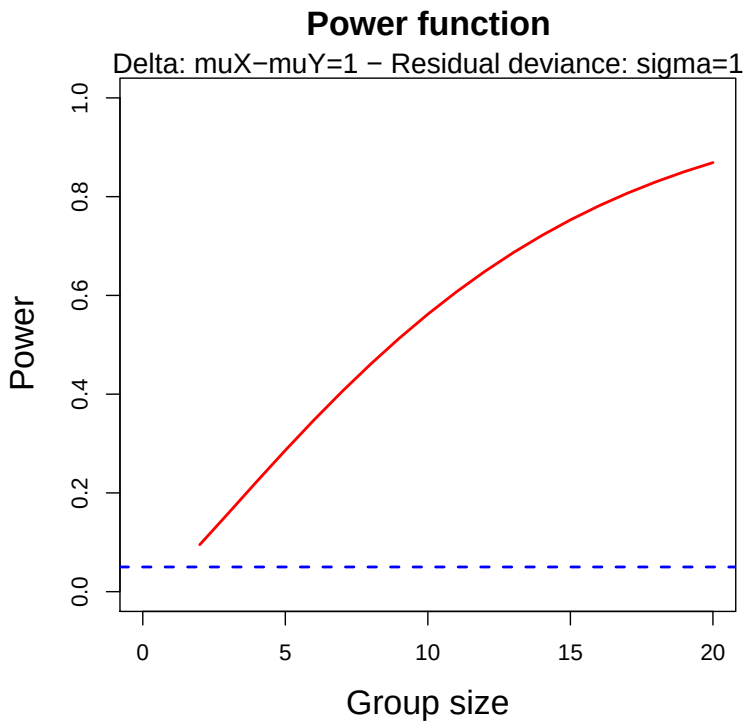
Udkendte størrelser

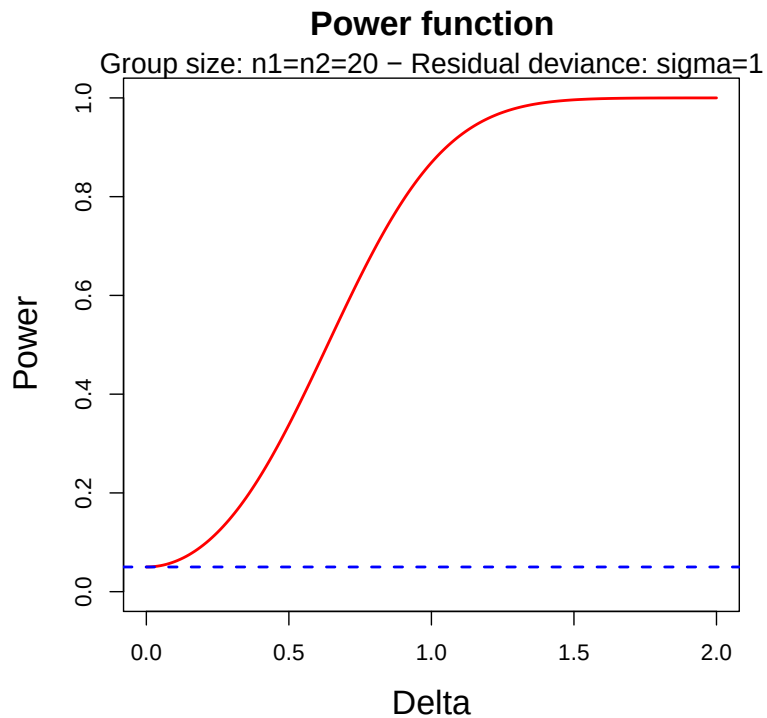
- størrelsen af behandlingseffekten, $\Delta = \mu_X - \mu_Y$
- residual spredningen, σ (inden for gruppe)
- signifikansniveau, α ($p < \alpha$ opfattes som signfikant)
- gruppestørrelse n
- sandsynlighed, β, for at opdage behandlingseffekten (styrken)

Hvis man angiver 4 af størrelserne kan den sidste beregnes.

Styrkefunktioner

Ofte optegnes styrken som funktion af en af de øvrige størrelser





- For $\sigma = 1$ vil der være over 80 % ssh. for at afsløre en forskel på $\mu_X - \mu_Y = 1$, med 17 personer i hver gruppe.
- For $\sigma = 1$ vil der med 20 personer i hver gruppe være 63.7 % ssh. for at afsløre en forskel på $\mu_X - \mu_Y = 0.75$.

Styrkeberegninger i R vha power.t.test

Ofte vælges $\alpha = 0.05$ og der ønskes en styrke på $\beta = 0.8$ eller 0.9.

$\Delta = \mu_X - \mu_Y$ bør afspejle den formodede effekt af behandlingen eller en minimal clinically relevant difference dvs en (nedre) grænse for, hvornår effekten vil være værd at skrive hjem om!

Det sværeste er tit at komme med bud på spredningen inden for grupper (σ), som bør begrundes ud fra pilotforsøg eller relevant faglitteratur.

```
power.t.test(power=0.8,delta=1,sd=1,sig.level=0.05)


##
##      Two-sample t test power calculation
##
##               n = 16.71477
##           delta = 1
##              sd = 1
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Randomisering

Når man har besluttet, hvilke forsøgsenheder som skal indgå i forsøget, så skal de to behandlinger allokeres til forsøgsenhederne.

Denne allokering bør foregå ved lodtrækning (randomisering) for at forhindre, at man systematisk (hvis eksperimentet gentages igen og igen!) tilordner 'de bedste' forsøgsenheder til den ene behandling.

I forsøg med 'behandlinger' (faktorer) som rent praktisk ikke kan randomiseres, fordi niveauerne er knyttet til forsøgsenhederne, handler randomisering om, at udvælge repræsentaterne i forsøget tilfældigt.

Tænk f.eks. på et forsøg, som skal sammenligne mænd og kvinder!