

# Matematisk Statistik

## PDF-version af web-bog

JENS LEDET JENSEN

Institut for Matematik  
2021



# Indhold

<b>Forord</b>	<b>v</b>
<b>1 Multinomialmodellen</b>	<b>1</b>
1.1 Multinomialmodellen . . . . .	2
1.1.1 Estimation i den fulde model . . . . .	3
1.2 Indledning til $G$ -test . . . . .	4
1.2.1 Eksempel på hypotese i multinomialmodellen . . . . .	7
1.3 $G$ -teststørrelsen . . . . .	7
1.4 Goodness of fit test . . . . .	11
1.5 Homogenitetstest: hypotesen . . . . .	14
1.6 Estimation og Test . . . . .	15
1.6.1 Test . . . . .	16
1.7 Test for uafhængighed . . . . .	18
1.8 Permutationstest og Fishers eksakte test . . . . .	21
1.8.1 Fishers eksakte test . . . . .	23
1.9 Diverse . . . . .	25
1.9.1 Notation for fordelingsfunktion og fraktil . . . . .	25
1.9.2 Egne funktioner i R . . . . .	26
1.10 Svar . . . . .	27
1.11 Opgaver til kapitel 1 . . . . .	27
<b>2 Normalfordelte data</b>	<b>37</b>
2.1 Normalfordelingen . . . . .	38
2.2 Normal-qqplot . . . . .	39
2.3 Model og estimation . . . . .	42
2.4 Test og konfidensinterval for middelværdi . . . . .	44
2.5 Kontrol af køkkenvægt . . . . .	46
2.6 Konfidensinterval for varians og spredning . . . . .	47
2.7 One sample $t$ -test i R . . . . .	48
2.8 Two-sample datasæt og boxplot . . . . .	50
2.9 Two-sample: Model og estimation . . . . .	52
2.9.1 Estimation i model $M_0$ . . . . .	52
2.9.2 Estimation i model $M_1$ . . . . .	53
2.10 Teste middelværdier ens når varianser er ens . . . . .	53
2.11 Teste middelværdier ens når varianser er forskellige . . . . .	55
2.12 Teste varianser ens . . . . .	59
2.13 Two sample tests i R . . . . .	61
2.13.1 Two samples: Teste varianser ens . . . . .	61
2.13.2 Two samples: Teste middelværdier ens . . . . .	61

2.13.3 Eksempel: log-data . . . . .	62
2.14 Standard Error . . . . .	63
2.15 Svar . . . . .	64
2.16 Opgaver til kapitel 2 . . . . .	66
<b>3 Lineær regression</b>	<b>71</b>
3.1 Model for lineær regression . . . . .	73
3.2 Estimation i den lineære regresionsmodel . . . . .	74
3.2.1 Modelkontrol . . . . .	77
3.3 Tests og konfidensintervaller i den lineære regressionsmodel . . . . .	79
3.4 Beregning i R via <i>lm</i> . . . . .	81
3.5 Linjens værdi og kalibrering . . . . .	83
3.5.1 Kalibrering (invers regression) . . . . .	86
3.6 Regression med kendt skæring . . . . .	88
3.6.1 Fordelingsresultater . . . . .	89
3.7 R-squared . . . . .	90
3.7.1 Relation til korrelation . . . . .	92
3.8 Svar . . . . .	92
3.9 Opgaver til kapitel 3 . . . . .	93
<b>4 En og tosidet variansanalyse</b>	<b>101</b>
4.1 Faktorer . . . . .	102
4.1.1 Produkt af faktorer . . . . .	103
4.2 Ensidet variansanalyse . . . . .	104
4.2.1 Estimation og fordelingsresultatet . . . . .	105
4.3 Teste middelværdierne ens . . . . .	106
4.4 Analyse i R . . . . .	108
4.4.1 Test af modelreduktion i R . . . . .	109
4.4.2 Analyse af data omkring metoder til håndvask . . . . .	109
4.5 Teste mere end to varianser ens . . . . .	111
4.6 Tosidet variansanalyse . . . . .	113
4.6.1 Analyse i R og parametrisering . . . . .	116
4.7 Det generelle F-test . . . . .	118
4.8 Estimation og t-test . . . . .	120
4.9 Svar . . . . .	121
4.10 Opgaver til kapitel 4 . . . . .	122
<b>5 Multipel regression</b>	<b>129</b>
5.1 Gruppespecifik regression . . . . .	131
5.2 Analyse af sediment transport . . . . .	133
5.3 Den multiple regressionsmodel . . . . .	136
5.3.1 Den multiple regressionsmodel . . . . .	138
5.4 Stepvis regression . . . . .	140
5.4.1 Eksempel . . . . .	142
5.5 Datasæt med et stort antal forklarende variable . . . . .	145
5.6 Cross Validation . . . . .	147
5.7 Beregning i R . . . . .	148
5.8 Prædiktion på nyt datasæt . . . . .	152
5.8.1 Estimation på fulde datasæt . . . . .	153

5.9 Svar . . . . .	155
5.10 Opgaver til kapitel 5 . . . . .	155
<b>6 Matematikken bag lineære modeller</b>	<b>161</b>
6.1 Den generelle lineære model via underrum . . . . .	163
6.2 Spaltningssætningen . . . . .	165
6.2.1 Likelihood ratio test . . . . .	167
6.3 T-test . . . . .	168
6.4 Opgaver til kapitel 6 . . . . .	170



# Forord

Her følger forord fra [webbogen](#) hørende til kurset *Matematisk Statistik*.

Denne webbog er beregnet til den anden halvdel af kurset *Matematisk Statistik*, hvor der fokuseres på modelbaseret inferens.

Et hovedelement i bogen er, at brugen af **R** er integreret direkte. Undervejs i læsningen kan man afprøve **R**-kommandoer uden at forlade bogen.

Et andet hovedelement er, at bogen er delt op i mange korte websider. Hver side har en navigationsbjælke i toppen, der tillader at komme rundt i hele bogen. For at øge overskueligheden er nogle elementer skjulte og foldes ud ved klik. Overskueligheden er også forsøgt øget ved brug af farvede elementer.

Bogen starter med et kapitel om multinomialfordelta data (tælledata). De resterende kapitler handler alle om forskellige modeller med normalfordelte data. Disse bliver sat ind i en generel ramme for lineære modeller via brugen af *faktorer* og *regressionsvariable*.

På kursushjemmesiden ligger der en pdf-version af bogen genereret fra web-bogen. De skjulte elementer i webbogen er foldet ud i pdf-versionen, og skjulte elementer med svar på spørgsmål er til sidst i hvert kapitel.



# Multinomialmodellen

Men ikke de fleste af os kender fornemmelsen af, at der aldrig kommer nogle seksere på terningen, når man spiller ludo. Man taler generelt om, at en terning er "ærlig", hvis de seks sider kommer op lige ofte. Hvis jeg skal teste, om min egen ludoterning er ærlig, vil jeg nok kaste den et stort antal gange og tælle op, hvor ofte side 1 kom op, hvor ofte side 2 og så videre. På denne måde får jeg seks observerede antal, der skal sammenlignes med seks ens forventede antal. Hvordan skal jeg foretage denne sammenligning?

Situationen med terningekspertementet er en generalisation af binomialmodellen. I binomialmodellen har hvert "kast" to mulige udfald. Vi siger, at kastet kan ramme ned i en af to "kasser". I terningesituationen kan man i hvert kast ramme ned i en af seks kasser. Modellen til at beskrive denne situation generelt hedder *multinomialmodellen* og beskrives i afsnit 1.1.

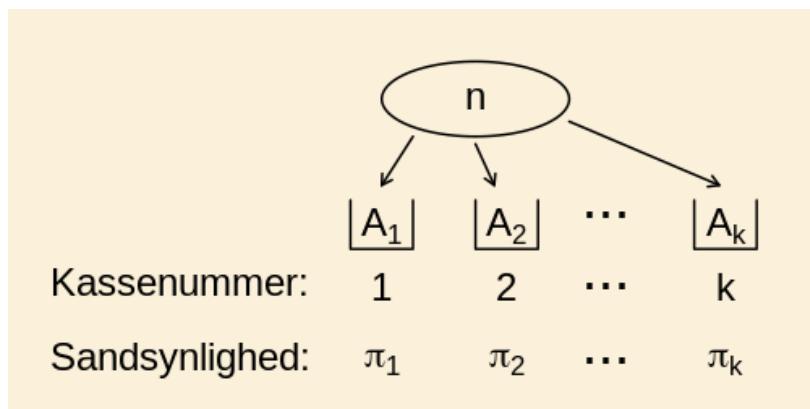
I eksemplet ovenfor med terningekast er hypotesen, at de seks sandsynligheder for de seks sider af terningen alle er  $\frac{1}{6}$ . Under hypotesen kender vi altså sandsynlighederne for at falde i de forskellige "kasser", og på denne måde ligner dette situationen med test for en given værdi af sandsynlighedsparameteren i en binomialmodel. Mere generelt kan en hypotese dog bestå i, at der lægges restriktioner på, hvordan sandsynlighederne kan variere, uden at angive sandsynlighederne med en numerisk værdi. Et eksempel på dette er, hvor en genetisk egenskab er bestemt af to [alleler](#). Hvis de to alleler kaldes  $a$  og  $A$ , har vi tre [genotyper](#)  $aa$ ,  $aA$  og  $AA$ . Hvis en population er i [Hardy-Weinberg ligevægt](#) er sandsynlighederne for de tre genotyper  $\theta^2$ ,  $2\theta(1-\theta)$  og  $(1-\theta)^2$ , hvor  $\theta$ ,  $0 \leq \theta \leq 1$ , er en ukendt parameter, der fortolkes som andelen af  $a$ -allelen i populationen. Jeg indfører i afsnittene 1.2 og 1.3 et generelt test (likelihood ratio test) for at håndtere denne type situation. Testet indføres først i binomialmodellen for at gøre beskrivelsen mere simpel.

Testet kan blandt andet bruges til at undersøge, om en række observationer  $x_1, \dots, x_n$  stammer fra en bestemt fordeling. Man taler i denne sammenhæng om et *goodness of fit test*. Dette beskrives og eksemplificeres i afsnit 1.4.

Kapitel 1 afsluttes med at se på situationen, hvor der er flere grupper af multinomialfordelte data, og man ønsker at sammenligne disse. For det biologiske eksempel ovenfor med en opdeling på tre genotyper  $aa$ ,  $aA$  og  $AA$ , kan vi have data fra forskellige lokationer og ønsker at se, om der er samme fordeling på de tre genotyper. Dette beskrives i afsnittene 1.5 og 1.6. Endelig omtales i afsnit 1.8 hvordan en  $p$ -værdi kan beregnes ved simulationer.

## 1.1 Multinomialmodellen

Jeg vil indføre multinomialmodellen ved først at vende tilbage til binomialmodellen. Hvis  $X \sim \text{binom}(n, p)$ , kan vi skrive  $X$  som  $X = B_1 + B_2 + \dots + B_n$ , hvor  $B_i$ 'erne er uafhængige og enten 0 eller 1 med sandsynlighederne  $1 - p$  og  $p$ . Dette kan billedeligt opfattes, som at data deles op i to kasser: alle  $B_i$ -erne med værdien 0 kommer i den ene kasse og alle med værdien 1 kommer i den anden kasse. I multinomialmodellen er der flere end to kasser, lad os sige  $k$  kasser, og vi kan tænke på modellen som en beskrivelse af  $n$  uafhængige kast med en generaliseret  $k$ -sidet terning. Hvert kast svarer til en stokastisk variabel,  $B_i$ ,  $i = 1, \dots, n$ , hvor de mulige værdier for  $B_i$ 'erne er  $1, 2, \dots, k$ . Jeg omtaler dette som at i hvert kast, kan man ramme ned i én ud af  $k$  kasser.



Den stokastiske variabel  $A_j$ ,  $j = 1, \dots, k$ , angiver, hvor mange af de  $n$  kast der lander i kasse  $j$ . Sandsynligheden i det enkelte kast for at lande i kasse  $j$  er  $\pi_j$ , hvor  $\pi_j \geq 0$  og  $\pi_1 + \dots + \pi_k = 1$ . Vektoren  $(A_1, \dots, A_k)$  af antallene i de  $k$  kasser siges at være *multinomialfordelt*:  $(A_1, \dots, A_k) \sim \text{multinom}(n, (\pi_1, \dots, \pi_k))$  med antalsværdi  $n$  og sandsynlighedsparameter  $(\pi_1, \dots, \pi_k)$ . For multinomialmodellen har vi følgende resultater:

Model:  $(A_1, \dots, A_k) \sim \text{multinom}(n, (\pi_1, \dots, \pi_k))$ ,  $\pi_j \geq 0$ ,  $\pi_1 + \dots + \pi_k = 1$ :

$$P((A_1, \dots, A_k) = (a_1, \dots, a_k)) = \binom{n}{a_1, \dots, a_k} \pi_1^{a_1} \cdots \pi_k^{a_k}, \quad a_j \geq 0, \quad a_1 + \dots + a_k = n,$$

$$A_j \sim \text{binom}(n, \pi_j), \quad E(A_j) = n\pi_j, \quad \text{Var}(A_j) = n\pi_j(1 - \pi_j).$$

Multinomialkoefficienten  $\binom{n}{a_1, \dots, a_k}$  er defineret som  $n!/(a_1! \cdots a_k!)$  og fortolkes, som antallet af måder man kan vælge  $b_1, \dots, b_n$ , således at  $a_1$  af disse har værdien 1,  $a_2$  har værdien 2 og så videre op til at  $a_k$  har værdien  $k$ . At  $A_j$  er binomialfordelt følger af, at vi kan reducere til, om det enkelte kast falder i kasse  $j$  eller ikke falder i kasse  $j$ .

### Showhide: Multinomialkoefficienten

Binomialkoefficienten  $\binom{n}{x} = n!/(x!(n-x)!)$  angiver, på hvor mange måder vi kan tage  $x$  ud af  $n$  elementer. Dette kan vises ved induktion. Hvis vi lader  $c_{n,x}$  være antallet af måder, vi kan tage  $x$  ud af  $n$  elementer, er det nemt at argumentere for, at  $c_{n+1,x} = c_{n,x-1} + c_{n,x}$ , idet man deler op efter,

om man blandt de  $n$  første har taget  $x$  eller  $x - 1$  elementer. Ved induktion kan man nu vise, at  $c_{n,x} = n!/(x!(n-x)!)$ .

Hvis vi nu i stedet betragter antallet af måder, hvorpå man kan dele  $n$  elementer op på  $k$  kasser med  $a_j$  i kasse  $j$ ,  $j = 1, \dots, k$ , kan man først vælge dem, der skal i kasse 1, dernæst dem der skal i kasse 2, og så videre. Dette giver at antallet af måder er

$$\begin{aligned} c_{n,a_1} c_{n-a_1,a_2} c_{n-a_1-a_2,a_3} \cdots c_{a_k,a_k} &= \frac{n!}{a_1!(n-a_1)!} \frac{(n-a_1)!}{a_2!(n-a_1-a_2)!} \frac{(n-a_1-a_2)!}{a_3!(n-a_1-a_2-a_3)!} \cdots 1 \\ &= \frac{n!}{a_1!a_2!\cdots a_k!}. \end{aligned}$$

Dette giver formlen for multinomialkoefficienten  $\binom{n}{a_1, a_2, \dots, a_k}$ . 

### Showhide: Multinomialfordelingen i R

I R kan man beregne sandsynlighederne i en multinomialfordeling med kommandoen `dmultinom((a1,...,ak),n,(π1,...,πk))`. Man kan simulere nye udfald som vist i følgende kodevindue.

#### Kodevindue

```
rmultinom(1,3,rep(1/6,6))
```

Her simuleres 1 udfald fra en multinomialfordeling, svarende til at en ærlig sekskantet terning kastes 3 gange. Kør koden, og bemærk at output skrives som en søjle. Prøv at ændre det første "1" til "4".

Prøv også at beregne sandsynligheden for hver af de tre udfald  $(a_1, \dots, a_6) = (1, 1, 1, 0, 0, 0)$ ,  $(a_1, \dots, a_6) = (1, 2, 0, 0, 0, 0)$  og  $(a_1, \dots, a_6) = (3, 0, 0, 0, 0, 0)$ , når en sædvanlig terning kastes 3 gange. Kan du på forhånd regne ud, hvilken af de tre sandsynligheder der er størst?

Kan du regne ud (dette er ikke et R-spørgsmål, men et tænke-spørgsmål), hvilken af følgende tre sandsynligheder der er størst: Sandsynligheden for at få tre forskellige tal når terning kastes 3 gange, sandsynligheden for kun at få to forskellige tal når terning kastes 3 gange, og endelig sandsynligheden for kun at få et tal når terning kastes 3 gange?

#### Svar 1.1. Multinomialsandsynligheder



## 1.1.1 Estimation i den fulde model

I multinomialmodellen  $(A_1, \dots, A_k) \sim \text{multinom}(n, (\pi_1, \dots, \pi_k))$  er likelihood-funktionen

$$L(\pi_1, \dots, \pi_k) = \binom{n}{a_1, \dots, a_k} \pi_1^{a_1} \cdots \pi_k^{a_k}, \quad (1.1)$$

og maksimum af denne funktion over området  $\pi_j \geq 0$ ,  $\pi_1 + \dots + \pi_k = 1$  (kaldet den *fulde model*), fås i punktet

$$\hat{\pi}_j = \frac{a_j}{n}, \quad j = 1, \dots, k.$$

Eftersom  $A_j \sim \text{binom}(n, \pi_j)$ , er dette helt i overensstemmelse med estimationen i binomialmodellen i Proposition 6.1.1 i MSRR. I ord estimeres sandsynligheden for at falde i kasse  $j$  med den observerede frekvens i kasse  $j$ .

### Showhide: Bevis for estimator

I ved fra binomialmodellen, at maksimum af  $p^x(1-p)^{n-x}$ ,  $0 \leq p \leq 1$  opnås for  $\hat{p} = x/n$  (MSRR side 152). Dette gælder også, hvis  $x = 0$  eller  $x = n$ . Når  $\pi_1^{a_1} \pi_2^{a_2} \dots \pi_k^{a_k}$  skal maksimeres over området  $\{\pi_j \geq 0, \pi_1 + \dots + \pi_k = 1\}$ , laver vi en omparametrering og skriver

$$\begin{aligned} \pi_1 &= p, \pi_2 = (1-p)v_2, \pi_3 = (1-p)v_3, \dots, \pi_k = (1-p)v_k, \\ (p, v_2, \dots, v_k) &\in [0, 1] \times \{v_j \geq 0, v_2 + \dots + v_k = 1\}. \end{aligned}$$

Med denne omparametrering opnås

$$\pi_1^{a_1} \pi_2^{a_2} \dots \pi_k^{a_k} = \left\{ p^{a_1} (1-p)^{n-a_1} \right\} \cdot \left\{ v_2^{a_2} \dots v_k^{a_k} \right\},$$

og maksimum findes ved at maksimere hvert led for sig. Det første led er som likelihoodfunktionen i binomialmodellen, og vi ved derfor, at

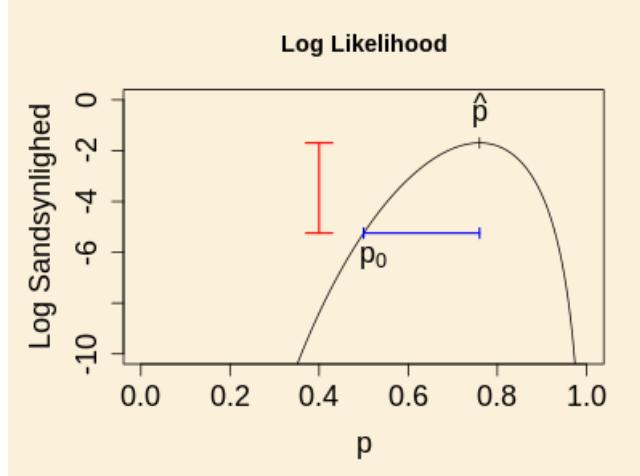
$$\hat{\pi}_1 = \hat{p} = \frac{a_1}{n}.$$

I ovenstående argument lavede vi omparametringen med udgangspunkt i  $\pi_1$ , men kunne have brugt et vilkårligt  $\pi_j$ ,  $j = 1, \dots, k$ , som udgangspunkt. Vi har derfor generelt, at  $\hat{\pi}_j = a_j/n$ ,  $j = 1, \dots, n$ .



## 1.2 Indledning til G-test

Hvis vi betragter  $X \sim \text{binom}(n, p)$ , er dette et specialtilfælde af multinomialmodellen, idet  $(X, n - X) \sim \text{multinom}(n, (p, 1-p))$ . I afsnit 6.1.1 i MSRR bliver likelihoodfunktionen  $L(p)$  brugt til at finde et skøn over  $p$ , idet vi bruger den værdi  $\hat{p}$ , der giver maksimum af likelihoodfunktionen. Dette er illustreret i følgende figur med logaritmen til likelihoodfunktionen baseret på observationen 19 fra en  $\text{binom}(25, p)$ -fordeling.



I afsnit 8.1.2 i MSRR bliver holdbarheden af hypotesen  $p = p_0$  vurderet ved at se på, hvor langt  $X$  ligger fra det forventede  $np_0$ , eller ækvivalent hermed, hvor langt  $\hat{p} = \frac{X}{n}$  ligger fra  $p_0$ . Dette svarer til afstand markeret med blåt på førsteaksen i ovenstående figur. Vi kan imidlertid også bruge likelihoodfunktionen til at konstruere et test af hypotesen  $p = p_0$ . Til dette betragtes forholdet  $Q = L(p_0)/L(\hat{p})$  (*likelihoodratio teststørrelsen*). Dette svarer til afstand markeret med rødt på andenaksen i figuren ovenfor med logaritmen til likelihoodfunktionen. Fordelen ved at bruge  $Q$  er, at denne metode nemt kan generaliseres til mere komplekse situationer, hvilket vi vil gøre i [næste afsnit](#) for test af hypotese i multinomialmodellen.

Per konstruktion ligger værdien af  $Q$  mellem 0 og 1, og små værdier er kritiske for hypotesen. En lille værdi betyder, at sandsynligheden for det observerede er meget mindre under  $p = p_0$  end under  $p = \hat{p}$ . Traditionelt transformerer man  $Q$  til  $G = -2\log(Q)$ , hvor det nu er store værdier, der er kritiske for hypotesen. Da  $\hat{p} = X/n$ , får man

$$Q = \frac{\binom{n}{X} p_0^X (1-p_0)^{n-X}}{\binom{n}{X} \left(\frac{X}{n}\right)^X \left(1 - \frac{X}{n}\right)^{n-X}} = \frac{1}{\left(\frac{X}{np_0}\right)^X \left(\frac{n-X}{n(1-p_0)}\right)^{n-X}},$$

og dermed

$$G = -2\log(Q) = 2 \left( X \log \left( \frac{X}{np_0} \right) + (n-X) \log \left( \frac{n-X}{n(1-p_0)} \right) \right).$$

Idet vi tænker på  $(X, n-X)$  som multinomialfordelt, er  $np_0$  og  $n(1-p_0)$  de forventede antal i de to kasser under hypotesen  $p = p_0$ . Ovenstående udtryk for  $G$  kan derfor læses som *2 gange summen over kasser af det observerede antal ganget med logaritmen til det observerede antal divideret med det forventede antal*. I [næste afsnit](#) genfinder vi dette udtryk mere generelt.

### Showhide: Inferens om fraktion

Betrægt binomialmodellen  $X \sim \text{binom}(n, p)$ , hvor vi ønsker at teste hypotesen  $p = p_0$  baseret på en observation  $x$ . Hvis alternativet er tosidet,  $p \neq p_0$ , angiver MSRR i afsnit 8.1.2  $p$ -værdien som

$$p\text{-værdi} = \begin{cases} 2P(X \leq x) & x \leq np_0, \\ 2P(X \geq x) & x > np_0. \end{cases}$$

Metoden med at gange med to skyldes, at man vil gøre beregningerne simple, men det betyder, at denne  $p$ -værdi ikke følger den generelle definition, hvor  $p$ -værdien er sandsynligheden for det, der er lige så kritisk eller mere kritisk end det observerede. Med **R** til rådighed kan vi sagtens

lave metoder, der følger definitionen. Jeg vil her nævne tre metoder. Med brug af R-notation for binomialsandsynligheder kan alle tre metoder skrives på formen

$$p\text{-værdi} = \sum_{z \in K(x)} \text{dbinom}(z, n, p_0),$$

hvor de tre metoder svarer til valgene

Afstand:  $K(x) = \{z : |z - np_0| \geq |x - np_0|\},$

Tæthed:  $K(x) = \{z : \text{dbinom}(z, n, p_0) \leq \text{dbinom}(x, n, p_0)\},$

Likelihoodratio:  $K(x) = \{z : Q(z) \leq Q(x)\}.$

Alle tre metoder er vist i kodevinduet nedenfor. Metode 2 er implementeret i **R** i funktionen *binom.test*. I kodevinduet har jeg kastet en terning 100 gange og fået en sekser 10 gange, og jeg tester, om dette er i overensstemmelse med en sandsynlighed på  $\frac{1}{6}$ . *P*-værdien ved metode 1 kan i dette eksempel beregnes på simpel vis som `pbinom(10, 100, 1/6)+1-pbinom(23, 100, 1/6)`. I opgaverne anbefaler jeg, at I bruger metode 1.

### Kodevindue

```
n=100
x=10
p0=1/6
relErr=1+1e-07
z=c(0:n)
a=abs(z-n*p0)
a0=a[x+1]
d=dbinom(z,n,p0)
d0=d[x+1]
Q=p0^z*(1-p0)^(n-z) / ((z/n)^z*(1-z/n)^(n-z))
Q0=Q[x+1]

pval1=sum(d[a>a0/relErr])
pval2=sum(d[d<d0*relErr])
pval3=sum(d[Q<Q0*relErr])

c(Afstand=pval1 , Taethed=pval2 , LikRat=pval3)
```

Som konfidensinterval for sandsynlighedsparameteren  $p$  i binomialmodellen bruger vi (7.15) og (7.16) i MSRR, som kan skrives kort som

$$\frac{x + \frac{u^2}{2} \pm u \sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}}{n + u^2},$$

hvor  $u = 1.96$  for at få et approksimativt 95%-konfidensinterval. Konfidensintervallet kan findes i **R** med kommandoen `prop.test(x, n)$conf.int`.



### 1.2.1 Eksempel på hypotese i multinomialmodellen

Data i nedenstående tabel viser for 100 kvinder, der alle er rygere og som alle prøver på at blive gravide, hvor mange menstruelle cykler der går, inden det lykkes at blive gravid. Der er 29 ud af de 100, der bliver gravide i første forsøg, 16 i andet forsøg, og så videre.

Cykelnummer	1	2	3	4	5	6	$\geq 7$
Antal kvinder	29	16	17	4	3	9	22

Det er naturligt at tænke på data i tabellen som et udfald fra en multinomialmodel,

$$(A_1, \dots, A_7) \sim \text{multinom}(100, (\pi_1, \dots, \pi_7)), \quad \pi_j \geq 0, \quad \pi_1 + \dots + \pi_7 = 1.$$

Hvis sandsynligheden for at blive gravid i et enkelt forsøg er  $\theta$  for alle kvinderne, er det relevant at betragte hypotesen

$$\pi_1 = \theta, \quad \pi_2 = (1 - \theta)\theta, \quad \pi_3 = (1 - \theta)^2\theta, \dots, \quad \pi_6 = (1 - \theta)^5\theta, \quad \pi_7 = (1 - \theta)^6.$$

For, som et eksempel, at blive gravid i det andet forsøg skal man ikke blive gravid i det første forsøg (sandsynlighed  $1 - \theta$ ) og blive gravid i det andet forsøg (sandsynlighed  $\theta$ ), hvorfor sandsynligheden er  $\pi_2 = (1 - \theta)\theta$ . Sandsynligheden for ikke at blive gravid i nogen af de 6 første forsøg er  $\pi_7 = (1 - \theta)^6$ .

Hypotesen beskrevet her, svarer til at sige, at antal forsøg indtil graviditet opnås er *geometrisk fordelt*. En stokastisk variabel  $X$  siges at være geometrisk fordelt med parameter  $\theta$ , hvis

$$P(X = x) = (1 - \theta)^{x-1}\theta, \quad x = 1, 2, 3, \dots$$

Data i dette eksempel stammer fra artiklen [The Beta-geometric distribution applied to comparative fecundability studies](#).

### 1.3 G-teststørrelsen

Jeg vil nu formulere en hypotese i multinomialfordelingen generelt. Udgangspunktet er modellen

$$M_0: (A_1, \dots, A_k) \sim \text{multinom}(n, (\pi_1, \dots, \pi_k)), \quad \pi_j \geq 0, \quad \pi_1 + \dots + \pi_k = 1. \quad (1.2)$$

Hypotesen lægger begrænsninger på variationsområdet for  $(\pi_1, \dots, \pi_k)$ , idet

$$\pi_j = p_j(\theta), \quad j = 1, \dots, k, \quad \theta \subseteq \Theta. \quad (1.3)$$

Her er  $p_j(\cdot)$  kendte funktioner,  $\theta$  er en ukendt parameter, der skal estimeres ud fra data, og  $\theta$  kan variere i området  $\Theta$ , som indeholder et åbent område af  $\mathbf{R}^d$ . Det sidste udtrykker vi sprogligt på

den måde, at  $\theta$  har  $d$  frie parametre. Under hypotesen betegnes den statistiske model med  $M_1$ , og man kan enten sige, at vi ønsker at teste hypotesen, eller at vi ønsker at teste reduktion fra model  $M_0$  til model  $M_1$ .

Som skøn over  $\theta$  bruges den værdi, der giver maksimum af likelihoodfunktionen  $L_1(\theta)$ :

$$L_1(\theta) = L(p_1(\theta), \dots, p_k(\theta)),$$

hvor  $L(\cdot)$  er givet i ligning (1.1).

Vi kan nu beregne *likelihoodratio teststørrelsen*  $Q$ , som er forholdet mellem den maksimale værdi af likelihoodfunktionen under model  $M_1$  og den maksimale værdi af likelihoodfunktionen under model  $M_0$ . Ved samme beregning som i [foregående afsnit](#) finder vi

$$Q = \frac{\binom{n}{A_1, \dots, A_k} p_1(\hat{\theta})^{A_1} \cdots p_k(\hat{\theta})^{A_k}}{\binom{n}{A_1, \dots, A_k} \left(\frac{A_1}{n}\right)^{A_1} \cdots \left(\frac{A_k}{n}\right)^{A_k}} = \frac{1}{\left(\frac{A_1}{np_1(\hat{\theta})}\right)^{A_1} \cdots \left(\frac{A_k}{np_k(\hat{\theta})}\right)^{A_k}},$$

og dermed

$$G = -2\log(Q) = 2 \sum_{j=1}^k A_j \log\left(\frac{A_j}{e_j}\right), \quad e_j = np_j(\hat{\theta}). \quad (1.4)$$

Her kaldes  $e_j$  det *forventede antal* i kasse  $j$  under hypotesen (under model  $M_1$ ).

En lille værdi af  $Q$  betyder, at data beskrives meget dårligere under model  $M_1$  end under model  $M_0$ . Jo mindre værdi af  $Q$  jo mere kritisk for hypotesen. Dette er det samme som, at jo større  $G$  er, jo mere kritisk.  $P$ -værdien for et test baseret på  $G$  er derfor sandsynligheden for ved gentagelse af eksperimentet at få en værdi af  $G$ , der er større end eller lig med den faktisk observerede værdi af  $G$ . Til beregning af  $p$ -værdien har vi følgende resultat.

### Resultat 1.1. (G-test)

Betrægt multinomialmodellen  $(A_1, \dots, A_k) \sim \text{multinom}(n, (\pi_1, \dots, \pi_k))$ ,  $\pi_j \geq 0$ ,  $\pi_1 + \cdots + \pi_k = 1$  (model  $M_0$ ) og hypotesen  $\pi_j = p_j(\theta)$ ,  $j = 1, \dots, k$ , hvor  $\theta$  har  $d$  frie parametre (model  $M_1$ ). Betragt teststørrelsen  $G = -2\log(Q) = 2 \sum_{j=1}^k A_j \log\left(\frac{A_j}{e_j}\right)$ ,  $e_j = np_j(\hat{\theta})$ , og lad  $G_{\text{obs}}$  være den observerede værdi af teststørrelsen. Hvis alle de forventede  $e_j = np_j(\hat{\theta})$  er større end eller lig med 5, har vi approksimativt

$$p\text{-værdi} = P(G \geq G_{\text{obs}}) = 1 - \chi_{\text{cdf}}^2(G_{\text{obs}}, k - 1 - d).$$

Beviset for dette resultat er ikke nemt. Intuitivt bygger det på den centrale grænseværdidisætning (se afsnit 4.3 i MSRR) og en andenordens taylorudvikling af likelihoodfunktionen. Antallet af frihedsgrader  $k - 1 - d$  i  $\chi^2$ -fordelingen er generelt  $d(M_0) - d(M_1)$ , hvor  $d(M_j)$  er antallet af frie parametre i model  $M_j$ . I model  $M_0$  har vi bindingen, at  $\pi_1 + \cdots + \pi_k = 1$ , hvorfor antallet af frie parametre er  $k - 1$ .

Sandsynligheden for at ligge til venstre for punktet  $z$  i en  $\chi^2$ -fordeling med  $f$  frihedsgrader,  $\chi_{\text{cdf}}^2(z, f)$ , beregnes i **R** med kommandoen `pchisq(z, f)`.

### Showhide: Likelihoodratio test

Her følger en generel definition på likelihoodratio teststørrelsen, når man vil teste en reduktion fra model  $M_1$  til model  $M_2$ . Vi betragter en statistisk model med likelihoodfunktion  $L(\theta; x)$ , hvor

$\theta$  er en parameter, og  $x$  er data. Under model  $M_1$  kan  $\theta$  variere i  $\Theta_1$  og under model  $M_2$  i  $\Theta_2 \subset \Theta_1$ . Så er likelihoodratio teststørrelsen  $Q$  givet ved

$$Q = \frac{\max_{\theta \in \Theta_2} L(\theta; x)}{\max_{\theta \in \Theta_1} L(\theta; x)}.$$

Da  $\Theta_2 \subset \Theta_1$ , er det klart, at  $Q \leq 1$ , hvorfor loglikelihoodratio teststørrelsen

$$G = -2 \log(Q) = 2 \left\{ \max_{\theta \in \Theta_1} \log(L(\theta; x)) - \max_{\theta \in \Theta_2} \log(L(\theta; x)) \right\}$$

er større end eller lig med 0, og små værdier af  $Q$  svarer til store værdier af  $G$ .

Hvis data stammer fra uafhængige og identisk fordelte stokastiske variable  $X_1, \dots, X_n$ , gælder der ofte, at fordelingen af  $G$  kan approksimeres med en  $\chi^2$ -fordeling, i grænsen hvor  $n$  går mod uendelig. Antallet af frihedsgrader i  $\chi^2$ -fordelingen er  $d_1 - d_2$ , hvor  $d_1$  og  $d_2$  er antallet af frie parametre i henholdsvis model  $M_1$  og model  $M_2$ . Hvis  $\Theta \subseteq R^d$ , og  $\theta$  indeholder en åben mængde, siger man, at  $\theta \in \Theta$  har  $d$  frie parametre.

I nogle situationer vil likelihoodratio testet (testet, hvor vi forkaster for store værdier af  $G$ ) være det "bedste" test, man kan lave. Dette skal forstås på den måde, at likelihoodratio testet har den største *styrke* blandt test med et niveau, der er mindre end eller lig med niveauet for likelihoodratio testet (Neyman-Pearson lemma side 273 i MSRR).



### Eksempel 1.2. (Tid indtil graviditet)

Vi vender tilbage til data omkring antal forsøg for at blive gravid i [foregående afsnit](#), og laver G-testet for hypotesen beskrevet der.

Først skal vi finde et skøn over parameteren  $\theta$ , hvor  $\pi_j = (1 - \theta)^{j-1}\theta$ ,  $j \leq 6$ , og  $\pi_7 = (1 - \theta)^6$ . Likelihoodfunktionen bliver

$$\begin{aligned} L_1(\theta) &= \binom{100}{29, \dots, 18} \theta^{29} ((1 - \theta)\theta)^{16} ((1 - \theta)^2\theta)^{17} ((1 - \theta)^3\theta)^4 ((1 - \theta)^4\theta)^3 ((1 - \theta)^5\theta)^9 ((1 - \theta)^6)^{22} \\ &= \binom{100}{29, \dots, 22} \theta^{78} (1 - \theta)^{251}. \end{aligned}$$

Ved sammenligning med likelihoodfunktionen i binomialmodellen (side 152 i MSRR) ses, at  $\hat{\theta} = 78/(78 + 251) = 0.2371$ .

Dernæst beregnes de forventede antal som  $e_j = 100 \cdot (1 - \hat{\theta})^{j-1}\hat{\theta}$ ,  $j \leq 6$ , og  $e_7 = 100 \cdot (1 - \hat{\theta})^6$ . Dette giver følgende tabel (forventede er afrundet til én decimal).

Cykelnummer	1	2	3	4	5	6	$\geq 7$
Antal kvinder	29	16	17	4	3	9	22
Forventede	23.7	18.1	13.8	10.5	8.0	6.1	19.7

Da alle de forventede er større end eller lig med 5, beregner vi G-teststørrelsen og den approksimative  $p$ -værdi fra en  $\chi^2$ -fordeling med  $7 - 1 - 1$  frihedsgrader. Ved beregning af antal frihedsgrader benyttes, at multinomialmodellen her deler op i 7 kasser, og den hypotese, der testes, har 1 parameter (nemlig  $\theta$ ). Beregningen i kodevinduet nedenfor giver  $G = 12.9$  og en  $p$ -værdi på 0.024. Da  $p$ -værdien er lille, er vi skeptiske over for holdbarheden af vores hypotese. En mulig forklaring på dette er, at hver kvinde har sin egen værdi af parameteren  $\theta$ , altså hver kvinde har sin egen

sandsynlighed for at blive gravid i et enkelt forsøg. I så fald vil data repræsentere en blanding, der ikke kan beskrives på samme måde som den enkelte kvinde.

### Showhide: Beregning i R

#### Kodevindue

```
a=c(29,16,17,4,3,9,22)
th=sum(a[1:6]) / (sum(a[1:6])+sum(c(1,2,3,4,5,6)*a[-1]))
n=sum(a)
ex=n*c(th , th*(1-th) , th*(1-th)^2 , th*(1-th)^3 , th*(1-th)^4 ,
th*(1-th)^5 ,(1-th)^6)
G=2*sum(a*log(a/ex))
c(Gteststørrelse=G, pværdi=1-pchisq(G,7-1-1))
```



### Showhide: $\chi^2$ -fordelingen i R

I det følgende kodevindue tegnes tætheden for en  $\chi^2(df)$  fordeling, og 95%-fraktilen markeres med en lodret streg. Fraktilen angiver punktet, hvor 95% af sandsynligheden i fordelingen ligger til venstre for punktet og 5% ligger til højre for punktet. Prøv at køre koden med forskellige valg af antallet af frihedegrader  $df$ . Prøv også i kodevinduet at beregne sandsynligheden for at ligge til højre for 5.99 i en  $\chi^2$ -fordeling med 2 frihedsgrader.

#### Kodevindue

```
df=2
mu=df
sigma=sqrt(2*df)
x=seq(max(c(0,mu-3*sigma)),(mu+3*sigma),length.out=1000)
plot(x,dchisq(x,df),type="l")
abline(v=qchisq(0.95,df),col=2)
```



## 1.4 Goodness of fit test

Vi skal nu bruge det generelle  $G$ -test i multinomialmodellen til at teste, at indsamlede data  $x_1, x_2, \dots, x_n$  følger en bestemt fordeling. Dette går under navnet *Goodness of fit test*.

Ideen er, at talaksen deles op i en række intervaller, lad os sige  $k$  intervaller,

$$(-\infty, z_1], (z_1, z_2], \dots, (z_{k-2}, z_{k-1}], (z_{k-1}, \infty),$$

hvorefter der tælles op, hvor mange af observationerne  $x_1, \dots, x_n$  der ligger i de forskellige intervaller

$$a_j = \{\text{antal } x_i\text{-er, der ligger i intervallet } (z_{j-1}, z_j]\}, \quad j = 1, \dots, k,$$

(her bruger vi  $z_0 = -\infty$  og  $z_k = \infty$ , og intervallet  $(a, b]$  går fra  $a$  til  $b$  med  $b$ , men ikke  $a$ , indeholdt i intervallet). Dette svarer til, at de  $n$  observationer er fordelt på  $k$  kasser, og de tilhørende stokastiske variable for antallene er derfor multinomialfordelt,

$$(A_1, \dots, A_k) \sim \text{multinom}(n, (\pi_1, \dots, \pi_k)).$$

På grund af den måde data er indsamlet på, kan vi skrive

$$\pi_j = P(z_{j-1} < X \leq z_j), \quad j = 1, \dots, k.$$

I modellen  $M_0$  er disse sandsynligheder vilkårlige:

$$M_0: \quad \pi_j \geq 0, \quad \sum_j \pi_j = 1.$$

Vi ønsker at teste, at  $X$  har en bestemt fordeling, der eventuelt afhænger af en parameter  $\theta$ , der kan variere i området  $\Theta$ . For at formulere dette betegnes fordelingsfunktionen (sandsynligheden for at ligge til venstre for et punkt) med  $F(x, \theta)$ . Vi kan nu formulere en ny model, eller specificere en hypotese, ved

$$M_1: \quad \pi_j = F(z_j, \theta) - F(z_{j-1}, \theta), \quad j = 1, \dots, k, \quad \theta \in \Theta.$$

Her skal  $F(-\infty, \theta)$  erstattes af 0, og  $F(\infty, \theta)$  skal erstattes af 1. Situationen her svarer til det generelle  $G$ -test med hypotesen  $p_j(\theta) = F(z_j, \theta) - F(z_{j-1}, \theta)$ , se ligning 1.3. Når skøn  $\hat{\theta}$  over  $\theta$  er fundet, bliver de forventede antal

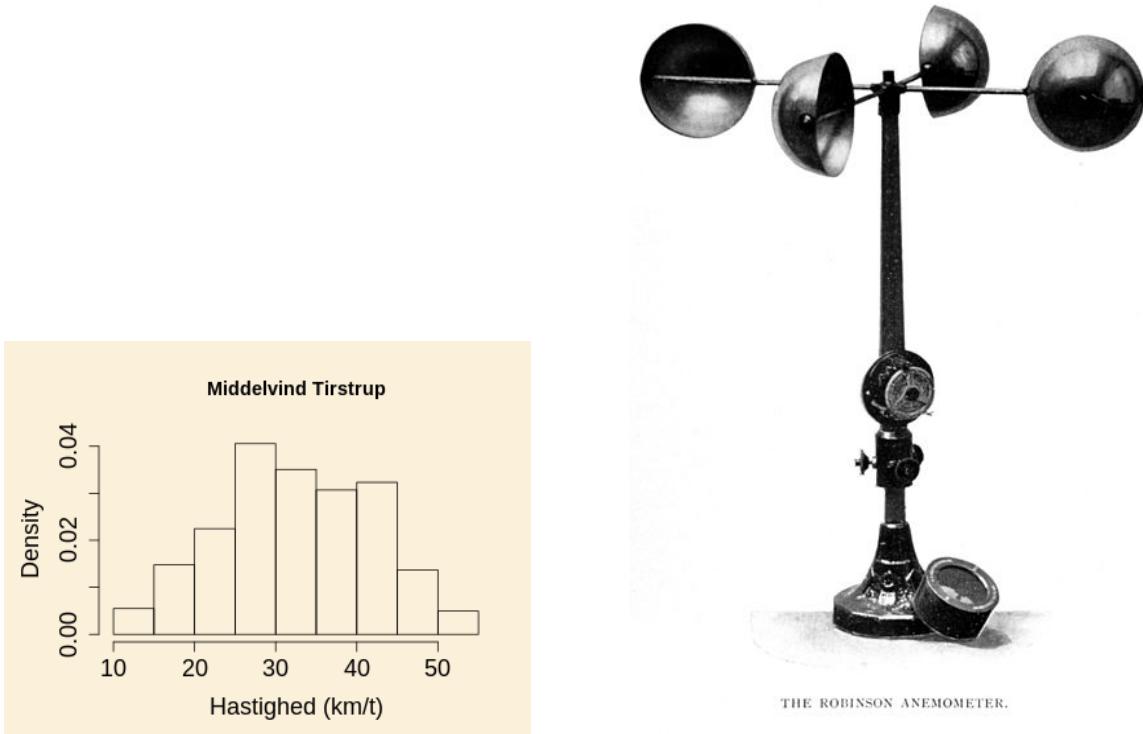
$$e_j = n(F(z_j, \hat{\theta}) - F(z_{j-1}, \hat{\theta})), \quad j = 1, \dots, k.$$

For at bruge Resultat 1.1 skal man have, at alle de forventede er større end eller lig med 5. Hvis dette ikke er opfyldt, gør man traditionelt det, at man slår kasser sammen for at få kravet opfyldt.

Typisk vil man, når man laver et goodness of fit test, lave en grafisk fremstilling af data i form af et histogram, og i dette histogram indtegne tæthedens for den fordeling, der undersøges.

### Showhide: Opgave med besvarelse: vindhastigheder

Data i denne opgave består af den daglige middelvind i Tistrup gennem hele 2019. Data er hentet hos [Iowa Environmental Mesonet](#), og de daglige middelvinde er givet i kilometer per time. Et tæthedshistogram er vist i nedenstående figur, og data er indskrevet i kodevinduet nedenfor.



Data af denne type beskrives ofte med [weibullfordelingen](#), og i opgaven her skal der laves et goodness of fit test for, om weibullfordelingen beskriver data. Hvis den stokastiske variabel  $X$  er weibullfordelt, gælder der

$$P(X > x) = e^{-(x/\lambda)^\alpha}, \quad x \geq 0,$$

hvor  $\alpha$  kaldes en formparameter og  $\lambda$  en skalaparameter. Tæthedsfunktionen og fordelingsfunktionen for en weibullfordeling beregnes i **R** med kommandoerne `dweibull(x, alpha, lambda)` og `pweibull(x, alpha, lambda)`. Til at lave goodness of fit testet skal der benyttes en intervalinddeling med intervaller af længde 3 startende i nul. Desuden skal der bruges, at maksimum likelihood skønnene baseret på antallene i de forskellige intervaller er  $\hat{\alpha} = 3.8851$  og  $\hat{\lambda} = 36.1116$ .

Idet den største værdi i data er 54, laver vi intervalinddelingen  $(0, 3], (3, 6], \dots, (48, 51], (51, \infty)$ . Antallene i de forskellige intervaller betegnes  $(a_1, \dots, a_{18})$  og findes i **R** med kommandoen `hist(vind, breaks=c(0:18)*3)$counts`. For de tilhørende stokastiske variable vælges modellen

$$(A_1, \dots, A_{18}) \sim \text{multinom}(365, (\pi_1, \dots, \pi_{18})), \quad \pi_j \geq 0, \quad \sum_j \pi_j = 1.$$

Vi ønsker at teste hypotesen

$$\begin{aligned} \pi_j &= F(3j, \alpha, \lambda) - F(3(j-1), \alpha, \lambda), \quad j = 1, \dots, 17, \\ \pi_{18} &= 1 - F(51, \alpha, \lambda), \quad \alpha, \lambda > 0, \end{aligned}$$

hvor  $F(x, \alpha, \lambda)$  er fordelingsfunktionen for en weibullfordeling. Fra opgaveformuleringen vides, at skønnene over de ukendte parametre er  $\hat{\alpha} = 3.8851$  og  $\hat{\lambda} = 36.1116$ . De forventede kan derfor beregnes som

$$\begin{aligned} e_j &= 365 \cdot (F(3j, 3.8851, 36.1116) - F(3(j-1), 3.8851, 36.1116)), \quad j = 1, \dots, 17, \\ e_{18} &= 1 - F(51, 3.8851, 36.1116). \end{aligned}$$

Fra **R**-beregningen får vi de observerede (første række) og forventede (anden række):

0	0	0	1	9	12	22	26	37	45	37	38	36	35	33	15	13	6
0.0	0.3	1.3	3.4	6.8	11.8	18.2	25.7	33.3	39.8	44.0	44.5	41.1	34.3	25.7	17.0	9.8	8.0

For at få alle de forventede større end eller lig med 5 slås de fire første kasser sammen. Dette giver det observerede antal 1 og det forventede antal 5.02. Efter denne sammenlægning er der 15 kasser, hvorfor antallet af frihedsgrader i  $\chi^2$ -fordelingen bliver  $15-1-2=12$ , idet vi under hypotesen har to frie parametre ( $\alpha$  og  $\lambda$ ). G-teststørrelsen for vores hypotese beregnes fra formlen  $G = 2 \sum_j \tilde{a}_j \log(\tilde{a}_j / \tilde{e}_j)$ , hvor  $\tilde{a}_j$  og  $\tilde{e}_j$  er de observerede og forventede, efter at kasser er slået sammen. Beregningen i R viser, at  $G = 13.72$ , og den tilhørende  $p$ -værdi er  $P(G \geq 13.72) = 1 - \chi^2_{\text{cdf}}(13.72, 12) = 0.32$ . Da  $p$ -værdien ligger langt over 0.05, strider data ikke mod hypotesen om, at de daglige middelvinde er weibullfordelt.

I R-kørslen nedenfor har jeg også indtegnet weibulltætheden i histogrammet. Desuden binder jeg de forskellige dele af output sammen ved at bruge R-kommandoen *list*.

### Showhide: Beregning i R

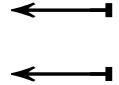
#### Kodevindue

```
vind=c(32,31,45,13,28,23,50,28,23,43,27,34,26,17,23,39,30,48,39,45,
38,39,53,40,42,47,33,31,29,35,31,37,54,37,34,23,53,28,33,23,
36,15,46,16,53,39,50,38,49,28,32,37,44,33,26,36,47,36,32,18,
41,34,47,29,31,44,35,24,27,37,31,37,40,34,30,20,20,35,44,33,
45,28,38,21,22,14,21,26,32,18,33,19,43,22,31,26,27,26,18,30,
28,20,15,26,19,18,30,23,22,19,27,23,25,18,37,30,35,33,39,30,
20,31,27,24,26,20,22,31,35,39,22,23,26,27,31,16,26,41,36,38,
41,27,41,30,15,27,39,28,28,45,43,35,44,45,38,42,30,41,30,22,
41,40,37,42,49,32,27,42,30,37,34,47,36,20,43,41,15,25,50,24,
30,23,19,18,14,30,25,16,43,45,21,33,30,38,39,46,39,42,37,34,
30,35,51,28,31,41,30,30,31,31,47,41,41,17,24,42,25,27,44,37,
52,46,30,32,31,40,34,41,45,38,44,39,30,43,31,27,32,37,25,36,
39,32,45,30,30,42,43,28,49,35,23,30,27,34,40,22,22,19,15,19,
21,50,25,30,26,23,39,43,39,36,51,42,28,36,40,34,35,30,22,21,
42,48,48,49,43,37,28,28,29,32,43,51,21,35,39,41,44,35,26,21,
32,28,30,41,41,45,26,12,37,36,27,32,27,35,44,44,42,36,19,25,
34,32,44,27,25,35,17,14,43,46,43,34,30,34,38,40,26,43,38,36,
30,20,24,26,31,35,22,48,28,52,46,33,34,40,33,49,47,40,50,41,
41,30,43,19,31)
```

```
Antal=hist(vind, breaks=c(0:18)*3, probability=TRUE)$counts
lines(c(0:55), dweibull(c(0:55), 3.8851, 36.1116), col=2)
```

```
br0=c(1:17)*3
pr=c(pweibull(br0,3.8851,36.1116),1)-c(0,pweibull(br0,3.8851,36.1116))
ex=365*pr
Antal1=c(sum(Antal[1:4]), Antal[5:18])
ex1=c(sum(ex[1:4]), ex[5:18])
G=2*sum(Antal1*log(Antal1/ex1))
```

```
pval=1-pchisq(G,15-1-2)
list(Observerede=Antal , forventede=ex , G_test=G, p_vaerdi=pval)
```



## 1.5 Homogenitetstest: hypotesen

Indtil nu har vi i dette kapitel udelukkende set på observationer fra én multinomialfordeling. Ofte vil man have observationer fra flere "populationer" og ønsker at sammenligne disse for at se, om der er samme forhold i populationerne. Her skal population forstås bredt. Det kan være biologiske populationer, men kan også være undersøgelser lavet på forskellige tidspunkter, eller for eksempel eksperimenter der gentages.

### Eksempel 1.3. (Aktivitet af delfingrupper)

Data i dette eksempel vedrører aktivitetsmønster for grupper af delfiner. Datasættet består af 72 delfingrupper observeret om morgen, og 79 delfingrupper observeret om aftenen tæt på Keflavik på Island.



Data er af Marianne Rasmussen (SDU) lagt op på [StatSci.org](http://StatSci.org). Delfingrupperne er klassificeret efter hovedaktivitet, som kan være enten *Rejse*, *Spise* eller *Leg*. Fordelingen på de tre aktivitetskategorier er som følger.

	Rejse	Spise	Leg	Total
Morgen	6	28	38	72
Aften	13	56	10	79

Biologerne ønsker at vurdere, om der er samme aktivitetsmønster om morgenen som om aftenen.

Jeg formulerer nu situationen generelt. Vi betragter  $r$  populationer, og i den  $i$ 'te er der i alt  $n_i$  observationer. For hver population kategoriseres data i  $k$  kasser, og antallene i disse kasser tælles:  $(A_{i1}, A_{i2}, \dots, A_{ik})$  er antallene i den  $i$ 'te population,  $A_{i1} + A_{i2} + \dots + A_{ik} = n_i$ . Som statistisk model benyttes

$$M_0: (A_{i1}, \dots, A_{ik}) \sim \text{multinom}(n_i, (\pi_{i1}, \dots, \pi_{ik})),$$

$$\pi_{ij} \geq 0, \quad \pi_{i1} + \dots + \pi_{ik} = 1, \quad i = 1, \dots, r,$$

og de  $r$  populationer er uafhængige.

Vi ønsker at teste hypotesen, at der er samme forhold i de  $r$  populationer. Med dette menes, at sandsynligheden for at falde i kasse  $j$  er den samme i de  $r$  populationer, og dette gælder for alle kasser  $j = 1, \dots, k$ . Dette kan skrives formelt som en ny model  $M_1$ , hvorunder der findes et sæt sandsynligheder  $\pi_1, \dots, \pi_k$ ,  $\pi_j \geq 0$ ,  $\pi_1 + \dots + \pi_k = 1$ , således at

$$M_1: \quad \pi_{ij} = \pi_j, \quad i = 1, \dots, r \quad \text{for alle } j = 1, \dots, k.$$

Hvis vi samler alle sandsynligheder i en  $r \times k$  matriks, kan hypotesen skrives på formen

$$\begin{array}{c|ccccc} & 1 & 2 & \cdots & k \\ \hline 1 & \pi_{11} & \pi_{12} & \cdots & \pi_{1k} \\ 2 & \pi_{21} & \pi_{22} & \cdots & \pi_{2k} \\ \vdots & \vdots & & & \vdots \\ r & \pi_{r1} & \pi_{r2} & \cdots & \pi_{rk} \end{array} = \begin{array}{c|ccccc} & 1 & 2 & \cdots & k \\ \hline 1 & \pi_1 & \pi_2 & \cdots & \pi_k \\ 2 & \pi_1 & \pi_2 & \cdots & \pi_k \\ \vdots & \vdots & & & \vdots \\ r & \pi_1 & \pi_2 & \cdots & \pi_k \end{array}$$

Hypotesen kaldes *homogenitetshypotesen* for kategoriske data.

## 1.6 Estimation og Test

Under model  $M_0$  kan estimation foretages for hver af de  $r$  multinomialmodeller under brug af resultatet i underafsnit 1.1.1. Dette giver

$$M_0: \quad \hat{\pi}_{ij} = \frac{A_{ij}}{n_i}, \quad j = 1, \dots, k, \quad i = 1, \dots, r.$$

Under model  $M_1$  skal det fælles sæt sandsynligheder  $(\pi_1, \dots, \pi_k)$  estimeres. Opstiller man likelihoodfunktionen, kan man indse, at estimaterne opnås ved at bruge

$$(A_{\bullet 1}, A_{\bullet 2}, \dots, A_{\bullet k}) \sim \text{multinom}(n_{\bullet}, (\pi_1, \dots, \pi_k)),$$

hvor  $A_{\bullet j} = A_{1j} + A_{2j} + \dots + A_{rj}$  er den  $j$ 'te søjlesum og  $n_{\bullet} = n_1 + n_2 + \dots + n_r$ . Igang kan vi bruge resultatet i underafsnit 1.1.1 og får

$$M_1: \quad \hat{\pi}_j = \frac{A_{\bullet j}}{n_{\bullet}}, \quad j = 1, \dots, k.$$

Vi kan nu beregne de forventede under model  $M_1$ . Idet  $e_{ij}$  er det forventede antal i kasse  $j$  for population  $i$ , er denne

$$e_{ij} = n_i \hat{\pi}_j = \frac{n_i A_{\bullet j}}{n_{\bullet}}, \quad j = 1, \dots, k, \quad i = 1, \dots, r.$$

Denne formel kan læses som "rækkesum gange søjlesum divideret med den totale sum".

### 1.6.1 Test

For at lave et test for reduktion fra model  $M_0$  til model  $M_1$  bruges igen likelihood ratio teststørrelsen på formen  $G = -2\log(Q)$ , hvor  $Q$  er forholdet mellem maksimum af likelihoodfunktionen under de to modeller:

$$Q = \frac{\max_{M_1} L}{\max_{M_0} L} = \prod_i \prod_j \frac{\hat{\pi}_j^{A_{ij}}}{\hat{\pi}_{ij}^{A_{ij}}} = \prod_i \prod_j \frac{1}{\left(\frac{A_{ij}}{(n_i A_{\bullet j})/n_{\bullet}}\right)^{A_{ij}}},$$

og dermed

$$G = 2 \sum_{i=1}^r \sum_{j=1}^k A_{ij} \log\left(\frac{A_{ij}}{e_{ij}}\right). \quad (1.5)$$

I ord kan vi sige dette, som at  $G$  er 2 gange sum over celler af det observerede antal ganget med logaritmen til det observerede antal divideret med det forventede antal. Med *celler* mener vi indgangene i  $r \times k$  matricen med antallene  $A_{ij}$ .

#### Resultat 1.4. (Homogenitetstest)

Betrægt modellerne  $M_0$  og  $M_1$  som beskrevet i dette afsnit. Hvis alle de forventede er større end eller lig med 5,  $e_{ij} \geq 5$ , kan vi approksimativt beregne  $p$ -værdien for test af reduktion fra model  $M_0$  til model  $M_1$  baseret på den observerede værdi  $G_{\text{obs}}$  af teststørrelsen  $G$  ved

$$p\text{-værdi} = P(G \geq G_{\text{obs}}) = 1 - \chi^2_{\text{cdf}}(G_{\text{obs}}, (r-1)(k-1)).$$

Antallet af frihedsgrader følger den generelle regel med antallet af frie parametre i  $M_0$  minus antallet af frie parametre i  $M_1$ :

$$(r(k-1)) - (k-1) = (r-1)(k-1).$$

#### Showhide: Beregning i R

I nedenstående kodevindue er *Obs* en  $3 \times 2$  matriks med følgende data:

	1	2	Sum
1	10	10	20
2	10	30	40
3	20	20	40
Sum	40	60	100

Kør koden, og forklar, hvad de forskellige dele af output indeholder.

### Kodevindue

```

Obs=rbind(c(10,10),c(10,30),c(20,20))
rs=rowSums(Obs)
cs=colSums(Obs)
rscs=outer(rowSums(Obs), colSums(Obs))
ex=rscs/sum(Obs)
G=2*sum(Obs*log(Obs/ex))
pval=1-pchisq(G,(3-1)*(2-1))
list(rs=rs, cs=cs, rscs=rscs, ex=ex, G=G, pval=pval)

```

### Svar 1.2. Homogenitetstest



### Eksempel 1.5. (Aktivitet af delfinggrupper)

Vi fortsætter med data omkring aktivitetsmønster for grupper af delfiner fra Eksempel 1.3.

Først opstilles en statistisk model for data. Lad  $Delf_{ij}$ ,  $i = M, A$  (Morgen,Aften),  $j = R, S, L$  (Rejse,Spise,Leg), være den stokastiske variabel, der angiver antal grupper i aktivitetskategori  $j$  til tidspunkt  $i$ . Vi benytter modellen

$$(Delf_{MR}, Delf_{MS}, Delf_{ML}) \sim \text{multinom}(72, (\pi_{MR}, \pi_{MS}, \pi_{ML})), \quad \pi_{Mj} \geq 0, \pi_{MR} + \pi_{MS} + \pi_{ML} = 1,$$

$$(Delf_{AR}, Delf_{AS}, Delf_{AL}) \sim \text{multinom}(79, (\pi_{AR}, \pi_{AS}, \pi_{AL})), \quad \pi_{Aj} \geq 0, \pi_{AR} + \pi_{AS} + \pi_{AL} = 1.$$

Under denne model ønsker vi at teste hypotesen om samme fordeling af aktivitet på de to tidspunkter,

$$(\pi_{MR}, \pi_{MS}, \pi_{ML}) = (\pi_{AR}, \pi_{AS}, \pi_{AL}).$$

Først findes de forventede antal under hypotesen som rækkesum gange søjlesum divideret med det totale antal. Dette giver følgende tabel (afrundet til to decimaler).

	Rejse	Spise	Leg	Total
Morgen	9.06	40.05	22.89	72
Aften	9.94	43.95	25.11	79

Dernæst beregnes  $G$ -teststørrelsen,

$$G = 2 \left\{ 6 \cdot \log \left( \frac{6}{9.06} \right) + \dots + 10 \cdot \log \left( \frac{10}{25.11} \right) \right\} = 29.25.$$

Da alle de forventede er større end fem (den mindste er 9.06), bruges  $\chi^2$ -approksimationen til fordelingen af  $G$ , og vi får

$$p\text{-værdi} = 1 - \chi^2_{\text{cdf}}(29.25, (2-1)(3-1)) = 4.5 \cdot 10^{-7}.$$

Denne  $p$ -værdi er meget lille, hvorfor data strider mod hypotesen om samme aktivitetsmønster på de to tidspunkter. Data tyder i grove træk på, at om morgenens spiser og leger grupperne, hvorimod grupperne er fokuseret på at spise om aftenen. Beregningerne er lavet i **R** som vist nedenfor.

### Showhide: Beregning i R

#### Kodevindue

```
Obs=rbind(c(6,28,38),c(13,56,10))
ex=outer(rowSums(Obs), colSums(Obs)) /sum(Obs)
G=2*sum(Obs*log(Obs/ex))
pval=1-pchisq(G, (dim(Obs)[1]-1)*(dim(Obs)[2]-1))
list(Forventede=ex, G=G, p_vaerdi=pval)
```



## 1.7 Test for uafhængighed

I artiklen [Upper Extremity Injuries in Homer's Iliad](#) har forfatterne læst Homers Illiade og registreret 146 personskader i kampene omkring Troja. Skaderne er delt op efter området på kroppen og om personen dør eller ikke dør. Her vil jeg nøjes med at se på skader, der vedrører arme og ben.

Sted	Død	Ikke Død
Hånd	0	4
Arm	3	6
Skulder	9	3
Ben	1	8

Er der uafhængighed mellem, hvor man får en skade, og om man overlever? Kunne de gamle grækere ved at bruge statistik have fået viden om, hvor de skulle forbedre kampuniformen?

I en generel formulering betragter jeg  $n$  individer, der inddeltes efter to kriterier. De oprindelige data skrives på formen  $(H_u, M_u)$ ,  $u = 1, \dots, n$ , hvor  $H_u$  er kategorien efter det første kriterie (med  $r$  katagorier) for observation nummer  $u$ , og  $M_u$  er kategorien efter det anden kriterie (med  $k$  katagorier). Fra disse data dannes en  $r \times k$  tabel med antal  $A_{ij}$ , hvor  $A_{ij}$  tæller op antallet af observationer, hvor  $H_u = i$  og  $M_u = j$ . Som startmodel bruger vi

$$M_{I0} : (A_{11}, A_{12}, \dots, A_{rk}) \sim \text{multinom}(n, (\pi_{11}, \pi_{12}, \dots, \pi_{rk})),$$

$$\pi_{ij} \geq 0, \quad \pi_{11} + \pi_{12} + \dots + \pi_{rk} = 1.$$

*Hypotesen om uafhængighed* siger, at sandsynligheden for at falde i den  $(i, j)$ 'te celle,  $\pi_{ij}$ , kan skrives som produktet af en sandsynlighed (kaldet  $\alpha_i$  nedenfor) for at falde i kasse  $i$  med hensyn til det første kriterie og en sandsynlighed (kaldet  $\beta_j$  nedenfor) for at falde i kasse  $j$  med hensyn til det andet kriterie. Dette giver modellen

$$M_{I1} : \pi_{ij} = \pi_{ij}(\alpha, \beta) = \alpha_i \beta_j, \quad i = 1, \dots, r, \quad j = 1, \dots, k, \quad (1.6)$$

$$\alpha_i \geq 0, \alpha_1 + \cdots + \alpha_r = 1, \beta_j \geq 0, \beta_1 + \cdots + \beta_k = 1.$$

Lad os starte med lidt notation. Vektoren med alle antallene  $A_{ij}$  kaldes  $A$ , og tilsvarende er  $\pi$  vektoren med alle indgangene  $\pi_{ij}$ . Den  $i$ 'te rækkesum er  $A_{i\bullet} = A_{i1} + \cdots + A_{ik}$ , og den  $j$ 'te søjlesum er  $A_{\bullet j} = A_{1j} + \cdots + A_{rj}$ , og vektorerne med disse summer betegnes  $A_{\star\bullet}$  og  $A_{\bullet\star}$ . Bemærk, at da rækkesummerne svarer til, at vi kun inddeler data efter det første kriterie, vil vektoren  $A_{\star\bullet}$  med disse summer være multinomialfordelt. Det samme gælder for søjlesummerne.

### Resultat 1.6. Uafhængighedstest

Et test for uafhængighedshypotesen  $M_{I1}$  foretages ved at beregne  $G$ -teststørrelsen fra (1.5) og beregne  $p$ -værdi som beskrevet i Resultat 1.4.

For at lave inferens om parameteren  $\alpha$  under uafhængighedsmodellen  $M_{I1}$  benyttes multinomialmodellen for rækkesummerne,  $A_{\star\bullet} \sim \text{multinom}(n, \alpha)$ , og for at lave inferens om  $\beta$  benyttes  $A_{\bullet\star} \sim \text{multinom}(n, \beta)$ .

Likelihoodfunktionen under model  $M_{I0}$  er  $L(\pi) = \binom{n}{A} \prod_{ij} \pi_{ij}^{A_{ij}}$ , og likelihoodfunktionen under uafhængighedshypotesen er

$$L(\pi(\alpha, \beta)) = \binom{n}{A} \prod_{ij} (\alpha_i \beta_j)^{A_{ij}} = \binom{n}{A} \left\{ \prod_i \alpha_i^{A_{i\bullet}} \right\} \left\{ \prod_j \beta_j^{A_{\bullet j}} \right\}.$$

De to led i krøllede parenteser har samme struktur som likelihoodfunktionen fra en multinomialmodel, hvorfor vi umiddelbart har

$$\hat{\alpha}_i = \frac{A_{i\bullet}}{n}, \quad i = 1, \dots, r, \quad \text{og} \quad \hat{\beta}_j = \frac{A_{\bullet j}}{n}, \quad j = 1, \dots, k.$$

Ved simpel indsættelse kan man nu se, at likelihood ratio tesstørrelsen  $Q$  vil være som i afsnit 1.6 for homogenitetstestet. I det næste skjulte punkt gives en dybere forklaring på, at de to test er ens.

### Showhide: Betingning

Vi starter med at lave en omparametrisering af model  $M_{I0}$ , idet vi skriver

$$\begin{aligned} \pi_{ij} &= \alpha_i \gamma_{ij}, \quad \alpha_i \geq 0, \quad \alpha_1 + \cdots + \alpha_r = 1, \\ \gamma_{ij} &\geq 0, \quad \gamma_{i1} + \cdots + \gamma_{ik} = 1, \quad i = 1, \dots, r, \end{aligned}$$

og alle parametrene kan variere uafhængigt af hinanden. Dette kan se lidt voldsomt ud, men her står blot, at  $\alpha_i$  er sandsynligheden for at falde i kategori  $i$  med hensyn til det første kriterie, og givet dette, er sandsynligheden for at falde i kategori  $j$  med hensyn til det andet kriterie givet som  $\gamma_{ij}$ . Hypotesen om uafhængighed,  $\pi_{ij} = \alpha_i \beta_j$ , bliver i denne formulering

$$(\gamma_{11}, \dots, \gamma_{1k}) = (\gamma_{21}, \dots, \gamma_{2k}) = \cdots = (\gamma_{r1}, \dots, \gamma_{rk}), \quad (1.7)$$

hvor den fælles værdi af disse sandsynlighedsvektorer svarer til  $(\beta_1, \dots, \beta_k)$ .

Den betingede sandsynlighed for hele tabellen  $A$ , givet rækkesummerne  $A_{\star\bullet}$ , er

$$\frac{P(A = a)}{P(A_{\star\bullet} = a_{\star\bullet})} = \frac{\binom{n}{a} \prod_{ij} (\alpha_i \gamma_{ij})^{a_{ij}}}{\binom{n}{a_{\star\bullet}} \prod_i \alpha_i^{a_{i\bullet}}}$$

$$= \prod_{i=1}^r \left\{ \binom{a_{i\bullet}}{a_{i1}, \dots, a_{ik}} \prod_j \gamma_{ij}^{a_{ij}} \right\}.$$

Her står to ting. For det første, at rækkerne er uafhængige givet rækkesummerne (på grund af produktstrukturen), og for det andet, at den  $i$ 'te række givet rækkesummen er multinomialfordelt med antalsværdi  $a_{i\bullet}$  og sandsynlighedsvektor  $(\gamma_{i1}, \dots, \gamma_{ik})$ . Sammenfattende har vi, at givet rækkesummerne svarer uafhængighedshypotesen (1.7) til hypotesen om homogenitet af  $r$  multinomialfordelinger fra afsnit 1.6.

Hvis man mere direkte skriver likelihood ratio teststørrelsen  $Q$  op for reduktion fra model  $M_{I0}$  til model  $M_{I1}$ , vil man se, at maksimum af ledet svarende til multinomialfordelingen for rækkesummerne forkorter ud, og tilbage er der produktet af multinomialfordelingerne for rækkerne givet rækkesummerne, og derfor bliver  $Q$  identisk med likelihood ratio tesstørrelsen udregnet i underafsnit 1.6.1.

På tilsvarende vis vil man finde, at likelihoodratio teststørrelsen for en hypotese om  $\alpha$ , under uafhængighedsmodellen  $M_{I1}$ , kun afhænger af rækkesummerne  $A_{\star\bullet}$ . Dette er baggrunden for, at der i Resultat 1.6 siges, at inferens om  $\alpha$  baseres på multinomialmodellen for rækkesummerne.

Under uafhængighedshypotesen har vi, at den  $i$ 'te række  $(A_{i1}, \dots, A_{ik})$  givet rækkesummerne  $A_{\star\bullet}$  er multinomialfordelt med antalsværdi  $A_{i\bullet}$  og sandsynlighedsvektor  $\beta$ . Da der er samme sandsynlighedsvektor for  $i = 1, \dots, r$ , bliver summen af rækkerne, som netop er vektoren  $A_{\star\bullet}$  af sjølesummer, multinomialfordelt med antalværdi  $n$  og sandsynlighedsvektor  $\beta$ . Dette er fordelingen givet rækkesummerne  $A_{\star\bullet}$ , men da denne betingede fordeling ikke afhænger af rækkesummerne, får vi udsagnet, at rækkesummer og sjølesummer er stokastisk uafhængige under uafhængighedshypotesen. Dette resultat vil jeg bruge i næste afsnit.



### Showhide: Chi-square test i stedet for G-test

Før G-testet, baseret på et likelihood ratio test, blev indført, benyttede man et andet test kaldet **chi-squared test**. Hvis vi kalder teststørrelsen for  $C$ , er de to teststørrelser

$$G = \sum \text{observeret} \cdot \log \left( \frac{\text{observeret}}{\text{forventet}} \right) \quad \text{og} \quad C = \sum \frac{(\text{observeret} - \text{forventet})^2}{\text{forventet}^2}.$$

$C$ -teststørrelsen vurderes i den samme  $\chi^2$ -fordeling som  $G$ -teststørrelsen og med samme krav om, at de forventede antal skal være større end eller lig med 5.

Chi-squared testet benyttes stadigt meget, men jeg foretrækker, at I bruger  $G$ -testet på grund af dets forbindelse til generelle metoder. I MSRR kapitel 10 benyttes næsten udelukkende  $C$ -testet, og dette er også standard i forskellige R-funktioner.



## 1.8 Permutationstest og Fishers eksakte test

Hvis vi prøver at lave uafhængighedstestet for data fra Homers Illiade, finder man, at  $G$ -teststørrelsen er 14.0, og de forventede antal er

Sted	Død	Ikke Død
Hånd	1.5	2.5
Arm	3.4	5.6
Skulder	4.6	7.4
Ben	3.4	5.6

Vi kan se her, at de forventede antal er ikke alle større end eller lig med 5, og vi kan ikke umiddelbart bruge Resultat 1.6. Ofte bruger man et mindre restriktivt krav, der kendes under betegnelsen **Cochran regel**. Denne regel siger, at alle de forventede skal være større end eller lig med 1, og højst 20 procent må være under 5. I eksemplet ovenfor med data fra Homers Illiade er der 5 ud af 8 forventede antal, der er under 1, hvilket er langt over grænsen på de 20 procent. Vi bør derfor ikke bruge  $\chi^2$ -fordelingen til at beregne en approksimativ  $p$ -værdi (hvis vi gør det alligevel, bliver  $p$ -værdien 0.0029).

Jeg vil nu beskrive en måde til at simulere en  $p$ -værdi for test af uafhængighedshypotesen  $M_{I1}$  fra (1.6). De underliggende data består af kategorierne  $(H_u, M_u)$  for  $n$  elementer  $u = 1, \dots, n$ , hvor  $H_u$  er kategorien efter det første kriterie, og  $M_u$  er kategorien efter det andet kriterie. Nedenfor viser jeg, at under uafhængighedshypotesen  $M_{I1}$  vil  $(H_1, M_1), \dots, (H_n, M_n)$  være uniformt fordelte i den betingede fordeling givet rækkesummerne  $A_{1\bullet}, \dots, A_{r\bullet}$  og søjlesummerne  $A_{\bullet 1}, \dots, A_{\bullet k}$ . Lad  $h_1^0, \dots, h_n^0$  være en fast følge, hvor antallet i de forskellige kategorier er  $a_{1\bullet}, \dots, a_{r\bullet}$ , og lad  $m_1^0, \dots, m_n^0$  være en fast følge, hvor antallet i de forskellige kategorier er  $a_{\bullet 1}, \dots, a_{\bullet k}$ . Så siger resultatet, at alle mulige ombytninger af  $h_1^0, \dots, h_n^0$  og alle mulige ombytninger af  $m_1^0, \dots, m_n^0$  er lige sandsynlige i den betingede fordeling givet rækkesummer og søjlesummer. En simuleret  $p$ -værdi kan nu findes ved for hver simuleret ombytning at finde antallene  $A_{ij}$  i de  $r \cdot k$  celler af tabellen, beregne  $G$ -teststørrelsen og se, hvor ofte vi får en værdi, der ligger over  $G$ -teststørrelsen for de oprindelige data,  $G_{obs}$ .

### Resultat 1.7. Permutationstest for uafhængighedshypotesen

Lad  $V = (V_1, \dots, V_n)$  være en tilfældig permutation af tallene  $1, \dots, n$ , og lad  $h_i^V = h_{V_i}^0$  være de ombyttede  $h^0$ -værdier. Dan ud fra data  $(h_1^V, m_1^0), \dots, (h_n^V, m_n^0)$  en tabel  $A^V$  med antallene i de  $r \cdot k$  celler, og dan ud fra denne tabel  $G$ -teststørrelsen betegnet med  $G^V$ . Så gælder, at

$$P(G \geq G_{obs} | (A_{\bullet\bullet}, A_{\bullet\bullet})) = P(G^V \geq G_{obs}).$$

### Show/hide: Beregning i R

En tilfældig ombytning (permutation) af elementerne i en vektor  $v$  kan i R simuleres ved kommandoen `sample(v)`. I det følgende kodevindue simuleres  $p$ -værdien for uafhængighedstestet for data fra Homers Illiade.

#### Kodevindue

```
Gfct=function(Amat){
```

```

ex=outer(rowSums(Amat), colSums(Amat)) /sum(Amat)
A1=ifelse(Amat==0, 1, Amat)
return(2*sum(Amat*log(A1/ex)))
}
h=rep(c(1,2,3,4), c(4,9,12,9))
m=rep(c(1,2), c(13,21))
Gobs=Gfct(rbind(c(0,4), c(3,6), c(9,3), c(1,8)))
N=10^4-1
res=numeric(N)
for (i in 1:N){
hperm=sample(h)
Ai=table(hperm,m)
res [i]=Gfct(Ai)
}
(sum(res>=Gobs)+1) / (N+1)

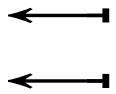
```

Her følger en række spørgsmål til forståelsen af koden.

1. Hvad beregnes i  $Gfct$ ?
2. Hvad er formålet med  $A1$  inde i  $Gfct$ ?
3. Hvad tror du funktionen *table* laver?

#### Showhide: Svar

1.  $Gfct$  beregner  $G$ -tesstørrelsen baseret på en matriks  $A$  med antallene i de forskellige kategorier.
2. I  $G$ -teststørrelsen vil vi få et problem, hvis det observerede antal i et af leddene er nul, idet  $\mathbf{R}$  ikke kan beregne  $0 \cdot \log(0)$ . For at få den rigtige værdi, nemlig nul, skriver vi i stedet  $0 \cdot \log(1)$ .
3. Funktionen *table* laver matricen med de observerede antal i kategorierne  $(i, j)$ ,  $i = 1, \dots, r$  og  $j = 1, \dots, k$ .



#### Showhide: Betingede sandsynlighed

Under uafhængighedshypotesen betragtes den betingede fordeling af  $(H_1, M_1), \dots, (H_n, M_n)$  givet rækkesummerne  $(A_{1\bullet}, \dots, A_{r\bullet})$  og søjlesummerne  $(A_{\bullet 1}, \dots, A_{\bullet k})$ . For at finde den betingede fordeling bruges at

$$P((H_1, M_1) = (h_1, m_1), \dots, (H_n, M_n) = (h_n, m_n)) = \prod_u \alpha_{h_u} \beta_{m_u} = \left\{ \prod_i \alpha_i^{a_{i\bullet}} \right\} \left\{ \prod_j \beta_j^{a_{\bullet j}} \right\}.$$

Idet rækkesummer og søjlesummer er uafhængige (se skjulte punkt i [foregående afsnit](#)), kan den betingede sandsynlighed nu skrives som

$$\frac{\prod_u \alpha_{h_u} \beta_{m_u}}{\left\{ \binom{n}{a_{**}} \prod_i \alpha_i^{a_{i*}} \right\} \left\{ \binom{n}{a_{**}} \prod_j \beta_j^{a_{*j}} \right\}} = \frac{1}{\binom{n}{a_{**}} \binom{n}{a_{**}}}.$$

Da denne sandsynlighed ikke afhænger af  $(h_1, m_1), \dots, (h_n, m_n)$ , har alle værdier af disse, der opfylder at rækkesummerne er  $a_{1*}, \dots, a_{r*}$  og søjlesummerne er  $a_{*1}, \dots, a_{*k}$ , lige stor sandsynlighed.



### 1.8.1 Fishers eksakte test

Den betingede fordeling, der simuleres i permutationstestet ovenfor, kan beregnes direkte i en  $2 \times 2$  tabel. Når vi betinger med de to rækkesummer og de to søjlesummer, er der kun én indgang tilbage i  $2 \times 2$  matricen  $A$ , der kan variere frit. Lad  $A_{1*} = b$ ,  $A_{2*} = c$ ,  $A_{*1} = d$  og  $A_{*2} = e$ , så kan vi skrive  $A$  som

$$\begin{array}{cc|c} x & b-x & b \\ d-x & x+c-d & c \\ \hline d & e & n \end{array} \quad \max\{0, d-c\} \leq x \leq \min\{b, d\}.$$

Den betingede sandsynlighed  $h(x, b, c, d)$  kan beregnes som

$$\begin{aligned} h(x, b, c, d) &= \frac{P(A_{11} = x, A_{12} = b-x, A_{21} = d-x, A_{22} = x+c-d)}{P(A_{1*} = b, A_{2*} = c) P(A_{*1} = d, A_{*2} = b+c-d)} \\ &= \frac{\binom{n}{x, b-x, d-x, x+c-d} (\alpha_1 \beta_1)^x (\alpha_1 \beta_2)^{b-x} (\alpha_2 \beta_1)^{d-x} (\alpha_2 \beta_2)^{x+c-d}}{\binom{n}{b} \binom{n}{d} \alpha_1^b \alpha_2^c \beta_1^d \beta_2^{b+c-d}} \\ &= \frac{n! b! c!}{x! (b-x)! (d-x)! (x+c-d)! \binom{n}{d}} = \frac{\binom{b}{x} \binom{c}{d-x}}{\binom{b+c}{d}}. \end{aligned}$$

Fordelingen, der optræder her, kendes under navnet den [hypergeometriske fordeling](#). For enhver mulig værdi af  $x$  beregner vi  $G$ -teststørrelsen,  $G(x)$ , og beregner  $p$ -værdien for uafhænghedstestet som

$$p\text{-værdi} = \sum_{x: G(x) \geq G_{\text{obs}}} h(x, b, c, d). \quad (1.8)$$

Testet baseret på den betingede fordeling kaldes [Fishers eksakte test](#). I R er dette test implementeret i funktionen `fisher.test`, men her beregnes  $p$ -værdien lidt anderledes end i (1.8). Når vi bruger (1.8), rangordnes de forskellige udfald  $x$ , det vil sige  $(1, 1)$ -indgangen i  $A$  matricen, ud fra  $G$ -teststørrelsen. I `fisher.test` rangordnes udfaldene efter værdien af den betingede sandsynlighed  $h(x, b, c, d)$ , således at  $p$ -værdien er

$$p\text{-værdi}_{\text{fisher.test}} = \sum_{x: h(x, b, c, d) \geq h(x_{\text{obs}}, b, c, d)} h(x, b, c, d). \quad (1.9)$$

### Eksempel 1.8. Homers Illiade

Vi betragter igen data fra Homers Illiade, men benytter kun data fra 13 hændelser, hvor skaden er på enten hånd eller arm:

Sted	Død	Ikke Død	Rækkesum
Hånd	0	4	4
Arm	3	6	9
Søjlesum	3	10	13

De forvente antal under hypotesen om uafhængighed mellem sted på kroppen og dødelighed er 0.9 og 3.1 i første række og 2.1 og 6.9 i anden række. Disse opfylder ikke Cochrans regel, og man kan derfor ikke stole på  $p$ -værdien beregnet fra en  $\chi^2$ -fordeling. Gennemfører vi beregningerne, fås  $G_{\text{obs}} = 2.59$  og  $1 - \chi^2_{\text{cdf}}(2.59, 1) = 0.11$ . I det næste kodevindue beregner jeg det eksakte test baseret på den betingede fordeling givet rækkesummer og sjølesummer. Det betingede test baseret på  $G$ -teststørrelsen giver en  $p$ -værdi på 0.31, hvorimod det betingede test baseret på (1.9) giver en  $p$ -værdi på 0.50. I begge tilfælde en noget højere  $p$ -værdi end den approksimative fra  $\chi^2$ -fordelingen. Forskellen mellem de to betingede tests ligger i, at den sidste  $p$ -værdi er beregnet ud fra de tre udfald 0, 2 og 3 for antallet af døde, hvor skaden er i hånden, i forhold til  $p$ -værdien baseret på  $G$ , hvor kun udfaldene 0 og 3 medtages.

#### Kodevindue

```
Aobs=rbind(c(0,4),c(3,6))
b=sum(Aobs[1,]); c=sum(Aobs[2,]); d=sum(Aobs[,1])

xvek=c(max(c(0,d-c)):min(b,d))
hxvek=dhyper(xvek,b,c,d)

Gfct=function(Amat){
  ex=outer(rowSums(Amat), colSums(Amat)) /sum(Amat)
  A1=ifelse(Amat==0,1,Amat)
  return(2*sum(Amat*log(A1/ex)))
}
Gobs=Gfct(Aobs)
hxobs=dhyper(Aobs[1,1],b,c,d)

Gxvek=c()
for (x in xvek){
  Gxvek=c(Gxvek,Gfct(rbind(c(x,b-x),c(d-x,x+c-d))))
}

c(Gtest=sum(hxvek[Gxvek>=Gobs]), RtestMin=sum(hxvek[hxvek<=hxobs]),
  Rtest=fisher.test(Aobs)$p.value)
```

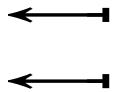
#### Showhide: Forstå koden

Her følger en række spørgsmål til forståelsen af koden.

1. Hvad indeholder vektoren  $xvek$ ?
2. Hvad tror du  $hxvek$  indeholder?
3. Lav et test for uafhængighed mellem sted og dodelighed, når du kun betragter hændelser i Homers Illiade, hvor skaden er på enten skulder eller ben.

**Showhide: Svar**

1. Vektoren  $xvek$  indeholder de mulige værdier af (1, 1)-indgangen i datamatricen  $A$ , når rækkesummer og søjlesummer holdes fast. I det konkrete eksempel bliver dette værdierne 0, 1, 2 og 3.
2. Funktionen  $dhyper$  beregner de betingede sandsynligheder  $h(x, b, c, d)$ .
3. Kodevinduet køres igen med `Aobs=rbind(c(9,3),c(1,8))`. Den betingede  $p$ -værdi bliver 0.0075.



## 1.9 Diverse

### 1.9.1 Notation for fordelingsfunktion og fraktil

Når vi for en stokastisk variabel vil udregne sandsynligheden  $P(X \leq x)$  (sandsynligheden for at ligge til venstre for  $x$ ), taler vi om at udregne fordelingsfunktionen i punktet  $x$ . Fordelingsfunktion hedder på engelsk *cumulative distribution function*, som forkortes *cdf*. I denne bog benytter jeg *cdf* som nedre fodtegn på et fordelingsnavn for at angive fordelingsfunktionen. Med denne notation betyder  $\text{binom}_{\text{cdf}}(138, 580, 0.25)$  således sandsynligheden for en værdi mindre end eller lig med 138 i en  $\text{binom}(580, 0.25)$ -fordeling.

I kender også normalfordelingen fra jeres sandsynlighedskurser. Hvis  $X$  er normalfordelt med middelværdi  $\mu$  og varians  $\sigma^2$ , skriver vi  $X \sim N(\mu, \sigma^2)$ . Sandsynligheden for at ligge til venstre for  $x$  i denne fordeling betegnes med  $N_{\text{cdf}}(x, \mu, \sigma)$ . For en given sandsynlighed  $p$  kan vi finde det punkt  $x_p$ , således at sandsynligheden for at ligge til venstre for dette punkt er  $p$ . Dette kaldes  $p$ -fraktilen i fordelingen. Notationsmæssigt angiver vi fraktiler ved at tilføje det nedre fodtegn *inv* til

fordelingsnavnet. Således er  $N_{\text{inv}}(0.95, 2, 1)$  95%-fraktilen i en normalfordeling med middelværdi 2 og spredning 1.

I **R** får man fordelingsfunktionen ved at sætte bogstavet  $p$  foran navnet på fordelingen, og fraktiler fås ved at sætte bogstavet  $q$  foran fordelingsnavnet. For en normalfordeling får man fordelingsfunktionen i **R** med kaldet `pnorm(x, mu, sigma)`. Bemærk at der bruges spredning  $\sigma$  og ikke varians  $\sigma^2$  i kaldet til *norm*. For standard normalfordelingen med middelværdi 0 og spredning 1 kan man udelade middelværdi og spredning i kaldet til *norm*.

### 1.9.2 Egne funktioner i R

Nogle få gange i dette kursus vil jeg bede jer om at lave nogle beregninger, der ikke laves nemt med standardfunktioner i **R**. Til dette har jeg lavet nogle nye funktioner, der alle er defineret i filen *Rfunktioner.txt*, som findes på kursushjemmesiden. Denne fil placeres naturligt i mappen *source*, I fik dannet gennem startpakken til **R** fra kursushjemmesiden. Funktionerne er til rådighed, når I har givet kommandoen `source("../source/Rfunktioner.txt")`. Filen *Rfunktioner.txt* indeholder følgende funktioner:

- (i) `errorbar`
- (ii) `qqnormFlere` (identisk med Utes funktion `qqnorm_mult`)
- (iii) `inversReg`
- (iv) `additivitetsPlot`
- (v) `FWstep`
- (vi) `FWcrossval`

I nedenstående kodevindue viser jeg koden for *errorbar* og et eksempel på brug af koden. Funktionen *errorbar* tager som input fire vektorer af samme længde, de første to angiver første- og andenkoordinaten for en række punkter, og de to sidste angiver nedre og øvre endepunkter for lodrette linjestykker gennem punkterne.

#### Kodevindue

```
errorbar=function(x,y,lower,upper){
  points(x,y)
  arrows(x,lower,x,upper,code=3,angle=90,length=0.05)
}

x=c((-2):2)
plot(x,x^2,type="l",ylim=c(-1,6))
errorbar(x,x^2,x^2-abs(x/4)-0.1,x^2+abs(x/4)+0.1)
```

## 1.10 Svar

### Svar 1.1. Multinomialsandsynligheder

1. Sandsynlighederne for de tre udfald er 0.0278 for  $(1, 1, 1, 0, 0, 0)$ , 0.0139 for  $(1, 2, 0, 0, 0, 0)$  og 0.0046 for  $(3, 0, 0, 0, 0, 0)$
2. Sandsynligheden for tre forskellige tal er antallet af måder at vælge 3 positioner ud af 6 og gange dette med 0.0278. Dette giver  $20 \cdot 0.0278 = 0.556$ . For at beregne sandsynligheden for to forskellige tal bruger vi, at der er 15 måder at vælge to positioner ud af 6, og for hver af disse er der to muligheder for at skrive 1 og 2 på de to positioner. Dette giver 30 muligheder der skal ganges med 0.0139 som giver 0.417. Endelig er der 6 måder at vælge 1 position, svarende til kun at få et tal, og ganger vi dette med 0.0046 får vi 0.028.

### Svar 1.2. Homogenitetstest

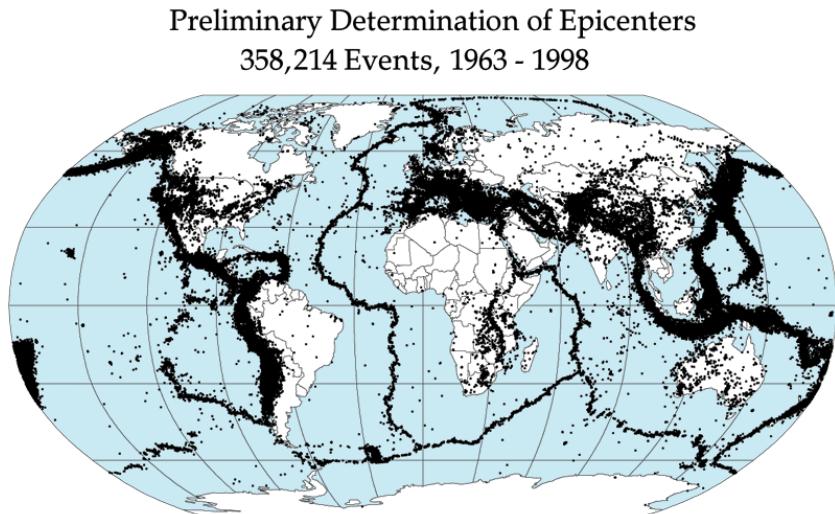
Funktionen `rowSums` beregner rækkesummer, og giver derfor en vektor af længde 3. Funktionen `colSums` beregner søjlesummer, og giver derfor en vektor af længde 2. Funktionen `outer` tager to vektorer som input og danner en matriks, hvor den  $(i, j)$ 'te indgang er den  $i$ 'te indgang i den første vektor ganget med den  $j$ 'te indgang i den anden vektor, hvorfor `ex` bliver matricen med de forventede antal. Endelig indeholder `G` og `pval` henholdsvis  $G$ -teststørrelsen og den tilhørende  $p$ -værdi. Koden i linjerne 2-5 kan skrives samlet som `ex=outer(rowSums(Obs), colSums(Obs))/sum(Obs)`.

## 1.11 Opgaver til kapitel 1

I opgaverne hørende til kapitel 1 skal I blive fortrolige med multinomialfordelte data og test af hypoteser om sandsynlighedsparametrene. Specielt skal I se metoden brugt til at lave goodness of fit test. Til sidst skal I sammenligne data fra flere multinomialfordelinger.

### Showhide: Opgave 1.1: Goodness of fit test: Uniform fordeling

I 2009 publicerede Pieter Vermeesch en lille note med den provokerende titel [Lies, damned lies, and statistics \(in geology\)](#). I noten betragter forfatteren 118415 jordskælv af styrke 4 eller over på Richterskalaen i perioden 1/1-1999 til 1/1-2009 (data fra [earthquake.usgs.gov](#)) og fordeler disse på ugedag. Billedet her viser, hvor jordskælv optræder.



Data kan ses i den følgende tabel og findes i filen *JordskælvDag.csv*.

Data	Mandag	Tirsdag	Onsdag	Torsdag	Fredag	Lørdag	Søndag
Observeret	16853	16553	16490	17399	16348	17019	17753

- (a) Opstil multinomialmodellen for disse data, hvor sandsynlighederne for at falde i de syv kasser er vilkårlige.
- (b) Opskriv, inden for den opstillede multinomialmodel, hypotesen om ligelig fordeling på de syv ugedage. Udregn de forventede antal, og lav G-testet for hypotesen. Hvad bliver konklusionen af testet?

Den overraskende konklusion er, at jordskælvene ikke fordeler sig ligeligt på de syv ugedage. Vermeesch ser dette resultat som udtryk for en svaghed i de statistiske metoder.

Har Vermeesch ret, eller bruger han metoderne forkert? Når vi laver goodness of fit test for en ligelig fordeling på de syv ugedage, sammenligner vi med *uafhængige* kast af en syvsidet terning. Som påpeget af Sornette og Pisarenko i artiklen [On the correct use of statistical tests: Reply to "Lies, damned lies and statistics \(in geology\)"](#), er det helt store problem mangel på uafhængighed på grund af efterskælv og klynger af små skælv. I datasættet er cirka 2/3 af skælvene efterskælv! Et andet stort problem er, at for de små jordskælv kan data indeholde menneskeskabte hændelser, og der kan være forskel i baggrundsstøj på de syv ugedage. Sornette og Pisarenko foreslår at betragte jordskælv med en styrke over 5 på Richterskalaen for at komme ud over det sidstnævnte problem, og foretager også en rensning af disse for at fjerne efterskælv (dette kan ikke gøres fuldstændigt sikkert). Data kan ses i den følgende tabel.

Data	Mandag	Tirsdag	Onsdag	Torsdag	Fredag	Lørdag	Søndag
Over 5	2374	2511	2291	2497	2153	2282	2360
Renset	780	847	793	831	785	821	779

- (c) Lav et G-test for en ligelig fordeling, både for alle jordskælv med en styrke over 5 og for delmængden, hvor efterskælv er fjernet.

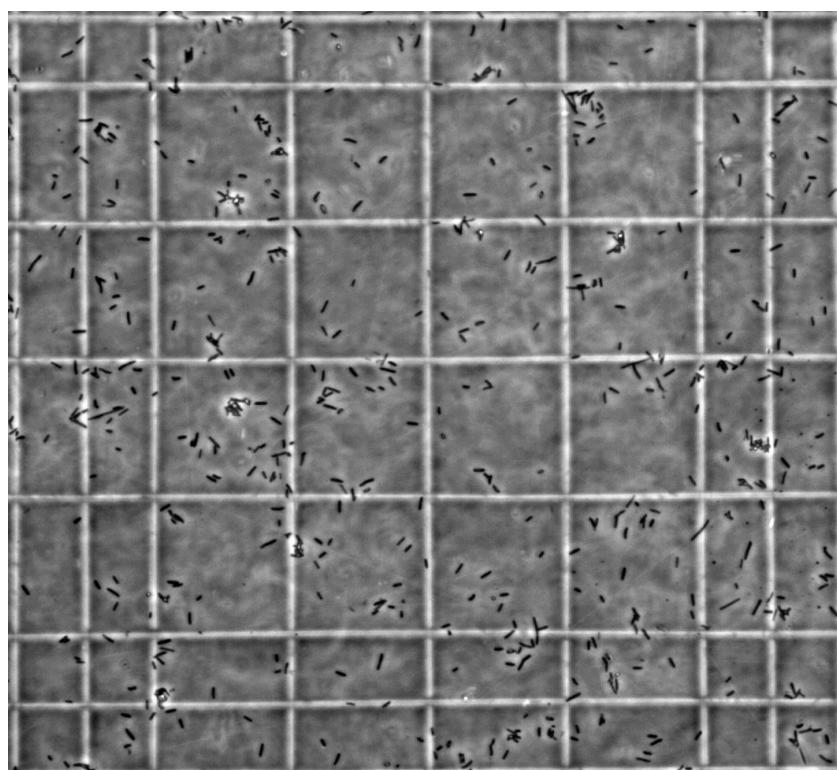
Ifølge Wikipedia blev udtrykket *lies, damned lies and statistics* gjort kendt af Mark Twain, som skrev "There are three kinds of lies: lies, damned lies, and statistics". Mark Twain mente, at udtrykket stammede fra den britiske premierminister Benjamin Disraeli, men dette kan tilsyneladende ikke bekræftes.

Vermesch var utilfreds med, at han fik en lav  $p$ -værdi for en hypotese, der synes oplagt sand. Dette er dog en mulighed, vi må være indstillet på, jævnfør omtalen af *fejl af type 1* i afsnit 8.4 i MSRR.



### Showhide: Opgave 1.2: Goodness of fit, poissonfordeling

Data i denne opgave tager udgangspunkt i forsøg med bakterieceller, hvor man ofte har behov for at tælle, hvor mange af disse man har i en given opløsning. Dette gøres ved at udtagte en mindre del af opløsningen og tage billede af denne i et mikroskop, hvor bakterierne så kan tælles. Et eksempel på et sådant billede er vist nedenfor, hvor de små sorte områder er enkelte *E. coli* bakterier. De store områder i midten af billedet er hver på  $50 \times 50 \mu\text{m}^2$ , og her inden for tælles antallet af bakterier. For at sikre konsistens tælles en bakteriecelle, der ligger ind over en kant, kun med hvis det er den venstre eller den øverste kant, der berøres.



I filen *CelleData.txt* ligger data fra optælling fra 14 sådanne billeder, hver med 16 områder. Data er indsamlet med henblik på opgaven her og stillet til rådighed af Morten Bormann Nielsen.

- (a) Indlæs de 224 kvadrattællinger med ordren `Ncoli=scan("CelleData.txt")`.

Lav et antalshistogram af data med intervalendepunkter `endePkt=c(4:32)-0.5` (det første interval er fra 3.5 til 4.5, svarende til at den mindste værdi i *Ncoli* er 4, og det sidste interval er fra 30.5 til 31.5, svarende til at den største værdi i *Ncoli* er 31). Indsæt titler på akserne i figuren ved at benytte `xlab` og `ylab` i kaldet til `hist`. Placer antallet af observationer i hvert interval i en vektor *antal*. Vælg et af intervallerne ud, og eftervis antallet i *antal* ved en direkte optælling blandt de 224 dataværdier.

- (b) Opskriv multinomialmodellen for den stokastiske antalsvektor *Antal*, hvor sandsynligheden for at falde i de forskellige kasser er vilkårlig.

I denne opgave skal I lave et goodness of fit test for at antallet af bakterier i et kvadrat er poissonfordelt. Lad  $\lambda$  være rateparametren i poissonfordelingen (enhed: antal per  $50^2 \mu\text{m}^2$ ). Da den første "kasse" i multinomialmodellen indeholder tælletal mindre end eller lig med fire, vil hypotesen om en poissonfordeling betyde at sandsynligheden for at falde i den første kasse er  $\sum_{j=0}^4 (\lambda^j / j!) e^{-\lambda}$ . Den anden kasse indeholder alle tælletal med værdien 5, og sandsynligheden for at falde i den anden kasse er  $(\lambda^5 / 5!) e^{-\lambda}$ , og så videre op til kasse 27. Sandsynligheden for at falde i den sidste kasse (kasse nummer 28) er 1 minus summen af sandsynlighederne for de første 27 kasser.

Som skøn over  $\lambda$  bruges gennemsnittet af de 224 observationer, se Proposition 6.1.2 i MSRR.

- (c) Opskriv, inden for din multinomialmodel, hypotesen om, at antallet af bakterier i et kvadrat er poissonfordelt.

Beregn de forventede antal under hypotesen. Hertil kan du benytte koden nedenfor. I R beregnes punktsandsynligheder i poissonfordelingen med `dpois(x,lambda)`, og sandsynligheden for en værdi mindre end eller lig med  $x$  beregnes med `ppois(x,lambda)`. Forklar, at koden giver de forventede værdier.

#### Kodevindue

```
lamhat=3324/224
Forv=224*c(ppois(4, lamhat), dpois(c(5:30), lamhat), 1-ppois(30, lamhat))
round(Forv, 2)
```

Indtægn de forventede antal i histogrammet fra spørgsmål (a) som en rød kurve med kommandoen `lines(c(4:31), Forv, col=2)`, hvor  $Forv$  er vektoren med de forventede antal.

- (d) Lav  $G$ -testet for hypotesen, at antal bakterier i et kvadrat er poissonfordelt. Slå kasser sammen, hvis de forventede ikke er større end 5 (slå kasser sammen fra hver sin ende, indtil det forventede antal er større end 5).

Hvad bliver konklusionen af dit goodness of fit test? Kan du give en forklaring på resultatet?

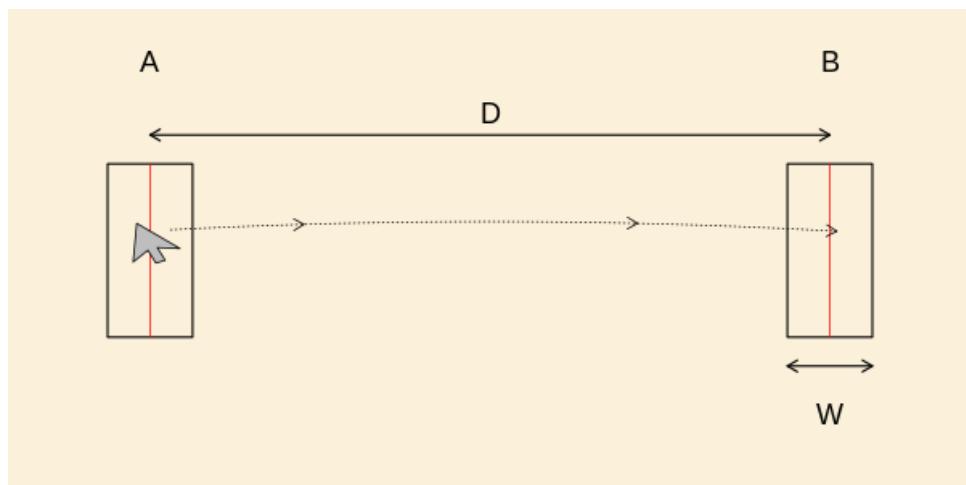
#### Showhide: Forklaring

Konklusionen af ovenstående analyse er, at poissonfordelingen ikke er en særlig god beskrivelse af data. Man kan indse, at de 224 tællingerne viser større spredning, end hvad man forventer i en poissonfordeling. Fortolkningen af dette er, at bakterierne ikke er tilfældigt spredt ud over området, nogle områder har større intensitet af bakterier end andre områder (bakterierne klumper).



#### Showhide: Opgave 1.3: Goodness of fit, normalfordeling

I menneske-maskine-interaktion betragtes blandt andet, hvordan man flytter pointeren på en computerskærm via musen. Figuren nedenfor viser en typisk opstilling, hvor en person skal flytte pointeren fra område A til område B.



Data er simulerede baseret på informationen i figur 1 i artiklen [An error model for pointing based on Fitts' Law](#). Der er 269 observationer målt relativt til midtpunkt af målområdet (enheden angives i figuren som "pixels"). Data ligger i filen *Position.txt*. Opgaven går ud på at lave et goodness of fit test for, at pointerpositionen kan beskrives med en normalfordeling.

- (a) Indlæs de 269 positioner med kommandoen `scan("Position.txt")`, og placer disse i variablen *pointer*. Lav et tæthedshistogram af data med intervalinddelingen `endePkt=c((-9):8)*2+0.5`. Placer antallet af observationer i hvert af de 17 intervaller i en vektor *antal*.

Hvis positionen af pointer skal beskrives med en normalfordeling, er det bedste valg af middelværdi  $\hat{\mu} = 0.1843$ , og det bedste valg af spredning er  $\hat{\sigma} = 4.9938$ . Indtegn normalfordelingstætheden i histogrammet med kommandoen

```
curve(dnorm(x, 0.1843, 4.9938), from=-20, to=20, add=TRUE)
```

- (b) Opskriv multinomialmodellen for den stokastiske vektor *Antal*, hvor sandsynlighederne for at falde i de forskellige intervaller er vilkårlige.

Opskriv dernæst hypotesen, at sandsynlighederne for at falde i de 17 intervaller er givet ved sandsynlighederne for intervallerne i en normalfordeling med middelværdi  $\mu$  og spredning  $\sigma$ . Husk at i denne sammenhæng skal det første interval opfattes som intervallet fra minus uendelig til -15.5, og det sidste interval skal opfattes som intervallet fra 14.5 til uendelig.

Beregn de forventede antal under hypotesen. Hertil kan du benytte koden nedenfor. I R beregnes sandsynligheden for en værdi mindre end eller lig med  $x$  i en normalfordeling med kommandoen `pnorm(x, mu, sigma)`. Forklar, at koden giver de forventede værdier.

#### Kodevindue

```

muhat=0.1843
sigmahat=4.9938
endePkt0=c((-8):7)*2+0.5
prob=pnorm(endePkt0 , muhat, sigmahat)
Forv=269*(prob [1] , prob[2:16]-prob[1:15], 1-prob [16])
round(Forv,2)

```

- (c) Lav G-testet for hypotesen, at pointerpositionen er normalfordelt. Kan disse data beskrives med en normalfordeling?



### Showhide: Opgave 1.4: Homogenitetstest

I bogen [Human-Computer Interaction: An Empirical Research Perspective](#) omtales kort et eksperiment, hvor kvinder og mænd observeres for at vurdere deres måde at scrolle i en tekst. Hver person klassificeres efter, om vedkommende bruger enten rullehjulet på musen til at scrollle i en tekst, bruger rullepanel på skærmen eller bruger tastaturtasterne. Der er 65 kvinder og 43 mænd i undersøgelsen. Fordelingen på de tre metoder for henholdsvis kvinder og mænd kan ses i tabelen nedenfor.

Køn	Rullehjul	Rullepanel	Tastatur	Total
Kvinder	37	16	12	65
Mænd	20	16	7	43

Vi ønsker med data at se, om der er kønsspecifikke måder at arbejde med computeren på.

- (a) Opstil den statistiske model, hvor tælletallene for hvert køn følger sin egen multinomialfordeling. Angiv inden for den opstillede model hypotesen, at der er samme sandsynlighedsvektor for kategorierne (Rullehjul, Rullepanel, Tastatur) for de to køn.
- (b) Undersøg, om data er i overensstemmelse med hypotesen om samme sandsynlighedsvektor for kategorierne (Rullehjul, Rullepanel, Tastatur) for de to køn (benyt eventuelt **R**-koden fra det skjulte kodevindue i eksempel 1.5).



### Showhide: Opgave 1.5: Poissonmodel med proportionale parametre

Antallet af jordskælv i et bestemt område og med en styrke i et givet interval beskrives ofte med en poissonfordeling. Et eksempel er artiklen [A Poisson model for earthquake frequency uncertainties in seismic hazard analysis](#). I artiklen betragtes blandt andet jordskælv i New Zealand. Information om disse kan findes på nettet under adressen [info.geonet.org.nz](http://info.geonet.org.nz). Data i tabellen nedenfor viser antallet for tre styrkeintervaller og for perioden 1930-2015. Styrken er på Richterskalaen, som er en logaritmisk skala. Hvis styrken stiger med 1, stiger den samlede energi i jordskælvet med  $10^{3/2} = 31.6$ . Gutenberg-Richter loven for jordskælv angiver forholdet mellem antallet af jordskælv af forskellig styrke. I tabellen er dette forhold angivet for de tre styrkeintervaller (med "b-value" i Gutenberg-Richter loven sat til 1).

Styrkeinterval	Antal	Forhold
6.0-6.3	72	1
6.3-6.6	41	1/2
6.6-6.9	25	1/4

I opgaven her skal I kun betragte de to første styrkeintervaller, 6.0-6.3 og 6.3-6.6. Idet vi vil bruge data til at vurdere holdbarheden af Gutenberg-Richter loven, skrives raterne i de to intervaller

som  $86\lambda_1$  og  $0.5 \cdot 86\lambda_2$ , eftersom en beregning viser, at under Gutenberg-Richter loven (med "b-value" lig med 1) er raten i det andet styrkeinterval halvt så stor som raten i det første styrkeinterval. Enheden på  $\lambda$  er antal forventede jordskælv per år. Gutenberg-Richter loven svarer således til hypotesen  $\lambda_1 = \lambda_2$ .

Lad os formulere situationen generelt gennem modellen

$$Y_1 \sim \text{pois}(t_1\lambda_1), \quad Y_2 \sim \text{pois}(t_2\lambda_2), \quad t_1 = 86, \quad t_2 = 0.5 \cdot 86,$$

hvor  $Y_1$  og  $Y_2$  er de stokastiske tælletal svarende til de to styrkeintervaller.

For at teste hypotesen  $\lambda_1 = \lambda_2$  kan man benytte følgende teoretiske resultat. Hvis vi forestiller os, at  $Y_1 + Y_2$  er fast (vi "betinger" med summen), så vil  $Y_1$  være binomialfordelt:

$$(Y_1 \text{ givet at } Y_1 + Y_2 = n) \sim \text{binom}(n, p), \quad p = \frac{t_1\lambda_1}{t_1\lambda_1 + t_2\lambda_2}. \quad (1.10)$$

Hvis  $\lambda_1 = \lambda_2$  bliver  $p = t_1/(t_1 + t_2)$ . Et test for hypotesen  $\lambda_1 = \lambda_2$  kan derfor laves som et test i binomialfordelingen for hypotesen, at sandsynlighedsparameteren  $p$  har værdien  $p = t_1/(t_1 + t_2)$ . I vores tilfælde bliver dette hypotesen, at  $p = 2/3$ .

- (a) Find  $p$ -værdien for et test af hypotesen  $\lambda_1 = \lambda_2$ , for data i de to første rækker af tabellen ovenfor, ved at teste  $p = 2/3$  i modellen (1.10). Hvad bliver konklusionen af dette test?

*Bemærkning:* Hvis data strider mod hypotesen  $\lambda_1 = \lambda_2$ , vil vi være interesseret i at indføre en parameter  $\theta$ , således at  $\lambda_2 = \theta\lambda_1$ . Parameteren  $\theta$  angiver, hvor mange gange større  $\lambda_2$  er i forhold til  $\lambda_1$ . Hvis vi lader  $p = t_1/(t_1 + \theta t_2)$ , kan vi løse for  $\theta$ , og får  $\theta = (1 - p)t_1/(pt_2)$ . Konfidensinterval for sandsynlighedsparameteren  $p$  i binomialmodellen (1.10) kan derfor oversættes til et konfidensinterval for forholdet  $\theta$ .

Hvis denne metode benyttes på data i første og tredje styrkeinterval i tabellen ovenfor, får vi intervallet  $[0.65, 0.82]$  for 95%-konfidensintervallet af sandsynlighedsparameteren  $p$ . Oversat til forholdet  $\theta$  mellem rateparametrene i de to styrkeintervaller giver dette intervallet  $[0.22, 0.55]$ . Dette interval indeholder værdien  $\frac{1}{4}$ , som svarer til Gutenberg-Richter loven.



### Showhide: Opgave 1.6: Ændring i stormmønster

DMI vedligeholder en side med alle [storme i Danmark](#) fra 1891 og fremefter. Stormene kalssificeres i fire styrkekategorier ud fra vindstyrken. I nedenstående tabel har jeg optalt antallet af storme i forskellige kategorier for fire 30-års perioder.

Periode	Stormstyrke 1	Stormstyrke 2	Stormstyrke 3 og 4
1891-1920	39	16	4
1921-1950	21	8	8
1951-1980	14	12	5
1981-2010	18	12	10

- (a) Opstil den statistiske model, hvor antallet af storme for hver periode følger sin egen multinomialfordeling på de tre kategorier 1, 2 og 3+4. Angiv inden for den opstillede model hypotesen, at der er samme sandsynlighedsvektor for de tre styrkekategorier 1, 2 og 3+4 for de fire tidsperioder.

- (b) Undersøg, om data er i overensstemmelse med hypotesen om samme sandsynlighedsvektor for kategorierne 1, 2 og 3+4 for de fire tidsperioder.

For storme i Danmark finder I således ikke en ændring i fordeling på styrkekategori. Den næste tabel viser fordelingen af [hurricanes](#) fra verdenshavene på styrkekategorierne 1-3 og 4-5. Der er data fra to tidsperioder: 1975-1989 og 1990-2004. Data er fra artiklen [Changes in Tropical Cyclone Number, Duration, and Intensity in a Warming Environment](#).

Periode	Hurricanes 1-3	Hurricanes 4-5
1975-1989	617	171
1990-2004	508	269

- (c) Undersøg om fordelingen af hurricanes på de to styrkekategorier er den samme for de to tidsperioder.

For hurricanes ser det således ud til, at der er sket en ændring. I det sidste spørgsmål i denne opgave skal du se på, om ændringen er den samme i de forskellige verdenshave. Den næste tabel viser fordelingen af de største hurricanes (kategori 4-5) på fem verdenshave for de to tidsperioder.

Verdenshav	1975-1989	1990-2004
East Pacific Ocean	36	49
West Pacific Ocean	85	116
North Atlantic	16	25
Southwestern Pacific	10	22
Indian	24	57

- (d) Opstil model for data, og undersøg, om fordeling på de fem verdenshave er den samme i de to perioder.



### Showhide: Opgave 3.7: Uafhængighedstesttest

I opgave 10.5 i MSRR benyttes et datasæt fra artiklen [Corporate Social Responsibility and Workers' Well-being in Nigerian Banks](#). Forfatterne har spurgt 137 personer, der arbejder i banksektoren, om de bruger de sundhedsmuligheder banken stiller til rådighed og om sundhedstilbuddene er tilstrækkelige. Data er i følgende tabel

Tilbud tilstrækkelige	Ja	Nej
Bruger muligheder: Ja	97	8
Bruger muligheder: Nej	26	6

- (a) Opstil en model for data i tabellen. Opstil dernæst en hypotese om sammenhængen mellem de to inddelingskriterier.
- (b) Lav både  $G$ -testet for den opstillede hypotese, såvel som permutationstetet og Fishers eksakte test. Kommenter på resultaterne.



### Showhide: Opgave 3.8: Betingning i poissonmodel

Lad  $X_i \sim \text{pois}(\lambda_i)$ ,  $i = 1, \dots, k$ , være uafhængige stokastiske variable.

- Angiv fordelingen af  $N = X_1 + X_2 + \dots + X_k$ .
- Find den betingede sandsynlighed  $P(X_1 = x_1, \dots, X_k = x_k | N = n)$ . Angiv i ord den betingede fordeling af  $(X_1, \dots, X_k)$  givet at  $N = n$ .



### Showhide: Opgave 3.9: Simulere styrke

I denne opgave skal I finde styrken af et test ved simulering. I skal betragte modellen  $X_1 \sim \text{binom}(n_1, p_1)$ ,  $X_2 \sim \text{binom}(n_2, p_2)$ , uafhængige, og test af hypotesen  $p_1 = p_2$ . Styrken skal beregnes i tilfældet, hvor vi forkaster hypotesen, når  $p$ -værdien fra Resultat 1.4 er mindre end 0.05. I kodevinduet nedenfor er vist det meste af den nødvendige kode.

#### Kodevindue

```

n1=20
n2=20
p1=0.4
p2=0.4
nsim=10000

x1=rbinom(nsim,)
x2=rbinom(nsim,)
phat=(x1+x2)/(n1+n2)

Q=(phat^(x1+x2)*(1-phat)^(n1+n2-x1-x2))/
((x1/n1)^x1*(1-x1/n1)^(n1-x1)*(x2/n2)^x2*(1-x2/n2)^(n2-x2))

G=
pvaerdi=1-pchisq(G,1)

sum() / nsim

```

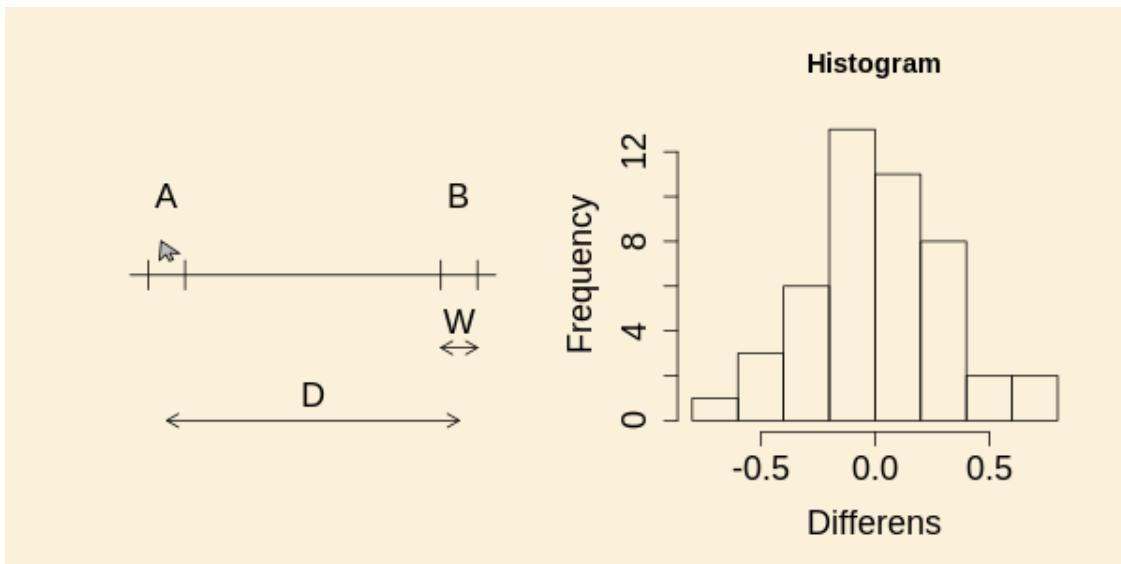
- Indsæt de manglende argumenter i de to kald til *rbinom*. Henvis til formel i webbogen med hensyn til beregningen af likelihood ratio teststørrelsen  $Q$ , og indsæt den manglende kode for  $G$ . Indsæt den manglende kode i *sum* i den sidste linje for at beregne den simulerede styrke.
- Benyt koden til at finde ud af, hvor stor  $n_1 = n_2$  skal være, for at styrken er mindst 0.80 i tilfældet, hvor  $p_1 = 0.4$  og  $p_2 = 0.5$ .





## Normalfordelte data

Figuren nedenfor viser et histogram for 46 differenser mellem tiden for at flytte computermusen fra punkt A til punkt B og tiden for at flytte musen retur fra B til A.



Der er flere ting at bemærke her. Data er kontinuerte i den forstand, at den målte differens kan antage alle mulige værdier. Den tilhørende stokastiske variabel beskrives gennem en tæthed, der kan fortolkes som sandsynlighed per længde. Vi kan tænke på de 46 målinger som et udsnit af en underliggende uendelig stor population af differenser, og det er denne population, vi ønsker at sige noget om ud fra de målte værdier i eksperimentet. Vores øjne vil automatisk ud fra histogrammet danne sig et billede af et centrum for fordelingen og et billede af, hvor stort et område målingerne spredte sig over. I forhold til den underliggende population svarer dette til middelværdi og spredning af fordelingen af differensen.

Det er naturligt at spørge, om middelværdien er nul svarende til, at man bruger lige lang tid på at flytte musen fra A til B som fra B til A. Det vil være oplagt her at se på, hvor langt gennemsnittet af målingerne ligger fra den forventede værdi nul. Her løber vi dog ind i et problem.

Sandsynligheden for at få noget, der ligger længere væk fra det forventede end det observerede gennemsnit, vil afhænge af spredningen i populationen.

I dette kapitel indfører jeg det nok mest udbredte statistiske test, nemlig  $t$ -testet (Students  $t$ -test). Testet tager hensyn til den ukendte spredning i populationen ved at standardisere afstanden mellem gennemsnit og den forventede værdi med en spredning beregnet ud fra målingerne. Testet indføres i afsnit 2.4. Testet tager udgangspunkt i en antagelse om, at data følger en normalfordeling. Afsnit 2.1 repeterer meget kort jeres viden om normalfordelingen fra jeres calculuskursus, og afsnit 2.2 giver en grafisk metode til at vurdere, om data kan beskrives med en normalfordeling. Det er en empirisk kendsgerning, at mange data kan beskrives med en normalfordelingsmodel, og kapitlerne 3 til 5 i bogen her omhandler forskellige modeller for normalfordelte data. Et teoretisk argument for, at data ofte kan beskrives med en normalfordeling, kan findes i den *centrale grænseværdidisætning* (afsnit 4.3 i MSRR), der siger, at hvis en stokastisk variabel kan tænkes på som fremkommet som en sum af mange små bidrag, så vil fordelingen ligne en normalfordeling.

I dette kapitel betragtes to normalfordelingsmodeller. Den første model er den grundliggende model med én gruppe af observationer, svarende til eksemplet ovenfor med differens af flyttetider, og den anden model er, hvor vi har to grupper af observationer og ønsker at sammenligne middelværdierne i de to underliggende populationer. De to modeller analyseres i afsnittene 2.3 og 2.9. Analyse af data ved brug af R beskrives i afsnittene 2.7 og 2.13.

Ovenfor har jeg fokuseret på at uddrage viden fra data om middelværdien i populationen. Spredningen kan dog også være af interesse, og inferens om denne er beskrevet i afsnittene 2.6 og 2.12.

## 2.1 Normalfordelingen

Hvis den stokastiske variabel  $X$  er **normalfordelt** med middelværdi  $\mu$  og spredning  $\sigma$ ,  $X \sim N(\mu, \sigma^2)$ , er tætheden givet ved

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right), \quad x \in \mathbf{R}, \quad (\mu, \sigma) \in \mathbf{R} \times \mathbf{R}_+. \quad (2.1)$$

Sandsynligheden for at ligge i et interval  $(a, b]$  er givet ved

$$P(a < X \leq b) = \int_a^b f(x; \mu, \sigma) dx.$$

Bemærk specielt, at dette viser  $P(X = x) = 0$  for enhver værdi  $x$ : for en kontinuert stokastisk variabel er enhver punktsandsynlighed lig med nul. Fordelingsfunktionen, det vil sige sandsynligheden for at ligge til venstre for et punkt, betegnes med  $N_{cdf}(x; \mu, \sigma)$ . Bemærk, at I fordelingsnotationen  $N(\mu, \sigma^2)$  bruges variansen  $\sigma^2$ , hvorimod i tætheden og fordelingsfunktionen bruges spredning  $\sigma$ . Det sidste er for at være i overensstemmelse med notationen i R. Normalfordelingen med middelværdi  $\mu = 0$  og spredning  $\sigma = 1$  kaldes *standard normalfordelingen*.

Der gælder følgende vigtige regneregler (disse er omtalt i jeres calculuskursus).

- (i) Hvis  $X \sim N(\mu, \sigma^2)$ , og  $a$  og  $b$  er givne tal, så vil  $a + bX \sim N(a + b\mu, b^2\sigma^2)$ . Specielt kan vi skrive  $X = \mu + \sigma U$ , hvor  $U \sim N(0, 1)$ .

- (ii) Hvis  $X$  og  $Y$  er normalfordelte og uafhængige, så vil også  $X + Y$  være normalfordelt.
- (iii) Hvis  $U_1, \dots, U_k$  er uafhængige og standard normalfordelte, så har den stokastiske variabel  $U_1^2 + \dots + U_k^2$  en  $\chi^2(k)$ -fordeling.

På et intuitivt niveau kan vi forstå de to første regneregler ud fra den centrale grænseværdidisætning (afsnit 4.3 i MSRR). Hvis  $X$  er en sum af mange små bidrag, så er dette også tilfældet for  $a + bX$ . Hvis både  $X$  og  $Y$  er summer af mange små led, så vil dette også gælde for summen af de to.

I kodevinduet nedenfor vises tæthed og fordelingsfunktion for en normalfordeling. Desuden er 97.5%-fraktilen markeret, det vil sige punktet, hvor der ligger 97.5% sandsynlighed til venstre for og 2.5% sandsynlighed til højre for. Tætheden i en normalfordeling beregnes med funktionen *dnorm* i R. Fordelingsfunktionen beregnes med *pnorm*, og fraktiler findes med funktionen *qnorm*. Hvis man ikke angiver middelværdien og spredningen i kaldet til de tre funktioner, benytter R standard normalfordelingen. Prøv at køre kommandoerne, og prøv at ændre på  $\mu$  og  $\sigma$ . Bemærk, at figuren har to andenakser, hvor aksen til venstre angiver tæthed og aksen til højre angiver sandsynlighed.

### Kodevindue

```
par(mar=c(5,5,2,5))
mu=0
sigma=1
x=5*c(-1000:1000)/1000
plot(x,dnorm(x,mu,sigma),ylab="Tæthed",type="l",
      xlim=c(-5,5),ylim=c(0,2))
abline(v=qnorm(0.975,mu,sigma),col=2)
par(new=TRUE)
plot(x,pnorm(x,mu,sigma),axes=F,xlab=NA,ylab=NA,col=4,
      type="l",xlim=c(-5,5))
axis(side=4)
mtext(side=4,line=3,"Fordelingsfunktion")
```

Prøv, om du kan beregne 95%-fraktilen i en standard normalfordeling. Beregn også 5%-fraktilen.

### Svar 2.1. Fraktiler

## 2.2 Normal-qqplot

### Eksempel 2.1. (Kontrol af køkkenvægt)

Som et eksempel ønskede jeg at måle vægten af 10 pakker af den muesli, jeg spiser til morgenmad, for at se, om pakkerne holder den lovede vægt på 600 gram. Til rådighed har jeg dog kun min simple køkkenvægt, og jeg ville ikke kunne sige, om en eventuel afvigelse skyldes indholdet af

muesli eller en fejlvisning af min køkkenvægt. I stedet lavede jeg et kalibreringeksperiment. Ti gange tændte jeg for køkkenvægten, mælte vægten af et målebæger, fyldte målebægeret op til 600 ml markeringen med vand, mælte vægten igen og trak de to vægtmålinger fra hinanden. De 10 differencer er skrevet ind i kodevinduet nedenfor. Det første spørgsmål, jeg ønsker at stille, er, om målingerne stemmer overens med det forventede 600 gram? Det andet spørgsmål, jeg er interesseret i, er, hvor meget variation der er i målingerne? Umiddelbart vil jeg bruge dette til at sige, hvor stabil min køkkenvægt er, men så simpelt er det ikke: variationen i målingerne kommer både fra køkkenvægten og fra min opmåling af de 600 ml vand.



Jeg vil gerne bruge normalfordelingen til at beskrive mine data, men er dette rimeligt? Jeg har ikke nok data til at lave et goodness of fit test som beskrevet i afsnit 1.4. I stedet vil jeg her beskrive en grafisk undersøgelse, der kan give en indikation af, om det er rimeligt at bruge en normalfordeling. I den grafiske metode laves en figur, hvor punkterne bør "sno sig" omkring en ret linje, i fald data stammer fra en normalfordeling. Med kun ganske få datapunkter, som i mit eksempel ovenfor, kan det være svært at afgøre, om data afviger fra at "sno sig" omkring en ret linje. Den grafiske undersøgelse er således af større værdi, hvor man har flere datasæt og kan se, om de alle viser den samme form for afvigelse fra "sno sig"-egenskaben. Det følgende kodevindue laver den grafiske undersøgelse for data i eksemplet ovenfor. R-kommandoen, der skal bruges, er `qqnorm` (den afsluttende "`c()`" er ikke relevant, når I kører R på jeres egen computer).

### Kodevindue

```
v600=c(579,583,586,601,576,559,609,572,567,587)
qqnorm(v600)
c()
```

Jeg beskriver nu den grafiske undersøgelse, lavet i kodevinduet ovenfor, som går under navnet *normal-qqplot*. Her står "q" for *quantile*, som på dansk er *fraktile*, og på dansk taler man om en *fraktilsammenligning*. For nemhed i notationen vil jeg fremover blot omtale metoden som et *qqplot*. For at beskrive metoden lader jeg  $u_p$  være  $p$ -fraktilen i en standard normalfordeling,  $N(0, 1)$ -fordelingen, det vil sige, at  $N_{cdf}(u_p, 0, 1) = p$ . I R beregnes  $p$ -fraktilen som `qnorm(p)`.

Vi betragter  $n$  datapunkter  $x_1, x_2, \dots, x_n$  og ordner disse efter størrelse,  $x_{[1]}$  betegner den mindste,  $x_{[2]}$  den næstmindste, og så videre op til  $x_{[n]}$  som er den største:  $x_{[1]} \leq x_{[2]} \leq \dots \leq x_{[n]}$ . Et *qqplot*

består i at tegne punkterne

$$(u_{(i-0.5)/n}, x_{[i]}), \quad i = 1, 2, \dots, n.$$

I R produceres denne figur med kommandoen `qqnorm(x)`, hvor  $x$  er en vektor med data.

### Showhide: QQplots af simulerede data

I nedenstående kodevinduer vises nogle eksempler på qqplots. Hvis vektoren  $x$  indholder data-værdierne, laves et (normal-) qqplot med kommandoen `qqnorm(x)`. R-kommandoen `qqline(x)` tilføjer en linje til figuren.

Her kommer først et kodevindue, hvor data er normalfordelt. Der laves en figur med fire qqplots, alle med det samme antal observationer. Prøv at køre koden et par gange. Prøv dernæst at ændre  $n$  fra 10 til 40, og dernæst til 100. Kommandoen `rnorm(n)` simulerer  $n$  observationer fra en standard normalfordeling.

#### Kodevindue

```
n=10
par(mfrow=c(2,2))
x=rnorm(n)
qqnorm(x, pch=20)
qqline(x)
x=rnorm(n)
qqnorm(x, pch=20)
qqline(x)
x=rnorm(n)
qqnorm(x, pch=20)
qqline(x)
x=rnorm(n)
qqnorm(x, pch=20)
qqline(x)
```

Nu følger et kodevindue, der danner en figur med fire qqplots, og hvor data ikke er normalfordelte for de sidste to qqplots. Prøv at køre koden et par gange. Prøv dernæst at ændre  $n$  fra 10 til 40, og dernæst til 100.

#### Kodevindue

```
n=10
x=rnorm(n)
par(mfrow=c(2,2))
qqnorm(x, main="Normalfordeling", pch=20)
qqline(x)
qqnorm(3+2*x, main="Normalfordeling", pch=20)
qqline(3+2*x)
qqnorm(exp(x), main="Log-normalfordeling", pch=20)
qqline(exp(x))
qqnorm(pnorm(x), main="Uniform_fordeling", pch=20)
```

**qqline (pnorm(x))**

1. I det øvre højre delplot betragtes data fra en normalfordeling. Hvad er middelværdi og spredning i denne normalfordeling?
2. I det nedre venstre delplot betragtes data fra en stokastisk variabel, der kun kan antage positive værdier. Hvad er fordelingen af logaritmen til den stokastiske variabel?

**Svar 2.2. Fordelinger**



Hvorfor giver et qqplot en figur, hvor normalfordelte data sno sig omkring en ret linje? Her er kort den tekniske ide bag et qqplot. Hvis data følger en  $N(\mu, \sigma^2)$ -fordeling, så bør der gælde, at for ethvert  $p$  mellem 0 og 1 vil  $p$ -fraktilen beregnet ud fra data  $x_1, \dots, x_n$  ligne  $p$ -fraktilen i en  $N(\mu, \sigma^2)$ -fordeling. Hvis  $X \sim N(\mu, \sigma^2)$ , kan vi skrive  $X = \mu + \sigma U$ , hvor  $U \sim N(0, 1)$ , hvorfor  $p$ -fraktilen for  $X$  kan skrives som  $\mu + \sigma u_p$ . Hvad mener jeg med fraktiler beregnet ud fra data  $x_1, \dots, x_n$ ? Per definition af de ordnede værdier vides, at i punktet  $x_{[i]}$  er andelen af data mindre end eller lig med denne værdi givet ved  $i/n$ , men hvis vi betragter en værdi  $x$  lidt mindre end  $x_{[i]}$  (men større end  $x_{[i-1]}$ ), er andelen af data mindre end eller lig med denne værdi i stedet  $(i-1)/n$ . Vi vælger derfor at sige, at  $x_{[i]}$  er et skøn over  $((i - \frac{1}{2})/n)$ -fraktilen. Vores argument er derfor, at hvis data er  $N(\mu, \sigma^2)$ -fordelt, så bør

$$x_{[i]} \approx \mu + \sigma u_{(i-\frac{1}{2})/n}, \quad i = 1, \dots, n,$$

hvor " $\approx$ " skal læses som "cirka lig med". I qqplottet bør punkterne derfor sno sig om en linje med hældning  $\sigma$  og skæring  $\mu$ .

I et qqplot, som beskrevet ovenfor, er data langs andenaksen. Hvis man ønsker data langs førsteakse, kan man bruge kommandoen `qqnorm(x, datax=TRUE)`, hvor punkterne nu vil sno sig om en linje med hældning  $1/\sigma$ .

## 2.3 Model og estimation

Med et *observationssæt* mener jeg en række stokastiske variable, der alle har den samme fordeling, og i dette afsnit betragter jeg situationen med et normalfordelt observationssæt. Der er  $n$  uafhængige stokastiske variable  $X_1, \dots, X_n$  med tilhørende målinger  $x_1, \dots, x_n$ . Normalfordelingsmodellen er på formen

$$\text{Model: } X_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, n, \quad (\mu, \sigma^2) \in \mathbf{R} \times \mathbf{R}_+. \quad (2.2)$$

Det første emne er estimation af middelværdien  $\mu$  og variansen  $\sigma^2$ . I kapitel 1 med tælledata bliver likelihoodfunktionen brugt til at finde estimater. Likelihoodfunktionen bliver der defineret som sandsynligheden for det observerede som funktion af parameteren i modellen, og estimatet

er den værdi af parameteren, der giver maksimum af likelihoodfunktionen. For kontinuerte data kan man ikke bruge punktsandsynligheder (disse er nul), men vi har tæthedens til rådighed, som repræsenterer sandsynligheden for at ligge i et lille område omkring et punkt. For kontinuerte data defineres *likelihoodfunktionen* til at være tæthedens for det observerede som funktion af de parametre, der indgår i modellen. For uafhængige målinger bliver tæthedens et produkt af tæthedene for de enkelte målinger. For vores normalfordelingsmodel bliver likelihoodfunktionen

$$\begin{aligned} L(\mu, \sigma^2) &= f(x_1; \mu, \sigma) \cdot f(x_2; \mu, \sigma) \cdots f(x_n; \mu, \sigma) \\ &= \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right). \end{aligned}$$

Heraf fremgår, at hvis denne funktion skal maksimeres med hensyn til  $\mu$ , så skal vi minimere

$$\sum_{i=1}^n (x_i - \mu)^2.$$

Derfor kendes estimationen også under navnet *mindste kvadraters metode*. Intuitivt kan man sige, at  $\mu$  findes ved at minimere den samlede kvadratiske afstand mellem  $\mu$  og observationerne. Differentieres det sidste udtryk med hensyn til  $\mu$ , og sættes den afledede lig med nul, fås estimatet

$$\hat{\mu} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \sum_i x_i / n = \bar{x},$$

hvor en streg over et bogstav betyder *gennemsnit* af de tilhørende værdier.

For at estimere variansparameteren  $\sigma^2$  kan vi indsætte  $\hat{\mu}$  i likelihoodfunktionen,  $L(\hat{\mu}, \sigma^2)$ , og maksimere denne med hensyn til  $\sigma^2$ . Dette giver  $\hat{\sigma}^2 = \sum_i (x_i - \bar{x})^2 / n$ . Dette skøn er ikke helt tilfredsstillende, idet man kan vise (se nedenfor), at betragtet som stokastisk variabel gælder der, at  $E(\hat{\sigma}^2) = \sigma^2(n-1)/n \neq \sigma^2$ . Man bruger derfor i stedet estimatet

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

som kaldes den *empiriske varians*. Kvadratroden,  $s$ , kaldes den *empiriske spredning*. Når vi betragter  $s^2$  (eller  $s$ ) som en stokastisk variabel, afviger vi fra vores generelle regel og betegner også denne med  $s^2$  og ikke med  $S^2$ .

For at kunne lave tests og konfidensintervaller for parametrene er det nødvendigt at kende fordelingen af vores skøn betragtet som stokastiske variable. Der gælder følgende resultat.

### Resultat 2.2. (Fordeling af parameterskøn)

I normalfordelingsmodellen (2.2) gælder der følgende fordelingsresultater:

$$\hat{\mu} = \bar{X} \sim N(\mu, \sigma^2/n), \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \sigma^2 \chi^2(n-1)/(n-1).$$

Desuden er de stokastiske variable  $\hat{\mu}$  og  $s^2$  uafhængige.

Fordelingsresultatet for  $\hat{\mu}$  følger umiddelbart fra regnereglerne for normalfordelte stokastiske variable i afsnit 4.1. Specielt bruges, at

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}, \quad \text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Fordelingsresultatet for den empiriske varians  $s^2$  er sværere at forstå. Vi ved, hvad en  $\chi^2(n-1)$ -fordeling er, men hvad er en  $\sigma^2\chi^2(n-1)/(n-1)$  fordeling? Dette skal faktisk forstås på den måde, at den stokastiske variabel  $(n-1)s^2/\sigma^2$  følger en  $\chi^2(n-1)$ -fordeling. Jeg vil ikke udlede fordelingen af  $s^2$ , men intuitivt bygger resultatet på, at  $\sum_i(X_i - \mu)^2 \sim \sigma^2\chi^2(n)$  ifølge regnereglerne i afsnit 4.1. Når  $\mu$  så erstattes med  $\bar{X}$ , viser det sig, at antallet af frihedsgrader går fra  $n$  til  $n-1$ .

### Showhide: Middelværdi af variansskøn

Idet  $Z_i$  defineres som  $Z_i = X_i - \bar{X}$ , kan man skrive  $E(s^2) = \sum_{i=1}^n E((X_i - \bar{X})^2)/(n-1) = \sum_{i=1}^n E(Z_i^2)/(n-1)$ . Lad os starte med at se på  $Z_i = X_i - \bar{X}$ . Denne kan skrives som

$$Z_i = (1 - \frac{1}{n})X_i - \frac{1}{n}X_1 - \cdots - \frac{1}{n}X_{i-1} - \frac{1}{n}X_{i+1} - \cdots - \frac{1}{n}X_n.$$

Fra regneregler for normalfordelingen findes, at  $Z_i \sim N(0, \sigma^2(1 - \frac{1}{n}))$ . Dette giver

$$E(Z_i^2) = \text{Var}(Z_i) = \sigma^2(1 - \frac{1}{n}).$$

For den empiriske varians  $s^2$  gælder der nu

$$E(s^2) = \frac{1}{n-1} \cdot n \cdot \sigma^2(1 - \frac{1}{n}) = \sigma^2.$$



## 2.4 Test og konfidensinterval for middelværdi

Vi ønsker nu at lave et test for, om middelværdien har en bestemt værdi  $\mu_0$ , det vil sige et test for hypotesen  $\mu = \mu_0$  inden for normalfordelingsmodellen (2.2). Intuitivt baseres testet på, om  $\hat{\mu} = \bar{x}$  ligger tæt på  $\mu_0$  eller langt fra, men tæt på eller langt fra skal ses i lyset af spredningen  $\sigma$  i fordelingen. Umiddelbart inddrages spredningen gennem en standardisering på formen  $(\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ , baseret på at  $\text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ . Da spredningen  $\sigma$  ikke kendes, erstattes denne med vores skøn, den empiriske spredning  $s$ . Dette giver teststørrelsen

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

Jeg har indført teststørrelsen ud fra et intuitivt argument, men en beregning viser, at store værdier af  $|T|$  er ækvivalent med små værdier af *likelihoodratio teststørrelsen*  $Q$ , som er forholdet mellem den maksimale værdi af likelihoodfunktionen under hypotesen  $\mu = \mu_0$  og den maksimale værdi af likelihoodfunktionen under den fulde model (2.2). For at kunne bruge teststørrelsen er det nødvendigt at kende fordelingen af denne. Hertil skal vi bruge følgende definition.

### Definition 2.3. ( $t$ -fordeling)

Betræt uafhængige stokastiske variable  $U \sim N(0, 1)$  og  $V \sim \chi^2(df)$ . Så siges  $T = U/\sqrt{V/df}$  at følge en  $t$ -fordeling med  $df$  frihedsgrader,  $T \sim t(df)$ . Fordelingsfunktionen i en  $t(df)$ -fordeling beregnes i R med kommandoen `pt(t, df)`, som er sandsynligheden for at ligge til venstre for  $t$ .

### Showhide: Vise $t$ -fordeling i R

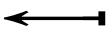
I kodevinduet nedenfor vises fordelingsfunktionen for tre  $t$ -fordelinger og for standard normalfordelingen. Desuden er 97.5%-fraktilen markeret, det vil sige punktet, hvor der ligger 97.5% sandsynlighed til venstre for og 2.5% sandsynlighed til højre for. Kan du på forhånd gætte, om  $t$ -fraktilerne ligger til højre eller til venstre for den tilsvarende fraktil i en standard normalfordeling? Fraktiler i en  $t(df)$ -fordeling findes i R med kommandoen `qt(p, df)`.

#### Kodevindue

```
df1=1; df2=2; df3=10
x=5*c(-100:100)/100
plot(x, pt(x, df1), type="l", ylim=c(0, 1), ylab="Fordelingsfunktion",
main="Sort: df=1, Rød: df=2, grøn: df=10, blå:N(0,1)")
lines(x, pt(x, df2), col=2)
lines(x, pt(x, df3), col=3)
lines(x, pnorm(x), col=4)
abline(h=0.975, lty=3)
lines(rep(qt(0.975, df2), 2), c(0.975, 2), col=2)
lines(rep(qt(0.975, df3), 2), c(0.975, 2), col=3)
lines(rep(qnorm(0.975), 2), c(0.975, 2), col=4)
```

- Hvad er 97.5%-fraktilen i en  $t(1)$ -fordeling, henholdsvis i en  $t(100)$ -fordeling?

#### Svar 2.3. T-fraktiler



#### Resultat 2.4. ( $t$ -test)

Betræt modellen (2.2) med  $\mu = \mu_0$ . I denne model er  $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$ . Hvis alternativet til hypotesen  $\mu = \mu_0$  er  $\mu \neq \mu_0$  udregner vi  $p$ -værdien for testet som

$$p\text{-værdi} = P(|T| \geq |t|) = 2 \cdot (1 - t_{\text{cdf}}(|t|, n-1)).$$

Desuden er et 95%-konfidensinterval for middelværdien  $\mu$  givet ved

$$\left[ \bar{x} - t_0 \frac{s}{\sqrt{n}}, \bar{x} + t_0 \frac{s}{\sqrt{n}} \right],$$

hvor  $t_0$  er 97.5%-fraktilen i en  $t(n-1)$ -fordeling,  $t_0 = t_{\text{inv}}(0.975, n-1)$ . Konfidensintervallet skrives ofte på kort form som  $\bar{x} \pm t_0 \frac{s}{\sqrt{n}}$

For at forstå at  $T$  følger en  $t(n - 1)$ -fordeling, skal vi blot skrive

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{(\bar{X} - \mu_0)/(\sigma/\sqrt{n})}{\sqrt{s^2/\sigma^2}},$$

og bruge definitionen på en  $t$ -fordeling med  $U = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$  og  $V = (n - 1)s^2/\sigma^2$ . Fordelingsresultaterne i Resultat 2.2 giver nu det ønskede.

Benyttes nu at  $T \sim t(n - 1)$ , og at  $t$ -fordelingen er symmetrisk omkring nul (hvilket følger af, at standard normalfordelingen er symmetrisk omkring nul), fås

$$\begin{aligned} P_\mu\left(\bar{X} - t_0 \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_0 \frac{s}{\sqrt{n}}\right) &= P_\mu\left(-t_0 \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq t_0\right) \\ &= P(T \leq t_0) - P(T \leq -t_0) = 0.975 - 0.025 = 0.95. \end{aligned}$$

## 2.5 Kontrol af køkkenvægt

Jeg vender tilbage til data omkring kontrol af min køkkenvægt i eksempel 2.1. Der er foretaget 10 uafhængige målinger af vægten af cirka 600 ml vand. Vi lader  $v_i$  være den  $i$ 'te måling af vægten og benytter modellen (2.2), med  $X_i$  erstattet af  $V_i$ :

$$\text{Model: } V_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, 10, \quad (\mu, \sigma^2) \in \mathbf{R} \times \mathbf{R}_+.$$

Ønsket er at teste hypotesen  $\mu = 600$  mod alternativet  $\mu \neq 600$ . Fra R-beregninger nedenfor fremgår, at  $\bar{v} = 581.9$ , og den empiriske spredning er  $s = 15.022$ . Herudfra beregnes  $t$ -teststørrelsen  $t = (581.9 - 600)/(15.022/\sqrt{10}) = -3.81$ .  $P$ -værdien i  $t$ -testet findes som  $p$ -værdi =  $2(1 - t_{\text{cdf}}(3.81, 9)) = 0.0042$ . Da  $p$ -værdien er langt under 0.05, konkluderer vi, at data strider mod hypotesen om en middelværdi på 600. Jeg må derfor erkende, at det måske er på tide at skifte min køkkenvægt ud!

Lad os også lave et 95%-konfidensinterval for middelværdien, som kan regnes om til et konfidensinterval for vægtens fejlvisning. Ved opslag finder vi, at 97.5%-fraktilen i en  $t(9)$ -fordeling er  $t_0 = 2.2622$ . Dermed bliver konfidensintervallet for middelværdien af målingerne  $581.9 \pm 2.2622 \cdot 15.022/\sqrt{10} = [571.2, 592.6]$ . Konfidensintervallet for middelværdien af fejlmålingen er dermed  $[571.2 - 600, 592.6 - 600] = [-28.8, -7.4]$ . Køkkenvægten viser altså *med 95% sikkerhed* et sted mellem 7 og 29 gram for lidt. Bemærk udtrykket *med 95% sikkerhed*, som blot er en anden måde at sige, at det er et 95%-konfidensinterval. Man må ikke sige *med 95% sandsynlighed*, hvilket giver indtryk af, at parameteren er stokastisk.

I kodevinduet nedenfor er vist de nødvendige beregninger i R (i afsnit 4.7 omtales en funktion i R, der kan lave alle beregningerne).

### Kodevindue

```
v600=c(579,583,586,601,576,559,609,572,567,587)
me=mean(v600)
s=sd(v600)
t=(me-600)/(s/sqrt(10))
```

```
pval=2*(1-pt(abs(t),9))
t0=qt(0.975,9)
KI=me+c(-1,1)*t0*s/sqrt(10)
c(gennemsnit=me, spred=s, t=t, pværdi=pval, tfrakttil=t0,
lower=KI[1], upper=KI[2])
```

### Showhide: Spørgsmål

1. Hvad beregner R-funktionerne *mean* og *sd*?
2. Hvilken *t*-fordeling benyttes i beregningen af  $t_0$ ?

#### Svar 2.4. Beregning af t-test



## 2.6 Konfidensinterval for varians og spredning

I normalfordelingsmodellen,  $X_i \sim N(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ , lavede vi ovenfor inferens om middelværdien, og spredningen  $\sigma$  var blot en nødvendig del af metoden. Spredningen kan imidlertid også være af interesse i sig selv. I mit eget dataeksempel, hvor jeg ønskede at kontrollere om mueslipakkerne levede op til en specifikation på 600 gram, er det ikke nok kun at se på middelværdien. Selvom middelværdien er 600 gram, er det ikke tilfredstillende, hvis jeg for eksempel kan få pakker med 500 gram eller 700 gram. I en medicinsk sammenhæng, hvor et nyt præparat testes, er spredningen i respons også vigtig. Mere generelt vil man ved undersøgelse af en population ofte også være interesseret i spredningen. Jeg vil her give et konfidensinterval for variansen  $\sigma^2$  og for spredningen  $\sigma$ .

Konfidensintervallerne baserer sig på Resultat 2.2, hvoraf det fremgår, at  $(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$ . Jeg betragter her en lidt mere generel situation for at kunne bruge resultatet i andre modeller. Jeg minder om, at fraktiler i en  $\chi^2(df)$ -fordeling betegnes med  $\chi_{\text{inv}}^2(p, df)$ .

#### Resultat 2.5. (Konfidensinterval for varians)

Lad  $s^2$  være en stokastisk variabel med tilknyttet antal frihedsgrader  $df$ , og hvor  $df \cdot s^2/\sigma^2 \sim \chi^2(df)$ . Så er et 95%-konfidensinterval for variansen  $\sigma^2$  givet ved

$$\left[ \frac{df \cdot s^2}{\chi_{\text{inv}}^2(0.975, df)}, \frac{df \cdot s^2}{\chi_{\text{inv}}^2(0.025, df)} \right],$$

og et 95%-konfidensinterval for spredningen  $\sigma$  er givet ved

$$\left[ \sqrt{\frac{df \cdot s^2}{\chi_{\text{inv}}^2(0.975, df)}}, \sqrt{\frac{df \cdot s^2}{\chi_{\text{inv}}^2(0.025, df)}} \right].$$

Det første resultat følger af, at

$$\begin{aligned} P_\sigma \left( \frac{df \cdot s^2}{\chi_{\text{inv}}^2(0.975, df)} \leq \sigma^2 \leq \frac{df \cdot s^2}{\chi_{\text{inv}}^2(0.025, df)} \right) \\ = P_\sigma \left( \chi_{\text{inv}}^2(0.025, df) \leq \frac{df \cdot s^2}{\sigma^2} \leq \chi_{\text{inv}}^2(0.975, df) \right) = 0.975 - 0.025 = 0.95. \end{aligned}$$

Det andet resultat følger ud fra en generel observation af, at hvis  $[\hat{\theta}_-, \hat{\theta}_+]$  er et 95%-konfidensinterval for en parameter  $\theta$ , så er  $[h(\hat{\theta}_-), h(\hat{\theta}_+)]$  et 95%-konfidensinterval for den transformerede parameter  $h(\theta)$ , hvor  $h$  er en voksende funktion. Dette ses af  $P(h(\hat{\theta}_-) \leq h(\theta) \leq h(\hat{\theta}_+)) = P(\hat{\theta}_- \leq \theta \leq \hat{\theta}_+)$ .

### Eksempel 2.6. (Kontrol af køkkenvægt)

Dette er en fortsættelse af Eksempel 2.1 og afsnit 2.5 med 10 uafhængige målinger af vægten af cirka 600 ml vand. Den empiriske spredning af de 10 målinger er  $s = 15.022$ . Ved opslag i tabel ses, at 0.025-fraktilen i en  $\chi^2(9)$ -fordeling er 2.700, og 0.975-fraktilen er 19.023. Et 95%-konfidensinterval for spredningen bliver derfor

$$\left[ \sqrt{\frac{9 \cdot 15.022^2}{19.023}}, \sqrt{\frac{9 \cdot s^2}{2.700}} \right] = [10.3, 27.4].$$

Beregningerne i R ser ud som følger.

#### Kodevindue

```
v600=c(579,583,586,601,576,559,609,572,567,587)
s=sd(v600)
df=length(v600)-1
sqrt(df*s^2/qchisq(c(0.975,0.025),df))
```

## 2.7 One sample $t$ -test i R

I afsnit 2.5 omkring kontrol af køkkenvægt blev  $t$ -testet for hypotese om middelværdien lavet ved at bruge R som en lommeregner. R har dog også en indbygget funktion beregnet til at lave dette test, nemlig funktionen  $t.test$ . Hvis data ligger i en vektor  $x$ , og vi ønsker at teste hypotesen, at middelværdien er for eksempel 9.81, bliver kaldet

```
t.test(x, mu=9.81)
```

Hvis  $\mu$  ikke specificres i kaldet, laves der et test for, at middelværdien er nul. Hvis output fra  $t.test$  placeres i  $TtestUD$ , vil denne blandt andet indeholde følgende:

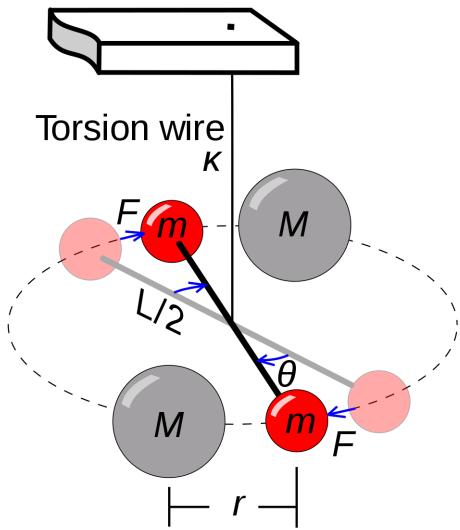
<code>TtestUD\$statistic</code>	$t$ -teststørrelsen
<code>TtestUD\$par</code>	frihedsgrader i $t$ -fordelingen
<code>TtestUD\$p.value</code>	$p$ -værdi for testet
<code>TtestUD\$conf.int</code>	95%-konfidensinterval for middelværdien

Prøv at gå tilbage til afsnit 2.5 omkring kontrol af køkkenvægt, og genfind værdierne beregnet der ved at bruge  $t.test$  i stedet.

### Svar 2.5. Brug t.test

#### Eksempel 2.7. (Cavendishs måling af jordens massetæthed)

I 1797 lavede Henry Cavendish en række eksperimenter for at måle jordens massetæthed.



I Cavendishs eksperiment bestemmes gravitationskonstanten i Newtons lov ud fra den kraft, hvormed to blykugler tiltrækker hinanden. Jordens masse bestemmes ved at sammenholde med tyngdeaccelerationen.

I kodevinduet nedenfor er 23 af Cavendishs målingerne gengivet. Den anerkendte værdi i dag for jordens massetæthed er 5.517. I eksemplet her undersøges, om "Cavendish målte rigtigt".

Lad massetaet<sub>i</sub> være den  $i$ 'te måling. Vi benytter modellen

$$\text{Model: Massetaet}_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, 23, \quad (\mu, \sigma) \in \mathbf{R} \times \mathbf{R}_+.$$

Middelværdien  $\mu$  repræsenterer her den massetæthed, som måles gennem Cavendishs eksperimenter. Der kan være fejl i den eksperimentelle opsætning, hvorfor det ikke er sikkert, at  $\mu$  er lig med den anerkendte værdi i dag. I kodevinduet nedenfor undersøges hypotesen  $\mu = 5.517$  mod alternativet at  $\mu \neq 5.517$ .

**Kodevindue**

```
massetaet=c(5.36,5.44,5.63,5.29,5.34,5.34,5.58,5.79,5.46,5.65,  
5.10,5.30,5.57,5.27,5.75,5.53,5.39,5.68,5.62,5.42,5.85,5.29,5.47)  
t.test(massetaet, mu=5.517)
```

**Showhide: Spørgsmål**

1. Aflæs skøn over middelværdien  $\mu$ .
2. Kan det antages, at middelværdien er  $\mu = 5.517$ ?
3. Angiv et 95%-konfidensinterval for middelværdien  $\mu$ .
4. Hvor mange frihedsgrader har  $t$ -fordelingen, der bruges i konstruktionen af konfidensintervallet?
5. Hvad beregnes i følgende kode:

```
(length(massetaet)-1)*var(massetaet)/qchisq(c(0.975,0.025),length(massetaet)-1)
```

**Svar 2.6. Cavendish**

## 2.8 Two-sample datasæt og boxplot

I artiklen [How the Horned Lizard Got Its Horns](#) beskrives data omkring længden af horn på den hornde tudseøgle for to populationer. Der er 30 målingerne fra tudseøgler dræbt af den amerikanske tornskade, hvor man har fundet tudseøglen ophængt på torne eller pigtråd, og 154 målingerne fra tilfældigt indfangede levende tudseøgler.



Forfatterne ønsker at sammenligne de to populationer og bruge en eventuel forskel til at diskutere selektion for lange horn. Data er indskrevet i kodevinduet nedenfor (enheden er millimeter), hvor der laves et qqplot af de to datasæt og en figur med boxplots. Bemærk koden til at lave flere qqplots i den samme figur.

Et *boxplot* består i midten af en kasse defineret ud fra de tre tal ( $q_{25}$ ,  $q_{50}$ ,  $q_{75}$ ), hvor  $q_{25}$  er den værdi, for hvilken 25% af dataværdierne ligger under og 75% ligger over værdien, og med en tilsvarende definition for  $q_{50}$  og  $q_{75}$  (**R** bruger lineær interpolation til beregning af disse). Den midterste værdi,  $q_{50}$ , kaldes *medianen* af data, og  $q_{75} - q_{25}$  kaldes *inter quartile range (IQR)*. Kassen går fra  $q_{25}$  til  $q_{75}$ , og medianen angives som en vandret streg inde i kassen. Over og under kassen er lavet en streg, der går fra kassen til den observation, der ligger længst væk fra kassen, men indenfor en afstand af  $1.5 \cdot \text{IQR}$  fra kassen. Endelig markeres de punkter, der ligger endnu længere væk fra kassen. Prøv at køre boxplot med datasættet 1, 2, 3, 4, 5, 6, 7, 13. Her er  $q_{25} = 2.5$ ,  $q_{50} = 4.5$ ,  $q_{75} = 6.5$ , værdierne 1 og 7 er punkterne længst væk, men indenfor  $1.5 \cdot \text{IQR} = 6$  fra kassen, og 13 ligger udenfor. I skal benytte kommandoen `boxplot(x)`, hvor  $x$  er en vektor med data. Boxplottet giver en meget simpel måde grafisk at sammenligne flere datasæt, som vist i det følgende kodevindue med data omtalt ovenfor (det afsluttende "`c()`" skal ikke med, når I kører **R** på jeres egen computer).

### Kodevindue

```
doede=c(21.4,23.9,23.2,22.6,22.5,19.3,23.5,23.4,19.0,21.7,20.2,26.7,21.7,
21.0,23.9,24.6,21.6,25.3,25.0,25.2,15.2,22.9,21.4,23.9,17.2,15.5,22.0,
22.0,23.1,20.7)
levende=c(25.2,26.9,26.6,25.6,25.7,25.9,27.3,25.1,30.3,25.6,26.0,24.6,
25.6,25.3,23.5,24.5,23.3,26.0,23.9,27.3,25.4,25.5,21.4,23.8,25.5,19.2,
20.7,19.2,25.5,20.5,20.6,24.9,23.7,23.4,25.6,26.6,27.7,25.7,27.0,26.5,
25.0,19.7,27.1,23.0,22.7,25.8,28.8,23.5,23.2,25.3,23.8,25.1,26.7,27.1,
22.5,25.5,24.4,25.6,20.6,25.5,24.5,23.0,27.6,27.3,20.7,26.0,25.1,19.9,
23.5,24.8,22.4,24.7,27.5,21.5,24.0,22.4,21.3,25.0,24.3,24.5,23.6,21.1,
25.6,26.1,23.7,24.2,23.2,26.5,28.1,24.1,29.5,24.0,26.8,25.8,23.3,22.4,
24.2,26.1,23.2,21.7,24.7,21.4,18.5,23.2,21.5,29.1,21.7,23.5,23.0,20.0,
21.9,27.4,27.1,23.1,23.8,26.5,27.0,26.4,26.3,20.6,24.5,22.8,25.5,25.3,
28.0,20.8,25.9,24.0,22.5,26.3,23.3,24.5,21.6,23.6,23.0,22.4,27.4,25.0,
```

```
23.9, 28.2, 27.2, 13.1, 22.9, 26.6, 25.8, 26.3, 20.9, 25.6, 24.8, 19.2, 27.4, 24.2,
15.7, 17.7)
```

```
# To qqplots i samme figur
par(mfrow=c(1,2))
qqnorm(doede, ylim=range(doede, levende))
points(qqnorm(levende, plot=FALSE), col=2, pch=20)
legend("topleft", legend=c(1,2), col=c(1,2), pch=c(1,20))

boxplot(doede, levende, names=c("Død", "Lev"))
c()
```

Prøv, i kaldet til boxplot ovenfor, at tilføje horizontal=TRUE. Ser data normalfordelt ud? Hvad viser boxplottene om forholdene mellem de to populationer?

**Svar 2.7. QQplot**

## 2.9 Two-sample: Model og estimation

Jeg beskriver nu en generel situation med  $n_1$  uafhængige målinger fra gruppe 1 (population 1) og  $n_2$  uafhængige målinger fra gruppe 2 (population 2). Målingerne betegnes med  $x_{ji}$ ,  $j = 1, 2$ ,  $i = 1, \dots, n_j$ . Målingerne i gruppe 1 er således  $x_{11}, x_{12}, \dots, x_{1n_1}$ , og fra gruppe 2 er målingerne  $x_{21}, x_{22}, \dots, x_{2n_2}$ . Vi betragter den statistiske model  $M_0$  beskrevet som

$$M_0: X_{ji} \sim N(\mu_j, \sigma_j^2), \quad j = 1, 2, \quad i = 1, \dots, n_j, \quad (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \in \mathbf{R}^2 \times \mathbf{R}_+^2. \quad (2.3)$$

Denne model siger, at data er normalfordelt, og hver gruppe har sin egen middelværdi og sin egen varians. Vi skal også betragte undermodellen  $M_1$ , hvor hver gruppe har sin egen middelværdi, men begge grupper har samme varians,

$$M_1: X_{ji} \sim N(\mu_j, \sigma^2), \quad j = 1, 2, \quad i = 1, \dots, n_j, \quad (\mu_1, \mu_2, \sigma^2) \in \mathbf{R}^2 \times \mathbf{R}_+. \quad (2.4)$$

### 2.9.1 Estimation i model $M_0$

For at estimere i model  $M_0$  kan man blot bruge resultaterne fra afsnit 2.3 på hver gruppe for sig. Dette giver skønnene

$$\hat{\mu}_1 = \bar{X}_1 \sim N(\mu_1, \sigma_1^2/n_1), \quad s_1^2 = \frac{1}{n_1 - 1} \sum_i (X_{1i} - \bar{X}_1)^2 \sim \sigma_1^2 \chi^2(n_1 - 1)/(n_1 - 1),$$

$$\hat{\mu}_2 = \bar{X}_2 \sim N(\mu_2, \sigma_2^2/n_1), \quad s_2^2 = \frac{1}{n_2-1} \sum_i (X_{2i} - \bar{X}_2)^2 \sim \sigma_2^2 \chi^2(n_2-1)/(n_2-1),$$

Her er  $\bar{X}_1 = \sum_i X_{1i}/n_1$  gennemsnit i gruppe 1 og  $\bar{X}_2 = \sum_i X_{2i}/n_2$  gennemsnit i gruppe 2.

## 2.9.2 Estimation i model $M_1$

Skøn over middelværdierne  $\mu_1$  og  $\mu_2$  bliver som i model  $M_0$ :

$$\hat{\mu}_1 = \bar{X}_1 \sim N(\mu_1, \sigma^2/n_1), \quad \hat{\mu}_2 = \bar{X}_2 \sim N(\mu_2, \sigma^2/n_2).$$

Som skøn over den fælles varians  $\sigma^2$  bruges

$$s^2 = \frac{1}{n_1+n_2-2} \left( \sum_i (X_{1i} - \bar{X}_1)^2 + \sum_i (X_{2i} - \bar{X}_2)^2 \right) \quad (2.5)$$

$$\sim \sigma^2 \chi^2(n_1+n_2-2)/(n_1+n_2-2). \quad (2.6)$$

At  $s^2$  følger en skaleret  $\chi^2$ -fordeling kan vises ud fra Resultat 2.2, som viser, at  $\sum_i (X_{ji} - \bar{X}_j)^2 \sim \sigma^2 \chi^2(n_j-1)$  og dermed at  $\sum_i (X_{1i} - \bar{X}_1)^2 + \sum_i (X_{2i} - \bar{X}_2)^2 \sim \sigma^2 \chi^2(n_1+n_2-2)$ .

## 2.10 Teste middelværdier ens når varianser er ens

### Resultat 2.8. (Test af ens middelværdier)

Til test af hypotesen  $\mu_1 = \mu_2$  i model  $M_1$  fra (2.4) med fælles varians i de to grupper bruges  $t$ -teststørrelsen

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t(n_1+n_2-2).$$

$P$ -værdien, når alternativet er  $\mu_1 \neq \mu_2$ , er  $2(1 - t_{\text{cdf}}(|t|, n_1+n_2-2))$ .

Endvidere er et 95%-konfidensinterval for forskel i middelværdi, det vil sige for parameteren  $\delta = \mu_1 - \mu_2$ , givet ved formlen

$$\bar{x}_1 - \bar{x}_2 \pm t_0 \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}, \quad t_0 = t_{\text{inv}}(0.975, n_1+n_2-2).$$

Disse resultater følger på samme måde som resultaterne i Resultat 2.4. Specielt benytter vi, at med  $\delta = \mu_1 - \mu_2$  er  $(\bar{X}_1 - \bar{X}_2 - \delta) \sim N(0, \sigma^2(1/n_1 + 1/n_2))$ , som sammen med fordelingen af  $s^2$  i formel (2.6) giver, at  $T \sim t(n_1+n_2-2)$ . Konfidensintervallet følger af  $P(-t_0 \leq (\bar{X}_1 - \bar{X}_2 - \delta)/\sqrt{s^2(1/n_1 + 1/n_2)} \leq t_0) = 0.95$ .

**Eksempel 2.9.** (Tudseøglens horn)

Vi vender tilbage til data omkring længden af tudseøglens horn beskrevet i afsnit 2.8. Lad  $Doe_{i,i}$  og  $Levende_i$  betegne de tilhørende stokastiske variable, og betragt den statistiske model

$$\begin{aligned} Doe_{i,i} &\sim N(\mu_1, \sigma^2), \quad i = 1, \dots, 30, \\ Levende_i &\sim N(\mu_2, \sigma^2), \quad i = 1, \dots, 154, \end{aligned}$$

hvor  $(\mu_1, \mu_2, \sigma^2)$  kan variere frit.

Fra data udregnes de følgende størrelser (beregningerne er lavet i kodevinduet længere fremme),

$$\begin{aligned} \overline{Doe_{i,i}} &= 21.9867, & s_{Doe_{i,i}} &= 2.7095, \\ \overline{Levende_i} &= 24.2812, & s_{Levende_i} &= 2.6308, \\ s^2 &= 6.9880, & s &= 2.6435, \end{aligned}$$

hvor  $s^2$  er det fælles variansskøn fra (2.6). Biologer ønsker at undersøge, om data peger på en forskel i de to populationer. Dette kan formuleres som hypotesen, at de to middelværdier er ens,  $\mu_1 = \mu_2$ , og vi ønsker ikke på forhånd at lægge os fast på et alternativ i en bestemt retning, således at alternativet er  $\mu_1 \neq \mu_2$ .  $T$ -teststørrelsen for denne hypotese bliver

$$t = \frac{21.9867 - 24.2812}{\sqrt{6.9880(1/30 + 1/154)}} = -4.3493,$$

og den tilhørende  $p$ -værdi fra en  $t(30 + 154 - 2)$ -fordeling er

$$p\text{-værdi} = 2 \cdot (0.0000114) = 0.0000227.$$

Da denne er meget mindre end 0.05, bliver konklusionen, at data strider mod samme middelværdi, og da  $\bar{x}_1 < \bar{x}_2$ , tyder data altså på, at de tudseøgler, der fanges af den amerikanske tornskade, i middel har mindre hornlængde end de levende tudseøgler.

Vi kan kvantificere forskellen i middelværdi mellem døde og levende tudseøgler ved at lave et 95%-konfidensinterval. Hertil bruges 97.5%-fraktilen i en  $t(182)$ -fordeling,  $t_0 = t_{inv}(0.975, 182) = 1.9731$ , og konfidensintervallet bliver

$$21.9867 - 24.2812 \pm 1.9731 \cdot \sqrt{6.9880(1/30 + 1/154)} = [-3.34, -1.25].$$

Med 95% sikkerthed ligger middelværdien af hornlængden for de døde tudseøgler mellem 1.25 og 3.34 millimeter under middelværdien for de levende.

Man taler somme tider om forskel i middelværdi divideret med spredning som standardiseret effektstørrelse, og en effektstørrelse større end 1 anses for stor. I tilfældet her er skønnet over den standardiserede effektstørrelse  $(24.2812 - 21.9867)/2.6435 = 0.87$ .

Forfatterne af artiklen med data i dette eksempel skriver til sidst: "clearly illustrates that defense against shrike predation is one factor driving the radical elongation of horns in some species of horned lizards". Andre biologer har kritiseret denne konklusion.

**Showhide: Beregninger i R**

**Kodevindue**

```

doede=c(21.4,23.9,23.2,22.6,22.5,19.3,23.5,23.4,19.0,21.7,20.2,26.7,21.7,
21.0,23.9,24.6,21.6,25.3,25.0,25.2,15.2,22.9,21.4,23.9,17.2,15.5,22.0,
22.0,23.1,20.7)
levende=c(25.2,26.9,26.6,25.6,25.7,25.9,27.3,25.1,30.3,25.6,26.0,24.6,
25.6,25.3,23.5,24.5,23.3,26.0,23.9,27.3,25.4,25.5,21.4,23.8,25.5,19.2,
20.7,19.2,25.5,20.5,20.6,24.9,23.7,23.4,25.6,26.6,27.7,25.7,27.0,26.5,
25.0,19.7,27.1,23.0,22.7,25.8,28.8,23.5,23.2,25.3,23.8,25.1,26.7,27.1,
22.5,25.5,24.4,25.6,20.6,25.5,24.5,23.0,27.6,27.3,20.7,26.0,25.1,19.9,
23.5,24.8,22.4,24.7,27.5,21.5,24.0,22.4,21.3,25.0,24.3,24.5,23.6,21.1,
25.6,26.1,23.7,24.2,23.2,26.5,28.1,24.1,29.5,24.0,26.8,25.8,23.3,22.4,
24.2,26.1,23.2,21.7,24.7,21.4,18.5,23.2,21.5,29.1,21.7,23.5,23.0,20.0,
21.9,27.4,27.1,23.1,23.8,26.5,27.0,26.4,26.3,20.6,24.5,22.8,25.5,25.3,
28.0,20.8,25.9,24.0,22.5,26.3,23.3,24.5,21.6,23.6,23.0,22.4,27.4,25.0,
23.9,28.2,27.2,13.1,22.9,26.6,25.8,26.3,20.9,25.6,24.8,19.2,27.4,24.2,
15.7,17.7)

n=c(length(doede),length(levende))
gns=c(mean(doede),mean(levende))
s=c(sd(doede),sd(levende))
s2=((n[1]-1)*var(doede)+(n[2]-1)*var(levende))/(n[1]+n[2]-2)
t=(gns[1]-gns[2])/sqrt(s2*(1/n[1]+1/n[2]))
pval=2*pt(-abs(t),n[1]+n[2]-2)
t0=qt(0.975,n[1]+n[2]-2)
KI=gns[1]-gns[2]+c(-1,1)*t0*sqrt(s2*(1/n[1]+1/n[2]))
list(Gennemsnit=gns, Spredning=s,
FaellesVarSpred=c(varians=s2, spredning=sqrt(s2)),
Test=c(t=t, pværdi=pval), KI=c(tfraktil=t0, lower=KI[1], upper=KI[2]))

```

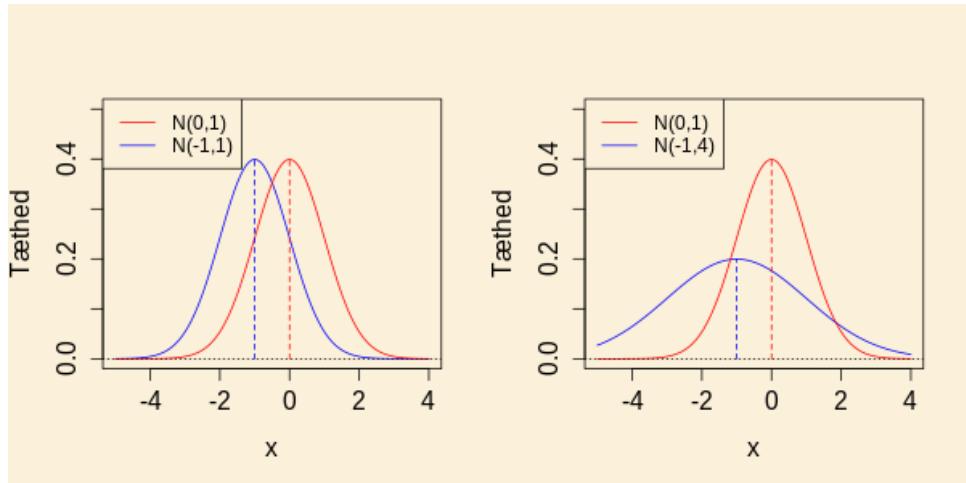


## 2.11 Teste middelværdier ens når varianser er forskellige

Vi betagter nu model  $M_0$  i (2.3) med to grupper af normalfordelte observationer med hver sin middelværdi og hver sin varians,  $X_{ji} \sim N(\mu_j, \sigma_j^2)$ . Vi ønsker stadig at teste de to middelværdier ens, det vil sige hypotesen  $\mu_1 = \mu_2$ .

Hvis middelværdierne er ens, er de to fordelinger centreret omkring det samme punkt, men den ene fordeling spreder sig ud over et større område end den anden fordeling. Alternativet, hvor middelværdierne er forskellige, skal man passe på, hvordan man fortolker. Hvis for eksempel  $\mu_1 <$

$\mu_2$ , vil man nemt danne sig det mentale billede, at observationerne fra gruppe 1 vil ligge under observationerne fra gruppe 2, men hvis samtidigt  $\sigma_1 > \sigma_2$ , vil der være et punkt  $x_0$ , således at for  $x > x_0$  vil der være større sandsynlighed i gruppe 1 end i gruppe 2 for at få en værdi over  $x$ . Som et konkret eksempel kan vi sige  $\sigma_1 = 2\sigma_2$  og  $\mu_1 = \mu_2 - \sigma_2$ , hvor der så gælder, at  $\mu_2 + \sigma_2 = \mu_1 + \sigma_1$ , hvilket er 85.1%-fraktilen i begge fordelinger. Den følgende figur med normalfordelingstætheder illustrerer eksemplet, hvor den højre del er situationen med forskellig varians.



Ligesom i konstruktionen af de to  $t$ -tests tidligere i dette kapitel starter vi med en standardisering:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1).$$

Da  $\sigma_1$  og  $\sigma_2$  ikke kendes, kan den standardiserede størrelse ikke bruges direkte, og i stedet erstattes  $\sigma_j^2$  med den empiriske varians  $s_j^2$ ,

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}. \quad (2.7)$$

I modsætning til tidligere er denne teststørrelse ikke  $t$ -fordelt på grund af de forskellige varianser. Faktisk afhænger fordelingen af  $T$  stadig af de ukendte varianser  $\sigma_1^2$  og  $\sigma_2^2$ . Man har dog kunnet vise matematisk, at som en approksimation kan man bruge en  $t$ -fordeling til beregning af  $p$ -værdi, hvor antallet af frihedsgrader i  $t$ -fordelingen afhænger af data gennem  $s_1^2$  og  $s_2^2$ .

### Resultat 2.10. (Welch's $t$ -test)

Til at teste hypotesen om ens middelværdier,  $\mu_1 = \mu_2$  mod alternativet  $\mu_1 \neq \mu_2$ , i model  $M_0$  hvor hver gruppe har sin egen varians, benyttes teststørrelsen  $T$  fra (2.7), og en approksimativ  $p$ -værdi beregnes som

$$p\text{-værdi} = 2(1 - pt(|t|, df_W)), \quad df_W = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Man kan vise, at  $df_W$  altid er større end eller lig med det mindste af de to frihedsgradsantal  $n_1 - 1$  og  $n_2 - 1$ , og mindre end eller lig med  $n_1 + n_2 - 2$ . Metoden givet ovenfor kaldes *Welch's t-test*.

Spørgmålet, om der findes et eksakt test for hypotesen  $\mu_1 = \mu_2$ , kendes under navnet **Brenrens–Fisher problem**.

### Eksempel 2.11. (Menneske-maskine interaktion)

Menneske-maskine interaktion ([human-computer interaction, HCI](#)) er et vigtigt område under informationsteknologi. I eksemplet her vil jeg se på tidsforbruget til at skrive en SMS. For en del år siden lavede en australsk skoleklasse et eksperiment, hvor en række personer blev bedt om at skrive en SMS med teksten [The quick brown fox jumps over the lazy dog](#). Teksten er et pangram, det vil sige, at alle alfabetets bogstaver optræder. Data var delt op på to aldersgrupper med en tydelig forskel i tidsforbrug mellem de to aldersgrupper. Desværre ser det ud til, at data ikke længere kan findes på nettet. Jeg vil i stedet bruge data inspireret af det australiske eksperiment, som jeg selv har indsamlet ved en forelæsning. Jeg brugte et dansk pangram: "Dansk jomfru på Ærø kyler halvsexet quizbog ned i wc", og delte op efter, om der blev brugt en smartphone med fuldt tastatur eller en "gammeldags" mobil, hvor hver knap koder for tre bogstaver (tripelkodende mobil).



Data i dette eksempel vedrører udelukkende kvindelige deltagere i eksperimentet. Der var 27 kvinder, der brugte en smartphone, og 33 der brugte en tripelkodende mobil. Data er givet i sekunder.

Vi lader  $Tid_{ji}$  være tidsforbrug for den  $i$ 'te kvinde i den  $j$ 'te gruppe (gruppe  $sm$ : smartphone, gruppe  $tr$ : tripelkodende mobil). Vi benytter modellen  $Tid_{ji} \sim N(\mu_j, \sigma_j^2)$ , idet data i qqplots snor sig nogenlunde om rette linjer (kør koden i kodevinduet nedenfor). Både qqplots og boxplots peger i retning af forskellige varianser i de to populationer.

Vi ønsker med eksperimentet at undersøge, om de to teknologier er lige gode med hensyn til at skrive en SMS-besked, formulert som hypotesen at de to middelværdier er ens  $\mu_{sm} = \mu_{tr}$ .  $T$ -teststørrelsen for denne hypotese, når varianserne er forskellige, bliver

$$t = \frac{32.926 - 42.333}{\sqrt{45.456/27 + 172.417/33}} = -3.579, \quad df_w = \frac{\left(\frac{45.456}{27} + \frac{172.417}{33}\right)^2}{\left(\frac{45.456}{27}\right)^2 + \left(\frac{172.417}{33}\right)^2} = 49.61,$$

baseret på følgende beregnede værdier

$$\overline{Tid}_{sm} = 32.926,$$

$$s_{sm}^2 = 45.456,$$

$$\bar{Tid}_{tr} = 42.333, \quad s_{tr}^2 = 172.417.$$

Den tilhørende  $p$ -værdi fra en  $t(49.61)$ -fordeling er

$$p\text{-værdi} = 2 \cdot (0.000391) = 0.000782.$$

Da denne er meget mindre 0.05, bliver konklusionen, at data strider mod samme middelværdi, og da  $\bar{Tid}_{sm} < \bar{Tid}_{tr}$ , tyder data altså på, at det er hurtigere at skrive en besked på en smartphone fremfor en tripelkodende mobil.

Forskellen i tidsforbruget ved brug af de to mobiltyper kan angives ved et 95%-konfidensinterval for forskellen i middelværdi. Hertil finder vi 97.5%-fraktilen i en  $t(49.61)$ -fordeling,  $t_0 = t_{inv}(0.975, 49.61) = 2.0090$ , og konfidensintervallet bliver

$$32.926 - 42.333 \pm 2.0090 \cdot \sqrt{45.456/27 + 172.417/33} = [-14.69, -4.13].$$

Middelværdien af tidsforbruget ved brug af smartphone ligger mellem 4.13 og 14.69 sekunder under middelværdien ved brug af en tripelkodende mobil med 95% sikkerthed.

### Showhide: Beregninger i R

Nedenstående kodevindue laver qqplots og boxplots af de to datasæt, og laver de beregnede værdier benyttet ovenfor.

#### Kodevindue

```
smart=c(34,33,24,31,35,35,42,47,35,28,42,43,32,23,31,38,
27,23,25,32,38,37,35,29,29,40,21)
tripel=c(80,79,46,50,27,31,23,27,35,45,33,30,28,43,53,39,
40,60,34,33,37,46,45,46,47,40,46,41,50,26,51,36,50)

par(mfrow=c(1,2))
qqnorm(smart, ylim=range(smart, tripel))
points(qqnorm(tripel, plot=FALSE), col=2, pch=20)
boxplot(smart, tripel, names=c("Smart", "Tripel"))

n=c(length(smart), length(tripel))
me=c(mean(smart), mean(tripel))
s2=c(var(smart), var(tripel))
dfw=(s2[1]/n[1]+s2[2]/n[2])^2/
((s2[1]/n[1])^2/(n[1]-1)+(s2[2]/n[2])^2/(n[2]-1))
t=(me[1]-me[2])/sqrt(s2[1]/n[1]+s2[2]/n[2])
pval=2*pt(-abs(t), dfw)
t0=qt(0.975, dfw)
KI=me[1]-me[2]+c(-1,1)*t0*sqrt(s2[1]/n[1]+s2[2]/n[2])
list(Gennemsnit=me, Varians=s2,
Test=c(t=t, frihedsgrader=dfw, pværdi=pval),
KI=c(tfraktil=t0, lower=KI[1], upper=KI[2]))
```

Kommenter, ud fra den dannede figur, på forholdet mellem de to varianser og forholdet mellem de to middelværdier.

**Svar 2.8.** [To datasæt](#)



## 2.12 Teste varianser ens

Jeg har ovenfor indført to tests af hypotesen om ens middelværdier i to normalfordelte populationer. Et test i situationen hvor varianserne er ens og et andet, når varianserne i de to grupper er forskellige. Hvorfor bruger vi to test i stedet for blot at nøjes med testet, hvor det ikke antages, at varianserne er ens? Svaret er, at hvis data ikke strider mod fælles varians, så får vi et stærkere test for hypotesen om samme middelværdi. Et stærkere test betyder, at man har nemmere ved at opdage en forskel i middelværdi, hvilket kan aflæses i, at konfidensintervallet for forskellen mellem de to middelværdier er smallere (97.5%-fraktilen i en  $t(df)$ -fordeling falder med antallet af frihedsgrader, og frihedsgraderne i tilfældet med forskellige varianser er  $df_W \leq n_1 + n_2 - 2$ ).

For at kunne afgøre hvilket af de to tests der skal brugs, skal man overveje, om de to varianser er ens. Vi betragter derfor hypotesen  $\sigma_1^2 = \sigma_2^2$  i model  $M_0$  i (2.3) med  $X_{ji} \sim N(\mu_j, \sigma_j^2)$ . Samme varians svarer i et qqplot af de to observationssæt til, at data snor sig om parallelle linjer. I et boxplot skal de to kasser være cirka lige store.

For at kunne bruge det test jeg nu vil indføre i andre modelsammenhænge, betragter jeg en lidt mere generel situation. Antag, at vi har to uafhængige variansskøn

$$s_1^2 \sim \sigma_1^2 \chi^2(df_1)/df_1, \quad s_2^2 \sim \sigma_2^2 \chi^2(df_2)/df_2. \quad (2.8)$$

Situationen under model  $M_0$  i (2.3) svarer til  $df_1 = n_1 - 1$  og  $df_2 = n_2 - 1$ . For at teste hypotesen om samme varians  $\sigma_1^2 = \sigma_2^2$ , vil jeg benytte forholdet  $s_1^2/s_2^2$ , som bør være tæt på 1 under hypotesen. Da

$$\frac{s_1^2}{s_2^2} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

under hypotesen, vil fordelingen af  $s_1^2/s_2^2$  være fordelingen af  $V_1/V_2$ , hvor  $V_1$  og  $V_2$  er uafhængige og  $V_j \sim \chi^2(df_j)/df_j$ .

### Definition 2.12. ( $F$ -fordeling)

Lad  $V_1$  og  $V_2$  være uafhængige,  $V_1 \sim \chi^2(df_1)/df_1$  og  $V_2 \sim \chi^2(df_2)/df_2$ . Så siges  $V_1/V_2$  at følge en  $F$ -fordeling med  $df_1$  frihedsgrader i tæller og  $df_2$  frihedsgrader i nævner. Fordelingsfunktionen betegnes  $F_{cdf}(\cdot, df_1, df_2)$  og fraktiler betegnes  $F_{inv}(\cdot, df_1, df_2)$ . I R er de tilsvarende funktioner  $pf(\cdot, df_1, df_2)$  og  $qf(\cdot, df_1, df_2)$ .

**Showhide:** [F-fordeling i R](#)

I nedenstående kodevindue tegnes tætheden for en  $F(df_1, df_2)$ -fordeling, og 2.5% og 97.5% fraktilerne markeres. Prøv at køre koden med forskellige valg af frihedsgradsantallene  $df_1$  og  $df_2$ . Ved det test, der laves nedenfor, bliver 2.5% og 97.5% fraktilerne grænserne for, hvornår vi accepterer, og hvornår vi forkaster.

### Kodevindue

```
df1=10
df2=10
x=c(1:600)/100
plot(x,df(x,df1,df2),type="l")
abline(v=qf(c(0.025,0.975),df1,df2),col=2)
lines(c(1,1),c(-1,df(1,df1,df2)),col=4,lty=3)
```



Når man laver et test for hypotesen  $\sigma_1^2 = \sigma_2^2$  mod alternativet  $\sigma_1^2 \neq \sigma_2^2$ , er både store og små værdier (værdier langt fra 1) af  $s_1^2/s_2^2$  kritiske. Hvis derfor den obsererede værdi  $F_{\text{obs}}$  af  $s_1^2/s_2^2$  er større end 1, bruger vi som  $p$ -værdi 2 gange sandsynlighed for at få en værdi over  $F_{\text{obs}}$ , og hvis  $F_{\text{obs}}$  er mindre end 1, bruger vi 2 gange sandsynlighed for at få en værdi mindre end  $F_{\text{obs}}$ . Med andre ord siger vi, at der er lige så stor en sandsynlighed for kritiske værdier på den anden side af 1 som på den side af 1, hvor  $F_{\text{obs}}$  ligger.

### Resultat 2.13. (Teste to varianser ens)

For test af hypotesen  $\sigma_1^2 = \sigma_2^2$  mod  $\sigma_1^2 \neq \sigma_2^2$  i modellen (2.8) benyttes  $s_1^2/s_2^2 \sim F(df_1, df_2)$ , og  $p$ -værdi beregnes som

$$p\text{-værdi} = \begin{cases} 2(1 - F_{\text{cdf}}(F_{\text{obs}}, df_1, df_2)), & F_{\text{obs}} > 1, \\ 2F_{\text{cdf}}(F_{\text{obs}}, df_1, df_2), & F_{\text{obs}} < 1, \end{cases}$$

hvor  $F_{\text{obs}}$  er den obsererede værdi af  $s_1^2/s_2^2$ .

### Eksempel 2.14. (Menneske-maskine-interaktion)

Jeg vender tilbage til Eksempel 2.11 omkring tidsforbruget til at skrive en SMS-tekst på enten en smartphone eller en tripelkodende mobil. Vi fandt i eksemplet, at de to variansskøn er

$$s_{\text{sm}}^2 = 45.46, \quad df_{\text{sm}} = 27 - 1 = 26, \quad s_{\text{tr}}^2 = 172.42, \quad df_{\text{tr}} = 33 - 1 = 32.$$

Herudfra kan man beregne  $F$ -teststørrelsen for hypotesen om samme varians,  $\sigma_{\text{sm}}^2 = \sigma_{\text{tr}}^2$ ,

$$F = \frac{45.46}{172.42} = 0.264, \quad p\text{-værdi} = 2 \cdot F_{\text{cdf}}(0.264, 26, 32) = 0.00085.$$

Da

$p$ -værdien

er langt under 0.05, bliver konklusionen, at data strider mod samme varians ved brug af de to metoder til at skrive SMS-teksten: der er større varians under brug af den tripelkodende mobil.

## 2.13 Two sample tests i R

I alle eksemplerne ovenfor omkring to normalfordelte observationssæt er de forskellige tests lavet ved at bruge **R** som en lommeregner. **R** har dog også indbyggede funktioner beregnet til at lave disse tests.

### 2.13.1 Two samples: Teste varianser ens

I model  $M_0$  i (2.3) med to normalfordelte observationssæt kan man lave  $F$ -testet for hypotesen om ens varianser med **R**-funktionen *var.test*. Hvis data ligger i to vektorer  $x1$  og  $x2$  bliver kaldet

```
var.test(x1, x2)
```

I output kan man finde  $F$ -teststørrelsen (*statistic*), de tilhørende frihedsgradsantal (*parameter*), og  $p$ -værdien (*p.value*). Der angives også et 95%-konfidensinterval for *forholdet* mellem de to variansparametre  $\sigma_1^2/\sigma_2^2$  (dette har jeg ikke omtalt ovenfor). Gå nu tilbage til Eksempel 2.14, og find de beregnede værdier der i output fra et kald af *vartest*.

**Svar 2.9.** Bruge *var.test*

### 2.13.2 Two samples: Teste middelværdier ens

For at teste at middelværdierne er ens i to normalfordelinger, skal man enten bruge  $t$ -testet, hvis de to varianser er ens, eller også bruge Welchs test, hvis de to varianser ikke er ens. Begge de to tests udregnes med **R**-funktionen *t.test*. Hvis data ligger i to vektorer  $x1$  og  $x2$  bliver kaldet

```
t.test(x1, x2, var.equal=TRUE)    hvis de to varianser er ens,
```

```
t.test(x1, x2, var.equal=FALSE)   hvis de to varianser er forskellige,
```

Output indeholder  $t$ -tesstørrelsen (*statistic*), antallet af frihedsgrader (*parameter*) og  $p$ -værdien (*pvalue*) for test af hypotesen, om at de to middelværdier er ens. Desuden angives et 95%-konfidensinterval for forskellen i middelværdi (*conf.int*), det vil sige for parameteren  $\delta = \mu_1 - \mu_2$ . Gå nu tilbage til Eksempel 2.9 og Eksempel 2.11 og gentag beregningerne ved hjælp af *t.test*.

**Svar 2.10.** Bruge t.test

### 2.13.3 Eksempel: log-data

I eksemplerne 2.11 og 2.14 så vi, at tidsforbruget ved at skrive en SMS-tekst både havde større middelværdi og større varians ved brug af tripelkodende mobil i forhold til en smartphone. Dette er ikke helt atypisk, når data vedrører en positiv størrelse (her tidsforbrug). I sådanne situationer vil der ofte ske det, at hvis data logaritmetransformeres, vil der efterfølgende være varianshomogenitet.

Lad  $\log Tid_{ji}$  være logaritmen til tidsforbruget for den  $i$ 'te person i den  $j$ 'te gruppe ( $sm$ : smartphone,  $tr$ : tripelkodende mobil). Vi betragter modellen

$$\text{Model: } \log Tid_{ji} \sim N(\nu_j, \tau_j^2), \quad (\nu_{sm}, \nu_{tr}, \tau_{sm}, \tau_{tr}) \in \mathbf{R}^2 \times \mathbf{R}_+^2,$$

hvor  $\nu_j$  er middelværdien af logaritmen til tidsforbruget. Man kan matematisk vise sammenhængen  $\mu_j = \exp(\nu_j + \frac{1}{2}\tau_j^2)$ . I kodevinduet nedenfor laves der qqplots for de logaritmetransformerede data, og disse giver ikke anledning til at forkaste modellen.

Først undersøges hypotesen om samme varians i de to grupper for de logaritmetransformerede værdier. Beregningen er vist i kodevinduet nedenfor:  $F$ -teststørrelsen er 0.519, og  $p$ -værdien (to gange sandsynlighed for værdi mindre end 0.519) fra en  $F(26, 32)$ -fordeling er 0.090. Da  $p$ -værdien er over 0.05, siger vi, at data ikke strider mod samme varians på logaritmeskalaen.

I kodevinduet laves der også et 95%-konfidensinterval for forskel i middelværdi,  $\delta = \nu_{sm} - \nu_{tr}$ , under antagelsen om samme varians. Konfidensintervallet er baseret på  $t(58)$ -fordelingen, og bliver  $[-0.364, -0.095]$ .

Da vi har samme varians  $\tau_{sm}^2 = \tau_{tr}^2 = \tau^2$  på logaritmeskalaen, giver sammenhængen  $\mu_j = \exp(\nu_j + \frac{1}{2}\tau^2)$ , at

$$\frac{\mu_{sm}}{\mu_{tr}} = \frac{\exp(\nu_{sm} + \frac{1}{2}\tau^2)}{\exp(\nu_{tr} + \frac{1}{2}\tau^2)} = \exp(\nu_{sm} - \nu_{tr}).$$

Her står, at forholdet mellem middelværdierne på den oprindelige skala er exponentialfunktionen taget på differensen mellem middelværdierne på logaritmeskalaen. Vi kan derfor umiddelbart oversætte konfidensintervallet for forskellen  $\delta = \nu_{sm} - \nu_{tr}$  til et konfidensinterval for forholdet  $\mu_{sm}/\mu_{tr}$ . For data omkring tidsforbruget for at skrive en SMS-tekst giver dette intervallet  $[e^{-0.364}, e^{-0.095}] = [0.69, 0.91]$ . Her står, at med 95% sikkerhed er middelværdien for tidsforbruget med smartphone mellem 69% og 91% af middelværdien ved brug af tripelkodende mobil.

**Showhide: Beregninger i R**

#### Kodevindue

```
smart=c(34,33,24,31,35,35,42,47,35,28,42,43,32,23,31,38,27,23,25,32,38,
37,35,29,29,40,21)
```

```

tripel=c(80,79,46,50,27,31,23,27,35,45,33,30,28,43,53,39,40,60,34,33,37,
46,45,46,47,40,46,41,50,26,51,36,50)
logSmart=log(smart)
logTripel=log(tripel)

par(mfrow=c(1,2))
qqnorm(smart, ylim=range(smart, tripel), main="Ikke_Log")
points(qqnorm(tripel, plot=FALSE), col=2, pch=20)
qqnorm(logSmart, ylim=range(logSmart, logTripel), main="Log")
points(qqnorm(logTripel, plot=FALSE), col=2, pch=20)

list(vartest=var.test(logSmart, logTripel),
ttest=t.test(logSmart, logTripel, var.equal=TRUE))

```



## 2.14 Standard Error

For en stokastisk variabel  $X$  er *varians* defineret som  $\text{Var}(X) = E((X - E(X))^2)$ , og *spredning* eller standardafvigelsen er defineret som  $\text{sd}(X) = \sqrt{\text{Var}(X)}$ . Notationen her refererer til det engelske navn *standard deviation* for spredning. Udover ordet spredning har vi på dansk også ordet *usikkerhed*. I fysik bruges dette ord i forbindelse med spredningen på en måling fra et måleapperat.

For en statistisk model med en parameter  $\theta$ , og et skøn  $\hat{\theta}$  over denne, kan vi tale om spredningen på den stokastiske variabel  $\hat{\theta}$ ,  $\text{sd}(\hat{\theta})$ . For binomialmodellen  $X \sim \text{binom}(n, p)$  er spredningen på skønnet  $\hat{p} = X/n$  givet ved  $\text{sd}(\hat{p}) = \sqrt{p(1-p)/n}$ . For normalfordelingsmodellen  $X_i \sim N(\mu, \sigma^2)$ ,  $i = 1, \dots, n$ , har vi skønnet  $\hat{\mu} = \bar{X}$  med spredning  $\text{sd}(\hat{\mu}) = \sigma/\sqrt{n}$ . Som det ses, vil spredningen på et parameterskøn ofte indeholde ukendte parametre. Hvis spredningen skal bruges i en udregning, må vi derfor indsætte skøn over disse parametre. Den resulterende værdi kaldes *standard error* for parameterskønnet, og betegnes i denne bog med  $\text{sd}_s(\hat{\theta})$ , hvor det nedre fodtegn  $s$  står for "skøn over". For normalfordelingsmodellen har vi  $\text{sd}_s(\hat{\mu}) = s/\sqrt{n}$  med  $s^2 = \sum_i (X_i - \bar{X})^2/(n-1)$ .

Med indførslen af standard error kan man udtrykke konfidensintervallet baseret på  $t$ -fordelingen på simpel vis. Konfidensintervallet fra Resultat 2.4 bliver  $\hat{\mu} \pm t_0 \cdot \text{sd}_s(\hat{\mu})$ ,  $t_0 = t_{\text{inv}}(0.975, n-1)$ . Konfidensintervallet for forskellen mellem to middelværdier  $\delta = \mu_1 - \mu_2$  fra Resultat 2.8 bliver  $\hat{\delta} \pm t_0 \cdot \text{sd}_s(\hat{\delta})$ ,  $t_0 = t_{\text{inv}}(0.975, n_1 + n_2 - 2)$ , med  $\text{sd}_s(\hat{\delta}) = s\sqrt{1/n_1 + 1/n_2}$  og  $s^2$  det fælles varianskøn. I tilfældet med to normalfordelinger med forskellig varians bliver konfidensintervallet også  $\hat{\delta} \pm t_0 \cdot \text{sd}_s(\hat{\delta})$ , hvor nu  $t_0 = t_{\text{inv}}(0.975, df_w)$  og  $\text{sd}_s(\hat{\delta}) = \sqrt{s_1^2/n_1 + s_2^2/n_2}$  (Resultat 2.10).

I de følgende kapitler skal vi bruge funktionen *lm* i **R**, hvor standard error automatisk er en del af output.

## 2.15 Svar

### Svar 2.1. Fraktiler

Fraktilerne beregnes som `qnorm(0.95)` og `qnorm(0.05)`.

### Svar 2.2. Fordelinger

- Der simuleres først data fra en  $N(0,1)$ -fordeling, hvorefter disse ganges med 2 og der lægges 3 til. Dette betyder, at de nye data stammer fra en  $N(3, 2^2)$ -fordeling, middelværdi er 3 og spredning er 2.
- Data kommer fra en stokastisk variabel  $Y = e^X$ , hvor  $X$  er  $N(0,1)$ -fordelt. Hvis vi tager logaritmen, får vi  $\log(Y) = \log(e^X) = X$ , som er normalfordelt. Man siger, at  $Y$  er *log-normalfordelt*.

### Svar 2.3. T-fraktiler

- Fraktilen i en  $t(1)$ -fordeling findes med kommandoen `qt(0.975, 1)` og har værdien 12.7.

### Svar 2.4. Beregning af t-test

- Funktionen `mean(x)` beregner gennemsnit af værdierne i vektoren  $x$  og funktionen `sd(x)` beregner den empiriske spredning af værdierne i  $x$ .
- Frihedssgraderne er det andet argument i kaldet til `qt`, hvorfor det er en  $t(9)$ -fordeling, der bruges.

### Svar 2.5. Bruge t.test

#### Kodevindue

```
v600=c(579,583,586,601,576,559,609,572,567,587)
t.test(v600, mu=600)
```

Vi aflæser i output  $t$ -teststørrelsen til -3.81 og  $p$ -værdien til 0.0042. Konfidensintervallet aflæses til [571.2, 592.6].

### Svar 2.6. Cavendish

- Skønnet er  $\hat{\mu} = \bar{x} = 5.4835$  fra aflæsning i output.

2. I output ses at  $p$ -værdien fra  $t$ -testet er 0.4076. Da denne er langt over 0.05, strider data ikke mod hypotesen om  $\mu = 5.517$ . Cavendish har derfor lavet et eksperiment, der mäter "korrekt".
3. Fra output aflæses 95%-konfidensinterval for middelværdien: [5.40, 5.57]. Da  $p$ -værdien for test af hypotesen  $\mu = 5.517$  er over 0.05, ligger 5.517 i konfidensintervallet.
4. Antallet af frihedsgrader er  $n - 1$ , som er 22 ( $df$  i output).
5. Koden beregner et 95%-konfidensinterval for variansen  $\sigma^2$ .

### Svar 2.7. QQplot

De to qqplots ser ud til at sno sig påent omkring en ret linje, hvorfor vi siger, at data kan beskrives med normalfordelingen. De to linjer i qqplottene ser parallelle ud, hvilket er en indikation af samme varians i de to grupper af observationer (hældningen i et qqplot er  $1/\sigma$ , hvor sigma er spredningen).

I boxplottene er de to kasser cirka lige høje, hvilket igen er en indikation af samme varians i de to grupper. De to kasser er forskudt i forhold til hinanden, hvilket tyder på forskel i middelværdi for de to populationer. Medianen ligger cirka midt i de to kasser, hvilket tyder på en symmetrisk fordeling i overensstemmelse med en normalfordeling.

### Svar 2.8. To datasæt

Der ser ud til at være forskellig hældning i de to qqplots, hvilket tyder på forskellig varians i de to grupper af observationer. Den samme tendens ses i boxplottene: det højre boxplot har større udstrækning. Boxplottene indikerer også, at der er forskel i middelværdi i de to grupper: kassen i det højre boxplot ligger højere end kassen i det venstre boxplot.

### Svar 2.9. Bruge var.test

#### Kodevindue

```
smart=c(34,33,24,31,35,35,42,47,35,28,42,43,32,23,31,38,27,23,25,32,38,
37,35,29,29,40,21)
tripel=c(80,79,46,50,27,31,23,27,35,45,33,30,28,43,53,39,40,60,34,33,37,
46,45,46,47,40,46,41,50,26,51,36,50)
var.test(smart, tripel)
```

Vi aflæser i output at  $F$ -tesstørrelsen er 0.264 og  $p$ -værdien er 0.00084. Hvilken  $F$ -fordeling bruges til beregningen af  $p$ -værdien?

**Svar:** I output fra *var.test* aflæses, at der er 26 frihedsgrader i tæller og 32 frihedsgrader i nævner. Den anvendte fordeling er derfor en  $F(26, 32)$ -fordeling.

### Svar 2.10. Bruge t.test

**Kodevindue**

```
x1=c(21.4,23.9,23.2,22.6,22.5,19.3,23.5,23.4,19.0,21.7,20.2,26.7,21.7,
21.0,23.9,24.6,21.6,25.3,25.0,25.2,15.2,22.9,21.4,23.9,17.2,15.5,22.0,
22.0,23.1,20.7)
x2=c(25.2,26.9,26.6,25.6,25.7,25.9,27.3,25.1,30.3,25.6,26.0,24.6,25.6,
25.3,23.5,24.5,23.3,26.0,23.9,27.3,25.4,25.5,21.4,23.8,25.5,19.2,20.7,
19.2,25.5,20.5,20.6,24.9,23.7,23.4,25.6,26.6,27.7,25.7,27.0,26.5,25.0,
19.7,27.1,23.0,22.7,25.8,28.8,23.5,23.2,25.3,23.8,25.1,26.7,27.1,22.5,
25.5,24.4,25.6,20.6,25.5,24.5,23.0,27.6,27.3,20.7,26.0,25.1,19.9,23.5,
24.8,22.4,24.7,27.5,21.5,24.0,22.4,21.3,25.0,24.3,24.5,23.6,21.1,25.6,
26.1,23.7,24.2,23.2,26.5,28.1,24.1,29.5,24.0,26.8,25.8,23.3,22.4,24.2,
26.1,23.2,21.7,24.7,21.4,18.5,23.2,21.5,29.1,21.7,23.5,23.0,20.0,21.9,
27.4,27.1,23.1,23.8,26.5,27.0,26.4,26.3,20.6,24.5,22.8,25.5,25.3,28.0,
20.8,25.9,24.0,22.5,26.3,23.3,24.5,21.6,23.6,23.0,22.4,27.4,25.0,23.9,
28.2,27.2,13.1,22.9,26.6,25.8,26.3,20.9,25.6,24.8,19.2,27.4,24.2,15.7,
17.7)

t.test(x1,x2, var.equal=TRUE)
```

Vi aflæser her  $t$ -teststørrelsen til  $-4.35$ ,  $p$ -værdien fra en  $t(182)$ -fordeling er  $0.000023$ , og et 95%-konfidensinterval er  $[-3.34, -1.25]$ .

**Kodevindue**

```
smart=c(34,33,24,31,35,35,42,47,35,28,42,43,32,23,31,38,27,23,25,32,38,
37,35,29,29,40,21)
tripel=c(80,79,46,50,27,31,23,27,35,45,33,30,28,43,53,39,40,60,34,33,37,
46,45,46,47,40,46,41,50,26,51,36,50)

t.test(smart, tripel, var.equal=FALSE)
```

Vi aflæser her  $t$ -teststørrelsen til  $-3.58$ ,  $p$ -værdien fra en  $t(49.606)$ -fordeling er  $0.00078$ , og et 95%-konfidensinterval er  $[-14.7, -4.1]$ .

Hvordan kan du i output se, om du betragter modellen med fælles varians i de to normalfordelinger, eller modellen med forskellig varians?

**Svar:** Output starter med enten "Two Sample t-test" eller "Welch Two Sample t-test".

## 2.16 Opgaver til kapitel 2

Øvelserne hørende til kapitel 2 vedrører situationen med et enkelt normalfordelt observationssæt og situationen med to normalfordelte observationssæt. I skal lave grafiske undersøgelser i form

af histogram, qqplot og boxplot. For ét observationssæt skal I lave inferens om middelværdien og variansen i normalfordelingen. For to observationssæt skal I både sammenligne varianser og sammenligne middelværdier, og specielt lave konfidensinterval for forskel i middelværdi.

### Showhide: Opgave 2.1: Inferens om middelværdi

I bjergegne danner floder en [alluvialkegle](#) når de aflejrer sedimenter ved foden af bjerget. Disse kegler beskrives ofte som værende symmetriske, men forfatterne til artiklen [Interactions between alluvial fans and axial rivers in Yukon, Canada and Alaska, USA](#) sætter spørgsmålstege ved dette. For 63 alluvialkegler har forfatterne målt en længde af keglen i hver sin side og dannet forholdet mellem de to længder, kegleforholdet:  $F = L_D / L_U$ , hvor  $D$  og  $U$  står for downstream og upstream for vandløbet nedenfor keglen. En symmetrisk kegle svarer til at kegleforholdet har værdien 1. Målingerne findes i filen *Alluvialkegle.txt*.

Indlæs data fra filen *Alluvialkegle.txt* med kommandoen `scan("Alluvialkegle.txt")`. Denne opgave kan nu formuleres kort som følger. Opstil en statistisk model for kegleforholdet, lav inferens for parametrene i modellen og overvej en hypotese, om at keglerne er symmetriske. Skrevet ud bliver dette til følgende spørgsmål.

- Undersøg grafisk, om kegleforholdet kan beskrives med en normalfordeling via et histogram og et qqplot. Overvej om det er bedre at beskrive logaritmen til kegleforholdet med en normalfordeling. Opskriv en statistisk model for data.
- Lav en tabel med skøn og 95%-konfidensinterval for middelværdien, variansen og spredningen i en normalfordelingsmodel.
- Overvej, om data er i overenstemmelse med teorien om symmetriske kegler.

Uanset, om I beskriver de oprindelige kegleforhold eller logaritmen til disse, vil I finde, at spredningen er så stor, at et kegleforhold under 1 vil have en sandsynlighed på cirka 16% i den estimerede normalfordeling (overvej dette). Forfatterne diskuterer ud fra fluiddynamiske betragtninger, både hvorfor kegleforholdet ofte er større end 1, men også hvorfor værdier mindre end 1 kan forekomme.



### Showhide: Opgave 2.2: Parret t-test

I artiklen [Effect of metallic iron from grinding on ferrous iron determinations](#) måles indholdet af jern ( $\text{Fe}^0$ ) i en række klippestykker ved to målemetoder betegnet som  $\text{HgCl}_2$  og  $\text{CuCl}_2$ . I tabellen nedenfor er kun medtaget de prøver, hvor jernindholdet er under 0.05 procent.

Nummer	$\text{HgCl}_2$	$\text{CuCl}_2$	Differens	Nummer	$\text{HgCl}_2$	$\text{CuCl}_2$	Differens
1	0.025	0.030	0.005	7	0.003	0.016	0.013
2	0.022	0.031	0.009	8	0.009	0.018	0.009
3	0.014	0.019	0.005	9	0.008	0.012	0.004
4	0.001	0.000	-0.001	10	0.004	0.013	0.009
5	0.002	0.004	0.002	11	0.031	0.037	0.006
6	0.003	0.016	0.013	12	0.020	0.041	0.021

I denne opgave skal I ud fra differenserne mellem de to målinger angive den viden, vi har om en eventuel forskel mellem de to målemetoder. Middelværdien af differensen siger noget om,

hvilken generel tendens der er i forskellen mellem de to metoder, og spredningen repræsenterer den kombinerede måleusikkerhed fra de to målinger på den samme prøve. Data findes i filen *Jern.csv*, som er organiseret i 12 rækker og tre søjler: første søje angiver prøvenummer, anden søje angiver  $\text{HgCl}_2$ -målingen og tredje søje angiver  $\text{CuCl}_2$ -målingen.

- Indlæs data, og lav en figur, hvor indholdet af jern fra  $\text{CuCl}_2$  metoden tegnes op mod indholdet fra  $\text{HgCl}_2$  metoden. Indtegn identitetslinjen i figuren. Prøv at beskrive i ord, hvad figuren viser om forskel i jernindhold mellem de to målemetoder.
- Betrægt nu de 12 differenser mellem jernindhold fra de to målemetoder. Lav et qqplot af data, og opskriv den statistiske model, hvor differensen er normalfordelt.
- Lav et test for hypotesen, at middelværdien af differensen er nul, svarende til hypotesen, at der ikke er forskel mellem de to målemetoder. Lav dernæst et 95%-konfidensinterval for middelværdien af differensen. Hvad bliver konklusionen af disse udregninger?

Når I laver et *t*-test for at middelværdierne af *differenserne* er nul, kaldes dette et *parret t-test*: observationerne fra de to målemetoder er parret, ved at der er målt på det samme klippestykke. For en given målemetode er der stor variation i jernindholdet mellem klippestykkerne, og det kan være svært at se en forskel mellem to målemetoder, hvis vi forestiller os et alternativt eksperiment, hvor der er indsamlet 12 klippestykker, der analyseres med den ene målemetode, og 12 andre klippestykker der analyseres med den anden målemetode. I kan se dette ved at prøve at lave et two-sample *t*-test for data i denne opgave, hvor det ene observationssæt er data fra den ene målemetode, og det andet observationssæt er data fra den anden målemetode (two-sample *t*-test skal I arbejde med i den næste opgave).



### Showhide: Opgave 2.3: Two-sample *t*-test, samme varians

Gøgen lægger sine æg i andre fugles reder. I artiklen [The eggs of \*Cuculus canorus\*. An Inquiry into the dimensions of the cuckoo's egg and the relation of the variations to the size of the eggs of the foster-parent, with notes on coloration](#) undersøges det, om der er sket en selektion, således at gøgen er tilpasset den fugl, der bruges som vært for æggene. For de to værtsarter Engpiber og Hvid vipstjert er gøgens æg indsamlet, og bredden af ægget divideret med længden af ægget er beregnet (kaldet æggets *form* fremover). Data ligger i filen *Goegen.csv* i form af to søjler, hvor første søje er værtsart, og anden søje er æggets form.

- Indlæs data og dan vektorerne *Art* og *form* ud fra søjlerne i de indlæste data. Dan dernæst to datasæt *formEng* og *formVip* med værdierne fra *form* hørende til henholdsvis Engpibe og Vipstjert.

Lav en figur med et qqplot for hvert af de to datasæt. Koden, til at lave flere qqplots i den samme figur, kan du se i kodevinduet i afsnit 2.8. Synes du, at gøgeæggernes form for hver værtsart kan beskrives med en normalfordeling?

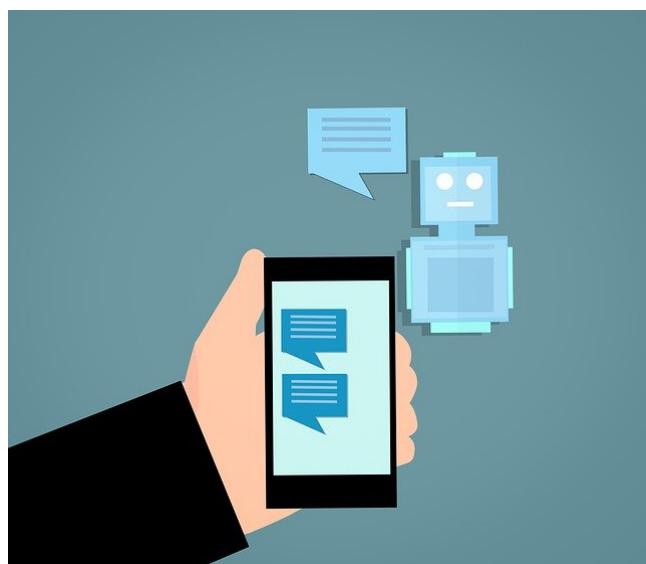
Lav også en figur med boxplot for hvert af de to datasæt. Flere boxplots i den samme figur kan laves som vist i kodevinduet i afsnit 2.8, men kan også laves med kommandoen `boxplot(form~Art)`. Hvilke ligheder og forskelle mellem de to datasæt kan du se i denne figur?

- (b) Opstil modellen, hvor hvert datasæt (*formEng* og *formVip*) følger sin egen normalfordeling. Opsummer de to datasæt i form af en tabel, som for hvert datasæt indeholder antallet af observationer, gennemsnit, empirisk spredning og et 95%-konfidensinterval for middelværdien. Antallet af elementer i en vektor kan i R findes med funktionen *length*.
- (c) Opskriv hypotesen, at de to varianser er ens, og lav *F*-testet for ens varianser. Er det rimeligt at antage, at variansen af æggets form er den samme for de to værtsarter?
- (d) Opstil nu modellen, hvor data er normalfordelt, og de to datasæt har hver sin middelværdi, men samme varians. Opstil hypotesen at de to middelværdier er ens, og lav et test af denne hypotese.
- Er det rimeligt at antage, at æggets form har samme middelværdi for de to værtsarter?
- (e) Angiv et 95%-konfidensinterval for forskellen i middelværdi af æggets form mellem værtsarten Engpiber og Hvid vipstjert.
- Synes du, at forskellen mellem de to middelværdier i denne opgave er stor (se begrebet *effektstørrelse* i eksempel 2.9)?



#### Showhide: Opgave 2.4: Two-sample *t*-test, forskellig varians

I artiklen [In the shades of the uncanny valley: An experimental study of human-chatbot interaction](#) undersøges, hvordan forsøgspersoner påvirkes af at interagere med en *chatbot*, enten en simpel tekst-chatbot eller en chatbot, der oplæser beskederne.



I artiklen betegnes de to situationer med *Text* og *Avatar*. Som et af målepunkterne i eksperimentet måles den gennemsnitlige puls af forsøgspersonerne: 16 personer i *Text*-gruppen og 15 personer i *Avatar*-gruppen. Data ligger i filen *Chatbot.csv* i form af to søjler, hvor første søjle angiver chatbotsituationen, og anden søjle er pulsen. Data i denne fil er simulerede på en sådan måde, at informationen i figur 8 i den ovennævnte artikel efterlignes.

- (a) Indlæs data fra filen *Chatbot.csv*. Lav to datasæt med puls svarende til grupperne *Text* og *Avatar*. Du skal i den samme figur lave et qqplot for begge datasæt.

Synes du, at pulsen for hver chatbotsituuation kan beskrives med en normalfordeling?

Lav en figur, der indeholder boxplot for de to chatbotsituationer. Hvilke ligheder og forskelle mellem de to datasæt kan du se i denne figur?

- (b) Opstil modellen, hvor hvert datasæt følger sin egen normalfordeling. Opsummer de to datasæt i form af en tabel, som for hvert datasæt indeholder antallet af observationer, gennemsnit, empirisk spredning og et 95%-konfidensinterval for middelværdien.
- (c) Opstil hypotesen, at de to varianser er ens. Eftervis, at data strider mod at sige, at variansen på pulsen er den samme for Text-gruppen som for Avatar-gruppen.
- (d) Angiv et 95%-konfidensinterval for forskellen i middelværdi af pulsen mellem Text-gruppen og Avatar-gruppen.

Synes du, at forskellen mellem de to middelværdier er stor?

- (e) Prøv til sidst at betragte logaritmen til pulsen. Lav qqplots for at se, om disse data kan beskrives med en normalfordeling. Lav et test, for at varianserne er ens, og lav et 95%-konfidensinterval for forskel i middelværdi af logaritmen til pulsen.

Oversæt det fundne konfidensinterval for forskel i middelværdi af logaritmen til pulsen til et 95%-konfidensinterval for forholdet mellem middelværdierne af pulsen, jævnfør underafsnit 2.13.3. Hvor mange gange større er middelværdien af pulsen for Avatar-gruppen i forhold til Text-gruppen?



#### Showhide: Opgave 2.5: “Standard error” kontra “standard deviation”

I skal i denne opgave lave en figur, der illustrerer standard deviation i forhold til standard error. Start med at dele plotvinduet op i to dele med ordren `par(mfrow=c(1, 2))`.

- (a) Simuler  $n = 20$  observationer  $x_1, \dots, x_{20}$  fra en standard normalfordeling (benyt `rnorm(20)` til dette). Beregn den empiriske spredning  $s$ , beregn skøn  $\hat{\mu} = \bar{x}$  over middelværdien og standard error for middelværdiskønnet,  $sd_s(\hat{\mu})$ .
- (b) Lav en figur med kaldet `boxplot(x, xlim=c(0, 3), ylim=c(-3, 3))`, hvor  $x$  er en vektor med de simulerede værdier.
- (c) Indsæt to lodrette linjestykker med yderpunkter henholdsvis  $\hat{\mu} \pm s$  og  $\hat{\mu} \pm s/\sqrt{n}$ . Disse skal placeres ud for 1.5 og 2.0 på førsteaksen. Dette kan gøres med funktionen `errorbar` omtalt i underafsnittet *Egne funktioner i R* i afsnit 1.9:

```
errorbar(c(1.5, 2.0), c(\hat{\mu}, \hat{\mu}), lower, upper)
```

hvor `lower=c(\hat{\mu}-s, \hat{\mu}-s/sqrt(n))`, og `upper` er tilsvarende med plus i stedet for minus. Indsæt endelig et vandret linjestykke til at markere værdien af  $\hat{\mu}$ .

- (d) Gentag ovenstående simulering og tegning med  $n = 200$  observationer. Hvilke dele skal ligge hinanden i de to tegninger, og hvilke skal ikke?



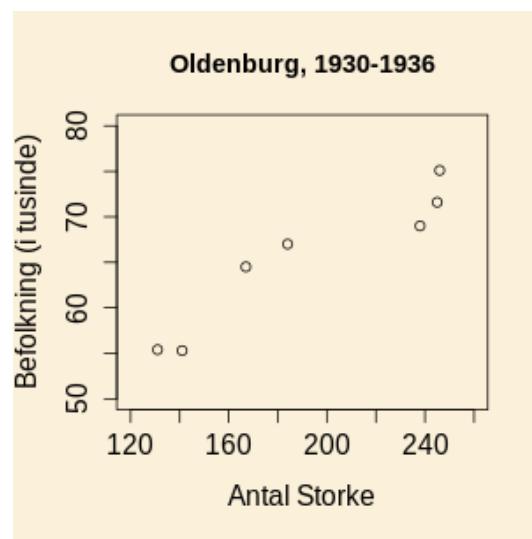
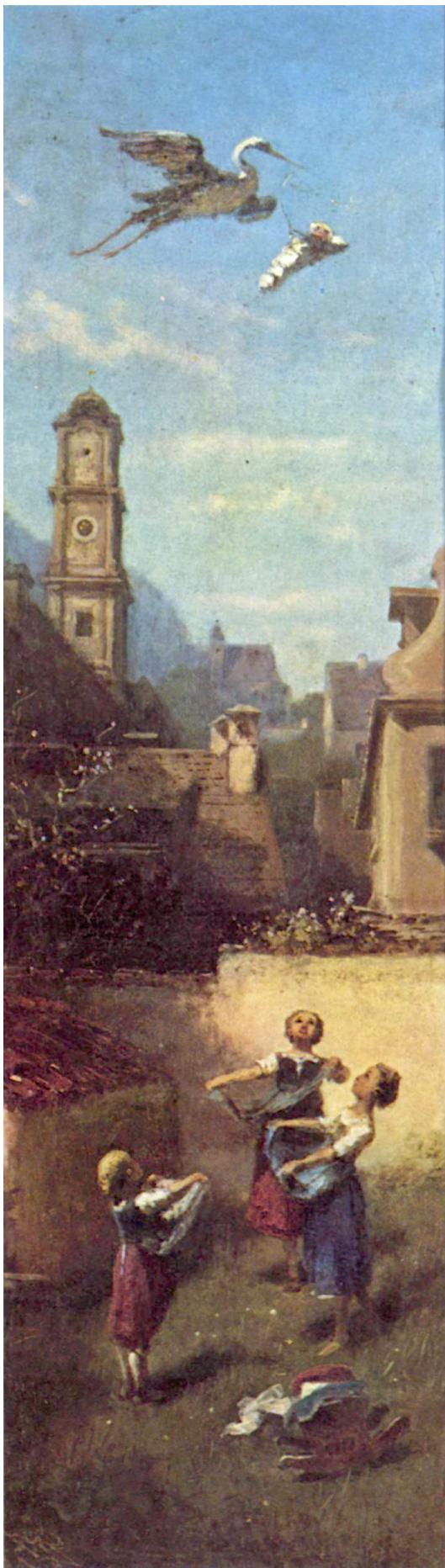
# Lineær regression

Mange undersøgelser går ud på at beskrive en sammenhæng mellem to variable, og i dette kapitel skal vi se på den (simple) lineære regressionsmodel til dette. Modellen indføres i afsnit 3.1 og analyseres i afsnit 3.2.

En sammenhæng kan være *kausal* forstået på den måde, at værdien af den ene variabel er årsag til værdien af den anden variabel. Dette kan illustreres med en fysisk lov som for eksempel Lambert-Beers lov. Denne siger, at mængden af lys, der absorberes ved passage af en glasbeholder med en opløsning af et bestemt stof, er proportional med koncentrationen af stoffet i opløsningen. Når man siger, at rygning øger risikoen for lungekræft, er der også tale om kausalitet, dog vedrører det kausale forhold ikke en større eller mindre grad af lungekræft hos den enkelte person, men sandsynligheden for at få lungekræft i en population.

En helt anden type sammenhæng er, når vi siger, at jo højere man er, jo mere vejer man. For det første kan man jo ikke ændre på højden af en person, så i udsagnet ligger der implicit, at det er forskellige personer der sammenlignes. For det andet er sammenhængen ikke kausal, selvom en person er 5 centimeter højere end en anden person, er det langt fra givet, at den højeste person vejer mest. Man taler i stedet for om en biologisk samvariation af de to variable højde og vægt og siger, at de to variable er korrelerede. Sammenhængen består i, at hvis vi ser på middelvægten for alle af en bestemt højde, så vil denne vokse med højden. Alt efter hvilke sammenhænge der betragtes, kan variationen omkring den linære sammenhæng være større eller mindre med en tilsvarende mindre eller større mulighed for at sige noget om den ene variabel ud fra den anden.

Når man kigger efter, om de to variable er korrelerede, skal man passe på, at det ikke er en falsk sammenhæng, man betragter (spurious correlation). Selvom to variable ikke har noget med hinanden at gøre, så kan de begge have en sammenhæng med en tredje variabel, og det er variationen i denne, der gør, at man ser en tilsyneladende sammenhæng i de to andre variable. Figuren nedenfor viser et eksempel med antallet af indbyggere i Oldenburg i perioden 1930-1936 og antallet af storke, med data aflæst fra figur i [Statistics for Experimenters: Design, Innovation, and Discovery](#), (oprindelsen af data kan findes samme sted). Maleriet, der er tilføjet til venstre i følgende figur, er "Der Klapperstorch" af [Carl Spitzweg](#).



Moralen er, at man skal passe på med at kigge "i blinde" efter sammenhænge. En sammenhæng skal helst være underbygget af en videnskabelig forståelse af emneområdet.

En lineær sammenhæng beskrives ved to parametre, nemlig hældning og skæring. Test og konfidensintervaller for disse parametre beskrives i afsnit 3.3, og brug af den estimerede linje til prædiktion og kalibrering omtales i afsnit 3.5. Analyse af data foretages i R via funktionen *lm*, der beskrives i afsnit 3.4, samt via funktionen *predict*, der omtales i afsnit 3.5.

### 3.1 Model for lineær regression

Jeg vil nu formulere den statistiske model, der er emnet for dette kapitel. Udgangspunktet er, at vi har  $n$  uafhængige par af observationer  $(t_i, x_i)$ ,  $i = 1, \dots, n$ , og ønsker at sige noget om, hvordan  $x$  afhænger af  $t$ . Med denne formulering tænker jeg på  $x$  som *responsvariabel* og på  $t$  som den *forklarende variabel*.

#### Eksempel 3.1. (Forurening i vandprøver)

Måling af mængden af E.coli bakterier i vandprøver bruges som mål for forureningsgraden af prøven. Målemetoden er langsom, og man er derfor interesseret i alternative målemetoder. Data i dette eksempel vedrører brugen af GLUase-aktivitet som et alternativ. Data er aflæst fra figur i artiklen [Rapid enzymatic detection of Escherichia coli contamination in polluted river water](#) og består af 98 sammenhørende værdier af variablene *logColi* og *logGlu* med henholdsvis logaritmen til mængden af E.coli (logaritmen til cfu E.coli per 100ml) og logaritmen til GLUase aktiviteten (logaritmen til NM MUF per minut). Data er fra vandprøver fra forskellige Østrigske floder indsamlet gennem 1998 og 1999. Følgende figur viser *logGlu* afsat mod *logColi*.

#### Showhide: Figur med *logGlu* afsat mod *logColi*

#### Kodevindue

```
logColi=c(2.80,2.86,2.88,2.86,3.07,3.10,3.19,3.33,3.32,3.32,3.41,3.26,  
3.28,3.35,3.34,3.33,3.40,3.47,3.44,3.50,3.36,3.61,3.65,3.75,3.72,  
3.72,3.74,3.79,3.84,3.84,3.82,3.88,3.85,3.81,3.87,3.87,3.88,3.98,  
3.97,4.02,4.04,4.06,4.23,4.52,4.28,4.35,4.43,4.52,4.36,4.68,4.67,  
4.88,4.98,4.88,5.02,5.14,5.23,5.24,5.35,5.33,5.31,5.39,5.43,5.44,  
5.48,5.51,5.53,5.56,5.64,5.66,5.73,5.79,5.90,5.83,5.86,5.85,5.79,  
5.77,5.99,6.11,6.33,6.32,6.34,6.19,6.29,6.74,6.89,5.49,5.50,5.52,  
5.55,5.79,5.79,4.00,3.79,3.32,3.38,2.83)  
logGlu=c(-1.05,-1.04,-1.10,-1.21,-1.28,-1.57,-0.56,-0.54,-0.59,-0.68,  
-0.77,-1.12,-1.02,-1.32,-1.28,-1.16,-1.10,-1.07,-0.99,-0.85,-1.05,  
-1.06,-0.78,-0.82,-0.90,-1.06,-1.17,-1.04,-1.03,-1.07,-0.26,-0.32,  
-0.54,-0.61,-0.68,-0.75,-0.80,-0.79,-0.65,-0.71,-0.66,-0.59,-0.79,
```

```
-0.86,-0.53,-0.34,-0.33,-0.40,-0.10,0.32,0.21,0.04,-0.29,0.37,  
0.45,0.56,0.14,0.89,0.78,0.89,1.01,1.08,1.22,1.10,1.15,1.08,1.04,  
0.97,1.12,1.35,1.49,1.28,1.09,1.06,1.00,0.83,0.77,0.64,1.56,1.73,  
1.79,1.42,1.22,1.03,1.01,1.83,1.53,1.12,1.06,1.06,0.99,0.67,0.70,  
-0.75,-1.08,-1.06,-1.01,-1.05)
```

```
plot(logColi, logGlu)
```



Som et første kig på data vil man typisk lave en figur, hvor respons  $x_i$  *afsættes mod* den forklarende værdi  $t_i$ , det vil sige, at  $t$ -værdierne er ud ad førsteaksen og  $x$ -værdierne op ad andenaksen i figuren.

Man tænker på værdierne  $t_i$  af den forklarende variabel som faste, og værdierne  $x_i$  af responsvariablen som udfald af de stokastiske variable  $X_1, \dots, X_n$ . Den sammenhæng, vi vil formulere, vedrører middelværdien af respons som funktion af den forklarende variabel. Vi betragter udelukkende en lineær sammenhæng givet som

$$E(X_i) = \alpha + \beta t_i.$$

Her er  $\beta$  hældningen i den lineære sammenhæng: en stigning på 1 i  $t$  medfører en stigning på  $\beta$  i middelværdien, og  $\alpha$  er skæringen med andenaksen. Uover middelværdispecifikationen forlanger vi, at alle de stokastiske variable har samme varians,  $\sigma^2$ , og endelig antages, at respons er normalfordelt.

### Definition 3.2. (Den lineære regressionsmodel)

I den lineære regresionsmodel har vi  $n$  faste tal  $t_1, \dots, t_n$  og  $n$  uafhængige stokastiske variable  $X_1, \dots, X_n$  med  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ . Variationsområdet for  $(\alpha, \beta, \sigma^2)$  er  $\mathbf{R} \times \mathbf{R} \times \mathbf{R}_+$ .

Vores mål er at lave inferens om parametrene i modellen. Nedenfor finder jeg skøn over parametrene og laver  $t$ -test for parametrene i middelværdispecifikationen.

Jeg bruger i denne bog betegnelserne *forklarende variabel* og *responsvariabel*. Man kan også støde på betegnelserne *uafhængige variabel* og *afhængige variabel*, som på engelsk er *independent* og *dependent variable*.

## 3.2 Estimation i den lineære regresionsmodel

For at lave skøn over  $\alpha$ ,  $\beta$  og  $\sigma^2$  i modellen  $M_0$ :  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ , opstiller vi som i afsnit 2.3 likelihoodfunktionen og maksimere denne. Da de  $n$  målinger antages uafhængige, er

likelihoodfunktionen produktet af tæthedener,

$$L(\alpha, \beta, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i - \alpha - \beta t_i)^2} = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \alpha - \beta t_i)^2}.$$

For at finde maksimum af denne med hensyn til  $\alpha$  og  $\beta$  skal vi minimere

$$SSD(\alpha, \beta) = \sum_{i=1}^n (x_i - \alpha - \beta t_i)^2,$$

hvorfor dette kaldes *mindste kvadraters metode* (som i afsnit 2.3).

### Showhide: Udledning af estimerter

Vi differentierer  $SSD(\alpha, \beta)$  med hensyn til  $\alpha$  og  $\beta$  og sætter de aflede lig med nul. Efter division med  $-2$  giver dette ligningerne

$$\sum_{i=1}^n (x_i - \hat{\alpha} - \hat{\beta} t_i) = 0 \quad \text{og} \quad \sum_{i=1}^n t_i (x_i - \hat{\alpha} - \hat{\beta} t_i) = 0.$$

Isolerer vi  $\hat{\alpha}$  i den første ligning får vi  $\hat{\alpha} = \bar{x} - \hat{\beta} \bar{t}$ , hvor  $\bar{x}$  og  $\bar{t}$  er gennemsnit. Indsættes dette i den anden ligning fås

$$\sum_{i=1}^n t_i (x_i - \bar{x} - \hat{\beta} (t_i - \bar{t})) = 0,$$

og løsningen til denne er

$$\hat{\beta} = \frac{\sum_{i=1}^n t_i (x_i - \bar{x})}{\sum_{i=1}^n t_i (t_i - \bar{t})}.$$

Da imidlertid  $\sum_i \bar{t} (x_i - \bar{x}) = 0$  og  $\sum_i \bar{t} (t_i - \bar{t}) = 0$ , kan vi skrive  $\hat{\beta}$  på følgende mere symmetriske måde

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(t_i - \bar{t})}{\sum_{i=1}^n (t_i - \bar{t})^2}.$$

At det fundne punkt er et minimumspunkt for  $SSD(\alpha, \beta)$  følger af, at  $SSD(\alpha, \beta)$  går mod uendelig, når  $\alpha$  og  $\beta$  går mod uendelig.



Som vist i det skjulte punkt ovenfor er skøn over hældning  $\beta$  og skæring  $\alpha$  givet ved

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(t_i - \bar{t})}{\sum_{i=1}^n (t_i - \bar{t})^2}, \quad \hat{\alpha} = \bar{x} - \hat{\beta} \bar{t}, \quad (3.1)$$

hvor de to gennemsnit er  $\bar{t} = \sum_i t_i / n$  og  $\bar{x} = \sum_i x_i / n$ . Det vil være bekvemt at bruge følgende notation

$$SSD_t = \sum_{i=1}^n (t_i - \bar{t})^2 \quad \text{og} \quad SSD(M_0) = \sum_{i=1}^n (x_i - (\hat{\alpha} + \hat{\beta} t_i))^2,$$

hvor  $SSD$  står for *Sum of Squared Deviations*. Værdien  $\hat{\xi}_i = \hat{\alpha} + \hat{\beta} t_i$  kaldes den *i'te forventede værdi* (middelværdien med parameterskøn indsat), og

$$r_i = x_i - \hat{\xi}_i$$

kaldes det *i'te residual*.

Indsættes  $\hat{\alpha}$  og  $\hat{\beta}$  i  $L(\alpha, \beta, \sigma^2)$ , og maksimeres med hensyn til  $\sigma^2$ , fås  $\hat{\sigma}^2 = \frac{1}{n} SSD(M_0)$ . Ligesom i afsnit 2.3 ændrer vi divisor her og bruger skønnnet  $s_r^2$  givet ved

$$s_r^2 = \frac{SSD(M_0)}{df(M_0)}, \quad df(M_0) = n - 2.$$

På denne måde opnås, at  $s_r^2$ , betragtet som en stokastisk variabel, har middelværdi  $\sigma^2$  (dette om-tales ofte som, at variansskønnnet er *unbiased*). Nedre indeks "r" står for *regression*.

I Eksempel 3.4 nedenfor laves en figur med data fra Eksempel 3.1, med den estimerede linje indtegnet og med to parallelle linjer i afstanden  $\pm 2s_r$ .

For at kunne bruge skønnene til at lave inferens om parametrene skal vi kende fordelingen af de tilhørende stokastiske variable.

### Resultat 3.3. (Fordeling af skøn i lineær regressionsmodel)

I modellen  $M_0$ :  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ , uafhængige, gælder der, at

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{SSD_t}\right), \quad \hat{\alpha} \sim N\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\bar{t}^2}{SSD_t}\right)\right), \quad s_r^2 \sim \sigma^2 \chi^2(n-2)/(n-2),$$

og  $s_r^2$  er uafhængig af  $(\hat{\alpha}, \hat{\beta})$ .

Jeg vil her kort forklare det første resultat, da det peger frem mod et generelt resultat i næste kapitel. Fra (3.1) kan vi skrive

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(t_i - \bar{t})}{SSD_t} = \frac{\sum_{i=1}^n X_i(t_i - \bar{t})}{SSD_t} = \sum_{i=1}^n X_i a_i, \quad a_i = \frac{t_i - \bar{t}}{SSD_t}.$$

Her er  $a_i$ -erne faste tal (ikke-stokastiske), og vi betragter altså en linearkombination af uafhængige normalfordelte variable. Fra regnereglerne i afsnit 2.1 gælder der således, at  $\hat{\beta}$  er normalfordelt. Vi skal nu blot eftervise, at middelværdi og varians er som angivet i Resultat 3.3. Dette følger af, at

$$\begin{aligned} E(\hat{\beta}) &= \sum_i (\alpha + \beta t_i) a_i = \sum_i (\alpha + \beta t_i) \frac{t_i - \bar{t}}{SSD_t} = 0 + \beta \sum_i t_i \frac{t_i - \bar{t}}{SSD_t} \\ &= \beta \sum_i (t_i - \bar{t}) \frac{t_i - \bar{t}}{SSD_t} = \beta, \end{aligned}$$

og

$$\text{Var}(\hat{\beta}) = \sigma^2 \sum_i a_i^2 = \sigma^2 \sum_i \frac{(t_i - \bar{t})^2}{SSD_t^2} = \sigma^2 \frac{SSD_t}{SSD_t^2} = \frac{\sigma^2}{SSD_t}.$$

Fordelingen for  $\hat{\alpha}$  findes på helt tilsvarende vis. For  $s_r^2$  gælder der, at  $\sum_i (x_i - (\alpha + \beta t_i))^2 \sim \sigma^2 \chi^2(n)$ , og man kan matematisk vise, at når vi indsætter  $\hat{\alpha}$  og  $\hat{\beta}$ , mister vi to frihedegrader (der estimeres to parametre), så  $\sum_i (x_i - (\hat{\alpha} + \hat{\beta} t_i))^2 \sim \sigma^2 \chi^2(n-2)$ .

### Showhide: Uafhængighed

Jeg vil her kort indikere, at  $(\hat{\alpha}, \hat{\beta})$  og  $s_r^2$  er uafhængige. Den simultane tæthed for  $X_1, \dots, X_n$  er på formen  $(2\pi\sigma^2)^{-n/2} \exp(-SSD(\alpha, \beta)/(2\sigma^2))$ . Idet vi antager, at alle  $t_i$ -erne ikke er ens, kan vi

uden tab af generalitet sige, at  $t_1 \neq t_2$ . Jeg vil transformere  $X_1, \dots, X_n$  til  $\hat{\alpha}, \hat{\beta}, R_3, \dots, R_n$  med  $R_i = X_i - \hat{\alpha} - \hat{\beta}t_i$  det  $i$ 'te residual. Jeg vil vise at  $(\hat{\alpha}, \hat{\beta})$  og  $(R_3, \dots, R_n)$  er uafhængige, og at  $s_r^2$  er en funktion af  $(R_3, \dots, R_n)$ . Da transformationen er lineær i  $X_1, \dots, X_n$ , afhænger jakobianten i den transformerede tæthed ikke af data, og har dermed ikke betydning for argumentet om uafhængighed.

For at vise uafhængigheden skal jeg vise, at  $SSD(\alpha, \beta)$  kan skrives som en sum, hvor det ene led kun afhænger af  $(R_3, \dots, R_n)$  og det andet led kun af  $(\hat{\alpha}, \hat{\beta})$ . Først skriver jeg

$$\begin{aligned} SSD(\alpha, \beta) &= \sum_{i=1}^n (X_i - \alpha - \beta t_i)^2 = \sum_{i=1}^n (R_i + (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) t_i)^2 \\ &= \sum_{i=1}^n R_i^2 + \sum_{i=1}^n ((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) t_i)^2, \end{aligned}$$

hvor jeg i det sidste lighedstegn har brugt  $\sum_i R_i = 0$  og  $\sum_i t_i R_i = 0$ , som er ligningerne, der bruges til at finde  $\hat{\alpha}$  og  $\hat{\beta}$  i det skjulte punkt ovenfor. De samme to ligninger kan også bruges til at indse, at  $R_1$  og  $R_2$  kan skrives som en lineær funktion af  $\sum_{i=3}^n R_i$  og  $\sum_{i=3}^n t_i R_i$ . Således har vi, at det første led i summen ovenfor kun afhænger af  $(R_3, \dots, R_n)$ , og vi har etableret uafhængigheden. Da  $s_r^2 = \sum_{i=1}^n R_i^2 / (n - 2)$ , overføres uafhængigheden til uafhængighed mellem  $(\hat{\alpha}, \hat{\beta})$  og  $s_r^2$ .



### 3.2.1 Modelkontrol

For at vurdere, om den lineære sammenhæng giver en god beskrivelse af data, laver man ofte et *residualplot*. I denne figur afsættes residualerne  $r_i$  mod de forklarende værdier  $t_i$ . Man kigger efter to ting. For det første om der er systematiske afvigelser fra nulllinjen, altså om der er områder, hvor de fleste af residualerne enten ligger over eller ligger under nulllinjen. Dette vil være et udtryk for, at sammenhængen er mere kompliceret end blot en lineær sammenhæng. For det andet kigger man efter, om der er områder, hvor residualerne spredes sig mere end i andre områder. Dette vil pege mod, at antagelsen om den samme varians  $\sigma^2$  på alle observationerne ikke er korrekt. Man kan også lave et qqplot af residualerne  $r_1, \dots, r_n$  for at vurdere, om normalfordelingsantagelsen er rimelig.

#### Eksempel 3.4. (Forurening i vandprøver)

Vi fortsætter med data fra Eksempel 3.1. Data beskrives med modellen

$$\text{LogGlu}_i \sim N(\alpha + \beta \cdot \log\text{Coli}_i, \sigma^2), \quad i = 1, \dots, 98, \quad (\alpha, \beta, \sigma) \in \mathbf{R}^2 \times \mathbf{R}_+.$$

Fra formlerne ovenfor fås  $\hat{\beta} = 0.8494$ ,  $\hat{\alpha} = -3.8872$  og  $s_r = 0.3094$ . Først laver vi en figur med regressionslinjen indtegnet og to parallelle linjer i afstanden  $\pm 2s_r$ .

**Showhide:** Figur med logGlu afsat mod logColi

**Kodevindue**

```

logColi=c(2.80,2.86,2.88,2.86,3.07,3.10,3.19,3.33,3.32,3.32,3.41,3.26,
3.28,3.35,3.34,3.33,3.40,3.47,3.44,3.50,3.36,3.61,3.65,3.75,3.72,
3.72,3.74,3.79,3.84,3.84,3.82,3.88,3.85,3.81,3.87,3.87,3.88,3.98,
3.97,4.02,4.04,4.06,4.23,4.52,4.28,4.35,4.43,4.52,4.36,4.68,4.67,
4.88,4.98,4.88,5.02,5.14,5.23,5.24,5.35,5.33,5.31,5.39,5.43,5.44,
5.48,5.51,5.53,5.56,5.64,5.66,5.73,5.79,5.90,5.83,5.86,5.85,5.79,
5.77,5.99,6.11,6.33,6.32,6.34,6.19,6.29,6.74,6.89,5.49,5.50,5.52,
5.55,5.79,5.79,4.00,3.79,3.32,3.38,2.83)
logGlu=c(-1.05,-1.04,-1.10,-1.21,-1.28,-1.57,-0.56,-0.54,-0.59,-0.68,
-0.77,-1.12,-1.02,-1.32,-1.28,-1.16,-1.10,-1.07,-0.99,-0.85,-1.05,
-1.06,-0.78,-0.82,-0.90,-1.06,-1.17,-1.04,-1.03,-1.07,-0.26,-0.32,
-0.54,-0.61,-0.68,-0.75,-0.80,-0.79,-0.65,-0.71,-0.66,-0.59,-0.79,
-0.86,-0.53,-0.34,-0.33,-0.40,-0.10,0.32,0.21,0.04,-0.29,0.37,
0.45,0.56,0.14,0.89,0.78,0.89,1.01,1.08,1.22,1.10,1.15,1.08,1.04,
0.97,1.12,1.35,1.49,1.28,1.09,1.06,1.00,0.83,0.77,0.64,1.56,1.73,
1.79,1.42,1.22,1.03,1.01,1.83,1.53,1.12,1.06,1.06,0.99,0.67,0.70,
-0.75,-1.08,-1.06,-1.01,-1.05)

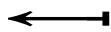
betahat=sum((logGlu-mean(logGlu))*(logColi-mean(logColi)))/
sum((logColi-mean(logColi))^2)
alphahat=mean(logGlu)-betahat*mean(logColi)
sr=sqrt(sum((logGlu-(alphahat+betahat*logColi))^2)/(98-2))

plot(logColi,logGlu)
abline(-3.8872,0.8494)
abline(-3.8872+2*0.3094,0.8494,lty=3)
abline(-3.8872-2*0.3094,0.8494,lty=3)
c(betahat=betahat, alphahat=alphahat, spredningsskøn=sr)

```

Hvor mange punkter forventer du ligger udenfor båndene givet ved de stribede linjer?

**Svar 3.1. Punkter udenfor**



Dernæst laver vi en figur med residualplot og en figur med qqplot af residualerne.

**Showhide: Figur med residualplot**

**Kodevindue**

```

logColi=c(2.80,2.86,2.88,2.86,3.07,3.10,3.19,3.33,3.32,3.32,3.41,3.26,
3.28,3.35,3.34,3.33,3.40,3.47,3.44,3.50,3.36,3.61,3.65,3.75,3.72,
3.72,3.74,3.79,3.84,3.84,3.82,3.88,3.85,3.81,3.87,3.87,3.88,3.98,

```

```

3.97,4.02,4.04,4.06,4.23,4.52,4.28,4.35,4.43,4.52,4.36,4.68,4.67,
4.88,4.98,4.88,5.02,5.14,5.23,5.24,5.35,5.33,5.31,5.39,5.43,5.44,
5.48,5.51,5.53,5.56,5.64,5.66,5.73,5.79,5.90,5.83,5.86,5.85,5.79,
5.77,5.99,6.11,6.33,6.32,6.34,6.19,6.29,6.74,6.89,5.49,5.50,5.52,
5.55,5.79,5.79,4.00,3.79,3.32,3.38,2.83)
logGlu=c(-1.05,-1.04,-1.10,-1.21,-1.28,-1.57,-0.56,-0.54,-0.59,-0.68,
-0.77,-1.12,-1.02,-1.32,-1.28,-1.16,-1.10,-1.07,-0.99,-0.85,-1.05,
-1.06,-0.78,-0.82,-0.90,-1.06,-1.17,-1.04,-1.03,-1.07,-0.26,-0.32,
-0.54,-0.61,-0.68,-0.75,-0.80,-0.79,-0.65,-0.71,-0.66,-0.59,-0.79,
-0.86,-0.53,-0.34,-0.33,-0.40,-0.10,0.32,0.21,0.04,-0.29,0.37,
0.45,0.56,0.14,0.89,0.78,0.89,1.01,1.08,1.22,1.10,1.15,1.08,1.04,
0.97,1.12,1.35,1.49,1.28,1.09,1.06,1.00,0.83,0.77,0.64,1.56,1.73,
1.79,1.42,1.22,1.03,1.01,1.83,1.53,1.12,1.06,1.06,0.99,0.67,0.70,
-0.75,-1.08,-1.06,-1.01,-1.05)

r=logGlu-(-3.8872+0.8495*logColi)
par(mfrow=c(1,2))
plot(logColi,r)
abline(0,0)
qqnorm(r)
c()

```

Residualplottet tyder hverken på systematiske afvigelser fra en lineær sammenhæng eller på områder med forskellig varians. QQplottet af residualerne giver ikke anledning til bekymring med hensyn til normalfordelingsantagelsen.



### 3.3 Tests og konfidensintervaller i den lineære regressionsmodel

Test og konfidensintervaller for skæring og hældning i den lineære regressionsmodel kan laves ud fra de samme principper som i afsnittene [2.4](#) og [2.10](#). Konfidensinterval for variansen følger principippet i afsnit [2.6](#).

#### Resultat 3.5. (Test af hypotese om regressionsparametre)

I den lineære regressionsmodel,  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ , fra Definition [3.2](#), kan vi teste hypotesen, at hældningen har en kendt værdi,  $\beta = \beta_0$  mod alternativet  $\beta \neq \beta_0$ , ved  $t$ -

teststørrelsen

$$T = \frac{\hat{\beta} - \beta_0}{s_r / \sqrt{SSD_t}} = \frac{\hat{\beta} - \beta_0}{\text{sd}_s(\hat{\beta})} \sim t(n-2), \quad \text{sd}_s(\hat{\beta}) = \frac{s_r}{\sqrt{SSD_t}},$$

og vi kan teste hypotesen, at skæringen har en kendt værdi,  $\alpha = \alpha_0$  mod alternativet  $\alpha \neq \alpha_0$ , ved  $t$ -teststørrelsen

$$T = \frac{\hat{\alpha} - \alpha_0}{s_r \sqrt{\frac{1}{n} + \frac{\bar{t}^2}{SSD_t}}} = \frac{\hat{\alpha} - \alpha_0}{\text{sd}_s(\hat{\alpha})} \sim t(n-2), \quad \text{sd}_s(\hat{\alpha}) = s_r \sqrt{\frac{1}{n} + \frac{\bar{t}^2}{SSD_t}}.$$

I begge tilfælde beregnes  $p$ -værdien som  $2(1 - t_{\text{cdf}}(|t|, n-2))$ .

Et 95%-konfidensinterval for hældningen  $\beta$  eller for skæringen  $\alpha$  beregnes som

$$\hat{\beta} \pm t_0 \cdot \text{sd}_s(\hat{\beta}), \quad \text{og} \quad \hat{\alpha} \pm t_0 \cdot \text{sd}_s(\hat{\alpha}),$$

hvor  $t_0$  er 97.5%-fraktilen i en  $t(n-2)$ -fordeling,  $t_0 = t_{\text{inv}}(0.975, n-2)$ .

Et 95%-konfidensinterval for variansen  $\sigma^2$  eller for spredningen  $\sigma$  beregnes som i Resultat 2.5, med  $s^2$  i resultatet erstattet af  $s_r^2$  og  $df$  i resultatet lig med  $n-2$ .

Resultaterne her følger direkte fra Resultat 3.3 på samme måde som at Resultat 2.4 følger fra Resultat 2.2.

### Eksempel 3.6. (Forurening i vandprøver)

I Eksempel 3.4 omkring GLUase aktivitetens afhængighed af mængden af E.coli bakterier er det naturligt at overveje proportionalitet mellem aktivitet og bakteriemængde. For logaritmen til værdierne betyder dette en lineær sammenhæng, hvor hældningen er lig med 1. I modellen  $\text{LogGlu}_i \sim N(\alpha + \beta \cdot \log\text{Coli}_i, \sigma^2)$ ,  $i = 1, \dots, 98$ , tester vi derfor hypotesen  $\beta = 1$ .  $T$ -teststørrelsen bliver, idet  $SSD_{\log\text{Coli}} = 117.0331$ ,

$$t = \frac{0.8494 - 1}{0.3094 / \sqrt{117.0331}} = -5.2657,$$

og den tilhørende  $p$ -værdi er  $2(1 - t_{\text{cdf}}(5.2657, 98-2)) = 8.5 \cdot 10^{-7}$ . Da  $p$ -værdien er meget lille, bliver konklusionen, at data strider mod hypotesen om proportionalitet. Forfatterne i artiklen, hvor data stammer fra, diskuterer selv mulige grunde til afvigelsen fra en hældning på 1.

Lad os dernæst se på, hvor meget viden vi har om skæringen  $\alpha$  ud fra de 98 målinger. Et 95%-konfidensinterval for  $\alpha$  bliver på formen

$$-3.8872 \pm 1.9850 \cdot 0.3094 \cdot \sqrt{\frac{1}{98} + \frac{4.5337^2}{117.0331}} = [-4.15, -3.62],$$

idet  $t_0 = t_{\text{inv}}(0.975, 96) = 1.9850$  og  $\overline{\log\text{Coli}} = 4.5337$ . Bredden på intervallet afspejler, at data-værdierne for  $\log\text{Coli}$  ligger fra 2.8 til 86.9, som er lidt væk fra nul ( $\alpha$  er linjens værdi i nul). I en situation som her vil skæringen  $\alpha$  sjældent være af interesse i sig selv. Det vil være mere relevant at se på linjens værdi  $\alpha + \beta t_*$  i et punkt  $t_*$  inden for dataområdet for den forklarende variabel. Dette gør vi i afsnit 3.5 nedenfor.

Lad os slutte eksemplet af med at se på, hvor meget vi ved om spredningen  $\sigma$  i den lineære sammenhæng. Skønnet over  $\sigma$  er  $s_r = 0.3094$ , og et 95%-konfidensinterval for  $\sigma$  er givet ved

$$\left[ 0.3094 \cdot \sqrt{\frac{96}{53.2034}}, 0.3094 \cdot \sqrt{\frac{96}{20.5694}} \right] = [0.271, 0.360].$$

Spredningen ligger altså med 95% sikkerhed i intervallet fra 0.27 til 0.36. Denne ret store værdi af spredningen kan skyldes stor måleusikkerhed i målingen af GLUase aktivitet og i målingen af mængden af E.coli bakterier, såvel som en biologisk variation i GLUase aktivitet for en given mængde af E.coli bakterier. En afgang på 0.31 på en log skala betyder en faktor 1.4 på GLUase aktiviteten. I afsnit 3.5 beskriver jeg, hvor velbestemt mængden af E.coli bakterier er ud fra en måling af GLUase aktiviteten.

### 3.4 Beregning i R via *lm*

I R analyseres en regressionsmodel  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ , ved hjælp af funktionen *lm*, som står for *lineær model* med normalfordelte data. Funktionen bruges både til regressionsmodellen i dette kapitel og til de generelle lineære modeller i de to følgende kapitler. Funktionen *lm* laver beregningsarbejdet, men skriver selv kun ganske få ting ud. Man anvender derfor normalt funktionen *summary* på output fra *lm* for at få en mere fyldig udskrift. Input til *lm* er en *modelformel*, der beskriver, hvordan respons *x* skal forbindes med den forklarende variabel *t*. I regressionsmodellen her, hvor vektoren med responsværdierne hedder *x* og vektoren med værdierne af den forklarende variabel hedder *t*, er modelformlen blot "x~t". På venstre side af tildesymbolet skal der stå navnet på responsvariablen og på højre side skal strukturen af middelværdien angives. For den lineære regressionsmodel angives middelværdien  $E(X) = \alpha + \beta t$  ved blot at skrive navnet på regressionsvariablen. Analysen af regressionmodellen kan nu laves i R med kommandoen

```
summary(lm(x~t))
```

I nogle situationer vil det være relevant at "gemme" output for at kunne regne videre på elementer af output. Hvis jeg kalder output fra *lm* for *lmUD* og output fra *summary* for *sumUD*, bliver kaldene

```
lmUD=lm(x~t)
```

```
sumUD=summary(lmUD)
```

Output fra *summary* består af en *parametertabel* med overskrift *Coefficients*, hvor der er fire søjler: parameterskøn, standard error, *t*-teststørrelse og en *p*-værdi, og tabellen indeholder en række for hver parameter i middelværdimodellen:

*Coefficients:*

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	—	—	—	—
t	—	—	—	—

Den simple lineære regressionsmodel  $E(X_i) = \alpha + \beta t_i$  har to parametre: skæring  $\alpha$  (intercept) og regressionskoefficienten  $\beta$ , som navngives efter navnet på den forklarende variabel. Søjlen med *Standard error* angiver et skøn over spredningen på parameterskønnet jævnfør formlerne for  $s_d$ ,

i Resultat 3.5. For at teste hypotesen, at den sande parameterværdi er nul, laves en  $t$ -teststørrelse, som netop er parameterskøn divideret med standard error. Endelig er søjlen med  $p$ -værdier fremkommet ved at slå op i en  $t$ -fordeling med det relevante antal frihedsgrader. Antallet af frihedsgrader kan aflæses i teksten under parametertabellen, hvor skønnet  $s_r$  over spredningen af punkterne omkring linjen er angivet (*Residual standard error*) sammen med frihedsgradsantallet. Man kan adressere de forskellige elementer i output, samt finde konfidensintervaller for parametrene, ved følgende kommandoer:

sumUD\$sigma	skøn $s_r$ over spredning
sumUD\$df[2]	frihedsgrader knyttet til $s_r^2$
sumUD\$coefficients	parametertabel
lmUD\$fitted.values	$\hat{\xi}_i = \hat{\alpha} + \hat{\beta} t_i$ (forventede værdier)
lmUD\$residuals	residualer $r_i = x_i - \hat{\xi}_i$
confint(lmUD)	95%-konfidensintervaller for parametre

Indsæt nu i det følgende kodevindue de nødvendige ordrer for at få lavet en parametertabel i regressionsmodellen for data omkring forurening af vandprøver fra Eksempel 3.1. Find skønnene over hældning, skæring og spredning, og sammenlign med værdierne beregnet i Eksempel 3.4. Find dernæst et 95%-konfidensinterval for hældning og skæring og sammenlign med beregningerne i Eksempel 3.6.

### Kodevindue

```
logColi=c(2.80,2.86,2.88,2.86,3.07,3.10,3.19,3.33,3.32,3.32,3.41,3.26,
3.28,3.35,3.34,3.33,3.40,3.47,3.44,3.50,3.36,3.61,3.65,3.75,3.72,
3.72,3.74,3.79,3.84,3.84,3.82,3.88,3.85,3.81,3.87,3.87,3.88,3.98,
3.97,4.02,4.04,4.06,4.23,4.52,4.28,4.35,4.43,4.52,4.36,4.68,4.67,
4.88,4.98,4.88,5.02,5.14,5.23,5.24,5.35,5.33,5.31,5.39,5.43,5.44,
5.48,5.51,5.53,5.56,5.64,5.66,5.73,5.79,5.90,5.83,5.86,5.85,5.79,
5.77,5.99,6.11,6.33,6.32,6.34,6.19,6.29,6.74,6.89,5.49,5.50,5.52,
5.55,5.79,5.79,4.00,3.79,3.32,3.38,2.83)
logGlu=c(-1.05,-1.04,-1.10,-1.21,-1.28,-1.57,-0.56,-0.54,-0.59,-0.68,
-0.77,-1.12,-1.02,-1.32,-1.28,-1.16,-1.10,-1.07,-0.99,-0.85,-1.05,
-1.06,-0.78,-0.82,-0.90,-1.06,-1.17,-1.04,-1.03,-1.07,-0.26,-0.32,
-0.54,-0.61,-0.68,-0.75,-0.80,-0.79,-0.65,-0.71,-0.66,-0.59,-0.79,
-0.86,-0.53,-0.34,-0.33,-0.40,-0.10,0.32,0.21,0.04,-0.29,0.37,
0.45,0.56,0.14,0.89,0.78,0.89,1.01,1.08,1.22,1.10,1.15,1.08,1.04,
0.97,1.12,1.35,1.49,1.28,1.09,1.06,1.00,0.83,0.77,0.64,1.56,1.73,
1.79,1.42,1.22,1.03,1.01,1.83,1.53,1.12,1.06,1.06,0.99,0.67,0.70,
-0.75,-1.08,-1.06,-1.01,-1.05)
#
```

### Svar 3.2. Parametertabel

## 3.5 Linjens værdi og kalibrering

I nogle situationer kan man være specielt interesseret i linjens værdi for bestemte værdier af den forklarende variabel  $t$ . I eksemplet, vi har betragtet, med en ny måde at måle forureningsgraden i vandprøver kan det være, at man vil oversætte en grænseværdi på antallet af E.coli bakterier til en grænseværdi på GLUase aktiviteten. Dette kan formuleres på den måde, at i modellen  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ , vil vi gerne sige noget om parameteren  $\theta = \alpha + \beta t_*$ , hvor  $t_*$  er en givet værdi af den forklarende variabel.

Som skøn over linjens værdi bruges  $\hat{\theta} = \hat{\alpha} + \hat{\beta}t_*$ . På samme måde som i Resultat 3.3 kan man indse, at  $\hat{\theta} \sim N(\theta, \sigma^2(\frac{1}{n} + \frac{(t_* - \bar{t})^2}{SSD_t}))$ . Som i Resultat 3.5 kan vi derfor lave test for værdien af  $\theta$ , og vi kan lave et 95%-konfidensinterval. Det sidste er på formen

$$\hat{\theta} \pm t_0 \cdot \text{sd}_s(\hat{\theta}), \quad \text{sd}_s(\hat{\theta}) = s_r \sqrt{\frac{1}{n} + \frac{(t_* - \bar{t})^2}{SSD_t}}, \quad t_0 = t_{\text{inv}}(0.975, n - 2).$$

Man omtaler ofte  $\hat{\theta}$  som en prædikteret værdi og intervallet som et konfidensinterval for prædiktionen.

Derudover taler man også om et [prædiktionsinterval](#), som er bredere end ovenstående konfidensinterval. Et 95%-prædiktionsinterval er et interval, der vil indeholde en *kommande observation* med sandsynlighed 0.95. Hvis  $X_* \sim N(\alpha + \beta t_*, \sigma^2)$ , skal der laves et interval, der indeholder  $X_*$  med sandsynlighed 0.95. Hertil benyttes, at  $\hat{\theta} - X_* \sim N(0, \sigma^2(1 + \frac{1}{n} + \frac{(t_* - \bar{t})^2}{SSD_t}))$ . Ud fra dette kan vi konstruere en "t-teststørrelse":

$$\frac{\hat{\theta} - X_*}{\text{sd}_{\text{præ}}}, \quad \text{sd}_{\text{præ}} = s_r \sqrt{1 + \frac{1}{n} + \frac{(t_* - \bar{t})^2}{SSD_t}},$$

og lave prædiktionsintervallet som

$$\hat{\theta} \pm t_0 \cdot \text{sd}_{\text{præ}}, \quad t_0 = t_{\text{inv}}(0.975, n - 2).$$

I R kan konfidensintervallet og prædiktionsintervallet beregnes med funktionen *predict*. Input til *predict* er output fra kald af *lm* og en såkaldt *dataframe* med de værdier af den forklarende variabel  $t$ , hvori vi ønsker at beregne linjens værdi. Lad output fra *lm* være `lmUD=lm(x~t)`. Så danner vi en *dataframe* med navnet *NyData* med kommandoen

```
NyData=data.frame(t=tNY)
```

hvor  $tNY$  er en vektor med de nye værdier af  $t$ . Det er vigtig, at der bruges det samme navn for den forklarende variabel (i vores tilfælde  $t$ ) i kaldet til *lm* og i konstruktionen af *NyData*. Nu beregnes konfidensintervallet for linjens værdi i de nye punkter med kommandoen

```
predict(lmUD,NyData,interval="confidence")
```

Output fra *predict* er en matrix med tre søjler og et antal rækker, der svarer til længden af  $tNY$ . Første søje er skøn over linjens værdi, anden søje er det nedre endepunkt i konfidensintervallet, og tredje søje er det øvre endepunkt.

For at beregne *prædiktionsintervallet* skal man erstatte "confidence" med "prediction" i kaldet til *predict*.

### Eksempel 3.7. (Forurening i vandprøver)

I Eksempel 3.1, omkring GLUase aktivitetens afhængighed af mængden af E.coli bakterier, kan vi være interesseret i at kunne skelne mellem badevand af *udmærket kvalitet*, *god kvalitet* eller *ringe kvalitet*, svarende til at mængden af E.coli bakterier er 250, 500 eller 1000 (cfu per 100 ml).



I det følgende kodevindue beregnes konfidensinterval for linjens værdi i disse punkter, det vil sige konfidensinterval for middelværdien af logaritmen til GLUase aktiviteten for de tre niveuaer af E.coli bakterier (vær opmærksom på, at søjleoverskrifterne, i det output I får her, er forskudt til venstre, hvilket ikke er tilfældet, når I laver beregningerne i jeres egen R-installation).

#### Kodevindue

```
logColi=c(2.80,2.86,2.88,2.86,3.07,3.10,3.19,3.33,3.32,3.32,3.41,3.26,
3.28,3.35,3.34,3.33,3.40,3.47,3.44,3.50,3.36,3.61,3.65,3.75,3.72,
3.72,3.74,3.79,3.84,3.84,3.82,3.88,3.85,3.81,3.87,3.87,3.88,3.98,
3.97,4.02,4.04,4.06,4.23,4.52,4.28,4.35,4.43,4.52,4.36,4.68,4.67,
4.88,4.98,4.88,5.02,5.14,5.23,5.24,5.35,5.33,5.31,5.39,5.43,5.44,
5.48,5.51,5.53,5.56,5.64,5.66,5.73,5.79,5.90,5.83,5.86,5.85,5.79,
5.77,5.99,6.11,6.33,6.32,6.34,6.19,6.29,6.74,6.89,5.49,5.50,5.52,
5.55,5.79,5.79,4.00,3.79,3.32,3.38,2.83)
logGlu=c(-1.05,-1.04,-1.10,-1.21,-1.28,-1.57,-0.56,-0.54,-0.59,-0.68,
-0.77,-1.12,-1.02,-1.32,-1.28,-1.16,-1.10,-1.07,-0.99,-0.85,-1.05,
-1.06,-0.78,-0.82,-0.90,-1.06,-1.17,-1.04,-1.03,-1.07,-0.26,-0.32,
-0.54,-0.61,-0.68,-0.75,-0.80,-0.79,-0.65,-0.71,-0.66,-0.59,-0.79,
-0.86,-0.53,-0.34,-0.33,-0.40,-0.10,0.32,0.21,0.04,-0.29,0.37,
0.45,0.56,0.14,0.89,0.78,0.89,1.01,1.08,1.22,1.10,1.15,1.08,1.04,
0.97,1.12,1.35,1.49,1.28,1.09,1.06,1.00,0.83,0.77,0.64,1.56,1.73,
1.79,1.42,1.22,1.03,1.01,1.83,1.53,1.12,1.06,1.06,0.99,0.67,0.70,
-0.75,-1.08,-1.06,-1.01,-1.05)

lmUD=lm(logGlu~logColi)
NyData=data.frame(logColi=log(c(250,500,1000)))
predict(lmUD,NyData, interval="confidence")
```

**Showhide:** Test dig selv

1. Hvorfor er konfidensintervallet i tilfældet med 250 bakterier noget smallere end intervallet i tilfældet med 1000 bakterier?
2. Lav prædiktionsintervaller i stedet for konfidensintervaller. Hvorfor er disse intervaller meget bredere end konfidensintervallerne?
3. Hvis du har fået en ny måling af GLUase aktiviteten, hvor log-værdien er 1.05, vil du så mene, det tyder på et E.coli bakterietal på 250 eller på 500? Kunne det tænkes, at bakteritallet er 1000?

### Svar 3.3. Prædiktionsinterval

I det følgende kodevindue beregnes konfidensintervallet for linjens værdi i mange punkter og indtegnes som en kurve i figur med data.

#### Kodevindue

```
logColi=c(2.80,2.86,2.88,2.86,3.07,3.10,3.19,3.33,3.32,3.32,3.41,3.26,
3.28,3.35,3.34,3.33,3.40,3.47,3.44,3.50,3.36,3.61,3.65,3.75,3.72,
3.72,3.74,3.79,3.84,3.84,3.82,3.88,3.85,3.81,3.87,3.87,3.88,3.98,
3.97,4.02,4.04,4.06,4.23,4.52,4.28,4.35,4.43,4.52,4.36,4.68,4.67,
4.88,4.98,4.88,5.02,5.14,5.23,5.24,5.35,5.33,5.31,5.39,5.43,5.44,
5.48,5.51,5.53,5.56,5.64,5.66,5.73,5.79,5.90,5.83,5.86,5.85,5.79,
5.77,5.99,6.11,6.33,6.32,6.34,6.19,6.29,6.74,6.89,5.49,5.50,5.52,
5.55,5.79,5.79,4.00,3.79,3.32,3.38,2.83)
logGlu=c(-1.05,-1.04,-1.10,-1.21,-1.28,-1.57,-0.56,-0.54,-0.59,-0.68,
-0.77,-1.12,-1.02,-1.32,-1.28,-1.16,-1.10,-1.07,-0.99,-0.85,-1.05,
-1.06,-0.78,-0.82,-0.90,-1.06,-1.17,-1.04,-1.03,-1.07,-0.26,-0.32,
-0.54,-0.61,-0.68,-0.75,-0.80,-0.79,-0.65,-0.71,-0.66,-0.59,-0.79,
-0.86,-0.53,-0.34,-0.33,-0.40,-0.10,0.32,0.21,0.04,-0.29,0.37,
0.45,0.56,0.14,0.89,0.78,0.89,1.01,1.08,1.22,1.10,1.15,1.08,1.04,
0.97,1.12,1.35,1.49,1.28,1.09,1.06,1.00,0.83,0.77,0.64,1.56,1.73,
1.79,1.42,1.22,1.03,1.01,1.83,1.53,1.12,1.06,1.06,0.99,0.67,0.70,
-0.75,-1.08,-1.06,-1.01,-1.05)

lmUD=lm(logGlu~logColi)
plot(logColi,logGlu)
abline(lmUD)
pktlogColi=2+c(0:100)*5/100
NyData=data.frame(logColi=pktlogColi)
predUD=predict(lmUD,NyData, interval="confidence")
lines(pktlogColi,predUD[,2],lty=3,col=2)
lines(pktlogColi,predUD[,3],lty=3,col=2)
```

Prøv at ændre "confidence" til "prediction" i ovenstående kørsel.



### 3.5.1 Kalibrering (invers regression)

Inden for kemi bruges lineære sammenhænge ofte til at lave et "måleapperat", hvor værdien af den forklarende variabel bestemmes ud fra respons (omtales ofte som "invers regression"). Når den lineære regressionsmodel etableres ud fra data, taler man om at kalibrere målemetoden. For eksempel kan man have lavet en række prøver med en kendt koncentration af et stof og målt intensiteten af lys efter passage af prøven. Typisk vil der være en lineær sammenhæng mellem logaritmen til lysintensiteten og koncentrationen. Efterfølgende kan man for en prøve med en ukendt koncentration måle lysintensiteten og lave skøn over koncentrationen ud fra den etablerede lineære sammenhæng.

I vores eksempel i dette afsnit med forurenede vandprøver ønsker vi, efter at sammenhængen mellem GLUase og antal bakterier er etableret, at bruge sammenhængen til ud fra en måling af GLUase at sige, hvad antallet af bakterier er.

Dette kan formuleres generelt på følgende vis. Ud fra de indsamlede data har vi estimeret parametrene i modellen  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ . For en ny værdi  $\tau$  af den forklarende variabel  $t$  betragtes  $m$  målinger  $Y_1, \dots, Y_m$  fra modellen  $Y_i \sim N(\alpha + \beta\tau, \sigma^2)$ . Vi ønsker at lave inferens om  $\tau$  baseret på både  $X_1, \dots, X_n$  og på  $Y_1, \dots, Y_m$ . Ved beregninger af samme type som i forbindelse med Resultat 3.5 kan man indse, at

$$t(\tau) = \frac{\bar{Y} - (\hat{\alpha} + \hat{\beta}\tau)}{s_r \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\tau - \bar{t})^2}{SSD_t}}} \sim \sim t(n-2),$$

hvor  $\bar{Y} = (Y_1 + \dots + Y_m)/m$ . Man kan derfor konstruere et 95%-konfidensinterval for  $\tau$  som de værdier af  $\tau$ , for hvilke

$$-t_0 \leq t(\tau) \leq t_0, \quad t_0 = t_{\text{inv}}(0.975, n-2).$$

For at finde de relevante værdier af  $\tau$  skal man løse en andengrads ligning. Beregningerne er vist i kodevinduet nedenfor for data omkring forurening af vandprøver fra Eksempel 3.1. Der laves først et konfidensinterval for log-værdien af antallet af E.coli bakterier i tilfældet, hvor log-værdien af GLUase aktiviteten er 1.05, hvorefter dette omregnes til et konfidensinterval for antallet af bakterier. Beregningerne laves på den måde, at der først defineres en funktion *inversReg* i R, hvorefter denne funktion kaldes. Funktionen *inversReg* findes i filen *Rfunktioner.txt* omtalt i underafsnittet *Egne funktioner i R* i afsnit 1.9.

#### Kodevindue

```
inversReg=function(lmUD, y) {
  sumUD=summary(lmUD)
  xbar=mean(lmUD$model[,1])
  m=length(y)
  ybar=mean(y)
```

```

t0=qt(0.975,sumUD$df[2])
ahat=sumUD$coefficients[1,1]
bhat=sumUD$coefficients[2,1]
s2r=(sumUD$sigma)^2
SSDt=s2r/(sumUD$coefficients[2,2])^2
A=bhat^2-t0^2*s2r/SSDt
if (A>0){
B=-2*t0^2*s2r*(ybar-xbar)/(bhat*SSDt)
C=-t0^2*s2r*(1/m+1/(sumUD$df[2]+2)+(ybar-xbar)^2/(bhat^2*SSDt))
thetahat=(ybar-ahat)/bhat
ci=thetahat+(-B+c(-1,1)*sqrt(B^2-4*A*C))/(2*A)
return(list(estimat=thetahat,konfidensinterval=ci))
} else {
return("Problem_er_ikke_veldefineret_da_beta_kan_være_nul")
}
}

logColi=c(2.80,2.86,2.88,2.86,3.07,3.10,3.19,3.33,3.32,3.32,3.41,3.26,
3.28,3.35,3.34,3.33,3.40,3.47,3.44,3.50,3.36,3.61,3.65,3.75,3.72,
3.72,3.74,3.79,3.84,3.84,3.82,3.88,3.85,3.81,3.87,3.87,3.88,3.98,
3.97,4.02,4.04,4.06,4.23,4.52,4.28,4.35,4.43,4.52,4.36,4.68,4.67,
4.88,4.98,4.88,5.02,5.14,5.23,5.24,5.35,5.33,5.31,5.39,5.43,5.44,
5.48,5.51,5.53,5.56,5.64,5.66,5.73,5.79,5.90,5.83,5.86,5.85,5.79,
5.77,5.99,6.11,6.33,6.32,6.34,6.19,6.29,6.74,6.89,5.49,5.50,5.52,
5.55,5.79,5.79,4.00,3.79,3.32,3.38,2.83)
logGlu=c(-1.05,-1.04,-1.10,-1.21,-1.28,-1.57,-0.56,-0.54,-0.59,-0.68,
-0.77,-1.12,-1.02,-1.32,-1.28,-1.16,-1.10,-1.07,-0.99,-0.85,-1.05,
-1.06,-0.78,-0.82,-0.90,-1.06,-1.17,-1.04,-1.03,-1.07,-0.26,-0.32,
-0.54,-0.61,-0.68,-0.75,-0.80,-0.79,-0.65,-0.71,-0.66,-0.59,-0.79,
-0.86,-0.53,-0.34,-0.33,-0.40,-0.10,0.32,0.21,0.04,-0.29,0.37,
0.45,0.56,0.14,0.89,0.78,0.89,1.01,1.08,1.22,1.10,1.15,1.08,1.04,
0.97,1.12,1.35,1.49,1.28,1.09,1.06,1.00,0.83,0.77,0.64,1.56,1.73,
1.79,1.42,1.22,1.03,1.01,1.83,1.53,1.12,1.06,1.06,0.99,0.67,0.70,
-0.75,-1.08,-1.06,-1.01,-1.05)

lmUD=lm(logGlu~logColi)
KalLog=inversReg(lmUD,1.05)
list(KIlog=KalLog$konfidensinterval,
KIantilog=exp(KalLog$konfidensinterval))

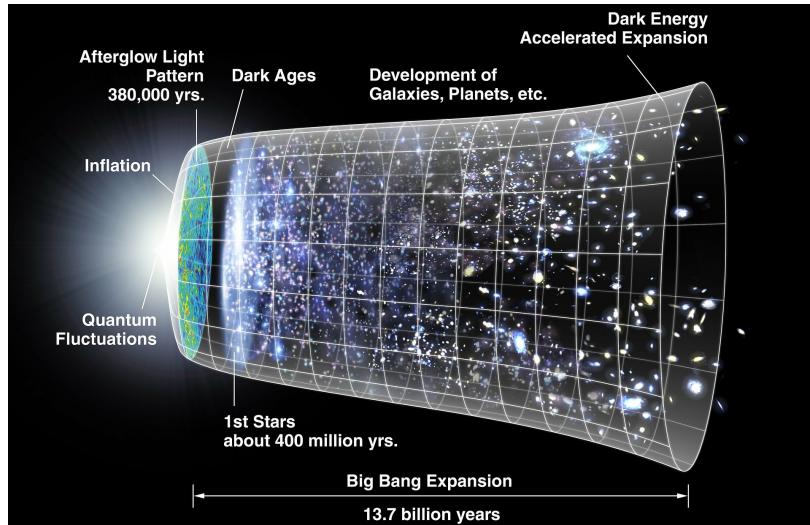
```

Kør koden og kommenter på resultatet. Prøv at ændre 1.05 i kaldet til `rep(1.05,4)`. Dette svarer til, at man har gentaget målingen af GLUase 4 gange (og tilfældigvis har målt den samme værdi alle fire gange). Kommenter også på dette resultat.

**Svar 3.4. Kalibrering**

### 3.6 Regression med kendt skæring

[Hubbles lov](#) siger, at den hastighed, hvormed galakser bevæger sig væk fra hinanden, er proportional med afstanden mellem galakserne. Formuleringen af loven af Edwin Hubble i 1929 er baseret på data indsamlet over en 10 års periode og vist i kodevinduet nedenfor. Loven danner baggrund for teorien om det ekspanderende univers.



Data består af værdierne (afstand, hast) for 24 galakser (afstand måles i megaparsecs og hastighed i kilometer per sekund). Som statistisk model bruger vi

$$\text{Hast}_i \sim N(\alpha + \beta \cdot \text{afstand}_i, \sigma^2), \quad i = 1, \dots, 24, \quad (\alpha, \beta, \sigma) \in \mathbf{R}^2 \times \mathbf{R}_+.$$

I kodevinduet analyseres denne model.

#### Kodevindue

```
afstand=c(0.032,0.034,0.214,0.263,0.275,0.275,0.450,0.500,0.500,
0.630,0.800,0.900,0.900,0.900,0.900,1.000,1.100,1.100,1.400,1.700,
2.000,2.000,2.000,2.000)
hast=c(170,290,-130,-70,-185,-220,200,290,270,200,300,-30,650,150,
500,920,450,500,500,960,500,850,800,1090)
```

```
lmUD=lm(hast~afstand)
plot(afstand,hast)
abline(lmUD)
summary(lmUD)
```

Når du kører denne kode, vil du se, at  $p$ -værdien for et test af hypotesen  $\alpha = 0$  er 0.629. Data strider altså ikke mod denne hypotese, som netop siger, at der er proportionalitet mellem afstand og (middelværdi af) hastighed.

Prøv at erstatte "summary" med "confint" i ovenstående kodevindue. Du vil da se, at et 95%-konfidensinterval for hældningen  $\beta$  er [298.1, 610.2]. Intervallet er meget bredt, hvilket afspejler, at der er stor variation i data omkring den lineære sammenhæng.

Modellen, der udtrykker proportionalitet, kan udtrykkes generelt som

$$X_i \sim N(\beta t_i, \sigma^2), \quad i = 1, \dots, n, \quad (\beta, \sigma^2) \in \mathbf{R} \times \mathbf{R}_+.$$

Analysen af denne model i **R** foretages som før med funktionen *lm*. For at fortælle at skæringen er nul, skal man tilføje "-1" i modelformlen. Kaldet til *lm* bliver således *lm(x~t-1)*. For Hubbles data er dette vist i det kommende kodevindue, hvor den røde linje i figuren er den estimerede linje i tilfældet med  $\alpha = 0$ .

### Kodevindue

```
afstand=c(0.032,0.034,0.214,0.263,0.275,0.275,0.450,0.500,0.500,
0.630,0.800,0.900,0.900,0.900,0.900,1.000,1.100,1.100,1.400,1.700,
2.000,2.000,2.000,2.000)
hast=c(170,290,-130,-70,-185,-220,200,290,270,200,300,-30,650,150,
500,920,450,500,500,960,500,850,800,1090)

lmUD1=lm(hast~afstand-1)
plot(afstand,hast)
abline(lm(hast~afstand))
abline(lmUD1, col=2)
confint(lmUD1)
```

I regressionsmodellen, hvor vi har antaget proportionalitet, altså at skæringen er nul,  $\alpha = 0$ , viser output fra *confint*, at konfidensintervallet for hældningen  $\beta$  er [336.7, 511.1]. Vi kan se, at konfidensintervallet bliver noget smallere sammenlignet med konfidensintervallet fra modellen, hvor  $\alpha$  er en ukendt parameter. Dette er et generelt fænomen: hvis man kan reducere en model ved at sætte nogle parametre til nul, vil de resterende parametre blive bedre bestemt. En del af den statistiske analyse går netop ud på at reducere en model for både at få en mere simpel model og for at få de resterende parametre bedre bestemt.

Intervallet for hældningen (= proportionalitetskonstanten = Hubbles konstant) er stadig stort og, som det har vist sig, fejlvisende. Den [anerkendte værdi](#) i dag ligger omkring 70. Et af problemerne med Hubbles data er, at strukturen af nogle af de stjerner, der blev brugt, blev fejltolket på daværende tidspunkt.

## 3.6.1 Fordelingsresultater

I modellen  $X_i \sim N(\beta t_i, \sigma^2)$  er

$$\hat{\beta} = \frac{\sum_i X_i t_i}{\sum_i t_i^2} \sim N\left(\beta, \sigma^2 / \sum_i t_i^2\right),$$

og skønnet over variansen  $\sigma^2$  er

$$s_{r0}^2 = \frac{1}{n-1} \sum_i (X_i - \hat{\beta} t_i)^2 \sim \sigma^2 \chi^2(n-1)/(n-1).$$

Ud fra disse resultater kan vi lave en  $t$ -teststørrelse for test af værdien af hældningen  $\beta$  og lave et 95%-konfidensinterval. Det sidstnævnte er på formen

$$\hat{\beta} \pm t_0 \frac{s_{r0}}{\sqrt{\sum_i t_i^2}}, \quad t_0 = t_{\text{inv}}(0.975, n-1).$$

Ovenfor har vi betragtet delmodellen af modellen  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ , hvor  $\alpha$  er kendt og lig med nul. Mere generelt kan vi se på situationen, hvor  $\alpha$  er kendt og lig med  $\alpha_0$ . Denne model kan analyseres ved at betragte  $\tilde{X}_i = X_i - \alpha_0$  og benytte resultaterne for situationen med  $\alpha = 0$ .

### 3.7 R-squared

Måske har I bemærket, at der i output fra `summary(lm(x~t))` i teksten under parametertabellen står *Multiple R-squared* og en værdi for denne størrelse. I vil opdage, at når I læser artikler, bliver denne værdi ofte angivet. Jeg vil ikke gøre brug af værdien i dette kursus, men vil her lige definere værdien.

Ideen bygger på, at vi tænker på respons som havende en variation, og noget af denne variation bliver forklaret ved vores model for middelværdien. Hvis responsværdierne er  $x_i$ ,  $i = 1, \dots, n$ , er den *totale variation* i respons givet ved

$$\text{SSD}_{\text{total}} = \sum_{i=1}^n (x_i - \bar{x})^2,$$

hvor  $\bar{x}$  er gennemsnit af responsværdierne. Betragt nu en model  $X_i \sim N(\xi_i, \sigma^2)$ , og lad  $\hat{\xi}_i$  være de forventede værdier, det vil sige middelværdien  $\xi_i$  med skøn over parametre indsats. I regressionsmodellen i dette kapitel er  $\xi_i = \alpha + \beta t_i$ , og de forventede værdier er  $\hat{\xi}_i = \hat{\alpha} + \hat{\beta} t_i$ . Med den del af variationen, der forklares af middelværdimodellen, menes

$$\text{SSD}_{\text{forklaret}} = \sum_{i=1}^n (\hat{\xi}_i - \bar{x})^2.$$

Den del af variationen i respons, der ikke forklares af modellen, er

$$\text{SSD}(M) = \sum_{i=1}^n (x_i - \hat{\xi}_i)^2,$$

som bruges i skønnet over variansen  $s^2(M) = \text{SSD}(M)/\text{df}(M)$ . Der gælder generelt at  $\text{SSD}_{\text{total}} = \text{SSD}_{\text{forklaret}} + \text{SSD}(M)$ , og *R-squared* værdien defineres som den fraktion af den totale variation, der forklares af modellen:

$$\text{R-squared} = \frac{\text{SSD}_{\text{forklaret}}}{\text{SSD}_{\text{total}}} = 1 - \frac{\text{SSD}(M)}{\text{SSD}_{\text{total}}}.$$

En R-squared værdi tæt på 1 er et udtryk for, at variansskønnet  $s^2(M)$  er lille relativt til den totale variation i respons.

Hvis R-squared skal være lig med 1, skal variansskønnet være nul. I regressionsmodellen betyder dette, at alle datapunkterne ligger præcist på en ret linje.

I definitionen af R-squared ovenfor kan vi dividere tæller og nævner med  $n$  og tænke på de to led som variansskøn,  $R\text{-squared} = 1 - (\text{SSD}(M)/n)/(\text{SSD}_{\text{total}}/n)$ . I vores normale arianskøn dividerer vi imidlertid med antallet af frihedsgrader. Hvis vi gør dette får vi den såkaldte *Adjusted R-squared*:  $1 - (\text{SSD}(M)/\text{df}(M))/(\text{SSD}_{\text{total}}/(n-1))$ .

I kodevinduet nedenfor er vist to eksempler med forskellig værdi af R-squared. Det første er med data fra Eksempel 3.1 og det andet med data fra afsnit 3.6.

### Showhide: Eksempler på R-squared værdier

#### Kodevindue

```
par(mfrow=c(1,2))
logColi=c(4.16,4.22,4.28,4.44,4.53,4.56,4.66,4.69,4.75,4.78,4.84,
4.84,4.84,4.88,4.88,5.00,5.00,5.09,5.34,5.66,5.78,5.81,5.88,6.03,
6.09,6.44,6.72,6.78,6.94,6.97,7.72,7.84,7.94,7.97,7.97,7.97,8.09)
logGlu=c(1.20,1.72,1.56,1.00,1.68,2.12,1.48,1.72,1.68,2.04,1.28,
1.60,1.96,1.52,2.28,2.00,1.64,1.60,2.56,2.52,2.40,2.04,2.84,3.08,
3.04,2.76,3.48,3.40,3.96,5.04,4.76,4.48,5.00,5.04,4.60,4.72,5.20)
lmUD=lm(logGlu~logColi)
plot(logColi,logGlu)
abline(lmUD)
text(4.2,4.8, pos=4, labels=round(summary(lmUD)$r.squared,2))

afstand=c(0.032,0.034,0.214,0.263,0.275,0.275,0.450,0.500,0.500,
0.630,0.800,0.900,0.900,0.900,0.900,1.000,1.100,1.100,1.400,1.700,
2.000,2.000,2.000,2.000)
hast=c(170,290,-130,-70,-185,-220,200,290,270,200,300,-30,650,150,
500,920,450,500,500,960,500,850,800,1090)
lmUD=lm(hast~afstand)
plot(afstand,hast)
abline(lmUD)
text(0.03,1000, pos=4, labels=round(summary(lmUD)$r.squared,2))
```



#### 3.7.1 Relation til korrelation

I kender fra jeres sandsynlighedskursus definitionen på kovarians og korrelation mellem to stokastiske variable  $X$  og  $Y$ :

$$\text{Cov}(X, Y) = E\{X - E(X)\}\{Y - E(Y)\}$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \text{Cov}\left(\frac{X - E(X)}{\text{sd}(X)}, \frac{Y - E(Y)}{\text{sd}(Y)}\right).$$

I kender også følgende regneregler

$$\text{Corr}(a + bX, c + dY) = \text{Corr}(X, Y), \quad |\text{Corr}(X, Y)| \leq 1,$$

og der gælder lighedstegn hvis og kun hvis der eksisterer konstanter  $a$  og  $b$  således at  $Y = a + bX$  (bevis for uligheden står også i MSRR Proposition 9.2.3 og Proposition 9.2.4).

Ligesom vi har indført empirisk varians  $s^2 = \sum_i(x_i - \bar{x})^2/(n - 1)$  kan vi indføre empirisk kovarians som  $s^2 = \sum_i(x_i - \bar{x})(y_i - \bar{y})/(n - 1)$ , og dermed empirisk korrelation  $r$  på formen

$$r = \frac{\frac{1}{n-1} \sum_i(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_i(x_i - \bar{x})^2 \frac{1}{n-1} \sum_i(y_i - \bar{y})^2}} = \hat{\beta} \sqrt{\frac{s_y^2}{s_x^2}}.$$

Her er  $\hat{\beta}$  skøn over hældning ved regression af  $x$  på  $y$ , og  $s_x^2$  og  $s_y^2$  er de empiriske varianser for henholdsvis  $x$  og  $y$ .

Lad os nu vende tilbage til  $R^2$ -værdien for regression af  $x$  på  $y$ . Lad  $r_i = x_i - \hat{\alpha} - \hat{\beta} = x_i - \bar{x} - \hat{\beta}(t_i - \bar{t})$  være det  $i$ 'te residual. Så har vi

$$\begin{aligned} SSD(M) &= \sum_i r_i^2 = \sum_i (x_i - \bar{x})^2 + \hat{\beta}^2 \sum_i (t_i - \bar{t})^2 - \hat{\beta} \sum_i (t_i - \bar{t})(x_i - \bar{x}) \\ &= \sum_i (x_i - \bar{x})^2 - \hat{\beta}^2 \sum_i (t_i - \bar{t})^2. \end{aligned}$$

Dermed bliver  $R^2$ -værdien

$$R^2 = 1 - \frac{\sum_i (x_i - \bar{x})^2 - \hat{\beta}^2 \sum_i (t_i - \bar{t})^2}{\sum_i (x_i - \bar{x})^2} = \hat{\beta}^2 \frac{s_t^2}{s_x^2} = r^2,$$

hvor  $r$  er en empirisk korrelation mellem  $x$  og  $t$ . Udover fortolkningen af  $R^2$ -værdien ovenfor, kan vi altså også tænke på denne som den kvadrerede empiriske korrelation.

### 3.8 Svar

#### Svar 3.1. Punkter udenfor

I en normalfordeling  $N(\mu, \sigma^2)$  er der cirka 5 procents sandsynlighed for at ligge udenfor  $\mu \pm 2\sigma$ . Jeg forventer derfor cirka  $0.05 \cdot 37 = 1.85$  punkter udenfor.

**Svar 3.2. Parametertabel**

For at få en parametertabel kan man bruge kommandoen `summary(lm(Lglu~Lcoli))`. For at få konfidensintervallerne kan man bruge kommandoen `confint(lm(Lglu~Lcoli))`.

**Svar 3.3. Prædiktionsinterval**

1. Med  $10^6$  bakterier ligger man i midten af dataområdet i forhold til  $10^4$  bakterier, hvor man ligger i yderkanten af dataområdet. Formelmæssigt ses denne forskel gennem bidraget  $(t_* - \bar{t})^2 / SSD_t$  til standard error for skønnet over linjens værdi.
2. Skifter blot "confidence" ud med "prediction" i kaldet til `predict`. Intuitivt skal prædiktionsintervaller "stikke" 1 til 2 gange spredningen længere ud end konfidensintervallerne, og da  $s_r$  er forholdsvis stor, bliver prædiktionsintervallerne væsentligt bredere end konfidensintervallerne.
3. Værdien 1.47 ligger både i prædiktionsintervallet for tilfældet  $10^4$  og i prædiktionsintervallet for tilfældet  $10^5$ , og i begge tilfælde lige langt fra en af grænserne i intervallet. Begge de to muligheder er derfor lige gode. Dog ligger 1.47 langt uden for prædiktionsintervallet i tilfældet med  $10^6$  bakterier, hvorfor det ikke er sandsynligt med så højt et bakterieantal.

**Svar 3.4. Kalibrering**

Kørsel af program viser, at antallet af E.coli bakterier ikke er særlig godt bestemt ud fra en enkelt måling af GLUases aktiviteten og ved brug af de data, vi har til rådighed. Konfidensintervallet er cirka fra 5000 til 200000.

Hvis vi har fire gentagne målinger af GLUase aktiviteten for den samme værdi af antallet af E.coli bakterier, får vi cirka halveret længden af konfidensintervallet på en  $\log_{10}$  skala.

Den overordnede konklusion er, at godt nok er der en lineær sammenhæng mellem GLUase aktiviteten og antallet af E.coli bakterier, men der er behov for forbedring af målemetoderne for at kunne bruge relationen i praksis til at vurdere antallet af bakterier ud fra GLUases aktiviteten.

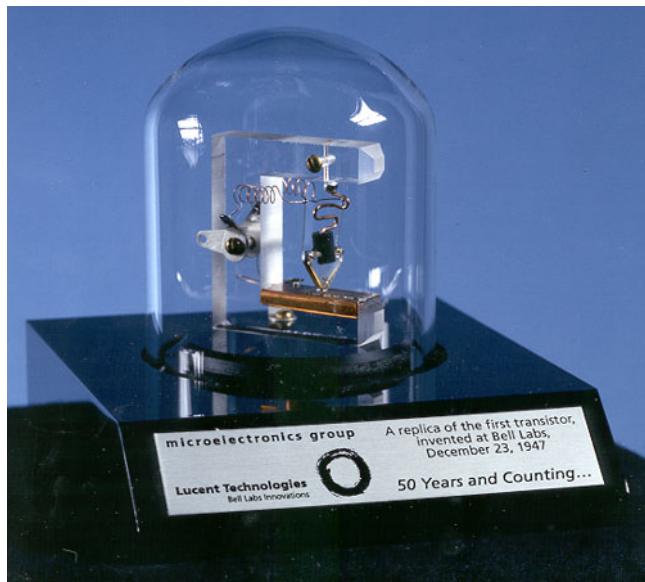
## 3.9 Opgaver til kapitel 3

I øvelserne hørende til kapitel 3 skal du blive fortrolig med den statistiske model for regressionsanalysen. Desuden skal du prøve at bruge funktionen `lm` i **R** til analysen. Denne funktion skal du også bruge i øvelserne hørende til kapitel 4 og kapitel 5.

**Showhide: Opgave 3.1: Regression**

Et mål for udviklingen af kompleksiteten af computerchips er antallet af transistorer på en chip. På adressen [Transistor count](#) kan man finde en tabel med antallet af transistorer for 108 chips produceret i perioden 1971-2016. [Moore's lov](#) siger, at antallet af transistorer på en chip fordobles

cirka hvert andet år. Dette blev formuleret af Gordon Moore, som var medstifter af *Intel* (Moore's udsagn går helt tilbage til 1965 og 1975).



I skal nedenfor undersøge dette udsagn ved at se på de  $\log_2$  transformerede antal transistorer afhængighed af produktionsår. Ved at bruge to-talslogaritmen opnår vi, at en fordobling af antal transistorer svarer til en stigning på 1.

Filen *Moore.csv* indeholder data. Filen har to søjler, hvor første søjle er produktionsår regnet med udgangspunkt i 1950 (en værdi på 24 svarer således til 1974), og anden søjle er antal transistorer på chippen.

- Indlæs data, og dan variablene *Aar* og *Antal* med indholdet af de to søjler. Dan endvidere variablen  $\log2Antal=\log2(Antal)$  med de  $\log_2$  transformerede antal transistorer. Lav en figur, hvor *log2Antal* tegnes op mod *Aar* (*Aar* langs førsteaksen og *log2Antal* langs andenaksen). Er det rimeligt at sige, at der er en lineær sammenhæng mellem de to variable?
- Opskriv den lineære regressionsmodel for data. Estimer parametrene i modellen, og lav figurer til modelkontrol. I residualplottet skal du indtegne to vandrette linjer, der skærer andenaksen i punkterne  $\pm 2s_r$ , hvor  $s_r$  er skønnet over spredningen i regressionsmodellen.
- Beregn 95%-konfidensintervaller for henholdsvis skæring og hældning i den lineære sammenhæng mellem middelværdien af *log2Antal* og *Aar*. Eftervis, at konfidensintervallet for hældningen, fundet gennem et kald til *confint*, er korrekt ved at bruge oplysninger i output fra *summary*.
- Overvej, om data er i overensstemmelse med en teori, der siger, at antallet af transitorer på en chip fordobles hvert andet år?



### Showhide: Opgave 3.2: Regressionsanalyse med prædiktion

I denne opgave skal I se på muligheden for at prædiktere den tid, der skal bruges til at teste et program ud fra oplysning om, hvor lang tid der er brugt på at kode programmet. Data i filen *Testtid.csv* indeholder oplysninger for 95 programmer. Filen har to søjler, hvor første søjle er tid

brugt på at kode, og anden søjle er tid brugt på at teste programmet (begge tider er i timer). Data er simulerede ud fra oplysningerne i figur 3 i artiklen [Software effort estimation with multiple linear regression: review and practical application](#).

I opgaven her skal I etablere en lineær sammenhæng mellem middelværdien af logaritmen til testtiden og logaritmen til kodningstiden og bruge denne sammenhæng til at prædiktere testtiden ved forskellige givne værdier af kodningstiden.

- (a) Lad  $\logKode$  være en vektor med logaritmen til kodningstiderne, og lad  $\logTest$  være en vektor med logaritmen til testtiderne. Lav en figur, hvor  $\logTest$  afsættes mod  $\logKode$ . Er det rimeligt at sige, at der er en lineær sammenhæng mellem de to variable?

Opskriv den lineære regressionsmodel for data, og estimer parametrene i denne via `lm` og `summary`. Indtegn den fundne linje i figuren med data.

Prøv i ord at beskrive sammenhængen i data, ud fra hvad du ser i figuren.

- (b) Lav et test, for hypotesen at hældningen er nul. Hvad bliver konklusionen af dit test? Lav dernæst et 95%-konfidensinterval for hældningen. Kommenter på betydningen af, at hældningen ser ud til at være væsentlig mindre end 1.

Lav et skøn over ændringen i middelværdi af  $\logTest$  mellem en værdi af  $\logKode$  på 2 og 4, og sammenhold denne med spredningen omkring regressionslinjen (jævnfør din egen beskrivelse af sammenhængen i data sidst i foregående spørgsmål).

- (c) Lav et 95%-konfidensinterval for middelværdien af  $\logTest$ , når  $\logKode$  er 6.

Lav dernæst et prædiktionsinterval for en kommende måling, når  $\logKode$  er 6.

Prøv at forklare, hvorfor prædiktionsintervallet er noget bredere end konfidensintervallet.

- (d) I dette spørgsmål skal du beregne konfidensintervallet og prædiktionsintervallet i mange punkter og indtegne disse som en kurve i figuren fra spørgsmål (a). Du kan finde inspiration til konstruktion af figuren i afsnit 3.5 i det skjulte punkt "Test dig selv". Til beregningen kan du kalde `predict` med nye datapunkter givet ved `data.frame(logKode=c(0:100)*0.07)`.



### Showhide: Opgave 3.3: Regression med kendt skæring

Data i denne opgave stammer oprindeligt fra artiklen [Rainfall erosivity over Rhodesia](#), men er her taget fra [Analysis of covariation and comparison of regression lines](#). Ønsket er at etablere en sammenhæng mellem årlig middelnedbør og erosionsraten, således at der kan laves et erosionskort over Rhodesia (nuværende Zambia og Zimbabwe).



Erosionsraten beregnes ud fra en mere detaljeret nedbørsregistrering end årlig middelnedbør. Data består af årlig middelnedbør (mm) og erosionsraten (joule mm/m<sup>2</sup>/hr) for 25 målepunkter i Highveld regionen af Rhodesia. Data findes i filen *RhodesiaHighveld.csv*, hvor første søje er årlig middelnedbør (*regnHigh*) og anden søje er erosionsraten (*erosionHigh*).

- (a) Lav en figur, hvor *erosionHigh* afsættes mod *regnHigh*, og hvor førsteaksen går fra 0 til 1000, og andenaksen går fra 0 til 15000.

Opstil regressionsmodellen  $M_1$ , hvor middelværdien af *erosionHigh* afhænger lineært af *regnHigh*. Estimer denne model via *lm*, og indtegn den fundne linje.

Angiv et 95%-konfidensinterval for hældning og for skæring.

Angiv et 95%-konfidensinterval for middelværdien af erosionen (linjens værdi), når den årlige nedbør er 500 mm.

- (b) Det er rimeligt at forestille sig, at hvis der ingen nedbør er, så er der heller ikke nogen erosion. Lav et *t*-test, for at skæringen med andenaksen er i punktet nul. Er det rimeligt at sige, at linjen går gennem (0, 0) ?

- (c) Opskriv regressionmodellen  $M_2$ , hvor middelværdien af *erosionHigh* er proportional med *regnHigh* (linjen har skæring med andenaksen i nul). Estimer denne model i **R** ved et passende kald til *lm*.

Angiv et 95%-konfidensinterval for hældningen i model  $M_2$ .

Angiv et 95%-konfidensinterval for middelværdien af erosionen (linjens værdi), når den årlige nedbør er 500 mm.

Prøv at beskrive i ord forskellen mellem de to konfidensintervaller i dette spørgsmål og de to konfidensintervaller i spørgsmål (a).



#### Showhide: Opgave 3.4: Prøve kalibreringsberegning

Denne opgave omhandler måden, hvorpå absorption af lys i en væske afhænger af koncentrationen af et absorberende molekyle i væsken, og hvordan vi kan bruge dette til at estimere koncentrationen ud fra en målt lysintensitet. Man mäter lysintensiteten  $I$  ved forskellige kendte koncentrationer af det absorberende molekyle. På denne måde får man etableret en kalibreringkurve, der

efterfølgende kan benyttes til at finde koncentrationen af molekylet i en prøve ud fra en måling af lysintensiteten efter lysets passage gennem prøven.

Absorption af denne type beskrives typisk via Lambert-Beers lov:

$$I = I_0 \exp\{-\varepsilon \nu c\}. \quad (3.2)$$

Her er  $\varepsilon$  absorptionskoefficienten for det absorberende molekyle,  $\nu$  er vejlængden gennem materialet,  $c$  er koncentrationen af molekylet og  $I_0$  er lysintensiteten når koncentrationen er nul.

I denne opgave betragter vi en serie målinger af lysintensiteten  $I$  som funktion af koncentrationen for en opløsning af Rhodamine 6G i ethanol. Den benyttede vejlængde gennem oplosningen er  $\nu = 1.00 \text{ cm}$ . Egentligt burde man i modelleringen af data også tage hensyn til, at koncentrationen af opløsningsmidlet ethanol ændrer sig, når koncentrationen af Rhodamine ændres, men denne effekt er så lille, at vi kan se bort fra den. Tager vi logartimen på begge sider i Lambert-Beers lov (3.2), får vi

$$H = \alpha - \varepsilon \nu c, \quad (3.3)$$

hvor  $\alpha = \log(I_0)$  og  $H = \log(I)$ .

Data i filen *LambertBeer.csv* giver den målte værdi af lysintensiteten  $I$  for 16 forskellige valg af koncentrationen. Filen har to søjler, hvor første søjle er koncentration, og anden søjle er lysintensiteten.

1. Dan en variabel *logLys* med logaritmen til de målte lysintensiteter og en variabel *konc* med koncentrationerne af Rhodamine 6G. Lav en figur, hvor *logLys* afsættes mod koncentrationen *konc*. Synes I, at der er en lineær sammenhæng i data? Synes I, at sammenhængen er god, med henblik på at estimere koncentration ud fra lysintensiteten?
2. Opskriv den lineære regressionsmodel, hvor respons er logaritmen til lysintensiteten, og den forklarende variabel er koncentration. Forklar, at regressionskoefficienten  $\beta$  i denne model er  $\beta = -\varepsilon \nu$ . Estimer modellen og indtegn den estimerede linje i figuren ovenfor.
3. Beregn et 95%-konfidensinterval for den ukendte koncentration af Rhodamine 6G i tre tilfælde med en enkelt ny måling af lysintensiteten *Lys*. Hertil kan du bruge funktionen *inversReg* omtalt i underafsnit 3.5.1. Funktionen findes i filen *Rfunktioner.txt*. De tre tilfælde er *Lys* = 2654, *Lys* = 4512 og *Lys* = 7688. Lav en tabel med resultaterne.



### Showhide: Opgave 3.5: Løvejagt

I denne opgave skal I bruge en lineær regressionsmodel til at sige noget om værdien af den forklarende variabel ud fra en målt responsværdi. I opgaven her er den forklarende variabel alderen af en løve, og respons er fraktion af sort pigment i løvens næsetip.



I artiklen [Sustainable trophy hunting of African lions](#) diskuteses hvordan trofæjagt af løver kan gøres bæredygtigt. Forfatternes konklusion er, at man skal sørge for, at de løver, der jages, er hanløver over en vis alder. Ofte bruger jægeren størrelsen og farven af løvens manke til at vurdere alderen, men dette er en meget usikker metode. En mere sikker metode består i at bruge andelen af sort pigment i løvens næsetip. I opgaven her skal I se på, hvordan andelen af sort pigment afhænger af alderen, og hvor godt vi kan estimere alderen ud fra dette.

Data for 32 hanløver fra Serengeti og Ngorongoro nationalparkerne ligger i filen *Loeve.csv*, der har to søjler med henholdsvis alder (år) og fraktion af sort.

- (a) Lav en figur, hvor *fraktion af sort* afsættes mod *alder*. Styr start og slut på andenaksen med tilføjelsen *ylim=c(0, 1)* til *plot*. Synes du, at der er en lineær sammenhæng i data? Synes du, at sammenhængen er god med henblik på at estimere alder ud fra fraktion af sort i næsetippet?
- (b) Opskriv den lineære regressionsmodel, hvor respons er fraktion af sort i næsetippet, og den forklarende variabel er alder.

Find skøn og 95%-konfidensinterval for hældning og skæring, og indtegn den skønnede linje i figuren fra foregående spørgsmål. Angiv også et skøn over spredningen  $\sigma$  omkring den lineære sammenhæng.

Lav figurer, der kan bruges til modelkontrol, og kommenter på disse figurer.

- (c) Betragt situationen, hvor en ny løve registreres, og fraktion af sort i næsen for denne løve er 0.2. Beregn et 95%-konfidensinterval for løvens alder i dette tilfælde.

Gentag beregningen i tre andre tilfælde, hvor en løve er observeret med henholdsvis 0.4, 0.6 og 0.8 for fraktionen af sort i næsen. Lav en tabel med resultaterne for de fire tilfælde.

Hvis vi kun ønsker at skyde løver, der er mindst 5 år gamle, hvor stor synes du så fraktionen af sort i næsen skal være, før du skyder?



### Showhide: Opgave 3.6: Udled skøn

Betræt regressionsmodellen  $X_i \sim N(\beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ ,  $(\beta, \sigma^2) \in \mathbf{R} \times \mathbf{R}_+$ , hvor skæringen er kendt og lig med nul.

- (a) Udled formlen for maksimum likelihood estimatet  $\hat{\beta}$ .
- (b) Udled fordelingen af  $\hat{\beta}$ .
- (c) Idet du må bruge, at  $\hat{\beta}$  og variansskøn er uafhængige, skal du udlede et *t*-test for hypotesen  $\beta = \beta_0$ .



### Showhide: Opgave 3.7: Kovarians

- (a) Vis, ud fra definitionen på kovarians, at  $\text{Cov}(a + bX, c + dY) = b \cdot d \cdot \text{Cov}(X, Y)$ , hvor  $a, b, c, d$  er konstanter.

**Showhide: Opgave 3.8: Fordeling af skøn**

Betrægt den lineære regressionsmodel  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ , og lad  $t_*$  være en fast kendt værdi af den forklarende variabel  $t$ . Som skøn over  $\theta = \alpha + \beta t_*$  bruger vi  $\hat{\theta} = \hat{\alpha} + \hat{\beta} t_*$ .

- (a) Vis, at  $\hat{\theta} \sim N\left(\theta, \sigma^2\left(\frac{1}{n} + \frac{(t_* - \bar{t})^2}{SSD_t}\right)\right)$ .





## En og tosidet variansanalyse

I kapitel 2 og 3 blev grunden lagt til forståelsen af en generel klasse af normalfordelingsmodeller. I afsnit 2.9 blev data inddelt i to undergrupper med hver sin middelværdi og i afsnit 3.1 så vi på regressionsmodeller, hvor middelværdien af respons afhænger lineært af en forklarende variabel. Endvidere stiftede vi i afsnit 3.4 bekendtskab med funktionen `lm` i **R** til analyse af modellen. I den generelle model inddeltes data i mere end to grupper og endvidere inddeltes data efter flere inddelingskriterier (som for eksempel i biologi efter køn og art). Vi vil også betragte regressionsmodeller med mere en én forklarende variabel og situationer, hvor regressionslinjen afhænger af hvilken undergruppe data tilhører.

Alle modellerne vil blive analyseret via `lm` i **R**, og et væsentligt aspekt i fremstillingen er at have en simpel notation for modellerne, således at vi nemt kan "kommunikere" med `lm`. Det første element i fremstillingen er at tænke på data på "matriksform"svarende til en *dataframe* i **R**. Hver søjle i matricen svarer til en variabel, som for eksempel responsvariabel, forklarende variabel eller variabel, der bruges til at inddеле data i undergrupper. Hver række svarer til et observationsnummer og indeholder værdierne for de forskellige variable knyttet til dette observationsnummer. Data i i afsnit 2.8 vedrørende længden af horn på den hornede tudseøgle vil på matrixform se ud som følger.

	Gruppe	Længde
1	doede	21.4
2	doede	23.9
:		
30	doede	20.7
31	levende	25.2
32	levende	26.9
183	levende	15.7
184	levende	17.7

Den generelle statistiske model bliver på formen

$$X_i \sim N(\xi_i, \sigma^2), \quad i = 1, \dots, n, \quad \text{uafhængige},$$

hvor den specifikke model fremkommer ved at beskrive, hvordan middelværdien  $\xi_i$  afhænger af værdierne for de forskellige variable i datasættet. Bemærk, at vi siger, at alle de stokastiske variable har samme varians  $\sigma^2$ . En specifik model  $M$  angiver, hvordan vektoren af middelværdier  $(\xi_1, \xi_2, \dots, \xi_n)$  afhænger af nogle ukendte parametre. Antallet af disse parametre betegnes med  $d(M)$ . Når de ukendte parametre i middelværdien er estimeret, indsættes disse i  $\xi_i$  og vi taler om de *forventede værdier*  $(\hat{\xi}_1, \hat{\xi}_2, \dots, \hat{\xi}_n)$  og om *residualerne*  $r_i = X_i - \hat{\xi}_i$ ,  $i = 1, \dots, n$ . Desuden bruger vi

$$df(M) = n - d(M), \quad SSD(M) = \sum_{i=1}^n (X_i - \hat{\xi}_i)^2 \quad \text{og} \quad s^2(M) = \frac{SSD(M)}{df(M)}, \quad (4.1)$$

hvor  $df(M)$  står for "degrees of freedom" (frihedsgradsantallet), og  $SSD(M)$  står for "Sum of Squared Deviations".

Afsnit 4.1 starter med en beskrivelse af, hvordan data inddeltes i undergrupper via *faktorer*. Begrebsmæssigt er en faktor meget simpel, men også meget nyttig når vi skal udvikle en sprogbrug for de generelle modeller. Et kendetegn ved den generelle model er, at alle de stokastiske variable har samme varians. Når data inddeltes i undergrupper, er det naturligt at lave en indledende undersøgelse, hvor vi vurderer, om der er samme varians i de forskellige undergrupper. Vi ved fra afsnit 2.12, hvordan to varianser sammenlignes, men hvordan sammenlignes varianser fra flere end to grupper? I afsnit 4.5 indføres *Bartletts test* til vurdering af mere end to varianser.

Afsnittene 4.2, 4.3 og 4.4 handler om ensidet variansanalyse, hvor data inddeltes i undergrupper ud fra en enkelt faktor. Når data inddeltes i undergrupper efter to faktorer, taler vi om tosidet variansanalyse, og denne model behandles i afsnittene 4.6 og 4.7.

## 4.1 Faktorer

Som omtalt i indledningen til dette kapitel, tænker vi på hele datasættet som organiseret i variable (søjler i datamatricen). En *faktor* er en variabel, der bruges til at inddеле data i undergrupper. Hvis der for eksempel laves en undersøgelse, hvor der deltager 4 kvinder og 3 mænd, og resultaterne for kvinderne angives først, kan vi angive dette med en variabel *køn\_ord* på formen

`køn_ord=(kvinde,kvinder,kvinder,kvinder,mand,mand,mand).`

Vi kunne også vælge at kode kvinde som 1 og mand som 2, og i stedet for *køn\_ord* benytte variablen

`køn_tal=(1,1,1,1,2,2,2).`

I den sidste udgave, *køn\_tal*, kan man ikke umiddelbart se, om denne variabel skal bruges til at inddеле data i undergrupper (altså som en faktor), eller skal bruges for eksempel som forklarende variabel i en regressionsmodel. I R laver man en variabel til en faktor med funktionen *factor*:

`køn_faktor=factor(køn_tal).`

I *køn\_faktor* står der ikke længere tal, men teksstrenge "1" og "2". I en faktor kaldes de forskellige værdier, der optræder, for *faktorniveauer*. Man kan se de forskellige faktorniveauer ved i R at anvende funktionen *levels* på faktoren.

**Showhide: Faktor i R**

Prøv at køre den følgende kode. Prøv dernæst at ændre "class" til "levels". Prøv til sidst at ændre "class(" til "lm(x~)". Den sidste kørsel viser, at *lm* fitter forskellige modeller, alt efter om højresiden i modelformlen er en numerisk variabel eller en faktor.

**Kodevindue**

```
Art=c(1,1,2,2,2,3,3,3,3)
fArt=factor(Art)
x=rnorm(9)
list(ud=fArt, udArt=class(Art), udfArt=class(fArt))
```



I har allerede brugt funktionen *lm* i **R** til analyse af regressionsmodellen. Input til *lm* er en såkaldt *modelformel*. For en regressionsmodel med respons i variablen  $x$  og den forklarende variabel  $t$  benyttede I  $lm(x \sim t)$ . En modelformel består af responsvariablen på venstre side af "tilde-symbolet og en angivelse af modellen på højre side af tilde (højresiden i sig selv kaldes også modelformlen). I regressionssituationen ved I, at  $x \sim t$  angiver modellen  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ . Hvis  $G$  er en faktor med indgangene  $G_1, \dots, G_n$ , angiver modelformlen  $X \sim G$  modellen med  $X_i \sim N(\mu_{G_i}, \sigma^2)$ ,  $i = 1, \dots, n$ . Det vil sige, at alle de observationer, der ligger på samme niveau af  $G$ , får den samme middelværdi. Modelformlen indfører dermed indirekte lige så mange middelværdiparametre som antallet af niveauer i faktoren. Mere konkret: for faktoren *køn\_faktor* ovenfor betragter vi modellen, hvor de 4 første observationer kommer fra en  $N(\mu_1, \sigma^2)$ -fordeling, og de 3 sidste observationer kommer fra en  $N(\mu_2, \sigma^2)$ -fordeling, med  $\mu_1$  og  $\mu_2$  parametre med ukendt værdi.

**4.1.1 Produkt af faktorer**

Ofte vil data være inddelt i undergrupper ud fra flere inddelingskriterier. Hvis der er to faktorer, for eksempel *Køn* og *Art*, kan vi danne en ny faktor *Køn\*Art*, der inddeler efter begge faktorer. Her er et eksempel:

Nummer	Køn	Art	Køn*Art
1	K1	A1	(K1,A1)
2	K1	A1	(K1,A1)
3	K1	A2	(K1,A2)
4	K1	A2	(K1,A2)
5	K2	A1	(K2,A1)
6	K2	A1	(K2,A1)
7	K2	A2	(K2,A2)
8	K2	A2	(K2,A2)

Vi kan se i dette eksempel, at *Køn\*Art* inddeler i fire grupper betegnet med (K1, A1), (K1, A2), (K2, A1) og (K2, A2).

### Showhide: Produkt af faktorer i R

Prøv at køre den følgende kode. I vil se, at **R** ikke accepterer "\*" mellem to faktorer direkte i kommandovinduet, hvorimod ":" mellem to faktorer ser ud til at give produktet af to faktorer som beskrevet ovenfor. Hvis det derimod er tale om en modelformel i **R**, må man gerne bruge "\*" mellem to faktorer, og colon er kun relevant i forbindelse med en mere præcis kontrol af den parametrisering af modellen, som **R** bruger.

#### Kodevindue

```
Kqn=factor(c("K1","K1","K1","K1","K2","K2","K2","K2"))
Art=factor(c("A1","A1","A2","A2","A1","A1","A2","A2"))

list(udStjerne=Kqn*Art ,udColon=Kqn:Art)
```



## 4.2 Ensidet variansanalyse

The Data And Story Library ([DASL](#)) indeholder datasæt, der kan bruges til at afprøve forskellige statistiske metoder. Specielt kan man finde et [datasæt](#), hvor en studerende har undersøgt, hvor effektiv forskellige former for håndvask er til at fjerne bakterier fra hænderne. Der undersøges fire metoder: vaske hænderne i vand, med almindelig sæbe, med antibakteriel sæbe og med antibakteriel spray (indeholdende 65% ethanol).



Hver morgen vælges en af metoderne, hænderne vaskes, og hånden placeres på en steril plade beregnet til at fremskynde bakterievækst. Antallet af bakteriekolonier tælles efter 2 dage. Proceduren er fulgt i 32 dage, således at hver af de fire metoder er afprøvet otte gange. I kodevinduet nedenfor udskrives data på tabelform, og der laves et boxplot for hver af håndvaskmetoderne. Bemærk, at input til *boxplot* er en modelformel, der deler data op i de fire undergrupper givet

ved faktoren *metode*. (I udskrift af datatabel, der dannes i outputvinduet, skal søjleoverskrifterne flyttes til højre.)

### Showhide: Data og boxplot

#### Kodevindue

```
bakt=c(74,84,70,51,135,51,164,5,102,110,88,19,124,67,111,18,
105,119,73,58,139,108,119,50,170,207,20,82,87,102,95,17)
metode=factor(rep(c("vand","saebe","antisaebe","antibakspray"),8))

boxplot(bakt~metode)
data.frame(Metode=metode, Bakterietal=bakt)
```

Kasserne i boxplottene er cirka lige høje, hvilket indikerer, at der er samme varians i de fire grupper af håndvaskmetode. I afsnit 4.5 viser jeg et formelt test for hypotesen om samme varians. Boxplottene tyder også på, at der er forskel i middelværdien af bakterietallet for de fire metoder. Jeg vil nu indføre en statistisk model for situationen med data opdelt i grupper og lave et test for hypotesen om samme middelværdi i grupperne.



Vi betragter  $n$  stokastiske variable  $X_1, \dots, X_n$  og en faktor  $G$ , der deler data op i grupper. Faktoren deler op i  $k$  grupper. Selvom faktorniveauerne er tekststrenge, vil det være bekvemt at ækvivalere disse med tallene  $1, 2, \dots, k$ , svarende til for eksempel en leksikografisk ordning af tekststrenge. Vi betragter den *ensidede variansanalysemødel*, hvor middelværdien af  $X_i$  er bestemt af faktorværdien  $G_i$ . På denne måde får modellen de  $k$  middelværdiparametre  $(\mu_1, \dots, \mu_k)$ . Præcist skrives modellen (her kaldet  $M_1$ ) som

$$M_1: X_i \sim N(\xi_i, \sigma^2), \quad \xi_i = \mu_{G_i}, \quad i = 1, \dots, n, \quad (\mu_1, \dots, \mu_k, \sigma^2) \in \mathbf{R}^k \times \mathbf{R}_+. \quad (4.2)$$

Hypotesen, om at der er samme middelværdi i de  $k$  grupper, kan skrives som

$$\mu_1 = \mu_2 = \dots = \mu_k,$$

og alternativet er blot, at ikke alle  $k$  middelværdier er ens. Under hypotesen befinder vi os i model  $M_2$ :

$$M_2: X_i \sim N(\xi_i, \sigma^2), \quad \xi_i = \mu, \quad i = 1, \dots, n, \quad (\mu, \sigma^2) \in \mathbf{R} \times \mathbf{R}_+. \quad (4.3)$$

## 4.2.1 Estimation og fordelingsresultatet

Situationen her er blot en lille udvidelse af situationen med to grupper i afsnit 2.9. Under model  $M_1$  gælder der, at

$$\hat{\mu}_g = \bar{X}_g = \frac{1}{n_g} \sum_{i \in I_g} X_i \sim N\left(\mu_g, \frac{\sigma^2}{n_g}\right), \quad g = 1, \dots, k,$$

$$s^2(M_1) = \frac{1}{n-k} \sum_i (X_i - \bar{X}_{G_i})^2 \sim \sigma^2 \chi^2(n-k)/(n-k),$$

hvor  $I_g$  er de indices blandt  $1, \dots, n$ , for hvilke  $G_i = g$  (alle observationsnumre tilhørende gruppe  $g$ ),  $n_g$  er antal elementer i gruppe  $g$ , og  $s^2(M_1)$  er skønnet over variansen  $\sigma^2$ . Uafhængigheden mellem  $\hat{\mu}_g$  og  $\sum_{i \in I_g} (X_i - \bar{X}_g)^2$  fra Resultat 2.2 giver, at variansskønnet  $s^2(M_1)$  er stokastisk uafhængig af skønnene over middelværdiparametrene ( $\hat{\mu}_1, \dots, \hat{\mu}_k$ ).

For model  $M_2$  er vi tilbage til en normalfordelt observationsrække fra afsnit 2.3. Vi har derfor

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_i X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

$$s^2(M_2) = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2 \sim \sigma^2 \chi^2(n-1)/(n-1).$$

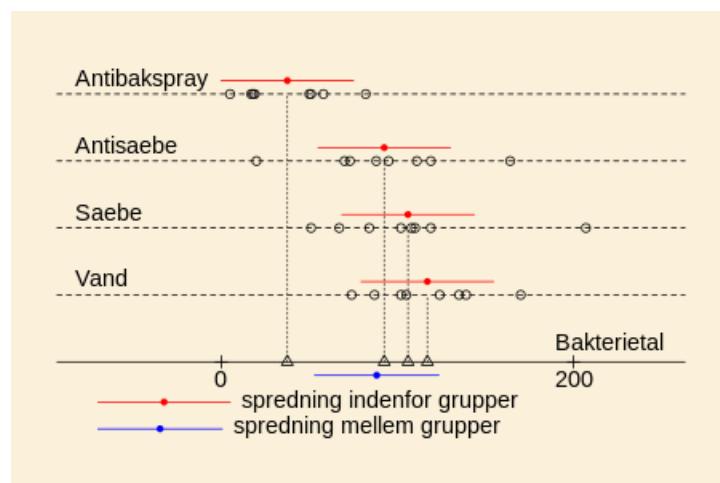
I næste afsnit skal vi også bruge, at under model  $M_2$  (samme middelværdi i alle grupperne) gælder der, at

$$s^2(M_1, M_2) = \frac{1}{k-1} \sum_i (\bar{X}_{G_i} - \bar{X})^2 = \frac{1}{k-1} \sum_g n_g (\bar{X}_g - \bar{X})^2 \sim \sigma^2 \chi^2(k-1)/(k-1),$$

og denne stokastiske variabel er uafhængig af  $s^2(M_1)$ . Uafhængigheden følger af, at  $s^2(M_1)$  er uafhængig af  $(\bar{X}_1, \dots, \bar{X}_k)$ , og  $\bar{X} = \sum_g n_g \bar{X}_g / n$ .

### 4.3 Teste middelværdierne ens

I dette afsnit indfører jeg på intuitiv vis et test for hypotesen om ens middelværdier i  $k$  grupper af observationer. Ideen er, at vi vil sammenligne variationen indenfor grupper, givet ved  $s^2(M_1)$ , og variationen mellem grupper givet ved  $s^2(M_1, M_2)$ . Den følgende figur illustrerer de to variationer for data fra afsnit 4.2 omkring fire måder at vaske hænder på.



Ideen er, at  $s^2(M_1)$  giver os viden om den ukendte varians  $\sigma^2$ , og med denne viden er vi i stand til at vurdere, om variationen mellem grupperne er større, end hvad der forventes, hvis middelværdierne er ens. Situationen ligner den i afsnit 2.12, hvor to uafhængige variansskøn sammelignes.

Vi vælger at betragte forholdet

$$F = \frac{s^2(M_1, M_2)}{s^2(M_1)} \sim F(k-1, n-k),$$

hvor fordelingen følger af Definition 2.12. En lille værdi er udtryk for, at gruppegennemsnittene ligger tæt på hinanden, som forventet under hypotesen om samme middelværdi i de  $k$  grupper, og en stor værdi tyder på, at de underliggende middelværdier ikke er ens. Formelt er store værdier af  $F$ -teststørrelsen kritiske for hypotesen om ens middelværdier, og  $p$ -værdien for testet er

$$p\text{-værdi} = 1 - F_{\text{cdf}}(F, k-1, n-k).$$

### Eksempel 4.1. (Vaske hænder)

For de 32 målinger af bakterietal ved fire metoder til håndvask fra afsnit 4.2, hvor data er vist i figuren ovenfor, finder man, at variansen inden for grupper er  $s^2(M_1) = 1410.1 = 37.55^2$ , og variansen mellem grupper er  $s^2(M_1, M_2) = 9960.7 = 8 \cdot 35.29^2$ , hvor 37.55 og 35.29 er spredningerne vist med rødt og blåt i figuren. Forholdet mellem disse er  $F = 9960.7/1410.1 = 7.06$ . Sandsynligheden for at få en værdi større end 7.06 i en  $F(4-1, 32-4)$ -fordeling er 0.0011. Da denne er langt under 0.05, siger vi, at data strider mod hypotesen om samme middelværdi af bakterietallet ved de fire metoder til håndvask. Direkte beregninger i R, uden brug af *lm*, fremgår af det følgende kodevindue.

#### Showhide: Direkte beregninger i R

##### Kodevindue

```
bakt=c(74,84,70,51,135,51,164,5,102,110,88,19,124,67,111,18,
105,119,73,58,139,108,119,50,170,207,20,82,87,102,95,17)
metode=factor(rep(c("vand", "saebe", "antisaebe", "antibakspray"), 8))

k=4
gnsj=tapply(bakt, metode, mean)
nj=tapply(bakt, metode, length)
vaj=tapply(bakt, metode, var)
n=sum(nj)
gnstotal=mean(bakt)
s2M1=sum((nj-1)*vaj)/(n-k)
s2M1M2=sum(nj*(gnsj-gnstotal)^2)/(k-1)
F=s2M1M2/s2M1
c(IndenforGruppe=sqrt(s2M1), MellemGruppe=sqrt(s2M1M2/8),
F=F, pværdi=1-pf(F, k-1,n-k))
```



## 4.4 Analyse i R

Den ensidede variansanalysemødel i (4.2) analyseres i R med kommandoen `lm(x~G)`, hvor  $G$  er en faktor, der deler data op i grupper, og  $x$  er en vektor med responsværdierne. For yderligere beregninger bruges `summary` og `confint` på output fra `lm`.

For at forstå output fra `summary` er det vigtigt at kende den parametrisering, som R anvender. For modellen i (4.2), med  $k$  grupper og tilhørende middelværdiparametre  $\mu_1, \dots, \mu_k$ , benytter R i forbindelse med modelformlen  $x \sim G$  følgende parametrisering og navngivning.

(Intercept)	G2	G3	...	Gk
$\mu_1$	$\mu_2 - \mu_1$	$\mu_3 - \mu_1$		$\mu_k - \mu_1$

Vi kan se, at `lm` bruger forskelle mellem parametre, og i mange tilfælde vil disse være af større interesse end parameterværdierne selv. Det  $t$ -test, der står i parametertabellen ud for et  $G_j$ , bliver således et test for at forskellen i parameterværdi er nul, eller sagt på en anden måde et test for, at de to parameterværdier er ens. De konfidensintervaller, der laves med `confint`, bruger den samme parametrisering og giver altså intervaller for forskel i parametrværdier. (Bemærk iøvrigt, at i tilfældet med to grupper vil `lm` betragte forskellen  $\mu_2 - \mu_1$ , hvormod `t.test` fra afsnit 2.13 betragter  $\mu_1 - \mu_2$ .)

### Showhide: Forstå parametrisering i R

I nedenstående kodevindue laves en analyse med `lm`, hvor spredningen  $\sigma$  er nul, hvorfor estimererne bliver lig med de sande værdier af parametrene. Kør kommandoerne, og sørge for at forstå output i parametertabellen i forhold til de sande værdier af  $\mu_A, \mu_B, \mu_C$  og  $\sigma$ .

#### Kodevindue

```
G=factor(c("B","B","A","A","C","C","C"))
muA=1
muB=3
muC=-2
sigma=0
x=c(muA,muA,muB,muB,muC,muC,muC)+sigma*rnorm(7)
summary(lm(x~G))
```

### Showhide: Forklaring

I output er Intercept =  $\mu_A$ , GB =  $\mu_B - \mu_A$  og GC =  $\mu_C - \mu_A$ . Vi kan også udtrykke dette omvendt:  $\mu_A$  = Intercept,  $\mu_B$  = Intercept+GB og  $\mu_C$  = Intercept+GC. I parametriseringen bruges en leksikografisk ordning af niveauerne i faktoren  $G$ , således at intercept svarer til "A".



Prøv at lægge mere og mere støj på data ved at vælge `sigma=0.5`, `sigma=1` og `sigma=2`. Bemærk, hvordan  $p$ -værdierne stiger i de tre  $t$ -test i parametertabellen.

Hvis man vil ændre på, hvilken gruppe **R** bruger som Intercept, kan man benytte funktionen *relevel*. Hvis kommandoen `G=relevel(G, "C")` indsættes lige efter definitionen af *G* i ovenstående kode, vil Intercept blive  $\mu_C$ . Prøv dette.



Hvis man ønsker, at *lm* skal bruge parametriseringen med  $\mu_1, \mu_2, \dots, \mu_k$  i stedet for forskelle i parameterværdierne, kan dette gøres med kaldet `lm(x~G-1)`. I modelformlen undertrykker -1" brugen af et intercept.

#### 4.4.1 Test af modelreduktion i R

For at beregne *F*-testet for reduktion fra model  $M_1$  i (4.2) til model  $M_2$  i (4.3), det vil sige *F*-testet for hypotesen, at alle middelværdierne er ens, skal man benytte funktionen *anova* i **R**. Input til *anova* er output fra to kald af *lm*, nemlig output fra analyse af model  $M_2$  og output fra analyse af model  $M_1$ . Vi kan skrive dette formelt som

```
anova(lm(modelformel2), lm(modelformel1))
```

I tilfældet med model  $M_2$  fra (4.3), hvor alle observationerne har samme middelværdi, foregår analysen med kaldet `lm(x~1)`. For at teste at alle middelværdierne er ens, bliver kaldet til *anova*:

```
anova(lm(x~1), lm(x~G))
```

Output fra *anova* er en *Testtabel*. Denne har 2 rækker og 7 søjler. Strukturen er som følger.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	—	—				
2	—	—	—	—	—	—

Søjlen *RSS* indeholder  $SSD(M)$  for de to modeller, og *Res.Df* de tilhørende frihedsgrader. Til beregning af *F*-testet skal vi bruge  $s^2(M_1, M_2)$ , som fremkommer som  $(\text{Sum of Sq})/\text{Df}$ , hvor *Df* kan beregnes som differensen mellem de to værdier under *Res.Df*, og *Sum of Sq* kan beregnes som differensen mellem de to værdier under *RSS* (dette beskrives nøjere i afsnit 4.7). Søjlen *F* indeholder *F*-teststørrelsen  $s^2(M_1, M_2)/s^2(M_1)$ , og søjlen *Pr(>F)* angiver den tilhørende *p*-værdi beregnet fra en  $F(Df, Res.Df[2])$ -fordeling.

#### 4.4.2 Analyse af data omkring metoder til håndvask

For datasættet beskrevet i starten af afsnit 4.2 lader vi  $bakt_i$  være bakterietallet for den  $i$ 'te måling og lader  $metode_i$  være den tilhørende metode til håndvask. Vi betragter modellen

$$Bakt_i \sim N(\mu_{metode_i}, \sigma^2), \quad i = 1, \dots, 32,$$

hvor middelværdiparametrene og  $\sigma^2$  kan variere frit. Kør følgende kode for at få lavet en parametertabel for modellen.

#### Kodevindue

```
bakt=c(74,84,70,51,135,51,164,5,102,110,88,19,124,67,111,18,
105,119,73,58,139,108,119,50,170,207,20,82,87,102,95,17)
metode=factor(rep(c("vand","saebe","antisaebe","antibakspray"),8))

summary(lm(bakt~metode))
```

Eftersom niveauet "antibakspray" kommer først i en leksikografisk ordning, er Intercept i parametertabellen  $\hat{\mu}_{\text{antibakspray}} = 37.50$ . Skønnet over forskellen mellem at bruge antibakteriel sæbe (antisaebe) og antibakteriel spray er  $\hat{\mu}_{\text{antisaebe}} - \hat{\mu}_{\text{antibakspray}} = 55.00$ , som står i rækken "metode-antisaebe". I samme række ses, at  $p$ -værdien er 0.0067 for et  $t$ -test af, at forskellen i middelværdi er nul (de to middelværdier er ens). Da  $p$ -værdien er langt under 0.05, tyder data altså på en forskel i de to metoder til håndvask. Hvis I erstatter "summary" med "confint", kan I se, at 95%-konfidensintervallet for forskel mellem de to middelværdier er [16.5, 93.5]. Dette er et bredt interval, hvilket afspejler, at der kun er 8 observationer i hver gruppe og spredningen i bakterietallet fra dag til dag er stort: skønnet over spredningen er  $s(M_1) = 37.55$ .

Parametertabellen indeholder tre  $t$ -test for forskel i middelværdier. Hvis nu de tre  $p$ -værdier alle havde været 0.06, skulle vi så konkludere, at data ikke strider mod at alle fire middelværdier er ens? Svaret er nej, for eksempel kunne  $\hat{\mu}_2$  ligge over  $\hat{\mu}_1$  og  $\hat{\mu}_3$  kunne ligge under  $\hat{\mu}_1$ , og så ville data tyde på en forskel mellem  $\mu_2$  og  $\mu_3$ . For at teste hypotesen om ens middelværdier

$$\mu_{\text{antibakspray}} = \mu_{\text{antisaebe}} = \mu_{\text{saebe}} = \mu_{\text{vand}},$$

benyttes kommandoen *anova* som vist i den følgende kode.

#### Kodevindue

```
bakt=c(74,84,70,51,135,51,164,5,102,110,88,19,124,67,111,18,
105,119,73,58,139,108,119,50,170,207,20,82,87,102,95,17)
metode=factor(rep(c("vand","saebe","antisaebe","antibakspray"),8))

anova(lm(bakt~1),lm(bakt~metode))
```

Genfind  $F$ -teststørrelsen og  $p$ -værdien fra afsnit 4.3 i testtabellen. Beregn også  $s^2(M_1)$  ud fra talene i testtabellen.

#### Showhide: Test dig selv

Betrægt igen analysen af modellen, hvor hver gruppe har sin egen middelværdi.

#### Kodevindue

```
bakt=c(74,84,70,51,135,51,164,5,102,110,88,19,124,67,111,18,
105,119,73,58,139,108,119,50,170,207,20,82,87,102,95,17)
metode=factor(rep(c("vand","saebe","antisaebe","antibakspray"),8))
```

```
summary(lm(bakt~metode))
```

Beregn ud fra parametertabellen skøn over de fire middelværdier. Kør så koden igen, hvor du tilføjer "-1" lige efter "metode" i modelformlen for at kontrollere dine beregninger.

Hvis vi gerne vil se forskellen mellem at bruge enten sæbe eller at bruge antibakteriel sæbe, tilføjer vi kommandoen `metode=relevel(metode, "saebe")` lige efter linjen, hvor `metode` indskrives, således at "saebe" bliver brugt som *Intercept*. Prøv dette. Er den antibakterielle sæbe bedre end almindelig sæbe?

**Svar 4.1. Bedre sæbe**



## 4.5 Teste mere end to varianser ens

I den ensidede variansanalysemodel (4.2) antager vi, at der er samme varians i alle grupperne. Det vil være naturligt at starte en analyse af data med at vurdere, om dette er rimeligt. Vi ved fra afsnit 2.12, hvordan man kan lave et test for, at to varianser er ens, men hvordan gør vi, når det skal vurderes, om mere end to varianser er ens. Der er ikke en intuitiv oplagt måde at gøre dette på. Vi kan imidlertid bruge det generelle princip til at konstruere en teststørrelse omtalt i afsnit 1.3 (Likelihoodratio test). Testet konstrueret på denne måde forbedrede M.S. Bartlett i 1937, og det kendes derfor i dag som [Bartletts test](#).

Antag, at der er  $k$  grupper af observationer, og for hver gruppe  $g = 1, \dots, k$  er der lavet et variansskøn  $s_g^2$  med  $df_g$  frihedsgrader:

$$s_g^2 \sim \sigma_g^2 \chi^2(df_g)/df_g,$$

og disse variansskøn er uafhængige. Vi ønsker at teste hypotesen, at varianserne er ens,

$$H: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2.$$

For at beskrive teststørrelsen indføres først et fælles variansskøn under hypotesen,

$$s^2 = \frac{\sum_{g=1}^k df_g s_g^2}{df}, \quad df = \sum_{g=1}^k df_g.$$

Bartletts test for ens varianser er på formen

$$Ba = \frac{1}{C} \left( df \cdot \ln(s^2) - \sum_{g=1}^k df_g \cdot \ln(s_g^2) \right), \quad C = 1 + \frac{1}{3(k-1)} \left( \sum_{g=1}^k \frac{1}{df_g} - \frac{1}{df} \right). \quad (4.4)$$

Store værdier af  $Ba$  er kritiske for hypotesen, og  $p$ -værdien for testet kan findes approksimativt som

$$p\text{-værdi} = 1 - \chi_{\text{cdf}}^2(Ba, k-1).$$

I R kan Bartletts test beregnes med funktionen `bartlett.test`. For den ensidede variansanalysemødel i (4.2) kan man blot benytte modelformlen  $x \sim G$  som input. Hvis data er inddelt efter to faktorer  $G$  og  $H$ , kan man først definere en ny faktor  $fGH=G:H$  og så bruge kaldet `bartlett.test(x ~ fGH)`. Endelig kan input også være en liste med output fra forskellige kørsler af `lm`, der hver især giver et variansskøn. I output fra `bartlett.test` angives teststørrelsen under *Bartlett's K-squared*.

### Eksempel 4.2. (Vaske hænder)

For datasættet beskrevet i starten af afsnit 4.2 lader vi  $bakt_i$  være bakteritallet for den  $i$ 'te måling og lader  $metode_i$  være den tilhørende metode til håndvask. Vi betragter modellen

$$M_0: Bakt_i \sim N(\mu_{metode_i}, \sigma_{metode_i}^2), \quad i = 1, \dots, 32,$$

hvor både middelværdien og variansen afhænger af gruppen. I det følgende kodevindue beregnes Bartletts test for hypotesen om ens varianser

$$\sigma_{antibakspray}^2 = \sigma_{antisaebe}^2 = \sigma_{saebe}^2 = \sigma_{vand}^2.$$

#### Kodevindue

```
bakt=c(74,84,70,51,135,51,164,5,102,110,88,19,124,67,111,18,
      105,119,73,58,139,108,119,50,170,207,20,82,87,102,95,17)
metode=factor(rep(c("vand", "saebe", "antisaebe", "antibakspray"), 8))

bartlett.test(bakt~metode)
```

I output ses, at teststørrelsen er  $B_a = 2.6325$ , og den approksimative  $p$ -værdi fra en  $\chi^2(3)$ -fordeling er 0.4518. Da denne er over 0.05, er konklusionen, at data ikke strider mod hypotesen om samme varians for de fire metoder til håndvask.

#### Showhide: Test dig selv

I det følgende kodevindue beregnes der variansskøn og frihedsgradsantal for hver af de fire metoder til håndvask. Tilføj kode til en direkte beregning af  $B_a$  fra (4.4).

#### Kodevindue

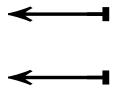
```
bakt=c(74,84,70,51,135,51,164,5,102,110,88,19,124,67,111,18,
      105,119,73,58,139,108,119,50,170,207,20,82,87,102,95,17)
metode=factor(rep(c("vand", "saebe", "antisaebe", "antibakspray"), 8))

varg=tapply(bakt, metode, var)
dfg=tapply(bakt, metode, length)-1
#
```

### Svar 4.2. Bartlett

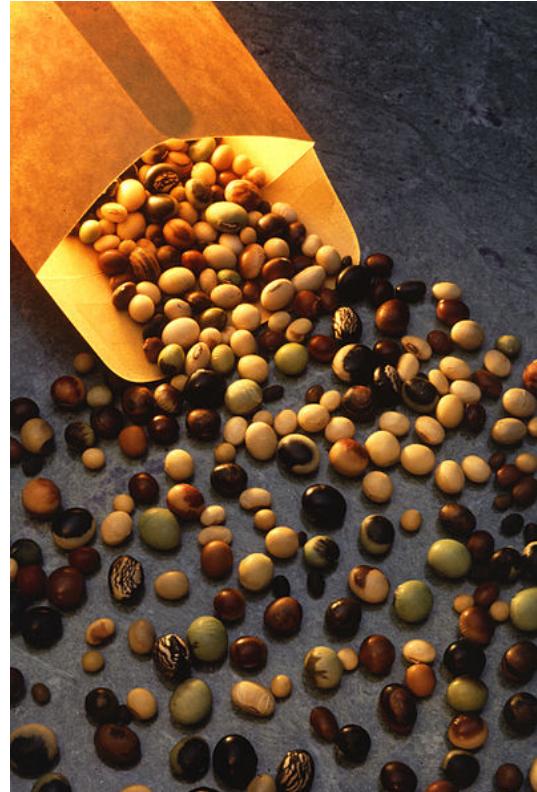
**Showhide: Svar: Bartlett**

Det fælles variansskøn beregnes som  $df = \text{sum}(dfg)$  og  $s^2 = \text{sum}(dfg * \text{varg}) / df$ . Tælleren i Bartletts teststørrelse beregnes som  $df * \log(s^2) - \text{sum}(dfg * \log(\text{varg}))$ , og nævneren beregnes som  $1 + (\text{sum}(1 / dfg) - 1 / df) / (4 - 1)$ .



## 4.6 Tosidet variansanalyse

Mekanisk "stress" som for eksempel vind påvirker planters vækst. For at studere dette i kontrollerede omgivelser er der i artiklen [Effects of seismic stress on the vegetative growth of Glycine max \(L.\) Merr. cv. Wells II](#) lavet et drivhuseksperiment, hvor væksten af soyabønner er undersøgt, når disse udsættes for stress i form af at blive rystet.



Efter 16 dages vækst er det totale bladareal målt. Planterne er delt op i to grupper, hvor den ene gruppe ikke udsættes for stress, og den anden gruppe udsættes for stress (faktoren *stress*). Planterne er desuden delt op i to grupper med hensyn til lysforhold i vækstperioden, hvor den ene gruppe vokser under en lav lysmængde og den anden under en højere lysmængde (faktoren *lys*).

På denne måde inddeltes data i fire grupper svarende til produktet  $stress*lys$  af de to faktorer. Rådata er ikke gengivet i artiklen, men den del, jeg vil bruge her, kan findes i bogen [Statistics for the Life Sciences](#).

I kodevinduet nedenfor laves en figur med boxplot for de fire grupper og en figur med qqplots for de fire grupper. Den sidste figur viser, at det er rimeligt at bruge en normalfordeling til beskrivelse af data i hver af de fire grupper, og den første figur peger på forskel i middelværdi mellem grupperne.

### Showhide: Boxplot og qqplot

#### Kodevindue

```
areal=c(264,200,225,268,215,241,232,256,229,288,253,288,230,
235,188,195,205,212,214,182,215,272,163,230,255,202,
314,320,310,340,299,268,345,271,285,309,337,282,273,
283,312,291,259,216,201,267,326,241,291,269,282,257)
stress=factor(rep(c("Uden", "Med", "Uden", "Med"), c(13,13,13,13)))
lys=factor(rep(c("Lav", "Høj"), c(26,26)))

stresslys=stress:lys
par(mfrow=c(1,2))
boxplot(areal~stresslys)

qqnorm(areal[(stress=="Med")&(lys=="Høj")], 
ylim=range(areal))
points(qqnorm(areal[(stress=="Med")&(lys=="Lav")]), 
plot=FALSE, col=2, pch=2)
points(qqnorm(areal[(stress=="Uden")&(lys=="Høj")]), 
plot=FALSE, col=3, pch=3)
points(qqnorm(areal[(stress=="Uden")&(lys=="Lav")]), 
plot=FALSE, col=4, pch=20)
legend("topleft", legend=c("MH", "ML", "UH", "UL"), col=c(1,2,3,4),
pch=c(1,2,3,20))
c()
```

Som I kan se i kodevinduet, er konstruktionen af et fælles qqplot for de fire undergrupper givet ved faktoren  $stress*lys$  lidt besværlig. Jeg har derfor lavet en R-funktion med navnet *qqnormFler* til dette. Kaldet til funktionen i dette eksempel bliver *qqnormFler(areal, stress:lys)*, hvor første del i input er variablen med responsværdierne, og det andet led er en variabel, der kan bruges til at dele data op i de ønskede undergrupper. Adgangen til denne funktion er som beskrevet i underafsnittet *Egne funktioner i R* i afsnit 1.9.



Jeg vil nu beskrive den tosidede variansanalysemødelen generelt. Data består af målinger fra  $n$  uafhængige stokastiske variable  $X_1, \dots, X_n$ . Disse inddeltes i grupper ved hjælp af to faktorer  $G = (G_1, \dots, G_n)$  og  $H = (H_1, \dots, H_n)$ , hvor  $G$  har  $k$  niveauer og  $H$  har  $m$  niveauer. Vi starter med en model, hvor både middelværdi og varians afhænger af, hvilken af de  $k \cdot m$  grupper observationen

tilhører

$$M_0: X_i \sim N(\mu_{G_i, H_i}, \sigma_{G_i, H_i}^2), \quad i = 1, \dots, n,$$

$$(\mu_{11}, \dots, \mu_{km}, \sigma_{11}^2, \dots, \sigma_{km}^2) \in \mathbf{R}^{k \cdot m} \times \mathbf{R}_+^{k \cdot m},$$

hvor niveauerne for de to faktorer for nemheds skyld betegnes med tal. Hvis det kan antages, at varianserne er ens, får vi modellen

$$M_1: X_i \sim N(\xi_i, \sigma^2), \quad \xi_i = \mu_{G_i, H_i}, \quad i = 1, \dots, n,$$

$$(\mu_{11}, \dots, \mu_{km}, \sigma^2) \in \mathbf{R}^{k \cdot m} \times \mathbf{R}_+, \quad d(M_1) = k \cdot m,$$

som kaldes den *tosidede variansanalysemød*el. I analysen af modellen betragtes følgende undermodeller:

$$M_2: \xi_i = \zeta_{G_i} + \eta_{H_i}, \quad (\zeta_1, \dots, \zeta_k, \eta_1, \dots, \eta_m, \sigma^2) \in \mathbf{R}^{k+m} \times \mathbf{R}_+, \quad d(M_2) = k + m - 1,$$

$$M_{3G}: \xi_i = \zeta_{G_i}, \quad (\zeta_1, \dots, \zeta_k, \sigma^2) \in \mathbf{R}^k \times \mathbf{R}_+, \quad d(M_{3G}) = k,$$

$$M_{3H}: \xi_i = \eta_{H_i}, \quad (\eta_1, \dots, \eta_m, \sigma^2) \in \mathbf{R}^m \times \mathbf{R}_+, \quad d(M_{3H}) = m,$$

$$M_4: \xi_i = \mu, \quad (\mu, \sigma^2) \in \mathbf{R} \times \mathbf{R}_+, \quad d(M_4) = 1.$$

Model  $M_2$  kaldes den *additive model*, og er vigtig på grund af fortolkningen af parametrene, som bliver beskrevet nedenfor. For modellerne  $M_{3G}$  og  $M_{3H}$  er vi tilbage ved den ensidede variansanalysemødel fra afsnit 4.2.

For at få en fornemmelse af om data kan beskrives med den additive model, kan man lave et *interaktionsplot*. Den indbyggede funktion i R er lidt mangelfuld på dette punkt, så i stedet anbefaler jeg en funktion *additivitetsPlot*, som findes i filen *Rfunktioner.txt*, jævnfør underafsnittet *Egne funktioner i R* i afsnit 1.9.

I et interaktionsplot beregner man gennemsnit i alle grupperne givet ved opdeling efter  $G * H$ . Gennemsnit afsættes mod niveauerne for den ene faktor, og alle gennemsnit, der ligger på det samme niveau af den anden faktor, forbinder. Hvis data kan beskrives med model  $M_2$  ovenfor, afspejler gennemsnittene i figuren altså  $\zeta_u + \eta_v$  afsat mod for eksempel  $u = 1, \dots, k$ , og punkterne med samme værdi af  $v$  forbinder. De kurver, der fremkommer, svarer altså til kurven  $(u, \zeta_u)$ , der parallelforskydes med værdierne fra  $\eta_v$ . I et interaktionsplot prøver vi derfor at vurdere, om kurverne ser ud til at være parallele.

### Showhide: Interaktionsplot

I det følgende kodevindue vises interaktionsplots baseret på den indbyggede funktion *interaction.plot* i R. Input til denne funktion er de to faktorer, der bruges til at dele data op i undergrupper, og vektoren med responsværdierne. Når I kører på jeres egen R-installation, kan I benytte *additivitetsPlot* fra filen *Rfunktioner.txt*, hvor input er som til *interaction.plot*.

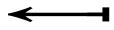
#### Kodevindue

```
areal=c(264,200,225,268,215,241,232,256,229,288,253,288,230,
235,188,195,205,212,214,182,215,272,163,230,255,202,
314,320,310,340,299,268,345,271,285,309,337,282,273,
283,312,291,259,216,201,267,326,241,291,269,282,257)
stress=factor(rep(c("Uden", "Med", "Uden", "Med"), c(13,13,13,13)))
```

```
lys=factor(rep(c("Lav","Høj"),c(26,26)))

par(mfrow=c(1,2))
interaction.plot(stress,lys,areal)
interaction.plot(lys,stress,areal)
```

Begge figurer viser approksimative parallelle kurver, hvilket tyder på, at data kan beskrives med den additive model.



### 4.6.1 Analyse i R og parametrisering

Model  $M_1$ , hvor hver gruppe har sin egen middelværdi, analyseres med kaldet  $x \sim G * H$ . Den additive model  $M_2$  analyseres med kaldet  $x \sim G + H$ . Lad os starte med at forstå output fra *summary* for den sidste model. Vi kan forstå output ved, for  $u = 1, \dots, k$  og  $v = 1, \dots, m$ , at skrive

$$\zeta_u + \eta_v = (\zeta_1 + \eta_1) + (\zeta_u - \zeta_1) + (\eta_v - \eta_1) = \text{Intercept} + Gu + Hv,$$

hvor højresiden viser de parametre, der bruges ved kaldet  $x \sim G + H$ . Vi ser her at modellen kan parametriseres med  $k + m - 1$  parametre, nemlig  $\zeta_1 + \eta_1, \zeta_2 - \zeta_1, \dots, \zeta_k - \zeta_1$  og  $\eta_2 - \eta_1, \dots, \eta_m - \eta_1$ . Går vi nu tilbage til model  $M_1$ , skriver vi i stedet

$$\begin{aligned}\mu_{u,v} &= \mu_{1,1} + (\mu_{u,1} - \mu_{1,1}) + (\mu_{1,v} - \mu_{1,1}) + (\mu_{u,v} - \mu_{u,1} - \mu_{1,v} + \mu_{1,1}) \\ &= \text{Intercept} + Gu + Hv + Gu:Hv,\end{aligned}$$

hvor den anden linje viser de parametre, der bruges ved kaldet  $x \sim G * H$ . Det sidste led kaldes interaktionen mellem de to faktorer. I den fulde model, model  $M_1$ , er  $Gu$  således forskel mellem niveau  $u$  og niveau 1 for faktoren  $G$ , når faktoren  $H$  ligger på niveau 1, og vice versa for  $Hv$ . Det nyttige ved den additive model  $M_2$  er, at  $Gu$  nu er forskel mellem niveau  $u$  og niveau 1 for faktoren  $G$ , *uanset* hvilket niveau faktoren  $H$  befinder sig på, og  $Hv$  er forskel mellem niveau  $v$  og niveau 1 for faktoren  $H$ , *uanset* hvilket niveau faktoren  $G$  befinder sig på.

#### Showhide: Parametrisering i R

Som for den ensidede variansanalyse i afsnit 4.4 betragtes simulerede data med spredning  $\sigma = 0$ , således at vi direkte kan se parametrene, der bruges i R. Vi betragter den additive model inden for den tosidede variansanalysemød.

#### Kodevindue

```
G=factor(c(1,1,1,1,1,1,2,2,2,2,2,2))
H=factor(c("A","A","B","B","C","C","A","A","B","B","C","C"))
eta=c(1,3)
```

```

zeta=c(2,0,1)
sigma=0
xi=c(1+2,1+2,1+0,1+0,1+1,1+1,3+2,3+2,3+0,3+0,3+1,3+1)
x=xi+sigma*rnorm(12)
summary(lm(x~G+H))

```

1. Hvad er middelværdien for en observation med  $G$  på niveau "2" og  $H$  på niveau "B"? Hvad er værdien af  $Intercept + G2 + HB$ ?
2. Udtryk R-parametrene  $Intercept$ ,  $G2$ ,  $HB$  og  $HC$  ud fra  $\eta$  og  $\zeta$ .
3. Opskriv de statistiske modeller svarende til henholdsvis kaldet  $lm(x \sim G * H)$  og til kaldet  $lm(x \sim G + H)$ .

**Svar 4.3. Forstå output**



**Eksempel 4.3.** (Soyabønner utsat for stress)

Vi betragter data omtalt i starten af dette afsnit omkring stresspåvirkning af soyaplanter. Lad  $Areal_i$  være den stokastiske variabel, der angiver respons (bladareal), og lad  $lys_i$  og  $stress_i$  være de tilhørende værdier for de to faktorer *lys* og *stress*. Lad os starte med modellen

$$M_0: Areal_i \sim N(\mu_{lys_i, stress_i}, \sigma_{lys_i, stress_i}^2), \quad i = 1, \dots, 42,$$

hvor hver gruppe bestemt af  $lys * stress$  har sin egen middelværdi og sin egen varians, og de fire middelværdier og varianser kan variere frit. Først undersøges hypotesen om fælles varians:

$$H: \sigma_{Hoej, Med}^2 = \sigma_{Hoej, Uden}^2 = \sigma_{Lav, Med}^2 = \sigma_{Lav, Uden}^2.$$

Beregningerne nedenfor i R viser, at Bartlett teststørrelsen er 1.16, og den tilhørende  $p$ -værdi fra en  $\chi^2(3)$ -fordeling er 0.76. Data strider således ikke mod hypotesen om samme varians i de fire grupper, og model  $M_0$  kan reduceres til model  $M_1$ :

$$M_1: Areal_i \sim N(\mu_{lys_i, stress_i}, \sigma^2), \quad i = 1, \dots, 42,$$

hvor de fire middelværdier og den fælles varians kan variere frit. Vi ønsker nu at teste reduktionen til den additive model

$$M_2: Areal_i \sim N(\zeta_{lys_i} + \eta_{stress_i}, \sigma^2), \quad i = 1, \dots, 42.$$

Vi har ovenfor lavet interaktionsplots, der viser overensstemmelse med den additive model. Parametertabellen hørende til model  $M_1$  viser, at der kun er en enkelt parameter, der vedrører interaktionen mellem *lys* og *stress* nemlig *lysLav:stressUden*. Test for, at denne parameter kan sættes lig med nul, aflæses under  $t$ -testet i parametertabellen, og giver en  $p$ -værdi på 0.86. Konklusionen er derfor, at data ikke strider mod hypotesen om additivitet.

Vi laver nu en parametertabel for den additive model  $M_2$ . Fra denne ses, at modellen ikke kan reduceres yderligere, idet et test for ingen effekt af *lys* giver  $p$ -værdi på  $7.1 \cdot 10^{-9}$ , og test for ingen

effekt af stress giver en  $p$ -værdi på 0.00015. Konfidensintervaller for de to effekter er  $[-74, -40]$  og  $[17, 50]$  for henholdsvis  $\zeta_{\text{Lav}} - \zeta_{\text{Høj}}$  og  $\eta_{\text{Uden}} - \eta_{\text{Med}}$ . Endelig er skønnet over spredningen 29.6. Der er således en tydelig effekt af både lys og stress.

### Showhide: Beregninger i R

#### Kodevindue

```
areal=c(264,200,225,268,215,241,232,256,229,288,253,288,230,
235,188,195,205,212,214,182,215,272,163,230,255,202,
314,320,310,340,299,268,345,271,285,309,337,282,273,
283,312,291,259,216,201,267,326,241,291,269,282,257)
stress=factor(rep(c("Uden", "Med", "Uden", "Med"), c(13,13,13,13)))
lys=factor(rep(c("Lav", "Høj"), c(26,26)))

list(Bartlett=bartlett.test(areal, lys:stress),
M1=summary(lm(areal~lys*stress)),
M2=summary(lm(areal~lys+stress)),
M2KI=confint(lm(areal~lys+stress)))
```



## 4.7 Det generelle F-test

I eksemplet omkring stresspåvirkning af soyabønner i foregående afsnit kunne vi teste interaktionen mellem *stress* og *lys* væk ved et *t*-test, da interaktionen kun bestod af en enkelt parameter. Generelt i den tosidede variansanalysemodel  $M_1$  fjernes  $d(M_1) - d(M_2) = km - (k + m - 1) = (k - 1)(m - 1)$  parametre, når vi går fra den fulde model til den addititive model, hvor  $k$  og  $m$  er antallet af niveauer i de to faktorer. Hvordan laver vi et test for denne reduktion? Ligesom i den ensidede variansanalysemodel skal varians "mellem grupper" sammenlignes med varians "inden for grupper". Jeg beskriver nu testet i en meget generel ramme.

En generel lineær normal model er på formen:

$$M: X_i \sim N(\xi_i, \sigma^2), \quad i = 1, \dots, n, \quad (\xi_1, \dots, \xi_n)^T \in L(M),$$

hvor  $L(M)$  er et lineært underrum af  $\mathbf{R}^n$ . Det sidste betyder blot, at der findes et  $k$  og faste vektorer  $\mathbf{v}_1, \dots, \mathbf{v}_k$ ,  $\mathbf{v}_j = (v_{j1}, \dots, v_{jn})^T$ , således at vektoren af middelværdier kan skrives på formen

$$(\xi_1, \dots, \xi_n)^T = \theta_1 \mathbf{v}_1 + \theta_2 \mathbf{v}_2 + \dots + \theta_k \mathbf{v}_k, \quad (4.5)$$

hvor  $(\theta_1, \dots, \theta_k)$  er ukendte parametre, som vi ønsker at estimere ud fra data. Dette kan virke noget abstrakt, men tænk på følgende to eksempler. For to grupper af observationer med hver sin middelværdi, hvor gruppe 1 kommer først, kan vi skrive

$$(\xi_1, \dots, \xi_n)^T = \mu_1(1, 1, \dots, 1, 0, 0, \dots, 0)^T + \mu_2(0, 0, \dots, 0, 1, 1, \dots, 1)^T,$$

eller for den simple regressionsmodel, kan vi skrive

$$(\xi_1, \dots, \xi_n)^T = \alpha(1, 1, \dots, 1)^T + \beta(t_1, t_2, \dots, t_n)^T.$$

I det generelle  $F$ -test ønsker vi at teste reduktionen fra en model  $M_1$  til en undermodel  $M_2$ , hvor  $L(M_2)$  er et underrum af  $L(M_1)$ . I praksis betyder det sidste typisk, at man tester en hypotese om, at nogle angivne parametre er nul.

#### Resultat 4.4. (Det generelle $F$ -test)

Betrægt to modeller  $M_1$  og  $M_2$ , hvor  $M_2$  er en undermodel af  $M_1$ . Lad  $\hat{\xi}_i(M_1)$  og  $\hat{\xi}_i(M_2)$ ,  $i = 1, \dots, n$ , være de forventede værdier i de to modeller, og definer  $s^2(M_1, M_2) = \sum_i (\hat{\xi}_i(M_1) - \hat{\xi}_i(M_2))^2 / (d(M_1) - d(M_2))$ . Så er  $F$ -teststørrelsen for reduktion fra model  $M_1$  til model  $M_2$  givet ved

$$F = \frac{s^2(M_1, M_2)}{s^2(M_1)} = \frac{(SSD(M_2) - SSD(M_1)) / (df(M_2) - df(M_1))}{s^2(M_1)}.$$

Under model  $M_2$  beregnes  $p$ -værdien for testet som

$$p\text{-værdi} = 1 - F_{\text{cdf}}(F, df(M_2) - df(M_1), df(M_1)).$$

Testet beregnes i **R** ved kommandoen

```
anova(lm(modelformel for M2), lm(modelformel for M1)).
```

Output fra kaldet af *anova* er en *testtabel* med 2 rækker og 7 søjler:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	—	—				
2	—	—	—	—	—	—

Første række vedrører model  $M_2$  og anden række model  $M_1$ . Søjlen *RSS* indeholder  $SSD(M)$  for de to modeller, og *Res.Df* de tilhørende frihedsgrader. Indholdet i anden række i søjlen *Df* er differensen mellem de to værdier under *Res.Df*, og *Sum of Sq* er differensen mellem de to værdier under *RSS*. De to sidste søjler indeholder selve  $F$ -teststørrelsen og den tilhørende  $p$ -værdi.

Måske har I bemærket, at i output fra `summary(lm(modelformel))` står der til sidst "F-statistics:". Hvis *modelformel* ikke indeholder `-1`, er dette  $F$ -testet fra kommandoen `anova(lm(x~1), lm(modelformel))`, hvor *x* er vektoren med responsværdier. Selvom slutmodellen i dette test altid er modellen, hvor alle de stokastiske variable har den samme middelværdi, vil  $p$ -værdien også afhænge af startmodellen givet gennem *modelformel*. I kan se dette konkret i output fra det skjulte kodevindue i eksempel 4.3.

#### Eksempel 4.5. (Stress af soyabønner)

I kodevinduet nedenfor vises beregningen af  $F$ -test for reduktion fra den tosidede variansanalysemødel til den additive mode. Derudover vises de to  $F$ -test for henholdsvis ingen effekt af stress og ingen effekt af lys.

### Showhide: Beregninger i R

#### Kodevindue

```
areal=c(264,200,225,268,215,241,232,256,229,288,253,288,230,
235,188,195,205,212,214,182,215,272,163,230,255,202,
314,320,310,340,299,268,345,271,285,309,337,282,273,
283,312,291,259,216,201,267,326,241,291,269,282,257)
stress=factor(rep(c("Uden", "Med", "Uden", "Med"), c(13,13,13,13)))
lys=factor(rep(c("Lav", "Høj"), c(26,26)))

list(M1tilM2=anova(lm(areal~lys+stress), lm(areal~lys*stress)),
M2tilM3lys=anova(lm(areal~lys), lm(areal~lys+stress)),
M2tilM3stress=anova(lm(areal~stress), lm(areal~lys+stress)))
```

Kør koden og genfind  $p$ -værdierne fra Eksempel 4.3.



## 4.8 Estimation og t-test

Estimation af parametrene i en generel lineær model som i (4.5) foretages ved at minimere

$$\sum_{i=1}^n (x_i - \xi_i(M))^2$$

over de parametre, der indgår i middelværdivektoren  $(\xi_1, \dots, \xi_n)$ . I praksis foretages denne beregning nemt i R ved matriksberegninger, se afsnit 6.0. Som skøn over variansen  $\sigma^2$  bruger vi altid

$$s^2(M) = \frac{SSD(M)}{df(M)} \sim \sigma^2 \chi^2(df(M))/df(M),$$

med notation fra (4.1). Variansskønnet er altid uafhængig af skønnet over de parametre, der indgår i middelværdien  $\xi_i(M)$ .

**Resultat 4.6.** (Fordeling af parameterskøn)

For en parameter  $\theta$ , der indgår i middelværdivektoren  $(\xi_1(M), \dots, \xi_n(M))$ , gælder der altid, at der findes en konstant  $C(M)$ , således at

$$\hat{\theta} \sim N(\theta, \sigma^2 C(M)).$$

Dermed er *standard error* for  $\hat{\theta}$  givet ved  $sd_s(\hat{\theta}) = \sqrt{s^2(M)C(M)}$ , og  $t = (\hat{\theta} - \theta_0)/sd_s(\hat{\theta}) \sim t(df(M))$  under hypotesen  $\theta = \theta_0$ . Et 95%-konfidensinterval for  $\theta$  er på formen

$$\hat{\theta} \pm t_0 \cdot sd_s(\hat{\theta}), \quad t_0 = t_{inv}(0.975, df(M)).$$

Den matematiske baggrund for dette resultat omtales i afsnit 6.3.

## 4.9 Svar

### Svar 4.1. Bedre sæbe

Fra output fra summary under indgangen *metodeantisæbe* ser vi at skøn over forkellen  $\mu_{antisæbe} - \mu_{sæbe}$  er -13.5, og et *t*-test for hypotesen at denne forskel er nul giver en *p*-værdi på 0.48. Den observerede forskel er derfor ikke stor nok til at vi kan påvise en forskel i middelværdi.

### Svar 4.2. Bartlett

Det fælles variansskøn beregnes som  $df = \text{sum}(dfg)$  og  $s2 = \text{sum}(dfg * varg) / df$ . Tælleren i Bartletts teststørrelse beregnes som  $df * \log(s2) - \text{sum}(dfg * \log(varg))$ , og nævneren beregnes som  $1 + (\text{sum}(1 / dfg) - 1 / df) / (4 - 1)$ .

### Svar 4.3. Forstå output

1. Fra *G* kommer bidraget 3 til middelværdien og fra *H* bidraget 0, hvorfor middelværdien er  $3 + 0 = 3$ . Intercept+G2+HB =  $3 + 2 - 2 = 3$ , som er middelværdien, vi lige har udregnet.
2. Intercept=eta[1]+zeta[1], G2=eta[2]-eta[1], HB=zeta[2]-zeta[1] og HC=zeta[3]-zeta[1].
3. Det *i*'te respons er  $X_i$ , og den *i*'te værdi af de to faktorer er  $G_i$  og  $H_i$ . Model svarende til det første kald siger, at hver undergruppe givet ved de to faktorer har sin egen middelværdi,  $X_i \sim N(\mu_{G_i, H_i}, \sigma^2)$ . Model svarende til det andet kald siger, at middelværdien består af et bidrag fra gruppen bestemt af faktoren *G* plus et bidrag fra gruppen bestemt af faktoren *H*,  $X_i \sim N(\eta_{G_i} + \zeta_{H_i}, \sigma^2)$ .

## 4.10 Opgaver til kapitel 4

Øvelserne til kapitel 4 har til formål at gøre jer fortrolige med den generelle lineære normale model gennem nogle grundlæggende eksempler. Efter øvelsen skal I vide, hvad en faktor er, og I skal have en forståelse af det generelle  $F$ -test for reduktion af middelværdimodellen.

### Showhide: Opgave 4.1: Ensidet variansanalyse

I denne opgave bruger vi data fra artiklen [Grain size discrimination between sands of desert and coastal dunes from northwestern Mexico](#). Forfatterne ønsker at studere, om der er forskel i størrelsesfordelingen af sandprøver indsamlet tre steder i Altar ørkenen i det nordlige Mexico på grænsen til USA.



Størrelsesfordelingen findes ved brug af en *Laser Particle Size Analyser*, og i denne opgave ser vi på middelkornstørrelsen for hver sandprøve. Der er 14 prøver fra *San Luis Rio Colorado* området (kodet som "SanLuis" i datafilen), 16 prøver fra *Pinacate north* (kodet som "PinaNorth"), og 8 prøver fra *Pinacate south* (kodet som "PinaSouth"). De tre områder ligger i størrelsesordenen 100 km fra hinanden. Data ligger i filen *Sandproeover.csv*. Der er 38 rækker i filen svarende til de 38 sandprøver i eksperimentet, og to søjler med henholdsvis område og middelkornstørrelse. Kornstørrelsen måles på Phi-skalaen, der fremkommer ved at tage to-tals logaritmen til kornstørrelsen i millimeter. Gennemsnit og empirisk spredning for hvert område er gengivet i den følgende tabel.

Område	San Luis Rio Colorado	Pinacate north	Pinacate south
Antal sandprøver	14	16	8
Gennemsnit	2.066	2.609	2.585
Empirisk spredning	0.154	0.084	0.159

- (a) Indlæs data, og dan de to variable *Omraade* og *kornstr* med indholdet af de to søjler. Undersøg, om *Omraade* er en faktor. Svaret afhænger af, hvilken version af **R** man kører. Hvis *Omraade* ikke er en faktor, skal du omdanne den til en faktor med funktionen *factor*. Lav en figur, hvor *kornstr* afsættes mod *Omraade* kodet som 1 til 3 (dette kan du gøre ved at bruge `as.numeric(Omraade)`). Prøv også at lave et qqplot med kommandoen `qqnormFlere(kornstr, Omraade)` (se omtalen af *qqnormFlere* i det skjulte punkt *Boxplot* og *qqplot* i afsnit 4.6) og dernæst et boxplot med kommandoen `boxplot(kornstr~Omraade)`. Overvej, hvad disse figurer viser om forholdet mellem spredningerne i de tre områder og forholdet mellem middelværdierne.
- (b) Opskriv den statistiske model, hvor data er delt ind i tre grupper svarende til de tre områder, og data er normalfordelt med en middelværdi og varians, der afhænger af gruppen.

Opskriv hypotesen, at de tre varianser er ens. Lav Bartletts test, for at de tre varianser er ens. Hvad bliver konklusionen af testet?

- (c) Opskriv den statistiske model, hvor middelværdien afhænger af området, men de tre varianser er ens. Find estimerater i denne model (både for middelværdierne og for spredningen). Benyt parametertabellen til at lave et *t*-test for hypotesen, at  $\mu_{\text{PinaNorth}} = \mu_{\text{PinaSouth}}$ . Angiv et 95%-konfidensinterval for  $\mu_{\text{SanLuis}} - \mu_{\text{PinaNorth}}$ .
- (d) Opskriv hypotesen, at de tre middelværdier er ens, og lav et test for denne hypotese ved et passende kald til *anova*.

Hvad bliver konklusionen i denne opgave: er det rimeligt at sige, at der er samme middelværdi af middelkornstørrelsen i de tre områder af Altar ørkenen?



### Showhide: Opgave 4.2: Tosidet variansanalyse

I artiklen [Environmental heterogeneity does not affect levels of phenotypic plasticity in natural populations of three Drosophila species](#) undersøges, hvordan forskellige arter og geografisk adskilte grupper af bananfluer reagerer på miljøpåvirkninger. I denne opgave skal I se på en delmængde af data, hvor der betragtes *D. melanogaster* indsamlet i henholdsvis Danmark og Italien. Fra ægstadiet og frem udsættes fluerne for tre forskellige behandlinger: C (constant), PF (predictable fluctuation) og UF (unpredictable fluctuation). For C-gruppen holdes temperaturen konstant på 23 (grader celsius) hen over døgnet, for PF følger temperaturen en sinuskurve med maksimum på 18 og minimum på 13, og for UF følger temperaturen også en sinuskurve, men med stokastiske maksimum og minimum for hvert døgn. Ved en given alder udsættes fluerne for en kritisk varmepåvirkning (37.5 grader celsius uafbrudt), og tiden (minutter) indtil fluen går i koma registreres. Med undersøgelsen ønsker man således at se, om forskel i opvæksten har betydning for deres evne til at klare varmepåvirkningen, og om der er forskel i de to geografiske populitioner med hensyn til denne evne. Gennemsnit og empirisk spredning for logaritmen til tiden indtil koma er gengivet for hver kombination af land og behandling i nedenstående tabel.

(Land, Behandling)	(DK,C)	(DK,PF)	(DK,UF)	(IT,C)	(IT,PF)	(IT,UF)
Antal fluer	20	20	20	20	20	20
Gennemsnit	4.09	4.69	4.26	3.89	4.28	4.06
Empirisk spredning	0.54	0.49	0.44	0.55	0.52	0.54

Data i artiklen kan findes under *Dryad Digital Repository*. Data til opgaven her er i filen *Bananfluer.csv*, hvor søjle 1 angiver land, søjle 2 angiver behandling og søjle 3 angiver tiden i minutter, indtil fluen går i koma.

- (a) Du skal først se på, om tidsmålingerne kan beskrives med en normalfordeling, eller om man først bør tage logaritmen til tidsmålingerne. Indlæs data og dan variablene *Land*, *Behandling* og *tid*. Hvis *Land* og *Behandling* ikke er faktorer efter indlæsningen, skal du omdanne dem til faktorer. Lav qqplots for *tid* for alle seks kombinationer af land og behandling med kommandoen `qqnormFlere(tid, Land:Behandling)` (se omtalen af *qqnormFlere* i det skjulte punkt *Boxplot og qqplot* i afsnit 4.6). Lav dernæst tilsvarende qqplot for logaritmen til *tid*.

Hvad bliver din konklusion ud fra de to figurer?

Lav dernæst interaktionsplot for logaritmen til tiden i forhold til de to faktorer land og behandling (se omtalen af funktionen *additivitetsPlot* i det skjulte punkt *Interaktionsplot* i afsnit 4.6).

- (b) Opskriv modellen, hvor logaritmen til tidsmålingen hørende til hver gruppe bestemt af land og behandling følger sin egen normalfordeling.

Opskriv hypotesen, at varianserne i de 6 grupper er ens, og lav Bartletts test for denne hypotese. Er det rimeligt at sige, at de seks varianser er ens?

- (c) Opskriv modellen, hvor logaritmen til tiden er normalfordelt, og hver gruppe bestemt af land og behandling har sin egen middelværdi, og alle har den samme varians. Opskriv inden for denne model additivitetshypotesen, hvor middelværdien består af et bidrag fra land og et bidrag fra behandling.

Lav et test, for at data kan beskrives med den additive model. Hvad bliver konklusionen af testet? Stemmer konklusionen, med hvad du kan se i interaktionsplottet?

- (d) Lav et test for henholdsvis ingen effekt af behandling og ingen effekt af land inden for den additive model.

- (e) Angiv inden for den addititive model et 95%-konfidensinterval for forskellen i middelværdi af logaritmen til tiden mellem de to lande.

Oversæt det fundne interval til et interval for forholdet mellem middelværdierne af tid indtil koma for de to lande, jævnfør underafsnit 2.13.3.



### Showhide: Opgave 4.3: Teste for en lineær sammenhæng

I hydrologi, når man skal beregne vandgennemstrømningen i jordlag, benyttes ofte [Darcys lov](#). Denne siger, at vandgennemstrømningen er proportional med trykforskellen (trykgradienten). Darcy formulerede loven i 1856 baseret på eksperimenter, hvor vand strømmer gennem et rør fyldt med sand. Der er siden lavet forskellige tilføjelser til loven, hvor flere aspekter af vandgennemstrømningen inddrages.

Darcys lov, formulert som  $Q = \alpha \cdot P$ , fører til en lineær sammenhæng for logaritmiske størrelse:  $\log(Q) = \gamma + \beta \cdot \log(P)$ , hvor  $Q$  er vandgennemstrømningen og  $P$  er trykforskellen. Her forventer vi så, at  $\beta = 1$ . I skal analysere data i denne opgave for at teste, om der er en lineær sammenhæng på den logaritmiske skala og dernæst se på om  $\beta = 1$ . Data I skal bruge er Darcys oprindelige data suppleret med 2 simulerede gentagelser af eksperimentet (i artiklen omtales det at chefingeniør Mr.Baumgarten har gentaget eksperimentet, men data herfra opgives ikke). Data findes i filen *Darcy.csv*. Filen har 30 rækker og 2 søjler, hvor hver række svarer til en måling, søjle 1 indeholder trykforskel (i meter vandsøjle), og søjle 2 indeholder vandgennemstrømningen (i liter per minut). For hver trykforskel er der tre målinger af vandgennemstrømningen.

- (a) Indlæs data, og dan de to variable *logTryk* og *logVand* med logaritmen til trykforskel og logaritmen til vandgennemstrømningen. Lav en figur, hvor logaritmen til vandgennemstrømningen afsættes mod logaritmen til trykforskel. Dan en faktor *FakTryk* ud fra variablen *Ltryk*. Beregn gennemsnit (benyt `tapply(logVand, FakTryk, mean)`) for hver trykforskelsgruppe og indtegn disse gennemsnit som en kurve i figuren. Indsæt endelig regressionslinjen fra en regression af *logVand* på *Ltryk*.

- (b) Opskriv den statistiske model  $M_1$ , hvor hver trykforskelsgruppe har sin egen middelværdi af logaritmen til vandgennemstrømningen, og variansen er ens.

Opskriv også den statistiske model  $M_2$ , hvor middelværdien af  $\log V_{and}$  afhænger lineært af  $Ltryk$ .

Lav nu  $F$ -testet for reduktion fra model  $M_1$  til model  $M_2$ . Hvad bliver konklusionen af testet: er det rimeligt at sige, at middelværdien af logaritmen til vandgennemstrømningen afhænger lineært af logaritmen til trykforskelse?

- (c) Angiv 95%-konfidensintervaller for skæring og hældning og for spredning omkring linjen i den lineære regressionsmodel. Kan det antages, at hældningen er 1 i overensstemmelse med Darcys lov?



#### Showhide: Opgave 4.4: Scanning and browsing

I artiklen [Effects of user age on smartphone and tablet use, measured with an eye-tracker via fixation duration, scan-path duration, and saccades proportion](#) studeres, hvordan brugen af smartphones og tablets varierer mellem forskellige aldersgrupper. Under brugen følges en persons øjenbevægelse, og herudfra dannes et mål SPD (scan-path duration, målt i millisekunder), der afspejler en persons evne til at bruge redskabet. I artiklen siges der: "SPD measures global processing of interfaces, where longer SPD indicates less efficient scanning and browsing". Personer deles op i tre aldersgrupper: *unge*, *midaldrende* og *aeldre*. Desuden fordeles personerne på to eksperimenter (*Ex1* og *Ex2*). Hvert eksperiment består af ni opgaver inden for brugen af forskellige smartphone apps, og opgaven hørende til en app er forskellig mellem de to eksperimenter. Data er i filen *Smartphone.csv* der har tre søjler med henholdsvis eksperiment, aldersgruppe og *SPD*-målingen.

- (a) Lav en figur med 3 delplots med qqplots af SPD for de tre aldersgrupper for eksperiment 1. Kommenter på figuren.

Opskriv den statistiske model, hvor hver gruppe bestemt af aldersgruppe og eksperiment har sin egen middelværdi og sin egen varians af *SPD*, og data er normalfordelt. Lav et test for hypotesen, at der er samme varians i de 6 grupper.

- (b) Lav et interaktionsplot, og kommenter på hvad du ser i figuren. Opskriv modellen, hvor hver gruppe bestemt af aldersgruppe og eksperiment har sin egen middelværdi af *SPD*, og alle grupperne har den samme varians. Opskriv hypotesen om en additiv struktur af middelværdien med et bidrag fra aldersgruppe og fra eksperiment. Lav  $F$ -testet for hypotesen om additivitet.

- (c) Undersøg, om det kan antages, at aldersgruppe ikke har nogen effekt på *SPD*. Undersøg også, om eksperiment har nogen effekt på *SPD*. Husk at skrive modellerne op.

- (d) Angiv skøn over middelværdien blandt de *aeldre* for eksperiment *Ex1*. Angiv et 95%-konfidensinterval for forskellen i middelværdi af *SPD* mellem gruppen af *unge* og gruppen af *aeldre* inden for den additive model.

Angiv skøn over spredningen på *SPD* i den additive model.



### Showhide: Opgave 4.5: Bartletts test

Betrægt  $k$  uafhængige variansskøn  $V_1, \dots, V_k$  med  $V_j \sim \sigma_j^2 \chi^2(f_j)/f_j$ . Skriver man tætheden op for  $V_j$  (jævnfør B.10 i MSRR), kan man se, at logaritmen til likelihoodfunktionen baseret på  $V_j$  er

$$l_j(\sigma_j^2; v_j) = \log\left(\frac{(f_j/2)^{f_j/2} v_j^{f_j/2-1}}{\Gamma(f_j/2)}\right) - \frac{f_j}{2} \log(\sigma_j^2) - \frac{f_j}{2\sigma_j^2} v_j.$$

- (a) Vis, at maksimum likelihoodskønnet for  $\sigma_j^2$  baseret på tætheden af  $V_j$  er  $\hat{\sigma}_j^2 = V_j$ . Vis dernæst, at maksimum af log-likelihoodfunktionen er

$$l_j(\hat{\sigma}_j^2; v_j) = \log\left(\frac{(f_j/2)^{f_j/2} v_j^{f_j/2-1}}{\Gamma(f_j/2)}\right) - \frac{f_j}{2} \log(v_j) - \frac{f_j}{2}.$$

- (b) Betrægt nu hypotesen, at de  $k$  varianser er ens, hvor vi betegner den fælles værdi med  $\sigma^2$ . Vis, at under hypotesen om ens varianser er maksimum likelihoodskønnet for den fælles varians  $\sigma^2$  givet ved  $\hat{\sigma}^2 = \sum_j f_j V_j / f_\bullet$ , hvor  $f_\bullet = f_1 + \dots + f_k$ .
- (c) Betrægt nu log-likelihoodratio teststørrelsen, her betegnet med  $\lambda_{Ba}$ , på formen

$$\lambda_{Ba} = -2 \left\{ \sum_j l_j(\hat{\sigma}; v_j) - \sum_j l_j(\hat{\sigma}_j; v_j) \right\}.$$

Vis, at

$$\lambda_{Ba} = f_\bullet \log(\hat{\sigma}^2) - \sum_j f_j \log(\hat{\sigma}_j^2).$$

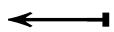
Dette er Bartlett teststørrelsen fra (4.4) pånær divisoren  $C$ . Faktoren  $C$  findes ved at lave en taylorudvikling af middelværdien af  $\lambda_{Ba}$ .



### Showhide: Opgave 4.6: Selv udregne $F$ -test

Gå tilbage til spørgsmål (d) i opgave 4.1, hvor der laves et  $F$ -test for at middelværdierne i de tre grupper er ens.

- (a) Udregn  $F$ -teststørrelsen ved kun at bruge output fra *summary(lm(modelformel))* for passende valg af modelformler.
- (b)  $F$ -teststørrelsen er på formen  $s^2(M_1, M_2)/s^2(M_1)$ . Udregn  $s^2(M_1, M_2)$  ved kun at benytte *lm(modelformel)\$fitted.values* for passende valg af modelformler.



### Showhide: Opgave 4.7: Selv udregne Bartletts teststørrelse

Gå tilbage til opgave 4.1, hvor Bartletts test for ens varianser betragtes. Beregn Bartletts teststørrelse baseret udelukkende på informationen i tabellen med gennemsnit og empiriske spredninger i de tre grupper.



**Showhide: Opgave 4.8: Selv udregne konfidensinterval**

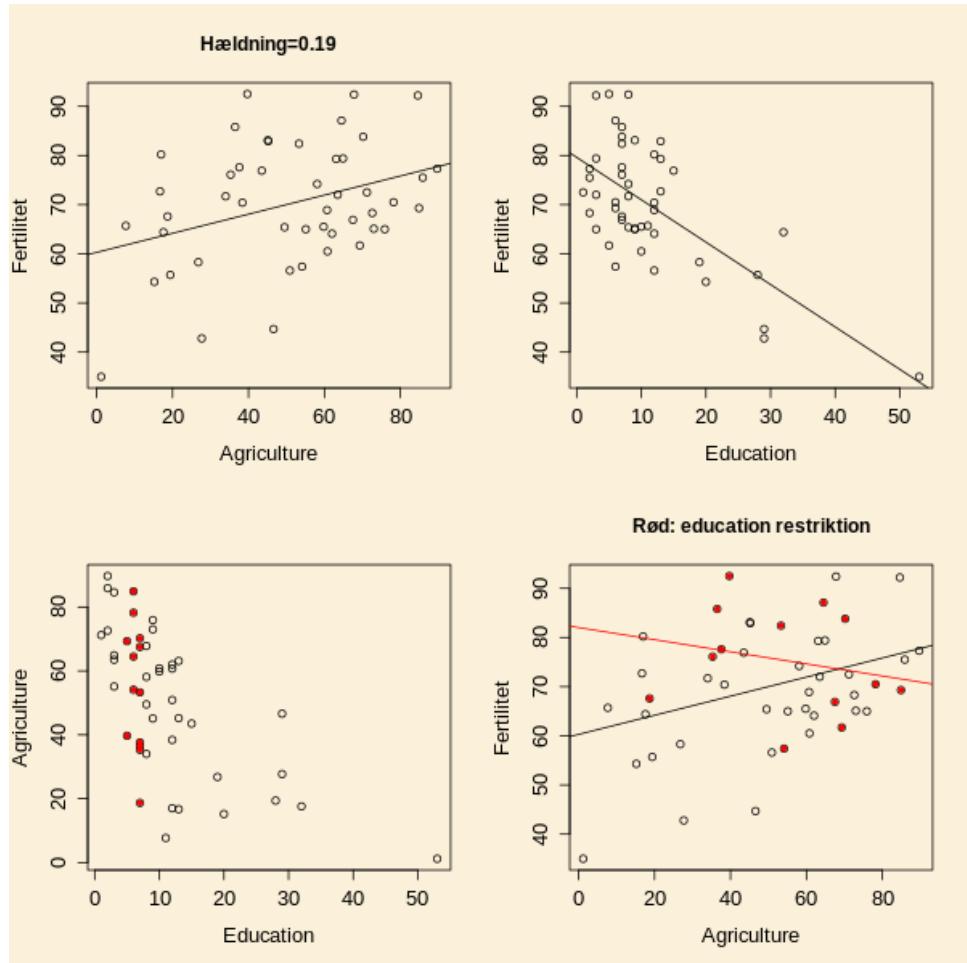
Gå tilbage til spørgsmål (e) i opgave 4.2, hvor der findes et konfidensinterval for forskellen i mid-delværdi af logaritmen til tiden mellem de to lande. Beregn dette interval baseret på den information du får fra et passende kald til *summary(lm(modelformel))*.





# Multipel regression

I indledningen til kapitel 3 viste jeg en figur, hvor indbyggerantal voksede med antallet af storkepar. Jeg omtalte det i kapitel 5 som *spurious correlation*. Man kan også udtrykke det på den måde, at det er vigtigt i en analyse af data at have inkluderet alle de relevante variable. Lad mig vise dette med et mere realistisk eksempel end eksemplet med storkepar. Datasættet `swiss` i R indeholder data omkring fertilitet i 47 kommuner i den fransktalende del af Sveits fra omkring 1888 (oprindelige kilde til data kan ses ved at følge ovenstående link). Hvis man afsætter fertiliteten mod procentdelen af befolkningen, der er beskæftiget i landbruget (*agriculture*), ses en positiv sammenhæng med en regressionskoefficient på 0.19 (*p*-værdi for at teste regressionskoefficienten lig med nul er 0.015). Afbildes i stedet fertiliteten mod procentdelen af befolkningen med en uddannelse (*education*), viser figuren en negativ sammenhæng. De to situationer er vist i de øverste delplots i figuren nedenfor. Laver vi en model, der tager hensyn til begge variable, en såkaldt *multipel regressionsmodel* som bliver indført i afsnit 5.3, viser det sig, at regressionskoefficienten hørende til *agriculture* nu er -0.07 (*p*-værdi for at teste regressionskoefficienten lig med nul er 0.41). At vi kan gå fra en regressionskoefficient på 0.19 til -0.07 skyldes en stærk korrelation mellem *agriculture* og *education*, som vist i det nedre venstre delplot. Når vi tager hensyn til både *education* og *agriculture*, kan man tale om afhængigheden af *agriculture* for fastholdt værdi af *education*. For at illustrere dette har jeg udvalgt data med en værdi af *education* mellem 4 og 8 og lavet regression af *fertilitet* på *agriculture* for de udvalgte data. De udvalgte data er vist med rødt i det nedre venstre delplot, og den tilhørende regressionslinje er vist i det nedre højre delplot. Det ses tydeligt i figuren, at afhængigheden af *agriculture* er anderledes når *education* holdes fast.



For 130 år siden havde folk i landbruget ikke nogen særlig uddannelse, og det er derfor naturligt, at en høj procentdel af uddannede peger på en lav procentdel inden for landbruget. Da der også har været en tendens til at uddannede får færre børn, må der være en sammenhæng mellem procentdel uddannede og fertiliteten. Når analysen tager hensyn til dette, kan vi ikke længere se en effekt af andelen, der arbejder i landbruget.

I dette kapitel ser jeg på to måder til at inddrage ekstra viden i en regressionssituation. Først ser jeg på situationen med en kategorisk variabel (en faktor), der inddeler data i undergrupper. For at undersøge indflydelsen af denne variabel ser vi på modellen, hvor hver gruppe har sin egen lineære sammenhæng mellem middelværdi af respons og den forklarende variabel i regressionsmodellen. Modellen indføres i afsnit 5.1, og et eksempel analyseres i afsnit 5.2. Dernæst ser jeg på situationen med flere forklarende variable, der kan bruges til beskrivelse af respons, som i fertilitetseksemplet ovenfor. Modellen indføres i afsnit 5.3, og i afsnit 5.4 diskuterer jeg, hvordan man kan vælge en delmængde af de forklarende variable til at danne en slutmodel for data.

Den sidste del af kapitel 7 handler også om den multiple regressionsmodel, i situationen hvor der er et meget stort antal forklarende variable, der kan bruges til at beskrive respons. Mange målemetoder indført gennem de sidste 30-50 år måler simultant et stort antal værdier knyttet til den samme prøve. Et eksempel, som bruges i 5.5, er et [near-infrared spektrum](#), hvor reflektionen af lys måles simultant ved et stort antal bølgelængder. Et andet eksempel er [microarray](#) målinger, hvor aktivitetsniveauet måles simultant for et stort antal gener i en celleprøve. Et tredje stort område er scanningsbilleder i diagnostiske medicinske sammenhænge. Vi skal kun berøre et lille hjørne af dette emne, nemlig hvordan man kan undgå "overfitting" ved at bruge *crossvalidation* til at vurdere, hvor godt modellen beskriver data. Dette beskrives i afsnit 5.6. Jeg har lavet nogle programmer i R, der gør det nemt for jer at "lege" med disse ting. Programmerne beskrives i 5.7.

## 5.1 Gruppесpecific regression

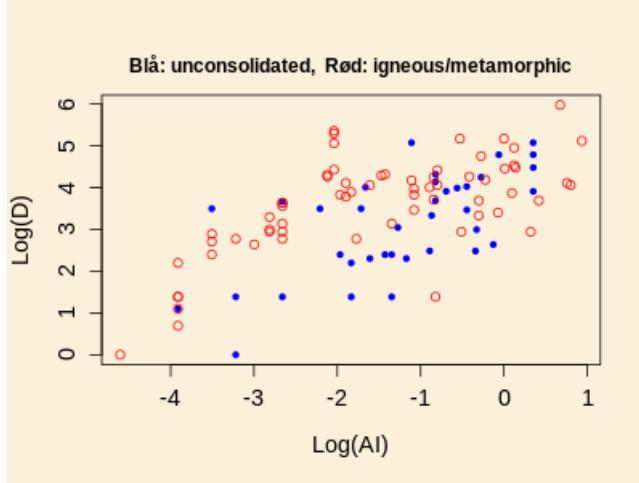
De fleste af jer har nok set reportager i nyhederne fra **mudderskred** rundt omkring i verdenen. For geologer er dette et af bidragene til **sedimenttransport** på skrânende flader. Det er forventeligt, at sedimenttransporten vil være større, jo større hældning en flade har, og geologerne formulerer dette som relationen

$$\text{transportrate} = D \cdot \text{hældning}.$$

Dette svarer lidt til en diffusionstypeligning. Det følgende billede viser et mudderskred i Virginia efter en orkan i 2004.



Jeg vil ikke her komme ind på, hvordan man estimerer en værdi af transportkoefficienten  $D$ , men i stedet se på en undersøgelse, hvor man prøver at beskrive  $D$  ud fra andre forhold såsom nedbørsmængde og jordforhold. I artiklen [Influences of climate and life on hillslope sediment transport](#) relateres data for  $D$  til et tørhedsindeks ([aridity index](#) AI) og til overfladestuktur ([lithology](#)) delt op på de to kategorier *unconsolidated* og *igneous/metamorphic*. Tørhedsindekset beregnes som gennemsnitlig årsnedbør divideret med et gennemsnitligt potentiel årsfordampningstal. Et tørhedsindeks på 1 svarer derfor til en form for "ligevægt" mellem nedbør og fordampning. Løseligt sagt, jo større tørhedsindeks jo mere vand er der til rådighed til sedimenttransport. Figuren nedenfor viser logaritmen til transportkoefficienten tegnet op mod logaritmen til tørhedsindekset for 102 områder delt op på 37 unconsolidated og 65 igneous/metamorphic. Vi vil betragte en model, hvor der for hver af de to overfladegrupper er en lineær sammenhæng mellem  $\log(D)$  og  $\log(AI)$ , og benytte denne model til at undersøge eventuelle forskelle mellem de to grupper.



I en generel formulering hr vi data fra  $n$  uafhængige stokastiske variable  $X_1, \dots, X_n$ , en forklarende variabel  $t = (t_1, \dots, t_n)$  og en faktor  $G$ , der inddeler data i  $k$  grupper (som her betegnes med tallene  $1, \dots, k$ ). Modellen, vi vil analysere, er

$$M_1: X_i \sim N(\xi_i, \sigma^2), \quad i = 1, \dots, n, \quad \xi_i = \alpha_{G_i} + \beta_{G_i} t_i, \\ (\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k, \sigma^2) \in \mathbf{R}^{2k} \times \mathbf{R}_+, \quad d(M_1) = 2k.$$

Denne model siger, at hver undergruppe, givet ved et bestemt niveau af  $G$ , har sin egen lineære sammenhæng mellem den forklarende variabel  $t$  og middelværdien af respons  $X$ . Model  $M_1$  har følgende naturlige undermodeller:

$$\begin{aligned} M_{2\alpha}: \quad & \xi_i = \alpha_{G_i} + \beta t_i, \quad d(M_{2\alpha}) = k + 1, \\ M_{2\beta}: \quad & \xi_i = \alpha + \beta_{G_i} t_i, \quad d(M_{2\beta}) = k + 1, \\ M_3: \quad & \xi_i = \alpha + \beta t_i, \quad d(M_3) = 2, \end{aligned}$$

hvor  $M_{2\alpha}$  er regressionsmodellen med fælles hældning og gruppespecifik skæring,  $M_{2\beta}$  er regressionsmodellen med fælles skæring og gruppespecifik hældning, og  $M_3$  er modellen med både fælles hældning og fælles skæring.

Den mest simple modelformel i **R** til analyse af model  $M_1$  er  $x \sim G * t$ . For at forstå den parametrisering, som **R** bruger, skal man vide, at **R** omskriver modelformen til  $x \sim G + t + G:t$ . Leddet  $G$  giver den gruppebestemte skæring  $\alpha_{G_i}$ , og i overensstemmelse med den ensidede variansanalysemødel fra afsnit 4.4 bruges parametrene Intercept =  $\alpha_1$  og forskellene  $\alpha_g - \alpha_1$ , der betegnes  $Gg$ ,  $g = 2, \dots, k$ . Leddet  $t$  giver regressionen for den første gruppe, det vil sige parameteren  $\beta_1$ , og  $G:t$  giver afvigelserne fra denne i de andre grupper, det vil sige  $\beta_g - \beta_1$ , som betegnes  $G:g:t$ . Den følgende tabel giver alternative måder at skrive modelformen på og de tilhørende parametriseringer i **R**.

Model	Modelformel	Parametre
$M_1$	$G * t$	$(\alpha_1, \alpha_2 - \alpha_1, \dots, \alpha_k - \alpha_1, \beta_1, \beta_2 - \beta_1, \dots, \beta_k - \beta_1)$
$M_1$	$G + t + G:t$	$(\alpha_1, \alpha_2 - \alpha_1, \dots, \alpha_k - \alpha_1, \beta_1, \beta_2 - \beta_1, \dots, \beta_k - \beta_1)$
$M_1$	$G - 1 + G:t$	$(\alpha_1, \alpha_2, \dots, \alpha_k, \beta_1, \beta_2, \dots, \beta_k)$
$M_1$	$G - 1 + t + G:t$	$(\alpha_1, \alpha_2, \dots, \alpha_k, \beta_1, \beta_2 - \beta_1, \dots, \beta_k - \beta_1)$
$M_{2\alpha}$	$G + t$	$(\alpha_1, \alpha_2 - \alpha_1, \dots, \alpha_k - \alpha_1, \beta)$
$M_{2\alpha}$	$G - 1 + t$	$(\alpha_1, \alpha_2, \dots, \alpha_k, \beta)$
$M_{2\beta}$	$t + G:t$	$(\alpha, \beta_1, \beta_2 - \beta_1, \dots, \beta_k - \beta_1)$
$M_{2\beta}$	$G:t$	$(\alpha, \beta_1, \beta_2, \dots, \beta_k)$
$M_3$	$t$	$(\alpha, \beta)$

Blandt undermodellerne  $M_{2\alpha}$  og  $M_{2\beta}$  er den første den vigtigste. Når  $E(X_i) = \alpha_{G_i} + \beta t_i$ , har vi en "additiv struktur" af  $G$  og  $t$ : uanset hvilken undergruppe der betragtes, er forskellen i middelværdier mellem to værdier af den forklarende variabel  $t$  den samme, og uanset hvilken værdi af den forklarende variabel der betragtes, er forskellen mellem to grupper den samme. I R laves  $F$ -testet fra Resultat 4.4 for reduktion fra model  $M_1$  til model  $M_{2\alpha}$  med kommandoen

```
anova(lm(x~G+t), lm(x~G*t))
```

$P$ -værdien for dette test findes fra en  $F(k - 1, n - 2k)$ -fordeling.

## 5.2 Analyse af sediment transport

Jeg vil i dette afsnit lave en analyse af data i det foregående afsnit omkring transportkoefficienten i sedimenttransport. Lad  $\log D_i$  være logaritmen til den beregnede transportkoefficient  $D_i$ , lad  $\log AI_i$  være logaritmen til tørhedsindeks  $AI_i$ , og lad  $G_i$  være overfladeforhold for det  $i$ 'te område, hvor  $G_i$  kan have de to værdier  $U$  og  $M$  for unconsolidated og metamorphic/igneous. Jeg starter med modellen

$$M_0: \log D_i \sim N(\alpha_{G_i} + \beta_{G_i} \cdot \log AI_i, \sigma_{G_i}^2), \quad i = 1, \dots, 102,$$

$$(\alpha_U, \alpha_M, \beta_U, \beta_M, \sigma_U, \sigma_M) \in \mathbf{R}^4 \times \mathbf{R}_+^2,$$

hvor målingerne er uafhængige, og alle parametrene kan variere frit. Denne model siger, at hver gruppe har sin egen lineære sammenhæng, og hver gruppe har sin egen spredning omkring den lineære sammenhæng. Et qqplot af residualer for hver af de to overfladegrupper viser ikke afvielse fra en normalfordeling. Et test for hypotesen om samme varians,  $\sigma_U^2 = \sigma_M^2$ , kan udføres som i afsnit 2.13 under brug af R-funktionen `var.test`. Input er de to output fra `lm` anvendt på hver af de to grupper af observationer. Kørsel af koden nedenfor viser, at  $p$ -værdien fra dette test er 0.56, og data strider derfor ikke mod hypotesen om samme varians. Koden viser også, hvordan man kan kalde `Bartlett.test` i denne situation.

### Showhide: QQplots

#### Kodevindue

```
D=c(50,70,33,55,12,32,56,54,63,4,4,4,33,9,3,10,11,11,
160,11,88,12,28,40,50,75,120,4,20,14,39,10,21,120,160,1,
33,55,70,41,46,58,46,44,49,158,212,38,38,35,20,19,27,14,
16,11,15,18,2,3,4,4,9,58,61,40,93,142,66,116,19,65,
32,53,75,73,58,61,16,71,74,84,200,23,19,40,28,86,48,88,
4,30,394,176,19,176,83,71,167,23,1,16)

AI=c(142,76,3,19,71,64,64,57,44,16,7,4,18,16,2,31,14,24,
```

```
33,26,142,41,42,44,50,44,94,26,72,88,7,20,28,142,142,4,
11,41,43,43,34,45,14,15,16,13,13,7,7,7,6,6,6,5,
4,3,3,3,2,2,2,2,2,223,213,152,113,113,80,76,60,33,
34,34,24,23,20,15,17,12,12,13,13,7,7,74,74,101,110,115,
44,93,196,100,138,59,45,66,255,26,1,7)/100
```

```
G=factor(rep(c("U", "M"), c(37, 65)))

logD=log(D)
logAI=log(AI)

lmUD1=lm(logD [G=="U"] ~ logAI [G=="U"])
lmUD2=lm(logD [G=="M"] ~ logAI [G=="M"])
qqnorm(lmUD1$residuals,
ylim=range(lmUD1$residuals ,lmUD2$residuals ))
points(qqnorm(lmUD2$residuals , plot=FALSE),
col=2,pch=20)

list(Ftest=var.test(lmUD1,lmUD2) ,
Bartlett=bartlett.test(list(lmUD1,lmUD2)))
```



Da vi kan antage, at de to varianser er ens, betragtes nu modellen

$$M_1: \text{Log}D_i \sim N(\alpha_{G_i} + \beta_{G_i} \cdot \text{log}AI_i, \sigma^2), \quad i = 1, \dots, 102,$$

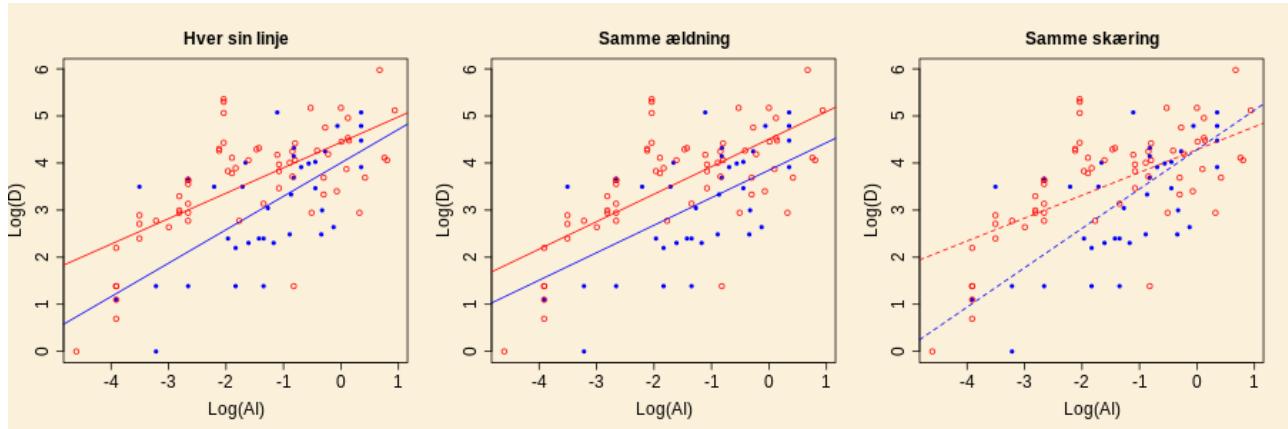
$$(\alpha_U, \alpha_M, \beta_U, \beta_M, \sigma) \in \mathbf{R}^4 \times \mathbf{R}_+.$$

Det venstre delplot i figuren nedenfor viser data med de to estimerede linjer indtegnet. Inden for dataområdet ligger linjen for unconsolidated sedimenter under linjen for igneous/metamorphic gruppen, selvom unconsolidated har en større hældning. Den større hældning er imidlertid ikke signifikant. Hvis vi laver  $F$ -testet for reduktion fra model  $M_1$  til den additive model  $M_{2\alpha}$ , hvor de to grupper af observationer har den samme hældning i den lineære sammenhæng, får vi en  $p$ -værdi på 0.28. Data strider altså ikke mod hypotesen  $\beta_U = \beta_M$ . Skøn og konfidensintervaller for parametrene i modellen med samme hældning er,

Parameter	Skøn	Konfidensinterval
Intercept	4.51	[4.20, 4.81]
$\alpha_U - \alpha_M$	-0.66	[-1.03, -0.29]
$\beta$	0.58	[0.45, 0.72]
$\sigma$	0.90	[0.83, 1.10]

Data med de estimerede linjer under modelen  $M_{2\alpha}$  er vist i det midterste delplot i figuren nedenfor. Da vores model er for logaritmetransformerede data, kan vi skrive de estimerede relationer kortfattet på formen  $D \approx e^{3.85} \cdot AI^{0.58}$  for unconsolidated gruppen, og  $D \approx e^{4.51} \cdot AI^{0.58}$  for igneous/-metamorphic gruppen. For samme værdi af tørhedsindekset har igneous/metamorphic gruppen en transportkoefficient, der næsten er dobbelt så stor som for unconsolidated gruppen (her taler vi om middelværdier). Et test, under model  $M_{2\alpha}$ , for hypotesen om samme skæring,  $\alpha_U = \alpha_M$ ,

giver en  $p$ -værdi på 0.0006. Modellen kan altså ikke reduceres til model  $M_3$ , hvor begge grupper har den samme lineære sammenhæng.



Vi reducerede ovenfor beskrivelsen af data fra model  $M_1$  til model  $M_{2\alpha}$ , hvor de to grupper har samme hældning. For disse data kan det imidlertid også give mening at betragte modellen med fælles skæring,  $M_{2\beta}$ . Dette skyldes, at når  $\log AI$  er nul, så er  $AI$  lig med 1, som betyder, at den årlige gennemsnitsnedbør er lig med den årlige gennemsnitsfordampning. En fælles skæring vil således være et udsagn om, at transportkoefficienten er uafhængig af overfladetype, når  $AI = 1$ . Et test for reduktion fra model  $M_1$  til model  $M_{2\beta}$  giver en  $p$ -værdi på 0.12, og vi siger derfor, at data ikke strider mod denne reduktion. De estimerede linjer under model  $M_{2\beta}$  er vist i det højre delplot i figuren ovenfor. Data i dette eksempel kan således lede frem til to vidt forskellige modeller, der begge synes fagligt rimelige. At dette kan lade sigøre hænger sammen med, at data udviser en meget stor spredning omkring de estimerede linjer.

I artiklen, hvor data stammer fra, deles data også op efter, hvordan transportkoefficienten beregnes, og der laves en analyse analog til analysen ovenfor, når data i stedet deles op efter beregningsmetoden. Desværre giver dette anledning til en stor usikkerhed omkring fortolkningen af resultaterne, vi kom frem til ovenfor. Det viser sig nemlig, at opdeling efter metode til beregning af transportkoefficienten i store træk er den samme som opdeling efter overfladetype (unconsolidated eller igneous/metomorphic). Vi kan derfor ikke vide, om de forskelle, der ses i data, skyldes overfladeforholdene eller skyldes beregningsmetoden!

### Showhide: F-test

#### Kodevindue

```
D=c(50,70,33,55,12,32,56,54,63,4,4,4,33,9,3,10,11,11,
160,11,88,12,28,40,50,75,120,4,20,14,39,10,21,120,160,1,
33,55,70,41,46,58,46,44,49,158,212,38,38,35,20,19,27,14,
16,11,15,18,2,3,4,4,9,58,61,40,93,142,66,116,19,65,
32,53,75,73,58,61,16,71,74,84,200,23,19,40,28,86,48,88,
4,30,394,176,19,176,83,71,167,23,1,16)

AI=c(142,76,3,19,71,64,64,57,44,16,7,4,18,16,2,31,14,24,
33,26,142,41,42,44,50,44,94,26,72,88,7,20,28,142,142,4,
11,41,43,43,34,45,14,15,16,13,13,7,7,7,6,6,6,5,
```

```
4,3,3,3,2,2,2,2,223,213,152,113,113,80,76,60,33,  
34,34,24,23,20,15,17,12,12,13,13,7,7,74,74,101,110,115,  
44,93,196,100,138,59,45,66,255,26,1,7)/100
```

```
G=factor(rep(c("U","M"),c(37,65)))  
logD=log(D)  
logAI=log(AI)  
  
list(M1TilM2alpha=anova(lm(logD~G+logAI),lm(logD~G*logAI)),  
M2alphaTilM3=anova(lm(logD~logAI),lm(logD~G+logAI)),  
M2alphaParameter=summary(lm(logD~G+logAI)),  
M2alphaKI=confint(lm(logD~G+logAI)),  
SigmaKI=sqrt(99*0.9028/qchisq(c(0.025,0.975),99)),  
M1TilM2beta=anova(lm(logD~G:logAI),lm(logD~G*logAI)))
```

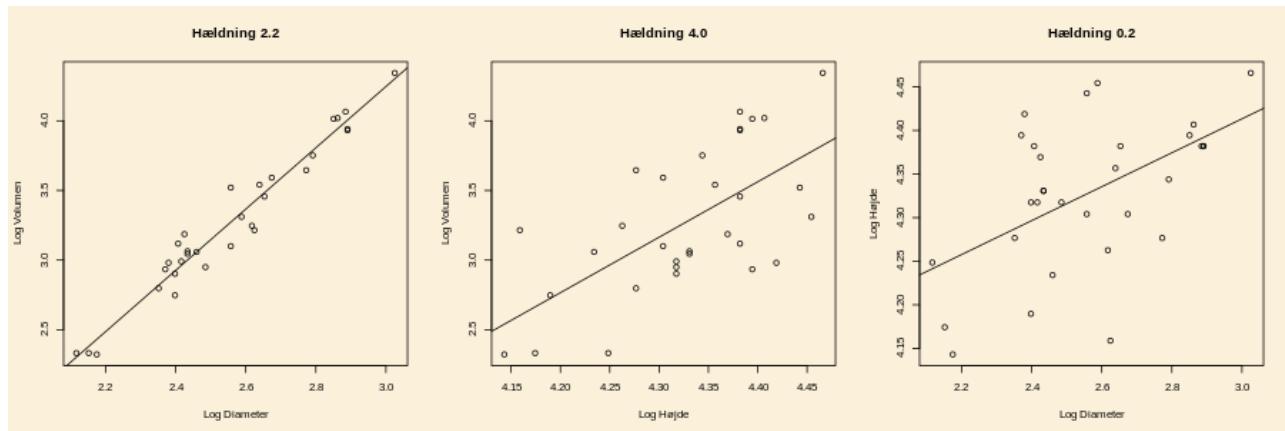


### 5.3 Den multiple regressionsmodel

Indbygget i R er et datasæt *trees*, der indeholder data for 31 [black cherry trees](#) fra Allegheny National Forest. Data stammer oprindeligt fra [The Minitab Student Handbook](#).



For hvert træ har man målt diameter (inches, målt 54 inches over jorden), højden (feet) og volumen af tømmer efter fældning (cubic feet). Figuren nedenfor viser logaritmen til volumen afsat mod henholdsvis logaritmen til diameter og logaritmen til højden. Den sidste delfigur viser logaritmen til højden afsat mod logaritmen til diameteren.



Det er tydeligt, at diameteren og højden hver især giver information om volumen, og diameteren er den, der bedst kan bruges til at forudsige volumen. Hældningen i en regression af log-volumen på log-diameter er 2.2, svarende til at volumen er proportional med diameter opløftet i 2.2. Dette kan umiddelbart være svært at fortolke, men hvis vi ser på regressionen af log-højde på log-

diameter, er hældningen her 0.2. Vi kan derfor sige, at  $diameter^{2.2}$  efterligner  $diameter^2 \cdot højde$ , hvilket intuitivt giver god mening. Spørgsmålet er, om man kan lave en model, hvor log-diameter og log-højde begge indgår, og dermed kan forbedre beskrivelsen af log-volumen? Den relevante modelklasse er *multipel regression*, som jeg vil beskrive i det følgende.

### 5.3.1 Den multiple regressionsmodel

Vi betragter målinger af  $n$  uafhængige stokastiske variable  $X_i$ ,  $i = 1, \dots, n$ . Til hvert observationsnummer  $i$  er der tilknyttet værdierne af  $d$  forklarende variable. I den simple lineære regressionsmodel i afsnit 3.1 blev værdien af den forklarende variabel betegnet med  $t_i$ . Når der er flere forklarende variable, lad os sige  $d$  af disse, betegnes værdierne med  $t_{ij}$ ,  $i = 1, \dots, n$  og  $j = 1, \dots, d$ . På denne måde passer index med en dataframestruktur, hvor  $i$  er rækkenummer og  $j$  er søjlenummer. Den  $j$ 'te forklarende variabel er vektoren  $t_j = (t_{1j}, t_{2j}, \dots, t_{nj})$ . I den *multiple regressionsmodel* er middelværdien af respons  $X_i$  en linearkombination af de  $d$  forklarende værdier. Vi skriver modellen på formen

$$\text{Model: } X_i \sim N(\alpha + \beta_1 t_{i1} + \beta_2 t_{i2} + \dots + \beta_d t_{id}, \sigma^2), \quad i = 1, \dots, n, \quad (5.1)$$

$$(\alpha, \beta_1, \dots, \beta_d, \sigma^2) \in \mathbf{R}^{d+1} \times \mathbf{R}_+.$$

De forklarende variable kaldes også *regressionsvariable*, og  $\beta_1, \dots, \beta_d$  kaldes *regressionskoefficienter*.

Modellen analyseres i R med kommandoen

```
lm(x~t1+t2+...+td)
```

hvor, i den konkrete situation,  $x$  skal erstattes af navnet på responsvariablen, og  $t1, t2, \dots, td$  skal erstattes med navnene på de forklarende variable, og summen af de  $d$  led skal skrives fuldstændigt ud. I parametertabellen fra *summary* er *Intercept* skønnet over  $\alpha$ , og skønnet  $\hat{\beta}_j$  over den  $j$ 'te regressionskoefficient står ud for navnet på den  $j$ 'te forklarende variabel (her  $t_j$ ). Den  $i$ 'te forventede værdi er  $\hat{\xi}_i = \hat{\alpha} + \hat{\beta}_1 t_{i1} + \dots + \hat{\beta}_d t_{id}$ , og skønnet over variansen i modellen er  $s^2 = \sum_i (x_i - \hat{\xi}_i)^2 / (n - d - 1)$ , idet middelværdimodellen har  $d + 1$  parametre.

Ligesom for den simple regressionsmodel i afsnit 5.5 kan vi være interesseret i middelværdien  $\xi^P = \alpha + \beta_1 t_{*1} + \dots + \beta_d t_{*d}$  for givne værdier  $t_{*1}, \dots, t_{*d}$  af de forklarende variable. Skønnet over denne,

$$\hat{\xi}^P = \hat{\alpha} + \hat{\beta}_1 t_{*1} + \dots + \hat{\beta}_d t_{*d}, \quad (5.2)$$

kaldes den *prædikterede værdi*. Et konfidensinterval for  $\xi^P$  beregnes i R med *predict* (med indstillingen *interval="confidence"*) som i afsnit 5.5, hvor der så skal bruges en dataframe

```
NyData=data.frame(t1=t_*1,...,td=t_*d)
```

Som i afsnit 5.5 kan man få et *prædiktionsinterval* i stedet, det vil sige et interval for en kommende observation, ved at lave indstillingen *interval="prediction"* i kaldet til *predict*.

**Eksempel 5.1.** (Cherry trees)

I det følgende kodevindue analyseres Black Cherry Trees datasættet omtalt ovenfor. Lad  $\logVol$  være logaritmen til volumen,  $\logDiam$  logaritmen til diameteren, og lad  $\logHoej$  være logaritmen til højden. Modellen, der analyseres, er

$$\text{LogVol}_i \sim N(\alpha + \beta_D \cdot \logDiam_i + \beta_H \cdot \logHoej_i, \sigma^2), \quad i = 1, \dots, 31,$$

hvor  $(\alpha, \beta_D, \beta_H, \sigma)$  kan variere frit. Kør koden.

#### Kodevindue

```
logDiam=log(trees[,1])
logHoej=log(trees[,2])
logVol=log(trees[,3])

summary(lm(logVol~logDiam+logHoej))
```

De estimerede regressionskoefficienter er henholdsvis 1.98 og 1.12, hvilket næsten svarer til modellen, hvor volumen beskrives som diameter<sup>2</sup> · højde.

Skønnet over spredningen er 0.081. Da dette er for logaritmetransformerede data, svarer dette cirka til en spredning på 8 procent på den oprindelige skala.

Hvis vi kun laver regression på logaritmen til diameteren, bliver spredningsskønnet 0.115. Dette kan formuleres på den måde, at inddragelsen af logaritmen til højden reducerer spredningen med knap 30 procent.

Det næste kodevindue laver plots af residualer mod henholdsvis logaritmen til diameteren og logaritmen til højden, og et normalt qqplot.

#### Showhide: Kodevindue

```
logDiam=log(trees[,1])
logHoej=log(trees[,2])
logVol=log(trees[,3])
r=lm(logVol~logDiam+logHoej)$residuals

layout(matrix(c(1,2,3,3),2,2,byrow=T))
plot(logDiam,r,xlab="Log_Diameter",ylab="Residualer")
abline(0,0,lty=3)
plot(logHoej,r,xlab="Log_Højde",ylab="Residualer")
abline(0,0,lty=3)
qqnorm(r,ylab="Residualer")
c()
```

Alle tre figurer understøtter den multiple regressionsmodel for disse data.



Lad os afslutte dette eksempel med at lave prædiktion af logaritmen til volumen for to nye træer med værdierne

	Log Diameter	Log Højde
Træ 1	2.5	4.3
Træ 2	2.2	4.1

Det følgende kodevindue lave konfidensintervaller for middelværdien for de to træer.

#### Showhide: Kodevindue

##### Kodevindue

```
logDiam=log(trees[,1])
logHoej=log(trees[,2])
logVol=log(trees[,3])
lmUD=lm(logVol~logDiam+logHoej)

predict(lmUD, data.frame(logDiam=c(2.5,2.2),logHoej=c(4.3,4.1)),
interval="confidence")
```

Kør koden. Kan du forklare, hvorfor det andet konfidensinterval er bredere end det første? Ændr koden, så der beregnes prædiktionsintervaller i stedet.

#### Svar 5.1. Cherry trees

##### Showhide: Svar: Cherry Trees

Det første af de to nye træer ligger midt i området for data, hvor middelværdien er velbestemt, hvorimod det andet træ ligger i udkanten af dataområdet.

For at lave et prædiktionsinterval skal man erstatte "confidence" med "prediction" i koden.



## 5.4 Stepvis regression

I en multipel regressionssituation ved man typisk ikke på forhånd, at alle de forklarende variable indeholder information om respons. Hvis man inkluderer mange variable, der ikke er relevante, kan dette give et forkert billede af afhængigheden af de relevante variable, og kan give et forkert

indtryk af, hvor godt respons kan beskrives (giver en for lille værdi af spredningsskøn  $s(M)$ ). Man taler i denne sammenhæng om "overfitting". Hvis for eksempel man har  $n$  observationer, og man har  $n$  eller flere forklarende variable, vil den multiple regressionsmodel give  $s(M) = 0$ , og de forventede værdier  $\hat{\xi}_i$  er lig med de observerede værdier. Vores mål må derfor være at finde en delmængde af de forklarende variable, der giver en god beskrivelse af respons, og som ikke overfitter. I en model med få forklarende variable er det nemmere at fortolke modellen, og parametrene vil være bedre bestemt end i en model med mange variable. Af to grunde er det dog ikke nemt at finde en passende delmængde af de forklarende variable. Hvis der er mange forklarende variablene, lad os sige  $d$ , hvor  $d$  er stor, så vil der være et meget stort antal mulige delmængder, nemlig  $2^d$ . Hvis  $d = 10$  giver dette 1024, og beregningsarbejdet begynder at være tungt. Selvom det er muligt at lave alle beregningerne, er det ikke oplagt, hvordan man vælger den bedste delmængde: hvordan skal vi vægte en lille værdi af  $s(M)$  i forhold til, hvor mange forklarende variable vi inddrager?

Nedenfor omtaler jeg to metoder til at vælge en delmængde af de forklarende variable, nemlig *backward selektion* og *forward selektion*. Der findes også metoder, der kombinerer de to, men dette vil jeg ikke beskrive her. De to metoder vil typisk ikke lede frem til den samme slutmodel, og ingen af metoderne forholder sig eksplisit til problematikken omkring overfitting. Spørgsmålet omkring overfitting vil jeg diskutere i afsnit 5.6. Når antallet af forklarende variable er lille, anvendes ofte *backward selektion*.

Ved backward selektion starter man med den størst mulige model (den fulde model), det vil sige modellen, hvor alle de forklarende variable indgår. Derefter fjerner man successivt en af de forklarende variable baseret på  $p$ -værdierne fra  $t$ -test af, at regressionskoefficienterne er nul.

### Definition 5.2. (Backward selektion)

I hvert trin køres `summary(lm(modelformel))` for den aktuelle model. Den største  $p$ -værdi blandt  $t$ -testene, for at en regressionskoefficient er nul, identificeres. Hvis denne  $p$ -værdi er større end en selvvalgt grænse (for eksempel 0.05), fjernes den tilhørende regressionsvariabel fra modellen. Proceduren stopper, når ingen af  $p$ -værdierne er over grænsen, og den tilhørende model kaldes slutmodellen.

Typisk vil man også supplere testene med en registrering af udviklingen af spredningsskønnnet  $s(M)$  i hvert trin, og til sidst lave et  $F$ -test for reduktion fra startmodellen til slutmodellen.

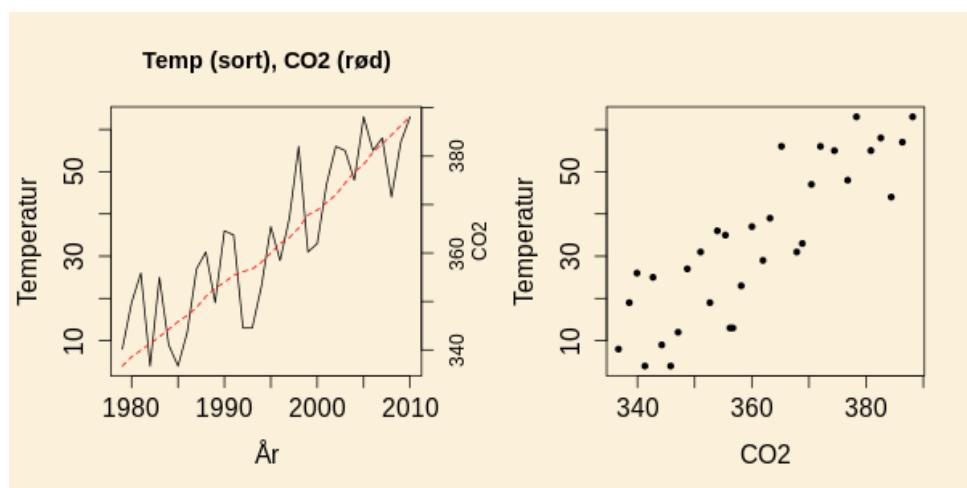
Ved forward selektion starter man med den mindst mulige model, det vil sige modellen uden nogen forklarende variabel, hvor alle de stokastiske variable har samme middelværdi. Man bygger dernæst modellen successivt op, ved i hvert trin at inkludere en ny forklarende variabel baseret på en væsentlig reduktion i spredningsskønnnet  $s(M)$ .

### Definition 5.3. (Forward selektion)

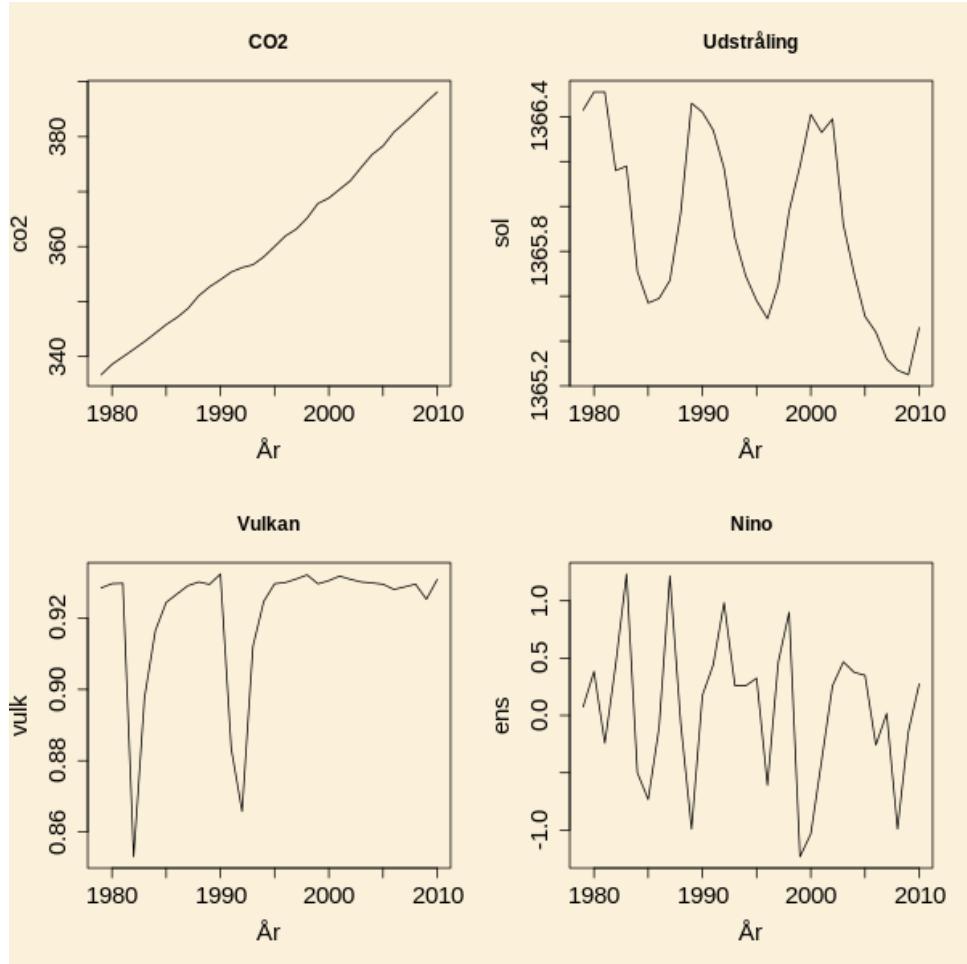
I hvert trin prøver man på skift at inkludere en af de variable, der endnu ikke er med i modellen. Fra `summary(lm(modelformel))`, for modellen med den ekstra variabel inkluderet, findes spredningsskønnnet  $s(M)$  og  $p$ -værdien for test af hypotesen, at regressionskoefficienten hørende til den ekstra variabel er nul. Når alle variable er afprøvet, vælges den variabel med det laveste spredningsskøn, og hvis den tilhørende  $p$ -værdi er under en selvvalgt grænse (for eksempel 0.05), inkluderes variablen i modellen, og forward selektionsproceduren fortsættes. Proceduren stopper, når  $p$ -værdien for den valgte variabel er over grænsen.

### 5.4.1 Eksempel

I dette eksempel vil jeg se på en beskrivelse af den globale årstemperatur ud fra en række forklarende variable. Dette er et emne, hvor man skal passe på med ikke at komme med "amatørudsagn", og jeg støtter mig da også til fremstillingen i artiklen [Using Data from Climate Science to Teach Introductory Statistics](#). Figuren nedenfor viser den globale temperaturs afvigelse fra et 30 års gennemsnit for årene 1979 til 2010 (afvigelse i grader celcius ganget med 100) og udviklingen af co2 i atmosfæren i den samme periode (ppm, parts per million). Begge grafer viser et stigende forløb hen over perioden, og derfor vil en regression af temperatur på co2 vise en tydelig sammenhæng, som i højre delfigur nedenfor.



Spørgsmålet er, om der er andre effekter end co2, der kan forklare temperaturstigningen? De kandidater, jeg vil inddrage, er solens udstråling (solar irradiance, watt/m<sup>2</sup>), vulkanaktivitet og et mål for el Nino effekten. Vulkanaktiviteten måles ved "MLO apparent transmission values", som angiver den procentdel af solen udstråling, der rammer jordoverfladen, og el Nino effekten måles som temperaturoafvigelse i grader celcius. I den næste figur er de fire forklarende variable vist som funktion af tidspunkt.



Umiddelbart er der ikke nogen af de tre nye forklarende variable, der viser en stigende eller faldende tendens hen over perioden, men måske er der et sammenspil mellem dem, som ikke umiddelbart kan ses. Dette kan vi få viden om ved at bruge en multipel regressionsmodel. Lad  $T$  være temperatur,  $C$  mængden af CO<sub>2</sub> i atmosfæren,  $S$  solens udstråling,  $V$  vulkanaktiviteten og  $E$  el Niño effekten. Jeg vil vise opbygningen af regressionsmodellen ved brug af forward selektion. Den følgende tabel viser for en række modeller skøn over spredning  $s(M)$  og  $p$ -værdi for test af  $\beta = 0$ , hvor  $\beta$  er regressionskoeficienten hørende til det sidste led i modelformlen.

Model	$C$	$S$	$V$	$E$
$s(M)$	9.5	17.5	16.8	18.6
$p$ -værdi	$2.6 \cdot 10^{-10}$	0.057	0.013	0.63

Model	$C + S$	$C + V$	$C + E$
$s(M)$	9.2	9.1	8.5
$p$ -værdi	0.104	0.081	0.007

Model	$C + E + S$	$C + E + V$
$s(M)$	8.0	7.1
$p$ -værdi	0.040	0.001

Model	$C + E + V + S$
$s(M)$	6.3
$p$ -værdi	0.006

Den første del af tabellen viser, at hvis vi kun laver regression på en af de forklarende variable, får vi den bedste beskrivelse ved at bruge co2 ( $s(M) = 9.5$ , og  $p$ -værdi for at fjerne co2 er  $2.6 \cdot 10^{-10}$ ). Prøver vi dernæst at udvide modellen med yderligere en forklarende variabel ud over co2, ser vi i den anden del af tabellen, at vi skal vælge el Nino ( $s(M) = 8.5$ , og  $p$ -værdi for at fjerne el Nino er 0.007). Prøver vi at udvide modellen med både co2 og el Nino, ses fra den tredje del af tabellen, at vi skal vælge vulkanaktivitet ( $s(M) = 7.1$ , og  $p$ -værdi for at fjerne vulkanaktivitet er 0.001). Endelig ses af den sidste del af tabellen, at vi også vil inkludere solens udstråling i modellen, idet spredningsskønnet reduceres til 6.3, og  $p$ -værdi for at fjerne solens udstråling fra modellen er 0.006.

I regressionsmodellen med alle fire forklarende variable

$$T_i \sim N(\alpha + \beta_C C_i + \beta_E E_i + \beta_V V_i + \beta_S S_i, \sigma^2), \quad i = 1, \dots, 32,$$

$$(\alpha, \beta_C, \beta_E, \beta_V, \beta_S, \sigma) \in \mathbf{R}^5 \times \mathbf{R}_+,$$

får vi følgende parametertabel med kaldet `summary(lm(T~C+E+V+S))`.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.412e+04	4.524e+03	-3.121	0.004264 **
co2	1.129e+00	9.150e-02	12.333	1.32e-12 ***
ens	1.014e+01	1.926e+00	5.263	1.50e-05 ***
vulk	2.808e+02	6.548e+01	4.288	0.000206 ***
sol	9.873e+00	3.297e+00	2.995	0.005817 **
<hr/>				
Residual standard error: 6.292 on 27 degrees of freedom				

Tabellen viser, at i slutmodellen fra forward selektion (som er den fulde model med alle de forklarende variable inkluderet) er alle de forklarende variable vigtige, alle fire  $p$ -værdier er ganske små. I dette eksempel har vi derfor også, at *backward selektion* vil give den fulde model, ingen led fjernes.

Det er usædvanligt, at med kun 32 datapunkter og med fire forklarende variable at alle variable er stærkt signifikante ( $p$ -værdi langt under 0.05). Det er også interessant, at når vi inkluderer flere og flere led, ændres betydningen af co2 kun lidt, og regressionskoefficienten hørende til co2 er mere velbestemt i slutmodellen end i startmodellen. Dette er vist i nedenstående tabel med skøn  $\hat{\beta}_C$  og standard error  $sd_s(\hat{\beta}_C)$  i de forskellige modeller. Tabellen viser samme tendens for estimation af de andre regressionskoefficienter.

Model	$C$	$C + E$	$C + E + V$	$C + E + V + S$
$\hat{\beta}_C$	1.03	1.08	0.99	1.13
$sd_s(\hat{\beta}_C)$	0.11	0.10	0.09	0.09
<hr/>				
$\hat{\beta}_E$		7.1	9.6	10.1
$sd_s(\hat{\beta}_E)$		2.5	2.2	1.9
<hr/>				
$\hat{\beta}_V$			269	281
$sd_s(\hat{\beta}_V)$			74	65

Vi kan se, at regressionskoefficienten for co2 ligger omkring 1, hvilket betyder, at hvis mængden af co2 får lov til at stige med 100 ppm, så tyder denne meget simple model på en stigning på 1 grad celcius i den globale temperatur.

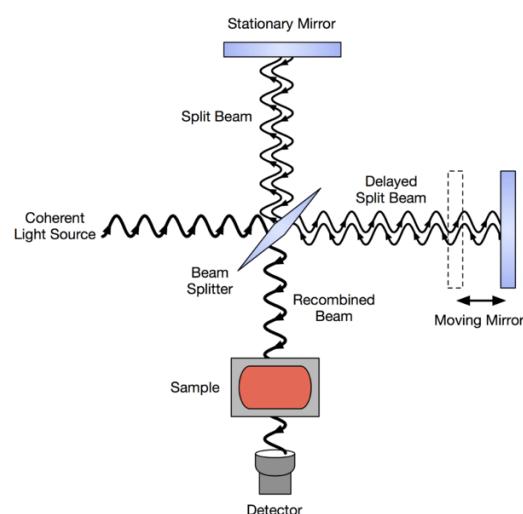
### Showhide: Datakilder

Fremstillingen i dette eksempel er som nævnt inspireret af artiklen [Using Data from Climate Science to Teach Introductory Statistics](#). Data for temperatur, solens udstråling og co2 er taget fra artiklen, men stammer oprindeligt fra temperaturens vedkommende fra [NASA Goddard Institute for Space Studies \(GISS\)](#), data for solens udstråling stammer fra [The Physikalisch-Meteorologisches Observatorium Davos/World Radiation Center](#), og data for co2 stammer fra [Earth System Research Laboratory of the National Oceanic and Atmospheric Administration \(NOAA\)](#). Data for vulkanaktivitet og el Nino effekten er ikke givet i artiklen, men der er givet links til data. Data for vulkanaktivitet er fra [Earth System Research Laboratories](#) og data for el Nino effekten er fra [Climate Prediction Center](#).



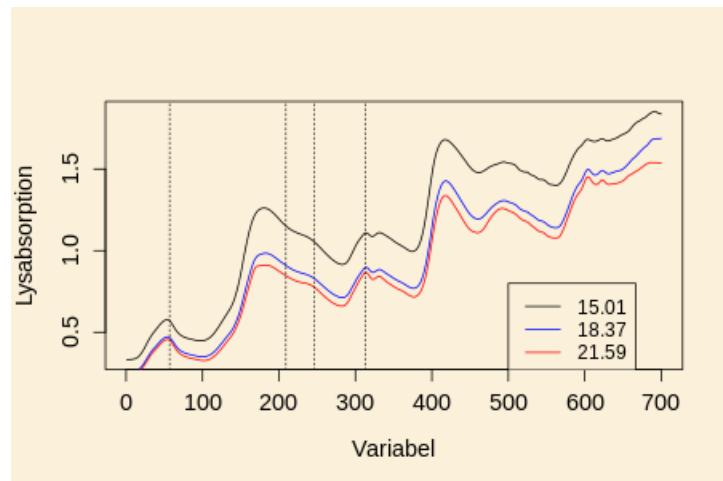
## 5.5 Datasæt med et stort antal forklarende variable

Near-infrared spectroscopy (NIRS) er en måleteknik, der benyttes mere og mere i fødevareindustrien såvel som i mange andre områder. Ideen er, at ved at sende lys af forskellige bølgelængder gennem en prøve kan man få viden om sammensætning af prøven. For eksempel kan vi være interesseret i at kunne vurdere mængden af fedtstoffer (respons) i en prøve ud fra lysabsorbsionen ved en række forskellige bølgelængder. Her er lysabsorbsionen ved en bestemt bølgelængde en forklarende variabel, og antallet af forklarende variable bliver det antal bølgelængde, som der måles ved. Lysabsorbsionen ved de forskellige bølgelængder kaldes tilsammen *spektrum* for prøven. Som et konkret eksempel vil jeg se på fedtprocenten i 40 prøver af småkagedej, hvor lysabsorbsionen er målt ved 700 bølgelængder i området fra 1100-2500 nanometer.

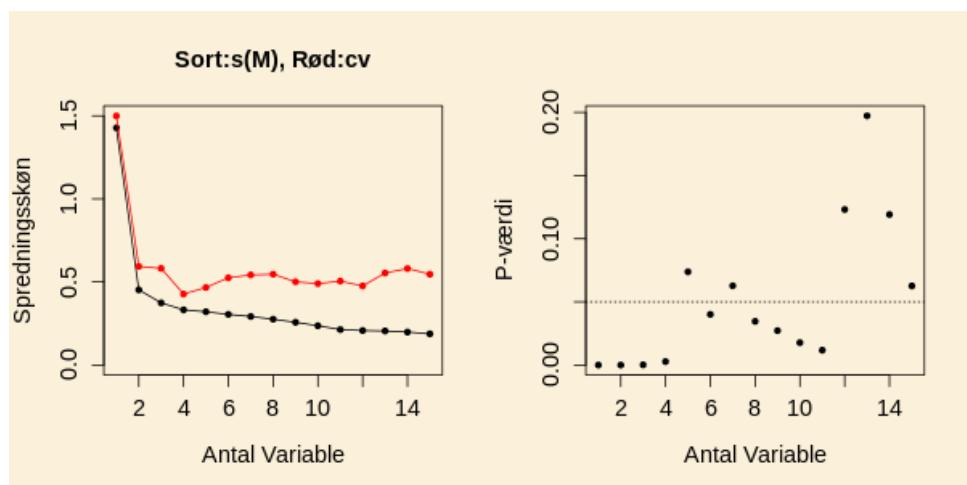


Data er oprindeligt fra artiklen [Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs](#), og er her hentet fra data hørende til artiklen [Comparison of Multi-response Prediction Methods](#). Oprindeligt er der 40 prøver, men forfatterne vurderer, at der er sket fejl i målingen af den ene prøve, hvorfor vi kun betragter de resterende 39.

I den følgende figur er vist spektrum for prøven med mindst fedtstof (15.01 procent), for prøven med mest fedtstof (21.59 procent) og for en prøve med fedtstofmængde midt mellem de to.



I det konkrete eksempel har vi således mange flere forklarende variable end antallet af datapunkter. Dette gør det svært at konstruere en multipel regressionsmodel. Hvis man prøver at lave regression på alle variablene, vil man fitte en model, hvor alle de forventede værdier bliver lig med de observerede responsværdier, og skønnet over spredningen bliver  $s(M) = 0$ . Man kan kalde dette en ekstrem grad af "overfitting". Hvis man prøver at etablere en model ved forward selektion, vil, qua metoden, spredningsskønnet  $s(M)$  typisk blive ved at falde jo flere variable, der inkluderes. Når vi har mange forklarende variable, vil der også være en del, der ved rene tilfældigheder ser ud til at være korreleret med respons, hvorfor disse inkluderes i modellen. Da vi hele tiden prøver at minimere  $s(M)$  gennem vores valg af forklarende variable, giver  $s(M)$  ikke et retvisende billede af, hvor god modellen er til at beskrive data. Den følgende figur viser udviklingen af spredningsskønnet  $s(M)$  under forward selektionsalgoritmen (sorte kurve i venstre delplot).



Figuren viser et kraftigt fald i spredningsskønnet  $s(M)$ , når modellen udvides fra 1 variabel til 4 variable ved forward selektion, og tilsvarende meget små  $p$ -værdier for test af, at regressionskoefficienten for det sidst inkluderede led i modellen er nul (højre delplot). Derefter er faldet i  $s(M)$  mindre, og  $p$ -værdierne tilsvarende større med værdier omkring 0.05 indtil 11 variable er inkluderet, hvor  $p$ -værdierne tager et yderligere spring opad. Med en grænse på  $p$ -værdien på 0.05 stopper forward proceduren, efter at 4 variable er inkluderet. Med 4 forklarende variable ved

forward selektion er spredningskønnet 0.33, og med 11 forklarende variable er spredninsskønnet 0.21.

Som nævnt, kan vi ikke regne med at spredningsskønnet er retvisende, og der er behov for en alternativ måde at lave et spredningsskøn på. Den røde kurve i figuren ovenfor viser netop sådan en alternativ metode, som bliver gennemgået i næste afsnit. Denne alternative metode peger på, at forward selektion med 4 variable er passende for disse data og giver et spredningsskøn (eller rettere et skøn over prædiktionsfejl: se næste afsnit) på 0.43. At der kun inkluderes 4 variable afspejler også, at vi kun har 40 prøver til rådighed for at etablere modellen. Med flere prøver til rådighed vil det være forventeligt, at forward selektion vil inkludere flere variable (se afsnit 5.8).

## 5.6 Cross Validation

Motivationen for at betragte NIR-spektret i det foregående afsnit er, at vi ønsker at konstruere en metode, hvormed man kan prædiktere respons (i vores tilfælde mængden af fedtstof). Det er derfor naturligt at tænke på at evaluere en metode ved at se på, hvor god metoden er til at prædiktere et nyt sæt observationer. Man taler generelt om *prædictionsspredningen* (root mean squared error of prediction: RMSEP) og *prædiktionsvariansen*, hvor man i den sidste tager gennemsnit over de kvadrerede prædiktionsfejl, det vil sige afstand mellem den observerede værdi og den prædikterede værdi. Problemet i denne tankegang er, at man typisk ikke har et nyt sæt observationer. I stedet deler vi det oprindelige datasæt op i et *træningssæt* og et *testsæt*, benytter træningssættet til at konstruere vores prædiktionsmetode og bruger testsættet til at se, hvor god metoden er. Typisk laver man så flere forskellige opdelinger i træningssæt og testsæt for, at testsættet kommer rundt i hele det oprindelige datasæt. Dette kendes under navnet *cross validation* (*krydsvalidering* på dansk). Når man som her lægger vægten på prædictionsspredningen, betragtes den multiple regressionsmodel også som en del af emneområdet *machine learning*.

Lad mig først prøve at beskrive cross validation abstrakt. Vi har  $n$  datapunkter og ønsker at evaluere en estimationsmetode. Cross validation kan beskrives gennem følgende punkter.

1. Del på tilfældig vis data op i et træningssæt og et testsæt.
2. Gennemfør estimationsmetoden på træningssættet, og find skøn over de parametre, der indgår i slutmodellen.
3. Lav for hver prøve i testsættet en prædikteret værdi ud fra de forklarende værdier hørende til prøven og parameterskønnene fra foregående punkt. Beregn dernæst *prædiktionsfejl* som responsværdi minus den prædikterede værdi.
4. Gentag punkt 1-3 med andre inddelinger i træningssæt og testsæt, således at de forskellige testsæt kommer rundt i hele datasættet.
5. Beregn *prædiktionsvarians* som gennemsnit over alle de kvadrerede værdier af prædiktionsfejlene. Beregn *prædictionsspredning* (root mean squared error of prediction, RMSEP) som kvadratroden af prædiktionsvariansen.

Lad os nu betragte cross validation i forbindelse med forward selektionsmetoden i den multiple regressionsmodel. Antag, at vi i træningsdelen har udvalgt de forklarende variable  $j_1, j_2, \dots, j_k$  og

har fået skønnene  $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k$  i den multiple regressionsmodel  $X_i \sim N(\alpha + \beta_1 t_{ij_1} + \dots + \beta_k t_{ijk}, \sigma^2)$ . For et observationsnummer  $i$  i *testsættet* bliver den prædikterede værdi

$$\hat{\xi}_i^{\text{cv}} = \hat{\alpha} + \hat{\beta}_1 t_{ij_1} + \dots + \hat{\beta}_k t_{ijk},$$

og bidraget til *prædiktionsvariansen* er  $(x_i - \hat{\xi}_i^{\text{cv}})^2$ .

I *Leave one out cross validation* (LOOCV) lader man testsættet bestå af kun en enkelt observation, og træningssættet er de resterende  $n - 1$  observationer. Dette gentager man  $n$  gange, hvor i det  $i$ 'te trin observation nummer  $i$  udgør testsættet. Hvis vi lader  $\hat{\xi}_i^{(-i)}$  betegne den prædikterede værdi for den  $i$ 'te observation, når træningssættet består af alle observationerne pånær den  $i$ 'te, kan vi skrive cross validation skønnet over prædiktionsspredningen som

$$s_{\text{cv}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{\xi}_i^{(-i)})^2}.$$

I  $K$ -fold cross validation deler man datasættet op i  $K$  cirka lige store dele, og hver del er så efter tur testsættet. Opdelingen er tilfældig og kan eventuelt gentages en række gange.

I figuren med spredningsskøn i det [foregående afsnit](#) er den røde kurve i venstre delplot prædiktionsspredningen fra leave one out cross validation. For  $j$  variable ( $j$  på førsteaksen i den omtalte figur) foregår beregningen som følger. I det  $i$ 'te trin i beregningen fjernes den  $i$ 'te observation fra datasættet. Dernæst laves der forward selektion med  $j$  variable, den multiple regressionsmodel med de  $j$  fundne variable estimeres, der laves prædiktion for den  $i$ 'te observation, som netop ikke var med i træningssættet, og bidraget til prædiktionsvariansen beregnes. Dette gentages for  $i = 1, \dots, n$ . Til sidst beregnes kvadratroden af gennemsnit af de  $n$  kvadrerede prædiktionsfejl. Figuren peger på, at der efter inklusion af fire variable ved forward selektion ikke længere sker en forbedring i evnen til at prædiktere fedtindholdet i dejprøver. Cross validation giver altså i eksemplet med dejprøver anledning til et antal variable, der passer med forward selektionsmetoden, hvor der bruges en grænse på  $p$ -værdien på 0.05. Det er dog sjældent, at vi ser en sådan overensstemmelse.

## 5.7 Beregning i R

Jeg har lavet en funktion *FWstep*, der laver beregninger til et enkelt trin i forward selektion. Koden til funktionen ligger i filen *Rfunktioner.txt* og er kopieret ind i kodevinduet nedenfor. Input til funktionen er en  $n \times d$  matrice med værdierne af de  $d$  forklarende variable for  $n$  prøver, en vektor med responsværdierne og en vektor med numrene på de variable, der allerede er inkluderet i modellen. Hvis matricen hedder  $T$ , og responsvektoren  $x$ , starter man med kaldet *FWstep*( $T, x$ ). Fra output aflæses hvilken variabel, man vil starte med at inkludere i modellen, og næste kald til funktionen bliver *FWstep*( $T, x, c(j_1)$ ), hvor  $j_1$  er nummeret på den variabel, der inkluderes. Efter at variablen  $j_1$  er inkluderet i modellen, findes det næste led ved kaldet *FWstep*( $T, x, c(j_1)$ ), dernæst *FWstep*( $T, x, c(j_1, j_2)$ ), og så videre. Output fra et kald til *FWstep* er spredningsskønnet  $s(M)$  fra den multiple regressionsmodel med et nyt led inkluderet (altså et led mere end i kaldet til *FWstep*), en vektor med numrene på de variable der indgår i modellen, og en vektor med  $p$ -værdierne for test af at en regressionskoefficient er nul. Det er den sidste indgang i vektoren med

*p*-værdier, der bruges til at vurdere, om man vil inkludere det sidst fundne led, og dermed om man vil fortsætte forward selektionsproceduren.

I kodevinduet nedenfor hentes datasættet fra afsnit 5.4, men i stedet for at betragte indholdet af fedtstof i dejprøverne ses på indholdet af sukker. Vi har 39 prøver og 700 forklarende variable.

### Showhide: Kodevindue

#### Kodevindue

```
FWstep=function(T,x,med=c()){
d=dim(T)[2]; n=length(x)
ma=ls.diag(lsfit(T[,1],x))$std.dev
res=rep(ma,d)
if (length(med)==0){lookup=c(1:d)} else {lookup=c(1:d)[-med]}
for (i in lookup){
med1=c(med, i)
res[i]=ls.diag(lsfit(T[,med1],x))$std.dev
}
sny=min(res)
medny=which.min(res)
med1=c(med, medny)
ud=lsfit(T[,med1],x)
uddiag=ls.diag(ud)
betahat=ud$coef
sds=uddiag$std.err
pval=2*pt(-abs(betahat/sds),n-1-length(med1))
return(list(sny=sny,med=med1,pval=pval))
}

load(url("https://data.mendeley.com/public-files/datasets/3ympjxywdm/files/9a54ab5b-"))
T00=NIR_Dough$NIR
x00=NIR_Dough$ingredient[,2] # sucrose
T=T00[1:40,]
x=x00[1:40]
# Fjerne outlier
T=T[-23,]
x=x[-23]

FWstep(T,x)
```

Det essentielle trin i beregningen er "for-løkken inde i *FWstep*-funktionen. Her prøver man for alle de forklarende variable, der endnu ikke er med i modellen, at tilføje en af disse og beregne spredningskønnet  $s(M)$ , når variablen er tilføjet. I den efterfølgende kommando *which.min(res)* finder man nummeret på den variabel, der giver den mindste værdi af  $s(M)$ . Estimation af den multiple regressionsmodel foregår ikke her med *lm*, som I ellers er vant til, men med funktionen

*lsfit*. I den sidste er input ikke en modelformel, men derimod en matrice med værdierne af de forklarende variable. Efter at den nye variabel er fundet, estimeres den multiple regressionsmodel, og  $p$ -værdier beregnes ud fra standard errors og opslag i en  $t$ -fordeling.



Kør koden, og se, at den første variabel, der medtages, er nummer 25, og at den tilhørende  $p$ -værdi er  $3.73 \cdot 10^{-4}$ . Kør nu koden igen, hvor FWstep(T, x) udskiftes med FWstep(T, x, c(25)). Her vil du se, at variablen nummer 488 inkluderes, og den tilhørende  $p$ -værdi er  $2.79 \cdot 10^{-6}$ . Fortsæt selv forward selektionsalgoritmen, og overvej, hvornår du vil stoppe.

Jeg har også lavet en funktion *FWcrossval* til at lave cross validation (leave one out) beregningerne for en forward selektionsmodel med et givet antal variable. Input til funktionen er  $n \times d$  matricen med værdierne af de forklarende variable, vektoren med responsværdierne og et tal, der angiver, hvor mange led der skal medtages i forward selektionsmodellen. I kodevinduet nedenfor kaldes funktionen med 8 som antal led i forward selektionsmodellen. Output fra *FWcrossval* er en vektor med prædiktionsspredningen beregnet ved leave one out cross validation, når der medtages henholdsvis 1 variabel i forward selektion, 2 variable og så videre op til de 8 variable, der blev angivet i kaldet til funktionen.

### Showhide: kodevindue

#### Kodevindue

```
FWcrossval=function(T,x,k){
d=dim(T)[2]
n=length(x)
Mr=matrix(0,n,k)
for (i in 1:n){
T0=T[-i,]
x0=x[-i]
res=rep(0,d)
for (j in 1:d){
res[j]=summary(lm(x0~T0[,j]))$sigma
}
med=which.min(res)
beta=lsfit(T0[,med],x0)$coef
Mr[i,1]=(x[i]-sum(beta*c(1,T[i,med])))^2
if (k>1){
for (j in 2:k){
res=rep(10,d)
for (r in c(1:d)[-med]){
med1=c(med,r)
res[r]=ls.diag(lsfit(T0[,med1],x0))$std.dev
}
med=c(med,which.min(res))
beta=lsfit(T0[,med],x0)$coef
Mr[i,j]=(x[i]-sum(beta*c(1,T[i,med])))^2
}}}
```

```
}
```

```
return (sqrt (apply (Mr, 2 ,sum) /n))
```

```
}
```

```
load (url ("https://data.mendeley.com/public-files/datasets/3ympjxywdm/files/9a54ab5b-
```

```
T00=NIR_Dough$NIR
x00=NIR_Dough$ingredient[,2] # sucrose
T=T00[1:40,]
x=x00[1:40]
# Fjerne outlier
T=T[-23,]
x=x[-23]

FWcrossval(T,x,8)
```

I funktionen *FWcrossval* indeholder matricen *Mr* alle de kvadrerede prædiktionsfejl, rækker svare til observationsnummer og såle angiver, hvor mange led der tages med i forward selektion. "For-løkken" over *i* er selve cross validation skridtet, hvor den *i*'te observation tages ud, og træningssættet består af de resterende *n*-1 observationer. For hvert træningssæt geneføres forward selektion, og hver gang en ny variabel er tilføjet, beregnes den prædikterede værdi for den observation, der er udeladt af træningssættet.

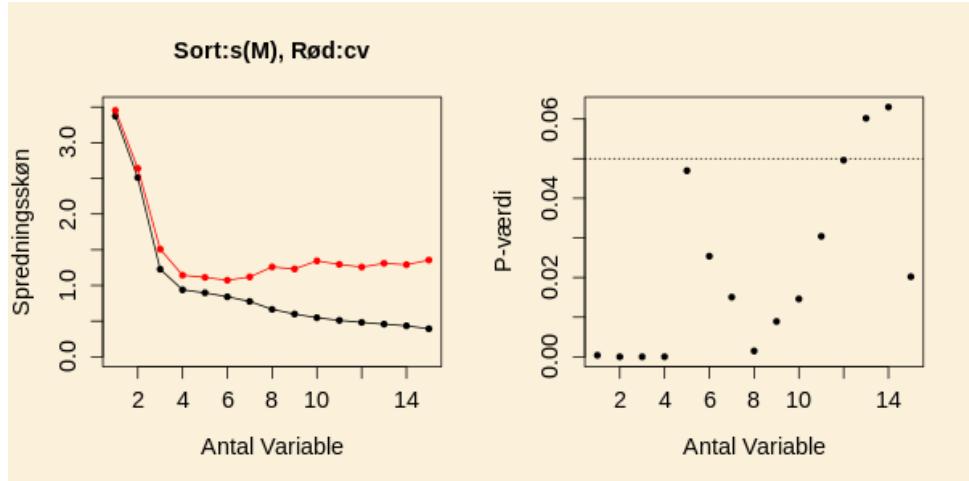


Kør koden (det tager nogle minutter), og vurder ud fra output, hvor mange led du vil medtage i din model.

### Showhide: Resultat

Resultaterne ved at køre *FWstep* 15 gange er vist i figuren nedenfor. Ud fra figuren med *p*-værdier vil man nok vælge forward selektionsmodellen med 4 variable, hvorimod den røde kurve med prædiktionspredningen fra leave one out cross validation peger på et sted mellem 4-7 variable (minimum af prædiktionspredningen er for 6 variable). Formelt vil man ved forward selektion med en grænseværdi på 0.05 medtage 12 variable, men cross validation viser, at dette ikke er en god model.

Skønnet over spredningen fra den multiple regressionsmodel med de 6 variable fra en forward selektion er  $s(M) = 0.84$ , hvorimod prædiktionspredningen er  $s_{cv} = 1.07$ . Sukkerprocenten ligger i disse data spredt jævnt ud mellem 10 og 23. Med en spredning på cirka 1 er der, set relativt til variationsområdet, en stor usikkerhed i bestemmelsen af sukkerindholdet.



## 5.8 Prædiktion på nyt datasæt

For data omkring småkagedej i de to foregående afsnit indsamlede man efter det første eksperiment yderligere 32 prøver (hvor igen en af prøverne blev vurderet til at være fejlbehæftet). Vi kan derfor tale om, at vi her har et uafhængigt testsæt. Hvor jeg i afsnit 5.6 beregnede en prædiktionsspredning baseret på leave one out cross validation, kan jeg her beregne en prædiktionsspredning baseret på de nye data. Jeg bruger således de oprindelige 39 observationer til at estimere en forward selektionsmodel og tester, hvor god denne er til at prædiktere observationerne i det nye testsæt med 31 observationer. I kodevinduet nedenfor er vist beregningen for data omkring fedtprocent i tilfældet med 4 forklarende variable fra forward selektion.

### Kodevindue

```
load(url("https://data.mendeley.com/public-files/datasets/3ympjxywdm/files/9a54ab5b-"))
T00=NIR_Dough$NIR
x00=NIR_Dough$ingredient[,1] # fedtprocent
T=T00[1:40,]
x=x00[1:40]
# Fjerne outlier
T=T[-23,]
x=x[-23]
# Testdata (øgte testsdata: indsamlet i et andet eksperiment)
Tt=T00[41:72,]
xt=x00[41:72]
# Fjerne outlier fra testdata
Tt=Tt[-21,]
```

```

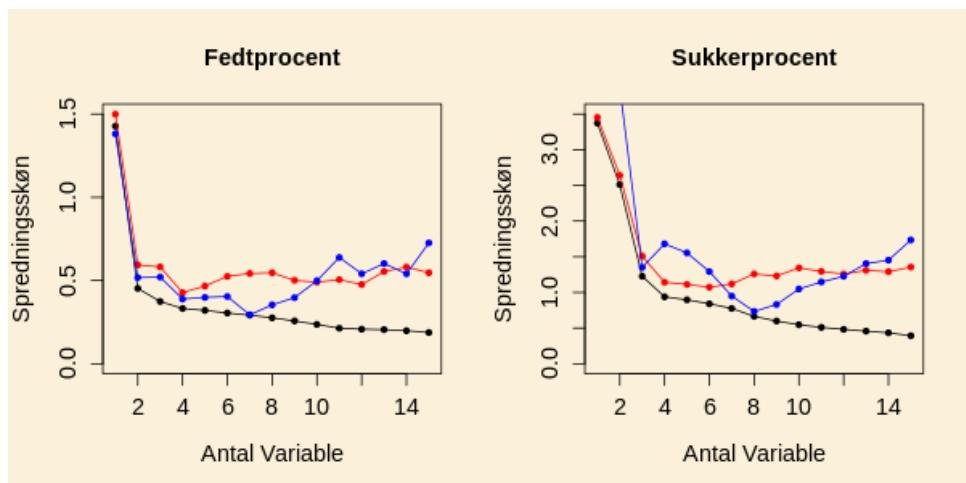
xt=xt[-21]

# De fire variable fra forward selektion
med=c(246,313,209,57)
# Estimation
beta=lsfit(T[,med],x)$coef
# prædiktion
pre=beta[1]+Tt[,med]%%beta[-1]
# Prædictionsspredning
sqrt(mean((xt-pre)^2))

```

Når I kører koden, vil I se, at prædictionsspredningen fra det uafhængige testsæt er 0.39. Vi fandt tidligere at prædictionsspredningen fra cross validation var 0.43, hvilket viser god overensstemmelse mellem de to tal.

Figuren nedenfor viser den tidligere figur med spredningskøn  $s(M)$  og prædictionsspredningen fra cross validation (sort og rød), og hvor nu også prædictionsspredningen fra det uafhængige testsæt er inkluderet (blå). Den venstre figur er for data omkring fedtprocenten og den højre figur er for data omkring sukkerprocenten.



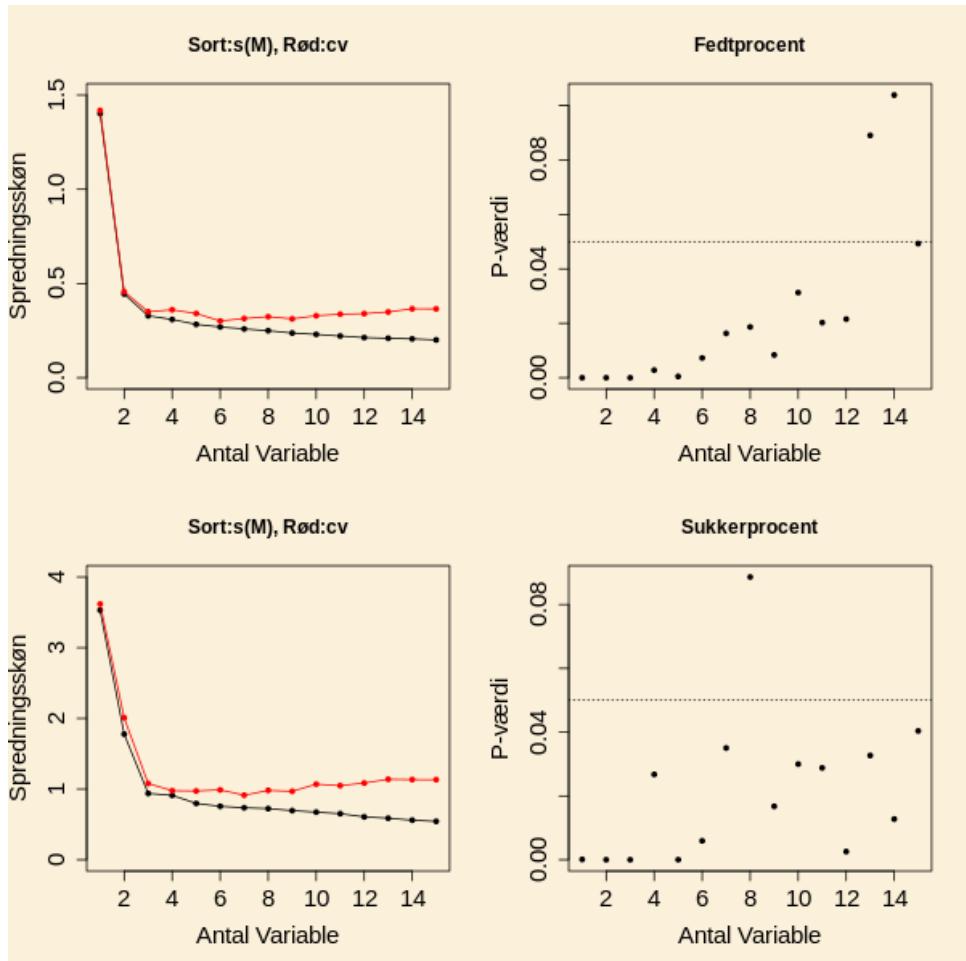
For data omkring fedtprocenten ser vi, at kurven (blå kurve) baseret på de uafhængige testdata også har et lokalt minimum ved 4 variable, hvilket understøtter beregningerne baseret på cross validation. Kurven dykker dog længere ned ved 7 og 8 variable, men dette skal nok betragtes som en tilfældighed.

For data omkring sukkerprocenten er kurven baseret på de uafhængige testdata noget mere variabel, og peger på 8 variable i stedet for de 6 variable ud fra cross validation resultaterne.

### 5.8.1 Estimation på fulde datasæt

Jeg slutter kapitlet af med at se på estimation i data med småkagedej, når alle prøverne inddrages. Det vil sige, at vi har  $40+32=72$  prøver, men da to betragtes som fejlbehæftede, bruger vi i alt 70

prøver. Figuren nedenfor er som tidligere med spredningskøn  $s(M)$  og prædiktionsspredningen fra cross validation, og med  $p$ -værdierne hørende til hvert trin i forward selektionsmetoden.



For data om fedtprocenten peger cross validation nu på 6 variable i stedet for de 4 variable baseret på den første del af datasættet. De fire oprindelige variable var numrene 246, 313, 209 og 57, og de nye 6 variable er 246, 314, 467, 216, 193 og 472. Her er den første variabel den samme, og nummer to stemmer formodentligt også overens, men derfra er det forskellige variable. Prædiktionsspredningen med 6 variable er skønnet til 0.30 og ved fire variable baseret på første del af datasættet til 0.43. Dette er forhåbentlig et udtryk for, at vi kan lave en bedre model baseret på 70 observationer i stedet for kun 39 observationer.

For data om sukkerprocenten peger cross validation nu på 7 variable i stedet for de 6 variable baseret på den første del af datasættet. De 6 oprindelige variable var numrene 25, 488, 539, 579, 535 og 96, og de nye 7 variable er 439, 487, 574, 696, 639, 656 og 515. Her stemmer 488 og 487 formodentlig overens og måske også 579 og 574, men derudover ser variablene forskellige ud. Prædiktionsspredningen med 7 variable er skønnet til 0.91 og ved 6 variable baseret på første del af datasættet til 1.07. Ingen ser vi en lille forbedring ved at bruge alle 70 observationer i estimationen af modellen.

## 5.9 Svar

**Svar 5.1. Cherry trees**

Det første af de to nye træer ligger midt i området for data, hvor middelværdien er velbestemt, hvorimod det andet træ ligger i udkanten af dataområdet.

For at lave et prædiktionsinterval skal man erstatte "confidence" med "prediction" i koden.

## 5.10 Opgaver til kapitel 5

I øvelserne hørende til kapitel 5 skal I arbejde med yderligere to modeller fra klassen af generelle lineære modeller. Den første model er en regression, hvor data er delt op i undergrupper. Den anden model er den multiple regressionsmodel, hvor man ønsker at beskrive respons ved hjælp af flere forklarende variable.

**Showhide: Opgave 5.1: Flere regressionslinjer**

Opgaven kan ses som en forlængelse af opgave 5.1, hvor vi så på længde og bredde af sprækker i jordoverfladen i de canadiske Rocky Mountains. Vi vil udvide undersøgelsen og inddrage sprækker fra to andre steder på jorden, nemlig fra Kyushu i Japan og fra Kilve i England. Som i opgave 5.1 er data aflæst fra figur i [A modern regression approach to determining fault displacement-length scaling relationships](#).

Data findes i filen *Sprækker.csv*, der har tre søjler. Den første søjle angiver området (med værdierne *RM*, *Kyushu* og *Kilve*), den anden søjle indeholder længden og den tredje søjle bredden af sprækkerne, begge målt i meter.

- (a) Indlæs data, og dan variablen *Omr* med område og variablene *logL* og *logB* med henholdsvis logaritmen til længden og logaritmen til bredden. Hvis *Omr* ikke er en faktor efter indlæsningen, skal du omdanne den til en faktor.

Lav en figur, hvor logaritmen til bredden afsættes mod logaritmen til længde, og hvor hver af de tre områder har sin egen farve (dette opnås med tilføjelsen `col=Omr` til plot-kommandoen).

Estimer for hver af de tre områder parametrene i modellen, hvor middelværdien af logaritmen til bredden afhænger lineært af logaritmen til længden. Du kan lade dig inspirere af koden i det første kodevindue i afsnit 5.2. Indtegn efterfølgende de tre estimerede linjer i jeres figur (jeg minder om, at en regressionslinje kan indtegnes ved at benytte *abline*, hvor input er output fra et kald til *lm*).

- (b) Opstil den statistiske model, hvor hvert område har sin egen lineære sammenhæng mellem middelværdien af *logB* og *logL*, og hvert område har sin egen varians omkring den lineære sammenhæng. Opstil hypotesen, at der er samme varians for de tre områder. Benyt Bartletts test for at vurdere denne hypotese. (I kan igen lade jer inspirere af koden i det første kodevindue i afsnit 5.2).

- (c) Opstil nu den reducerede model, hvor der er samme varians i de tre regressionsmodeller (selvom  $p$ -værdien i Bartletts test var lidt under 5%, vælger vi at sige samme varians). Undersøg, om det kan antages, at de tre hældninger er ens.  
Undersøg dernæst, om det kan antages, at de tre skæringer er ens.
- (d) For modellen, hvor der er den samme hældning for de tre områder, skal du angive skøn og konfidensinterval for de parametre, der indgår i modellen.  
Kan det antages, at hældningen er 1, svarende til at bredden er proportional med længden ?



### Showhide: Opgave 5.2: Multipel regression

Hvordan mäter man vægten af en bjørn? Umiddelbart kan man mene, at svaret er simpelt: man tager en stor vægt med ud i felten og beder bjørnen om at træde op på denne! I praksis er dette ikke så nemt, og inden for vildtpflege vil man gerne have mulighed for at vurdere vægten ud fra mål, der er nemmere at opnå. I denne opgave skal I bruge en multipel regressionsmodel til at beskrive vægten ud fra morfometriske mål på bjørnen. Disse sidste mål kan nemt foretages, efter at bjørnen er blevet bedøvet. Hvorfor vil man kende bjørnens vægt? I artiklen [Estimating the Live Body Weight of American Black Bears in Florida](#) siger forfatteren "Collecting body weight measurement is therefore recommended during handling because demographic and reproductive variables are functionally dependent on weight rather than age." I opgaven her skal I ikke bruge data fra denne artikel, men derimod et datasæt der kan findes i R-pakken *Bolstad*. Oprindelsen til datasættet beskrives i R-pakken som "This data set was supplied by [Gary Alt](#)". Data fra R-pakken ligger på kursushjemmesiden i filen *Bear.csv*. Datasættet har søjlerne *headlen* (hovedlængde i tommer), *headwid* (hovedbredde i tommer), *neck* (nakkeomkreds i tommer), *length* (bjørnens længde i tommer), *chest* (brystomkreds i tommer) og *weight* (vægt i pund). En rumfangsbetratning gør, at vi kan forestille os en relation, der siger, at vægt er proportional med længde gange brystomkreds. Denne type tankegang gør, at det er en fordel at bruge log-transformerede data i den multiple regressionsmodel.

- (a) Indlæs data fra filen *Bear.csv*, og dan variable med logaritmen til værdierne i de seks søjler.  
Opskriv den fulde regressionsmodel, model  $M_1$ , hvor middelværdien af logaritmen til vægten afhænger lineært af de fem logaritmer til de morfometriske mål.  
Lav et qqplot af residualerne i denne model.
- (b) Reducer den fulde multiple regressionsmodel ved successivt at fjerne led i modellen (backward selektion). Lav en tabel som for hver model i den successive procedure indeholder model, spredningskøn  $s(M)$ , den største  $p$ -værdi for test af hypotese om at en regessionskoefficient er nul og angivelse af den tilhørende hypotese.  
Lav desuden et  $F$ -test for reduktion fra den fulde model til slutmodellen ved backward selektionsproceduren.
- (c) Lav figurer med residualerne for slutmodellen afsat mod hver af de forklarende variable og med nullinjen indsæt (linjen med skæring nul og hældning nul). Lav desuden et qqplot af residualerne. Lav endelig en figur hvor logaritmen til vægten afsættes mod de forventede værdier og indsæt identitetslinjen (linjen med skæring i nul og hældning 1) i denne figur.

Synes du, at slutmodellen giver en god beskrivelse af data? Inddrag eventuelt sprednings-skønnet  $s(M)$  for slutmodellen i din diskussion. Husk at din model er for logaritmen til vægten, således at en spredning på for eksempel 0.05 svarer til en 5 procents spredning på vægten.

- (d) Lav et 95%-konfidensinterval for middelværdien af logaritmen til vægten og et 95%-prædiktionsinterval for logaritmen til vægten for en ny bjørn med  $length=45$ ,  $chest=25$  og  $neck=13$ . Benyt *predict* i R som beskrevet i afsnit 5.3

Oversæt det sidste interval til et interval for vægten.

- (e) I en model på formen

$$\text{weight} = \text{constant} \cdot \text{length}^\alpha \cdot \text{chest}^\gamma \cdot \text{neck}^\tau$$

vil vi ud fra en "dimensionsanalyse" forvente at  $\alpha + \gamma + \tau = 3$ . Synes du at dette passer med skønnene over regressionscoefficienterne i din multiple regressionsmodel?



### Showhide: Opgave 5.3: BMI for fisk

Data i filen *Tampere.csv* giver tre længdemål samt højde, bredde og vægt for 126 fisk fordelt på fem arter. De tre længdemål adskiller sig ved, hvor langt ud langs halen der måles. Data er oprindeligt publiceret i artiklen *Bidrag till kaennedom on fiskbestonet i vaera sjoeare. Laengelmaeveni*, men er her hentet på adressen [JSE Data Archive](#).

- (a) Indlæs data og dan variablene *Art* (første søjle), *logL1*, *logL2*, *logL3*, *logH*, *logB*, *logV* med logaritmen til værdierne i søjlerne 2 til 8, og dan til sidst variablen  $\logBMI = \logV - 3 * \logL3 + \log(1000)$ . Den sidste variabel er logaritmen til et body mass index for fiskene på formen  $BMI = V / (L3/10)^3$ .

Lav to deldatasæt for henholdsvis Aborre og Brasen med kommandoerne

```
logAborre=logBMI [Art=="Aborre"] og logBrasen=logBMI [Art=="Brasen"]
```

- (b) For de 56 aborre i datasættet er der 9, der har et BMI over 12. Opstil en statistisk model til beskrivelse af observationen 9, og lav et 95%-konfidensinterval for sandsynligheden, for at en aborre har et BMI over 12.
- (c) For de 34 brasen i datasættet er der 4, der har et BMI over 12. Undersøg, om der er samme frekvens af fisk med BMI over 12 blandt de to arter aborre og brasen.
- (d) Opstil en statistisk model til beskrivelse af data i *logAborre* og *logBrasen*. Lav et test for hypotesen, at der er samme middelværdi af logaritmen til BMI for de to fiskearter.
- (e) Betragt nu logaritmen til BMI for alle fem fiskearter. Opstil en statistisk model for data, og undersøg først, om der er samme varians for de fem arter, og dernæst, om der er samme middelværdi for de fem arter.

- (f) Opstil en multipel regressionsmodel til beskrivelse af logaritmen til vægten ( $\log V$ ) for aborre ud fra aborre-værdierne for de fem forklarende variable  $\log L1$ ,  $\log L2$ ,  $\log L3$ ,  $\log H$  og  $\log B$ . Reducer modellen ved brug af backward selektion og lav grafisk kontrol af slutmodellen. Lav et test for reduktion fra startmodel til slutmodel og angiv 95%-konfidensintervaller for parametrene i slutmodellen.



#### Showhide: Opgave 5.4: Multipel regression baseret på NIR-spektrum

I artiklen [Near-infrared spectroscopy as a novel non-invasive tool to assess Spiny Lobster nutritional condition](#) undersøges muligheden af at bruge near-infrared (NIR) spektrometri til at vurdere den ernæringsmæssige tilstand i [languster](#). Forfatterne skriver selv om motivationen: "A practical, rapid and non-invasive technique to analyse lobster nutritional condition has considerable potential to assist with the management of wild stocks,...". I denne opgave skal I se på muligheden for at beskrive "abdominal muscle dry matter content" (AMDM) ud fra NIR-spektret. Der er data for 89 languster i filen *Lobster.txt*. Hver række svarer til en languster, de første 495 søger er NIR-spektret for bølgelængder i området 1063-1334 nm, og søjle 496 indeholder værdierne af AMDM. Spektret er oprindeligt målt for bølgelængder i området 1063-2354 nm, men forfatterne vælger kun at bruge området 1063-1334 til beskrivelse af AMDM.

- (a) Indlæs de 89 spektre og de 89 AMDM-værdier med kommandoerne

```
Dat=matrix(scan("Lobster.txt"), 89, 496, byrow=TRUE)
Spek=Dat[, -496]; AMDM=Dat[, 396]
```

Benyt **R**-funktionen *FWstep* fra afsnit 7.7 til at opbygge en multipel regressionsmodel ved forward selektion med op til 15 forklarende variable. Lav en figur med to delfigurer. Den venstre delfigur skal vise skøn over spredning  $s(M_j)$ ,  $j = 1, \dots, 15$ , hvor  $M_j$  er modellen med  $j$  forklarende variable, og den højre delfigur skal vise  $p$ -værdien for test af hypotesen  $\beta_j = 0$ , hvor  $\beta_j$  er regresionskoefficienten hørende til det sidste led i modellen  $M_j$ . Lav endvidere en tabel med numrene på de forklarende variable i den rækkefølge som de inkluderes.

Vurder ud fra disse figurer, hvor mange led du vil medtage i din multiple regressionsmodel.

- (b) Du skal nu vurdere kvaliteten af de forskellige multiple regressionsmodeller fra en forward selektionsprocedure ved brug af crossvalidation. Til dette skal du bruge **R**-funktionen *FWcrossval* beskrevet i afsnit 7.7. Lav en figur, hvor både spredningsskøn  $s(M_j)$  og cross-validation prædictionsspredningen  $s_{cv}$  afsættes mod  $j$ , med hver sin farve.

Hvilken model vil du vælge til beskrivelse af AMDM ud fra NIR-spektret?

- (c) For din valgte multiple regressionsmodel skal du lave en figur, hvor den målte værdi af AMDM afsættes mod den forventede værdi, og identitetslinjen indtegnes. Indtegn desuden to linjer med hældning 1 i afstanden  $\pm 2s_{cv}$ , hvor  $s_{cv}$  hører til din valgte model. Ser denne figur ud, som du forventer?

Lav desuden et qqplot af residualerne i modellen.

Diskuter størrelsen af crossvalidation prædictionsspredningen  $s_{cv}$  for den valgte model i forhold til variationsområdet for AMDM (forfatterne nævner en værdi på 1.41 for prædictionsspredningen for en noget mere kompliceret model end den I betragter i opgaven her).





# Matematikken bag lineære modeller

I har nu flere gange set, hvordan der dannes en  $t$ -teststørrelse ved at bruge at et skøn over en middelværdiparameter er normalfordelt, variansskønnet følger en skaleret  $\chi^2$ -fordeling, og de to skøn er stokastisk uafhængige. I dette kapitel vil jeg gå lidt ind på matematikken bag denne type resultat, såvel som resultatet om det generelle  $F$ -test i afsnit 4.7.

I analysen af den generelle lineære model tænker vi ofte på data som organiseret i en dataframe, hvor responsvektoren er en af søjlerne, og de andre søjler er faktorer og regressionsvariable. Det er denne vektortilgang, vi vil udnytte i dette kapitel. Så i stedet for at have fokus på den enkelte variabel  $X_i$  vil vi have fokus på hele vektoren  $(X_1, \dots, X_n)^T \in \mathbf{R}^n$ .

Jeg starter med at indføre lidt vektor- og matrixnotation. Vektorer kan enten være rækkevektorer eller søjlevektorer. Hvis vi transponerer en rækkevektor (notation:  $\mathbf{v}^T$ , hvor  $\mathbf{v}$  er rækkevektoren) får vi en søjlevektor, og vice versa.

## Showhide: Middelværdi og varians af vektor

Middelværdien af en stokastisk (søjle-) vektor  $\mathbf{Z} = (Z_1, \dots, Z_n)^T$  defineres som vektoren af middelværdier af de enkelte indgange

$$E(\mathbf{Z}) = (E(Z_1), \dots, E(Z_n))^T.$$

Variansen  $\text{Var}(\mathbf{Z})$  defineres som en  $n \times n$  matrix, hvor den  $i$ 'te diagonalindgang er variansen  $\text{Var}(Z_i)$ , og den  $(i, j)$ 'te indgang er kovariansen  $\text{Cov}(Z_i, Z_j)$ :

$$\text{Var}(\mathbf{Z}) = \begin{pmatrix} \text{Var}(Z_1) & \text{Cov}(Z_1, Z_2) & \cdots & \text{Cov}(Z_1, Z_n) \\ \text{Cov}(Z_2, Z_1) & \text{Var}(Z_2) & \cdots & \text{Cov}(Z_2, Z_n) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(Z_n, Z_1) & \text{Cov}(Z_n, Z_2) & \cdots & \text{Var}(Z_n) \end{pmatrix}.$$

I kender formodentligt følgende regneregler

$$E\left(\sum_i a_i Z_i\right) = \sum_i a_i E(Z_i), \quad \text{Var}\left(\sum_i a_i Z_i\right) = \sum_{i,j} a_i a_j \text{Cov}(Z_i, Z_j),$$

$$\text{Cov}\left(\sum_i a_i Z_i, \sum_i b_i Z_i\right) = \sum_{i,j} a_i b_j \text{Cov}(Z_i, Z_j).$$

Med den indførte matrixnotation kan vi samle disse regneregler på følgende vis. Lad  $\mathbf{B}$  være en  $k \times n$  ikke-stokastisk matrix. Så er

$$E(\mathbf{BZ}) = \mathbf{BE}(\mathbf{Z}) \quad \text{og} \quad \text{Var}(\mathbf{BZ}) = \mathbf{BVar}(\mathbf{Z})\mathbf{B}^T.$$



### Showhide: Matriks-regneregler

For fuldstændighedens skyld samler jeg her nogle vigtige regneregler for matricer.

#### Resultat 6.1. (Regneregler for matricer)

1. Hvis  $\mathbf{B}$  er en  $n \times k$  matrix, med  $(i, j)$ 'te indgang  $B_{ij}$ , så er  $\mathbf{B}^T$  en  $k \times n$  matrix med  $(i, j)$ 'te indgang  $B_{ji}$ .
2. Hvis  $\mathbf{A}$  er  $n \times k$  og  $\mathbf{B}$  er  $k \times m$ , så er  $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ .
3. Hvis  $\mathbf{X}$  er  $n \times k$ , så er  $k \times k$  matricen  $\mathbf{X}^T\mathbf{X}$  symmetrisk ( $(\mathbf{X}^T\mathbf{X})^T = \mathbf{X}^T\mathbf{X}$ ), og dette gælder også for den inverse matrix  $(\mathbf{X}^T\mathbf{X})^{-1}$ .
4. Hvis  $\mathbf{A}$  er en  $k \times k$  symmetrisk matrix, er den inverse matrix  $\mathbf{A}^{-1}$  bestemt ved  $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$  og  $\mathbf{AA}^{-1} = \mathbf{I}$ .



For analysen af normalfordelingsmodellerne skal vi også vide noget om projektioner. Når vi skal finde skøn over  $(\beta_1, \dots, \beta_k)$  ved at minimere

$$\sum_{i=1}^n (x_i - \beta_1 h_{i1} - \dots - \beta_k h_{ik})^2,$$

er dette ækvivalent med at minimere den kvadrerede  $L^2$ -norm

$$|\mathbf{x} - \mathbf{H}\boldsymbol{\beta}|^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\beta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\beta}), \quad \mathbf{x} = (x_1, \dots, x_n)^T, \quad \boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^T,$$

hvor  $\mathbf{H}$  er  $n \times k$  matricen med  $(i, j)$ 'te indgang  $h_{ij}$ . Her står, at vi skal finde det punkt, udspændt af søjlerne i  $\mathbf{H}$ , som er tættest på  $\mathbf{x}$ , men dette er netop projektionen af  $\mathbf{x}$  på rummet udspændt af søjlerne i  $\mathbf{H}$ .

#### Resultat 6.2. (Projektion)

Projektionen af vektoren  $\mathbf{x}$  ned på underrummet udspændt af søjlerne i  $\mathbf{H}$  er givet ved  $\mathbf{Px}$ , hvor  $n \times n$  matricen  $\mathbf{P}$  er givet som  $\mathbf{P} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ . I ovenstående minimeringsproblem giver dette skønnet  $\hat{\boldsymbol{\beta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$ .

For at eftervise dette resultat skal man vise, at  $\mathbf{x} - \mathbf{Px}$  står vinkelret på søjlerne i  $\mathbf{H}$ , men dette følger af

$$\mathbf{H}^T(\mathbf{x} - \mathbf{Px}) = \mathbf{H}^T(\mathbf{I} - \mathbf{P})\mathbf{x} = \mathbf{H}^T(\mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T)\mathbf{x} = (\mathbf{H}^T - \mathbf{H}^T)\mathbf{x} = 0,$$

hvor  $\mathbf{I}$  er en diagonalmatrix med 1 langs diagonalen.

#### Showhide: Vektor af normalfordelte variable

Hvis  $Z_i \sim N(\mu_i, \sigma^2)$ ,  $i = 1, \dots, n$ , er uafhængige, skriver vi dette kort som

$$\mathbf{Z} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}),$$

hvor  $\mathbf{I}$  er en diagonalmatriks med 1 langs diagonalen, og  $\boldsymbol{\mu}$  er søjlevektoren med middelværdierne,  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ . I notationen  $N_n(\cdot, \cdot)$  er det første argument middelværdien  $E(\mathbf{Z})$ , og det andet argument er variansen  $\text{Var}(\mathbf{Z})$ .

Hvis vi laver linearkombinationer af koordinaterne i  $\mathbf{Z}$ , bliver disse igen normalfordelte (regneregler for normalfordelingen!), men ikke nødvendigvis uafhængige. Vi vil stadig bruge notationen med  $N_n(\cdot, \cdot)$  og skriver

$$\mathbf{BZ} \sim N_n(\mathbf{B}\boldsymbol{\mu}, \sigma^2 \mathbf{B}\mathbf{B}^T),$$

idet  $\mathbf{B}\mathbf{B}^T = \mathbf{B}\mathbf{B}^T$ . Hvis vi har en søjlevektor  $\mathbf{U}$ , der er fremkommet ved linearkombinationer af uafhængige normalfordelte variable, og  $\mathbf{U} \sim N(\boldsymbol{\mu}, \Sigma)$ , så vil

$$\mathbf{BU} \sim N_n(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\Sigma\mathbf{B}^T).$$

Dette skyldes, at koordinaterne i  $\mathbf{BU}$  vil også være linearkombinationer af uafhængige normalfordelte variable.

Ved hjælp af den ovenfor etablerede matriksnotation kan vi nemt lave beregninger baseret på normalfordelte variable.



## 6.1 Den generelle lineære model via underrum

I har nu i de tidligere kapitler set forskellige eksempler på normalfordelingsmodeller specificeret gennem faktorer og regressionsvariable. Det har været en stor bekvemmelighed, at modellen kan angives gennem en *modelformel*. Dette giver en meget simpel og kompakt sprogbrug, der også gør det nemt at kommunikere med **R**, specielt med funktionen *lm*.

Det der karakteriserer modellerne, ud fra en matematisk synsvinkel, er, at *vektoren* af middelværdier kan variere frit i et lineært underrum af  $\mathbf{R}^n$ , hvor  $n$  er antallet af observationer. Lad  $X_i \sim N(\xi_i, \sigma^2)$ ,  $i = 1, \dots, n$ , være uafhængige stokastiske variable, og lad  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$ . En lineær model siger, at  $\boldsymbol{\xi}$  kan variere frit i et givet lineært underrum  $L$ . I det næste skjulte punkt vises det relevante underrum for nogle af de modeller, vi har betragtet i de tidligere kapitler.

#### Showhide: Kendte eksempler

#### Showhide: To grupper

Vi starter med modellen, hvor vi har to grupper af normalfordelte variable med hver sin middelværdi  $\mu_1$  og  $\mu_2$ . Betragt en ordning, hvor de  $n_1$  observationer fra gruppe 1 kommer først, og de  $n_2$  observationer fra gruppe 2 kommer sidst. Betragt de to vektorer

$$\mathbf{v} = (1, \dots, 1, 0, \dots, 0)^T \quad \text{og} \quad \mathbf{w} = (0, \dots, 0, 1, \dots, 1)^T,$$

og lad  $\mathbf{H}$  være matricen med søjlerne  $\mathbf{v}$  og  $\mathbf{w}$ , og lad  $L$  være underrummet udspændet af de to vektorer. Så kan modellen skrives på formen  $\xi = \mu_1 \mathbf{v} + \mu_2 \mathbf{w} = \mathbf{H}\boldsymbol{\mu}$ , hvor  $\boldsymbol{\mu}$  er søjlevektoren med indgangene  $\mu_1$  og  $\mu_2$ , og kan også formuleres på den måde, at  $\xi$  kan variere frit i  $L$ .

Det er nemt at se, at  $\mathbf{H}^T \mathbf{H}$  er diagonalmatricen med indgangene  $n_1$  og  $n_2$ ,  $\mathbf{H}^T \mathbf{x}$  er søjlevektoren med summen over gruppe 1 og summen over gruppe 2 som indgange, og derfor

$$\hat{\boldsymbol{\mu}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix},$$

hvor  $\bar{x}_1$  og  $\bar{x}_2$  er gennemsnittene i de to grupper.



### Showhide: Simpel lineær regression

Vi betragter den lineære regressionsmodel med  $X_i \sim N(\alpha + \beta t_i, \sigma^2)$ ,  $i = 1, \dots, n$ , med  $t_1, \dots, t_n$  kendte tal. Betragt de to vektorer

$$\mathbf{e} = (1, \dots, 1)^T \quad \text{og} \quad \mathbf{t} = (t_1, \dots, t_n)^T,$$

og lad  $\mathbf{H}$  være matricen med søjlerne  $\mathbf{e}$  og  $\mathbf{t}$ , og lad  $L$  være underrummet udspændet af de to vektorer. Så kan modellen skrives på formen  $\xi = \alpha \mathbf{e} + \beta \mathbf{t} = \mathbf{H} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ , og kan også formuleres på den måde, at  $\xi$  kan variere frit i  $L$ .

Hvis vi lader  $\mathbf{v} = \mathbf{t} - \bar{t} \mathbf{e}$ , så er  $L$  også udspændet af  $\mathbf{e}$  og  $\mathbf{v}$ . Lad  $\mathbf{K}$  være matricen med søjlerne  $\mathbf{e}$  og  $\mathbf{v}$ . Så viser en simpel beregning at

$$(\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \mathbf{x} = \begin{pmatrix} \bar{x} \\ \hat{\beta} \end{pmatrix},$$

med  $\hat{\beta}$  givet i (3.1). Da

$$\bar{x} \mathbf{e} + \hat{\beta} \mathbf{v} = (\bar{x} - \hat{\beta} \bar{t}) \mathbf{e} + \hat{\beta} \mathbf{t},$$

ses at  $\hat{\alpha} = \bar{x} - \hat{\beta} \bar{t}$ .



### Showhide: One way anova

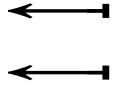
Vi betragter normalfordelingsmodellen med  $X_i \sim N(\mu_{G_i}, \sigma^2)$ ,  $i = 1, \dots, n$ , hvor  $G$  er en faktor, der deler op i  $k$  grupper. For hver gruppe  $j = 1, \dots, k$  defineres en vektor  $\mathbf{v}_j$  ved at den  $i$ 'te indgang er 1, hvis  $G_i = j$ , og nul ellers. Så kan vi skrive modellen på formen

$$\xi = \mu_1 \mathbf{v}_1 + \dots + \mu_k \mathbf{v}_k = \mathbf{H} \boldsymbol{\mu},$$

hvor  $\mathbf{H}$  er matricen med søjlerne  $\mathbf{v}_j$ ,  $j = 1, \dots, k$ , og  $\boldsymbol{\mu}$  er søjlevektoren indeholdende  $\mu_1, \dots, \mu_k$ .

Bemærk at  $\mathbf{H}^T \mathbf{H}$  bliver en diagonalmatriks, hvoraf det let ses, at  $\hat{\boldsymbol{\mu}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{x}$  giver, at  $\hat{\mu}_j$  er gennemsnittet i den  $j$ 'te gruppe.

Matricen  $\mathbf{H}$  giver parametriseringen med  $\mu_1, \dots, \mu_k$ . Det samme middelværdirum kan imidlertid også udspændes af søjlevektorerne  $\mathbf{e}, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_k$ , hvor  $\mathbf{e}$  er vektoren med lutter 1-taller, som svarer til den parametrisering, der bruges i  $\mathbf{R}$  med  $\mu_1, \mu_2 - \mu_1, \dots, \mu_k - \mu_1$ .



I en generel lineær model kan middelværdien skrives som en sum af bidrag fra enten en faktor eller en regressionsvariabel. Vi har set i eksemplerne ovenfor, hvordan henholdsvis en faktor og en regressionsvariabel definerer et linaært underrum. Når vi i en model betragter sum af bidrag, svarer dette til sum af de lineære underrum, og dette er i sig selv et nyt linaært underrum. Vi ender derfor med følgende generelle setup,

$$\begin{aligned} \mathbf{X} &\sim N_n(\boldsymbol{\xi}, \sigma^2 \mathbf{I}) \\ \text{Model } M_1: \quad \boldsymbol{\xi} &\in L_1, \\ \text{Model } M_2: \quad \boldsymbol{\xi} &\in L_2, \quad L_2 \subset L_1, \end{aligned} \tag{6.1}$$

hvor  $L_1$  er et  $d_1$ -dimensionalt linaært underrum af  $\mathbf{R}^n$ , og  $L_2$  er et  $d_2$ -dimensionalt linaært underrum af  $L_1$ . Model  $M_2$  fremkommer typisk ved, at man sætter nogle af parametrene i model  $M_1$  lig med nul, eller man sætter nogle parametre lig med hinanden. Vi skal i næste afsnit bruge følgende matematiske resultat.

### Resultat 6.3. (Ortogonalitet af projektioner)

Lad  $\mathbf{P}_1$  og  $\mathbf{P}_2$  være projekionsmatricer hørende til de to underrum  $L_1$  og  $L_2 \subset L_1$ . Så gælder der

$$\begin{aligned} \mathbf{P}_1 \mathbf{P}_2 &= \mathbf{P}_2, \quad (\mathbf{I} - \mathbf{P}_1)(\mathbf{P}_1 - \mathbf{P}_2)^T = 0, \\ (\mathbf{I} - \mathbf{P}_1) \mathbf{P}_2^T &= 0, \quad (\mathbf{P}_1 - \mathbf{P}_2) \mathbf{P}_2^T = 0. \end{aligned}$$

For en projekionsmatriks  $\mathbf{P}$  har vi  $\mathbf{P}^T = \mathbf{P}$  og  $\mathbf{P}^2 = \mathbf{P}$ . Hvis den første ligning ovenfor er vist, kan ligning nummer to reduceres som følger

$$\begin{aligned} (\mathbf{I} - \mathbf{P}_1)(\mathbf{P}_1 - \mathbf{P}_2)^T &= (\mathbf{I} - \mathbf{P}_1)(\mathbf{P}_1 - \mathbf{P}_2) \\ &= (\mathbf{P}_1 - \mathbf{P}_1^2) - (\mathbf{P}_2 - \mathbf{P}_1 \mathbf{P}_2) = (\mathbf{P}_1 - \mathbf{P}_1) - (\mathbf{P}_2 - \mathbf{P}_2) = 0. \end{aligned}$$

Ligning nummer 3 og 4 følger på samme vis. Vi skal derfor blot argumentere for korrektheden af den første ligning. For en vektor  $\mathbf{v} \in L_2$  gælder der, at  $\mathbf{P}_1 \mathbf{v} = \mathbf{v}$ , eftersom  $L_2 \subset L_1$ . Da søjlerne i projektionmatricen  $\mathbf{P}_2$  ligger i  $L_2$ , følger det nu, at  $\mathbf{P}_1 \mathbf{P}_2 = \mathbf{P}_2$ .

## 6.2 Spaltningssætningen

Vi vil bruge følgende sætning uden bevis (sætningen bevises i kurset *Multivariat statistisk analyse*).

**Resultat 6.4.** (Spaltningssætningen)

Lad  $Y_1, \dots, Y_n$  være uafhængige og identisk fordelte med  $Y_i \sim N(0, \sigma^2)$ , og lad  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ .

- (i) Lad  $\mathbf{B}_1$  være en  $k_1 \times n$  matrix, og lad  $\mathbf{B}_2$  være en  $k_2 \times n$  matrix. Definer

$$\mathbf{Z}_1 = (Z_{11}, \dots, Z_{1,k_1})^T = \mathbf{B}_1 \mathbf{Y} \quad \text{og} \quad \mathbf{Z}_2 = (Z_{21}, \dots, Z_{2,k_2})^T = \mathbf{B}_2 \mathbf{Y}.$$

Hvis  $\mathbf{B}_1 \mathbf{B}_2^T = \mathbf{0}$  så er  $\mathbf{Z}_1$  og  $\mathbf{Z}_2$  stokastisk uafhængige.

- (ii) Lad  $\mathbf{B}$  være en  $n \times n$  matrix, der opfylder  $\mathbf{B}\mathbf{B}^T = \mathbf{B}$  og  $\mathbf{B}^T = \mathbf{B}$  (matematisk betyder dette, at  $\mathbf{B}$  er en orthogonal projektion). Definer  $\mathbf{Z} = (Z_1, \dots, Z_n)^T = \mathbf{B}\mathbf{Y}$ . Så gælder der, at

$$\sum_{i=1}^n Z_i^2 = |\mathbf{B}\mathbf{Y}|^2 = (\mathbf{B}\mathbf{Y})^T(\mathbf{B}\mathbf{Y}) \sim \sigma^2 \chi^2(k),$$

hvor  $k$  er rangen af matricen  $\mathbf{B}$  (dimensionen af det rum  $\mathbf{B}$  projektører ned på).

Kombinere vi ovenstående spaltningssætning med ortogonalitetsresultaterne 6.3 kan vi vise de fordelingsresultater, vi allerede har brugt nogle gange.

**Resultat 6.5.** (Uafhængighed mellem middelværdiskøn og variansskøn)

Betrægt en generel lineær model med  $\xi \in L$ , hvor det  $d$ -dimensionale underrum  $L$  er udspændt af søjlerne i matricen  $\mathbf{H}$ , og vi benytter parametriseringen  $\xi = \mathbf{H}\boldsymbol{\theta}$ . I denne model er skønnet  $\hat{\boldsymbol{\theta}}$  over middelværdiparametrene og skønnet over variansen  $s^2(M) = \sum_{i=1}^n (X_i - \hat{\xi}_i(M))^2 / (n - d)$  stokastisk uafhængige. Desuden er  $s^2(M) \sim \sigma^2 \chi^2(n - d) / (n - d)$ .

**Showhide: Bevis**

Lad  $\mathbf{P} = \mathbf{H}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$  være projektionsmatricen, og lad  $\mathbf{Y} = \mathbf{X} - \xi$  med  $\xi \in L$ . Så har vi

$$\hat{\boldsymbol{\theta}} = \mathbf{B}_1 \mathbf{X} = \mathbf{B}_1 \mathbf{Y} + \xi, \quad \mathbf{B}_1 = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T,$$

og

$$(n - d) s^2(M) = (\mathbf{B}_2 \mathbf{X})^T (\mathbf{B}_2 \mathbf{X}) = (\mathbf{B}_2 \mathbf{Y})^T (\mathbf{B}_2 \mathbf{Y}), \quad \mathbf{B}_2 = \mathbf{I} - \mathbf{P}.$$

Da  $\mathbf{B}_1 = \mathbf{B}_1 \mathbf{P}$  får vi

$$\mathbf{B}_1 \mathbf{B}_2^T = \mathbf{B}_1 \mathbf{P} (\mathbf{I} - \mathbf{P}) = \mathbf{B}_1 (\mathbf{P} - \mathbf{P}) = 0,$$

og uafhængighedsresultatet følger af spaltningssætningen. Fordelingen af variansskønnet følger også af spaltningssætningen, idet  $\mathbf{B}_2 = \mathbf{I} - \mathbf{P}$  er matricen for projektionen på et rum af dimension  $n - d$ .



Det generelle  $F$ -test for reduktion fra en generel lineær model  $M_1$  til en model  $M_2$  er baseret på teststørrelsen  $s^2(M_1, M_2) / s^2(M_1)$ , hvor  $s^2(M_1, M_2) = |\mathbf{P}_1 \mathbf{X} - \mathbf{P}_2 \mathbf{X}|^2 / (d_1 - d_2)$ .

**Resultat 6.6.** (Uafhængighed mellem tæller og nævner i generelt  $F$ -test)

Under model  $M_2$  i den generelle model (6.1) er  $s^2(M_1)$  og  $s^2(M_1, M_2)$  stokastisk uafhængige,  $s^2(M_1) \sim \sigma^2 \chi^2(n - d_1)/(n - d_1)$  og  $s^2(M_1, M_2) \sim \sigma^2 \chi^2(d_1 - d_2)/(d_1 - d_2)$ . Desuden har vi, at  $s^2(M_1, M_2) = (SSD(M_2) - SSD(M_1))/(df(M_2) - df(M_1))$

### Showhide: Bevis

Vi lader  $\mathbf{Y} = \mathbf{X} - \boldsymbol{\xi}$  med  $\boldsymbol{\xi} \in L_2$ , og lader  $\mathbf{P}_1$  og  $\mathbf{P}_2$  være matricerne, der projektører ned på henholdsvis  $L_1$  og  $L_2$ . Da  $\boldsymbol{\xi} \in L_2$  gælder der, at  $\mathbf{P}_1\boldsymbol{\xi} = \mathbf{P}_2\boldsymbol{\xi} = \boldsymbol{\xi}$  og  $(\mathbf{I} - \mathbf{P}_j)\mathbf{X} = (\mathbf{I} - \mathbf{P}_j)\mathbf{Y}$ ,  $j = 1, 2$ . Vi har følgende udtryk

$$(n - d_1)s^2(M_1) = |(\mathbf{I} - \mathbf{P}_1)\mathbf{Y}|^2 \quad \text{og} \quad (d_1 - d_2)s^2(M_1, M_2) = |(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{Y}|^2.$$

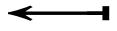
Uafhængigheden følger nu af spaltingssætningen og  $(\mathbf{I} - \mathbf{P}_1)(\mathbf{P}_1 - \mathbf{P}_2)^T = 0$  ifølge Resultat 6.3.

Fordelingsresultaterne følger også af spaltingssætningen, idet både  $\mathbf{I} - \mathbf{P}_1$  og  $\mathbf{P}_1 - \mathbf{P}_2$  er projektionsmatricer.

Endelig har vi fra Resultat 6.3, at

$$SSD(M_2) = |(\mathbf{I} - \mathbf{P}_2)\mathbf{X}|^2 = |(\mathbf{I} - \mathbf{P}_1 + \mathbf{P}_1 - \mathbf{P}_2)\mathbf{X}|^2 = |(\mathbf{I} - \mathbf{P}_1)\mathbf{X}|^2 + |(\mathbf{P}_1 - \mathbf{P}_2)\mathbf{X}|^2 = SSD(M_1) + SSD(M_1, M_2),$$

hvor  $s^2(M_1, M_2) = SSD(M_1, M_2)/(d_1 - d_2)$ . Da også  $d_1 - d_2 = df(M_2) - df(M_1)$ , er den sidste del af resultatet vist.



## 6.2.1 Likelihood ratio test

I normalfordelingsmodellen  $X_i \sim (\xi_i, \sigma^2)$ ,  $i = 1, \dots, n$ , og de  $n$  variable er uafhængige, er likelihoodfunktionen

$$L(\boldsymbol{\xi}, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(X_i - \xi_i)^2} = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}|\mathbf{X} - \boldsymbol{\xi}|^2\right).$$

Betrægt to lineære modeller  $M_1 : \boldsymbol{\xi} \in L_1$  og  $M_2 : \boldsymbol{\xi} \in L_2$  med  $L_2 \subset L_1$ , og lad de tilhørende projekionsmatricer være  $\mathbf{P}_1$  og  $\mathbf{P}_2$ , og de tilhørende dimensioner af de to rum være  $d_1$  og  $d_2$ . Vi lader

$$\hat{\boldsymbol{\xi}}(M_1) = \mathbf{P}_1\mathbf{X}, \quad \hat{\boldsymbol{\xi}}(M_2) = \mathbf{P}_2\mathbf{X}, \quad SSD_1 = |\mathbf{X} - \mathbf{P}_1\mathbf{X}|^2 \quad \text{og} \quad SSD_2 = |\mathbf{X} - \mathbf{P}_2\mathbf{X}|^2.$$

Hvis vi for eksempel maksimerer over  $(\boldsymbol{\xi}, \sigma^2) \in L_1 \times \mathbf{R}_+$ , får vi middelværdiskønnet  $\hat{\boldsymbol{\xi}}(M_1)$  og  $\hat{\sigma}^2 = SSD_1/n$ .

Likelihood ratio teststørrelsen for reduktion fra model  $M_1$  til model  $M_2$  er

$$\begin{aligned} Q &= \frac{\max_{(\boldsymbol{\xi}, \sigma^2) \in L_2 \times \mathbf{R}_+} L(\boldsymbol{\xi}, \sigma^2)}{\max_{(\boldsymbol{\xi}, \sigma^2) \in L_1 \times \mathbf{R}_+} L(\boldsymbol{\xi}, \sigma^2)} = \frac{\left(2\pi \frac{SSD_2}{n}\right)^{-n/2} e^{-n/2}}{\left(2\pi \frac{SSD_1}{n}\right)^{-n/2} e^{-n/2}} \\ &= \left(\frac{SSD_1}{SSD_2}\right)^{n/2} = \left(\frac{|\mathbf{X} - \mathbf{P}_1\mathbf{X}|^2}{|\mathbf{X} - \mathbf{P}_2\mathbf{X}|^2}\right)^{n/2} = \left(\frac{|\mathbf{X} - \mathbf{P}_1\mathbf{X}|^2}{|\mathbf{X} - \mathbf{P}_1\mathbf{X}|^2 + |\mathbf{P}_1\mathbf{X} - \mathbf{P}_2\mathbf{X}|^2}\right)^{n/2} \end{aligned}$$

$$= \left( \frac{1}{1 + \frac{|\mathbf{P}_1\mathbf{X} - \mathbf{P}_2\mathbf{X}|^2}{|\mathbf{X} - \mathbf{P}_1\mathbf{X}|^2}} \right)^{n/2} = \left( \frac{1}{1 + \frac{d_1 - d_2}{n - d_1} F} \right)^{n/2},$$

hvor vi i det tredje lighedstegn i den anden linje har brugt Resultat 6.3 (ortogonalitet af projektorer), og til sidst har brugt definitionen på  $F$ -teststørrelsen i Resultat 4.4.

Lad mig til sidst i dette afsnit argumentere for, at et  $t$ -test, for at en middelværdiparameter er nul, er ækvivalent med  $F$ -testet for den tilsvarende reduktion af modellen.

### Showhide: Argument

Lad os skrive modellen på formen

$$\boldsymbol{\xi} = \mathbf{H}\boldsymbol{\theta},$$

hvor  $\mathbf{H}$  er en  $n \times d$  matrix og  $\boldsymbol{\theta}$  er en  $d$ -dimensional søjle. Vi ønsker at teste hypotesen  $\theta_1 = 0$ .

Vi ændrer nu på søjlerne i  $\mathbf{H}$  for at gøre beregningerne nemmere. Først ændrer vi den første søjle i  $\mathbf{H}$  ved at trække en linearkombination af søjlerne 2 til  $d$  fra, således at den nye søjle bliver vinkelret på de andre søjler. Dette ændrer ikke på parameteren  $\theta_1$ , og hypotesen  $\theta_1 = 0$  er den samme. Dernæst vælger vi nye søjler 2 op til  $d$ , således at de nye søjler udspænder det samme rum og er vinkelrette på hinanden. Vi betegner den således ændrede  $\mathbf{H}$ -matrix med  $\mathbf{K}$  og tillader os at kalde den nye tilhørende parameter for  $\boldsymbol{\theta}$  (fordi  $\theta_1$  er den samme). Kalder vi søjlerne i  $\mathbf{K}$  for  $\mathbf{v}_1, \dots, \mathbf{v}_d$ , har vi

$$M_1: \hat{\theta}_1(M_1) = \frac{\mathbf{v}_1^T \mathbf{X}}{|\mathbf{v}_1|^2}, \quad \hat{\theta}_j(M_1) = \frac{\mathbf{v}_j^T \mathbf{X}}{|\mathbf{v}_j|^2}, \quad j = 2, \dots, d,$$

$$M_2: \hat{\theta}_1(M_2) = 0, \quad \hat{\theta}_j(M_2) = \frac{\mathbf{v}_j^T \mathbf{X}}{|\mathbf{v}_j|^2}, \quad j = 2, \dots, d.$$

Heraf får vi

$$|\mathbf{P}_1\mathbf{X} - \mathbf{P}_2\mathbf{X}|^2 = |\hat{\theta}_1(M_1)\mathbf{v}_1|^2 = \hat{\theta}_1(M_1)^2 |\mathbf{v}_1|^2,$$

og  $F$ -teststørrelsen er

$$\frac{\hat{\theta}_1(M_1)^2}{s^2(M_1)/|\mathbf{v}_1|^2}.$$

Da  $\hat{\theta}_1(M_1) \sim N(0, \sigma^2/|\mathbf{v}_1|^2)$ , ser vi, at  $F$ -teststørrelsen er den kvadrerede  $t$ -teststørrelse for hypotesen  $\theta_1 = 0$ . 

## 6.3 T-test

Betrægt den generelle lineære model  $X_i \sim N(\xi_i, \sigma^2)$ ,  $i = 1, \dots, n$ , hvor søjlevektoren  $\boldsymbol{\xi}$  af middelværdier ligger i det lineære underrum  $L$  af dimension  $d$ . Betragt en parametrisering  $\boldsymbol{\theta}$  givet ved  $\boldsymbol{\xi} = \mathbf{H}\boldsymbol{\theta}$ , hvor  $\mathbf{H}$  er en  $n \times d$  matrix og  $\boldsymbol{\theta}$  er en søjlevektor.

Vi har, at  $\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{X}$ , hvilket specielt viser, at enhver koordinat  $\hat{\theta}_j$  er en linearkombination af  $X_1, \dots, X_n$ . Dette er baggrunden for resultaterne i de tidligere kapitler, hvor alle skønnene

over parametrene i middelværdien er normalfordelte. Når alle disse skøn også har haft den egen-skab, at middelværdien er lig med den sande værdi af parameteren, følger dette af

$$E\left((\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{X}\right) = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T E(\mathbf{X}) = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{H} \boldsymbol{\theta} = \boldsymbol{\theta}.$$

### Resultat 6.7. (T-test i lineær model)

Betrægt en lineær normal model  $M$ , hvor  $\gamma$  er en af koordinaterne i middelværdiparametriseringen. Ifølge ovenstående kan vi så skrive  $\hat{\gamma} = \sum_i a_i X_i$  for passende konstanter  $a_1, \dots, a_n$ . Lad endvidere variansskønnet være  $s^2(M)$  med  $df(M)$  frihedsgrader.

- (i) Under hypotesen  $\gamma = \gamma_0$  gælder der, at

$$t = \frac{\hat{\gamma} - \gamma_0}{\text{sd}_s(\hat{\gamma})} \sim t(df(M)), \text{ hvor } \text{sd}_s(\hat{\gamma}) = s(M) \sqrt{\sum_i a_i^2}.$$

- (ii) Lad  $t_0 = t_{\text{inv}}(0.975, df(M))$ , så er et 95%-konfidensinterval for  $\gamma$  på formen

$$[\hat{\gamma} - t_0 \cdot \text{sd}_s(\hat{\gamma}), \hat{\gamma} + t_0 \cdot \text{sd}_s(\hat{\gamma})].$$

### Showhide: Bevis

Med  $C = \sum_i a_i^2$  har vi

$$\frac{\hat{\gamma} - \gamma_0}{\sigma \sqrt{C}} \sim N(0, 1), \quad \frac{s^2(M)}{\sigma^2} \sim \chi^2(df(M)/df(M)),$$

og de to stokastiske variable er uafhængige ifølge Resultat 6.5. Fra Definition 2.3 har vi derfor

$$t = \frac{\hat{\gamma} - \gamma_0}{s(M) \sqrt{C}} = \frac{(\hat{\gamma} - \gamma_0)/(\sigma \sqrt{C})}{\sqrt{s^2(M)/\sigma^2}} \sim t(df(M)),$$

hvilket viser (i).

For at vise konfidensintervallet laver vi omskrivningen

$$\begin{aligned} P_{\gamma_0}([\hat{\gamma} - t_0 \cdot \text{sd}_s(\hat{\gamma}), \hat{\gamma} + t_0 \cdot \text{sd}_s(\hat{\gamma})] \ni \gamma) &= P_{\gamma_0}(\hat{\gamma} - t_0 \cdot \text{sd}_s(\hat{\gamma}) \leq \gamma \leq \hat{\gamma} + t_0 \cdot \text{sd}_s(\hat{\gamma})) \\ &= P_{\gamma_0}\left(-t_0 \leq \frac{\hat{\gamma} - \gamma_0}{\text{sd}_s(\hat{\gamma})} \leq t_0\right) = t_{\text{cdf}}(t_0, df(M)) - t_{\text{cdf}}(-t_0, df(M)) \\ &= 0.975 - 0.025 = 0.95. \end{aligned}$$



## 6.4 Opgaver til kapitel 6

### Showhide: Opgave 6.1: Lineær transformation

Vis, at

$$E(\mathbf{BZ}) = \mathbf{BE}(\mathbf{Z}),$$

hvor  $\mathbf{B}$  er en  $k \times n$  ikke-stokastisk matrix, og  $\mathbf{Z}$  er en  $n$ -dimensional stokastisk vektor.



### Showhide: Opgave 6.2: Ortogonale vektorer

Betrægt en generel lineær model  $X_i \sim N(\xi_i, \sigma^2)$ ,  $i = 1, \dots, n$ , hvor  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$  er på formen

$$\boldsymbol{\xi} = \theta_1 \mathbf{v}_1 + \dots + \theta_d \mathbf{v}_d,$$

med  $\mathbf{v}_1, \dots, \mathbf{v}_d$   $n$ -dimensionale søjlevektorer. Antag at vektorerne  $\mathbf{v}_1, \dots, \mathbf{v}_d$  er ortogonale, det vil sige  $\mathbf{v}_i^T \mathbf{v}_j = 0$  for alle  $i \neq j$ .

- (a) Vis, at  $\hat{\theta}_j = \frac{\mathbf{v}_j^T \mathbf{X}}{|\mathbf{v}_j|^2}$ ,  $j = 1, \dots, d$ , hvor  $\mathbf{X} = (X_1, \dots, X_n)^T$ .



### Showhide: Opgave 6.3: T-test og F-test

Betrægt en generel lineær model  $M_1$  med  $X_i \sim N(\xi_i, \sigma^2)$ ,  $i = 1, \dots, n$ , hvor  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$  er på formen

$$\boldsymbol{\xi} = \theta_1 \mathbf{v}_1 + \theta_2 \mathbf{v}_2,$$

med  $\mathbf{v}_1$  og  $\mathbf{v}_2$   $n$ -dimensionale søjlevektorer.

- (a) Lad  $\mathbf{w} = \mathbf{v}_1 - a\mathbf{v}_2$ ,  $a = \frac{\mathbf{v}_1^T \mathbf{v}_2}{|\mathbf{v}_2|^2}$ , og vis, at  $\boldsymbol{\xi} = \theta_1 \mathbf{w} + \gamma \mathbf{v}_2$  med  $\gamma = \theta_2 + a\theta_1$ .
- (b) Vis, at  $\hat{\theta}_1 = \frac{\mathbf{w}^T \mathbf{X}}{|\mathbf{w}|^2}$  og  $\hat{\gamma} = \frac{\mathbf{v}_2^T \mathbf{X}}{|\mathbf{v}_2|^2}$ .
- (c) Vis, at  $\hat{\theta}_1 \sim N(\theta_1, \sigma^2 / |\mathbf{w}|^2)$ .
- (d) Betragt nu delmodellen  $M_2$  med  $\theta_1 = 0$ ,  $\boldsymbol{\xi} = \theta_2 \mathbf{v}_2$ . Vis, at under  $M_2$  er skøn over  $\theta_2$  givet ved  $\hat{\theta}_{20} = \frac{\mathbf{v}_2^T \mathbf{X}}{|\mathbf{v}_2|^2} = \hat{\gamma}$ . Vis dernæst, at  $|\hat{\boldsymbol{\xi}}(M_1) - \hat{\boldsymbol{\xi}}(M_2)|^2 = \hat{\theta}_1^2 |\mathbf{w}|^2$ .
- (e) Vis, at  $F$ -teststørrelsen for reduktion fra  $M_1$  til  $M_2$  er identisk med den kvadrerede  $t$ -teststørrelse for hypotesen  $\theta_1 = 0$ .

