

Introduktion til Statistik

5. udgave

Susanne Ditlevsen og Helle Sørensen

Susanne Ditlevsen, susanne@math.ku.dk
Helle Sørensen, helle@math.ku.dk

Institut for Matematiske Fag
Københavns Universitet
Universitetsparken 5
2100 København Ø

5. udgave, november 2018

Copyright Susanne Ditlevsen og Helle Sørensen


ISBN 978-87-7078-882-3

Forord

Dette notesæt er udarbejdet med henblik på statistikdelen af kurset *Sandsynlighedsregning og Statistik* (SS) på Københavns Universitet.

Notesættet er inspireret af Inge Henningsens noter til tidligere kurser (Henningsen, 2006a,b). Emnemæssigt afviger de fra Inges noter ved at næsten alt vedrørende modeller på diskrete udfaldsrum er skåret væk. Vi har også ladet os inspirere af bøgerne *Basal Biostatistik* (del I og II) som tidligere blev benyttet på Det Biovidenskabelige Fakultet på Københavns Universitet (Skovgaard *et al.*, 1999; Skovgaard, 2004) og af bogen *Introduction to Statistical Data Analysis for the Life Sciences* (Ekstrøm and Sørensen, 2010).

Notesættet omhandler kun en lille klasse af modeller, nemlig en simpel binomialfordelingsmodel, normalfordelingsmodeller for en enkelt eller to stikprøver samt lineær regression. Givet den mængde sandsynlighedsregning vi har til rådighed fra sandsynlighedsregningsdelen, er den nødvendige matematik ikke svær, men det betyder ikke nødvendigvis at stoffet er let. Vores erfaring er at statistikbegreberne er svære at få ind under huden, og vi gør derfor et stort nummer ud af forsøge at forklare meningen med og betydningen af de indførte begreber.

Alle kapitler på nær kapitel 2 indeholder et afsnit hvor vi viser hvordan R kan bruges til at udføre analyserne. For at få udbytte af disse afsnit er det nødvendigt med et basalt kendskab til R, herunder hvordan man indlæser data. Der findes en kort introduktion til R på Absalonsiden for kurset *Sandsynlighedsregning og Statistik*. Filer med data som bruges i eksempler eller opgaver, ligger samme sted. Opgaver der kræver brug af R, er mærket med symbolet . På kurset regnes i øvrigt mange andre opgaver.

Den vigtigste ændring i forhold til 4. udgave er at der nu henvises til bogen *Introduction to Probability* (Blitzstein and Hwang, 2015), der bruges på sandsynlighedsregningsdelen af kurset. Vi bruger forkortelsen BH til disse henvisninger. I tidligere

udgaver blev der henvist til Michael Sørensens bog *En Introduktion til Sandsynlighedsregning* (Sørensen, 2011). I samme ombæring har vi tilføjet appendix A med diverse resultater fra sandsynlighedsregning som kan være svære at finde direkte i BH. Appendix B indeholder en beskrivelse af profilmaksimering. Begge appendikser og en mindre del af kapitel 6 er omskrivninger af notater af vores kollega Ernst Hansen. Derudover har vi rettet trykfejl og foretaget andre mindre rettelser.

København, november 2018

Susanne Ditlevsen, Helle Sørensen

Indhold

Forord	3
1 Binomialfordelingen	9
1.1 Statistisk model	10
1.2 Maksimum likelihood estimation	12
1.3 Modeller med endeligt udfaldsrum	18
1.4 Sammenfatning og perspektiv	21
1.5 R	21
1.6 Opgaver	22
2 Normalfordelingsmodeller	27
3 En stikprøve med kendt varians	31
3.1 Statistisk model	31
3.2 Maksimum likelihood estimation	33
3.3 Konfidensinterval for middelværdien	36
3.4 Test af hypotese om middelværdien	40
3.5 Sammenfatning og perspektiv	48
3.6 R	49
3.7 Opgaver	51

4	En stikprøve med ukendt varians	55
4.1	Statistisk model	55
4.2	Maksimum likelihood estimation	56
4.3	Konfidensinterval for middelværdien	60
4.4	Test af hypotese om middelværdien	62
4.5	Kontrol af normalfordelingsantagelse	67
4.6	Sammenfatning og perspektiv	71
4.7	R	72
4.8	Opgaver	76
5	To stikprøver	81
5.1	Statistisk model	81
5.2	Maksimum likelihood estimation	83
5.3	Konfidensintervaller	86
5.4	Hypotesetest	89
5.5	Modelkontrol	95
5.6	Eksempel: Energiforbrug	98
5.7	Sammenfatning og perspektiv	101
5.8	R	101
5.9	Opgaver	104
6	Lineær regression	111
6.1	Statistisk model	112
6.2	Maksimum likelihood estimation	114
6.3	Konfidensintervaller	120
6.4	Hypotesetest	122
6.5	Regressionslinjen og prædiktions	126
6.6	Residualer og modelkontrol	130

INDHOLD	7
6.7 Eksempel: CAPM	134
6.8 Sammenfatning og perspektiv	142
6.9 R	143
6.10 Opgaver	147
A Resultater fra sandsynlighedsregning	153
B Profilmaksimering	158
Referencer	160
Indeks	160

Kapitel 1

Binomialfordelingen

I mange sammenhænge er man interesseret i hyppigheden for et givet fænomen, og man vil så indsamle data der indeholder information om denne hyppighed. Antag for eksempel at man er interesseret i risikoen for en given bivirkning (hovedpine) af et medicinsk præparat. Hvis man giver 100 patienter medicinen og undersøger hvor mange der får hovedpine (passende ofte og passende kraftigt), så vil andelen af patienter med hovedpine sige noget om denne risiko. Eller antag at man vil undersøge en persons evne til at smage forskel på Coca-cola og Pepsi. Personen får serveret to glas cola, et af hver slags, og skal så efter smagning udpege hvilket glas der indeholder Pepsi. Eksperimentet gentages 10 gange, og den relative hyppighed af gange hvor personen svarer korrekt indeholder information om hvorvidt personen kan smage forskel.

Det er ikke svært at beregne relative hyppigheder — problemet er hvor meget vi kan “stole på dem”. Hvis vi udførte eksperimentet på ny (med 100 nye patienter, eller med 10 nye smagstest), så ville vi næppe få præcis det samme resultat, så hvor pålidelige er de relative hyppigheder beregnet fra de data der nu engang er til rådighed? En vigtig pointe med en statistisk analyse er netop at den beskriver usikkerheden i de opnåede resultater!

Eksperimenterne ovenfor kan beskrives ved hjælp af binomialfordelingen, og vi skal i dette kapitel introducere de statistiske begreber *statistisk model*, *likelihoodfunktion* og *estimator* for en simpel binomialfordelingsmodel. Matematisk set er det ganske simpelt. Det vanskelige ligger snarere i at forstå selve begreberne og hvad de skal gøre godt for. Hovedformålet med dette kapitel er netop at give et indtryk af dette.

1.1 Statistisk model

En statistisk model skal bruges til at beskrive den usikkerhed der er forbundet med data. Modellen specificeres ved at angive udfaldsrummet samt de fordelinger som med rimelighed kan antages at have frembragt data. Vi vil i dette afsnit opstille en simpel statistisk model baseret på binomialfordelingen.

Lad os antage at vores observation (eller data) x er antallet af gange en given hændelse er indtruffet i n uafhængige gentagelser af samme forsøg. Sandsynligheden p for at hændelsen indtræffer er den samme i hvert forsøg. Forsøget kan være et smagsforsøg hvor den interessante hændelse er om personen kan udpege glasset med Pepsi, og p er sandsynligheden for at dette sker. Eller forsøget kan være medicinering af en patient hvor den interessante hændelse er om patienten får hovedpinebivirkninger, og p er sandsynligheden for at dette er tilfældet for en tilfældig patient.

Dette kan formaliseres ved hjælp af binomialfordelingen (BH, afsnit 3.3) idet vi kan tænke på observationen x som en realisation af en stokastisk variabel X der er binomialfordelt med antalsparameter n og sandsynlighedsparameter p . Udfaldsrummet for X er $E = \{0, 1, \dots, n\}$. Antalsparameteren n er et kendt tal (antallet af gentagelser), men sandsynlighedsparameteren p er ukendt. Det eneste vi ved, er at den ligger i intervallet $[0, 1]$.

For ethvert $p \in [0, 1]$ er der en tilhørende fordeling, og den statistiske model består af udfaldsrummet for X samt denne samling — eller familie — af fordelinger, altså alle binomialfordelinger med antalsparameter n . Sandsynlighedsparameteren p er som sagt ikke et kendt tal. Vi siger at p er en ukendt *parameter* som skal *estimeres* fra data. Det vil vi gøre i næste afsnit. Mængden af mulige værdier for parameteren kaldes *parametermængden* og benævnes Θ . Hvis der ikke er yderligere restriktioner på p så er $p \in \Theta = [0, 1]$, men Θ kan også være en mindre delmængde af $[0, 1]$.

Formelt kan vi specificere den statistiske model ved at angive udfaldsrummet samt familien af fordelinger, betegnet \mathcal{P} . Alternativt kan vi bruge en formulering der involverer den stokastiske variabel X . Hvis vi bruger notationen $\text{bin}(n, p)$ for binomialfordelingen med parametre n og p har vi altså følgende definition.

Definition 1.1. *Modellen for en enkelt binomialfordelt observation består af udfaldsrummet $E = \{0, 1, \dots, n\}$ samt familien*

$$\mathcal{P} = \{\text{bin}(n, p) : p \in \Theta\}$$

hvor $\Theta \subseteq [0, 1]$. *Alternativ formulering: Lad X være en stokastisk variabel med udfaldsrum $\{0, 1, \dots, n\}$, og antag at $X \sim \text{bin}(n, p)$ hvor $p \in \Theta$.*

Typen af fordeling, de enkelte fordelinger i modellen og den ukendte parameter formaliserer forskellige aspekter af vores viden/uvidenhed om det (videnskabelige) problem som data skal belyse. Vi kan fortolke ingredienserne på følgende måde:

- Valget af fordelingstype formaliserer vores forhåndsviden eller forhåndsantagelser. I situationen med uafhængige gentagelser af et forsøg med to udfald er binomialfordelingen det naturlige valg.
- De enkelte fordelinger formaliserer den usikkerhed der er forbundet med observationerne. Mere specifikt: for en fast værdi af p angiver sandsynlighedsfunktionen for $\text{bin}(n, p)$ fordelingen af X :

$$P(X = x) = f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n.$$

Bemærk fodtegnet på f der understreger at sandsynlighedsfunktionen afhænger af p .

- Mængden af sandsynlighedsfordelinger — specificeret ved mængden af mulige parametre — i modellen formaliserer den uvidenhed vi har om de mekanismer der har frembragt observationerne. Vi ved ikke hvilken værdi af p der kan antages at have frembragt x . Det er ikke nødvendigvis altid rimeligt at bruge hele $[0, 1]$ som parametermængde. I eksemplet med smagstesten er det svært at fortolke sandsynligheder der er mindre end $1/2$ — det svarer til at personen vælger det korrekte glas sjældnere end hvis han gætter — så man kan hævde at den naturlige parametermængde er $\Theta = [1/2, 1]$. Dette vil vi dog ikke gøre mere ud af i det følgende.

I situationen med uafhængige gentagelser af samme forsøg virkede det oplagt at bruge binomialfordelingen, men normalt er det en vanskelig sag at vælge en statistisk model. Hvis gentagelserne ikke er uafhængige — for eksempel fordi forsøgspersonen ikke skyller munden mellem smagstestene, eller fordi nogle af patienterne er i familie og dermed har fælles gener, så er antallet ikke binomialfordelt. Tilsvarende hvis sandsynligheden ikke er den samme i de enkelte gentagelser, for eksempel fordi der kan være forskel på mænds og kvinders tendens til hovedpine.

I virkeligheden tror vi ikke nødvendigvis at alle forudsætningerne der ligger til grund for en given model, er opfyldt. Vi bruger snarere modellen som en approksimation til virkeligheden fordi vi mener at den giver en god beskrivelse af usikkerheden i data og samtidig beskriver vores mangel på fuldstændig viden. Det skal selvfølgelig

undersøges nærmere om modellen giver en rimelig beskrivelse af data fordi konklusionerne — resultaterne af den statistiske analyse — afhænger kritisk af forudsætningerne i modellen.

1.2 Maksimum likelihood estimation

Hvis $X \sim \text{bin}(n, p)$ for et givet p så beskriver sandsynlighedsfunktionen

$$f_p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n \quad (1.1)$$

sandsynlighederne for de mulige udfald af X : *hvis sandsynlighedsparameteren er p så er sandsynligheden for at observere x som angivet.* Det er sådan vi tænker når vi laver sandsynlighedsregning.

Vores situation er imidlertid den modsatte: vi *har* en observation x , men kender ikke sandsynlighedsparameteren p . Udfra observationen ønsker vi at *estimere* parameteren p . Det betyder løst sagt at finde den værdi af p der “passer bedst muligt” med observationen x . Det kan jo betyde hvad som helst og skal præciseres nærmere: som estimat vil vi bruge den værdi af p der gør det mest sandsynligt at observere netop den værdi af X som vi har observeret. Tankegangen er altså at beregne $f_p(x) = P(X = x)$ — for den observerede værdi x — for alle mulige værdier af p og så vælge den værdi af p der giver den største værdi.

Dette formaliseres ved hjælp af *likelihoodfunktionen*. Likelihoodfunktionen er identisk med sandsynlighedsfunktionen — bortset fra at den nu opfattes som funktion af p for fast x snarere end omvendt. Hvis parametermængden er Θ , så er likelihoodfunktionen hørende til observationen x defineret ved

$$L_x: \Theta \rightarrow [0, 1]$$

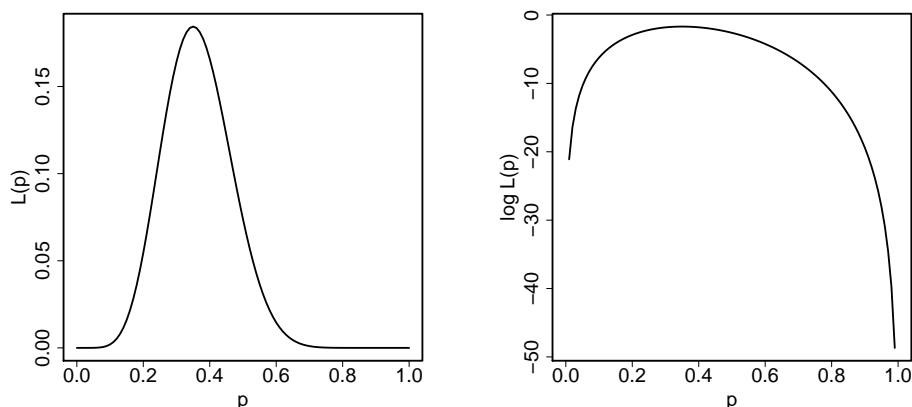
$$L_x(p) = f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad p \in \Theta.$$

Som estimat for p vil vi bruge den værdi i Θ der gør L_x størst mulig, hvor x altså holdes fast i observationsværdien. Vi søger således en værdi $\hat{p} \in \Theta$ så

$$L_x(\hat{p}) \geq L_x(p), \quad p \in \Theta,$$

og kalder \hat{p} for et *maksimum likelihood estimat* eller et maksimaliseringsestimater for p . Man bruger også forkortelsen MLE. Maksimum likelihood estimatet afhænger af den observerede værdi x og for at understrege dette skriver vi sommetider $\hat{p}(x)$.

Maksimum likelihood estimation er illustreret i venstre side af figur 1.1. Likelihood-funktionen er tegnet som funktion af p for $n = 20$ og $x = 7$. Det følger af sætningen nedenfor at funktionen har maksimum for $p = 7/20 = 0.35$.



Figur 1.1: Likelihoodfunktionen (til venstre) og log-likelihoodfunktionen (til højre) som funktion af p for $x = 7$ i en binomialfordeling med $n = 20$. Maksimum antages for $p = x/n = 0.35$.

Sætning 1.2. For den statistiske model fra definition 1.1 med $\Theta = [0, 1]$ er maksimum likelihood estimatet for p entydigt bestemt og givet ved $\hat{p}(x) = x/n$.

Bevis Da x er fast, er binomialkoefficienten uden betydning for optimeringsproblemet. Vi definerer derfor funktionen $g : [0, 1] \rightarrow \mathbb{R}$ ved

$$g(p) = p^x(1-p)^{n-x}.$$

Bemærk først at hvis $x = 0$ så har g maksimum for $p = 0$, og hvis $x = n$ så har g maksimum for $p = 1$. Altså er $\hat{p}(x) = x/n$ i disse tilfælde.

Antag dernæst at $x \in \{1, \dots, n-1\}$. Så er $g(p) = 0$ for $p \in \{0, 1\}$, men $g(p) > 0$ for $p \in (0, 1)$, så en løsning skal søges blandt stationære punkter. Funktionen h givet ved

$$h(p) = \log g(p) = x \log(p) + (n-x) \log(1-p)$$

er veldefineret på $(0, 1)$ og har maksimum samme sted som g da \log er strengt voksende. Desuden er h to gange kontinuert differentiabel med

$$h'(p) = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x-np}{p(1-p)}$$

$$h''(p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}.$$

Specielt er $h'(p) = 0$ hvis og kun hvis $p = x/n$ og $h''(p) < 0$ for alle $p \in (0, 1)$. Således har h og dermed g maksimum for $p = x/n$. \square

Bemærk at vi med det samme fjernede binomialkoefficienten fra optimeringsproblemet: der er ikke nogen grund til at slæbe rundt på led der ikke afhænger af parameteren p . Bemærk også at vi lavede funktionsundersøgelse for funktionen h , defineret som logaritmen til likelihoodfunktionen (på nær en konstant), snarere end likelihoodfunktionen selv. Vi taler også om *log-likelihoodfunktionen*. Den er illustreret i højre side af figur 1.1. Dette “trick” benyttes ofte, blandt andet fordi produkter derved bliver omsat til summer der er meget nemmere at regne med.

Resultatet fra sætning 1.2 er ikke særligt overraskende: sandsynligheden for at en given hændelse indtræffer skal estimeres ved den relative hyppighed af gange hændelsen indtræffer i n uafhængige eksperimenter. Det er faktisk svært at forestille sig nogen anden estimator for p , men der er alligevel nogle vigtige pointer at notere sig.

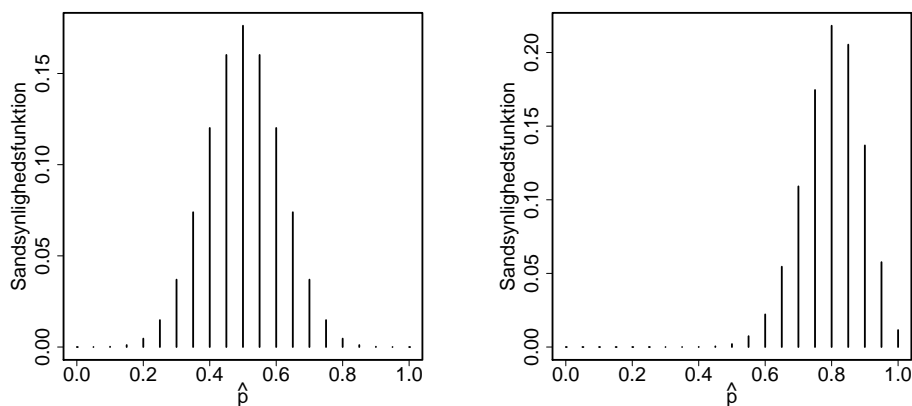
Den vigtigste er fortolkningen af $\hat{p} = x/n$ som realisationen af den stokastiske variabel $\hat{p}(X) = X/n$. Denne variabel kaldes maksimum likelihood estimatoren. Vi skelner således mellem estimatet x/n som er et tal og estimatoren X/n som er en stokastisk variabel — og derfor har en fordeling. Da X kan antage værdierne $0, 1, \dots, n$ kan \hat{p} antage værdierne $0, 1/n, 2/n, \dots, 1$ og sandsynlighedsfunktionen for \hat{p} er givet ved

$$P\left(\hat{p} = \frac{x}{n}\right) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n.$$

Fordelingen af \hat{p} er illustreret i figur 1.2 for $n = 20$, til venstre for $p = 0.5$ og til højre for $p = 0.8$. Det er nok nemmest at forstå hvad fordelingen af \hat{p} betyder hvis vi forestiller os dataindsamlingen — for eksempel et smageeksperiment med 20 gentagelser — gentaget mange gange. Hver dataindsamling giver anledning til en observation x og dermed et estimat $\hat{p} = x/n$. Hvis den sande værdi af sandsynlighedsparameteren er 0.5 vil vi for eksempel i cirka 12% af tilfældene få estimatet 0.6 (venstre side af figur 1.2). Hvis den sande værdi af sandsynlighedsparameteren derimod er 0.8 vil dette kun ske i cirka 2% af tilfældene (højre side af figur 1.2).

En anden måde at udtrykke fordelingen af \hat{p} er ved at sige at $n\hat{p}$ — som jo netop er X — er binomialfordelt med antalsparameter n og sandsynlighedsparameter p . Hvis den sande parameter er p således at $X \sim \text{bin}(n, p)$, følger det af eksempel 4.2.2 og 4.6.5 i BH, at $n\hat{p}$ har middelværdi og varians

$$E(n\hat{p}) = E(X) = np, \quad \text{Var}(n\hat{p}) = \text{Var}(X) = np(1-p).$$



Figur 1.2: Sandsynlighedsfunktionen for \hat{p} for $n = 20$. Sandsynlighedsparameteren er $p = 0.5$ (til venstre) og $p = 0.8$ (til højre).

Det følger derefter fra BH, sætning 4.2.1 og næstøverste bulletpoint på side 159 i BH, at \hat{p} har middelværdi

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} np = p \quad (1.2)$$

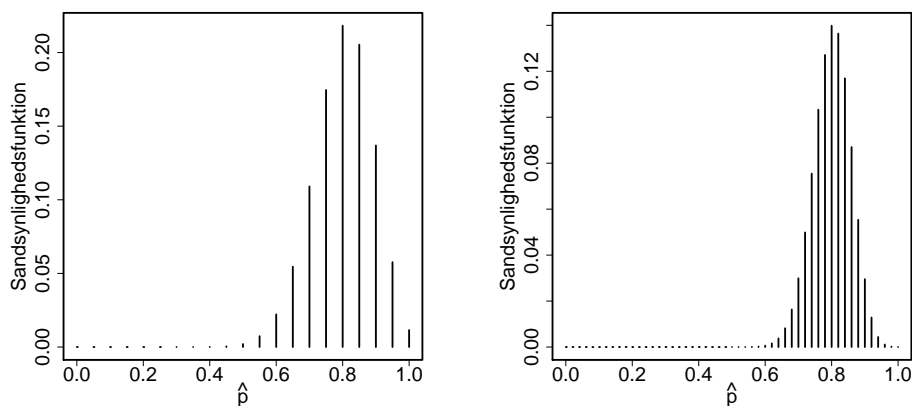
og varians

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(X) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}. \quad (1.3)$$

Egenskaben (1.2) udtrykker at middelværdien af maksimum likelihood estimatoren er lig den sande værdi, og vi siger at \hat{p} er en *central estimator* for p . Dette illustreres af figur 1.2 hvor middelværdierne er 0.5 henholdsvis 0.8. At \hat{p} er central betyder løst sagt at estimatoren “i gennemsnit” rammer den sande værdi, dvs. at gennemsnittet af estimer fra mange uafhængige forsøg vil nærme sig den sande værdi i passende forstand.

Egenskaben (1.3) udtrykker blandt andet at variansen af \hat{p} er aftagende i n . Dette giver god mening: flere gentagelser giver anledning til større præcision. Dette er illustreret i figur 1.3 hvor sandsynlighedsfunktionen for \hat{p} er tegnet for $(n, p) = (20, 0.8)$ til venstre og $(n, p) = (50, 0.8)$ til højre. Specielt er p altså ens i de to figurer. Fordelingen af \hat{p} er tydeligtvis smallere for $n = 50$ end for $n = 20$.

Lad os formulere egenskaberne ved fordelingen af \hat{p} i en sætning:



Figur 1.3: Sandsynlighedsfunktionen for \hat{p} for $n = 20$ (til venstre) og $n = 50$ (til højre). Sandsynlighedsparameteren er $p = 0.8$ i begge figurer.

Sætning 1.3. Lad $\hat{p} = X/n$ være maksimum likelihood estimatoren for den statistiske model fra definition 1.1 med $\Theta = [0, 1]$. Så er $n\hat{p}$ binomialfordelt,

$$n\hat{p} \sim \text{bin}(n, p).$$

Specielt er $E(\hat{p}) = p$ og $\text{Var}(\hat{p}) = p(1-p)/n$.

Der er en ikke ubetydelig hage ved fordelingsresultatet fra sætning 1.3: vi kender ikke den sande værdi af p . Ikke desto mindre er vi glade for resultatet: estimatoren har en kendt fordeling og er oven i købet central med en varians der aftager med antalsparameteren. Desuden har vi jo et estimat for \hat{p} og vi kan derfor få et estimat for fordelingen ved at indsætte dette estimat: den estimerede fordeling for $n\hat{p}$ er $\text{bin}(n, x/n)$.

Bemærk specielt at den estimerede spredning for \hat{p} er $\sqrt{\hat{p}(1-\hat{p})/n}$, jf. (1.3). Vi vil sommetider skrive $s(\hat{p})$ for denne estimerede spredning, altså

$$s(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}}.$$

Eksempel 1.4. (Smagsforsøg) En forsøgsperson får serveret to glas cola (Coca-cola og Pepsi) og bliver bedt om at udpege glasset med Pepsi. Dette gentages 20 gange og personen udvælger det rigtige glas $x = 15$ gange. Under passende antagelser — overvej selv hvilke — er det rimeligt at antage at x er en realisation af en $\text{bin}(20, p)$ -fordelt stokastisk variabel hvor p er sandsynligheden for at personen kan udpege

glasset med Pepsi i en tilfældig smagsprøve. Estimatet for p er således $\hat{p} = 15/20 = 0.75$, og hvis vi bruger $\Theta = [0, 1]$ som parametermængde, så er $n\hat{p} = X \sim \text{bin}(20, p)$. Den estimerede fordeling af $n\hat{p}$ er $\text{bin}(20, 0.75)$, og \hat{p} har estimeret spredning $s(\hat{p}) = 0.0968$.

Bemærk at værdien $p = 0.5$ svarer til at forsøgspersonen ikke kan smage forskel: han eller hun gætter, og gætter derfor rigtigt med sandsynlighed 0.5 hver gang. Værdier større end 0.5 svarer derimod til at personen i en vis udstrækning kan smage forskel. Hvis $p = 0.5$, så er $X \sim \text{bin}(20, 0.5)$ og så er sandsynlighedsfunktionen for \hat{p} den som er tegnet i den venstre del af figur 1.2. Her kan vi se at det er ret usædvanligt at observere værdier af \hat{p} der er 0.75 eller større, dvs. værdier af X der er 15 eller større. Der er således et vist belæg for at hævde at forsøgspersonen faktisk kan smage forskel. \square

Eksempel 1.5. (*Mendelsk spaltning*) For at undersøge arvelighed udførte Gregor Mendel i midten af 1800-tallet en lang række eksperimenter med ærteblomster. I et af forsøgene undersøgte Mendel farvefordelingen for 1238 såkaldte andengenerationsfrø (se nedenfor): 949 var gule og 289 var grønne. Hvis vi antager at hvert af frøene har samme sandsynlighed for at blive gult og at ærtefrøene ikke har noget med hinanden at gøre, kan vi antage at antallet af gule frø er binomialfordelt med antalsparameter $n = 1238$ og sandsynlighedsparameter p .

Estimatet for p er dermed $\hat{p} = 949/1238 = 0.767$. Estimatorens fordeling er givet ved $n\hat{p} \sim \text{bin}(1238, p)$, den estimerede fordeling af $n\hat{p}$ er $\text{bin}(1238, 0.767)$, og \hat{p} har estimeret spredning $s(\hat{p}) = 0.012$.

Farven på frøet bestemmes af hvad vi i dag ville kalde et gen. Farvegenet forekommer i to varianter: A der er dominant og giver gul farve og a der er recessiv og giver grøn farve. I eksperimentet krydsede Mendel individer med genotype AA og individer med genotype aa . I første generation er alle individerne af type Aa og dermed gule. I anden generation er genotyperne givet ved følgende skema:

Køns-celle	A	a
A	AA	Aa
a	aA	aa

Hvis de mendelske regler for arvelighed gælder, vil forekomsten af fænotyperne — altså ærternes udseende — være i forholdet 3:1 mellem gule og grønne idet gul forekommer for kombinationerne AA , Aa og aA , mens grøn kun forekommer for kombinationen aa . Dette svarer til at sandsynlighedsparameteren i den statistiske model er $p = 0.75$.

Hvis den sande værdi af p er 0.75, så er $n\hat{p} = X \sim \text{bin}(1238, 0.75)$. Vi kan så beregne

$$P(\hat{p} \leq 0.767) = P(X \leq 949) = 0.927$$

$$P(\hat{p} \geq 0.767) = P(X \geq 949) = 0.094$$

hvilket indikerer at den observerede værdi af \hat{p} ligger rimeligt centralt i fordelingen. Data er således ikke umiddelbart i modstrid med de mendelske regler. \square

Sommetider er man interesseret i hvorvidt en specifik værdi af sandsynlighedsparameteren, p_0 , er rimelig eller ej, data taget i betragtning. Som antydnet i eksemplerne ovenfor undersøger man så hvor ekstremt den observerede værdi af \hat{p} ligger i fordelingen af \hat{p} hvis sandsynlighedsparameteren faktisk er p_0 . Hvis estimatet ligger ekstremt i fordelingen, svarende til at de observerede data er usandsynlige, så konkluderer man at værdien p_0 næppe er den rigtige. Omvendt, hvis estimatet ligger rimeligt centralt i fordelingen konkluderer man at p_0 ikke kan afvises at være den rigtige. Som tommelfingerregel kan man sige at værdien p_0 er i god overensstemmelse med data hvis p_0 ligger i intervallet fra $\hat{p} \pm 2 \cdot s(\hat{p})$. Mere formelt kan man udføre et hypotesetest. Vi vil ikke sige yderligere om hypotesetest for binomialdata, men vender tilbage til det i kapitel 3.

Inden vi gør situationen lidt mere generel er det værd at dvæle ved det princip som vi brugte til at finde \hat{p} : Maksimum likelihood estimatoren $\hat{p}(x)$ er den værdi af p som maksimerer likelihoodfunktionen, dvs. den værdi af p der gør den observerede værdi x mest sandsynlig. Det virker ikke helt tåbeligt. Antag et øjeblik at der kun er to mulige sandsynligheder, for eksempel 0.15 og 0.50, svarende til $\Theta = \{0.15, 0.50\}$, og at vi har observeret værdien $x = 2$ i en binomialfordeling med antalsparameter 10. Så er

$$P_{0.15}(X = 2) = 0.276; \quad P_{0.50}(X = 2) = 0.044$$

hvor vi har brugt fodtegn til at markere værdien af sandsynlighedsparameteren, og det virker fornuftigt at tro mere på at den "sande" sandsynlighed er 0.15 end 0.50. Det er denne tankegang der er generaliseret til tilfældet hvor p tillades at variere i hele intervallet $[0, 1]$.

1.3 Modeller med endeligt udfaldsrum

I dette afsnit beskriver vi maksimum likelihood estimation for statistiske modeller med endeligt udfaldsrum. Binomialfordelingsmodellen fra definition 1.1 er et specialtilfælde, og formålet med at se på den mere generelle klasse af modeller er at understrege at maksimum likelihood metoden er et generelt estimationsprincip.

Antag at data kan beskrives ved hjælp af en fordeling på en endelig mængde E med en sandsynlighedsfunktion som er kendt, bortset fra at den afhænger af en ukendt parameter. Lad os kalde parameteren θ og antage at den varierer i parametermængden Θ . Parameteren θ kan være flerdimensional, for eksempel d -dimensional, således at Θ er en delmængde af \mathbb{R}^d . For hvert $\theta \in \Theta$ har vi altså en sandsynlighedsfunktion $f_\theta : E \rightarrow [0, 1]$ hvor $f_\theta(x)$ er sandsynligheden for at observere x hvis parameteren er θ .

Vi forestiller os nu at vi har en observation x og tænker på x som en realisation af en stokastisk variabel X med sandsynlighedsfunktion f_θ . Vi opfatter sandsynlighedsfunktionen som funktion af den ukendte parameter θ , for den observerede værdi x . Dette giver os likelihoodfunktionen, $L_x : \Theta \rightarrow [0, 1]$,

$$L_x(\theta) = f_\theta(x), \quad \theta \in \Theta,$$

og en maksimum likelihood estimator er en værdi $\hat{\theta} \in \Theta$ der gør L_x størst mulig:

$$L_x(\hat{\theta}) \geq L_x(\theta), \quad \theta \in \Theta.$$

Som for binomialfordelingsmodellen vil estimatoren $\hat{\theta}$ afhænge af observationen x . Vi skriver således $\hat{\theta}(x)$ og kan også betragte estimatoren $\hat{\theta}(X)$ som en stokastisk variabel og tale om dens fordeling.

Bemærk at det ikke på forhånd er givet at estimatet eksisterer og er entydigt bestemt. Det skal undersøges for en given model ligesom vi gjorde det for binomialmodellen.

Eksempel 1.6. (*Legetøjseksempel*) Antag at observationen x er et udfald af en stokastisk variabel der kan antage værdierne 0, 1 og 2, og at fordelingen af X har sandsynlighedsfunktion

$$f_\theta(x) = \begin{cases} \theta/4, & x = 0 \\ 3\theta/4, & x = 1 \\ 1 - \theta, & x = 2 \end{cases}$$

for en ukendt parameter θ . Overvej selv at dette definerer et sandsynlighedsmål hvis og kun hvis $\theta \in [0, 1]$. Således er $\Theta = [0, 1]$ den naturlige parametermængde.

Likelihoodfunktionen fås ved at betragte sandsynlighedsfunktionen som funktion af θ for fast x , altså $L_x(\theta) = f_\theta(x)$ for $\theta \in [0, 1]$. Det er klart at L_x har maksimum for $\theta = 1$ hvis $x = 0, 1$ og for $\theta = 0$ hvis $x = 2$. Således eksisterer maksimum likelihood estimatet og er entydigt givet ved

$$\hat{\theta}(x) = \begin{cases} 1, & x = 0, 1 \\ 0, & x = 2 \end{cases}$$

Den tilhørende estimator $\hat{\theta} = \hat{\theta}(X)$ er en stokastisk variabel med værdier i $\{0, 1\}$ og fordeling givet ved

$$\begin{aligned} P(\hat{\theta}(X) = 1) &= P(X \in \{0, 1\}) = \frac{\theta}{4} + \frac{3\theta}{4} = \theta \\ P(\hat{\theta}(X) = 0) &= P(X = 2) = 1 - \theta. \end{aligned}$$

Specielt er $E(\hat{\theta}) = \theta$, så $\hat{\theta}$ er en central estimator for θ . □

Eksempel 1.7. (*Ventetid*) Betragt et forsøg med to udfald (succes og fiasko), og antag at det gentages indtil succesudfaldet indtræffer, dog højst 4 gange. Hvis X er en stokastisk variabel der tæller antallet af gange forsøget gentages, så har X udfaldsrum $\{1, 2, 3, 4\}$, og hvis successandsynligheden er p , så har X sandsynlighedsfunktion

$$f_p(x) = \begin{cases} p(1-p)^{x-1}, & x = 1, 2, 3 \\ (1-p)^3, & x = 4. \end{cases}$$

Se også opgave 1.7.

Vi antager at sandsynlighedsparameteren $p \in [0, 1]$ er ukendt og skal estimeres på baggrund af en observation x . Som for binomialmodellen opstiller vi likelihoodfunktionen ved at betragte sandsynlighedsfunktionen som funktion af p snarere end x :

$$L_x(p) = f_p(x), \quad p \in [0, 1].$$

Maksimum likelihood estimatet er så en værdi af p der gør $L_x(p)$ størst mulig. Det viser sig — se igen opgave 1.7 — at

$$\hat{p}(x) = \begin{cases} 1/x, & x = 1, 2, 3 \\ 0, & x = 4 \end{cases}$$

Udfaldsrummet for estimatoren $\hat{p} = \hat{p}(X)$ er altså $\{1, 1/2, 1/3, 0\}$, og sandsynlighedsfunktion er givet ved

$$P(\hat{p} = y) = \begin{cases} p, & y = 1 \\ p(1-p), & y = 1/2 \\ p(1-p)^2, & y = 1/3 \\ (1-p)^3, & y = 0 \end{cases}$$

Specielt kan vi regne på middelværdien af \hat{p} :

$$E(\hat{p}) = p + \frac{1}{2}p(1-p) + \frac{1}{3}p(1-p)^2 = p \left(\frac{11}{6} - \frac{7}{6}p + \frac{1}{3}p^2 \right)$$

der er lig p når $p \in \{0, 1\}$, men ellers skarpt større end p . Det er altså ikke alle estimatører der er centrale. □

1.4 Sammenfatning og perspektiv

Vi har studeret en situation hvor data kan tænkes at komme fra uafhængige gentagelser af et eksperiment med to mulige udfald. I denne ramme har vi defineret og undersøgt følgende:

En statistisk model er en familie af binomialfordelinger hvor sandsynlighedsparameteren er ukendt og skal estimeres ved hjælp af data.

Maksimum likelihood estimatet er den værdi af p der gør den observerede værdi mest sandsynlig.

Maksimum likelihood estimatoren er den tilhørende stokastiske variabel forstået på den måde at estimatet er den observerede værdi af estimatoren. Fordelingen af estimatoren beskriver den usikkerhed der er forbundet med estimatet, og vi kan specielt interessere os for estimatorens middelværdi, varians og spredning.

Maksimum likelihood estimation er et meget generelt estimationsprincip, og vi beskrev metoden for statistiske modeller med endeligt udfaldsrum. Senere i bogen skal vi se hvordan samme princip kan bruges for statistiske modeller baseret på normalfordelingen.

Der findes andre estimationsprincipper, for eksempel momentestimation. I den givne binomialfordelingsmodel betyder det at estimere p således at $E(X)$ er lig den observerede værdi x . Når X er binomialfordelt med parametre n og p er $E(X) = np$ så kravet er at $np = x$ eller $p = x/n$. I dette tilfælde giver de to estimationsprincipper altså den samme estimator, men dette er ikke altid tilfældet. Generelt set foretrækker vi estimators der er centrale, dvs. som opfylder $E(\hat{p}) = p$, og har lille varians. Man kan for en meget generel klasse af modeller vise at maksimum likelihood estimatoren har lignende egenskaber (for n stor nok) således at vi normalt foretrækker den, men det ligger langt udenfor dette kursus at indse disse ting.

1.5 R

Beregningerne i dette kapitel er så simple at de nemt kan udføres på en lommeregner eller “manuelt” i R. Det kan dog være nyttigt at kende funktionerne `dbinom` og `pbinom` der beregner værdier af sandsynlighedsfunktionen og fordelingsfunktionen for binomialfordelingen.

Antag for eksempel at X er binomialfordelt med antalsparameter 20 og sandsynlighedsparameter 0.3. Vi kan beregne $P(X = 3)$ og $P(X \leq 3)$ således:

```
> dbinom(3, size=20, p=0.3)          # P(X=3), X~bin(20,0.3)
[1] 0.07160367
> dbinom(0:3, size=20, p=0.3)        # P(X=x) for x=0,1,2,3
[1] 0.0007979227 0.0068393371 0.0278458725 0.0716036722
> sum(dbinom(0:3, size=20, p=0.3)) # Summen, dvs. P(X <= 3)
[1] 0.1070868
> pbinom(3, size=20, p=0.3)          # P(X <= 3) igen
[1] 0.1070868
```

Funktionen `rbinom` bruges til simulation af udfald fra binomialfordelingen. Følgende kommando simulerer 10 udfald fra $\text{bin}(20, 0.3)$:

```
> rbinom(10, size=20, p=0.3) # 10 udfald fra bin(20,0.3)
[1] 4 4 8 5 9 6 7 5 7 6
```

Hvis kommandoen gentages, fås et andet output da kommandoen genererer tilfældige tal.

Bemærk at man ikke behøver skrive `size=` og `p=`. Kommandoerne

```
> dbinom(3, 20, 0.3)
> pbinom(3, 20, 0.3)
> rbinom(10, 20, 0.3)
```

er således identiske med de ovenstående.

1.6 Opgaver

1.1 Et opgavesæt består af 50 spørgsmål af vekslende sværhedsgrad. Hvert spørgsmål kan besvares enten rigtigt eller forkert.

1. Kan binomialfordelingen bruges til at beskrive antallet af rigtige svar for en enkelt person?
2. Kan binomialfordelingen bruges til at beskrive antallet af gange 50 personer besvarer prøvens første spørgsmål rigtigt?

1.2 En valutahandler registrerer i en periode på 21 dage om renten på en bestemt obligation stiger i forhold til den foregående dag. Under hvilke omstændigheder kan binomialfordelingen bruges til at beskrive antallet af dage hvor renten er steget?

1.3 For at undersøge udviklingen på aktiemarkedet en bestemt dag udvælges 10 aktier, og det registreres hvor mange af aktierne der er faldet i kurs den pågældende dag.

1. Under hvilke omstændigheder kan binomialfordelingen bruges til at beskrive antallet af aktier hvor kursen er faldet? Hvad er fortolkningen af sandsynlighedsparameteren p ?

Antag at omstændighederne er opfyldt og at kursen faldt for otte af aktierne, dvs. $x = 8$.

2. Opstil en statistisk model der kan bruges til at beskrive eksperimentet. Angiv et estimat for p , den tilhørende estimators fordeling, og den estimerede spredning for estimatoren.
3. Værdien 0.5 af sandsynlighedsparameteren er særligt interessant. Hvorfor?
4. Antag at sandsynlighedsparameteren er 0.5. Hvad er så sandsynligheden for at mindst 8 aktier faldt i kurs, og hvad er sandsynligheden for at højst 8 aktier faldt i kurs?
5. Tyder data på at der har været en generel udvikling i aktiekurserne den pågældende dag?

Vink: Vi har ikke de præcise redskaber til at svare på dette, men overvej følgende: Hvis der ikke har været en generel ændring, hvor usædvanligt er det så at have fået data hvor estimatet ligger så langt væk fra 0.5, som det vi fik? Du kan antage at aktiekurser nødvendigvis stiger eller falder (ikke er uændrede) fra dag til dag.

1.4 Kødprøver analyseres med kemiske test for tilstedeværelsen af bestemte typer bakterier. Ideelt set er prøven positiv hvis bakterietypen er i kødet og negativ hvis bakterietypen ikke er i kødet. Tabellen nedenfor viser resultaterne for 62 kødprøver med bakterien *E. coli* O157 og 131 kødprøver uden bakterien *E. coli*-O157. Som det ses er testen ikke perfekt.

	Positiv test	Negativ test	Total
Kød med <i>E. coli</i> -O157	57	5	62
Kød uden <i>E. coli</i> -O157	4	127	131

Sensitiviteten af testen defineres som sandsynligheden for at testen er positiv hvis bakterien er tilstede, mens specificiteten defineres som sandsynligheden for at testen er negativ hvis bakterien ikke er tilstede.

1. Angiv et estimat for sensitiviteten af testen og et estimat for specificiteten af testen.
2. Beregn den estimerede spredning for estimatoren for sensitiviteten og den estimerede spredning for estimatoren for specificiteten.
3. Antag at man planlægger et nyt forsøg og at man ønsker en estimeret spredning for sensitiviteten på 0.02. Hvor mange kødprøver bør man bruge?

1.5 Antag at en mønt enten har sandsynligheden $p = 1/2$ eller $p = 1/4$ for at vise krone. Mønten kastes n gange og viser krone x gange.

1. Opskriv en statistisk model der beskriver forsøget. Specielt: hvad er parametermængden?
2. Vis at $L_x(0.5) = L_x(0.25)$ hvis og kun hvis $x = x_0$ hvor

$$x_0 = \frac{n \log(3/2)}{\log(3)}.$$

3. Vis at $\hat{p}(x) = 0.25$ hvis $x < x_0$ og at $\hat{p}(x) = 0.75$ hvis $x > x_0$ (bemærk at x stadig er et heltal mellem 0 og n).
4. Antag at $n = 5$, og bestem $P_{1/2}(\hat{p} = 1/2)$ og $P_{1/4}(\hat{p} = 1/2)$, dvs. sandsynligheden for at $\hat{\theta} = 1/2$ når $p = 1/2$ henholdsvis $p = 1/4$. Kommenter resultatet.

1.6 Betragt eksempel 1.6. Vis at f_θ definerer en sandsynlighedsfunktion hvis og kun hvis $\theta \in [0, 1]$, se evt. BH, sætning 3.2.7 (støtten er endelig, så der kun er endeligt mange led i summen).

1.7 Betragt eksempel 1.7 om ventetid.

1. Vis at X har sandsynlighedsfunktion f_p som angivet i eksemplet.

2. Vis at maksimum likelihood estimatet $\hat{p}(x)$ er som angivet i eksemplet.
3. Gør rede for at maksimum likelihood estimatoren har sandsynlighedsfunktion som angivet i eksemplet.
4. Vis at middelværdien af \hat{p} er som påstået i eksemplet og at den er større end p for $p \in (0, 1)$. Forklar hvad det betyder.

1.8 Lad $\theta \in \{1, 2, \dots\}$ være en ukendt parameter, og antag at X er en stokastisk variabel med udfaldsrum $\{1, 2, \dots, \theta\}$ og punktsandsynligheder

$$f_{\theta}(x) = P(X = x) = \frac{1}{\theta}, \quad x \in \{1, \dots, \theta\}. \quad (1.4)$$

1. Gør rede for at (1.4) faktisk definerer en sandsynlighedsfunktion for en vilkårlig værdi $\theta \in \{1, 2, \dots\}$.
2. Opstil likelihoodfunktionen for θ og find derefter maksimum likelihood estimatet. *Vink:* For et givet x , hvad er de mulige værdier af θ ?

1.9 Lad X_1, \dots, X_n være uafhængige stokastiske variable hvor X_i er binomialfordelt med antalsparameter m_i og sandsynlighedsparameter p . Bemærk at sandsynlighedsparameteren er den samme for alle X_i . Specielt er de mulige værdier for X_i værdierne $0, 1, \dots, m_i$, så fordelingen af (X_1, \dots, X_n) er koncentreret på $M = \{0, 1, \dots, m_1\} \times \dots \times \{0, 1, \dots, m_n\}$.

1. Vis at sandsynlighedsfunktionen for $X = (X_1, \dots, X_n)$ er givet ved

$$p(x_1, \dots, x_n) = \left[\prod_{i=1}^n \binom{m_i}{x_i} \right] p^s (1-p)^{m-s}, \quad (x_1, \dots, x_n) \in M$$

hvor $s = \sum_{i=1}^n x_i$ og $m = \sum_{i=1}^n m_i$.

Antag nu at vi har observeret (x_1, \dots, x_n) og vil estimere p .

2. Opskriv likelihoodfunktionen og log-likelihoodfunktionen.
3. Find maksimum likelihood estimatet for p .
4. Angiv fordelingen af maksimum likelihood estimatoren.

Antag i stedet at vi kun har observeret summen $s = x_1 + \dots + x_n$ (i stedet for alle x_i 'erne).

5. Opstil en statistisk model der beskriver s . Angiv estimatet for p baseret på denne observation og estimatorens fordeling. Sammenlign med spørgsmål 3 og 4 og forklar resultatet.

1.10 Dette er en fortsættelse af opgave 1.9. For at undersøge tilfredsheden med bibliotekerne har man i en kommune tre dage i træk spurgt 25 biblioteksgængere om de er tilfredse med serviceniveauet. Der var kun to svarmuligheder: tilfreds eller ikke tilfreds. På de tre dage svarede henholdsvis 16, 18 og 13 borgere at de var tilfredse.

1. Opstil en statistisk model der beskriver data.
2. Bestem et estimat for andelen af tilfredse biblioteksgængere i kommunen.
3. Angiv fordelingen af estimatoren samt den estimerede spredning for estimatoren.

Kapitel 2

Normalfordelingsmodeller

I dette og de følgende kapitler skal vi beskæftige os med statistisk analyse af data der kan antages at være normalfordelte. Vi skal diskutere statistiske modeller, maksimum likelihood estimatorer, konfidensintervaller, hypotesetest, og modelkontrol.

Vi vil overalt antage at data består af n observationer y_1, \dots, y_n og tænke på dem som realisationer eller udfald af stokastiske variable Y_1, \dots, Y_n . Den statistiske model består så af udfaldsrummet og de mulige simultane fordelinger for (Y_1, \dots, Y_n) . Tre antagelser går igen for alle de normalfordelingsmodeller vi skal kigge på i disse noter.

Uafhængighed Den første antagelse er at Y_1, \dots, Y_n er uafhængige. Dette letter opgaven med at opstille en statistisk model betragteligt fordi det så er nok at beskrive de marginale fordelinger: Tætheden for den simultane fordeling er lig produktet af de marginale tætheder (sætning A.1 i appendiks A eller afsnit 7.1.2 i BH).

Normalfordeling Den anden antagelse er at den marginale fordeling af Y_i er en normalfordeling for alle $i = 1, \dots, n$, således at vi kun mangler at angive de mulige middelværdier og varianser.

Varianshomogenitet Den tredje antagelse er at alle Y_i har samme varians. Dette kaldes varianshomogenitet.

Så er der kun middelværdierne tilbage at lege med. Vi starter med den simpleste situation i kapitel 3 og 4 hvor antagelsen er at alle observationer har samme middelværdi og dermed samme fordeling. Vi taler om en enkelt stikprøve. I kapitel 3 antager vi

desuden at variansen er kendt. Dette er som regel urealistisk, men de forskellige begreber kan med fordel introduceres i denne ramme fordi modellen matematisk set er nem at gå til. I kapitel 4 diskuterer vi tilfældet hvor både middelværdi og varians er ukendte.

I kapitel 5 fortsætter vi med to stikprøver hvor antagelsen er at observationerne stammer fra to forskellige normalfordelinger svarende til en opdeling af observationerne i to forskellige grupper. Det kunne for eksempel være opdeling efter køn, efter aktietype, eller efter behandlingstype. Hovedformålet med en sådan analyse er ofte at undersøge om der er forskel på de to grupper i den forstand at de to normalfordelingers middelværdier er forskellige, og at kvantificere en eventuel forskel.

Endelig handler kapitel 6 om lineær regression. Her antages det at der til hver observation y_i er knyttet et tal x_i , og at middelværdien i normalfordelingen svarende til y_i afhænger lineært af x_i . Som regel er man interesseret i sammenhængen mellem x og y .

I dette kursus vil vi kun beskæftige os med disse tre specifikke tilfælde, men I vil møde en mere generel formulering i senere kurser.

Umiddelbart kan de tre antagelser om uafhængighed, normalfordeling og varianshomogenitet lyde restriktive. Det er de også, men de giver alligevel anledning til en meget nyttig klasse af modeller som har en enorm udbredelse. Det er der forskellige grunde til. Dels viser det sig at forbavsende mange data med rimelighed kan beskrives ved hjælp af normalfordelingen. Dels er det typisk middelværdistrukturen der er af interesse, og på det punkt er der stadig stor frihed. Endelig har normalfordelingen pæne matematiske/sandsynlighedsteoretiske egenskaber således at vi får pæne og eksakte fordelingsresultater for estimatorer og teststørrelser.

På den anden side er det vigtigt at understrege at modellerne ikke kan klare alt. De forskellige resultater vedrørende estimation, konfidensintervaller og hypotesetest gælder hvis Y_i 'erne opfylder modelantagelserne. Men hvis antagelserne ikke er opfyldt, ved vi ikke hvad der sker, og så kan vi ikke stole på resultaterne af den statistiske analyse. Det er derfor essentielt at undersøge om antagelserne er rimelige hver gang man udfører statistiske analyser.

Vi vil diskutere antagelser og modelkontrol i eksemplerne undervejs, men lad os komme med nogle generelle betragtninger allerede nu. Uafhængighedsantagelsen er ofte rimelig hvis observationerne stammer fra forskellige individer, men næppe rimelig hvis der er flere observationer fra samme individ, hvis nogle af individerne er i familie med hinanden, eller hvis observationerne er målinger af den samme størrelse over en årrække. Antagelsen om ens varians er heller ikke altid rimelig. Det er for eksem-

pel ret almindeligt at variansen er større for observationer med store middelværdier end for observationer med små middelværdier. Endelig er det naturligvis ikke alle data der med rimelighed kan beskrives ved hjælp af normalfordelingen.

Nogle gange kan problemer med varianshomogenitet og normalfordelingsantagelsen afhjælpes ved at transformere observationerne og analysere de transformerede data i stedet for de oprindelige, dvs. analysere $f(y_1), \dots, f(y_n)$ for en passende funktion f . Dette illustreres med data i eksempler og opgaver i det følgende.

Kapitel 3

En stikprøve med kendt varians

I dette kapitel skal vi betragte situationen med en enkelt normalfordelt stikprøve eller observationsrække og yderligere antage at den fælles varians er kendt. Det er kun rimeligt i få situationer — som regel vil vi bruge data til at estimere variansen som i kapitel 4 — men der er en pædagogisk pointe i at gå grundigt til værks. Sagen er at vi nemt kan vise forskellige egenskaber i denne model, og derfor kan koncentrere os om at forstå de forskellige begreber og meningen med dem. Dette vil komme os til gavn i de senere kapitler hvor strukturen af modellerne bliver lidt mere kompliceret.

3.1 Statistisk model

Lad os starte med et eksempel.

Eksempel 3.1. (*Kobbertråd*) Til kontrol af en løbende produktion af kobbertråd udtages med passende mellemrum ni stykker tråd af ens længde. De ni stykker tråd vejes, og erfaringerne viser at man kan antage at vægten er normalfordelt med en varians på $\sigma^2 = 0.000074 \text{ g}^2$, dvs. en spredning på $\sigma = 0.0086 \text{ g}$. En stikprøve gav følgende vægte (også i gram):

18.459	18.461	18.452
18.434	18.453	18.436
18.449	18.447	18.443

Vi antager at de ni målinger y_1, \dots, y_9 er realisationer af stokastiske variable Y_1, \dots, Y_9

der er uafhængige og normalfordelte med en ukendt middelværdi (som vi er interesseret i) og en varians på 0.000074 g^2 .

Man tilstræber en produktionsstandard svarende til at den gennemsnitlige vægt af trådstykkerne i produktionen er 18.441 g , og vi skal i det følgende beskrive en metode til at undersøge hvorvidt data er i modstrid med dette mål. \square

Udgangspunktet er at vi antager at de stokastiske variable Y_1, \dots, Y_n er uafhængige og allesammen $N(\mu, \sigma_0^2)$ -fordelte. Variansen er et kendt tal — vi har understreget dette ved at betegne den σ_0^2 — mens middelværdien μ ikke er kendt. Middelværdien er med andre ord en parameter i modellen, ganske som sandsynligheden p er en parameter i binomialfordelingsmodellen givet i definition 1.1.

Den simultane tæthed for (Y_1, \dots, Y_n) er lig produktet af de marginale tætheder (sætning A.1 i appendiks A). Når vi indsætter tætheden for den relevante normalfordeling (BH, side 216), får vi derfor den simultane tæthed

$$\begin{aligned} f_{\mu}(y) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(y_i - \mu)^2\right) \\ &= \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (y_i - \mu)^2\right), \quad y = (y_1, \dots, y_n) \in \mathbb{R}^n. \end{aligned} \quad (3.1)$$

Hvis vi lader N_{μ}^n betegne fordelingen på \mathbb{R}^n med denne tæthed, kan vi definere den statistiske model som mængden af sådanne fordelinger hvor μ varierer i en parametermængde $\Theta \subseteq \mathbb{R}$. Vi vil antage $\mu \in \mathbb{R}$, altså $\Theta = \mathbb{R}$, men Θ kunne også være en ægte delmængde af \mathbb{R} .

Definition 3.2. *Modellen for en enkelt stikprøve med kendt varians består af udfaldsrummet \mathbb{R}^n samt familien*

$$\mathcal{P} = \{N_{\mu}^n : \mu \in \mathbb{R}\}$$

af fordelinger på \mathbb{R}^n hvor N_{μ}^n har tæthed (3.1) for et givet $\sigma_0^2 > 0$.

Alternativ formulering: Lad Y_1, \dots, Y_n være uafhængige og identisk normalfordelte stokastiske variable, $Y_i \sim N(\mu, \sigma_0^2)$ hvor $\sigma_0^2 > 0$ er kendt mens $\mu \in \mathbb{R}$ er ukendt.

Ganske som i binomialtilfældet afspejler den statistiske model vores viden og uvidenhed om de mekanismer der har frembragt data.

- Vores antagelser om uafhængighed og marginale normalfordelinger formaliserer vores forhåndsviden eller forhåndsantagelser. Det skal kontrolleres om

disse antagelser er opfyldt — eller rettere om de giver en rimelig beskrivelse af usikkerheden i data.

- Den enkelte normalfordeling, $N(\mu, \sigma_0^2)$, beskriver usikkerheden der er forbundet med dataindsamlingen hvis μ er den sande parameter.
- De forskellige mulige værdier af μ formaliserer vores uvidenhed om hvilken normalfordeling der har frembragt data.

3.2 Maksimum likelihood estimation

Tætheden $f_\mu(y)$ fra (3.1) angiver sandsynlighedsmassen per volumenenhed omkring punktet $y \in \mathbb{R}^n$. Når vi laver sandsynlighedsregning tænker vi altså på $f_\mu(y)$ som udtryk for hvor sandsynligt det er at få data “i nærheden af” $y = (y_1, \dots, y_n)$ når vi ved at middelværdien er μ . Når vi laver statistik er situationen den modsatte: vi *har* data y og *antager* at de stammer fra uafhængige $N(\mu, \sigma_0^2)$ -fordelte variable, men vi *kender ikke* μ . Vi skal bruge vores observationer til at *estimere* μ .

Husk at vi for binomialfordelingen lavede maksimum likelihood estimation og estimerede sandsynlighedsparameteren med den værdi der gjorde vores observation mest sandsynlig. Alle udfald i normalfordelingen har sandsynlighed nul fordi det er en kontinuert fordeling, så vi kan ikke gøre helt det samme. På den anden side udtrykker tætheden noget lignende, og maksimum likelihood estimation går ud på at estimere μ med den værdi der maksimerer tætheden $f_\mu(y)$. Vi vil stadig tænke på estimatet som den værdi af μ der gør de observerede værdier mest sandsynlige, selvom vi skal huske at tænke på sandsynligheder for områder snarere end punktsandsynligheder. På engelsk ville man tale om “the likelihood of the data” eller om “how likely the data is” — vi mangler tilsvarende formuleringer på dansk.

Formelt set definerer vi *likelihoodfunktionen* som tætheden, nu opfattet som funktion af μ for fast $y \in \mathbb{R}$ snarere end omvendt, og søger en værdi $\hat{\mu}$ der gør funktionen størst mulig. Likelihoodfunktionen hørende til observationen $y = (y_1, \dots, y_n) \in \mathbb{R}$ defineres derfor ved

$$L_y : \mathbb{R} \rightarrow \mathbb{R}$$

$$L_y(\mu) = f_\mu(y) = \frac{1}{(2\pi\sigma_0^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (y_i - \mu)^2\right) \quad (3.2)$$

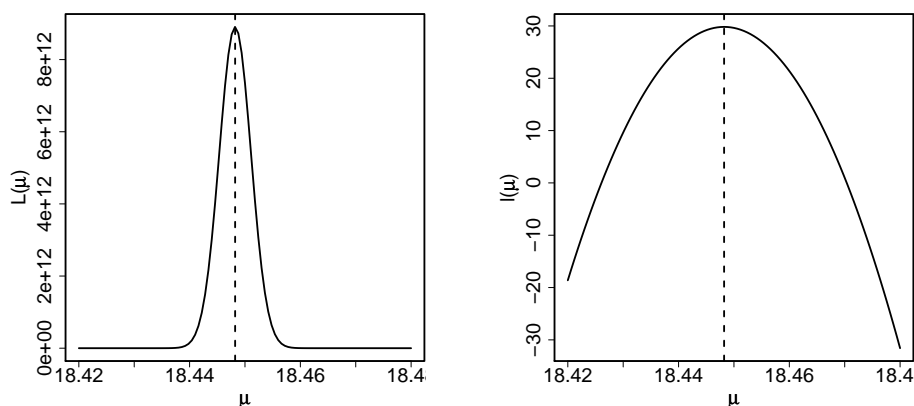
og et maksimum likelihood estimat $\hat{\mu} \in \mathbb{R}$ opfylder

$$L_y(\hat{\mu}) \geq L_y(\mu), \quad \mu \in \mathbb{R}. \quad (3.3)$$

Det er klart fra strukturen af L_y at det er mere hensigtsmæssigt at arbejde med logaritmen til likelihoodfunktionen, også kaldet log-likelihoodfunktionen. Det skyldes at likelihoodfunktionen er defineret som et produkt af tætheder, som så bliver til en sum af log-tætheder. Vi vil sommetider bruge betegnelsen l for log-likelihoodfunktionen, dvs.

$$l_y(\mu) = \log L_y(\mu) = -\frac{n}{2} \log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (y_i - \mu)^2.$$

Da logaritmen er en strengt voksende funktion kan vi erstatte L_y med l_y i (3.3). Figur 3.1 viser likelihoodfunktionen og log-likelihoodfunktionen for de ni observationer af kobbertrådsvægte (eksempel 3.1, side 31).



Figur 3.1: Likelihoodfunktionen (til venstre) og log-likelihoodfunktionen (til højre) for data fra eksempel 3.1. Den stiplede linje svarer til gennemsnittet $\bar{y} = 18.44822g$.

Sætning 3.3. For den statistiske model fra definition 3.2 er maksimum likelihood estimatet for μ entydigt bestemt og givet ved $\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Estimatoren $\hat{\mu} = \bar{Y}$ er normalfordelt med middelværdi μ og varians σ_0^2/n .

Bevis Hvis vi differentierer log-likelihoodfunktionen med hensyn til μ får vi

$$l'_y(\mu) = \frac{1}{\sigma_0^2} \sum_{i=1}^n (y_i - \mu)$$

$$l''_y(\mu) = -\frac{n}{\sigma_0^2} < 0.$$

Vi ser at $l'_y(\mu) = 0$ hvis og kun hvis $\sum_{i=1}^n y_i = n\mu$, altså hvis og kun hvis $\mu = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}$, så \bar{y} er det eneste stationære punkt for l_y . Desuden er $l''_y(\bar{y}) < 0$ så l_y

har maksimum i \bar{y} som ønsket. Fordelingsresultatet om $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ følger direkte af sætning A.5 i appendiks A.

Estimatet for middelværdien er altså blot gennemsnittet af observationerne. Det kan næppe siges at være ret overraskende. Estimatet \bar{y} er et tal, mens estimatoren \bar{Y} er en stokastisk variabel. Estimatet er en realisation af estimatoren. Bemærk at vi ofte bruger samme notation, nemlig $\hat{\mu}$, for begge dele. Hvis vi ønsker at fremhæve at de er funktioner af y_1, \dots, y_n henholdsvis Y_1, \dots, Y_n , kan vi skrive $\hat{\mu} = \hat{\mu}(y_1, \dots, y_n) = \bar{y}$ for estimatet og $\hat{\mu} = \hat{\mu}(Y_1, \dots, Y_n) = \bar{Y}$ for estimatoren.

Maksimum likelihood estimatoren \bar{Y} er en stokastisk variabel, og som angivet i sætningen har vi $\bar{Y} \sim N(\mu, \sigma_0^2/n)$. Specielt har vi altså

$$E(\hat{\mu}) = \mu, \quad \text{Var}(\hat{\mu}) = \frac{\sigma_0^2}{n}, \quad \text{SD}(\hat{\mu}) = \frac{\sigma_0}{\sqrt{n}} \quad (3.4)$$

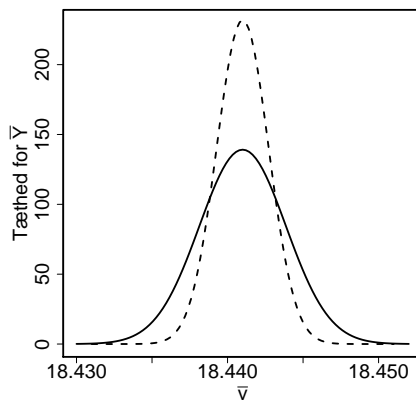
hvor vi bruger notationen SD for spredning (standard deviation). Bemærk specielt at $\hat{\mu} = \bar{Y}$ er en central estimator for μ fordi middelværdien er den sande værdi.

Fordelingen af $\hat{\mu} = \bar{Y}$ udtrykker den usikkerhed der er forbundet med estimatet. For at forstå hvad det betyder, kan det være hensigtsmæssigt at forestille sig forsøget gentaget mange gange (for eksempel måling af ni stykker kobbertråd). For hver dataindsamling får vi et nyt gennemsnit \bar{y} , og tætheden for $N(\mu, \sigma_0^2/n)$ fortæller os hvilke gennemsnit der er sandsynlige at observere. Specielt udtrykker (3.4) at vi i gennemsnit — over mange dataindsamlinger — vil få den sande værdi, og at flere observationer i stikprøven giver anledning til større præcision. Dette er illustreret i Figur 3.2 hvor tætheden for \bar{Y} 's fordeling er tegnet for $\mu = 18.441$ og $\sigma_0^2 = 0.000074$. Antallet af observationer er $n = 9$ for den fuldt optrukne kurve og $n = 25$ for den stiplede kurve. Værdier langt fra 18.441 er tydeligvis mindre sandsynlige når $n = 25$ sammenlignet med når $n = 9$.

Fordelingen af $\hat{\mu} = \bar{Y}$ er $N(\mu, \sigma_0^2/n)$, men husk at middelværdien μ er ukendt, uanset at vi har et estimat for den. Vi taler sommetider om fordelingen som den “sande” eller den “teoretiske” fordeling.

Eksempel 3.4. (*Kobbertråd, fortsættelse af eksempel 3.1, side 31*) Gennemsnittet for de ni observerede vægte af kobbertrådsstykker er $\bar{y} = 18.44822$, så $\hat{\mu} = 18.44822$. Dette er en realisation af \bar{Y} hvis teoretiske eller sande fordeling er $N(\mu, 0.000074/9)$. Specielt er spredningen i lig fordelingen $\text{SD}(\hat{\mu}) = 0.002867$. \square

Vi fandt maksimum likelihood estimatet ved at maksimere likelihoodfunktionen. Fra udtrykket (3.2) for likelihoodfunktionen kan vi se at dette er ækvivalent med at mini-



Figur 3.2: Tætheden for $N(18.441, 0.000074/n)$ for $n = 9$ (fuldt optrukket) og $n = 25$ (stiplet).

mere

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

Derfor er $\hat{\mu} = \bar{y}$ den værdi der gør summen af de kvadrerede afstande fra observationerne til middelværdien mindst mulig. Vi taler om “mindste kvadraters metode” eller “least squares method”, og i dette tilfælde giver mindste kvadraters metode og maksimum likelihood estimation det samme estimat.

3.3 Konfidensinterval for middelværdien

Hvis vi gentog dataindsamlingen ville vi få nogle andre observationer og dermed en anden værdi af \bar{y} , så hvor meget kan vi stole på vores estimat? Fordelingen af $\hat{\mu} = \bar{Y}$ beskriver netop denne usikkerhed, men man opsummerer ofte usikkerheden i et *konfidensinterval*.

Et $1 - \alpha$ konfidensinterval for μ er et interval $(L(Y), U(Y))$ som indeholder den sande værdi med sandsynlighed mindst $1 - \alpha$:

$$P(\mu \in (L(Y), U(Y))) \geq 1 - \alpha.$$

I de modeller vi skal se på, kan vi endda opnå lighedstegn i stedet for ulighedstegn. Man bruger ofte 95% konfidensintervaller svarende til $\alpha = 0.05$, men 90% og 99%

konfidensintervaller rapporteres også af og til. Bogstaverne L og U står for “lower” og “upper”, og med notationen $L(Y)$ og $U(Y)$ understreger vi at endepunkterne i konfidensintervallet er stokastiske variable, afledt af $Y = (Y_1, \dots, Y_n)$. For en given observation indsætter vi y og får det observerede konfidensinterval $(L(y), U(y))$.

Spørgsmålet er hvordan vi skal vælge intervalendepunkterne $L(Y)$ og $U(Y)$. Husk at $\bar{Y} \sim N(\mu, \sigma_0^2/n)$ således at

$$\frac{\bar{Y} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1).$$

Lad $z_{1-\alpha/2}$ betegne $1 - \alpha/2$ fraktilen i $N(0, 1)$. Der er sandsynlighedsmasse $\alpha/2$ til venstre for $-z_{1-\alpha/2}$ og sandsynlighedsmasse $\alpha/2$ til højre for $z_{1-\alpha/2}$, så

$$\begin{aligned} 1 - \alpha &= P\left(-z_{1-\alpha/2} < \frac{\bar{Y} - \mu}{\sigma_0/\sqrt{n}} < z_{1-\alpha/2}\right) \\ &= P\left(\mu - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} < \bar{Y} < \mu + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}\right). \end{aligned}$$

Hvis vi omroterer leddene så den sande værdi μ optræder “i midten”, får vi i stedet

$$P\left(\bar{Y} - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} < \mu < \bar{Y} + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha. \quad (3.5)$$

Dette svarer til at vælge

$$L(Y) = \bar{Y} - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}; \quad U(Y) = \bar{Y} + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}.$$

Vi har således vist følgende sætning.

Sætning 3.5. *Betragt den statistiske model fra definition 3.2. Så er*

$$\bar{Y} \pm z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} = \left(\bar{Y} - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}, \bar{Y} + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}\right) \quad (3.6)$$

et $1 - \alpha$ konfidensinterval for μ .

Husk fra (3.4) at spredningen for $\hat{\mu} = \bar{Y}$ er σ_0/\sqrt{n} . Således har konfidensintervallet formen

$$\hat{\mu} \pm \text{fraktil} \cdot \text{spredning for } \hat{\mu}. \quad (3.7)$$

Specielt er konfidensintervallet symmetrisk om $\hat{\mu}$. Dette synes at være mest naturligt, men man kan godt konstruere konfidensintervaller uden symmetriegenskaben.

For et datasæt bestående af observationerne y_1, \dots, y_n erstattes den stokastiske variabel \bar{Y} af det observerede gennemsnit \bar{y} . For eksempel beregnes 95% konfidensintervallet som

$$\bar{y} \pm 1.96 \frac{\sigma_0}{\sqrt{n}} \quad (3.8)$$

da $z_{1-0.05/2}$ er lig 97.5% fraktilen i $N(0, 1)$, dvs. 1.96.

Eksempel 3.6. (*Kobbertråd, fortsættelse af eksempel 3.1, side 31*) Husk estimerterne $\bar{y} = 18.44822$ og $\sigma_0^2 = 0.000074$. Vi beregner således et 95% konfidensinterval for μ til

$$18.44822 \pm 1.96 \sqrt{\frac{0.000074}{9}} = 18.44822 \pm 0.00562 = (18.44260, 18.45384).$$

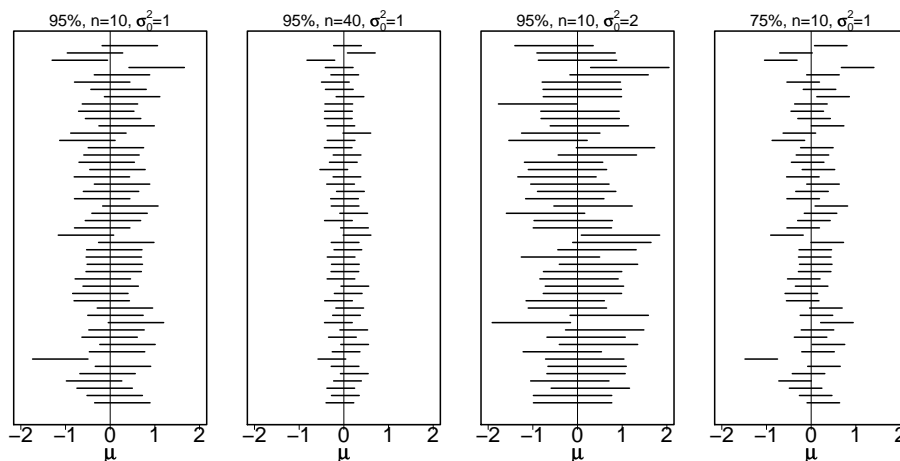
Bemærk at konfidensintervallet ikke indeholder værdien 18.441 som var den ønskede gennemsnitsvægt af kobbertrådene i produktionen. \square

Det er nemt at få fortolkningen af konfidensintervaller galt i halsen. Som vi kan se af (3.5), er endepunkterne i intervallet stokastiske variable, og (3.5) er et udsagn om intervallet snarere end om μ . Det forstås nok bedst ved at tænke på gentagelser af eksperimentet: Hvis vi forestiller os at dataindsamlingen gentages mange gange (med samme μ og samme σ_0^2) og at intervallet beregnes for hvert nyt datasæt, så vil omtrent andelen $1 - \alpha$ af disse intervaller indeholde den sande værdi af μ .

Dette er illustreret i figur 3.3 for $\mu = 0$ og forskellige kombinationer af n , σ_0^2 og $1 - \alpha$. For at lave figuren til venstre har vi simuleret 50 datasæt, hver bestående af $n = 10$ uafhængige observationer fra $N(0, 1)$. Når vi simulerer data beder vi computeren trække dem tilfældigt fra en given fordeling. For hver af de 50 datasæt har vi beregnet 95% konfidensintervallet (3.8) og tegnet det som en vandret streg i figuren. Den lodrette streg viser den sande værdi, $\mu = 0$. Vi kan se at nul ligger i alle konfidensintervallerne på nær tre. Dette svarer nogenlunde til 95%.

Konfidensintervallet afhænger af variansen σ_0^2 , antallet af observationer n og graden af konfidens, $1 - \alpha$. Det ses nemt fra (3.6) hvad der sker hvis vi varierer på disse størrelser:

- Hvis n vokser bliver konfidensintervallet smallere. Dette giver god mening: jo flere observationer, jo mere præcist er estimatet bestemt, og et smallere interval giver os samme grad af konfidens. Dette er illustreret i plot 2 fra venstre i figur 3.3 hvor $n = 40$, mens σ_0^2 og $1 - \alpha$ er som i plottet yderst til venstre. Intervallerne til højre er som ventet smallere end til venstre.



Figur 3.3: Konfidensintervaller for simulerede datasæt for forskellige værdier af n , σ_0^2 og $1 - \alpha$.

- Hvis σ_0^2 vokser bliver konfidensintervallet bredere. Dette giver også god mening: stor variation på de enkelte observationer giver stor variation på gennemsnittet og dermed et mindre præcist estimat, således at et bredere interval er nødvendigt for at fastholde graden af konfidens. Dette er illustreret i plot 3 fra venstre i figur 3.3 hvor $\sigma_0^2 = 2$ mens n og $1 - \alpha$ er uændret i forhold til plottet længst til venstre. Konfidensintervallerne er tydeligvis blevet bredere.
- Hvis vi ønsker et større $1 - \alpha$ (dvs. et mindre α) så vokser fraktilen $z_{1-\alpha/2}$ og konfidensintervallet bliver bredere: en høj grad af konfidens kræver et bredt interval. Dette illustreres ved sammenligning af venstre og højre plot i figur 3.3 hvor konfidensgraden er henholdsvis 95% og 75%. Konfidensintervallerne er bredest til venstre. For $\alpha = 0.25$, svarende til konfidensgrad 75%, skal vi bruge 87.5% fraktilen i $N(0, 1)$, som er 1.15, og den sande værdi er indeholdt i 41 af de 50 konfidensintervaller (82%) til højre. Hvis vi foretog øvelsen med et større antal gentagelser ville vi komme tættere på 75%.

Set fra et praktisk synspunkt er tankegangen omkring gentagelser problematisk: vi har jo kun et enkelt datasæt til rådighed og kan kun beregne et enkelt konfidensinterval. Enten ligger μ i intervallet eller også ligger μ ikke i intervallet, men vi ved det ikke. Alligevel kan vi bruge konfidensintervallet som indikation af hvilke værdier af μ der med rimelighed kan antages at være sande. Hvis den sande middelværdi er μ_0 og

$\alpha = 0.05$, så gælder:

- sandsynligheden for at observere data y som opfylder at μ_0 ligger i det tilhørende konfidensinterval er 95%
- sandsynligheden for at observere data y som opfylder at μ_0 ikke ligger i det tilhørende konfidensinterval er 5%

Hvis den sande værdi er μ_0 er det altså ret usædvanligt at observere et konfidensinterval der ikke indeholder μ_0 . I eksempel 3.6 (side 38) konstaterede vi at værdien 18.441 ikke var indeholdt i 95% konfidensintervallet. Hvis den sande middelværdi faktisk er 18.441, er de observerede data altså temmelig usædvanlige. Vi skal bygge videre på denne tankegang i næste afsnit om hypotesetest.

For at konstruere konfidensintervallet benyttede vi (3.5). Formlen giver os en egenkab ved fordelingen af \bar{Y} , nemlig et interval som \bar{Y} rammer med sandsynlighed 95%. Dette er bare ét aspekt af \bar{Y} 's fordeling. Sagt på en anden måde: konfidensintervallet opsummerer kun visse aspekter af den usikkerhed der er forbundet med estimatet — selve fordelingen indeholder mere information. Alligevel benyttes konfidensintervallet ofte til at opsummere usikkerheden fordi det er simple end en beskrivelse af hele fordelingen, samtidig med at det i ret høj grad giver os den relevante information. Blot skal vi huske at tænke os grundigt om når vi fortolker konfidensintervallet.

3.4 Test af hypotese om middelværdien

Sommetider er man interesseret i at undersøge om middelværdien i fordelingen af Y 'erne med rimelighed kan antages at have en bestemt værdi — måske er det endda derfor man har indsamlet data. Vi betragter et fast tal, $\mu_0 \in \Theta = \mathbb{R}$ og tester hypotesen om at middelværdien af Y_1, \dots, Y_n netop er μ_0 . Løst sagt betyder det at vi undersøger om data er i modstrid med hypotesen eller ej, dvs. om data med rimelighed kan tænkes at være fremkommet hvis hypotesen er sand. Som regel betegner vi hypotesen H og skriver

$$H : \mu = \mu_0. \quad (3.9)$$

Eksempel 3.7. (*Kobbertråd, fortsættelse af eksempel 3.1, side 31*) Man ønsker at den gennemsnitlige vægt af kobbertråde i produktionen er 18.441 g. For at undersøge om dette kan antages at være tilfældet har man udtaget stikprøven bestående af n kobbertråde. Den relevante hypotese er således $H : \mu = 18.441$, og spørgsmålet er om stikprøven tyder på at populationsgennemsnittet afviger fra 18.441 g. \square

Mere generelt defineres en hypotese ved at lægge restriktioner på parameteren (eller parametrene), og kræve at den ligger i en delmængde Θ_0 af den oprindelige parametermængde Θ . Således kan vi skrive $H : \mu \in \Theta_0$. Hypotesen (3.9) svarer til at vælge $\Theta_0 = \{\mu_0\}$, og vi siger at hypotesen er simpel fordi parametermængden under hypotesen kun indeholder et enkelt punkt. I dette kapitel vil vi kun betragte den simple hypotese (3.9).

Hypotesetest handler om at afgøre hvorvidt den afvigelse fra hypotesen som data udviser, er et udtryk for at hypotesen faktisk er falsk eller om den lige så godt kan skyldes tilfældig variation. Ideen er at spørge: *Hvis* hypotesen er sand, hvor sandsynligt er det så at observere de data som vi faktisk observerede, eller nogle der passer endnu dårligere med hypotesen? Dette skal selvfølgelig præciseres nærmere: hvad betyder det at nogle data “passer dårligere med hypotesen” end andre?

I vores situation med en enkelt stikprøve er svaret intuitivt ret klart: data passer godt med hypotesen hvis \bar{y} ligger tæt på μ_0 , så vi kan måle hvor godt hypotesen passer til data ved hjælp af afstanden $|\bar{y} - \mu_0|$. Det er da også præcis det vi vil gøre, men vi vil gå en lille omvej og introducere et generelt testprincip, nemlig *kvotienttestet* eller, på engelsk, *likelihood ratio testet*.

Likelihoodfunktionen $L_y(\mu)$ udtrykker hvor sandsynligt det er at observere y når middelværdien er μ . Specielt er $L_y(\mu_0)$ et udtryk for hvor sandsynligt det er at observere y under hypotesen (3.9), og $L_y(\hat{\mu})$ er et udtryk for hvor sandsynligt det er at observere y i modellen uden den ekstra restriktion givet ved hypotesen.

Således giver det mening at fortolke *kvotientteststørrelsen* (engelsk: the likelihood ratio test statistic)

$$Q(y) = \frac{L_y(\mu_0)}{L_y(\hat{\mu})}$$

som mål for hvor meget dårligere hypotesen $\mu = \mu_0$ passer til data end den oprindelige model $\mu \in \Theta$. Estimatet $\hat{\mu} \in \mathbb{R}$ er valgt så L_y er størst mulig, specielt gælder $L_y(\hat{\mu}) \geq L_y(\mu_0)$. Således er $Q(y) \in (0, 1]$. Store og små værdier af $Q(y)$ fortolkes på følgende måde:

- Hvis $Q(y)$ er lille (tæt på nul) er det langt mindre sandsynligt at observere y under hypotesen end i den oprindelige model. Dette tyder på at hypotesen er falsk, og vi siger at små værdier af Q er *kritiske* for hypotesen.
- Hvis $Q(y)$ er stor (tæt på en) er det næsten lige så sandsynligt at observere y under hypotesen som i den oprindelige model. Dette tyder på at hypotesen er sand — i hvert fald tyder det ikke på at hypotesen er falsk.

Vi kan med andre ord bruge $Q(y)$ til at måle hvor godt hypotesen passer til data, selvom det stadig er uklart hvad “lille” og “stor” betyder i ovenstående udsagn. Det kommer vi tilbage til om lidt. I Sætning 3.8 nedenfor viser vi at

$$Q(y) = \exp\left(-\frac{1}{2\sigma_0^2}n(\bar{y} - \mu_0)^2\right).$$

Fortolkningerne ovenfor kan derfor oversættes til følgende: hvis \bar{y} og μ_0 ligger langt fra hinanden, så tyder det på at hypotesen er falsk, mens det tyder på at hypotesen er sand hvis \bar{y} og μ_0 ligger tæt på hinanden. Det giver jo god mening!

Værdien $Q(y)$ er en realisation af den stokastiske variabel

$$Q(Y) = \frac{L_Y(\mu_0)}{L_Y(\hat{\mu})} = \exp\left(-\frac{1}{2\sigma_0^2}n(\bar{Y} - \mu_0)^2\right), \quad (3.10)$$

som er en transformation af de oprindelige stokastiske variable Y_1, \dots, Y_n . p -værdien eller *testsandsynligheden* for hypotesen $H : \mu = \mu_0$ defineres som sandsynligheden for — givet at hypotesen er sand — at observere en værdi af $Q(Y)$ der passer lige så dårligt eller dårligere med hypotesen end værdien $Q(y)$ som vi faktisk observerede:

$$\varepsilon(y) = P(Q(Y) \leq Q(y)).$$

For at beregne p -værdien har vi brug for at kende fordelingen af $Q(Y)$ under hypotesen. Eftersom vi kender fordelingen af \bar{Y} , kunne vi i princippet finde tætheden af $Q(Y)$ ved hjælp af transformationsætningen (BH, sætning 8.1.1), men vi kan gøre livet lidt nemmere for os selv. Ved at kaste et blik på udtrykket (3.10) bliver det klart at det er hensigtsmæssigt at betragte

$$U = \frac{\bar{Y} - \mu_0}{\sigma_0/\sqrt{n}}; \quad u = \frac{\bar{y} - \mu_0}{\sigma_0/\sqrt{n}}.$$

Her er u en observeret værdi og $Q(y) = \exp(-\frac{1}{2}u^2)$, mens U er en stokastisk variabel og $Q(Y) = \exp(-\frac{1}{2}U^2)$. Da funktionen der fører u over i $Q(y)$ (eller U over i $Q(Y)$) er aftagende, får vi

$$\varepsilon(y) = P(U^2 \geq u^2) = P(|U| \geq |u|).$$

Det er således nok at kende fordelingen af U under hypotesen. Her er vi på sikker grund: under hypotesen er $\bar{Y} \sim N(\mu_0, \sigma_0^2/n)$ så $U \sim N(0, 1)$ og $U^2 \sim \chi_1^2$ (BH, definition 10.4.1). Vi siger at vi udfører testet på u eller at vi udfører et u -test.

Lad os samle resultaterne i en sætning og vise den formelt.

Sætning 3.8. *Betragt den statistiske model givet i definition 3.2 og hypotesen $H : \mu = \mu_0$ for et fast $\mu_0 \in \mathbb{R}$. Kvotientteststørrelsen er givet ved*

$$Q(y) = \exp\left(-\frac{1}{2\sigma_0^2}n(\bar{y} - \mu_0)^2\right),$$

og vi kan udføre testet på

$$u = \frac{\bar{y} - \mu_0}{\sigma_0/\sqrt{n}}.$$

p -værdien er givet ved $\varepsilon(y) = 2(1 - \Phi(|u|))$ hvor Φ er fordelingsfunktionen for standardnormalfordelingen, $N(0, 1)$.

Bevis Hvis vi indsætter μ_0 og $\hat{\mu} = \bar{y}$ i (3.1) og bemærker at normeringskonstanten forkorter ud, så får vi

$$\begin{aligned} Q(y) &= \frac{L_y(\mu_0)}{L_y(\hat{\mu})} = \frac{\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(y_i - \mu_0)^2\right)}{\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(y_i - \bar{y})^2\right)} \\ &= \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n((y_i - \mu_0)^2 - (y_i - \bar{y})^2)\right) \\ &= \exp\left(-\frac{1}{2\sigma_0^2}n(\bar{y} - \mu_0)^2\right) \end{aligned}$$

hvor sidste lighedstegn følger ved at bruge formlen for kvadratet på en toleddet størrelse:

$$\begin{aligned} \sum_{i=1}^n((y_i - \mu_0)^2 - (y_i - \bar{y})^2) &= \sum_{i=1}^n(y_i^2 + \mu_0^2 - 2\mu_0 y_i - y_i^2 - \bar{y}^2 + 2\bar{y}y_i) \\ &= \sum_{i=1}^n(\mu_0^2 - 2\mu_0 y_i - \bar{y}^2 + 2\bar{y}y_i) \\ &= n\mu_0^2 - 2n\mu_0\bar{y} - n\bar{y}^2 + 2n\bar{y}^2 \\ &= n(\bar{y} - \mu_0)^2. \end{aligned}$$

p -værdien er

$$\varepsilon(y) = P(Q(Y) \leq Q(y)) = P\left(\frac{n}{\sigma_0^2}(\bar{Y} - \mu_0)^2 \geq \frac{n}{\sigma_0^2}(\bar{y} - \mu_0)^2\right) = P(U^2 \geq u^2)$$

hvor $U \sim N(0, 1)$. Således får vi

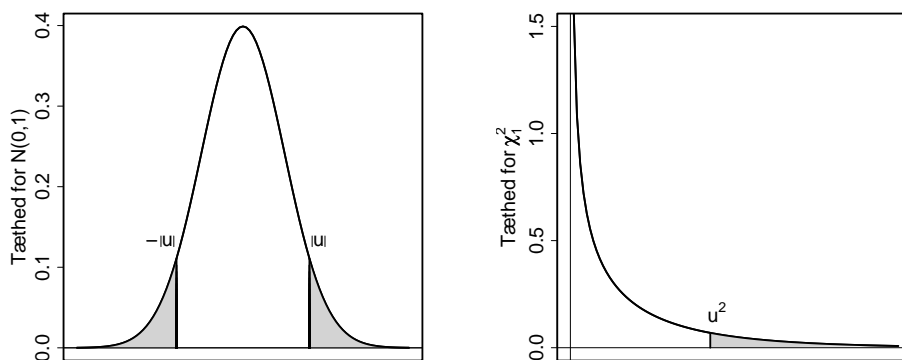
$$\varepsilon(y) = 2P(U \geq |u|) = 2(1 - P(U \leq |u|)) = 2(1 - \Phi(|u|)),$$

og vi har vist det ønskede. \square

Bemærk at $U^2 \sim \chi_1^2$ så vi kan også beregne p -værdien som en sandsynlighed i χ_1^2 -fordelingen:

$$\varepsilon(y) = P(U^2 \geq u^2) = 1 - F_{\chi_1^2}(u^2),$$

hvor $F_{\chi_1^2}$ er fordelingsfunktionen for χ_1^2 -fordelingen. p -værdien er illustreret i figur 3.4 som arealet af de grå områder. Den venstre del af figuren viser tætheden for $N(0, 1)$ sammen med en fiktiv værdi af $|u|$ og $-|u|$. Den højre del af figuren viser tætheden for χ_1^2 -fordelingen sammen med den fiktive værdi af u^2 . Det samlede areal af de to grå områder i venstre del af figuren er det samme som arealet af det grå område i højre del — nemlig $\varepsilon(y)$ — selvom det på grund af skalering af figurerne er svært at se.



Figur 3.4: Tætheden for $N(0, 1)$ til venstre sammen med værdier af $\pm|u|$ (til venstre) og tætheden for χ_1^2 sammen med u^2 (til højre). De grå områder har areal lig $\varepsilon(y)$.

Vi mangler stadig at afgøre hvorvidt hypotesen skal afvises eller ej. p -værdien $\varepsilon(y)$ måler hvor sandsynligt det er — hvis hypotesen er sand — at få data der passer lige så dårligt eller dårligere med hypotesen end de observerede data y , målt ved $Q(y)$ eller u . Små værdier er kritiske: en lille værdi af $\varepsilon(y)$ tyder på at hypotesen er falsk mens store værdier tyder på at hypotesen er sand. Men hvad skal vi mene med “stor” og “lille”? Inden analysen vælges et *signifikansniveau* α . Det betyder at vi vælger at forkaste eller afvise hypotesen hvis $\varepsilon(y) \leq \alpha$. Vi siger at μ er signifikant forskellig fra μ_0 på niveau α . Hvis $\varepsilon(y) > \alpha$ så kan vi ikke afvise hypotesen. Med andre ord: hypotesen afvises hvis $|u| \geq z_{1-\alpha/2}$ hvor $z_{1-\alpha/2}$ er $1 - \alpha/2$ fraktilen i $N(0, 1)$, dvs. hvis $|\bar{y} - \mu_0| \geq z_{1-\alpha/2} \cdot \sigma_0 / \sqrt{n}$. Dette giver god mening: vi afviser hypotesen hvis \bar{y} afviger meget fra μ_0 .

Ofte vælges $\alpha = 0.05$, men der er ingen dybere mening med den værdi. Man kan også vælge for eksempel 1% eller 10%, men man skal have besluttet sig inden man udfører testet.

Eksempel 3.9. (Kobbertråd, fortsættelse af eksempel 3.1, side 31) Vi har $n = 9$, $\bar{y} = 18.44822$ og $\sigma_0^2 = 0.000074$ og får derfor

$$u = \frac{\sqrt{9}(18.44822 - 18.441)}{\sqrt{0.000074}} = 2.52.$$

Ved opslag i normalfordelingen får vi $\Phi(2.52) = 0.994$ så

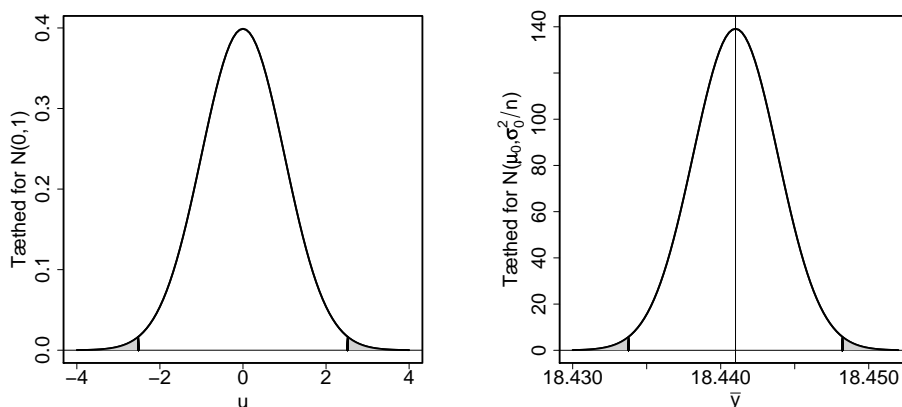
$$\varepsilon(y) = 2 \cdot (1 - 0.994) = 0.012.$$

Alternativt kunne vi slå p -værdien op i χ_1^2 -fordelingen og få $\varepsilon(y) = 1 - F_{\chi_1^2}(6.35) = 0.012$. Da p -værdien er mindre end 5% afviser vi hypotesen. Data tyder således på at den gennemsnitlige vægt af kobbertråde i produktionen afviger fra det ønskede.

Beregningen af p -værdien er illustreret i venstre del af figur 3.5. Det grå område har areal $\varepsilon(y) = 0.012$. Vi kan også skrive $\varepsilon(y)$ som sandsynligheden for at afstanden mellem \bar{Y} og $\mu_0 = 18.441$ er større end den observerede afstand:

$$\varepsilon(y) = P(|\bar{Y} - \mu_0| \geq |\bar{y} - \mu_0|)$$

hvor $\bar{Y} \sim N(\mu_0, \sigma_0^2/n)$. Dette er illustreret i den højre del af figuren. □



Figur 3.5: Tætheden for standardnormalfordelingen, $N(0, 1)$, til venstre. Tætheden for $N(18.441, 0.000074/9)$ til højre.

Nu følger nogle vigtige kommentarer omkring sprogbrugen vedrørende konklusionen på et hypotesetest, sammenhængen mellem konfidensintervaller og hypotesetest, og forskellige fejltyper:

Afvisning og accept Med et hypotesetest kan vi strengt taget kun afvise hypoteser, ikke acceptere hypoteser. Fortolkningen af en lille p -værdi er at det er usandsynligt at have observeret y (eller noget endnu værre) hvis hypotesen er sand, og den er derfor formentlig falsk. Fortolkningen af en stor p -værdi er at det er sandsynligt at observere y (eller noget endnu værre) hvis hypotesen er sand, men derfor behøver hypotesen jo ikke være sand. Der kan være mange hypoteser der gør de observerede værdier sandsynlige. Hvis man er meget nøjeregnende, bruger man derfor som regel en formulering som 'hypotesen kan ikke afvises' snarere end 'hypotesen kan accepteres'.

Angivelse af p -værdi Man bør altid angive den observerede p -værdi i stedet for blot at angive hvorvidt hypotesen kan afvises eller ej: En meget lille p -værdi (for eksempel 0.001) er udtryk for en kraftigere evidens mod hypotesen end en p -værdi tæt på α (for eksempel 0.04), og to tætte p -værdier på hver sin side af α (for eksempel 0.04 og 0.06) er udtryk for cirka samme grad af modstrid med hypotesen.

Konfidensinterval og hypotesetest Det er ikke nogen tilfældighed at vi har benyttet notationen α om signifikansniveauet og $1 - \alpha$ om konfidensgraden i et konfidensinterval. Tværtimod er der en tæt sammenhæng mellem konfidensintervaller og hypotesetest: $1 - \alpha$ konfidensintervallet for μ består netop af de værdier μ_0 for hvilke hypotesen $H : \mu = \mu_0$ ikke kan afvises på signifikansniveau α . Dette vigtige resultat er vist i sætning 3.10 nedenfor.

Type I og type II fejl Hypotesetest er baseret på sandsynligheder, og konklusionen på testet kan være forkert. Vi siger at man begår *fejl af type I* hvis man afviser en sand hypotese. Signifikansniveauet fastsætter sandsynligheden for denne type fejl. Antag igen at vi gentager eksperimentet/dataindsamlingen mange gange og for hvert datasæt udfører hypotesetestet som beskrevet. Hvis hypotesen er sand vil vi for andelen α af datasættene afvise hypotesen.

Hvis man ikke afviser (dvs. accepterer) en falsk hypotese siger vi at man har begået en fejl af type II. Vi har ikke styr på fejlraten af type II fejl på samme måde som for type I fejl, men der er selvfølgelig en sammenhæng: hvis vi sænker signifikansniveauet fra 5% til 1%, for eksempel, så gør vi det sværere at afvise hypotesen. Derfor falder sandsynligheden for type I fejl, til gengæld

vokser sandsynligheden for type II fejl. Valget af signifikansniveau repræsenterer altså en afvejning af de to fejltyper, og ved fastsættelsen af α skal man således overveje hvilken type af fejl man helst vil sikre sig mod. Dette kan være forskelligt fra anvendelse til anvendelse. Bemærk at sandsynligheden for at begå type I fejl er fastlagt ved signifikansniveauet og derfor ikke afhænger af n , mens sandsynligheden for fejl af type II falder når n vokser.

Som lovet viser vi nu sammenhængen mellem konfidensinterval og hypotesetest.

Sætning 3.10. *Betragt den statistiske model fra definition 3.2 og konfidensintervallet*

$$C_{1-\alpha}(y) = \bar{y} \pm z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} = \left(\bar{y} - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}}, \bar{y} + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \right)$$

for μ med konfidensgrad $1 - \alpha$ beregnet ved hjælp af observationen y . Så er

$$C_{1-\alpha}(y) = \{ \mu_0 \in \mathbb{R} \mid \varepsilon(y) > \alpha \text{ for hypotesen } H : \mu = \mu_0 \}.$$

Bevis Vi bruger definitionen af $C_{1-\alpha}(y)$ og rykker rundt på leddene:

$$\begin{aligned} \mu_0 \in C_{1-\alpha}(y) &\Leftrightarrow \bar{y} - z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} < \mu_0 < \bar{y} + z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \\ &\Leftrightarrow -z_{1-\alpha/2} < \frac{\bar{y} - \mu_0}{\sigma_0/\sqrt{n}} < z_{1-\alpha/2} \\ &\Leftrightarrow -z_{1-\alpha/2} < u < z_{1-\alpha/2}. \end{aligned}$$

Dette er ensbetydende med at $P(|U| \geq |u|) > \alpha$ hvor $U \sim N(0, 1)$, og det følger af Sætning 3.8 at dette er ensbetydende med at $\varepsilon(y) > \alpha$ hvor $\varepsilon(y)$ er p -værdien for hypotesen $\mu = \mu_0$. \square

Eksempel 3.11. *(Kobbertråd, fortsættelse af eksempel 3.1, side 31)* Vi beregnede i eksempel 3.6 (side 38) et 95% konfidensinterval for μ og konstaterede at værdien 18.441 ikke er inkluderet. I eksempel 3.9 (side 45) testede vi $H : \mu = 18.441$ og afviste den på 5%-niveau. De to konklusioner er konsistente. \square

Eksempel 3.12. *(Læsetest)* Antag at skalaen for en national læsetest er konstrueret således at resultaterne er normalfordelte med middelværdi 100 og spredning 12, dvs. varians 144. På en bestemt skole blev 55 elever testet og opnåede i gennemsnit en score på 97 point. Spørgsmålet er om dette resultat er udtryk for at skolens elever er dårligere end landsgennemsnittet eller om det lige så godt kan skyldes tilfældig variation.

Vi antager at de 55 elevers scorer, y_1, \dots, y_{55} er realisationer af uafhængige stokastiske variable Y_1, \dots, Y_{55} der alle er normalfordelte med middelværdi μ og varians $\sigma_0^2 = 144$. Det observerede gennemsnit er $\bar{y} = 97$. Estimatet og estimator for μ er således givet ved

$$\hat{\mu} = \bar{y} = 97, \quad \bar{Y} \sim N(\mu, \sigma_0^2/55).$$

Vi beregner et 95% konfidensinterval til

$$97 \pm 1.96 \frac{12}{\sqrt{55}} = 97 \pm 3.2 = (93.8, 100.2)$$

som lige netop indeholder værdien 100. Den relevante hypotese er $\mu = 100$ og giver anledning til

$$u = \frac{97 - 100}{12/\sqrt{55}} = -1.85, \quad \varepsilon(y) = 2 \cdot (1 - \Phi(1.85)) = 0.06.$$

Hypotesen kan således ikke afvises på 5% signifikansniveau, men både konfidensinterval og test indikerer at data er svagt usædvanlige hvis skolens elever læser lige så godt som landsgennemsnittet. \square

3.5 Sammenfatning og perspektiv

Vi har i dette kapitel diskuteret statistisk analyse af normalfordelte data med kendt varians. Modellen er ikke særligt anvendelig i praksis fordi det kun sjældent er rimeligt at antage at variansen er kendt på forhånd. Det vigtige i kapitlet er først og fremmest introduktionen og diskussionen af de vigtige statistiske begreber. Lad os opsummere:

Statistisk model En statistisk model beskriver vores antagelser om frembringelsen af data. I modellen indgår en eller flere parametre som skal estimeres ved hjælp af data.

Maksimum likelihood estimation Som estimator bruger vi den værdi af parameteren der gør de observerede data mest sandsynlige, målt med den simultane tæthed. Dette formaliseres med likelihoodfunktionen, dvs. tætheden opfattet som funktion af parameteren (eller parametrene). Vi skelner mellem estimatet som er et tal og estimatoren som er en stokastisk variabel. Estimatet er en realisation af estimatoren, og fordelingen af estimatoren beskriver usikkerheden på estimatet.

Konfidensinterval Fordelingen af estimatoren kan opsummeres af et konfidensinterval med konfidensgrad der er specificeret på forhånd. Konfidensintervallet er et interval omkring estimatoren, og konfidensgraden er sandsynligheden for at intervallet indeholder den sande værdi. Man skal være varsom med fortolkningen.

Hypotesetest Kvotienttestet hører naturligt sammen med maksimum likelihood estimation. Testet består af flere ingredienser: opstilling af en hypotese, beregning af kvotientteststørrelsen der ved hjælp af likelihoodfunktionen måler hvor godt modellen passer til data, beregning af p -værdi og konklusion. p -værdien er sandsynligheden for at få en kvotientteststørrelse der er mindre end eller lig den observerede værdi, beregnet under antagelse af at hypotesen er sand. For at beregne p -værdien skal vi kende fordelingen af kvotientteststørrelsen under hypotesen, eller i det mindste fordelingen af en transformation af kvotientteststørrelsen. Hypotesen afvises hvis p -værdien er mindre end eller lig det på forhånd fastsatte signifikansniveau.

I de følgende kapitler skal vi diskutere den statistiske analyse af andre typer data, men analysen består af de samme trin som ovenfor. Det er derfor vigtigt at forstå meningen med og betydningen af begreberne. Der mangler en vigtig brik i listen ovenfor: modelkontrol. Hvordan kontrollerer man at antagelserne i modellen er rimelige, specielt om det er rimeligt at antage at data er normalfordelte? Vi vender tilbage til dette spørgsmål i afsnit 4.5.

3.6 R

I tilfældet med kendt varians er der ingen nemme genveje i R, men man kan nemt beregne alle de værdier man har brug for til analysen, og bruge dem til at lave konfidensintervaller, udføre hypotesetest osv. Gennemnittet \bar{y} beregnes med funktionen `mean`. Fraktiler og sandsynligheder i $N(0,1)$ beregnes med `qnorm` og `pnorm` — q for quantile og p for probability.

For kobberdata fra eksempel 3.1 får vi for eksempel følgende:

```
> vgt <- c(18.459, [Flere tal her], 18.443) # Indlæsning
> ybar <- mean(vgt) # Gennemsnit
> ybar
```

```
[1] 18.44822

> se <- sqrt(0.000074/9)      # Spredning på estimator
> se
[1] 0.002867442

> ybar - 1.96 * se          # Nedre grænse i 95% KI
[1] 18.44260
> ybar + 1.96 * se          # Øvre grænse i 95% KI
[1] 18.45384

> u <- (ybar-18.441) / se    # Teststørrelsen
> u
[1] 2.518699

> pnorm(2.519)              # P(U <= 2.519) hvis  $U \sim N(0,1)$ 
[1] 0.9941156
>
> 2*(1-pnorm(u))           # p-værdien
[1] 0.01177894
```

Check selv at tallene stemmer overens med tallene fra eksempel 3.4 (side 35), 3.6 (side 38) og 3.9 (side 45).

Ovenfor har vi gemt de værdier der skal bruges senere, for eksempel gennemsnittet og spredningen på estimatoren, i variable, som vi bruger i de senere beregninger. Det er naturligvis ikke nødvendigt — man kan for eksempel sagtens beregne teststørrelsen med en enkelt kommando:

```
> (mean(vgt) - 18.441) / sqrt(0.000074) * sqrt(9)
[1] 2.518699
```

Det er en smagssag om man foretrækker det ene eller det andet.

Vi brugte ovenfor at 97.5% fraktilen i $N(0,1)$ er 1.96. Hvis vi vil beregne konfidensintervaller med en anden konfidensgrad har vi brug for andre fraktiler. Til et 90% konfidensinterval skal vi bruge 95% fraktilen, og konfidensintervallet kan beregnes som følger:

```
> qnorm(0.95)              # 95%-fraktil i  $N(0,1)$ 
```

```
[1] 1.644854

> ybar - 1.645 * se          # Nedre grænse i 90% KI
[1] 18.44351
> ybar + 1.645 * se          # Øvre grænse i 90% KI
[1] 18.45294
```

Som det fremgår beregner `pnorm` værdier af fordelingsfunktionen for $N(0, 1)$. Hvis man i stedet har brug for sandsynligheder i normalfordelingen med en anden middelværdi og/eller varians, skal middelværdien og *spredningen* (ikke variansen!) angives som argumenter til `pnorm`. For eksempel er $P(Y \leq 0) = 0.0368$ hvis $Y \sim N(4, 5)$:

```
> pnorm(0, mean=4, sd=sqrt(5)) # P(Y<=0), Y~N(4,5)
[1] 0.03681914
```

eller blot

```
> pnorm(0, 4, sqrt(5))          # P(Y<=0), Y~N(4,5)
[1] 0.03681914
```

På tilsvarende måde kan `qnorm` bruges til beregning af fraktiler i normalfordelinger med vilkårlig middelværdi og varians.

Der findes to funktioner mere der er relateret til normalfordelingen: `dnorm` der beregner tætheder og `rnorm` der simulerer udfald:

```
> dnorm(1, mean=2, sd=0.5)     # Tæthed i 1 for N(2,0.25)
[1] 0.1079819
> rnorm(4, mean=2, sd=0.5)     # 4 udfald fra N(2,0.25)
[1] 2.013622 2.000236 1.846199 1.926197
```

Vi ser at tætheden for $N(2, 0.25)$ evalueret i punktet 1 er 0.1080, mens den sidste kommando har simuleret 4 observationer fra $N(2, 0.25)$.

3.7 Opgaver

3.1 I et medicinsk studie blev kropstemperaturen målt for 130 raske personer. Gennemsnittet af de 130 temperaturmålinger var 36.805. Det kan antages at observationerne er normalfordelte med spredning 0.4°C .

1. Beregn et 95% konfidensinterval for den gennemsnitlige kropstemperatur for raske mennesker. Beregn også et 90% konfidensinterval.
2. Det antages ofte at den gennemsnitlige kropstemperatur for raske mennesker er 37° C. Bekræfter eller afkræfter data denne hypotese?

3.2 En løber er interesseret i at undersøge om hendes løbeur er kalibreret korrekt. Hun udmåler derfor en strækning på præcis 1000 m og løber den 16 gange. For hver løbetur noterer hun den distance som løbeuret registrerer som løbet distance. Gennemsnittet af de 16 målinger er 1013 meter. Fabrikanten af løbeuret siger at variationen af løbeurets distancemålinger kan beskrives med en spredning på 30 meter for en strækning på 1000 meter.

1. Opstil en statistisk model til beskrivelse af forsøget.
2. Angiv et estimat og et 95% konfidensinterval for middelværdien af løbeurets distancemålinger.
3. Udfør et test for hypotesen om at løbeuret er kalibreret korrekt. Du kan bruge at $P(U \leq 1.733) = 0.958$ hvis $U \sim N(0, 1)$. *Vink:* Hvad er den relevante hypotese?
4. Angiv et estimat og et 95% konfidensinterval for den forventede fejl i løbeurets distancemåling. *Vink:* Hvad er fejlen som funktion af middelværdien af distancemålingerne? Hvad kunne være et fornuftigt estimat for fejlen?

3.3 I eksempel 3.6 (side 38) blev 95% konfidensintervallet for gennemsnitsvægten af kobbertråde beregnet til (18.4426, 18.45384). Specielt har konfidensintervallet længden 0.01124. Dette var baseret på en stikprøve på 9 kobbertråde.

1. Hvor stor skal stikprøven være for at længden af konfidensintervallet bliver halvt så langt?
2. Udled et generelt resultat: i tilfældet med en enkelt stikprøve med kendt varians, hvor meget skal stikprøvestørrelsen øges for at længden af konfidensintervallet for middelværdien bliver halveret? Afhænger resultatet af konfidensgraden for konfidensintervallet?
3. Udled et andet generelt resultat: i tilfældet med en enkelt stikprøve med kendt varians σ_0^2 , hvor stor skal stikprøven være for at konfidensintervallet for middelværdien med konfidensgrad $1 - \alpha$ får en længde der er højst l ?

3.4 Betragt den statistiske model hvor Y_1, \dots, Y_n er uafhængige og normalfordelte med ukendt middelværdi μ og kendt varians σ_0^2 . Vi skal i denne opgave interessere os for den såkaldte styrke af testet for hypotesen $H : \mu = 0$. Overalt testes med signifikansniveau 5%.

1. Gør rede for at hypotesen forkastes hvis og kun hvis

$$|\bar{Y}| \geq 1.96 \frac{\sigma_0}{\sqrt{n}}.$$

2. Betragt en fast men vilkårlig værdi af den sande middelværdi μ . Nedenfor er sandsynligheden for at hypotesen $H : \mu = 0$ forkastes, som funktion af μ , beregnet. Overbevis dig selv om at beregningerne er korrekte.

$$\begin{aligned} g(\mu) &= P\left(|\bar{Y}| \geq 1.96 \cdot \frac{\sigma_0}{\sqrt{n}}\right) \\ &= P\left(\left|\frac{\sqrt{n}\bar{Y}}{\sigma_0}\right| \geq 1.96\right) \\ &= P\left(\frac{\sqrt{n}\bar{Y}}{\sigma_0} \leq -1.96\right) + P\left(\frac{\sqrt{n}\bar{Y}}{\sigma_0} \geq 1.96\right) \\ &= \Phi\left(-1.96 - \frac{\sqrt{n}}{\sigma_0}\mu\right) + 1 - \Phi\left(1.96 - \frac{\sqrt{n}}{\sigma_0}\mu\right) \end{aligned} \quad (3.11)$$

hvor Φ er fordelingsfunktionen for standardnormalfordelingen.

3. Beregn $g(0)$ og forklar hvad relationen er til fejl af type I. *Vink:* Er hypotesen sand eller falsk?
4. Antag at $\mu \neq 0$ og forklar hvad relationen er mellem $g(\mu)$ og fejl af type II. *Vink:* Er hypotesen sand eller falsk?
5. Hvad sker der med $g(\mu)$ når $|\mu|$ vokser? Relatér til type II fejl.
6. Betragt nu g som funktion af n for fast μ . Forklar hvad der sker når n vokser. Relatér til type II fejl.
7. Sæt $\sigma_0 = 1$. Tegn grafen for g som funktion af μ på intervallet $(-1.5, 1.5)$ for $n = 10$ og for $n = 25$, gerne i samme figur. Forklar hvad du ser.
Følgende R-kode kan evt. benyttes — sørg for at forstå hvad de enkelte kommandoer gør!

```
## definerer funktionen både som funktion af mu og n
g <- function(mu,n)
  pnorm(-1.96-sqrt(n)*mu) + 1-pnorm( 1.96-sqrt(n)*mu)

## mu-værdier og tilhørende funktionsværdier
x <- seq(-1.5,1.5,0.05)
y10 = g(x,10)
y25 = g(x,25)

## Selve figuren
plot(x,y10,type="l")
lines(x,y25,col=2)
```

8. Antag at vi gerne vil være i stand til at opdage en afvigelse på 0.3 fra 0 af middelværdien med en sikkerhed på 80%. Hvor stor skal stikprøven være for at dette er opfyldt? *Vink:* Du skal finde n så $g(0.3) = 0.8$. Hvorfor? Prøv dig frem, for eksempel med R-funktionen g .

Kapitel 4

En stikprøve med ukendt varians

I kapitel 3 betragtede vi modellen for en enkelt stikprøve med kendt varians. I eksempel 3.1 om kobbertråd gav det god mening fordi man på fabrikken har lang erfaring med variationen i produktionen. I langt de fleste tilfælde har man imidlertid ikke nogen ide om størrelsen af variansen, og vi vil nu betragte det mere realistiske tilfælde hvor både middelværdi og varians er ukendte. Tingene bliver en smule mere komplicerede fordi der er to ukendte parametre, men begreberne er de samme, så vi kan trække på vores erfaring fra det simple tilfælde.

4.1 Statistisk model

Udgangspunktet for modellen er stadig uafhængige og normalfordelte stokastiske variable Y_1, \dots, Y_n med middelværdi μ og varians σ^2 . Den simultane fordeling betegnes N_{μ, σ^2}^n og har tæthed

$$\begin{aligned} f_{\mu, \sigma^2}(y) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right), \quad y = (y_1, \dots, y_n) \in \mathbb{R}^n. \end{aligned} \quad (4.1)$$

Notationen f_{μ, σ^2} understreger at både middelværdi og varians er ukendte parametre. Vi har med andre ord en todimensional parameter (μ, σ^2) . Vi antager at parametermængden er $\Theta = \mathbb{R} \times (0, \infty)$, men det kunne også være en delmængde af denne mængde.

Definition 4.1. Modellen for en enkelt stikprøve med ukendt varians består af udfaldsrummet \mathbb{R}^n samt familien

$$\mathcal{P} = \{N_{\mu, \sigma^2}^n : (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$$

af fordelinger på \mathbb{R}^n hvor N_{μ, σ^2}^n har tæthed (4.1).

Alternativ formulering: Lad Y_1, \dots, Y_n være uafhængige og identisk normalfordelte stokastiske variable, $Y_i \sim N(\mu, \sigma^2)$ hvor $\mu \in \mathbb{R}$ og $\sigma^2 > 0$ er ukendte parametre.

Eksempel 4.2. (Prothrombinindeks) En persons prothrombinindeks er en markør for leversvigt hvor et lavt indeks indikerer leversvigt. For at undersøge effekten af en behandling fik 40 personer målt deres prothrombinindeks både før og efter behandling. Som observationer bruger vi forskellen mellem de to målinger således at en positiv værdi af y_i indikerer en positiv effekt af behandlingen. Data består altså af 40 observationer y_1, \dots, y_{40} . Observationerne betragtes som realisationer af Y_1, \dots, Y_{40} som antages at være uafhængige og normalfordelte med middelværdi μ og varians σ^2 . \square

4.2 Maksimum likelihood estimation

Vi skal estimere (μ, σ^2) på basis af data, $y = (y_1, \dots, y_n)$. Vi definerer igen likelihoodfunktionen som tætheden, opfattet som funktion af parameteren,

$$L_y : \mathbb{R} \times (0, \infty) \rightarrow \mathbb{R}$$

$$L_y(\mu, \sigma^2) = f_{\mu, \sigma^2}(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right). \quad (4.2)$$

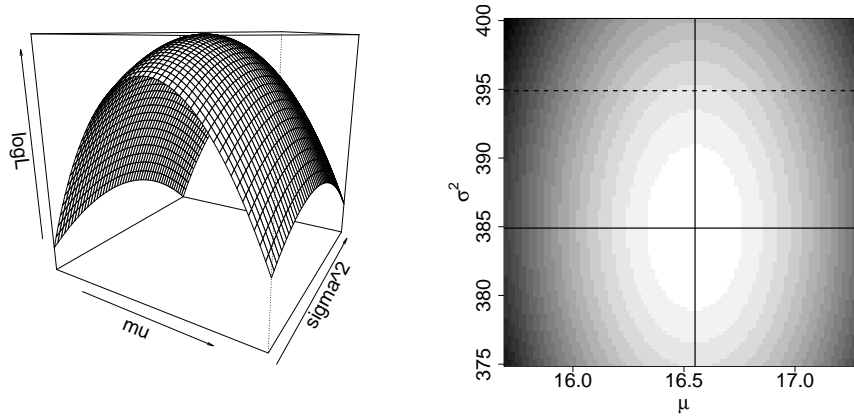
Et maksimum likelihood estimat for $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ opfylder

$$L_y(\hat{\mu}, \hat{\sigma}^2) \geq L_y(\mu, \sigma^2), \quad (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty). \quad (4.3)$$

Man ser ofte på log-likelihoodfunktionen, dvs.

$$l_y(\mu, \sigma^2) = \log L_y(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2.$$

Da log er strengt voksende er maksimering af L_y ækvivalent med maksimering af l_y . Den venstre del af figur 4.1 viser et 3D-plot for log-likelihoodfunktionen for prothrombindata fra eksempel 4.2. Det ser ud til at log-likelihoodfunktionen — og dermed likelihoodfunktionen — har et entydigt maksimum.



Figur 4.1: Log-likelihoodfunktionen for prothrombindata (eksempel 4.2). Figuren til venstre er et 3D-plot, mens figuren til højre viser værdierne af l_y på en gråtoneskala (se side 60 for detaljer).

Eksistens og entydighed af et maksimum er netop et af udsagnene i sætning 4.3 nedenfor. Vi har brug for lidt notation for at formulere sætningen: husk kvadratafvigelsessummen $SSD_Y = \sum_{i=1}^n (Y_i - \bar{Y})^2$, se sætning A.5 i appendiks A, og indfør den tilsvarende observerede størrelse, $SSD_y = \sum_{i=1}^n (y_i - \bar{y})^2$. SSD står for “sum of squared deviations”, og fodtegnet viser om det er den stokastiske eller den observerede version der er tale om. Bemærk at sætningen bør læses sammen med bemærkning 4.5.

Sætning 4.3. For den statistiske model fra definition 4.1 er maksimum likelihood estimatet for (μ, σ^2) entydigt bestemt og givet ved

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \hat{\sigma}^2 = \frac{1}{n} SSD_y = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Estimatorerne $\hat{\mu} = \bar{Y}$ og $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n} SSD_Y$ er uafhængige, og deres marginale fordelinger er

$$\bar{Y} \sim N(\mu, \sigma^2/n), \quad \hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi_{n-1}^2.$$

Inden vi beviser sætningen, bemærker vi at $\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi_{n-1}^2$ betyder at $\hat{\sigma}^2$ er χ^2 -fordelt med $n - 1$ frihedsgrader og skalaparameter σ^2/n . Med andre ord: $\frac{n}{\sigma^2} \hat{\sigma}^2$ er “ægte” χ_{n-1}^2 -fordelt. Det viser specielt at $E(\hat{\sigma}^2) < \sigma^2$. Vi korrigerer derfor $\hat{\sigma}^2$ og dividerer med $n - 1$ i stedet for n , se bemærkning 4.5.

Bevis Vi skal maksimere en funktion af to variable og benytter os af en metode der kaldes profilering, se appendiks B. Ideen er at erstatte det todimensionale maksimeringsproblem med to endimensionale maksimeringsproblemer som hver for sig er nemme at løse.

Betragt først en fast positiv værdi af σ^2 . Funktionen $\mu \rightarrow L_y(\mu, \sigma^2)$ er identisk med likelihoodfunktionen fra modellen med kendt varians, så det følger af sætning 3.3 at der er entydigt maksimum for $\mu = \bar{y}$.

Dette gælder for alle $\sigma^2 > 0$, dvs.

$$L_y(\bar{y}, \sigma^2) \geq L_y(\mu, \sigma^2), \quad \mu \in \mathbb{R}, \sigma^2 > 0.$$

Vi betragter derfor funktionen $\tilde{L}_y : (0, \infty) \rightarrow \mathbb{R}$ defineret ved

$$\tilde{L}_y(\sigma^2) = L_y(\bar{y}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\text{SSD}_y}{2\sigma^2}\right).$$

Denne funktion kaldes for profillikelihoodfunktionen for σ^2 . Lemma 4.4 nedenfor — anvendt med $x = \sigma^2$, $a = \text{SSD}_y/2$ og $b = n/2$ — viser at \tilde{L} har maksimum for $\sigma^2 = \frac{1}{n} \text{SSD}_y$. Vi har således vist at

$$L_y\left(\bar{y}, \frac{1}{n} \text{SSD}_y\right) = \tilde{L}_y\left(\frac{1}{n} \text{SSD}_y\right) \geq \tilde{L}_y(\sigma^2) = L_y(\bar{y}, \sigma^2) \geq L_y(\mu, \sigma^2)$$

for alle $\mu \in \mathbb{R}$ og $\sigma^2 > 0$ så $(\bar{y}, \frac{1}{n} \text{SSD}_y)$ er et maksimumpunkt for L_y .

Resultatet vedrørende fordelingen af $(\bar{Y}, \hat{\sigma}^2)$ følger direkte af sætning A.5 i appendiks A. \square

I beviset brugte vi følgende lemma, som kommer os til nytte flere gange i de følgende kapitler:

Lemma 4.4. *Lad a og b være positive, reelle tal, og definer funktionen f ved*

$$f(x) = x^{-b} e^{-\frac{a}{x}}, \quad x \in (0, \infty).$$

Så har f entydigt maksimum for $x = \frac{a}{b}$.

Bevis Definer funktionen g ved

$$g(x) = \log(f(x)) = -b \log(x) - \frac{a}{x}, \quad x \in (0, \infty).$$

Da log er strengt voksende har f og g maksimum samme sted, men g er nemmere at regne på. Vi ser at g er to gange kontinuert differentiabel med

$$g'(x) = -\frac{b}{x} + \frac{a}{x^2}, \quad g''(x) = \frac{b}{x^2} - \frac{2a}{x^3}.$$

Specielt er $g'(x) = 0$ hvis og kun hvis $x = \frac{a}{b}$ så dette er det eneste stationære punkt. Desuden er $g''(\frac{a}{b}) = -\frac{b^3}{a^2} < 0$ så der er tale om et maksimumpunkt. \square

Bemærk at $E(\bar{Y}) = \mu$ således at \bar{Y} er en central estimator for μ . Derimod er

$$E(\hat{\sigma}^2) = \frac{n-1}{n} \sigma^2$$

så $\hat{\sigma}^2$ er ikke en central estimator for σ^2 . Det følger af at middelværdien i χ_k^2 -fordelingen er k således at $E(\text{SSD}_Y) = n-1$. I gennemsnit estimeres σ^2 således for lavt hvis vi benytter maksimum likelihood estimatoren. Det er imidlertid nemt at korrigere $\hat{\sigma}^2$ og opnå en central estimator: vi skal blot normere med $n-1$ i stedet for n i definitionen af $\hat{\sigma}^2$, og i stedet bruge

$$\tilde{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1} \text{SSD}_Y$$

som estimator. Så er $E(\tilde{\sigma}^2) = \sigma^2$, \bar{Y} og $\tilde{\sigma}^2$ er uafhængige, og $(n-1)\tilde{\sigma}^2 \sim \sigma^2 \chi_{n-1}^2$. Altså er $\tilde{\sigma}^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$. Det tilsvarende estimat, hvor observationerne sættes ind, betegnes som regel s^2 , dvs.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2, \quad (4.4)$$

og det er så godt som altid dette estimat vi bruger for variansen. Størrelsen kaldes også for den empiriske varians. For god ordens skyld samler vi resultatet vedrørende estimatorerne i en bemærkning:

Bemærkning 4.5. I den statistiske model fra definition 4.1 bruger vi estimatorerne $\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ og $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$.

Den sande eller teoretiske fordeling af $(\bar{Y}, \tilde{\sigma}^2)$ er givet ovenfor, men afhænger af de ukendte parametre. Spredningen i fordelingen af $\hat{\mu}$ er særligt vigtig fordi den giver os information om præcisionen af vores estimat. Sammen med selve estimatet, angiver man derfor som regel også den estimerede spredning for estimatoren. Den

estimerede spredning fås ved at erstatte σ med dets estimat, s , og man bruger som regel forkortelsen SE (standard error). Vi har altså

$$\text{SE}(\hat{\mu}) = \frac{s}{\sqrt{n}}.$$

Eksempel 4.6. (*Prothrombinindeks, fortsættelse af eksempel 4.2, side 56*) For de 40 observationer y_1, \dots, y_{40} af forskellen i prothrombinindeks før og efter behandling viste det sig at

$$\bar{y} = 16.55, \quad \text{SSD}_y = 15395.9$$

således at estimaterne er

$$\hat{\mu} = 16.55, \quad s^2 = \frac{15395.9}{39} = 394.8, \quad s = 19.87.$$

Estimatorerne \bar{Y} og $\tilde{\sigma}^2$ er uafhængige, $\bar{Y} \sim N(\mu, \sigma^2/n)$ og $\tilde{\sigma}^2 \sim \frac{\sigma^2}{39} \chi_{39}^2$. Den estimerede spredning for $\hat{\mu}$ er $\text{SE}(\hat{\mu}) = s/\sqrt{40} = 3.14$. \square

Estimationen er illustreret i den højre del af figur 4.1. Figuren viser log-likelihood-funktionen for prothrombindata på en gråtoneskala, hvor lyse pixels svarer til store værdier af l_y og mørke pixels svarer til små værdier af l_y . Den lodrette linje svarer til $\mu = 16.55$, mens de lodrette linjer svarer til $\sigma^2 = 384.9$ (maksimum likelihood estimatet) og $\sigma^2 = 394.8$ (det centrale estimat, s^2). Som beskrevet ovenfor foretrækker vi s^2 selvom likelihoodfunktionen er mindre.

Husk i øvrigt at vi i starten af beviset for sætning 3.3 maksimerede funktionen $\mu \rightarrow L_y(\mu, \sigma^2)$ for fast værdi af σ^2 . Vi indså at denne funktion har maksimum for $\mu = \bar{y}$ uanset værdien af σ^2 . Det kan vi godt fornemme på figuren. Uanset hvilken vandrette linje vi ser på, er figuren lysest for $\mu = 16.55$. Derefter maksimerede vi profillikelihoodfunktionen $\sigma^2 \rightarrow \tilde{L}_y(\sigma^2) = L_y(\bar{y}, \sigma^2)$. Dette svarer til at følge den lodrette linje i figuren og finde stedet hvor funktionen er størst mulig (lysest).

4.3 Konfidensinterval for middelværdien

I afsnit 3.3 udledte vi konfidensintervallet

$$\bar{Y} \pm z_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \tag{4.5}$$

for middelværdien i tilfældet med kendt varians. Det er oplagt at erstatte den kendte varians med estimatoren $\tilde{\sigma}^2$, men der skal tages højde for at det er en estimator i

stedet for en kendt værdi. Det forøger usikkerheden og ændrer fordelingerne. Derfor får vi brug for fraktiler i t -fordelingen. Lad $t_{k,1-\alpha/2}$ betegne $1 - \alpha/2$ fraktilen i t -fordelingen med k frihedsgrader.

Sætning 4.7. *Betragt den statistiske model fra definition 4.1. Så er*

$$\bar{Y} \pm t_{n-1,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n}} = \left(\bar{Y} - t_{n-1,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n}}, \bar{Y} + t_{n-1,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n}} \right) \quad (4.6)$$

et $1 - \alpha$ konfidensinterval for μ .

Bevis Det følger fra sætning A.5 i appendiks A (påstand 4) at

$$T = \frac{\sqrt{n}(\bar{Y} - \mu)}{\tilde{\sigma}} = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{\text{SSD}_Y / (n-1)}}$$

er t -fordelt med $n - 1$ frihedsgrader. Således er

$$P\left(-t_{n-1,1-\alpha/2} < \frac{\sqrt{n}(\bar{Y} - \mu)}{\tilde{\sigma}} < t_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

eller, hvis vi isolerer μ i midten,

$$P\left(\bar{Y} - t_{n-1,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n}} < \mu < \bar{Y} + t_{n-1,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n}}\right) = 1 - \alpha. \quad (4.7)$$

Dette viser som ønsket at $\bar{Y} \pm t_{n-1,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n}}$ er et konfidensinterval for μ med konfidensgrad $1 - \alpha$. \square

Bemærk at strukturen af konfidensintervallet er den samme som i tilfældet med kendt varians, bortset fra at spredningen for $\hat{\mu}$ ikke er kendt, men skal estimeres:

$$\hat{\mu} \pm \text{fraktil} \cdot \text{estimeret spredning for } \hat{\mu},$$

sammenlign med (3.7). Mere specifikt så består forskellen mellem konfidensintervalterne (4.5) og (4.6) i at den kendte værdi σ_0 er erstattet med estimatoren $\tilde{\sigma}$ og at normalfordelingsfraktilen er udskiftet med en t -fordelingsfraktil. t -fordelingsfraktilen er altid større end den tilsvarende normalfordelingsfraktil (se figur A.1 i appendiks A eller figur 10.6 i BH), så konfidensintervallet er (lidt) bredere for modellen med ukendt varians sammenlignet med modellen med kendt varians (for ens værdier af σ_0^2 og $\tilde{\sigma}^2$). Dette giver god mening: vores uvidenhed om σ^2 giver anledning til ekstra usikkerhed om estimatet på μ . Der er dog ikke stor forskel på fraktilerne når n ikke er alt for lille.

For et datasæt bestående af observationerne y_1, \dots, y_n erstattes de stokastiske variable \bar{Y} og $\bar{\sigma}$ af de observerede størrelser \bar{y} og s . Hvis der for eksempel er ti observationer beregnes 95% konfidensintervallet som

$$\bar{y} \pm 2.262 \frac{s}{\sqrt{n}},$$

da 97.5% fraktilen i t -fordelingen med 9 frihedsgrader er 2.262.

Diskussionerne fra afsnit 3.3 vedrørende konfidensintervallet er stadig gyldige:

- Fortolkningen af (4.7) er et udsagn om intervallet snarere end om parameteren og forstås bedst hvis man tænker på gentagelser af dataindsamlingen.
- Konfidensintervallet bliver smallere hvis n vokser og bredere hvis $1 - \alpha$ stiger. Hvis den sande varians σ^2 øges, vil s^2 typisk øges og konfidensintervallet vil blive bredere.

Bemærk at vi kun har konstrueret et konfidensinterval for middelværdien, μ . Ved at udnytte at SSD_Y er χ^2 -fordelt, kan man også lave et konfidensinterval for variansen, σ^2 , men i praksis er det sjældent det man interesserer sig for, så det undlader vi her.

Eksempel 4.8. (*Prothrombinindeks, fortsættelse af eksempel 4.2, side 56*) Husk at $n = 40$, $\bar{y} = 16.55$ og $s = 19.87$. Desuden er 97.5% fraktilen i t -fordelingen med 39 frihedsgrader lig 2.023, således at

$$16.55 \pm 2.023 \cdot \frac{19.87}{\sqrt{40}} = 16.55 \pm 6.35 = (10.20, 22.90)$$

er et 95% konfidensinterval for μ . Bemærk at nul ikke ligger i konfidensintervallet. Hvis $\mu = 0$ — svarende til at der ikke er en effekt af behandlingen — er det således ret usandsynligt at vi skulle have observeret de data vi faktisk har til rådighed. \square

4.4 Test af hypotese om middelværdien

Ligesom i afsnit 3.4 vil vi interessere os for hypotesen $H : \mu = \mu_0$ om middelværdien for en fast værdi $\mu_0 \in \mathbb{R}$. Der er ingen restriktioner på variansen, så vi kan også skrive hypotesen som

$$H : (\mu, \sigma^2) \in \Theta_0 = \{\mu_0\} \times (0, \infty)$$

Hypotesen er ikke en simpel hypotese da parametermængden under hypotesen, Θ_0 , indeholder mere end et enkelt punkt. Vi kan derfor ikke kopiere fremgangsmåden fra afsnit 3.4 fuldstændigt. I stedet er planen at gøre følgende:

- Estimere parameteren (μ, σ^2) under hypotesen, dvs. bestemme $(\hat{\mu}, \hat{\sigma}^2) \in \Theta_0$ så

$$L_y(\hat{\mu}, \hat{\sigma}^2) \geq L_y(\mu, \sigma^2), \quad (\mu, \sigma^2) \in \Theta_0.$$

Det er klart at $\hat{\mu} = \mu_0$ da det er den eneste mulige værdi, så det er kun et spørgsmål om at finde $\hat{\sigma}^2$.

- Opskrive kvotientteststørrelsen

$$Q(y) = \frac{L_y(\hat{\mu}, \hat{\sigma}^2)}{L_y(\hat{\mu}, \hat{\sigma}^2)}.$$

Som i kapitel 3 har vi $Q(y) \in (0, 1]$, og $Q(y)$ kan fortolkes som et mål for hvor sandsynlige data er under hypotesen i forhold til den oprindelige model. Små værdier er kritiske, dvs. passer dårligt med hypotesen. Som notationen antyder, er $Q(y)$ er en funktion af $y = (y_1, \dots, y_n)$. Den tilhørende stokastiske variabel betegnes $Q(Y)$.

- Bestemme p -værdien eller testsandsynligheden

$$\varepsilon(y) = P(Q(Y) \leq Q(y)),$$

dvs. sandsynligheden for at få en værdi af $Q(Y)$ der passer lige så dårligt som eller dårligere med hypotesen end den værdi vi har fået fra de observerede data, givet at hypotesen er sand.

- Afvise hypotesen hvis $\varepsilon(y) \leq \alpha$ for et på forhånd fastsat signifikansniveau og i givet fald konkludere at μ er signifikant forskellig fra μ_0 .

Det viser sig at vi kan udføre testet som et såkaldt t -test:

Sætning 4.9. *Betragt den statistiske model givet i definition 4.1 og hypotesen $H : \mu = \mu_0$ for et fast $\mu_0 \in \mathbb{R}$. Under hypotesen er maksimum likelihood estimatet $(\hat{\mu}, \hat{\sigma}^2)$ givet ved*

$$\hat{\mu} = \mu_0, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mu_0)^2.$$

Under hypotesen er $n\sigma^2 = \sum_{i=1}^n (Y_i - \mu_0)^2 \sim \sigma^2 \chi_n^2$. Kvotientteststørrelsen er givet ved

$$Q(y) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2} \right)^{n/2}.$$

Kvotienttestet kan udføres på

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}},$$

og p -værdien er givet ved

$$\varepsilon(y) = 2P(T \geq |t|) = 2 \cdot (1 - F_{t_{n-1}}(|t|))$$

hvor T er t -fordelt med $n - 1$ frihedsgrader og $F_{t_{n-1}}$ er fordelingsfunktionen for denne fordeling.

Bevis Det er klart at $\hat{\mu} = \mu_0$ da μ_0 er den eneste mulige værdi. Vi mangler så at finde maksimum for

$$L_y(\mu_0, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_0)^2\right).$$

som funktion af σ^2 . Det følger af lemma 4.4 — denne gang med $a = \frac{1}{2} \sum_{i=1}^n (y_i - \mu_0)^2$ — at maksimum antages for

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mu_0)^2.$$

Vi regner derefter på kvotientteststørrelsen $Q(y)$. Det følger af (4.2) og udtrykkene for $\hat{\sigma}^2$ og $\hat{\mu}$ at

$$-\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (y_i - \bar{y})^2 = -\frac{n}{2}, \quad -\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (y_i - \mu_0)^2 = -\frac{n}{2}$$

således at eksponentialleddene i tælleren og nævneren af $Q(y)$ er ens. Vi får derfor

$$Q(y) = \frac{L_y(\hat{\mu}, \hat{\sigma}^2)}{L_y(\hat{\mu}, \hat{\sigma}^2)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2}\right)^{n/2} = \left(\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \mu_0)^2}\right)^{n/2}. \quad (4.8)$$

Vi mangler at vise at kvotienttestet kan udføres som et test på t , så vi regner videre på $Q(y)$. For at lette notationen, indfører vi

$$u = \frac{\sqrt{n}(\bar{y} - \mu_0)}{\sigma}, \quad z = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{\sigma^2} \text{SSD}_y.$$

Så er

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{y} - \mu_0)}{\sqrt{\text{SSD}_y/(n-1)}} = \frac{\sigma u}{\sqrt{\sigma^2 z/(n-1)}} = \frac{u}{\sqrt{z/(n-1)}}.$$

Nævneren i (4.8) omskrives til

$$\begin{aligned} \sum_{i=1}^n (y_i - \mu_0)^2 &= \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \mu_0)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu_0)^2 + 2 \sum_{i=1}^n (y_i - \bar{y})(\bar{y} - \mu_0) \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu_0)^2 \\ &= \sigma^2 z + \sigma^2 u^2 \end{aligned}$$

så

$$(Q(y))^{2/n} = \frac{\sigma^2 z}{\sigma^2 z + \sigma^2 u^2} = \left(1 + \frac{u^2}{z}\right)^{-1} = \left(1 + \frac{t^2}{n-1}\right)^{-1}.$$

Vi kan således skrive $Q(y)$ som en aftagende funktion af t^2 .

Hvis vi definerer den stokastiske variabel

$$T = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{\sqrt{\text{SSD}_Y/(n-1)}},$$

har vi den samme sammenhæng mellem $Q(Y)$ og T^2 , og t er et udfald af T . Vi får derfor

$$\varepsilon(y) = P(Q(Y) \leq Q(y)) = P(T^2 \geq t^2) = 2P(T \geq |t|)$$

som ønsket. Bemærk endelig at det følger af sætning A.5 i appendiks A at T er t -fordelt med $n-1$ frihedsgrader, således at p -værdien kan beregnes som angivet i sætningen. \square

Sætningen siger at testet består i at beregne den observerede værdi af T -teststørrelsen, dvs. t , og beregne hvor ekstremt værdien ligger i t -fordelingen med $n-1$ frihedsgrader. Intuitivt giver dette god mening: Hypotesen bør afvises hvis \bar{y} og μ_0 afviger meget og bør således baseres på $|\bar{y} - \mu_0|$. Division med s/\sqrt{n} kan opfattes som en normering der transformerer teststørrelsen til en kendt skala og således tager højde for variationen i data. Testet kaldes et t -test.

Bemærk sammenhængen med testet på u fra sætning 3.8 i tilfældet med kendt varians. Den kendte spredning er erstattet med estimatet s . Derfor ændres fordelingen af den tilhørende teststørrelse fra standardnormalfordelingen til en t -fordeling. Det betyder at værdien af t skal være større for at blive signifikant end den tilsvarende u -størrelse. Det skyldes den ekstra usikkerhed der er introduceret i modellen når variansen ikke

er kendt. Bemærk dog at hvis n ikke er alt for lille — svarende til at variansen er rimeligt præcist estimeret — så ligner $N(0, 1)$ og t_{n-1} -fordelingen hinanden, og det gør ikke den store forskel om vi benytter normalfordelingen eller t -fordelingen.

Hvis hypotesen ikke kan afvises, plejer man at opdatere estimerterne, dvs. angive $\hat{\mu} = \mu_0$ og $\hat{\sigma}^2$. Rationalet er at der ikke er belæg for at den oprindelige model med ukendt middelværdi giver en bedre beskrivelse af data end modellen svarende til hypotesen.

Kommentarerne fra afsnit 3.4 vedrørende sprogbrug, fejltyper og sammenhængen mellem konfidensintervaller og hypotesetest gælder uændret. Specielt indeholder $1 - \alpha$ konfidensintervallet for μ netop de værdier μ_0 for hvilke hypotesen $H : \mu = \mu_0$ ikke kan afvises på signifikansniveau α .

Eksempel 4.10. (*Prothrombinindeks, fortsættelse af eksempel 4.2, side 56*) Husk at observationerne er forskellen mellem målinger af prothrombinindekset før og efter en behandling. Hvis der ikke er nogen effekt af behandlingen, må vi forvente at niveauet i gennemsnit er ens før og efter behandling. Ingen effekt af behandlingen svarer således til hypotesen $H : \mu = 0$. Værdien af t -teststørrelsen er

$$t = \frac{\sqrt{n}(\bar{y} - 0)}{s} = \frac{\sqrt{40} \cdot 16.55}{19.87} = 5.27$$

og p -værdien er

$$\varepsilon(y) = 2P(T \geq 5.27) < 0.0001$$

hvor $T \sim t_{39}$. Der er således stærk evidens mod hypotesen som afvises, og det er påvist at behandlingen har en effekt. Stigningen i prothrombinindekset er estimeret til 16.55, med 95% konfidensinterval (10.2, 22.90). \square

Testet i ovenstående eksempel kaldes et parret t -test, fordi data består af par af observationer, nemlig målinger af prothrombinindeks før og efter behandling (se eksempel 4.2, side 56). Før- og eftermålingerne for den samme person kan næppe antages at være uafhængige, hvorimod det er rimeligt at antage at differenserne for de forskellige personer er uafhængige. Analysen gennemføres derfor på differenserne. Det følgende eksempel giver også anledning til et parret t -test.

Eksempel 4.11. (*Dagligvarepriser*) Dagbladet Politiken laver med jævne mellemrum sammenligninger af dagligvarepriser i forskellige butikskæder. I juni 2009 undersøgte man priserne på 34 veldefinerede varer — for eksempel 1 liter letmælk, 500 g skiveskåret rugbrød, 1 kg gulerødder — i fem discountkæder, bla. Netto og Fakta. Den samlede pris for de 34 varer var 343.38 kr i Netto og 354.54 kr i Fakta. Fakta er

altså dyrere for netop dette udvalg af varer, men spørgsmålet er om dette skyldes det specifikke udvalg af varer eller om resultatet kunne tænkes at være anderledes for et andet udvalg af varer (en anden stikprøve).

For hver af de 34 varer ser vi på forskellen mellem log-prisen i Fakta og log-prisen i Netto, dvs.

$$y_i = \log(f_i) - \log(n_i)$$

hvor f_i er prisen i Fakta og n_i er prisen i Netto. Bemærk at y_i approksimativt er lig $(f_i - n_i)/n_i$, dvs. den relative prisforskel, da $\log(1 + x) \approx x$ når x er lille. Fordelen ved at bruge ovenstående definition er at de to sæt af priser indgår symmetrisk på nær fortegn. Den rå forskel $f_i - n_i$ kan være uheldig fordi den i højere grad afhænger af prisniveauet på varerne.

Vi antager at y_1, \dots, y_{34} er udfald af uafhængige stokastiske variable Y_1, \dots, Y_{34} og at $Y_i \sim N(\mu, \sigma^2)$. Gennemsnit og empirisk varians og spredning for de 34 observationer viste sig at være

$$\hat{\mu} = \bar{y} = 0.025, \quad s^2 = 0.0285, \quad s = 0.169.$$

95% konfidensintervallet for μ beregnes til $(-0.034, 0.084)$. Endepunkterne svarer til 3.4% besparelse i Fakta henholdsvis 8.4% besparelse i Netto. Bemærk at nul ligger i konfidensintervallet. At varerne i gennemsnit koster det samme i de to butikker svarer til $\mu = 0$ så den relevante hypotese er $H : \mu = 0$. Den observerede værdi af t -teststørrelsen er $t = 0.87$ og skal vurderes i t -fordelingen med 33 frihedsgrader. p -værdien er 0.39. De indsamlede priser giver således ikke belæg for at sige at prisniveauet er forskelligt i de to butikskæder. Da hypotesen ikke kan afvises, opdaterer vi estimererne: $\hat{\mu} = 0$, $\hat{\sigma}^2 = 0.0282$, og $\hat{\sigma} = 0.168$. \square

4.5 Kontrol af normalfordelingsantagelse

I de foregående afsnit har vi udledt estimer, konfidensintervaller og test og diskuteret deres egenskaber. Alt dette gælder hvis den statistiske model er sand, altså hvis variationen i data kan beskrives ved hjælp af en normalfordeling, dvs. ved hjælp af tætheden (4.1) for passende værdier af μ og σ^2 . Hvis data ikke er normalfordelt kender vi ikke egenskaberne og kan derfor ikke stole på resultaterne fra analysen. Det er derfor vigtigt at lave modelkontrol, dvs. kontrollere om antagelserne i den statistiske model er rimelige for de givne data.

Vi fokuserer her på *normalfordelingsantagelsen*: givet data y_1, \dots, y_n , hvordan undersøger vi om variationen med rimelighed kan beskrives med en normalfordeling?

Vi vil ikke udføre et egentligt test — selvom sådanne findes — men derimod lave grafisk modelkontrol på to måder:

Histogram og normalfordelingstæthed Hvis datasættet er tilstrækkeligt stort, laver man ofte et histogram hvor skalaen er normeret således at arealet af kasserne tilsammen er en, og sammenligner med tætheden for $N(\bar{y}, s^2)$, dvs. normalfordelingen med middelværdi og varians givet ved de estimerede (empiriske) værdier.

Hvis observationerne er normalfordelte, så bør tætheden være en god approksimation til histogrammet, da arealer under tætheden kan fortolkes som sandsynligheder.

QQ-plot I et QQ-plot sammenlignes de empiriske fraktiler med normalfordelingsfraktilerne. På engelsk hedder fraktile 'quantile' — heraf navnet QQ-plot. På dansk kaldes et QQ-plot sommetider for et fraktilplot.

Antag først at vi vil undersøge om z_1, \dots, z_n kommer fra $N(0, 1)$. Brug notationen $z_{(j)}$ for den j 'te mindste observation således at

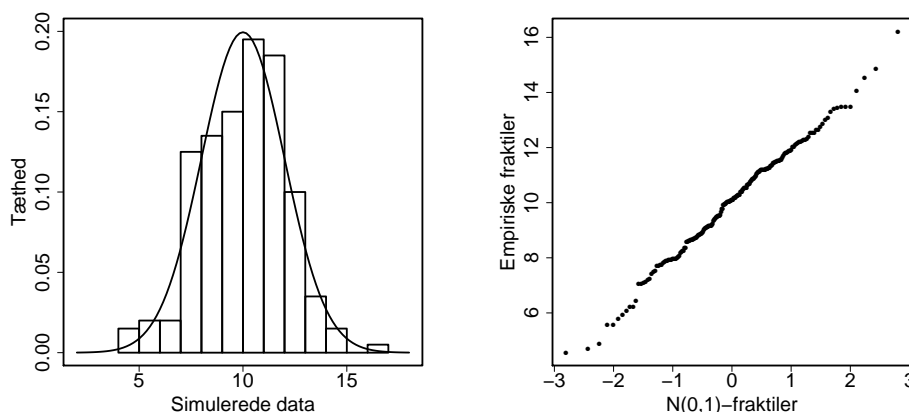
$$z_{(1)} < z_{(2)} < \dots < z_{(n)}.$$

Vi inddeler enhedsintervallet $(0, 1)$ i n lige store dele. Midtpunktet i det j 'te interval er så $p_j = (j - 0.5)/n$. Den empiriske p_j -fraktile defineres som den j 'te mindste observation, dvs. $z_{(j)}$. Den tilsvarende fraktile i $N(0, 1)$ er $u_j = \Phi^{-1}(p_j)$, hvor Φ er fordelingsfunktionen for $N(0, 1)$. Et QQ-plot er et scatterplot af $z_{(j)}$ mod u_j for $j = 1, \dots, n$. Hvis observationerne z_1, \dots, z_n er genereret af $N(0, 1)$, så stemmer de empiriske fraktiler og normalfordelingsfraktilerne overens på nær tilfældig variation, så punkterne bør ligge omkring en ret linje med skæring 0 og hældning 1.

Antag i stedet at vi vil undersøge om y_1, \dots, y_n kommer fra $N(\mu, \sigma^2)$ for et eller andet sæt af værdier (μ, σ^2) . Ligesom før ordner vi observationerne så $y_{(1)} < y_{(2)} < \dots < y_{(n)}$, og tegner $y_{(j)}$ mod fraktilerne u_j fra $N(0, 1)$. Husk at hvis $Y \sim N(\mu, \sigma^2)$ så kan vi skrive $Y = \mu + \sigma Z$ hvor $Z \sim N(0, 1)$. Hvis y_1, \dots, y_n er genereret fra $N(\mu, \sigma^2)$, forventer vi derfor at punkterne ligger omkring en ret linje med skæring μ og hældning σ .

Når vi laver et QQ-plot, dvs. optegner de empiriske fraktiler mod $N(0, 1)$ -fraktilerne, skal vi således kigge efter om punkterne — på nær tilfældig variation — ligger omkring en ret linje. Systematiske afvigelser fra en ret linje, tyder på at data ikke er normalfordelt.

De to modelkontrolmetoder er illustreret i figur 4.2 for 200 værdier simuleret fra $N(10,4)$. Tætheden er en god approksimation til histogrammet (venstre figur), og punkterne ligger omkring en ret linje i QQ-plottet (højre figur).



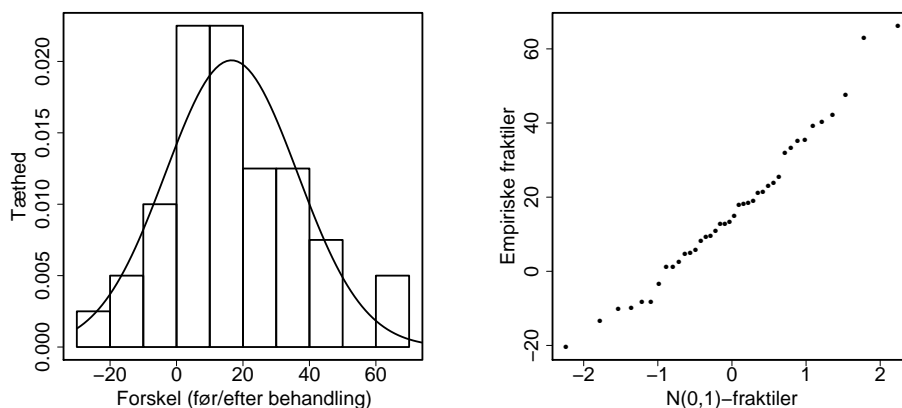
Figur 4.2: Histogram og normalfordelingstæthed (til venstre) og QQ-plot (til højre) for 200 værdier simuleret fra $N(10,4)$.

Eksempel 4.12. (*Prothrombinindeks, fortsættelse af eksempel 4.2, side 56*) Figur 4.3 viser et histogram og et QQ-plot for de 40 observationer af forskellen i prothrombinindex før og efter behandling. Bemærk at vi laver modelkontrollen for forskellen — som jo er den variabel vi analyserer — ikke for de originale prothrombinindeksmålinger. Begge figurer tyder på at normalfordelingsantagelsen er rimelig for disse data. □

Eksempel 4.13. (*Vægt af hjerner*) P. Topinard publicerede i 1888 data vedrørende størrelsen af menneskehjerner. Vi vil her bruge data der består af vægten af hjernen for 108 mænd (Samuels and Witmer, 2003, eksempel 2.12). Kontrol af normalfordelingsantagelsen er illustreret i figur 4.4. Hverken histogrammet til venstre eller QQ-plottet til højre giver anledning til bekymring vedrørende normalfordelingsantagelsen.

Gennemsnit og empirisk spredning for de 108 observationer er $\bar{y} = 1270.7$ og $s = 129.2$. Dette giver et 95% konfidensinterval for middelværdien på $(1246.1, 1295.4)$. Regn selv efter! Bemærk at der ikke er nogen naturlig hypotese at teste i dette tilfælde. □

Eksempel 4.14. (*Malaria*) En medicinsk forsker tog blodprøver fra 31 børn inficeret med malaria og bestemte for hvert barn antallet af malariaparasitter i 1 ml blod (Samuels and Witmer, 2003, opgave 2.75).

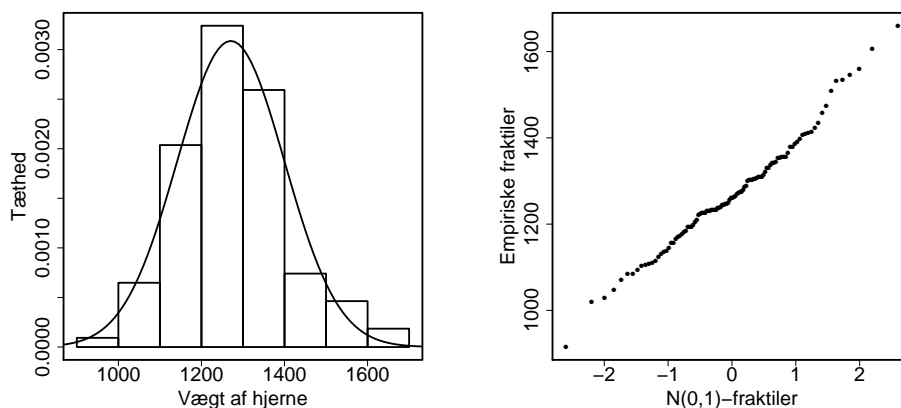


Figur 4.3: Histogram og normalfordelingstæthed (til venstre) og QQ-plot (til højre) for forskellen i prothrombinindeks før og efter behandling, se eksempel 4.12.

QQ-plottet for observationerne er vist i venstre side af figur 4.5. Punkterne afviger voldsomt fra en ret linje, så det er urimeligt at antage at antallet af malariaparasitter er normalfordelt blandt malariainficerede børn. Problemet er en meget tung hale af høje observationer. Dette kommer også til udtryk ved at gennemsnittet (12890) er meget højere end medianen (3672). Højre side af figuren viser QQ-plottet for de logaritmetransformerede antal. Dette plot giver ikke anledning til bekymring, så det vil være rimeligt at analysere de logaritmetransformerede data ved hjælp af en normalfordelingsmodel. \square

Eksemplet med malariaparasitter illustrerer en vigtig pointe, nemlig at det sommetider er nødvendigt at transformere data før en normalfordelingsantagelse er rimelig. I eksemplet gjorde logaritmetransformationen nytte fordi problemet var en tung hale til højre i fordelingen: intuitionen er at logaritmen “trækker skalaen sammen” således at ekstremt høje værdier på den oprindelige skala er knapt så høje på log-skalaen. Det er imidlertid ikke altid at man kan finde en passende transformation. I så fald må man ty til helt andre metoder, men det skal vi ikke komme yderligere ind på her.

QQ-plottene i figurerne ovenfor var alle ret nemme at fortolke: der var enten klar overensstemmelse eller klar uoverensstemmelse med den rette linje. Sådan er det desværre ikke altid — faktisk kan det være ret svært at vurdere hvorvidt en afvigelse fra en ret linje kan tilskrives tilfældig variation eller at normalfordelingsantagelsen er urimelig, især for små datasæt. Figur 4.6 viser QQ-plots for fire simulerede datasæt hver bestående af 10 observationer fra $N(10, 4)$. Som det ses er der ret store afvigelser fra en ret linje — selvom vi ved at data er trukket fra normalfordelingen. Moralen er at



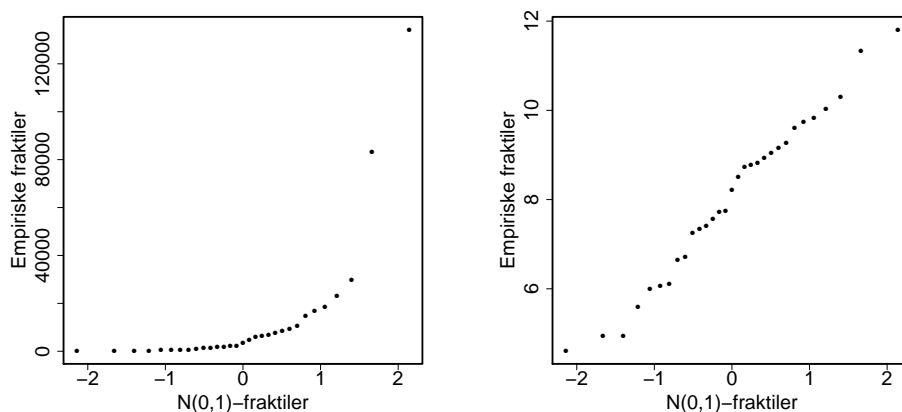
Figur 4.4: Histogram og normalfordelingstæthed (til venstre) og QQ-plot (til højre) for vægten af 108 hjerner, se eksempel 4.13.

der for små datasæt skal være ganske kraftige afvigelser fra en ret linje før man med sikkerhed kan skyde normalfordelingsantagelsen i sænk. En anden morale er at det kan være nyttigt at lave lignende simulerede QQ-plots hvis man for et givet datasæt er i tvivl om hvorvidt afvigelsen fra en ret linje kan skyldes tilfældig variation eller ej. Simulationerne giver en ide om størrelsesordenen af de naturlige afvigelser fra en ret linje når normalfordelingsantagelsen faktisk er sand.

4.6 Sammenfatning og perspektiv

Vi har diskuteret statistisk analyse af uafhængige normalfordelte observationer med ukendt middelværdi og varians. Modellen kan bruges når observationerne kan antages at være frembragt af samme normalfordeling. Som regel er man først og fremmest interesseret i at estimere middelværdien, og analysen sammenfattes ofte med estimatet og et konfidensinterval. Der er ikke altid en naturlig hypotese der ønskes testet.

Modellen kan også bruges til analyse af parrede data, hvor den samme størrelse er målt to gange, men under forskellige omstændigheder, på samme forsøgsgenhed (samme person, plante, maskine eller lignende). Man kan så analysere differenserne ved hjælp af modellen fra dette kapitel, og man er specielt interesseret i om middelværdien er lig nul, svarende til at der ikke er forskel i niveauet på den målte variabel under de to omstændigheder. I nogle situationer er det mere naturligt at se på forholdet mellem de to observationer i et observationspar (eller en anden funktion af dem). Pointen er først og fremmest at de to målinger fra et par bliver reduceret til en enkelt



Figur 4.5: QQ-plots for antallet af malariaparasitter (til venstre) og for logaritmen til dette antal (til højre), se eksempel 4.14.

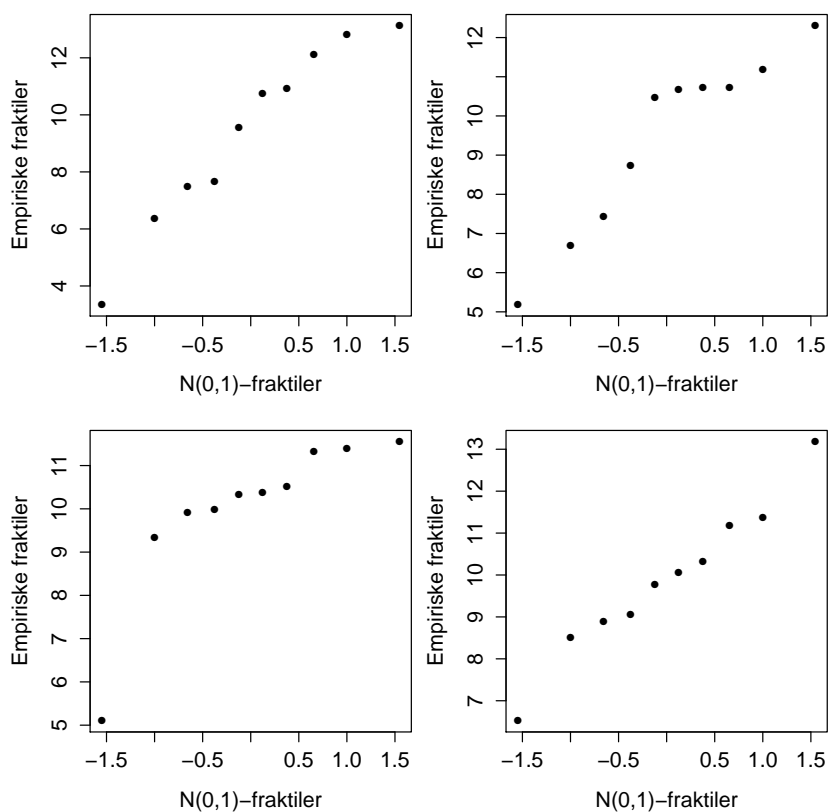
observation.

Konstruktionen af konfidensintervaller og udførelsen af hypotesetest var som for modellen med kendt varians. Teknisk set blev den kendte varians erstattet med et estimat, og fraktiler og sandsynligheder fra normalfordelingen blev erstattet med de tilsvarende størrelser fra en t -fordeling. Herved tages der hensyn til den ekstra usikkerhed om middelværdiestimatet forårsaget af den ukendte varians. Fortolkningen af estimater, estimatorers fordeling, konfidensintervaller og hypotesetest er helt ækvivalent med fortolkningerne fra situationen med kendt varians.

4.7 R

For en stikprøve med ukendt varians kan vi beregne estimater og konfidensintervaller samt udføre test af hypoteser om middelværdien “manuelt” på tilsvarende måde som vi gjorde det i tilfældet med kendt varians i afsnit 3.6. Alternativt kan vi lade funktionen `t.test` gøre arbejdet for os.

Vi illustrerer begge dele med data fra prothrombineksemplet, som er gemt i filen `prothrombin.txt`. Filen indeholder en linje med teksten `forskel` samt 40 linjer med de observerede forskelle. Vi indlæser data og har så adgang til variabelen med `$`-syntaksen. For at gøre kommandoerne nedenfor lidt simplere, definerer vi en variabel, `dif`, ”udenfor” datasættet, som indeholder forskellene. Denne variabel bruges nedenfor, men vi kunne lige så godt have skrevet `leverdata$forskel` alle vegne



Figur 4.6: QQ-plots for fire simulerede datasæt med 10 observationer fra $N(10,4)$ i hver.

(og det ville på mange måder være bedre programmeringsstil).

```
> leverdata <- read.table("prothrombin.txt", header=T)
> dif <- leverdata$forskel
> dif
 [1] 12.690540  1.079137 12.872840 -8.173629 13.299110
 [6] 17.925140  4.896743  5.885729 22.991870 23.847640
[11] 15.064980  2.586553 -10.217300 19.120460 21.150610
[16] 39.119960 -9.988888 32.043970 21.473210 -3.281252
[21] 66.244900 18.325120 47.469950 25.376020  9.282428
[26] 42.210290  9.562957 -8.341821 -13.350730  4.673018
```

```
[31] 35.265890 33.431440 10.861930 -20.428800 40.407760
[36] 62.998670 1.309416 35.535210 18.464130 8.314870
```

Analyse med `t.test` Det nemmeste er at bruge `t.test`:

```
> t.test(dif) # Analyse af variabelen dif

One Sample t-test

data: dif
t = 5.2681, df = 39, p-value = 5.357e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 10.19567 22.90433
sample estimates:
mean of x
 16.55000
```

Outputtet giver os næsten alt det vi har brug for: gennemsnittet, et 95% konfidensinterval, testet incl. værdien af t , antallet af frihedsgrader og p -værdien. Check selv at værdierne er de samme som i eksempel 4.6 (side 60), 4.8 (side 62) og 4.10 (side 66).

Outputtet giver ikke estimatet for variansen eller spredningen, men de kan beregnes ved hjælp af `var` og `sd`:

```
> var(dif) # s^2
[1] 394.7667
> sd(dif) # s
[1] 19.86874
```

Ved at ændre argumenter til `t.test` kan man ændre på konfidensgraden i konfidensintervallet og værdien i hypotesen. For eksempel ville nedenstående kommandoer føre til output med 90% konfidensinterval, henholdsvis output med test af hypotesen $H: \mu = 4$ (hvilket i prothrombineksemplet er en komplet uninteressant hypotese):

```
> t.test(dif, conf.level=0.90) # 90% KI
> t.test(dif, mu=4) # Test af H:mu=4
```

Ovenfor brugte vi *forskellen* mellem observationerne før og efter behandling som argument til `t.test` — det er jo den variabel vi har opstillet en model for. Hvis før-

og eftermålingerne var tilgængelige som to variable, `foer` og `efter`, kunne vi også have udført analysen med kommandoen

```
t.test(efter, foer, paired=TRUE) # Parret analyse
```

Det er helt essentielt at tilføje koden `paired=TRUE` for ellers opfatter R de to variable som uafhængige og laver analysen som en analyse af to uafhængige stikprøver, se kapitel 5.

Analyse med manuelle beregninger Til illustration viser vi nu hvordan beregningerne kunne foretages manuelt — det ville man normalt næppe gøre. Bemærk specielt funktionerne `qt` og `pt` der beregner fraktiler og sandsynligheder i t -fordelinger.

```
> ybar <- mean(dif)           # Gennemsnit
> s <- sd(dif)                # Estimeret spredning, s

> qt(0.975, df=39)           # 97.5% fraktil i t(39)
[1] 2.022691

> ybar - 2.0223 * s / sqrt(40) # Nedre grænse i 95% KI
[1] 10.1969
> ybar + 2.0223 * s / sqrt(40) # Øvre grænse i 95% KI
[1] 22.90310

> ybar / s * sqrt(40)         # t
[1] 5.268146
> 2*(1-pt(5.27, df=39))      # p-værdien
[1] 5.325305e-06
```

Modelkontrol Til sidst illustrerer vi hvordan figurerne til kontrol af normalfordelingsantagelsen kan laves. De følgende kommandoer laver tegningen til venstre i figur 4.3, bortset fra nogle layoutmæssige ting:

```
hist(dif, prob=T) # Hist. på ssh-skala
f <- function(x) dnorm(x, ybar, s) # Tætheden som funktion
plot(f, -30, 80, add=T) # Tilføj graf for f
```

Først laves selve histogrammet med `hist`. Koden `prob=T` sørger for at histogrammet kommer på “sandsynlighedsskala”, dvs. at det samlede areal under rektanglerne

er 1. Man kan styre inddelingen af x -aksen på forskellig vis, men defaultværdierne er ofte ganske gode. Anden linje definerer funktionen f der er tætheden for normalfordelingen med middelværdi og varians lig de estimerede værdier (se evt. afsnit 3.6). Til sidst tegnes grafen for denne funktion; koden `add=T` sørger for at grafen tegnes oven i det eksisterende plot i stedet for i en ny figur.

QQ-plot laves nemt ved hjælp af funktionen `qqnorm`. Den følgende kommando laver tegningen til højre i figur 4.3 (pånær layout):

```
qqnorm(dif) ## QQ-plot for variabelen dif
```

4.8 Opgaver

4.1 For at undersøge om der er forskel på visuel og auditiv reaktionshastighed hos mennesker målte man begge slags reaktionshastigheder hos 15 basketballspillere. Den visuelle reaktionstid blev målt som den tid der går før forsøgspersonen reagerer på et lyssignal, mens den auditive reaktionstid blev målt som den tid der går før forsøgspersonen reagerer på en bestemt lyd. Alle målinger er i millisekunder.


Spiller	Visuel	Auditiv	Forskel
1	161	157	4
2	203	207	-4
3	235	198	37
4	176	161	15
5	201	234	-33
6	188	197	-9
7	228	180	48
8	211	165	46
9	191	202	-11
10	178	193	-15
11	159	173	-14
12	227	187	40
13	193	182	11
14	192	159	33
15	212	186	26
\bar{y}	197	185.4	11.6
s	23.11	20.99	25.67

I denne opgave skal du analysere data hørende til visuel og auditiv reaktionstid hver for sig.

1. Opstil en statistisk model for data svarende til visuel reaktionshastighed, og angiv estimatorne for parametrene i modellen.
2. Angiv den teoretiske fordeling af de tilhørende estimatorer og den estimerede spredning for middelværdiestimatet.
3. Beregn et 95% og et 90% konfidensinterval for den forventede visuelle reaktionstid. Du kan benytte at 97.5% fraktilen i t_{14} -fordelingen er 2.145, mens 95% fraktilen er 1.761.
4. Beregn tilsvarende et 95% og et 90% konfidensinterval for den forventede auditive reaktionstid.

4.2 I denne opgave skal du bruge data fra opgave 4.1 til at undersøge om der er forskel på visuel og auditiv reaktionstid.

1. Opstil en statistisk model der kan bruges til dette formål. *Vink:* hvilken variabel skal du analysere?
2. Angiv estimator for parametrene i modellen samt estimatorernes fordeling. Bestem også den estimerede spredning for middelværdiestimatoren.
3. Test hypotesen om at der ikke er forskel på de to slags reaktionstider. Du kan benytte at $P(T \leq 1.75) = 0.949$ hvis $T \sim t_{14}$.
4. Beregn et 95% konfidensinterval for den gennemsnitlige forskel mellem visuel og auditiv reaktionstid for en tilfældig spiller. Du kan benytte at 97.5% fraktilen i t_{14} -fordelingen er 2.145.

4.3  I denne opgave skal du bruge R til at udføre analysen fra opgave 4.2. Data ligger i filen `reaktionstid.txt`.

1. Indlæs data til et datasæt i R, kald det fx `reaktionData`. Brug derefter kommandoen `t.test(reaktionData$forskel)`, og check at R giver dig de samme resultater som du fik da du regnede det igennem i hånden.
2. Hvilken variabel antages at være normalfordelt? Lav det relevante QQ-plot, og vurdér om normalfordelingsantagelsen er rimelig (men vær opmærksom på at det er vanskeligt at vurdere når der er få observationer).

3. Tegn også QQ-plot for variablene `visuel` og `auditiv`. Ser de ud til at være normalfordelte?


4.4 Ved studentereksamen i 2002 i skriftlig dansk udførtes et såkaldt standardforsøg hvor elever i forsøgsklasserne på forhånd blev sat sammen i grupper der diskuterede opgaverne i en time før den egentlige eksamen. Datamaterialet omfatter alle grupper fra forsøgsklasserne med netop tre deltagere. I tabellen nedenfor er angivet gennemsnittet for gruppens medlemmer for henholdsvis årskarakterer i 3.g og eksamenskarakterer i skriftlig dansk.

Data er stillet til rådighed af Marianne Hansen, Haslev Gymnasium og HF, og ligger i filen `skrdansk.txt`.

Gruppe	Årskarakter	Eksamenskarakter	Forskel
1	7.67	5.67	2.00
2	9.33	7.67	1.67
3	8.67	8.33	0.33
4	9.67	8.33	1.33
5	7.33	7.00	0.33
6	8.67	8.33	0.33
7	7.33	7.33	0.00
8	8.00	8.33	-0.33
9	8.33	7.00	1.33
10	9.00	7.67	1.33
11	7.33	7.00	0.33
12	7.67	6.33	1.33
13	8.67	9.33	0.67
14	6.33	5.33	1.00
15	8.00	7.67	0.33
16	8.00	8.00	0.00
17	9.67	9.00	0.67
\bar{y}	8.216	7.548	0.744
s	0.914	1.086	0.662

- Opstil en statistisk model der kan bruges til at undersøge om der er forskel på årskarakterer og eksamenskarakterer.
- Angiv estimater for parametrene i modellen. Angiv også estimatorernes fordelinger og den estimerede spredning for middelværdiestimatoren.

3. Undersøg om der er niveauforskel mellem de to slags karakterer.

4.5  Denne opgave handler om analyse af malariadata fra eksempel 4.14. Data ligger i filen `malaria.txt`; variabelen med antallet af parasitter hedder `parasites`.

1. Indlæs data til et datasæt i R, kald det fx `malariaData`. Konstruér en variabel `logparasites` der indeholder den naturlige logaritme til parasitantallene. Brug `log`-funktionen. Brug `qqnorm` til at lave QQ-plots for `parasites` og for `logparasites`.
2. Opstil en statistisk model der kan bruges til at beskrive data. *Vink*: Hvilken variabel kan du lave en model for?
3. Angiv et estimat og et konfidensinterval for den forventede værdi af logaritmen til parasittallet for børn inficeret med malaria. Brug funktionen `t.test`. Angiv også fordelingen af estimatorerne.
4. Angiv et estimat for medianen af parasittallet for børn inficeret med malaria. *Vink*: Er der forskel på middelværdi og median i en normalfordeling? Hvad sker der med medianen ved transformation med en voksende funktion?
5. Forklar hvorfor outputtet fra `t.test(malariaData$parasites)` ikke bør benyttes til analyse af disse data.

Kapitel 5

To stikprøver

I kapitel 3 og 4 diskuterede vi normalfordelingsmodellen for en enkelt stikprøve med kendt og ukendt varians. Mange statistiske undersøgelser er dog sammenlignende, hvor man ønsker at sammenligne to eller flere grupper. Det kan for eksempel være sammenligninger af forskellige behandlinger eller en behandling overfor ingen behandling, sammenligninger af forskellige produkter, eller sammenligninger af forskellige investeringsstrategier. Vi vil i dette kursus kun se på sammenligninger af to grupper. I dette kapitel er der begrebsmæssigt intet nyt i forhold til kapitel 4, vi regner blot på en udvidet model.

5.1 Statistisk model

Udgangspunktet er stadig uafhængige og normalfordelte stokastiske variable, men vi har nu to grupper af observationer: x_1, \dots, x_{n_1} og y_1, \dots, y_{n_2} . Observationerne er realisationer af de stokastiske variable X_1, \dots, X_{n_1} med middelværdi μ_1 og varians σ^2 , og Y_1, \dots, Y_{n_2} med middelværdi μ_2 og varians σ^2 . Observationerne er identisk fordelte indenfor gruppen, men hver gruppe har sin egen middelværdi, og vi antager at der er samme varians i begge grupper.

Det er en antagelse i modellen at varianserne er ens. Dette kaldes varianshomogenitet, og det bør altid kontrolleres om denne antagelse er fornuftig. Vi vil dog ikke teste for varianshomogenitet i dette kursus, men i det mindste vil vi grafisk vurdere om varianserne med rimelighed kan antages at være ens. Man kan også analysere en model hvor varianserne er forskellige, men det vil vi ikke komme ind på her.

Bemærk at der ikke nødvendigvis er samme antal observationer i hver stikprøve. Der er n_1 observationer i den første gruppe og n_2 observationer i den anden gruppe, og således $n = n_1 + n_2$ observationer i alt. Den simultane fordeling af alle X_i 'erne og Y_j 'erne betegnes $N_{\mu_1, \mu_2, \sigma^2}^{n_1, n_2}$ og har tæthed

$$\begin{aligned} f_{\mu_1, \mu_2, \sigma^2}(x, y) &= \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_1)^2\right) \prod_{j=1}^{n_2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_j - \mu_2)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{j=1}^{n_2}(y_j - \mu_2)^2\right)\right), \end{aligned} \quad (5.1)$$

hvor $x = (x_1, \dots, x_{n_1}) \in \mathbb{R}^{n_1}$ og $y = (y_1, \dots, y_{n_2}) \in \mathbb{R}^{n_2}$. Notationen $f_{\mu_1, \mu_2, \sigma^2}$ understreger at parameteren (μ_1, μ_2, σ^2) er tredimensional. Vi antager at parameterområdet er $\Theta = \mathbb{R} \times \mathbb{R} \times (0, \infty)$, men det kunne også være en delmængde af denne mængde.

Definition 5.1. Modellen for to stikprøver med samme varians består af udfaldsrummet \mathbb{R}^n samt familien

$$\mathcal{P} = \left\{ N_{\mu_1, \mu_2, \sigma^2}^{n_1, n_2} : (\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty) \right\}$$

af fordelinger på \mathbb{R}^n hvor $N_{\mu_1, \mu_2, \sigma^2}^{n_1, n_2}$ har tæthed (5.1).

Alternativ formulering: Lad X_1, \dots, X_{n_1} og Y_1, \dots, Y_{n_2} være uafhængige normalfordelte stokastiske variable, hvor $X_i \sim N(\mu_1, \sigma^2)$ og $Y_j \sim N(\mu_2, \sigma^2)$, og hvor $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}$ og $\sigma^2 > 0$ er ukendte parametre.

Bemærk forskellen fra situationen med parrede observationer i afsnit 4. I eksempel 4.2 om prothrombinindeks er der to observationer for hver person, nemlig målinger før og efter behandling. Man kan næppe antage at observationer fra samme person er uafhængige så disse data passer ikke ind i modellen fra definition 5.1. I stedet analyserer man differenserne som en enkelt stikprøve og udfører et parret t -test.

Eksempel 5.2. (*Tuberkulosevaccine*) For at sammenligne BCG-vaccine (mod tuberkulose) fra to forskellige produktionscentre har man vaccineret grupper af skolebørn med vaccinerne og undersøgt deres reaktioner. Tuberkulinreaktionen måles 3 dage efter indsprøjtning af 5 tuberkulinenheder ved at måle diameteren i mm af det hævede område omkring indsprøjtningstedet. Data består af 130 målinger fra Statens Seruminstitut i København, x_1, \dots, x_{130} , og 116 målinger fra Nationalforeningens

BCG-laboratorium i Oslo, y_1, \dots, y_{116} . Observationerne betragtes som realisationer af X_1, \dots, X_{130} og Y_1, \dots, Y_{116} som antages at være uafhængige og normalfordelte med varians σ^2 og middelværdier μ_1 henholdsvis μ_2 . \square

5.2 Maksimum likelihood estimation

Vi skal estimere (μ_1, μ_2, σ^2) på basis af samtlige data, dvs. $x = (x_1, \dots, x_{n_1})$ og $y = (y_1, \dots, y_{n_2})$. Vi definerer igen likelihoodfunktionen som tætheden, opfattet som funktion af parameteren,

$$\begin{aligned} L_{x,y} &: \mathbb{R} \times \mathbb{R} \times (0, \infty) \rightarrow \mathbb{R} \\ L_{x,y}(\mu_1, \mu_2, \sigma^2) &= f_{\mu_1, \mu_2, \sigma^2}(x, y) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n_1} (x_i - \mu_1)^2 + \sum_{j=1}^{n_2} (y_j - \mu_2)^2 \right)\right). \end{aligned}$$

Et maksimum likelihood estimat for $(\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$ opfylder

$$L_{x,y}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) \geq L_{x,y}(\mu_1, \mu_2, \sigma^2), \quad (\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty).$$

Nedenfor benyttes notationen $SSD_x = \sum_{i=1}^{n_1} (x_i - \bar{x})^2$ og $SSD_y = \sum_{j=1}^{n_2} (y_j - \bar{y})^2$. De tilsvarende stokastiske variable er $SSD_X = \sum_{i=1}^{n_1} (X_i - \bar{X})^2$ og $SSD_Y = \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2$.

Sætning 5.3 giver os maksimum likelihood estimatorerne samt fordelingen af de tilhørende estimatorer. Ligesom for en enkelt stikprøve korrigerer vi senere variansestimateret så den tilhørende estimator er central, se bemærkning 5.4.

Sætning 5.3. For den statistiske model fra definition 5.1 er maksimum likelihood estimatet for (μ_1, μ_2, σ^2) entydigt bestemt og givet ved

$$\hat{\mu}_1 = \bar{x}, \quad \hat{\mu}_2 = \bar{y}, \quad \hat{\sigma}^2 = \frac{1}{n} (SSD_x + SSD_y).$$

Estimatorerne $\hat{\mu}_1 = \bar{X}$, $\hat{\mu}_2 = \bar{Y}$ og $\hat{\sigma}^2 = \frac{1}{n} (SSD_X + SSD_Y)$ er uafhængige, og deres marginale fordelinger er

$$\hat{\mu}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), \quad \hat{\mu}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right), \quad \hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi_{n-2}^2.$$

Bevis Vi skal maksimere en funktion af tre variable og benytter et profileringsargument, see appendiks B. Betragt først en fast positiv værdi af σ^2 . Funktionen

$$(\mu_1, \mu_2) \rightarrow L_{x,y}(\mu_1, \mu_2, \sigma^2)$$

splitter op i et produkt af to funktioner, en der kun afhænger af μ_1 og en der kun afhænger af μ_2 . På nær konstanter er disse funktioner identiske med likelihoodfunktionen for modellen med kendt varians, så det følger af sætning 3.3 at de har entydigt maksimum for $\mu_1 = \bar{x}$ og $\mu_2 = \bar{y}$.

Dette gælder for alle $\sigma^2 > 0$, dvs.

$$L_{x,y}(\bar{x}, \bar{y}, \sigma^2) \geq L_{x,y}(\mu_1, \mu_2, \sigma^2), \quad \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}, \sigma^2 > 0.$$

Vi betragter derfor profillikelihoodfunktionen $\tilde{L}_{x,y} : (0, \infty) \rightarrow \mathbb{R}$ for σ^2 defineret ved

$$\tilde{L}_{x,y}(\sigma^2) = L_{x,y}(\bar{x}, \bar{y}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\text{SSD}_x + \text{SSD}_y)\right).$$

Vi kan igen benytte lemma 4.4 — med $x = \sigma^2$, $a = (\text{SSD}_x + \text{SSD}_y)/2$ og $b = n/2$ — til at indse at \tilde{L} har maksimum for $\sigma^2 = \frac{1}{n}(\text{SSD}_x + \text{SSD}_y)$. Vi har således vist at

$$\begin{aligned} L_{x,y}\left(\bar{x}, \bar{y}, \frac{1}{n}(\text{SSD}_x + \text{SSD}_y)\right) &= \tilde{L}_{x,y}\left(\frac{1}{n}(\text{SSD}_x + \text{SSD}_y)\right) \\ &\geq \tilde{L}_{x,y}(\sigma^2) \\ &= L_{x,y}(\bar{x}, \bar{y}, \sigma^2) \\ &\geq L_{x,y}(\mu_1, \mu_2, \sigma^2) \end{aligned}$$

for alle $\mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}$ og $\sigma^2 > 0$ så $(\bar{x}, \bar{y}, \frac{1}{n}(\text{SSD}_x + \text{SSD}_y))$ er entydigt maksimumpunkt for $L_{x,y}$.

Resultatet vedrørende fordelingen af $(\bar{X}, \bar{Y}, \hat{\sigma}^2)$ beviser vi i flere trin. Fra sætning A.5 i appendix A benyttet på X_i 'erne og Y_j 'erne hver for sig ved vi at

$$\hat{\mu}_1 \sim N\left(\mu_1, \frac{\sigma^2}{n_1}\right), \text{SSD}_X \sim \sigma^2 \chi_{n_1-1}^2, \quad \hat{\mu}_2 \sim N\left(\mu_2, \frac{\sigma^2}{n_2}\right), \text{SSD}_Y \sim \sigma^2 \chi_{n_2-1}^2.$$

Da \bar{X} og SSD_X kun afhænger af X_1, \dots, X_{n_1} og \bar{Y} og SSD_Y kun afhænger af Y_1, \dots, Y_{n_2} , er (\bar{X}, SSD_X) og (\bar{Y}, SSD_Y) uafhængige. Det følger af en generalisering af sætning A.3 i appendix A til funktioner $g : \mathbb{R}^{n_1} \rightarrow \mathbb{R}^2$ og $h : \mathbb{R}^{n_2} \rightarrow \mathbb{R}^2$. Sætning A.5 fortæller at $\hat{\mu}_1$ og SSD_X er uafhængige, og at $\hat{\mu}_2$ og SSD_Y er uafhængige, og det følger endelig fra sætning A.3 benyttet på disse fire variable at at $\hat{\mu}_1, \hat{\mu}_2$ og $\hat{\sigma}^2$ er uafhængige.

Det følger umiddelbart af definitionen af χ^2 -fordelingen (BH, definition 10.4.1) at summen af to uafhængige χ^2 -fordelte variable er χ^2 -fordelt med antal frihedsgrader lig summen af frihedsgrader. Derfor er

$$n\hat{\sigma}^2 = \text{SSD}_X + \text{SSD}_Y \sim \sigma^2 \chi_{n_1-1+n_2-1}^2 = \sigma^2 \chi_{n-2}^2,$$

da χ_k^2 jo er en Γ -fordeling med formparameter $k/2$ og skalaparameter 2. □

Bemærk at $E(\bar{X}) = \mu_1$ og $E(\bar{Y}) = \mu_2$ således at \bar{X} og \bar{Y} er centrale estimatorer for μ_1 og μ_2 . Derimod er

$$E(\hat{\sigma}^2) = \frac{n-2}{n} \sigma^2$$

så $\hat{\sigma}^2$ er ikke en central estimator for σ^2 , på samme måde som vi så i forrige kapitel, idet middelværdien i χ_k^2 -fordelingen er k således at $E(n\hat{\sigma}^2) = (n-2)\sigma^2$. I gennemsnit estimeres σ^2 altså for lavt hvis vi benytter maksimum likelihood estimatoren. Det er imidlertid nemt at korrigere $\hat{\sigma}^2$ og opnå et centralt estimat: vi skal blot normere med $n-2$ i stedet for n i definitionen af $\hat{\sigma}^2$, og i stedet bruge

$$\tilde{\sigma}^2 = \frac{1}{n-2} \left(\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2 \right) = \frac{1}{n-2} (\text{SSD}_X + \text{SSD}_Y)$$

som estimator. Så er $(n-2)\tilde{\sigma}^2 \sim \sigma^2 \chi_{n-2}^2$, og specielt er $E(\tilde{\sigma}^2) = \sigma^2$ som ønsket. Læg dog mærke til at jo større n er, jo mindre betyder korrektionen af variansestimateret. Det tilsvarende estimat, hvor observationerne sættes ind, betegnes som regel s^2 ,

$$\begin{aligned} s^2 &= \frac{1}{n-2} \left(\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right) \\ &= \frac{\text{SSD}_x + \text{SSD}_y}{n-2} \\ &= \frac{(n_1-1)s_x^2 + (n_2-1)s_y^2}{n-2}, \end{aligned} \tag{5.2}$$

hvor s_x^2 og s_y^2 er de empiriske varianser for de to stikprøver. Det er dette estimat man benytter. Bemærk at dette variansestimater har samme struktur som (4.4): den totale SSD-størrelse divideret med en konstant som er antallet af observationer minus antallet af middelværdiparametre, der i dette tilfælde er to: μ_1 og μ_2 . Den totale SSD-størrelse er kvadratsummen af observationerne minus deres estimerede middelværdi. Lad os samle resultatet vedrørende estimatorerne i en bemærkning:

Bemærkning 5.4. I den statistiske model fra definition 5.1 bruger vi estimererne $\hat{\mu}_1 = \bar{x}$, $\hat{\mu}_2 = \bar{y}$, og $s^2 = \frac{1}{n-2} (\text{SSD}_x + \text{SSD}_y)$.

Den sande eller teoretiske fordeling af estimatorerne \bar{X} , \bar{Y} og $\bar{\sigma}^2$ er givet i sætning 5.3, men afhænger af de ukendte parametre. Hvis vi erstatter den sande spredning σ med estimatet s i spredningerne for $\hat{\mu}_1$ og $\hat{\mu}_2$, får vi de estimerede spredninger (standard errors):

$$\text{SE}(\hat{\mu}_1) = \frac{s}{\sqrt{n_1}}, \quad \text{SE}(\hat{\mu}_2) = \frac{s}{\sqrt{n_2}}.$$

Eksempel 5.5. (Tuberkulosevaccine, fortsættelse af eksempel 5.2, side 82) For de 130 observationer x_1, \dots, x_{130} af turberkulinreaktioner fra København og de 116 observationer y_1, \dots, y_{116} fra Oslo viste det sig at

$$\begin{aligned} \bar{x} &= 17.13; & \frac{1}{129} \sum_{i=1}^{130} (x_i - \bar{x})^2 &= 11.03 \\ \bar{y} &= 16.84; & \frac{1}{115} \sum_{j=1}^{116} (y_j - \bar{y})^2 &= 12.66 \end{aligned}$$

således at estimererne er

$$\hat{\mu}_1 = 17.13; \quad \hat{\mu}_2 = 16.84$$

og

$$s^2 = \frac{129 \cdot 11.03 + 115 \cdot 12.66}{130 + 116 - 2} = 11.80, \quad s = 3.43.$$

Estimatorerne er uafhængige, $\hat{\mu}_1 \sim N(\mu_1, \sigma^2/130)$, $\hat{\mu}_2 \sim N(\mu_2, \sigma^2/116)$ og $\bar{\sigma}^2 \sim \frac{\sigma^2}{244} \chi_{244}^2$. De estimerede spredninger (standard errors) for middelværdiestimatorerne er

$$\text{SE}(\hat{\mu}_1) = \frac{s}{\sqrt{130}} = 0.301, \quad \text{SE}(\hat{\mu}_2) = \frac{s}{\sqrt{116}} = 0.319.$$

Bemærk at præcisionen af estimererne er forskellig for Oslo og København. Dette skyldes at antallet af observationer er forskellige i de to grupper. \square

5.3 Konfidensintervaller

I afsnit 4.3 udledte vi konfidensintervallet

$$\bar{Y} \pm t_{n-1, 1-\alpha/2} \frac{\bar{\sigma}}{\sqrt{n}}$$

for middelværdien i en enkelt stikprøve. På samme måde kan vi lave konfidensintervaller for de to middelværdier i to stikprøver. Forskellen er at vi nu bruger begge stikprøver til at estimere den fælles varians, der derfor er bedre bestemt fordi estimatet bygger på flere observationer.

Sætning 5.6. *Betragt den statistiske model fra definition 5.1. Så er*

$$\bar{X} \pm t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_1}} = \left(\bar{X} - t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_1}}, \bar{X} + t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_1}} \right) \quad (5.3)$$

et $1 - \alpha$ konfidensinterval for μ_1 , og

$$\bar{Y} \pm t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_2}} = \left(\bar{Y} - t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_2}}, \bar{Y} + t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_2}} \right) \quad (5.4)$$

er et $1 - \alpha$ konfidensinterval for μ_2 .

Bevis Vi beviser kun (5.3), da (5.4) bevises på samme måde. Det følger af sætning 5.3 at $U = (\bar{X} - \mu_1)/(\sigma/\sqrt{n_1})$ er standard normalfordelt, at $Z = (SSD_X + SSD_Y)/\sigma^2 \sim \chi_{n-2}^2$, og at U og Z er uafhængige. Det følger da af definitionen af t -fordelingen (definition A.4 i appendiks A), at

$$T = \frac{U}{\sqrt{Z/(n-2)}} = \frac{\sqrt{n_1}(\bar{X} - \mu_1)}{\tilde{\sigma}}$$

er t -fordelt med $n - 2$ frihedsgrader. Således er

$$P\left(-t_{n-2,1-\alpha/2} < \frac{\sqrt{n_1}(\bar{X} - \mu_1)}{\tilde{\sigma}} < t_{n-2,1-\alpha/2}\right) = 1 - \alpha$$

eller, hvis vi isolerer μ_1 i midten,

$$P\left(\bar{X} - t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_1}} < \mu_1 < \bar{X} + t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_1}}\right) = 1 - \alpha. \quad (5.5)$$

Dette viser at $\bar{X} \pm t_{n-2,1-\alpha/2} \frac{\tilde{\sigma}}{\sqrt{n_1}}$ er et konfidensinterval for μ_1 med konfidensgrad $1 - \alpha$. \square

Bemærk at for en givet værdi af $\tilde{\sigma}$ er dette konfidensinterval smallere end hvis man kun havde benyttet den ene stikprøve til at estimere variansen, idet t_{n-2} -fordelingsfraktilen altid er mindre end t_{n_1-1} -fordelingsfraktilen fordi $n > n_1$. Dette giver god

mening: vi har bestemt σ^2 mere præcist og der er således mindre usikkerhed om estimatet på μ_1 .

Når man har to stikprøver er man ofte interesseret i forskellen mellem deres middelværdier, og det er derfor interessant at have estimat, estimeret spredning og konfidensinterval for forskellen i middelværdier. Det synes naturligt at bruge forskellen mellem de middelværdiestimater som estimat for forskellen mellem middelværdierne, dvs.

$$\widehat{\mu_1 - \mu_2} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{x} - \bar{y}.$$

Da \bar{X} og \bar{Y} er uafhængige og er fordelt som angivet i sætning 5.3, følger det af BH, eksempel 6.6.3 at

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right). \quad (5.6)$$

Ved at erstatte σ med estimatet s , fås den estimerede spredning for estimatoren

$$\text{SE}\left(\widehat{\mu_1 - \mu_2}\right) = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Sætning 5.7. *Betragt den statistiske model fra definition 5.1. Så er*

$$\begin{aligned} \bar{X} - \bar{Y} \pm t_{n-2, 1-\alpha/2} \tilde{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \\ \left(\bar{X} - \bar{Y} - t_{n-2, 1-\alpha/2} \tilde{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X} - \bar{Y} + t_{n-2, 1-\alpha/2} \tilde{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \end{aligned}$$

et $1 - \alpha$ konfidensinterval for $\mu_1 - \mu_2$.

Bevis Fra fordelingen i (5.6) ser vi at vi kan betragte konfidensintervallet for $\mu_1 - \mu_2$ som et konfidensinterval for middelværdien af en normalfordelt variabel med ukendt varians, og derfor benytte samme argumenter som i beviset for sætning 4.7. Definer de stokastiske variable

$$\begin{aligned} U &= \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ Z &= \frac{1}{\sigma^2} (\text{SSD}_X + \text{SSD}_Y). \end{aligned}$$

Så er U standard normalfordelt, Z er χ_{n-2}^2 fordelt og U og Z er uafhængige. Vi har da fra definitionen af en t -fordeling (definition A.4 i appendiks A), at

$$T = \frac{U}{\sqrt{Z/(n-2)}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\tilde{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

er t -fordelt med $n - 2$ frihedsgrader. Således er

$$P\left(-t_{n-2,1-\alpha/2} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\tilde{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{n-2,1-\alpha/2}\right) = 1 - \alpha.$$

Ved at isolere $\mu_1 - \mu_2$ får vi at

$$\bar{X} - \bar{Y} \pm t_{n-2,1-\alpha/2} \tilde{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

er et konfidensinterval for $\mu_1 - \mu_2$ med konfidensgrad $1 - \alpha$. \square

Diskussionerne fra afsnit 3.3 og 4.3 vedrørende konfidensintervaller er stadig gyldige.

Eksempel 5.8. (*Tuberkulosevaccine, fortsættelse af eksempel 5.2, side 82*) Husk at

$$n_1 = 130, \quad n_2 = 116, \quad \bar{x} = 17.13, \quad \bar{y} = 16.84, \quad s = 3.43.$$

Desuden er 97.5% fraktilen i t -fordelingen med 244 frihedsgrader lig 1.97, således at

$$17.13 \pm 1.97 \cdot \frac{3.43}{\sqrt{130}} = 17.13 \pm 0.59 = (16.54, 17.72)$$

$$16.84 \pm 1.97 \cdot \frac{3.43}{\sqrt{116}} = 16.84 \pm 0.63 = (16.21, 17.46)$$

er 95% konfidensintervaller for μ_1 og μ_2 . Vi får følgende 95% konfidensinterval for $\mu_1 - \mu_2$:

$$17.13 - 16.84 \pm 1.97 \cdot 3.43 \sqrt{\frac{1}{130} + \frac{1}{116}} = 0.29 \pm 0.86 = (-0.57, 1.16).$$

Bemærk at konfidensintervallet for μ_1 indeholder punktestimatet for μ_2 , og at konfidensintervallet for μ_2 indeholder punktestimatet for μ_1 . Intuitivt passer det med at konfidensintervallet for forskellen indeholder nul. Hvis $\mu_1 - \mu_2 = 0$ — svarende til at der ikke er forskel mellem de to grupper — er det således ikke usandsynligt at vi skulle have observeret de data vi har til rådighed. \square

5.4 Hypotesetest

Vi vil nu betragte hypotesen om at middelværdien er den samme i de to grupper. Det er det samme som hypotesen $H : \mu_1 - \mu_2 = 0$ om at forskellen i middelværdier

mellem de to grupper er nul. Der er ingen restriktioner på variansen, og vi kan skrive hypotesen som

$$H : \mu_1 = \mu_2 = \mu$$

eller

$$H : (\mu_1, \mu_2, \sigma^2) \in \Theta_0 = \{(\mu_1, \mu_2) \in \mathbb{R}^2 \mid \mu_1 = \mu_2\} \times (0, \infty).$$

Ligesom i afsnit 4.4 er hypotesen ikke en simpel hypotese da parametermængden under hypotesen, Θ_0 , indeholder mere end et enkelt punkt. Vi vil gøre følgende:

- Estimere (μ, σ^2) under hypotesen, dvs. bestemme $(\hat{\mu}, \hat{\sigma}^2) \in \mathbb{R} \times (0, \infty)$ så

$$L_{x,y}(\hat{\mu}, \hat{\mu}, \hat{\sigma}^2) \geq L_{x,y}(\mu_1, \mu_2, \sigma^2), \quad (\mu_1, \mu_2, \sigma^2) \in \Theta_0.$$

- Opskrive kvotientteststørrelsen

$$Q(x, y) = \frac{L_{x,y}(\hat{\mu}, \hat{\mu}, \hat{\sigma}^2)}{L_{x,y}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)}.$$

- Bestemme p -værdien eller testsandsynligheden

$$\varepsilon(x, y) = P(Q(X, Y) \leq Q(x, y)).$$

- Afvise hypotesen hvis $\varepsilon(x, y) \leq \alpha$ for et på forhånd fastsat signifikansniveau og i givet fald konkludere at μ_1 er signifikant forskellig fra μ_2 .

Sætning 5.9. *Betragt den statistiske model givet i definition 5.1 og hypotesen $H : \mu_1 = \mu_2 = \mu$. Under hypotesen er maksimum likelihood estimatet $(\hat{\mu}, \hat{\sigma}^2)$ givet ved*

$$\hat{\mu} = \frac{1}{n} \left(\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j \right), \quad \hat{\sigma}^2 = \frac{1}{n} \left(\sum_{i=1}^{n_1} (x_i - \hat{\mu})^2 + \sum_{j=1}^{n_2} (y_j - \hat{\mu})^2 \right).$$

Fordelingerne af de tilsvarende stokastiske variable er $\hat{\mu} \sim N(\mu, \sigma^2/n)$ og $\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi_{n-1}^2$, og de er uafhængige. Kvotientteststørrelsen er givet ved

$$Q(x, y) = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2} \right)^{n/2}$$

og kvotienttestet kan udføres på

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

p -værdien er givet ved

$$\varepsilon(x, y) = 2P(T \geq |t|) = 2 \cdot (1 - F_{t_{n-2}}(|t|))$$

hvor T er t -fordelt med $n - 2$ frihedsgrader og $F_{t_{n-2}}$ er fordelingsfunktionen for t_{n-2} -fordelingen.

Inden vi beviser sætningen, bemærk da at $\hat{\mu}$ blot er gennemsnittet af *alle* målingerne (fra begge grupper), og at $n\hat{\sigma}^2$ er kvadratafvigelsessummen af alle målingerne, en størrelse som vi nedenfor vil betegne $SSD_{x,y}$.

Bevis Under hypotesen har vi model 4.1 for en enkelt stikprøve med ukendt varians, og får derfor direkte fra sætning 4.3 estimatorerne og deres fordeling.

Vi regner derefter på kvotientteststørrelsen $Q(x, y)$. Bemærk først at

$$-\frac{1}{2\hat{\sigma}^2} \left(\sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (y_j - \hat{\mu}_2)^2 \right) = -\frac{n}{2}$$

og at

$$-\frac{1}{2\hat{\hat{\sigma}}^2} \left(\sum_{i=1}^{n_1} (x_i - \hat{\mu})^2 + \sum_{j=1}^{n_2} (y_j - \hat{\mu})^2 \right) = -\frac{n}{2}$$

således at eksponentialleddene i tælleren og nævneren af $Q(x, y)$ er ens. Vi får således

$$Q(x, y) = \frac{L_{x,y}(\hat{\mu}, \hat{\mu}, \hat{\sigma}^2)}{L_{x,y}(\hat{\mu}_1, \hat{\mu}_2, \hat{\hat{\sigma}}^2)} = \left(\frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} \right)^{n/2}$$

som er en af påstandene i sætningen.

Det følger umiddelbart af definitionen af $\hat{\sigma}^2$ og $\hat{\hat{\sigma}}^2$ at

$$(Q(x, y))^{2/n} = \frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} = \frac{SSD_x + SSD_y}{SSD_{x,y}}, \quad (5.7)$$

hvor $SSD_{x,y}$ er kvadratafvigelsessummen for hele datasættet. Det meste af resten af beviset går ud på at vise at

$$(Q(x, y))^{2/n} = \left(1 + \frac{t^2}{n-2} \right)^{-1}. \quad (5.8)$$

Vi regner i første omgang på den totale kvadratafvigelsessum, $SSD_{x,y}$:

$$\begin{aligned}
 SSD_{x,y} &= \sum_{i=1}^{n_1} (x_i - \hat{\mu})^2 + \sum_{j=1}^{n_2} (y_j - \hat{\mu})^2 \\
 &= \sum_{i=1}^{n_1} (x_i - \bar{x} + \bar{x} - \hat{\mu})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y} + \bar{y} - \hat{\mu})^2 \\
 &= \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_1} (\bar{x} - \hat{\mu})^2 + 2(\bar{x} - \hat{\mu}) \underbrace{\sum_{i=1}^{n_1} (x_i - \bar{x})}_{=0} \\
 &\quad + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 + \sum_{j=1}^{n_2} (\bar{y} - \hat{\mu})^2 + 2(\bar{y} - \hat{\mu}) \underbrace{\sum_{j=1}^{n_2} (y_j - \bar{y})}_{=0} \\
 &= SSD_x + SSD_y + n_1(\bar{x} - \hat{\mu})^2 + n_2(\bar{y} - \hat{\mu})^2. \tag{5.9}
 \end{aligned}$$

Det totale gennemsnit $\hat{\mu}$ er et vægtet gennemsnit af \bar{x} og \bar{y} ,

$$\hat{\mu} = \frac{1}{n} \left(\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j \right) = \frac{1}{n} (n_1 \bar{x} + n_2 \bar{y}).$$

Hvis vi samtidig benytter at $n_1 + n_2 = n$, så får vi

$$\begin{aligned}
 n_1(\bar{x} - \hat{\mu})^2 + n_2(\bar{y} - \hat{\mu})^2 &= n_1 \left(\bar{x} - \frac{n_1 \bar{x} + n_2 \bar{y}}{n} \right)^2 + n_2 \left(\bar{y} - \frac{n_1 \bar{x} + n_2 \bar{y}}{n} \right)^2 \\
 &= n_1 \left(\left(1 - \frac{n_1}{n}\right) \bar{x} - \frac{n_2}{n} \bar{y} \right)^2 + n_2 \left(-\frac{n_1}{n} \bar{x} + \left(1 - \frac{n_2}{n}\right) \bar{y} \right)^2 \\
 &= \frac{n_1 n_2^2}{n^2} (\bar{x} - \bar{y})^2 + \frac{n_1^2 n_2}{n^2} (\bar{y} - \bar{x})^2 \\
 &= \frac{n_1 n_2}{n} (\bar{x} - \bar{y})^2 \\
 &= \frac{(\bar{x} - \bar{y})^2}{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{5.10}
 \end{aligned}$$

I sidste lighedstegn har vi benyttet at $n_1 n_2 / n = (1/n_1 + 1/n_2)^{-1}$.

Lad os nu indføre størrelserne

$$u = \frac{\bar{x} - \bar{y}}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad z = \frac{1}{\sigma^2} (SSD_x + SSD_y) = (n-2) \frac{s^2}{\sigma^2}.$$

Bemærk at den eneste forskel på u og t er at der er divideret med σ (deterministisk, men ukendt) henholdsvis s (kendt udfald af en stokastisk variabel). Forholdet mellem s^2 og σ^2 er givet ved $z/(n-2)$ således at

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{u}{\sqrt{z/(n-2)}}.$$

Lad os samle stumperne fra beviset sammen. Tilsammen giver (5.9), (5.10) og definitionen af z og u at

$$\text{SSD}_{x,y} = \text{SSD}_x + \text{SSD}_y + \frac{(\bar{x} - \bar{y})^2}{\frac{1}{n_1} + \frac{1}{n_2}} = \sigma^2 z + \sigma^2 u^2,$$

og ved indsættelse i (5.7) får vi

$$(Q(x,y))^{2/n} = \frac{\sigma^2 z}{\sigma^2 z + \sigma^2 u^2} = \left(1 + \frac{u^2}{z}\right)^{-1} = \left(1 + \frac{t^2}{n-2}\right)^{-1}.$$

Dette er netop (5.8). Vi har dermed vist at $Q(x,y)$ er en aftagende funktion af t^2 .

Vi mangler stadig at vise udtrykket for p -værdien. Lad os indføre de stokastiske variable $Q(X,Y)$, U , Z og T svarende til $Q(x,y)$, u , z og t :

$$\begin{aligned} Q(X,Y) &= \left(\frac{\text{SSD}_X + \text{SSD}_Y}{\text{SSD}_{X,Y}}\right)^{n/2}, \\ U &= \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \\ Z &= \frac{1}{\sigma^2} (\text{SSD}_X + \text{SSD}_Y), \\ T &= \frac{U}{\sqrt{Z/(n-2)}}. \end{aligned}$$

Så er der den samme relation mellem $Q(X,Y)$ og T som mellem $Q(x,y)$ og t , dvs.

$$(Q(X,Y))^{2/n} = \frac{\sigma^2 Z}{\sigma^2 Z + \sigma^2 U^2} = \left(1 + \frac{U^2}{Z}\right)^{-1} = \left(1 + \frac{T^2}{n-2}\right)^{-1}.$$

Vi kan derfor skrive p -værdien som

$$\varepsilon(x,y) = P(Q(X,Y) \leq Q(x,y)) = P(T^2 \geq t^2).$$

Under hypotesen følger det af (5.6) at $U \sim N(0, 1)$, og af sætning 5.3 at $Z \sim \chi_{n-2}^2$. Da \bar{X} , \bar{Y} , $SSD_{\bar{X}}$ og $SSD_{\bar{Y}}$ er uafhængige, følger det af sætning A.3 i appendix A at U og Z er uafhængige. Det følger således af definitionen på en t -fordeling (definition A.4 i appendix A), at T er t -fordelt med $n - 2$ frihedsgrader, og p -værdien kan derfor beregnes i t -fordelingen som angivet i sætningen. \square

Testet kaldes et t -test. På samme vis som vi har set i de foregående kapitler, består testet altså i at beregne den observerede værdi af T -teststørrelsen, dvs. t , og beregne hvor ekstremt værdien ligger i t -fordelingen med $n - 2$ frihedsgrader. Også intuitivt giver dette god mening: hypotesen bør afvises hvis \bar{x} og \bar{y} afviger meget og bør således baseres på $|\bar{x} - \bar{y}|$. Division med den estimerede spredning af forskellen i gruppegennemsnit kan opfattes som en normering der transformerer teststørrelsen til en kendt skala og således tager højde for variationen i data.

Ligesom ved test i en enkelt stikprøve, plejer man at opdatere estimererne hvis hypotesen ikke kan afvises, dvs. angive estimatet for μ_1 og μ_2 til $\hat{\mu}$ og estimatet for σ^2 til $SSD_{x,y}/(n - 1)$.

Kommentarerne fra afsnit 3.4 vedrørende sprogbrug, fejltyper og sammenhængen mellem konfidensintervaller og hypotesetest gælder uændret. Specielt vil $1 - \alpha$ konfidensintervallet for $\mu_1 - \mu_2$ indeholde værdien 0 hvis og kun hvis hypotesen $H : \mu_1 = \mu_2$ ikke kan afvises på signifikansniveau α .

Eksempel 5.10. (*Tuberkulosevaccine, fortsættelse af eksempel 5.2, side 82*) Hvis der ikke er forskel i turberkulinreaktionerne på vacciner foretaget i København eller Oslo, må vi forvente at niveauet i gennemsnit er ens for de to produktionscentre. Ingen forskel svarer således til hypotesen $H : \mu_1 = \mu_2$. Værdien af t -teststørrelsen er

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{17.13 - 16.84}{3.43 \sqrt{\frac{1}{130} + \frac{1}{116}}} = 0.671$$

og p -værdien er

$$\varepsilon(x, y) = 2P(T \geq 0.671) = 0.501$$

hvor $T \sim t_{244}$. Der er således ingen evidens mod hypotesen som derfor accepteres. Bemærk at dette stemmer overens med at konfidensintervallet for forskellen indeholder 0. Turberkulinreaktionen estimeres til 16.99 mm uanset produktionscenter, med konfidensinterval (16.56, 17.42). \square

5.5 Modelkontrol

I hele dette kapitel om analyse af to stikprøver har vi antaget følgende:

- Observationerne er uafhængige
- Der er samme varians i de to grupper
- Observationerne er normalfordelte

Uafhængigheden følger ofte af måden data er indsamlet på. Hvis data stammer fra tilfældigt udvalgte forsøgseenheder, der i øvrigt ikke formodes at have noget med hinanden at gøre, er der ikke grund til at betvivle uafhængigheden. For data vedrørende tuberkulosevaccine fra eksempel 5.2 synes antagelsen at være rimelig hvis børnene i undersøgelsen ikke er søskende, ikke hører sammen i grupper der får samme behandling, eller lignende.

Eksempel 5.11. (*Produktivitetsscore*) En produktivitetsscore er blevet målt på 3 forskellige slags maskiner af de samme tilfældigt udvalgte fabriksarbejdere, dvs. alle arbejdere har testet alle 3 maskiner (Pinheiro and Bates, 2000). Hver af de 6 fabriksarbejdere har testet hver maskine 3 gange, dvs. der er 18 målinger per maskine.

Data er indtegnet til venstre i figur 5.1 for to af maskinerne. Der ses en tydelig forskel mellem grupperne (maskintype). I dette tilfælde vil antagelsen om uafhængighed mellem målinger ikke være opfyldt, idet det må forventes at målinger foretaget af samme arbejder vil ligne hinanden mere end målinger foretaget af forskellige arbejdere. Vi får først redskaber til at håndtere den slags data på et senere kursus (se dog opgave 5.6). □

Antagelsen om *varianshomogenitet*, altså antagelsen om at variansen er ens i de to grupper, kan formelt testes med et såkaldt F -test. Det er dog udenfor pensum i dette kursus, og vi vil i stedet grafisk vurdere om antagelsen virker rimelig. For produktivitetsscorerne fra eksempel 5.11 lader til at variansen er den samme indenfor de to grupper fordi observationerne spreder sig nogenlunde lige meget fra gennemsnittet i de to grupper (til gengæld var der problemer med uafhængigheden).

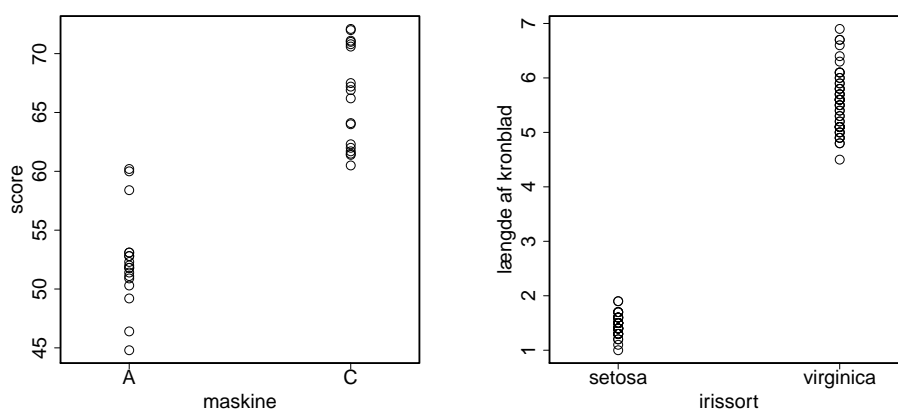
Eksempel 5.12. (*Længde af kronblade*) Længden af kronbladene i cm på 50 irisblomster af forskellige sorter er blevet målt (Venables and Ripley, 1999). Data er indtegnet til højre i figur 5.1. Der ses en tydelig forskel mellem grupperne (blomstersort), og der er tydeligvis også stor forskel på variansen indenfor hver gruppe. Variansen lader til at vokse med middelværdien, hvilket er et fænomen man ofte ser.

Hvis man ønskede at sammenligne de to sorter ville man ikke umiddelbart kunne benytte metoderne fra dette kapitel. En log-transformation ville formentlig afhjælpe problemet, og analysen skulle i så fald foretages på de transformerede data (se opgave 5.8). \square

I afsnit 4.5 diskuterede vi hvordan man grafisk kan vurdere om en enkelt stikprøve kan antages at komme fra en normalfordeling. Det gør vi på samme måde her, bortset fra at vi nu har to grupper. Vi skal derfor tjekke *normalfordelingsantagelsen* i begge grupper — ikke i det samlede datasæt. Hvis de to grupper har forskellig middelværdi, vil fordelingen i det samlede datasæt være topuklet. Vi vil således tegne histogrammer og QQ-plots for begge grupper.

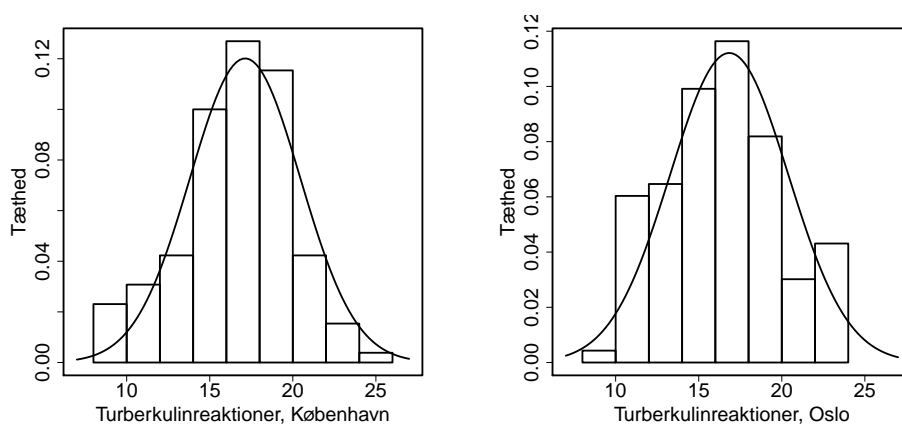
Eksempel 5.13. (*Tuberkulosevaccine, fortsættelse af eksempel 5.2, side 82*) Figur 5.2 viser histogrammer med normalfordelingstæthed indtegnet for tuberkulinreaktionsmålingerne fra København (til venstre) og fra Oslo (til højre). Tætheden er en god approksimation til histogrammet i begge figurer — det er faktisk sjældent at man ser så god overensstemmelse med normalfordelingen.

Figur 5.3 viser de tilsvarende QQ-plots. Punkterne ligger nogenlunde omkring en ret linje. Læg mærke til hvordan punkterne ligger som på en trappe. Det skyldes at datamålingen er forholdsvis upræcis, og kun opgivet i hele antal mm. Der vil således være mange ens målinger, som aldrig ville ske ved en “sand” normalfordeling. De “sande” hævselser er jo heller ikke et præcist antal hele mm, og formentlig er to hævselser aldrig helt ens. Data er trunkeret, og i dette datasæt er denne trunkering temmelig voldsom. Det betyder dog ikke noget for analysen.

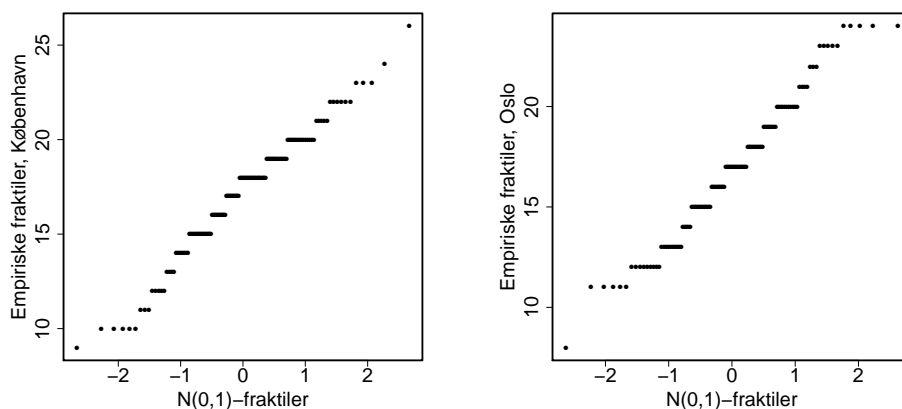


Figur 5.1: Produktivitetsscore for 2 forskellige maskintyper (til venstre) og længden af kronbladene på forskellige sorter af irisblomsten (til højre).

Vi kan således godt acceptere normalfordelingsantagelsen. Bemærk også at histogrammerne er nogenlunde lige brede, hvilket antyder at antagelsen om samme varians i begge grupper er acceptabel. Bemærk at en figur svarende til dem i figur 5.1 ikke er særligt nyttig i dette tilfælde på grund af trunkeringen. Der ville være mange punkter oven i hinanden, og det ville derfor være vanskeligt at vurdere variabiliteten i data. \square



Figur 5.2: Histogram og normalfordelingstæthed for turberkulinreaktionsmålingerne i København (til venstre) og i Oslo (til højre).



Figur 5.3: QQ-plots for turberkulinreaktionsmålingerne i København (til venstre) og i Oslo (til højre).

	Undervægtig ($n = 13$)	Overvægtig ($n = 9$)
	7.53 7.48 8.08	9.21 11.51
	8.09 10.15 8.40	12.79 11.85
	10.88 6.13 7.90	9.97 8.79
	7.05 7.48 7.58	9.69 9.68
	8.11	9.19
Gennemsnit	8.066	10.298
Spredning	1.238	1.398

Tabel 5.1: Energiforbruget i MJ/dag over 24 timer i grupper af undervægtige og overvægtige kvinder (Altman, 1999).

5.6 Eksempel: Energiforbrug

I dette afsnit analyseres et datasæt vedrørende det daglige energiforbrug for under- og overvægtige kvinder. Formålet med eksemplet er at få samlet trådene fra resten af kapitlet sammen og set hvordan de sættes sammen til en (mere eller mindre) fuldstændig analyse.

Eksempel 5.14. (*Energiforbrug*) Energiforbruget i løbet af 24 timer er blevet målt i MJ/dag hos to grupper af henholdsvis undervægtige og overvægtige kvinder (Altman, 1999). Data består af 13 målinger af undervægtige kvinder, x_1, \dots, x_{13} , og 9 målinger af overvægtige kvinder, y_1, \dots, y_9 . Observationerne betragtes som realisationer af X_1, \dots, X_{13} og Y_1, \dots, Y_9 som antages at være uafhængige og normalfordelte med varians σ^2 og middelværdier μ_1 og μ_2 . Data er angivet i tabel 5.1 og plottet i figur 5.4.

Udfra figur 5.4 lader det til at antagelsen om samme varians i begge grupper godt kan accepteres. Med kun 13 og 9 observationer i hver gruppe er der ikke data nok til at lave histogrammer, og det er svært at kontrollere normalfordelingsantagelsen. I figur 5.4 er QQ-plots indtegnet, og vi kan udfra disse godt acceptere normalfordelingsantagelsen, dog med forbehold fordi der er så få observationer. Bemærk at med så få punkter er det almindeligt at der er store afvigelser fra en ret linje, selv når data faktisk er normalfordelt, som beskrevet og illustreret i afsnit 4.5.

For de 13 observationer x_1, \dots, x_{13} for de undervægtige kvinder og de 9 observationer

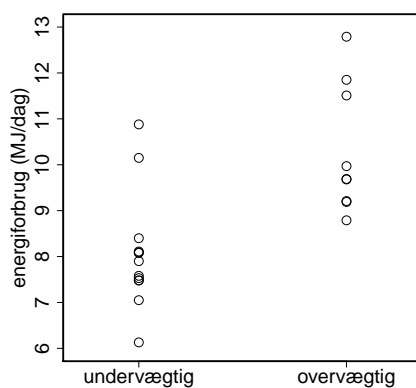
y_1, \dots, y_9 for overvægtige kvinder har vi følgende:

$$\bar{x} = 8.066; \quad \frac{1}{12} \sum_{i=1}^{13} (x_i - \bar{x})^2 = 1.5326 = 1.238^2$$

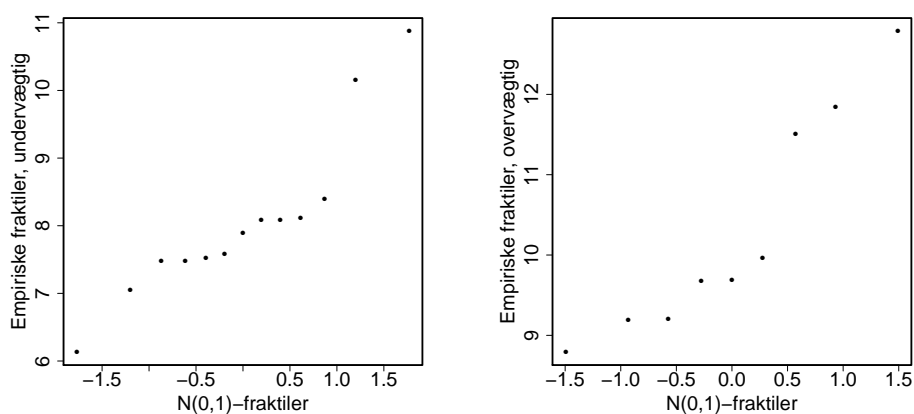
$$\bar{y} = 10.298; \quad \frac{1}{8} \sum_{j=1}^9 (y_j - \bar{y})^2 = 1.9544 = 1.398^2.$$

Således er estimerterne

$$\bar{\mu}_1 = 8.066, \quad \bar{\mu}_2 = 10.298$$



Figur 5.4: Energiforbruget i MJ/dag over 24 timer i grupper af undervægtige og overvægtige kvinder (Altman, 1999).



Figur 5.5: QQ-plots for energiforbrug for undervægtige kvinder (til venstre) og for overvægtige kvinder (til højre).

og

$$s^2 = \frac{12 \cdot 1.5326 + 8 \cdot 1.9544}{13 + 9 - 2} = 1.7014, \quad s = 1.3044.$$

Estimatorerne er uafhængige, $\hat{\mu}_1 \sim N(\mu_1, \sigma^2/13)$, $\hat{\mu}_2 \sim N(\mu_2, \sigma^2/9)$ og $20\hat{\sigma}^2 \sim \sigma^2\chi_{20}^2$. De estimerede spredninger (standard errors) er $SE(\hat{\mu}_1) = s/\sqrt{13} = 0.36$ for $\hat{\mu}_1$ og $SE(\hat{\mu}_2) = s/\sqrt{9} = 0.44$ for $\hat{\mu}_2$.

For at konstruere 95% konfidensintervaller for μ_1 og μ_2 skal vi bruge 97.5% fraktilen i t -fordelingen med 20 frihedsgrader der er lig 2.086. Vi får følgende:

$$8.066 \pm 2.086 \cdot \frac{1.3044}{\sqrt{13}} = 8.066 \pm 0.755 = (7.311, 8.821)$$

$$10.298 \pm 2.086 \cdot \frac{1.3044}{\sqrt{9}} = 10.298 \pm 0.907 = (9.391, 11.205)$$

I virkeligheden er forskellen mellem de to grupper mere interessant. Forskellen mellem middelværdierne, $\mu_1 - \mu_2$, estimeres til $8.066 - 10.298 = -2.232$ med estimeret spredning 0.567. Vi får derfor 95% konfidensinterval for forskellen:

$$-2.232 \pm 2.086 \cdot 0.567 = -2.232 \pm 1.180 = (-3.412, -1.052).$$

Bemærk at konfidensintervallerne for μ_1 og μ_2 er disjunkte. Intuitivt passer det med at konfidensintervallet for forskellen ikke indeholder nul. Hvis $\mu_1 - \mu_2 = 0$ — svarende til at der ikke er forskel mellem de to grupper — er det således usandsynligt at vi skulle have observeret de data vi har til rådighed.

Hvis der ikke er forskel i energiforbruget hos undervægtige og overvægtige kvinder, må vi forvente at niveauet i gennemsnit er ens for de to grupper. Ingen forskel svarer således til hypotesen $H : \mu_1 = \mu_2$. Værdien af t -teststørrelsen er

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{8.066 - 10.298}{1.3044\sqrt{\frac{1}{130} + \frac{1}{116}}} = -3.9456$$

og p -værdien er

$$\varepsilon(x, y) = 2P(T \geq 3.9456) = 0.000799$$

hvor $T \sim t_{20}$. Fortolkningen af p -værdien er at hvis der ikke er forskel mellem grupperne, dvs. hypotesen er sand, da vil sandsynligheden for at observere disse data, eller noget der er længere væk fra hypotesen, være 0.000799. Da denne sandsynlighed er meget lille, afviser vi hypotesen. Bemærk at dette stemmer overens med konfidensintervallet for forskellen, der ikke indeholder nul. Vi konkluderer således at der er evidens i data for at energiforbruget hos overvægtige kvinder er højere end energiforbruget hos undervægtige kvinder. \square

5.7 Sammenfatning og perspektiv

Vi har diskuteret statistisk analyse af uafhængige normalfordelte observationer fra to grupper med samme varians, men muligvis med forskellige middelværdier. Modellen kan bruges når observationerne indenfor hver gruppe kan antages at være frembragt af samme normalfordeling. Som regel er man først og fremmest interesseret i at estimere forskellen i middelværdier, eller undersøge om middelværdien er den samme i de to grupper. Analysen sammenfattes ofte med estimater og konfidensintervaller for middelværdierne i hver gruppe, og for forskellen i middelværdier. Den naturlige hypotese at teste er at grupperne har samme middelværdi.

Konstruktionen af konfidensintervaller og udførelsen af hypotesetest er begrebsmæssigt den samme som for modellen for en enkelt stikprøve.

Antagelserne for at lave analysen bør altid tjekkes før man drager nogle konklusioner. Antagelserne er at observationerne er uafhængige, normalfordelte og med samme varians i de to grupper.

5.8 R

Vi bruger data fra eksempel 5.2 (side 82) om tuberkulosevaccine som illustration. Data er tilgængelige i filen `tb.txt`, med variable `by` og `diameter`. Vi indlæser først datasættet i R:

```
> tbdata <- read.table("tb.txt", header=T)
> tbdata      # Datasæt med alle observationer
  by diameter
1 kbh         9
2 kbh        10
.
.      [Flere dataliner her]
.
245 oslo      24
246 oslo      24
```

Til nogle af analyserne har vi brug for variable der kun indeholder observationer fra enten Oslo eller København. De kan fx laves på følgende måde:

```
diameterKbh <- subset(tbdata, by=="kbh")$diameter
```

```
diameterOslo <- subset(tbdata, by=="oslo")$diameter
```

Her laver `subset(tbdata, by="kbh")` et deldatasæt af `tbdata` der kun indeholder de observationer (datalinjer) hvor variabelen `by` har værdien `kbh`. Fra dette datasæt udtrækkes variabelen `diameter` med `$`. Variablen med tallene fra København ser nu således ud:

```
> diameterKbh          # Observationer fra København
 [1]  9 10 10 10 10 10 11 11 11 12 12 12 12 12 13 13 13 13
[19] 14 14 14 14 14 14 14 15 15 15 15 15 15 15 15 15 15 15
  .
  .
[127] 23 23 24 26
```

Estimator, konfidensinterval for $\mu_1 - \mu_2$ og hypotesetest Analysen vedrørende forskellen mellem middelværdierne μ_1 og μ_2 laves nemmest ved at bruge funktionen `t.test`. Det kan gøres på to måder, afhængig af om man bruger det oprindelige datasæt eller de to variable der udgør stikprøverne. Kommandoerne er følgende:

```
t.test(diameter~by, data=tbdata, var.equal=TRUE)
t.test(diameterKbh, diameterOslo, var.equal=T)
```

Bemærk at argumentet `var.equal` skal ændres fra defaultværdien `FALSE` til `TRUE` fordi vi antager at varianserne er ens i de to grupper. Bemærk også symbolet `~` i den første kommando; det kalder en "tilde". Outputtet fra de kommandoer er ens (bortset fra lidt tekst der beskriver variablene) og ser således ud:

```
> t.test(diameterKbh, diameterOslo, var.equal=T) # Analyse

Two Sample t-test

data: diameterKbh and diameterOslo
t = 0.6714, df = 244, p-value = 0.5026
alternative hypothesis: true difference in means is
                                not equal to 0
95 percent confidence interval:
 -0.5695597  1.1586844
sample estimates:
```

```
mean of x mean of y
 17.13077  16.83621
```

Øverst i outputtet aflæses den observerede værdi af t -teststørrelsen (0.6714), antallet af frihedsgrader (244) og p -værdien (0.5026) for hypotesen $H : \mu_1 = \mu_2$. Derefter følger 95% konfidensintervallet for forskellen mellem middelværdierne, $\mu_1 - \mu_2$, nemlig $(-0.5695597, 1.1586844)$. Nederst angives de to middelværdiestimer (17.13077 og 16.83621). På nær afrundingsfejl er disse værdier de samme som vi beregnede i eksempel 5.5, 5.8 og 5.10 (side 86, 89 og 94).

Variansestimater Outputtet giver ikke variansestimateret s^2 så det må beregnes manuelt. Empiriske varianser for de to stikprøver hver for sig kan beregnes ved hjælp af `var`. Hvis vi bruger (5.2) kan s^2 og s således beregnes på følgende måde:

```
> (129*var(diameterKbh) + 115*var(diameterOslo)) /
+ (129+115) # Beregning af s^2
[1] 11.79781
> sqrt((129*var(diameterKbh) + 115*var(diameterOslo)) /
+ (129+115)) # Beregning af s
[1] 3.434794
```

Konfidensintervaller for μ_1 og μ_2 Konfidensintervallerne fra afsnit 5.6 for μ_1 og μ_2 (ikke deres forskel) angives ikke som output fra `t.test`-kommandoen ovenfor, men kan naturligvis beregnes manuelt. Gennemsnittene beregnes med `mean` mens t -fordelingsfraktilen beregnes med `qt`, se afsnit 4.7.

Modelkontrol Histogrammer og QQ-plots for de enkelte variable laves ved hjælp af `hist` og `qqnorm` som forklaret i afsnit 4.7. Følgende kommandoer giver (pånær layout) plottene i figur 5.2 og 5.3:

```
hist(diameterKbh, prob=T) # Histogram for København
hist(diameterOslo, prob=T) # Histogram for Oslo
qqnorm(diameterKbh) # QQ-plot for København
qqnorm(diameterOslo) # QQ-plot for Oslo
```

Figuren svarende til figur 5.1 er som nævnt ikke så nyttig for disse data fordi de samme værdier er observeret mange gange, men figuren kunne laves således:

```
stripchart(list(diameterKbh, diameterOslo), vertical=T)
```

5.9 Opgaver

5.1 I en undersøgelse offentliggjort i artiklen *Are Women Really More Talkative Than Men?* (Mehl *et al.*, 2007) blev antallet af ord som 396 kvindelige og mandlige universitetsstuderende i USA og Mexico taler på en dag målt. Resultaterne er opsummeret i følgende tabel.

	Kvinder ($n = 210$)	Mænd ($n = 186$)
Gennemsnit	16215	15669
Spredning	7301	8633

1. Opstil en statistisk model der gør det muligt at undersøge om der er forskel på antallet af ord en kvinde og en mand taler på en dag.
2. Angiv estimater for samtlige parametre i modellen, og angiv også de tilhørende estimatorers fordeling. Bestem desuden den estimerede spredning for estimatorerne for middelværdiparametrene.
3. Beregn et estimat for den forventede forskel mellem antallet af ord de to køn taler på en dag. Bestem også den estimerede spredning for den tilhørende estimator samt et 95% konfidensinterval for forskellen.
4. Udfør et hypotesetest der undersøger om der er forskel på antallet af ord en kvinde og en mand taler på en dag.

5.2 Værdistigningen for 15 investeringsforeninger er blevet undersøgt over en fem-årsperiode. Værdien af aktieporteføljen blev sat til 100 ved periodens start, og tallene i tabellen viser værdien ved periodens slutning. Seks af foreningerne investerer hovedsageligt i danske aktier, de øvrige ni hovedsageligt i udenlandske aktier.

Danske	213.41	228.50	214.16	217.94	230.01	203.52
Udenlandske	148.40	217.42	205.98	221.83	164.19	224.09
	193.56	205.27	218.44			

1. Udfør et test for hypotesen om at der ikke er forskel på værdistigningen for de to typer investeringsforeninger. Du kan bruge nedenstående R-output:

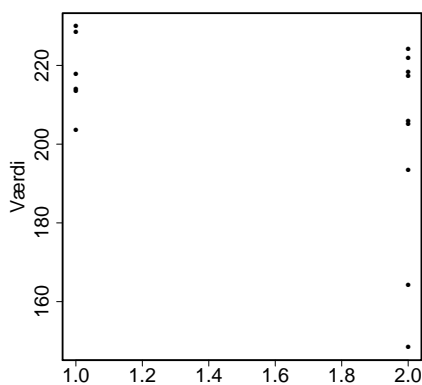

```
> dk <- c(213.41, 228.50, 214.16, ... , 203.52)
> udl <- c(148.40, 217.42, 205.98, ..., 218.44)


> t.test(dk, udl, var.equal=T)
```

Two Sample t-test

```
data: dk and udl
t = 1.5588, df = 13, p-value = 0.1430
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
-6.95164 42.98053
sample estimates:
mean of x mean of y
217.9233 199.9089
```

2. Gør rede for forudsætningerne for testet. Er der grund til at tro at nogle af forudsætningerne er problematiske? Benyt evt. tegningen nedenfor.



5.3  Nedenstående data stammer fra en undersøgelse af 2 typer organiske opløsningsmidler, dels aromatiske forbindelser og dels klorerede hydrocarboner (Ortego *et al.*, 1995). I uafhængige prøver fra hver af de to stoffer målt bindingsraten. Resultaterne er angivet i tabellen.

Opløsningsmiddel	Bindingsrate
Aromatiske forbindelser	1.06 0.79 0.82 0.89 1.05 0.95 0.65 1.15 1.12
Klorerede hydrocarb.	1.28 1.35 0.57 1.16 1.12 0.91 0.83 0.43

Data ligger i filen `oplosningsmiddel.txt`.

Det kan i det følgende antages at observationer stammer fra uafhængige og normalfordelte stokastiske variable.

1. Opstil en statistisk model til beskrivelse af forsøget.
2. Konstruér to vektorer, `arom` og `klor`, der indeholder målingerne.

Vink: Brug `read.table` til at indlæse data til et R-datasæt, fx med navnet `oplosdata`. Derefter kan du bruge følgende kommandoer:

```
arom <- subset(oplosdata, oplos=="Arom")$rate
klor <- subset(oplosdata, oplos=="Klor")$rate
```

Alternativt kan du indtaste vektorerne manuelt.

3. Brug `mean` og `var` til at beregne gennemsnit og empiriske varianser for `arom` og `klor` hver for sig. Beregn derefter det sammenevejede variansestimater s^2 .
4. Beregn 95% konfidensintervaller for middelværdien af bindingsraten ved hver af de to opløsningsmidler. Brug `qt` til at bestemme den relevante fraktil.
5. Prøv kommandoerne

```
t.test(arom)
t.test(klor)
```

og sammenlign konfidensintervallerne med dem fra spørgsmål 4. Hvorfor er de forskellige?

6. Bestem estimatet og 95% konfidensintervallet for forskellen mellem bindingsraten i de to grupper. Regn det både i hånden og med følgende kommando:

```
t.test(arom, klor, var.equal=TRUE)
```

7. Undersøg med et hypotesetest om de to bindingsrater kan antages at ligge på samme niveau (brug outputtet fra spørgsmål 6). Sammenlign med resultaterne fra spørgsmål 6.
8. Redegør for forudsætningerne for analysen. Det er vanskeligt at lave modelkontrol når der er så få observationer, men gør dig alligevel nogle overvejelser om hvorvidt antagelserne med rimelighed kan antages at være opfyldt.

5.4 Betragt den statistiske model fra definition 5.1.


1. Lad n være et givet lige tal. Hvordan vælges n_1 og n_2 således at $n = n_1 + n_2$ og således at $\text{Var}(\bar{X} - \bar{Y})$ er mindst mulig?
2. Forklar hvad dette betyder i forbindelse med forsøgsplanlægning: Antag at vi ønsker at sammenligne to behandlinger og har n forsøgspersoner til rådighed. Hvordan fordeler vi bedst muligt de n personer på de to behandlinger? Og hvad betyder 'bedst muligt' i denne sammenhæng?

Betragt desuden $1 - \alpha$ konfidensintervallet for $\mu_1 - \mu_2$ fra sætning 5.6. Længden af konfidensintervallet er

$$L = 2 \cdot t_{n-2, 1-\alpha/2} \cdot \tilde{\sigma} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

som er en stokastisk variabel.

3. Gør rede for at middelværdien af $\tilde{\sigma}$ eksisterer og kun afhænger af n_1 og n_2 gennem n . Det er ikke meningen at du skal beregne $E(\tilde{\sigma})$.
4. Gør rede for at middelværdien af L eksisterer.
5. Hvad siger resultatet fra spørgsmål 1–2, udtrykt ved hjælp af længden af konfidensintervallet?

5.5  Data til denne opgave stammer fra to eksperimenter, hvor man målte fluers reaktionstid efter de var blevet udsat for nervegas (Blæsild and Granfeldt, 2003). Målingen for den enkelte flue består i den tid — reaktionstiden — der går fra fluen bringes i kontakt med giften og indtil den ikke længere kan stå på benene. I det første eksperiment blev fluerne udsat for giften i 30 sekunder og i det andet i 60 sekunder. Målingerne af reaktionstiden ses nedenfor.

	Reaktionstid i sekunder							
kontakttid	3	5	5	7	9	9	10	12
30 sekunder	20	24	24	34	43	46	58	140
kontakttid	2	5	5	7	8	9	14	18
60 sekunder	24	26	26	34	37	42	90	

1. Indtast data i to variable, `reak30` og `reak60`. Lav derefter to nye variable, `logreak30` og `logreak60`, bestående af de log-transformerede data.


2. Undersøg rimeligheden af normalfordelingsantagelsen både på de oprindelige data og de log-transformerede data.
3. Beregn den empiriske varians for hver af de fire variable. På hvilken skala virker det mest rimeligt at antage at variansen er den samme?


Antag at fra nu af at observationerne er uafhængige, og at logaritmen til observationstiderne er normalfordelte med samme varians.

2. Vis at data ikke tyder på at fordelingen af reaktionstiden afhænger af om kontakttiden er 30 eller 60 sekunder.
3. Vi kan altså betragte alle data som en enkelt stikprøve. Angiv et estimat og et konfidensinterval for de logaritmetransformerede observationers middelværdi.
4. Angiv et estimat og et konfidensinterval for den forventede reaktionstid. *Vink:* Transformer estimatet og konfidensgrænserne tilbage til den oprindelige skala. Bliver konfidensintervallet symmetrisk?

5.6 Læs eksempel 5.11 igen. Som det fremgår er der problemer med antagelsen om uafhængighed. Vi skal nu overveje hvordan man alligevel kunne undersøge om der er forskel på produktivitetsscoren for de to maskiner.

1. Der er tre gentagelser for hver kombination af person og maskine. Hvordan kan disse på en hensigtsmæssig måde reduceres til en enkelt observation, således at data består af kun 12 tal (et per kombination af person og maskine)?
2. Hvilken model kan bruges til at analysere disse 12 tal? *Vink:* Er der tale om et parret eller et uparret set-up?

5.7  Kør analysen fra eksempel 5.14 om energiforbrug hos under- og overvægtige kvinder i R. Check at du får de samme resultater som i eksemplet.

5.8  Data til eksempel 5.12 ligger allerede i datasættet `iris` i R. Følgende kommandoer konstruerer de to stikprøver fra eksemplet:

```
x <- with(iris, Petal.Length[Species=="setosa"])
y <- with(iris, Petal.Length[Species=="virginica"])
```

1. Kør kommandoerne og skriv `x` og `y` ud på skærmen.

2. Prøv følgende kommandoer, en ad gangen:

```
stripchart(list(x,y), vertical=T)
stripchart(list(log(x),log(y)), vertical=T)
```

Første graf er (pånær layout) identisk med højre del af figur 5.1. Forklar hvad du ser på den anden graf. Hvad er konklusionen med hensyn til varianshomogenitet?

3. Overvej hvordan du ville undersøge om der er forskel på længden af kronblade for de to sorter. Udfør evt. analysen med `t.test`. *Vink*: Hvilken variabel ville du bruge?

5.9 Betragt data x_1, \dots, x_n og y_1, \dots, y_n og antag at de er udfald af uafhængige stokastiske variable X_1, \dots, X_{n_1} og Y_1, \dots, Y_{n_2} hvor $X_i \sim N(\mu_1, \sigma_0^2)$ og $Y_j \sim N(\mu_2, \sigma_0^2)$. Her er $\mu_1, \mu_2 \in \mathbb{R}$ ukendte parametre mens $\sigma_0^2 > 0$ er et kendt tal. Modellen er altså identisk med modellen fra definition 5.1 bortset fra at variansen er kendt. Vi skal interessere os for test af hypotesen $H : \mu_1 = \mu_2$.

1. Gør rede for at estimererne for μ_1 og μ_2 i modellen og under hypotesen er givet ved

$$\hat{\mu}_1 = \bar{x}, \quad \hat{\mu}_2 = \bar{y}, \quad \hat{\mu}_1 = \hat{\mu}_2 = \frac{1}{n}(n_1\bar{x} + n_2\bar{y})$$

hvor $n = n_1 + n_2$.

2. Vis at kvotientteststørrelsen er givet ved

$$Q(x,y) = \frac{L_{x,y}(\hat{\mu}_1, \hat{\mu}_2)}{L_{x,y}(\hat{\mu}_1, \hat{\mu}_2)} = \exp\left(-\frac{1}{2\sigma_0^2} \frac{n_1 n_2}{n} (\bar{x} - \bar{y})^2\right)$$

og gør rede for at likelihood ratio testet derfor kan udføres på

$$u = \frac{\bar{x} - \bar{y}}{\sigma_0 \sqrt{1/n_1 + 1/n_2}}.$$

3. Gør rede for at

$$U = \frac{\bar{X} - \bar{Y}}{\sigma_0 \sqrt{1/n_1 + 1/n_2}}$$

er standard normalfordelt under hypotesen, og at hypotesen derfor accepteres (ikke forkastes) på 5% niveau hvis og kun hvis $|u| < 1.96$.

Antag nu at $\Delta = \mu_1 - \mu_2 > 0$ således at hypotesen er falsk.

4. Vis at sandsynligheden for at hypotesen $H : \mu_1 = \mu_2$ accepteres på 5% signifikansniveau er

$$q(k) = \Phi(1.96 - k) - \Phi(-1.96 - k), \quad k = \frac{\Delta}{\sigma_0 \sqrt{1/n_1 + 1/n_2}}$$

Vink: Skriv U som

$$U = \frac{\bar{X} - \bar{Y} - \Delta}{\sigma_0 \sqrt{1/n_1 + 1/n_2}} + k$$

og vis at første led er standard normalfordelt når den sande forskel $\mu_1 - \mu_2$ er Δ .

5. Vis at q er en aftagende funktion på $(0, \infty)$. *Vink:* Differentier.
6. Hvad sker der med sandsynligheden for at begå fejl af type II (acceptere en falsk hypotese) når
- den sande forskel Δ mellem μ_1 og μ_2 vokser?
 - variansen σ_0^2 vokser?

Argumentér både intuitivt og ved hjælp af funktionen q .

7. For et givet lige tal n , hvordan vælges n_1 og n_2 således at $n = n_1 + n_2$ og således at sandsynligheden for at begå fejl af type II er mindst mulig. *Vink:* Du kan benytte resultatet fra opgave 5.2 hvis du har lavet den.

Kapitel 6

Lineær regression

I de forrige kapitler har vi set på normalfordelingsmodeller der involverer en enkelt variabel. Ofte er man dog interesseret i at beskrive sammenhænge mellem flere variable eller, mere specifikt, at beskrive en variabel som funktion af en anden variabel. Funktionen kan være givet ud fra en teori om årsagssammenhænge. Dette gælder for eksempel banen som et projektil gennemløber som funktion af tiden, da denne bane kan beskrives ved en parabel bestemt ud fra tyngdeaccelerationen og den hastighed projektilet afskydes med.

Ofte kender man dog ikke de bagvedliggende fysiske love eller biologiske mekanismer, og den statistiske analyse skal netop sandsynliggøre eller afvise forskellige forklaringsmodeller for observerede sammenhænge. I mangel af teoretisk viden om årsagssammenhænge baseres analysen således på empiriske sammenhænge, dvs. sammenhænge baseret på observerede data. Data indsamles for at få viden om sammenhængen.

Regressionsmodeller anvendes til at beskrive sammenhænge mellem en stokastisk responsvariabel og en eller flere forklarende variable, der formodes at have indflydelse på niveauet af responsvariablen. De forklarende variable kaldes også regressionsvariable, baggrundsvariable eller kovariater.

Vi vil i disse noter kun se på en enkelt forklarende variabel og desuden kun på lineære sammenhænge mellem responsvariablen og den forklarende variabel. Dette kaldes i nogle sammenhænge en simpel lineær regression og danner udgangspunkt for mere avancerede og realistiske modeller. Teorien for simpel lineær regression gennemgås i afsnit 6.1–6.6, og bliver illustreret af et eksempel fra medicinsk forskning. I afsnit 6.7 ser vi nærmere på en berømt model fra finansiering, nemlig CAPM. Dette afsnit skal

blot ses som et eksempel på lineær regression i en økonomisk/finansieringsmæssig sammenhæng.

6.1 Statistisk model

Den simpleste beskrivelse af sammenhængen mellem en responsvariabel y og en forklarende variabel x er en lineær funktion

$$y(x) = \alpha + \beta x. \quad (6.1)$$

Udgangspunktet er stadig uafhængige og normalfordelte stokastiske variable, men middelværdien kan nu afhænge af værdien af en anden variabel. Vi betragter n par af sammenhørende observationer $(x_1, y_1), \dots, (x_n, y_n)$. Observationerne y_1, \dots, y_n er realisationer af de stokastiske variable Y_1, \dots, Y_n , hvor vi antager at Y_1, \dots, Y_n er stokastisk uafhængige, og at Y_i er normalfordelt med middelværdi $\alpha + \beta x_i$ og varians σ^2 . Værdierne x_1, \dots, x_n antages derimod at være kendte tal. Vi antager også at mindst to af x 'erne er forskellige — ellers vil vi jo ikke kunne udtale os om hvordan middelværdien ændrer sig som funktion af x , og $\alpha + \beta x_i$ er blot en konstant for alle $i = 1, \dots, n$. Modellen ville således svare til modellen for en enkelt stikprøve med ukendt varians, som blev behandlet i kapitel 4.

Linjen (6.1) kaldes regressionslinjen. Parameteren β beskriver hvordan middelværdien ændrer sig når x ændrer sig. Hvis $\beta > 0$ vil middelværdien af Y vokse når x vokser. Hvis $\beta < 0$ vil en højere værdi af x gøre middelværdien af Y mindre. Parameteren β kaldes også effektparameteren af x på y og kan fortolkes direkte: når x vokser en enhed, ændres middelværdien af Y med β enheder. En ændring af x med en enhed er ikke meningsfuld i alle sammenhænge; tænk for eksempel på en situation hvor x naturligt varierer mellem 0 og 1. Men fortolkningen kan skaleres: når x vokser med værdien Δ , ændres middelværdien af Y med $\beta\Delta$ enheder. Parameterværdien $\beta = 0$ er særlig interessant, fordi middelværdien af Y i dette tilfælde ikke afhænger af x . Ofte er formålet med den statistiske analyse netop at teste om Y afhænger af x , og den naturlige hypotese er i så fald $H : \beta = 0$.

Parameteren α angiver middelværdien svarende til $x = 0$, dvs. skæringen med y -aksen. Bemærk dog at værdien $x = 0$ ikke giver mening i alle sammenhænge. Tænk for eksempel på en situation hvor man interesserer sig for sammenhængen mellem højde (x) og vægt (y). Her ville $x = 0$ svare til en person der er 0 m høj, hvilket er meningsløst. Man skal således være en smule varsom med fortolkningen af α .

Udgangspunktet er således uafhængige stokastiske variable Y_1, \dots, Y_n , hvor

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2).$$

Sommetider skriver man i stedet $Y_i = \alpha + \beta x_i + \varepsilon_i$ hvor $\varepsilon_1, \dots, \varepsilon_n$ er uafhængige og $N(0, \sigma^2)$ -fordelte. Den centrale antagelse for lineær regression er

$$E(Y_i) = \alpha + \beta x_i.$$

Den simultane fordeling af (Y_1, \dots, Y_n) betegnes $N_{\alpha, \beta, \sigma^2}^n$ og har tæthed

$$\begin{aligned} f_{\alpha, \beta, \sigma^2}(y) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \alpha - \beta x_i)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right), \end{aligned} \quad (6.2)$$

hvor $y = (y_1, \dots, y_n) \in \mathbb{R}^n$. Vi antager at parameterområdet er $\Theta = \mathbb{R} \times \mathbb{R} \times (0, \infty)$, men det kunne også være en delmængde af denne mængde.

Definition 6.1. Modellen for en lineær regression består af udfaldsrummet \mathbb{R}^n samt familien

$$\mathcal{P} = \left\{ N_{\alpha, \beta, \sigma^2}^n : (\alpha, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty) \right\}$$

af fordelinger på \mathbb{R}^n hvor $N_{\alpha, \beta, \sigma^2}^n$ har tæthed (6.2).

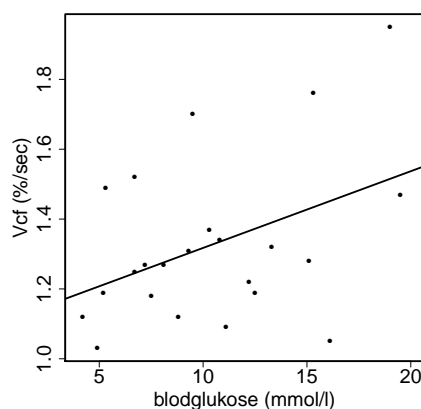
Alternativ formulering: Lad Y_1, \dots, Y_n være uafhængige normalfordelte stokastiske variable, $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ hvor $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ og $\sigma^2 > 0$ er ukendte parametre.

Eksempel 6.2. (Vcf og blodglukose) Et ekkokardiogram bruger ultralyd til at observere hjertets kamre og klapper og benyttes til at diagnosticere en række forskellige hjerteproblemer. For at undersøge om middelhastigheden hvormed det venstre hjertekammer trækker sig sammen (Vcf) målt ved et ekkokardiogram afhænger af blodglukosen under faste, blev der foretaget målinger af 23 patienter med type 1 diabetes (Altman, 1999). Patienter med diabetes har højere blodglukose under faste end raske personer, og diabetes er en risikofaktor for forskellige hjertesygdomme.

Data består af 23 sammenhørende målinger af Vcf (målt i % per sekund), betegnet y_1, \dots, y_{23} , og fasteblodglukosen (målt i mmol per liter), betegnet x_1, \dots, x_{23} . Det første man bør gøre, er altid at plotte data, både for at få en fornemmelse af data, og for at se om det er fornuftigt at beskrive data ved en lineær regressionsmodel. Det naturlige plot for sådanne sammenhørende par af målinger er et *scatterplot*, hvor

responsvariablen afsættes mod baggrundsvariablen i et koordinatsystem, således at y -variablen bestemmer værdien på den vertikale akse, og x -variablen bestemmer værdien på den horisontale akse. I figur 6.1 ses et scatterplot for blodglukose og Vcf. Det lader til at Vcf stiger når blodglukosen stiger. Det er ikke umiddelbart klart at sammenhængen er lineær, men det kan heller ikke afvises.

I den lineære regressionsmodel betragter vi blodglukosemålingerne x_1, \dots, x_{23} som faste og Vcf-målingerne y_1, \dots, y_{23} som udfald af stokastiske variable Y_1, \dots, Y_{23} der antages at være uafhængige og normalfordelte med varians σ^2 og middelværdier $\alpha + \beta x_i$. \square



Figur 6.1: Sammenhæng mellem blodglukose under faste og middelhastigheden hvormed det venstre hjertekammer trækker sig sammen (Vcf). Data stammer fra 23 type 1 diabetikere (Altman, 1999). Den rette linje er regressionslinjen beregnet i eksempel 6.8 på side 120.

6.2 Maksimum likelihood estimation

Vi skal estimere $(\alpha, \beta, \sigma^2)$ på basis af data, $x = (x_1, \dots, x_n)$ og $y = (y_1, \dots, y_n)$. Vi definerer igen likelihoodfunktionen som tætheden, men opfattet som funktion af parameteren. Vi ser altså på $L_y : \mathbb{R} \times \mathbb{R} \times (0, \infty) \rightarrow \mathbb{R}$ givet ved

$$L_y(\alpha, \beta, \sigma^2) = f_{\alpha, \beta, \sigma^2}(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right). \quad (6.3)$$

Vi vil maksimere likelihoodfunktionen ved hjælp af profilmethoden og skal se at det kan gøres ved at trække på maksimeringsresultaterne fra modellen for en enkelt

stikprøve (kapitel 4) på en smart måde, når vi kombinerer med følgende resultat:

Lemma 6.3. For talsæt $z_1, \dots, z_n \in \mathbb{R}$ og $s_1, \dots, s_n \in \mathbb{R}$, hvor ikke alle s_i 'erne er lig nul, vil funktionen

$$g(\beta) = \sum_{i=1}^n (z_i - \beta s_i)^2$$

minimeres entydigt af

$$\hat{\beta} = \frac{\sum_{i=1}^n s_i z_i}{\sum_{i=1}^n s_i^2}.$$

Bevis Vi differentierer ind i summen og ser at

$$\begin{aligned} g'(\beta) &= \sum_{i=1}^n (-s_i) 2(z_i - \beta s_i) = -2 \sum_{i=1}^n s_i z_i + 2\beta \sum_{i=1}^n s_i^2 \\ &= 2 \left(\sum_{i=1}^n s_i^2 \right) \left(\beta - \frac{\sum_{i=1}^n s_i z_i}{\sum_{i=1}^n s_i^2} \right). \end{aligned}$$

Heraf aflæser vi fortegnforholdene for $g'(\beta)$, og vi konstaterer let det ønskede. \square

For at formulere maksimeringsresultatet for (6.3) på en overskuelig måde, indfører vi nogle forkortelser. Som sædvanlig betegner \bar{x} og \bar{y} gennemsnittene, mens

$$\text{SSD}_x = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Derudover definerer vi

$$\text{SPD}_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Her står SPD for *sums of products of deviation*. Ligesom i de forrige kapitler korrigerer vi estimatoren for variansen, så den er central, se bemærkning 6.7.

Sætning 6.4. For den statistiske model fra definition 6.1 er maksimum likelihood estimatet for $(\alpha, \beta, \sigma^2)$ entydigt bestemt og givet ved

$$\hat{\alpha} = \bar{y} - \bar{x} \frac{\text{SPD}_{xy}}{\text{SSD}_x}, \quad \hat{\beta} = \frac{\text{SPD}_{xy}}{\text{SSD}_x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Bevis Vi skal maksimere en funktion af tre variable og benytter profileringsmetoden, see appendiks B. For fastholdt værdi af β og σ^2 er maksimeringsproblemet for (6.3) identisk med maksimeringsproblemet for en enkelt stikprøve med kendt varians når vi opfatter $y_i - \beta x_i$ som “observationerne”. Dette problem blev løst i afsnit 3.2, og vi ser derfra at maksimum antages i gennemsnittet af “observationerne”, dvs.

$$\hat{\alpha}(\beta, \sigma^2) = \frac{1}{n} \sum_{i=1}^n (y_i - \beta x_i) = \bar{y} - \beta \bar{x}.$$

Bemærk hvordan vi i notationen gør opmærksom på at vi har fundet et maksimum-punkt mht. α for fastholdt (β, σ^2) , men at løsningen faktisk kun afhænger af β .

Indsættes $\hat{\alpha}(\beta, \sigma^2)$ for α i (6.3), fås en funktion der kun afhænger af β og σ^2 :

$$\tilde{L}_y(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \bar{y} - \beta(x_i - \bar{x}))^2\right).$$

Vi siger at vi har profileret α ud. For fastholdt σ^2 kan vi maksimere $\tilde{L}_y(\beta, \sigma^2)$ mht. β ved at minimere eksponenten. Men det svarer netop til det problem der blev løst i lemma 6.3 med $z_i = y_i - \bar{y}$ og $s_i = x_i - \bar{x}$. Vi ser derfor at vi for fastholdt σ^2 maksimerer $\tilde{L}_y(\beta, \sigma^2)$ i

$$\hat{\beta}(\sigma^2) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{SPD}_{xy}}{\text{SSD}_x}.$$

Vi konstaterer at denne størrelse slet ikke afhænger af σ^2 . Det følger at for fast σ^2 vil (6.3) blive maksimeret af

$$\hat{\alpha} = \bar{y} - \bar{x} \frac{\text{SPD}_{xy}}{\text{SSD}_x}, \quad \hat{\beta} = \frac{\text{SPD}_{xy}}{\text{SSD}_x}.$$

Indsættes disse værdier i (6.3), fås en likelihood hvor både α og β er profileret ud og som altså kun har σ^2 som argument,

$$\tilde{\tilde{L}}_y(\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2\right).$$

Men en funktion af denne type blev maksimeret i lemma 4.4, så vi kan direkte aflæse at maksimum bliver antaget i

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

Dermed har vi alt i alt vist det ønskede. \square

Ved at indsætte estimaterne får vi den estimerede regressionslinje

$$\hat{y}(x) = \hat{\alpha} + \hat{\beta}x = \bar{y} - \frac{\text{SPD}_{xy}}{\text{SSD}_x} \bar{x} + \frac{\text{SPD}_{xy}}{\text{SSD}_x} x.$$

Bemærk specielt at $\hat{y}(\bar{x}) = \bar{y}$. Det betyder at den estimerede regressionslinje går gennem punktet (\bar{x}, \bar{y}) bestående af gennemsnittene af de to variable. Bemærk også at der i det vigtige specialtilfælde hvor $\bar{x} = 0$, gælder at $\hat{\alpha} = \bar{y}$.

Næste trin i analysen af den lineære regressionsmodel er at forstå hvordan $\hat{\alpha}$, $\hat{\beta}$ og $\hat{\sigma}^2$ opfører sig når vi betragter dem som stokastiske variable, dvs. når vi tænker på den som afledt af Y_1, \dots, Y_n snarere end y_1, \dots, y_n .

Sætning 6.5. *De marginale fordelinger af maksimaliseringsestimaterne for middelværdiparametrene i en lineær regressionsmodel er*

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x}\right)\right), \quad \hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\text{SSD}_x}\right).$$

Bevis Vi starter med at finde fordelingen af den stokastiske variabel

$$\text{SPD}_{xY} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}).$$

Idet vi observerer at

$$\sum_{i=1}^n (x_i - \bar{x}) = 0, \tag{6.4}$$

har vi at

$$\text{SPD}_{xY} = \sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y} \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})Y_i. \tag{6.5}$$

Vi ser således at SPD_{xY} er en linearkombination af de uafhængige, normalfordelte variable Y_1, \dots, Y_n . Det følger af BH, eksempel 6.6.3 at spd_{xY} er normalfordelt. Ved

hjælp af regnereglerne for middelværdi, får vi

$$\begin{aligned}
 E(\text{SPD}_{xY}) &= \sum_{i=1}^n (x_i - \bar{x}) E Y_i = \sum_{i=1}^n (x_i - \bar{x}) (\alpha + \beta x_i) \\
 &= \alpha \sum_{i=1}^n (x_i - \bar{x}) + \beta \sum_{i=1}^n (x_i - \bar{x}) x_i \\
 &= \beta \sum_{i=1}^n (x_i - \bar{x}) (x_i - \bar{x}) + \beta \bar{x} \sum_{i=1}^n (x_i - \bar{x}) \\
 &= \beta \cdot \text{SSD}_x
 \end{aligned}$$

ved gentagen brug af (6.4). Vi kan også udregne variansen som

$$\text{Var}(\text{SPD}_{xY}) = \sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}(Y_i) = \text{SSD}_x \cdot \sigma^2$$

idet alle Y_i 'erne jo har samme varians σ^2 . Opsummerende er

$$\text{SPD}_{xY} \sim N(\text{SSD}_x \cdot \beta, \text{SSD}_x \cdot \sigma^2),$$

og dermed er

$$\hat{\beta} = \frac{\text{SPD}_{xY}}{\text{SSD}_x} \sim N\left(\beta, \frac{\sigma^2}{\text{SSD}_x}\right)$$

som ønsket. Her har vi brugt at en skalafaktor går direkte ind på middelværdien, mens variansen skal kvadreres.

Vi finder fordelingen af $\hat{\alpha}$ på helt tilsvarende vis. Vi starter med at indse at

$$\hat{\alpha} = \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}}{\text{SSD}_x} (x_i - \bar{x}) \right) Y_i \quad (6.6)$$

hvor vi har brugt opskrivningen af SPD_{xY} fra før. Dermed er $\hat{\alpha}$ en linearkombination af Y_i 'erne, og den er derfor normalfordelt. Vi ser at

$$\begin{aligned}
 E(\hat{\alpha}) &= \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}}{\text{SSD}_x} (x_i - \bar{x}) \right) (\alpha + \beta x_i) \\
 &= \alpha + \beta \bar{x} - \frac{\bar{x}}{\text{SSD}_x} \sum_{i=1}^n (x_i - \bar{x}) (\alpha + \beta x_i) \\
 &= \alpha.
 \end{aligned}$$

Tilsvarende ser vi at

$$\begin{aligned}\text{Var}(\hat{\alpha}) &= \sum_{i=1}^n \left(\frac{1}{n} - \frac{\bar{x}}{\text{SSD}_x} (x_i - \bar{x}) \right)^2 \sigma^2 \\ &= \sum_{i=1}^n \left(\frac{1}{n^2} + \frac{\bar{x}^2}{\text{SSD}_x^2} (x_i - \bar{x})^2 - \frac{2\bar{x}}{n \text{SSD}_x} (x_i - \bar{x}) \right) \sigma^2 \\ &= \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x} \right) \sigma^2\end{aligned}$$

hvilket præcis var hvad vi ønskede. \square

Sætning 6.5 giver os de marginale fordelinger af $\hat{\alpha}$ og $\hat{\beta}$, men ikke den simultane fordeling. Man kan vise at $\hat{\alpha}$ og $\hat{\beta}$ uafhængige stokastiske variable hvis og kun hvis $\bar{x} = 0$. Man kan desuden vise at \bar{Y} og $\hat{\beta}$ er uafhængige uanset værdien af \bar{x} .

Næste sætning udtaler sig om fordelingen af estimatoren for variansen. Beviset springes over.

Sætning 6.6. *Den marginale fordeling af maksimaliseringsestimatoren for variansparameteren i en lineær regressionsmodel er givet ved*

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi_{n-2}^2$$

Der gælder endvidere at den todimensionale variabel $(\hat{\alpha}, \hat{\beta})$ er uafhængig af $\hat{\sigma}^2$.

Det fremgår direkte af sætning 6.5 at $\hat{\alpha}$ og $\hat{\beta}$ er centrale estimators for α og β . Derimod er

$$E(\hat{\sigma}^2) = \frac{n-2}{n} \sigma^2$$

så $\hat{\sigma}^2$ er ikke en central estimator for σ^2 . I gennemsnit estimeres σ^2 for lavt hvis vi benytter maksimum likelihood estimatoren. Dette svarer til hvad vi så i kapitel 4 og 5, og det er også i dette tilfælde nemt at korrigere $\hat{\sigma}^2$ og opnå et centralt estimat: vi skal blot normere med $n-2$ i stedet for n i definitionen af $\hat{\sigma}^2$, og i stedet bruge

$$\tilde{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

som estimator. Så er $\tilde{\sigma}^2 \sim \frac{\sigma^2}{n-2} \chi_{n-2}^2$, og specielt er $E(\tilde{\sigma}^2) = \sigma^2$ som ønsket. Det tilsvarende estimat hvor observationerne sættes ind betegnes som regel s^2 , dvs.

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

Den følgende bemærkning præciserer at det er dette estimat man benytter.

Bemærkning 6.7. *I den statistiske model fra definition 6.1 bruger vi estimererne*

$$\hat{\alpha} = \bar{y} - \bar{x} \frac{\text{SPD}_{xy}}{\text{SSD}_x}, \quad \hat{\beta} = \frac{\text{SPD}_{xy}}{\text{SSD}_x}, \quad s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

De sande eller teoretiske fordelinger af $\hat{\alpha}$, $\hat{\beta}$ og $\hat{\sigma}^2$ er beskrevet ovenfor, men afhænger som altid af de ukendte parametre. Vi får estimerede spredninger (standard errors) for middelværdiestimatorerne hvis vi erstatter den sande spredning σ med estimatet s i udtrykket for estimatorernes spredning:

$$\text{SE}(\hat{\alpha}) = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x}}, \quad \text{SE}(\hat{\beta}) = \frac{s}{\sqrt{\text{SSD}_x}}$$

Eksempel 6.8. *(Vcf og blodglukose, fortsættelse af eksempel 6.2, side 113)* For de 23 observationer x_1, \dots, x_{23} af blodglukosen og de tilsvarende 23 observationer y_1, \dots, y_{23} af Vcf, får man de summariske størrelser

$$\bar{x} = 10.374; \quad \bar{y} = 1.326; \quad \text{SSD}_x = 429.704; \quad \text{SPD}_{xy} = 9.437,$$

således at estimererne er

$$\hat{\alpha} = 1.326 - 10.374 \frac{9.437}{429.704} = 1.098; \quad \hat{\beta} = \frac{9.437}{429.704} = 0.0220$$

og

$$s^2 = \frac{\sum_{i=1}^{23} (y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))^2}{23 - 2} = 0.0470, \quad s = 0.2167.$$

Den estimerede regressionslinje, $\hat{y}(x) = \hat{\alpha} + \hat{\beta}x$, er indtegnet på figur 6.1 på side 114. Den estimerede spredning for $\hat{\alpha}$ kan beregnes til 0.1175, mens den estimerede spredning for $\hat{\beta}$ er 0.0105. \square

6.3 Konfidensintervaller

Konfidensintervaller kan findes på samme måde som vi allerede har set det i afsnit 3.3, 4.3 og 5.3. Diskussionerne fra de tidligere afsnit vedrørende konfidensintervaller er selvfølgelig stadig gyldige.

Sætning 6.9. *Betragt den statistiske model fra definition 6.1. Så er*

$$\hat{\alpha} \pm t_{n-2, 1-\alpha^*/2} \cdot \tilde{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x}} \quad (6.7)$$

et $1 - \alpha^*$ konfidensinterval for α , og

$$\hat{\beta} \pm t_{n-2, 1-\alpha^*/2} \frac{\tilde{\sigma}}{\sqrt{\text{SSD}_x}} \quad (6.8)$$

et $1 - \alpha^*$ konfidensinterval for β .

Bemærk at vi nu skriver α^* for signifikansniveauet for at skelne denne fra parameteren α .

Bevis Vi beviser kun (6.8), da (6.7) bevises på nøjagtig samme måde. Det følger af sætning 6.5 og 6.6 at

$$U = \frac{\hat{\beta} - \beta}{\sigma / \sqrt{\text{SSD}_x}} \sim N(0, 1), \quad Z = \frac{n-2}{\sigma^2} \tilde{\sigma}^2 \sim \chi_{n-2}^2,$$

og at U og Z er uafhængige. Det følger da af definitionen af t -fordelingen (definition A.4 i appendiks A) at

$$T = \frac{U}{\sqrt{Z/(n-2)}} = \frac{\sqrt{\text{SSD}_x}(\hat{\beta} - \beta)}{\tilde{\sigma}}$$

er t -fordelt med $n - 2$ frihedsgrader. Således er

$$P\left(-t_{n-2, 1-\alpha^*/2} < \frac{\sqrt{\text{SSD}_x}(\hat{\beta} - \beta)}{\tilde{\sigma}} < t_{n-2, 1-\alpha^*/2}\right) = 1 - \alpha^*$$

eller, hvis vi isolerer β i midten,

$$P\left(\hat{\beta} - t_{n-2, 1-\alpha^*/2} \frac{\tilde{\sigma}}{\sqrt{\text{SSD}_x}} < \beta < \hat{\beta} + t_{n-2, 1-\alpha^*/2} \frac{\tilde{\sigma}}{\sqrt{\text{SSD}_x}}\right) = 1 - \alpha^*.$$

Dette viser netop at (6.8) er et konfidensinterval for α med konfidensgrad $1 - \alpha^*$. \square

Eksempel 6.10. (Vcf og blodglukose, fortsættelse af eksempel 6.2, side 113) Vi skal bruge 97.5% fraktilen i t -fordelingen med $n - 2 = 21$ frihedsgrader. Den viser sig at være 2.08. Således er

$$1.098 \pm 2.08 \cdot 0.2167 \cdot \sqrt{\frac{1}{23} + \frac{10.374^2}{429.704}} = 1.098 \pm 0.244 = (1.232, 1.420)$$

$$0.0220 \pm 2.08 \cdot \frac{0.2167}{\sqrt{429.704}} = 0.0220 \pm 0.0217 = (0.0002, 0.0437)$$

95% konfidensintervaller for α og β .

Hvis $\beta = 0$, svarende til at Vcf ikke afhænger af blodglukosen, er det således lidt usandsynligt at vi skulle have observeret de data vi har til rådighed. Bemærk dog at konfidensintervallet for β er tæt på at indeholde nul. \square

6.4 Hypotesetest

I en lineær regression er man ofte interesseret i at teste om responsvariablen overhovedet afhænger af den målte baggrundsvARIABLE x . Vi vil derfor betragte hypotesen om at middelværdien af Y ikke afhænger af x . Det er det samme som at teste om $\beta = 0$. Vi skriver hypotesen som

$$H : \beta = 0, \text{ eller } (\alpha, \beta, \sigma^2) \in \Theta_0 = \mathbb{R} \times \{0\} \times (0, \infty).$$

Under hypotesen er alle $Y_i \sim \mathcal{N}(\alpha, \sigma^2)$, dvs. vi er tilbage i situationen fra afsnit 4 med en enkelt stikprøve.

Ligesom i afsnit 4.4 og 5.4 er hypotesen ikke en simpel hypotese da parametermængden under hypotesen, Θ_0 , indeholder mere end et enkelt punkt. Vi vil på nøjagtig samme måde som i de tidligere afsnit gøre følgende:

- Estimere $(\alpha, \beta, \sigma^2)$ under hypotesen, dvs. bestemme $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2) \in \Theta_0$ så

$$L_y(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2) \geq L_y(\alpha, \beta, \sigma^2), \quad (\alpha, \beta, \sigma^2) \in \Theta_0.$$

Det er klart at $\hat{\beta} = 0$ da det er den eneste mulige værdi.

- Opskrive kvotientteststørrelsen

$$Q(y) = \frac{L_y(\hat{\alpha}, 0, \hat{\sigma}^2)}{L_y(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)}.$$

- Bestemme testsandsynligheden

$$\varepsilon(y) = P(Q(Y) \leq Q(y)).$$

- Afvise hypotesen hvis $\varepsilon(y) < \alpha^*$ for et på forhånd fastsat signifikansniveau og i givet fald konkludere at β er signifikant forskellig fra 0 — med andre ord at responsvariablen afhænger af baggrundsvARIABLEN.

Sætning 6.11. *Betragt den statistiske model givet i definition 6.1 og hypotesen $H : \beta = 0$. Under hypotesen er maksimum likelihood estimatet $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ givet ved*

$$\hat{\alpha} = \bar{y}, \quad \hat{\beta} = 0, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2,$$

og fordelingerne af de tilsvarende stokastiske variable er $\hat{\alpha} \sim N(\alpha, \sigma^2/n)$ og $\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi_{n-1}^2$, og de er uafhængige. Kvotientteststørrelsen er givet ved

$$Q(y) = \left(\frac{\hat{\sigma}^2}{\hat{\alpha}} \right)^{n/2}$$

og kvotienttestet kan udføres på

$$t = \frac{\hat{\beta}}{s/\sqrt{\text{SSD}_x}}.$$

p -værdien er givet ved

$$\varepsilon(y) = 2P(T \geq |t|) = 2 \cdot (1 - F_{t_{n-2}}(|t|))$$

hvor T er t -fordelt med $n - 2$ frihedsgrader og $F_{t_{n-2}}$ er fordelingsfunktionen for denne fordeling.

Bevis Under hypotesen har vi modellen fra definition 4.1 for en enkelt stikprøve med ukendt varians, og vi får derfor direkte fra sætning 4.3 estimatorerne og deres fordeling.

Vi regner derefter på kvotientteststørrelsen $Q(y)$. Bemærk at

$$-\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = -\frac{n}{2}$$

og at

$$-\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (y_i - \bar{y})^2 = -\frac{n}{2}$$

således at eksponentialleddene i tælleren og nævneren af $Q(y)$ er ens. Vi får således

$$Q(y) = \frac{L_y(\hat{\alpha}, 0, \hat{\sigma}^2)}{L_y(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2} \right)^{n/2}.$$

Vi mangler at vise at kvotienttestet kan udføres som et test på t , dvs. at vise udtrykket for p -værdien. For at lette notationen, indfører vi størrelserne

$$u = \frac{\sqrt{\text{SSD}_x} \hat{\beta}}{\sigma}, \quad z = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Så er

$$t = \frac{u}{\sqrt{z/(n-2)}},$$

og

$$(Q(y))^{2/n} = \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sigma^2 z}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6.9)$$

Husk at $\hat{\alpha} + \hat{\beta} \bar{x} = \bar{y}$. Derfor er

$$y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}) = y_i - \hat{\alpha} - \hat{\beta} x_i,$$

og summen i nævneren af (6.9) kan omskrives til

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n \left(y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}) + \hat{\beta}(x_i - \bar{x}) \right)^2 \\ &= \sum_{i=1}^n \left(y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}) \right)^2 + \sum_{i=1}^n \hat{\beta}^2 (x_i - \bar{x})^2 \\ &\quad + 2 \sum_{i=1}^n \left(y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}) \right) \hat{\beta}(x_i - \bar{x}) \\ &= \sigma^2 z + \hat{\beta}^2 \text{SSD}_x + 2\hat{\beta} \sum_{i=1}^n \left(y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}) \right) (x_i - \bar{x}) \\ &= \sigma^2 z + \sigma^2 u^2 + 2\hat{\beta} (\text{SPD}_{xy} - \hat{\beta} \text{SSD}_x) \\ &= \sigma^2 z + \sigma^2 u^2. \end{aligned}$$

Vi får således, nøjagtigt som i de forrige kapitler, at

$$(Q(y))^{2/n} = \frac{\sigma^2 z}{\sigma^2 z + \sigma^2 u^2} = \left(1 + \frac{u^2}{z}\right)^{-1} = \left(1 + \frac{t^2}{n-2}\right)^{-1},$$

dvs. at $Q(y)$ er en aftagende funktion af t^2 . Hvis vi betegner de tilhørende stokastiske variable med $Q(Y)$ og T , har vi derfor

$$\varepsilon(y) = P(Q(Y) \leq Q(y)) = P(T^2 \geq t^2) = 2P(T \geq |t|)$$

Her er

$$T = \frac{U}{\sqrt{Z/(n-2)}},$$

hvor U og Z er de stokastiske variable hvis udfald er u og z , dvs.

$$U = \frac{\sqrt{\text{SSD}_x} \hat{\beta}}{\sigma}, \quad Z = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2,$$

Under hypotesen, dvs. når $\beta = 0$, følger det af sætning 6.5 og 6.6 at $U \sim N(0, 1)$ og $Z \sim \chi_{n-2}^2$, og at de er uafhængige. Det følger således af definitionen på en t -fordeling (definition A.4 i appendiks A) at T er t -fordelt med $n-2$ frihedsgrader, således at p -værdien skal beregnes i t -fordelingen. \square

På samme vis som vi har set i de foregående kapitler består testet altså i at beregne den observerede værdi af T -teststørrelsen, dvs. t , og beregne hvor ekstremt værdien ligger i t -fordelingen med $n-2$ frihedsgrader. Som før giver dette intuitivt god mening: Hypotesen bør afvises hvis $\hat{\beta}$ afviger meget fra nul og bør således baseres på $|\hat{\beta}|$. Division med den estimerede spredning kan opfattes som en normering der transformerer teststørrelsen til en kendt skala og således tager højde for variationen i data.

Ligesom ved de tidligere hypotesetest, plejer man at opdatere estimerne hvis hypotesen ikke kan afvises, dvs. angive estimatet for α til \bar{y} , β til 0 og estimatet for σ^2 til $\text{SSD}_y / (n-1)$.

Kommentarerne fra afsnit 3.4 vedrørende sprogbug, fejltyper og sammenhængen mellem konfidensintervaller og hypotesetest gælder uændret. Specielt vil $1 - \alpha^*$ konfidensintervallet for β indeholde værdien 0 hvis og kun hvis hypotesen $H: \beta = 0$ ikke kan afvises på signifikansniveau α^* .

Eksempel 6.12. (Vcf og blodglukose, fortsættelse af eksempel 6.2, side 113) At teste om Vcf afhænger af blodglukosen svarer til hypotesen $H : \beta = 0$. Værdien af t -teststørrelsen er

$$t = \frac{\hat{\beta}}{s/\sqrt{\text{SSD}_x}} = \frac{0.022}{0.217/\sqrt{429.704}} = 2.101$$

og p -værdien er

$$\varepsilon(y) = 2P(T \geq 2.101) = 0.0479$$

hvor $T \sim t_{21}$. Der er således svag evidens mod hypotesen som afvises på 5% signifikansniveau. Bemærk dog at p -værdien er tæt på 0.05, hvilket stemmer overens med at den nedre grænse i konfidensintervallet er tæt på nul. Konklusionen er derfor at Vcf formentlig afhænger af blodglukosen, selvom evidensen ikke er stor. Dette kan enten skyldes at vi rent tilfældigt har observeret data der er lineært i blodglukosen selvom der ikke er en virkelig sammenhæng, men det kan også skyldes at datasættet er for lille til at give statistisk signifikans for en reel sammenhæng. \square

6.5 Regressionslinjen og prædiktion

Som allerede vist, er estimatet for regressionslinjen givet ved

$$\hat{y}(x) = \hat{\alpha} + \hat{\beta}x.$$

Hvis vi betragter $\hat{\alpha}$ og $\hat{\beta}$ som stokastiske variable giver dette for fast x en ny stokastisk variabel

$$\hat{Y}(x) = \hat{\alpha} + \hat{\beta}x.$$

Vi skal nu undersøge fordelingen af denne variabel. Fra de sædvanlige regneregler for middelværdi følger det at

$$E(\hat{Y}(x)) = E(\hat{\alpha}) + E(\hat{\beta})x = \alpha + \beta x.$$

Det er sværere at bestemme variansen fordi $\hat{\alpha}$ og $\hat{\beta}$ ikke er uafhængige (medmindre $\bar{x} = 0$), men det faktisk muligt at bestemme varians og fordeling af $\hat{Y}(x)$.

Sætning 6.13. For givet x er fordelingen af $\hat{Y}(x)$ givet ved

$$\hat{Y}(x) \sim N\left(\alpha + \beta x, \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\text{SSD}_x}\right)\right).$$

Bevis Husk at $\hat{\alpha} + \hat{\beta}\bar{x} = \bar{Y}$ (regressionslinjen går gennem gennemsnitspunktet), således at

$$\hat{Y}(x) = \hat{\alpha} + \hat{\beta}\bar{x} + \hat{\beta}(x - \bar{x}) = \bar{Y} + \hat{\beta}(x - \bar{x}).$$

Efter sætning 6.5 blev det bemærket at \bar{Y} og $\hat{\beta}$ er uafhængige. Begge stokastiske variable er desuden normalfordelte, så det følger at $\hat{Y}(x)$ også er normalfordelt. Vi har allerede fundet middelværdien, men mangler at bestemme variansen. På grund af uafhængigheden får vi

$$\text{Var}(\hat{Y}(x)) = \text{Var}(\bar{Y}) + \text{Var}\left(\hat{\beta}(x - \bar{x})\right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{\text{SSD}_x}(x - \bar{x})^2,$$

hvilket fuldender beviset.

Man kan faktisk godt slippe udenom at benytte resultatet om uafhængighed mellem \bar{Y} og $\hat{\beta}$. Ved at kombinere (6.5) og (6.6) og reducere, får vi

$$\hat{Y}(x) = \sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x - \bar{x})}{\text{SSD}_x} \right) Y_i.$$

Dette er en linearkombination af de uafhængige stokastiske variable Y_1, \dots, Y_n . Linearkombinationen er igen normalfordelt, og de sædvanlige regneregler giver efter lidt regneri middelværdi og varians. \square

Sætningen viser at $\hat{Y}(x)$ er en central estimator for $\alpha + \beta x$, dvs. for middelværdien af Y for fast x . Bemærk at variansen for $\hat{Y}(x)$ afhænger af x og er mindst for $x = \bar{x}$, hvilket giver god mening: regressionslinjen er bedst bestemt i det område hvor vi har flest observationer, hvorimod usikkerheden stiger jo længere vi kommer væk fra \bar{x} .

På samme måde som vi så i afsnit 6.3, kan vi konstruere et konfidensinterval for værdierne på regressionslinjen ved at betragte de stokastiske variable

$$\begin{aligned} U &= \frac{\hat{Y}(x) - \alpha - \beta x}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\text{SSD}_x}}} \sim N(0, 1), \\ Z &= \frac{(n-2)s^2}{\sigma^2} \sim \chi_{n-2}^2, \\ T &= \frac{U}{\sqrt{Z/(n-2)}} \sim t_{n-2}. \end{aligned}$$

Prøv selv at gennemføre argumenterne. Vi får følgende konfidensinterval for regressionslinjen med konfidensgrad α^* :

$$\hat{Y}(x) \pm t_{n-2, 1-\alpha^*/2} \tilde{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\text{SSD}_x}}.$$

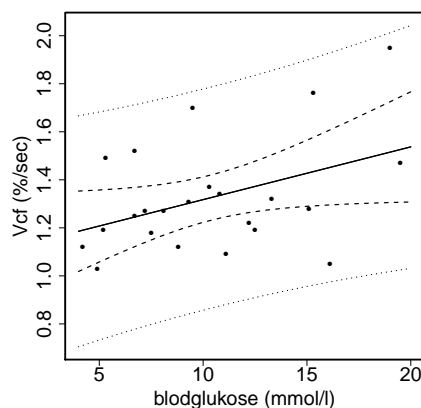
Eksempel 6.14. (*Vcf og blodglukose, fortsættelse af eksempel 6.2, side 113*) Den estimerede regressionslinje er

$$\hat{y}(x) = \hat{\alpha} + \hat{\beta}x = 1.098 + 0.0220x.$$

Et 95% konfidensinterval for regressionslinjen i punktet x er givet ved

$$1.098 + 0.0220x \pm 2.08 \cdot 0.2167 \sqrt{\frac{1}{23} + \frac{(x - 10.374)^2}{429.704}}.$$

I figur 6.2 ses regressionslinjen indtegnet med punktvisse 95% konfidensgrænser som



Figur 6.2: Regressionslinje (fuldt optrukket), punktvisse konfidensintervaller (stiplet) og punktvisse prædiktionsintervaller for blodglukose-Vcf data fra eksempel 6.2. Se også eksempel 6.8, 6.10, 6.12, 6.14 og 6.16.

stiplede kurver. Bemærk hvordan konfidensgrænserne bliver bredere mod siderne i figuren, hvor vi er længere væk fra \bar{x} . At linjen er mere usikker længere væk fra midten er naturligt: hvis man vipper linjen en smule er effekten størst i siderne, hvor vi kun har lidt information. \square

Hvis forsøget havde været større, dvs. hvis n havde været større, ville også SSD_x være større (medmindre de nye x 'er alle er lig \bar{x}). Under alle omstændigheder ville regressionslinjen være mere sikkert bestemt. Dette ses i udtrykket for variansen af estimatet for regressionslinjen. I figur 6.2 ville det betyde at konfidensgrænserne ville ligge tættere omkring regressionslinjen. Dette sker selvom de enkelte observationers variation omkring linjen er den samme. Konfidensgrænserne siger således *intet* om hvor tæt på regressionslinjen vi kan forvente at finde de enkelte observationer, men angiver kun hvor sikre vi kan være på estimatet for middelværdien.

Nogle gange ønsker man faktisk at forudsige værdier af Y for en given værdi af x , herunder at angive et interval hvor en ny observation af Y med en given sandsynlighed vil ramme. Dette kaldes at *prædiktere*, der egentlig betyder at forudsige. Vi kan for eksempel være interesserede i at angive et interval hvor vi forventer at en ny observation vil falde med sandsynligheden 0.95, for en given værdi af x .

Sætning 6.15. *Lad $(x_1, y_1), \dots, (x_n, y_n)$ være sammenhørende observationer fra den statistiske model fra definition 6.1. Som prædiktør for en ny observation Y med tilhørende værdi x , hvor Y er uafhængig af de tidligere observationer, benyttes den estimerede middelværdi,*

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x.$$

Et tilhørende prædiktionsinterval på niveau $1 - \alpha^$ er givet ved*

$$\hat{Y} \pm t_{n-2, 1-\alpha^*/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\text{SSD}_x}}.$$

Bevis Ifølge modellen er $Y \sim N(\alpha + \beta x, \sigma^2)$, og den nye observation antages uafhængig af Y_1, \dots, Y_n , og dermed også af $\hat{\alpha}$ og $\hat{\beta}$, dvs. af \hat{Y} . Vi betragter nu den stokastiske variabel $Y - \hat{Y}$, der umiddelbart ses at have middelværdi nul og varians

$$\text{Var}(Y - \hat{Y}) = \text{Var}(Y) + \text{Var}(\hat{Y}) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\text{SSD}_x} \right).$$

Den er desuden normalfordelt, da det jo er en linearkombination af uafhængige normalfordelte variable. Vi kan således konstruere de stokastiske variable

$$\begin{aligned} U &= \frac{Y - \hat{Y}}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\text{SSD}_x}}} \sim N(0, 1) \\ V &= \frac{(n-2)s^2}{\sigma^2} \sim \chi_{n-2}^2 \\ T &= \frac{U}{\sqrt{V/(n-2)}} \sim t_{n-2}. \end{aligned}$$

Resten af argumenterne er overladt til læseren, da det følger nøjagtig de samme principper som vi allerede har set flere gange. \square

Af formelen for prædiktionsintervallet fremgår det at ligegyldigt hvor stor stikprøven er, vil længden af konfidensintervallet aldrig blive mindre end to gange fraktilen gange den estimerede spredning, s , og s vil for en stor stikprøve være tæt på den sande

værdi σ . Det skyldes at vi kun kan reducere den variation, der er knyttet til vores forsøg, hvorimod vi ikke kan reducere den naturlige variation. Der er således stor forskel på at lave konfidensintervaller for middelværdier og parametre og på at lave prædiktionsintervaller for udfaldet af nye stokastiske variable.

Eksempel 6.16. (*Vcf og blodglukose, fortsættelse af eksempel 6.2, side 113*) Antag at der kommer en ny diabetespatient ind til lægen med en blodglukose på 15 mmol/l. Lægen vil gerne prædiktere patientens Vcf. Vi får

$$\hat{Y} = \hat{\alpha} + \hat{\beta}x = 1.098 + 0.0220 \cdot 15 = 1.427.$$

Et prædiktionsinterval for denne patient er givet ved

$$\begin{aligned} \hat{Y} \pm t_{n-2, 1-\alpha^*/2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SSD_x}} \\ = 1.427 \pm 2.08 \cdot 0.2167 \sqrt{1 + \frac{1}{23} + \frac{(15-10.374)^2}{429.704}} \\ = (0.956, 1.898). \end{aligned}$$

I figur 6.2 ses regressionslinjen indtegnet med punktvis 95% prædiktionsgrænser som prikkede kurver. Som forventet, og som det fremgår af formlerne, er prædiktionsintervallerne altid bredere end konfidensintervallet (de stiplede kurver). Bemærk at alle observationerne falder indenfor kurverne, men i gennemsnit vil vi forvente at 95% af observationerne falder indenfor prædiktionskurverne. \square

6.6 Residualer og modelkontrol

I den lineære regressionsmodel fra definition 6.1 er der gjort nogle antagelser, og vores konklusioner omkring estimatorer, konfidensintervaller og test gælder kun hvis antagelserne er rimelige. Hvis data ikke er genereret af modellen kender vi ikke egenskaberne og kan derfor ikke stole på resultaterne fra analysen. Det er derfor vigtigt at foretage modelkontrol. Antagelserne kan opdeles i antagelser omkring middelværdistrukturen, dvs. den systematiske del af modellen, og antagelser vedrørende den tilfældige del. Vi har følgende antagelser, hvor de tre sidste vedrører den tilfældige variation:

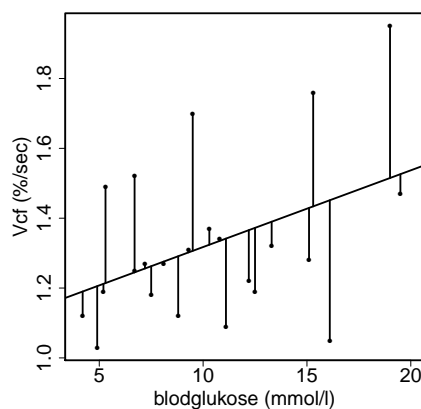
- Middelværdien af Y_i er en lineær funktion af x_i .

- Y_1, \dots, Y_n er uafhængige.
- Y_i er normalfordelt.
- Spredningen af Y_i afhænger ikke af x .

For at kontrollere disse antagelser (undtaget uafhængigheden), definerer vi *residualerne*, der er observationernes afvigelser fra den estimerede regressionslinje:

$$e_i = y_i - \hat{\alpha} - \hat{\beta}x_i.$$

I figur 6.3 er residualerne fra eksempel 6.14 på side 128 om diabetes angivet ved de lodrette streger, der forbinder de observerede værdier med de prædikterede værdier.



Figur 6.3: Residualerne er angivet ved de lodrette afstande til den estimerede regressionslinje. Fra eksempel 6.14 på side 128.

Vi bør kontrollere om residualerne har systematiske afvigelser fra nul, hvilket ikke må forveksles med om punkterne ligger tæt på linjen eller ej. Afstanden fra linjen er et spørgsmål om størrelsen af spredningen. Vi kan også se på om de øvrige antagelser synes at være opfyldt. Vi kan opfatte residualerne som stokastiske variable,

$$E_i = Y_i - \hat{\alpha} - \hat{\beta}x_i.$$

Da residualerne er linearkombinationer af normalfordelte variable er de igen normalfordelte, og de har middelværdi

$$E(E_i) = E(Y_i - \hat{\alpha} - \hat{\beta}x_i) = \alpha + \beta x_i - \alpha - \beta x_i = 0.$$

Variansen har vi ikke redskaber til at udregne på dette kursus, da Y_i jo hverken er uafhængig af $\hat{\alpha}$ eller $\hat{\beta}$. Det betyder at

$$\text{Var}(Y_i - \hat{\alpha} - \hat{\beta}x_i) \neq \text{Var}(Y_i) + \text{Var}(\hat{\alpha}) + \text{Var}(\hat{\beta}x_i).$$

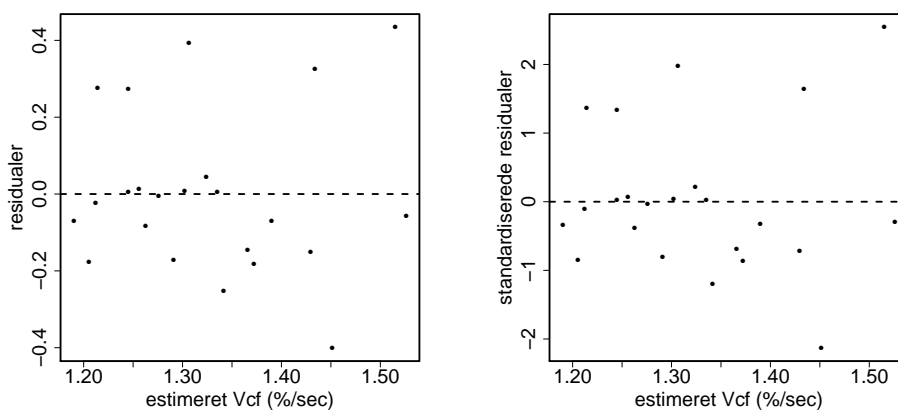
Vi nøjes derfor med at postulere at variansen er givet ved

$$\text{Var}(E_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\text{SSD}_x} \right).$$

Bemærk at variansen aftager med afstanden mellem x_i og \bar{x} , modsat hvad der sker med variansen af den estimerede regressionslinje. Dette skyldes at estimatet for regressionslinjen er mere følsomt overfor punkter, der ligger langt fra \bar{x} end punkter i midten af intervallet. Selv med den samme tilfældige variation over hele intervallet af x -værdier, vil de beregnede residualer i yderpunkterne derfor blive mindre. For at tage højde for det, betragter man i stedet de *standardiserede residualer*:

$$R_i = \frac{E_i}{\sigma \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\text{SSD}_x}}}$$

der er standard normalfordelte, $R_i \sim N(0, 1)$, og således har samme spredning uanset



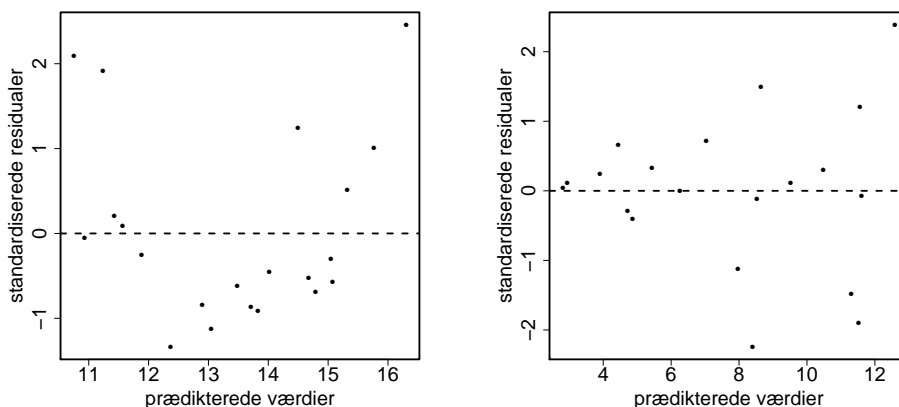
Figur 6.4: Residualer plottet mod prædikterede værdier fra analysen i eksempel 6.14 side 128. Til venstre er det de rå residualer, til højre standardiserede residualer.

værdien af x . Dette gælder vel at mærke hvis antagelserne i modellen er korrekte. I praksis indsættes s som estimat for den ukendte spredning σ .

I figur 6.4 er residualerne, henholdsvis de standardiserede residualer tegnet op mod de prædikterede værdier fra eksempel 6.14 på side 128. Bortset fra enhederne på y-aksen ligner de to figurer hinanden meget, og i dette tilfælde betyder det ikke noget om man ser på de rå residualer eller standardiserer dem først. Det skyldes at der er mange punkter over hele intervallet og linjen derfor er godt bestemt.

Residualerne er ikke uafhængige, men de er “næsten uafhængige”, og man kan i modelkontrollen godt antage tilnærmelsesvis uafhængighed. Modelkontrol kan foretages ved at vurdere om de standardiserede residualer kan antages at være standard normalfordelte. Der er forskellige ting, man skal være opmærksom på, relateret til hver af antagelserne ovenfor.

Linearitetsantagelsen kontrolleres ved at plotte de standardiserede residualer mod de prædikterede værdier som i figur 6.4. Hvis linearitetshypotesen holder, skal punkterne ligge tilfældigt omkring nul, og sprede sig lodret som standard normalfordelte variable uanset hvor på førsteaksen man kigger, idet man ser bort fra afhængigheden mellem residualerne. Det ser i dette tilfælde ud til at passe meget godt. Afvigelser fra dette mønster fortæller noget om, hvad der er galt med hypotesen. Hvis for eksempel residualerne typisk er positive for små og store værdier af de prædikterede værdier, men negative for værdier midt i intervallet, tyder det på at sammenhængen ikke er lineær, men måske kvadratisk eller eksponentiel. Dette er illustreret til venstre i figur 6.5.

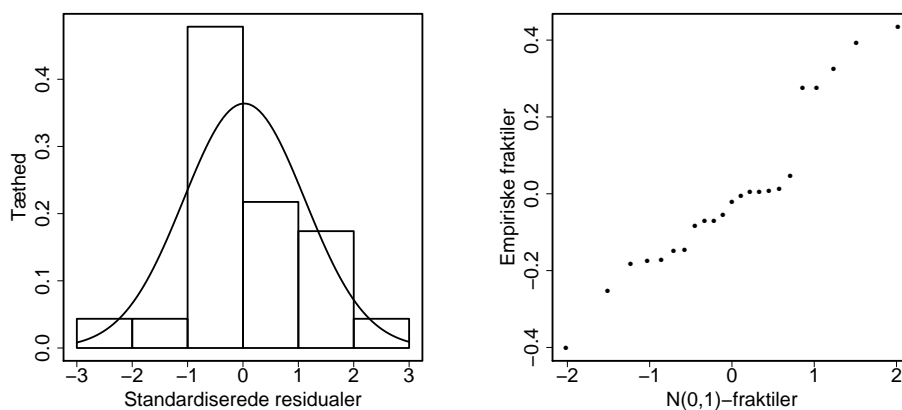


Figur 6.5: Eksempler på residualer plottet mod prædikterede værdier. Til venstre ses et systematisk mønster omkring nul, til højre ses at variansen vokser med middelværdien.

Antagelsen om at spredningen ikke afhænger af middelværdien kan kontrolleres ved

at se på om residualerne fordeler sig i en lige bred sky over hele intervallet. Hvis de for eksempel har "trompetform", tyder det på at spredningen vokser med middelværdien, og antagelsen om samme σ^2 for alle x kan ikke accepteres. Dette er illustreret til højre i figur 6.5. Det kan indimellem løses ved en transformation af data, for eksempel således at $\log(y)$ benyttes som respondvariabel i stedet for y . Sommetider bør den foreklarende variabel x også transformeres. Man skal selvfølgelig huske at kontrollere om antagelserne holder for modellen for de transformerede variable.

Antagelsen om at data er normalfordelte kan kontrolleres ved et histogram eller et QQ-plot af de standardiserede residualer, på samme måde som i afsnit 4.5 og 5.5. Dette er illustreret i figur 6.6 for analysen i eksempel 6.14 side 128 om diabetes. Nor-



Figur 6.6: Histogram og QQ-plot for de standardiserede residualer fra analysen i eksempel 6.14 side 128.

malfordelingsantagelsen er acceptabel i dette eksempel. Bemærk at det ikke giver mening at lave histogrammer og QQ-plots af y 'erne, idet de ikke har samme middelværdi.

6.7 Eksempel: CAPM

I dette afsnit gennemgår vi et eksempel mere om lineær regression. Den interessante hypotese er om regressionslinjen skærer y -aksen i nul.

Eksempel 6.17. (*Capital Asset Pricing Model*) I finansielle sammenhænge benyttes *the Capital Asset Pricing Model* (CAPM) til at bestemme det forventede afkast for et givet aktiv, såsom aktier i en bestemt virksomhed. Modellen blev introduceret i

1960'erne af flere forskellige forskere uafhængigt af hinanden, og udløste Nobelprisen i økonomi i 1990 til Harry Markowitz, Merton Miller og William Sharpe.

Det forventede afkast modelleres som funktion af markedets bevægelser og det såkaldte *risikofri aktiv*. Det risikofri aktiv er afkastet på et aktiv, hvor man på forhånd kender afkastet. Det svarer til at sætte pengene i banken til en kendt rente i stedet for at investere dem. Det siger sig selv at afkastet er mindre end hvad man vil forvente fra en investering, da man jo ikke risikerer noget — ellers er der ingen grund til at investere!

Lad r betegne afkastet af det risikofri aktiv, markedsafkastet betegnes med M , og afkastet af et bestemt aktiv betegnes med R . CAPM antager at

$$E(R - r) = \beta E(M - r). \quad (6.10)$$

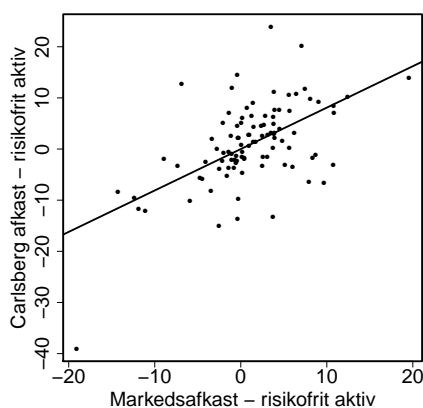
Her er β specifik for det konkrete aktiv vi er interesseret i og repræsenterer hvor kraftigt aktivet reagerer på markedets bevægelser. Man kan dagligt finde estimater af β for en lang række aktiver i finansielle aviser. Estimaterne benyttes af investorer til at sammensætte deres investeringer så hensigtsmæssigt som muligt.

Sammenhørende værdier af de tre størrelser r , M og R kan måles til forskellige tidspunkter, og opgives typisk som månedlige afkast. Vi definerer nu variablene $Y_i = R_i - r_i$ og $x_i = M_i - r_i$, hvor subindex i angiver tidspunktet. Modellen (6.10) passer da ind i modellen for en lineær regression, definition 6.1, bortset fra to ting:

- Det statistiske udsagn i CAPM er at regressionslinjen skærer y -aksen i nul, hvilket svarer til at $\alpha = 0$. Der er gode finansieringsteoretiske argumenter for dette — men dem må I vente med til senere kurser. Vi vil teste om antagelsen virker rimelig udfra data. Bemærk at dette er et andet test end det vi har behandlet tidligere i kapitlet, hvor vi kun har testet for om hældningen β er forskellig fra nul.
- Der er desværre en anden afvigelse, som er sværere at håndtere. Der er ingen grund til at tro at målinger til tætliggende tidspunkter skulle være uafhængige! Hvis for eksempel aktivets afkast har været højt i marts, vil vi også forvente at det ligger højt i april, også udover hvad der kan forklares med markedsafkastet. Vi vil derfor kun analysere data med tre måneders mellemrum og smide de mellemliggende datapunkter væk, i håb om at disse data er nogenlunde uafhængige. Det er ikke nogen optimal løsning, men det bedste vi kan gøre med de redskaber vi har til rådighed på dette kursus. På senere kurser vil metoder til at håndtere afhængighed blive behandlet.

I figur 6.7 er data for Carlsberg aktien i perioden fra juli 1985 til oktober 2009 plottet. Data er månedlige afkast i % — men kun opgivet med tre måneders mellemrum. Data består af 98 sammenhørende målinger af $Y_i = R_i - r_i$, betegnet y_1, \dots, y_{98} , og $x_i = M_i - r_i$, betegnet x_1, \dots, x_{98} . Observationerne y_1, \dots, y_{98} betragtes som realisationer af stokastiske variable Y_1, \dots, Y_{98} som antages at være uafhængige og normalfordelte med varians σ^2 og middelværdier $\alpha + \beta x_i$.

Ud fra figuren kan sammenhængen mellem Carlsberg aktiens afkast og markedsafkastet udmærket være lineær. Derudover ser det ud til at regressionslinjen kunne skære y-aksen i nul, da punktet $(0,0)$ lader til at ligge meget tæt på den estimerede regressionslinje. Bemærk en ekstrem observation nede i venstre hjørne, hvor både Carlsberg aktiens afkast og markedsafkastet er meget negativt. Dette er målingen i oktober 2008 — det tidspunkt hvor den finansielle krise var ved at ramme Danmark, efter at være begyndt i USA.



Figur 6.7: Sammenhæng mellem det månedlige afkast af Carlsberg aktien og markedsafkastet. Den rette linje er den estimerede regressionslinje.

For de 98 observationer x_1, \dots, x_{98} af markedsafkastet og de tilsvarende 98 observationer y_1, \dots, y_{98} af Carlsberg aktiens afkast, viste det sig at

$$\bar{x} = 1.1620; \quad \text{SSD}_x = \sum_{i=1}^{98} (x_i - \bar{x})^2 = 3174.6$$

$$\bar{y} = 0.9392; \quad \text{SPD}_{xy} = \sum_{i=1}^{98} (y_i - \bar{y})(x_i - \bar{x}) = 2571.1$$

således at estimerterne er

$$\hat{\alpha} = -0.001858; \quad \hat{\beta} = \frac{2571.1}{3174.6} = 0.8099$$

og

$$s^2 = \frac{\sum_{i=1}^{98} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{98 - 2} = 45.7966, \quad s = 6.7673.$$

Estimatorernes fordeling er som angivet i sætning 6.5 og 6.6. De estimerede spredninger (standard errors) er

$$SE(\hat{\alpha}) = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SSD_x}} = 6.7673 \cdot \sqrt{\frac{1}{98} + \frac{1.1620^2}{3174.6}} = 0.6977,$$

og

$$SE(\hat{\beta}) = \frac{s}{\sqrt{SSD_x}} = \frac{6.7673}{\sqrt{3174.6}} = 0.1201.$$

Den estimerede regressionslinje, $\hat{y}(x) = \hat{\alpha} + \hat{\beta}(x - \bar{x})$, er indtegnet på figur 6.7.

I figur 6.8 er de sædvanlige modelkontroltegninger plottet. I venstre plot er de standardiserede residualer tegnet op mod de prædikterede værdier. Punkterne lader til at ligge tilfældigt omkring nul, og sprede sig lodret som standard normalfordelte variable uanset hvor på førsteaksen man kigger, som de bør. Bemærk at der er flere punkter tæt ved nul end langt fra, og det er derfor naturligt at se en lidt større spredning her. Der er et enkelt ekstremt residual, som stammer fra den føromtalte måling fra oktober 2008. I højre plot er tegnet et QQ-plot af residualerne. Normalfordelingsantagelsen er acceptabel i dette eksempel, og vi kan roligt fortsætte vores analyser.

For at beregne 95% konfidensintervaller for parametrene, behøver vi 97.5% fraktilen i t -fordelingen med $n - 2 = 96$ frihedsgrader. Den er 1.98. Således er

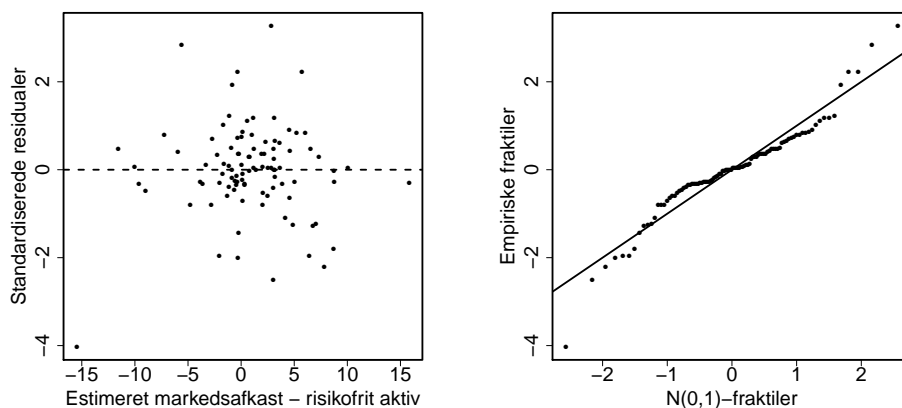
$$-0.001858 \pm 1.98 \cdot 0.6977 = -0.001858 \pm 1.38145 = (-1.3833, 1.3796)$$

og

$$0.8099 \pm 1.98 \cdot 0.1201 = 0.8099 \pm 0.2384 = (0.5715, 1.0483)$$

95% konfidensintervaller for α og β .

Hvis $\beta = 0$ — svarende til at Carlsberg aktien ikke afhænger af markedets øvrige bevægelser — er det således usandsynligt at vi skulle have observeret de data vi har



Figur 6.8: Standardiserede residualer plottet mod prædikterede værdier og et QQ-plot for de standardiserede residualer fra analysen af Carlsberg aktien.

til rådighed, da nul jo ikke er indeholdt i konfidensintervallet for β . Vi ved derfor allerede at et test for om hældningen er nul vil være statistisk signifikant, men vi kan ikke umiddelbart se p -værdien ud fra konfidensintervallet. Resultatet er ikke overraskende — det ville være mærkeligt hvis Carlsberg aktien overhovedet ikke fulgte markedets øvrige bevægelser.

Vi vil nu teste om hældningen kan være nul, som vi har gjort tidligere i kapitlet. Dette svarer til hypotesen $H : \beta = 0$. Værdien af t -teststørrelsen er

$$t = \frac{\hat{\beta}}{s/\sqrt{\text{SSD}_x}} = \frac{0.8099}{6.7673/\sqrt{3174.7}} = 6.7430$$

og p -værdien er

$$\varepsilon(y) = 2P(T \geq 6.7430) = 10^{-9}$$

hvor $T \sim t_{96}$. Der er således stærk evidens mod hypotesen som afvises på 5% signifikansniveau, som vi allerede vidste fra konfidensintervallet. Konklusionen er at Carlsberg aktien følger markedets øvrige bevægelser, som angivet i modellen.

Husk at CAPM antager at regressionslinjen skærer y -aksen i nul, svarende til at midelværdien af afkastet af aktivet ikke er større end det risikofri aktiv, hvis markedets generelle afkast heller ikke er større end det risikofri aktiv. Det er således interessant at teste om α kan antages at være lig $\beta\bar{x}$. For at teste hypotesen $H : \alpha = \beta\bar{x}$ kan vi som vi allerede har set flere gange gøre følgende: estimere parametrene under hypotesen, opskrive kvotientteststørrelsen, bestemme p -værdien og til sidst vurdere om hypotesen kan accepteres eller skal afvises, afhængigt af den fundne p -værdi. Vi vil

skyde en genvej, og den flittige læser kan selv udføre de relevante beregninger og se at man kommer frem til det samme resultat. Dette er stillet som en opgave.

I afsnit 6.5 fandt vi fordelingen af den stokastiske variabel $\hat{Y}(x)$ for fast x , dvs. den estimerede regressionslinje i et givet punkt x . Vi er interesserede i punktet $x = 0$ og har at

$$\hat{Y}(0) \sim N\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x}\right)\right).$$

Vi har derfor at $\hat{Y}(0) \sim N\left(0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x}\right)\right)$ under hypotesen $H : \alpha = 0$. Betragt nu de stokastiske variable

$$\begin{aligned} U &= \frac{\hat{Y}(0)}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x}}} \sim N(0, 1), \\ Z &= \frac{(n-2)s^2}{\sigma^2} \sim \chi_{n-2}^2, \\ T &= \frac{U}{\sqrt{Z/(n-2)}} = \frac{\hat{\alpha}}{s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x}}} \sim t_{n-2}, \end{aligned}$$

hvor de angivne fordelinger er under hypotesen. Fordelingen af T følger fordi U og Z er uafhængige. Vi kan derfor udføre et test på den observerede værdi af T og vurdere den i t -fordelingen med $n - 2$ frihedsgrader, og finder at p -værdien er givet ved

$$\varepsilon(y) = 2P(T \geq |t|) = 2 \cdot (1 - F_{t_{n-2}}(|t|)).$$

I eksemplet med Carlsberg aktien fås t -teststørrelsen

$$t = \frac{\hat{\alpha}}{s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\text{SSD}_x}}} = \frac{-0.001858}{6.7673 \sqrt{\frac{1}{98} + \frac{1.1620^2}{3174.6}}} = -0.0027.$$

og p -værdien er

$$\varepsilon(y) = 2P(T \geq 0.0027) = 0.998.$$

Dette er en meget høj p -værdi! Vi accepterer derfor hypotesen — dvs. disse data giver evidens til CAPM.

Da vi har accepteret hypotesen skal vi opdatere vores estimater. Likelihoodfunktionen

under hypotesen er

$$L_y : \mathbb{R} \times (0, \infty) \rightarrow \mathbb{R}$$

$$L_y(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta x_i)^2\right).$$

På tilsvarende måde som i afsnit 6.2 kan et maksimum likelihood estimat for β findes ved at minimere

$$\sum_{i=1}^n (y_i - \beta x_i)^2$$

der har løsningen

$$\hat{\beta} = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}.$$

Ligeledes kan man nemt finde maksimum likelihood estimatet for σ^2 ,

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta} x_i)^2.$$

Regn selv efter! Som sædvanlig benyttes i stedet det centrale estimat

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{\beta} x_i)^2.$$

Det kan vises at estimatorerne

$$\hat{\beta} = \frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2}, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \hat{\beta} x_i)^2$$

er uafhængige, og at deres marginale fordelinger er givet ved

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right), \quad \hat{\sigma}^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2.$$

Bemærk at vi nu dividerer med $n-1$ i variansestimateret fordi $n\hat{\sigma}^2 \sim \sigma^2 \chi_{n-1}^2$. Vi estimerer kun en middelværdiparameter og mister derfor kun en frihedsgrad. I eksemplet med Carlsberg aktien fås

$$\sum_{i=1}^{98} y_i x_i = 2678.1; \quad \sum_{i=1}^{98} x_i^2 = 3307.0$$

således at estimerterne er

$$\hat{\beta} = \frac{2678.1}{3307.0} = 0.8098, \quad s^2 = \frac{\sum_{i=1}^{98} (y_i - \hat{\beta}x_i)^2}{98 - 1} = 45.3244, \quad s = 6.7323.$$

Den estimerede spredning af $\hat{\beta}$ er $SE(\hat{\beta}) = s/\sqrt{\sum x_i^2} = 0.1171$. Bemærk at estimerterne stort set er de samme som før. Dette er endnu et udtryk for at hypotesen om at regressionslinjen skærer y-aksen i nul er meget plausibel. Den estimerede regressionslinje, $\hat{y}(x) = \hat{\beta}x$ kan faktisk ikke skelnes fra regressionslinjen fra den fulde model indtegnet på figur 6.7.

Det er let at vise at et $1 - \alpha^*$ konfidensinterval for β er givet ved

$$\hat{\beta} \pm t_{n-1, 1-\alpha^*/2} \frac{s}{\sqrt{\sum_{i=1}^n x_i^2}}.$$

For Carlsberg aktien fås at

$$0.8098 \pm 1.98 \frac{6.732}{\sqrt{3307.0}} = (0.5775, 1.0422)$$

er et 95% konfidensinterval for β .

Den estimerede regressionslinje er $\hat{y}(x) = \hat{\beta}x = 0.8098 \cdot x$ og den estimerede fordeling af $\hat{Y}(x)$ er

$$N\left(\hat{\beta}x, s^2 \left(\frac{x^2}{\sum_{i=1}^n x_i^2}\right)\right) = N\left(0.8098 \cdot x, 45.3244 \left(\frac{x^2}{3307.0}\right)\right).$$

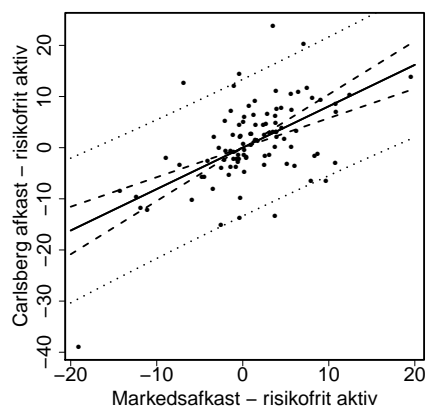
Et 95% konfidensinterval for regressionslinjen i punktet x er givet ved

$$0.8098 \cdot x \pm 1.98 \cdot 6.7323 \sqrt{\frac{x^2}{3307.0}}.$$

I figur 6.9 ses regressionslinjen indtegnet med punktvis 95% konfidensgrænser som stiplede kurver. Bemærk at der ingen usikkerhed er for estimatet af regressionslinjen i $x = 0$. Det skyldes at modellen antager at værdien her er nul — dvs. vi ikke estimerer noget i dette punkt.

Antag at der kommer en ny måling af markedsafkastet og det risikofri aktiv således at differensen er 10%. Vi vil da gerne prædiktere Carlsberg aktiens afkast. Vi får

$$\hat{Y} = \hat{\beta}x = 0.8098 \cdot 10 = 8.098.$$



Figur 6.9: Regressionslinje (fuldt optrukket), punktvisse konfidensintervaller (stiplet) og punktvisse prædiktionsintervaller for Carlsberg aktien.

Et 95% prædiktionsinterval for afkastet af Carlsberg aktien fratrukket det risikofri aktiv er til dette tidspunkt givet ved

$$\begin{aligned}\hat{Y} \pm t_{n-1,0.975} \cdot s \sqrt{1 + \frac{x^2}{SSD_x}} &= 8.098 \pm 1.98 \cdot 6.7323 \sqrt{1 + \frac{10^2}{3307.0}} \\ &= (-5.4642, 21.6606).\end{aligned}$$

I figur 6.9 er de punktvisse 95% prædiktionsgrænser vist som prikkede kurver. Læg mærke til at 7 observationer falder udenfor grænserne. I gennemsnit vil vi forvente at 5% af observationerne falder udenfor prædiktionskurverne, hvilket passer udmærket da vi har 98 observationer i alt.

Til sidst bør bemærkes at modellen ikke kan fange ekstreme begivenheder såsom pludseligt opståede finansielle kriser, der her giver sig udtryk i den ekstreme måling fra oktober 2008. Man kunne overveje at gentage analysen hvor denne måling udelades for at se hvor stor indflydelse den har på resultatet. Hvis resultatet ikke ændrer sig nævneværdigt kan man stadig stole på konklusionerne. \square

6.8 Sammenfatning og perspektiv

Vi har diskuteret statistisk analyse af uafhængige normalfordelte observationer, hvor middelværdien afhænger lineært af en forklarende variabel. Ofte er man interesseret i at estimere sammenhængen, vurdere om der faktisk er en sammenhæng mellem

den forklarende variabel og responsvariablen, eller foretage prædiktioner. Analysen sammenfattes med estimater og konfidensintervaller for parametrene, der indgår i beskrivelsen af middelværdierne. Den naturlige hypotese er ofte om parameteren der angiver sammenhængen med den forklarende variabel er lig nul.

Konstruktionen af konfidensintervaller og udførelsen af hypotesetest er begrebsmæssigt den samme som for modellerne for en og to stikprøver.

Antagelserne for at lave analysen bør altid tjekkes før man drager konklusioner. Antagelserne er at middelværdien afhænger lineært af en forklarende variabel, og at observationerne er uafhængige, normalfordelte og med samme varians uanset værdien af den forklarende variabel.

Til sidst er Capital Asset Pricing modellen blevet gennemgået. Vi har testet for om interceptet kan antages at være nul, og estimater og deres fordelinger er fundet i den reducerede model der skærer y-aksen i nul.

6.9 R

Vi bruger datasættet om sammenhængen mellem blodglukose og Vcf (eksempel 6.2, side 113) som illustration. Data er tilgængelige i filen `vcfdata.txt` med variabelnavne `gluk` og `vcf`. Vi indlæser det i R, så de første linier ser således ud:

```
> vcfdata <- read.table("vcfdata.txt", header=T)
> head(vcfdata)
  gluk  vcf
1 15.3 1.76
2 10.8 1.34
3  8.1 1.27
4 19.5 1.47
5  7.2 1.27
6  5.3 1.49
```

Parameterestimer, konfidensintervaller og test Arbejdshesten ved lineær regression — og også ved mange andre modeller — er funktionen `lm`. Vi kan bruge forskellige kommandoer,

```
lm(vcf ~ gluk, data=vcfdata)
lm(vcfdata$vcf ~ vcfdata$gluk)
```

hvor datasættet enten angives en gang for alle (det nemmeste) eller for hver variabel. Hvis to variable blot lå som x og y , ville koden være `lm(y ~ x)`. Outputtet er følgende:

```
> lm(vcf ~ gluk, data=vcfdata)

Call:
lm(formula = vcf ~ gluk, data = vcfdata)

Coefficients:
(Intercept)      gluk
  1.09781      0.02196
```

På venstre side af `~` skrives responsvariablen, altså den variabel der skal bruges som y , på højre side skrives den forklarende variabel, altså den variabel der skal bruges som x . Vi aflæser estimerne fra outputtet: $\hat{\alpha} = 1.09781$ og $\hat{\beta} = 0.02196$. Estimatet for α står under `(Intercept)` for α er jo netop skæringen med x -aksen. Estimatet for β står under `gluk` fordi hældningsparameteren beskriver effekten af denne variabel.

Umiddelbart giver `lm` kun estimerne, men i virkeligheden laver kaldet et modelobjekt som kombineret med andre funktioner nemt giver os de størrelser vi måtte have brug for. Det nemmeste er at give modelobjektet et navn og så arbejde videre med det. Nedenfor defineres for eksempel modelobjektet `model`. Funktionen `summary` er særligt vigtig:

```
> model <- lm(vcf ~ gluk, data=vcfdata)
> summary(model)

Call:
lm(formula = vcf ~ gluk)

Residuals:
    Min       1Q   Median       3Q      Max
-0.40141 -0.14760 -0.02202  0.03001  0.43490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.09781    0.11748   9.345 6.26e-09 ***
gluk         0.02196    0.01045   2.101  0.0479 *
---

```



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

```
Residual standard error: 0.2167 on 21 degrees of freedom
```

```
Multiple R-squared: 0.1737, Adjusted R-squared: 0.1343
```

```
F-statistic: 4.414 on 1 and 21 DF, p-value: 0.0479
```

Den vigtigste del af outputtet står under `Coefficients`. Der er en linje for α , (`Intercept`), og en linje for β , `gluk`. For hver af parametrene angives estimatet, men også den estimerede spredning for estimatoren (standard error), værdien af t -teststørrelsen for hypotesen om at den tilhørende parameter er nul, og p -værdien for dette test. For β aflæser vi den estimerede spredning til 0.01045. Hypotesen $H: \beta = 0$ giver anledning til $t = 2.101$ og en p -værdi på 0.0479. Vi genkender tallene fra eksempel 6.8 og 6.12 (side 120 og 126). Tilsvarende giver hypotesen $H: \alpha = 0$ anledning til $t = 9.345$ og en p -værdi der er mindre end $6 \cdot 10^{-9}$. Dette er uinteressant — vi interesserer os slet ikke for denne hypotese — men det kan R jo ikke vide.

Nederst i outputtet finder vi `Residual standard error`, dvs. estimatet s for σ , her $s = 0.2167$ og antallet af frihedsgrader, her 21, sammenlign igen med eksempel 6.8 (side 120). Øverst i outputtet finder vi summariske oplysninger om residualerne.

Konfidensintervaller for α og β kan beregnes manuelt ved hjælp af estimator, estimerede spredninger og en t -fraktil. For β får vi for eksempel:

```
> 0.02196 - qt(0.975, 21)*0.01045 # Nedre grænse i KI
[1] 0.0002280353
> 0.02196 + qt(0.975, 21)*0.01045 # Øvre grænse i KI
[1] 0.04369196
```

Endnu nemmere er det at bruge funktionen `confint`, som giver konfidensintervallet for begge parametre. Konfidensgraden kan ændres med argumentet `level`.

```
> confint(model)
                2.5 %      97.5 %
(Intercept) 0.8534993816 1.34213037
gluk         0.0002231077 0.04370194
```

Plot af data Et scatterplot med data laves med `plot`. Regressionslinjen tilføjes nemmest bruge `abline`. Følgende kommandoer giver (pånær layout) grafen i figur 6.1:

```
plot(vcfdata$gluk, vcfdata$vcf) # Scatterplot
plot(vcf~gluk, data=vcfdata)   # Scatterplot igen
abline(model)                  # Estimeret regr.linje
```

Bemærk at scatterplottet kan laves på to måder, og at rækkefølgen af de to variable er forskellig afhængig af om hvilken syntaks man benytter.

Prædiktion Værdier på den fittede regressionslinje og tilhørende konfidens- eller prædiktionsintervaller fås med funktionen `predict`. Hvis vi er interesserede i prædiktioner for glukoseværdierne 8 og 15 kan vi bruge følgende kommandoer:

```
> newdata <- data.frame(gluk=c(8,15))      # Nyt datasæt
> newdata
  gluk
1    8
2   15
> predict(model, newdata, interval="confidence") # KI
      fit      lwr      upr
1 1.273515 1.166310 1.380720
2 1.427253 1.289617 1.564888
> predict(model, newdata, interval="predict")   # PI
      fit      lwr      upr
1 1.273515 0.8102958 1.736734
2 1.427253 0.9560598 1.898446
```

Den første kommando konstruerer et nyt R-datasæt med en enkelt variabel, `gluk`, og to værdier af denne variabel, nemlig 8 og 15. Derefter beregnes konfidensintervallerne og prædiktionsintervallerne i den fittede model. Sammenlign med figur 6.2, og genkend specielt prædiktionsintervallet for glukoseniveau 15 fra eksempel 6.16 (side 130).

Residualer og modelkontrol De estimerede værdier, de rå residualer og de standardiserede residualer fås ved hjælp af `fitted`, `residuals` og `rstandard`. De derved genererede vektorer kan derefter bruges med grafikfunktioner på sædvanlig måde:

```
fit <- fitted(model)      # Estimerede værdier
```

```

rawres <- residuals(model) # Rå residualer
stdres <- rstandard(model) # Standardiserede residualer

plot(fit, stdres)          # Residualplot
qqnorm(stdres)            # QQ-plot af std. residualer


```

På nær layout laver disse kommandoer plottene fra figur 6.6.

6.10 Opgaver

6.1 Antag at vi har observationer $x = (x_1, \dots, x_n)$ og $y = (y_1, \dots, y_n)$, og lad som sædvanlig $\hat{\alpha}$ og $\hat{\beta}$ være estimerne fra den lineære regressionsmodel af y på x , se sætning 6.4.

1. Antag at hvert x_i erstattes med $x'_i = 2x_i$ og at vi udfører regressionen af y på x' (altså bruger x' i stedet for x i modellen). Hvilken indflydelse har det på estimerne? Hvilken indfyldelse har det på testet for hypotesen om at der ingen sammenhæng er mellem de to variable?
2. Antag i stedet at hvert y_i erstattes med $y'_i = 3y_i$ og at vi udfører regressionen af y' på x . Besvar samme spørgsmål som før.

6.2  For et fødevarerprodukt ønsker man at undersøge hvordan koncentrationen af et bestemt stof ændrer sig som funktion af den temperatur som produktet opbevares ved. Derfor har man for fem forskellige temperaturer omkring frysepunktet målt koncentrationen af stoffet. Data er gengivet nedenfor.

Temp.	Konc.	Temp.	Konc.
-5°C	6.6	1°C	8.8
-3°C	7.6	3°C	10.5
-1°C	8.8		

Vi skal betragte den lineære regressionsmodel hvor middelværdien af koncentrationen beskrives som en lineær funktion af temperaturen, dvs.

$$E(Y) = \alpha + \beta x,$$

1. Forklar i termer af temperatur og koncentration hvad fortolkningen er af parametrene α og β .

2. Indtast data med følgende kommandoer:

```
temp <- c(-5, -3, -1, 1, 3)
konc <- c(6.6, 7.6, 8.8, 8.8, 10.5)
```

3. Udfør følgende kommandoer og forklar hvorfor de beregner SSD_x , SPD_{xy} , $\hat{\beta}$ og $\hat{\alpha}$:

```
SSDx <- var(temp)*4
SPDxy <- sum((temp-mean(temp)) * (konc-mean(konc)))
betahat <- SPDxy / SSDx
alphahat <- mean(konc) - mean(temp)*betahat
```

4. Brug følgende kommando og sammenlign med resultatet af de “manuelle” beregninger fra spørgsmål 3:

```
summary(lm(konc ~ temp))
```

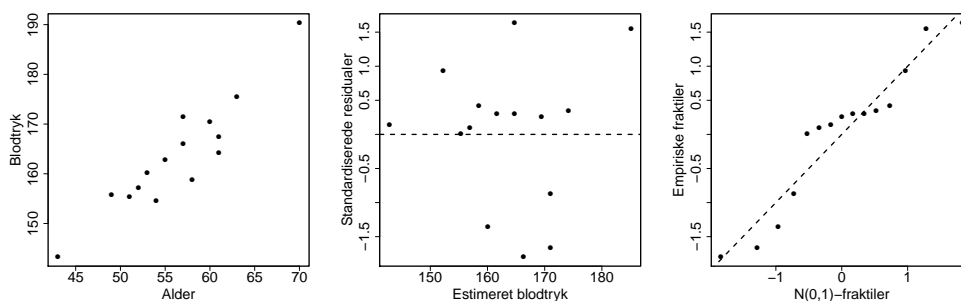
Aflæs desuden de estimerede spredninger for $\hat{\alpha}$ og $\hat{\beta}$, og bestem også variansestimatet s^2 .

6.3 Det antages almindeligvis at blodtrykket stiger med alderen. I tabellen nedenfor er angivet sammenhørende værdier af alder og blodtryk for 15 universitetslærere (Blæsild and Granfeldt, 2003).

Alder	Blodtryk
55	162.9
49	155.9
60	170.5
54	154.5
58	158.9
51	155.4
43	143.3
52	157.3
53	160.2
61	164.3
61	167.5
57	171.5
57	166.0
70	190.3
63	175.5

1. I R-udskriften nedenfor er data analyseret ved hjælp af en lineær regressionsmodel. Opstil den statistiske model. Redegør for forudsætningerne for analysen, og diskuter om disse kan antages at være opfyldt i det foreliggende tilfælde.
2. Hvad er den præcise fortolkning af hældningsparameteren i modellen (forklaret ved hjælp af alder og blodtryk)?
3. Angiv estimater for parametrene i regressionsmodellen og estimatorernes marginale fordeling. Angiv også de estimerede spredninger for $\hat{\alpha}$ og $\hat{\beta}$.
4. Kan man på grundlag af disse observationer opretholde en hypotese om at der ikke er sammenhæng mellem alder og blodtryk? Du kan bruge at 97.5% fraktilen i t_{13} -fordelingen er 2.16.

Ved besvarelsen kan nedenstående uddrag af et R-udskrift samt figurerne benyttes. Data antages at ligge i datasættet `tryk` med de to variable `alder` og `blodtryk`.



Call:

```
lm(formula = blodtryk ~ alder)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	75.5408	9.9976	7.556	4.15e-06 ***
alder	1.5650	0.1766	8.863	7.17e-07 ***

Residual standard error: 4.285 on 13 degrees of freedom

6.4 Data til denne opgave er de samme som til opgave 6.3, og vi skal stadig betragte den lineære regressionsmodel med blodtryk som responsvariabel (y) og alderen som forklarende variabel (x).

1. Find estimaterne for α , β og σ i R-outputtet fra opgave 6.3.
2. Betragt en tilfældig universitetslærer på 50 år. Beregn den prædikterede værdi for vedkommendes forventede blodtryk.
3. Beregn både et 95% konfidensinterval for den forventede værdi og et 95% prædiktionsinterval for den nye observation. Du kan benytte at $\bar{x} = 56.26667$, $SSD_x = 588.9333$, og at 97.5% fraktilen i t_{13} -fordelingen er 2.160.
4. En universitetslærer på 50 får målt sit blodtryk til 170. Er det usædvanligt?
Vink: Hvilken type interval skal du bruge?
5. Antag at tidligere undersøgelser har vist at 50-årige tjenere i gennemsnit har et blodtryk på 170. Giver vores data belæg for at hævde at 50-årige universitetslærere har lavere blodtryk end 50-årige tjenere? *Vink:* Hvilken type interval skal du bruge?

6.5 Denne opgave handler om test af hypotesen $H : \alpha = 0$ (se eksempel 6.17, specielt side 138).

1. Vis at kvotientteststørrelsen for test af hypotesen $H : \alpha = 0$ er givet ved

$$Q(y) = \left(\frac{\hat{\sigma}^2}{\hat{\alpha}^2} \right)^{n/2}$$

og kan udføres på

$$t = \frac{\hat{\alpha}}{s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SSD_x}}}.$$

Du kan bruge at estimaterne i modellen uden intercept er som angivet i delspørgsmål 2.

2. Vis at maksimum likelihood estimatorerne i modellen uden intercept er givet ved


$$\hat{\beta} = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta} x_i)^2.$$

3. Vis at

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right).$$

4. Vis at et $1 - \alpha^*$ konfidensinterval for β er givet ved

$$\hat{\beta} \pm t_{n-1, 1-\alpha^*/2} \frac{s}{\sqrt{\sum_{i=1}^n x_i^2}}.$$

6.6  Data i denne opgave er månedlige afkast i % af Danske Bank aktien, af markedsafkastet og det risikofri aktiv, opgivet hver tredje måned fra juli 1985 til oktober 2009. Data ligger i filen `danskebank.txt`. Foretag samme analyse som analysen af Carlsberg aktien i eksempel 6.17, dvs. svar på følgende spørgsmål.

1. Plot afkastet af Danske Bank aktien minus det risikofri aktiv mod markedsafkastet minus det risikofri aktiv. Diskuter ud fra figuren om en lineær regression virker rimelig.
2. Beregn maksimum likelihood estimaterne for parametrene i modellen fra definition 6.1, og angiv estimatorernes marginale fordelinger.
3. Indtegn den estimerede regressionslinje i scatterplottet fra spørgsmål 1.
4. Beregn 95% konfidensintervaller for α og β .
5. Foretag modelkontrol ved henholdsvis et plot af de standardiserede residualer mod de prædikterede værdier, og et QQ-plot af de standardiserede residualer. Diskuter om modellen synes rimelig.
6. Test hypotesen $H : \beta = 0$.
7. Test hypotesen $H : \alpha = 0$. Kan CAPM modellen accepteres?
8. Opdater estimaterne for parametrene afhængig af konklusionerne i de to forrige spørgsmål, angiv estimatorernes marginale fordelinger. Beregn et 95% konfidensinterval for middelværdiparameteren.
9. Hvad er den estimerede fordeling af regressionslinjen i punktet x ?
10. Antag at der kommer en ny måling af markedsafkastet minus det risikofri aktiv på 15%. Prædikter Danske Bank aktiens afkast og angiv et 95% prædiktionsinterval.

Vink: Det er en god ide at definere nye variable, $x_i = M_i - r_i$ og $Y_i = R_i - r_i$. Det kan i R gøres på følgende måde, hvis data er indlæst i datasættet DBdata:

```
DBdata <- transform(DBdata, Yi = Ri-ri)
DBdata <- transform(DBdata, xi =Mi-ri)
```

De relevante modeller kan fites i R på følgende måde:

- Sædvanlig linear regression $y(x) = a + \beta x$: `lm(Yi ~ xi)`
- Uden intercept $y(x) = \beta x$: `lm(Yi ~ xi-1)`

Man kan få mere information ud ved kommandoen

```
summary(lm(Yi ~ xi, data=DBdata))
```

Konfidensintervaller kan fås ved

```
confint(lm(Yi ~ xi))
```

Et scatterplot med regressionslinjen indtegnet kan konstrueres på følgende måde:

```
plot(DBdata$xi, DBdata$Yi,
      xlab="Markedsafkast - risikofrit aktiv",
      ylab="Danske bank afkast - risikofrit aktiv")
abline(lm(Yi ~ xi, data=DBdata))
```


Appendiks A

Resultater fra sandsynlighedsregning

I dette appendiks har vi samlet nogle resultater fra sandsynlighedsregning. Det meste står også i BH, men sommetider godt gemt og uden argumenter.

Den første sætning præciserer et resultat om tætheden for uafhængige kontinuerte stokastiske variable. Resultatet fremgår af teksten umiddelbart under definition 7.1.18 i BH, men er absolut vigtigt nok til at det fortjener sin egen sætning. I disse noter har vi brug for sætningen hver gang vi opskriver en likelihoodfunktion for uafhængige variable.

Sætning A.1. *Lad X og Y være to kontinuerte stokastiske variable med marginale tætheder f_X og f_Y og simultan tæthed f . Så er X og Y uafhængige hvis og kun hvis*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad x,y \in \mathbb{R} \quad (\text{A.1})$$

Bevis. Antag først at X og Y er uafhængige (BH definition 7.1.18), altså at

$$F(x,y) = F_X(x)F_Y(y), \quad x,y \in \mathbb{R}$$

hvor F , F_X og F_Y er notation for den simultane og de marginale fordelingsfunktioner. Så er den simultane tæthed i punktet (x,y) lig

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y) = \frac{\partial}{\partial x} \left(\frac{\partial}{\partial y} F_X(x)F_Y(y) \right) = \frac{\partial}{\partial x} (F_X(x)f_Y(y)) = f_X(x)f_Y(y)$$

hvor vi har benyttet at $F'_X = f_X$ og $F'_Y = f_Y$.

Antag omvendt at (A.1) gælder. Så er den simultane fordelingsfunktion i (x_0, y_0) givet ved

$$\begin{aligned}
 F(x_0, y_0) &= P(X \leq x_0, Y \leq y_0) \\
 &= \int_{-\infty}^{y_0} \int_{-\infty}^{x_0} f(x, y) \, dx \, dy \\
 &= \int_{-\infty}^{y_0} \int_{-\infty}^{x_0} f_X(x) f_Y(y) \, dx \, dy \\
 &= \int_{-\infty}^{y_0} f_Y(y) \left(\int_{-\infty}^{x_0} f_X(x) \, dx \right) \, dy \\
 &= F_X(x_0) \int_{-\infty}^{y_0} f_Y(y) \, dy \\
 &= F_X(x_0) F_Y(y_0)
 \end{aligned}$$

som ønsket. □

Bemærk at Sætning 7.1.20 i BH siger at X og Y er uafhængige hvis den simultane tæthed $f(x, y)$ splitter op i et produkt $g(x)h(y)$. Funktionerne g og h er ikke nødvendigvis de marginale tætheder, men beviset i BH viser at de er proportionale med de marginale tætheder. Sætning A.1 og sætning 7.1.20 generaliserer på den naturlige måde til n stokastiske variable: Kontinuerte stokastiske variable X_1, \dots, X_n er uafhængige hvis og kun hvis den simultane tæthed f kan skrives som et produkt af funktioner der kun afhænger af en variabel,

$$f(x_1, \dots, x_n) = g_1(x_1) \cdots g_n(x_n).$$

Den næste sætning handler om *separate transformationer* og siger at hvis vi transformerer to uafhængige kontinuerte variable hver for sig, så får vi igen uafhængige stokastiske variable. Resultatet er nævnt (omend ikke skrevet præcist) for diskrete variable på side 118 i BH, men det gælder altså også for kontinuerte variable.

Sætning A.2. *Lad X_1 og X_2 være reelle, kontinuerte stokastiske variable med marginale tætheder f_1 og f_2 . Lad desuden $g_1, g_2 : \mathbb{R} \rightarrow \mathbb{R}$ være givne funktioner, og definer to nye stokastiske variable som*

$$Z_1 = g_1(X_1), \quad Z_2 = g_2(X_2).$$

Hvis X_1 og X_2 er uafhængige, så er Z_1 og Z_2 uafhængige.

Bevis. Vi antager at X_1 og X_2 er uafhængige. Så er den simultane tæthed for (X_1, X_2) er givet ved

$$f(x_1, x_2) = f_1(x_1)f_2(x_2).$$

Vi skal vise at Z_1 og Z_2 er uafhængige, dvs.

$$P(Z_1 \leq z_1, Z_2 \leq z_2) = P(Z_1 \leq z_1)P(Z_2 \leq z_2)$$

for alle $z_1, z_2 \in \mathbb{R}$. Sandsynligheden på venstre side fås ved at integrere den simultane tæthed f over mængden

$$A = \{(x_1, x_2) \in \mathbb{R}^2 \mid g_1(x_1) \leq z_1 \text{ og } g_2(x_2) \leq z_2\}$$

Denne mængde kan skrives som $A = A_1 \times A_2$ hvor

$$A_1 = \{x_1 \in \mathbb{R} \mid g_1(x_1) \leq z_1\}, \quad A_2 = \{x_2 \in \mathbb{R} \mid g_2(x_2) \leq z_2\}.$$

Vi får derfor

$$\begin{aligned} P(Z_1 \leq z_1, Z_2 \leq z_2) &= \int_A f(x_1, x_2) d(x_1, x_2) \\ &= \int_{A_1 \times A_2} f_1(x_1)f_2(x_2) d(x_1, x_2) \\ &= \left(\int_{A_1} f_1(x_1) dx_1 \right) \left(\int_{A_2} f_2(x_2) dx_2 \right) \\ &= P(Z_1 \leq z_1)P(Z_2 \leq z_2) \end{aligned}$$

som ønsket. □

Sætningen kan udvides til n uafhængige stokastiske variable på to måder, som den næste sætning siger.

Sætning A.3. *Lad X_1, \dots, X_n være uafhængige reelle, kontinuerte stokastiske variable. Så gælder følgende.*

1. *Lad $g: \mathbb{R}^k \rightarrow \mathbb{R}$ og $h: \mathbb{R}^{n-k} \rightarrow \mathbb{R}$ være givne funktioner for et $k \in \{1, \dots, n-1\}$, og definer to nye stokastiske variable som*

$$Z_1 = g(X_1, \dots, X_k), \quad Z_2 = h(X_{k+1}, \dots, X_n).$$

Så er Z_1 og Z_2 uafhængige.

2. Lad $g_1, \dots, g_n : \mathbb{R} \rightarrow \mathbb{R}$ være givne funktioner. De transformerede stokastiske variable $g_1(X_1), \dots, g_n(X_n)$ er uafhængige.

Bevis. Vi nøjes med at skitsere beviserne. For at bevise den første påstand, skal man benytte at den simultane tæthed for (X_1, \dots, X_n) kan skrives som produkt af de marginale tætheder og at sandsynligheden $P(Z_1 \leq z_1, Z_2 \leq z_2)$ kan skrives som et integral over mængden $A_1 \times A_2$ hvor

$$A_1 = \{(x_1, \dots, x_k) \in \mathbb{R}^k \mid g(x_1, \dots, x_k) \leq z_1\}$$

$$A_2 = \{(x_{k+1}, \dots, x_n) \in \mathbb{R}^{n-k} \mid h(x_{k+1}, \dots, x_n) \leq z_2\}.$$

Den anden påstand vises ved induktion. For $n = 2$ er udsagnet identisk med sætning A.2, og i induktionstrinnet benyttes første del af sætningen samt induktionsantagelsen. \square

Vi får brug for *t-fordelingen* når vi konstruerer konfidensintervaller og udfører hypotesetest. Den følgende definition af *t-fordelingen* er identisk med Definition 10.4.4 i BH.

Definition A.4. Lad Z og V være uafhængige stokastiske variable hvor $Z \sim N(0, 1)$ og $V \sim \chi_k^2$, og definér

$$T = \frac{Z}{\sqrt{V/k}}.$$

Fordelingen af T kaldes *t-fordelingen med k frihedsgrader*, og vi skriver $T \sim t_k$.

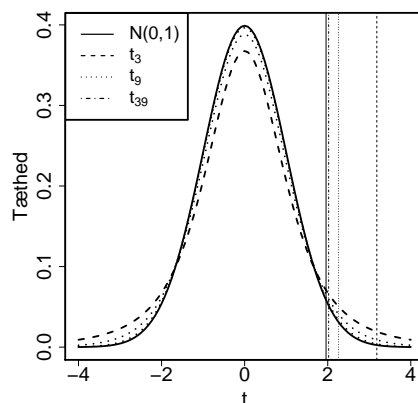
Figur A.1 viser tæthederne for tre forskellige *t-fordelinger* sammen med tætheden for standardnormalfordelingen. Det fremgår at *t-fordelingen* har tungere haler end $N(0, 1)$, men også at *t-fordelingen* ligner $N(0, 1)$ når antallet af frihedsgrader er stort. De lodrette linjer viser 97.5% fraktilerne, som vi bruger når vi beregner 95% konfidensintervaller. Når antallet af frihedsgrader vokser, bliver fraktilen mindre og nærmer sig 1.96 som er 97.5% fraktilen for $N(0, 1)$.

Sætningen nedenfor samler og præciserer en række resultater fra BH om uafhængige og identisk normalfordelte variable.

Sætning A.5. Lad Y_1, \dots, Y_n være uafhængige reelle stokastiske variable, og antag at hvert Y_i er $N(\mu, \sigma^2)$ -fordelt. Definer

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad SSD = \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad T = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{SSD/(n-1)}}$$

Da gælder:



Figur A.1: Tætheder for t -fordelingen med 3, 9 og 39 frihedsgrader sammen med tætheden for standardnormalfordelingen. De lodrette linjer angiver 97.5% fraktiler i de fire fordelinger: 1.960 for $N(0, 1)$, 3.182 for t_3 , 2.262 for t_9 og 2.023 for t_{39} .

1. \bar{Y} er $N\left(\mu, \frac{\sigma^2}{n}\right)$ -fordelt.
2. SSD/σ^2 er χ^2 -fordelt med $n - 1$ frihedsgrader
3. \bar{Y} og SSD er uafhængige
4. T er t -fordelt med $n - 1$ frihedsgrader

Bevis. Det følger af foldningsegenskaben for normalfordelinger (BH eksempel 6.6.3) at

$$\sum_{j=1}^n Y_j \sim N(n\mu, n\sigma^2).$$

Påstand 1 fremkommer nu af de sædvanlige regneregler for affine transformationer af normalfordelinger, se BH definition 5.4.3.

Påstand 2 er indholdt af BH eksempel 10.4.3.

Påstand 3 er indholdt af BH eksempel 7.5.9 (som vi ikke beviser på dette kursus).

Påstand 4 følger af påstand 1–3 samt definitionen på t -fordelingen (BH definition 10.4.4 eller definition A.4 ovenfor). \square

Appendiks B

Profilmaksimering

Det kan være overraskende vanskeligt at maksimere en funktion af to variable, lad os sige $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. På indledende matematikkurser lærer man ofte en teknik, hvor man finder de stationære punkter, altså punkter hvor de partielle afledede er nul, og undersøger opførslen af f i disse punkter samt “ude ved randen” af definitionsområder. Nogle gange virker teknikken glimrende, men langt fra altid: Det kan være vanskeligt at bestemme de stationære punkter, og det kan være svært at få kontrol over hvad der sker ved randen.

Nedenfor præsenteres en anden teknik, kaldet *profilering* eller *profilmaksimering*. Metoden består i at erstatte det oprindelige problem hvor der skal maksimeres over både x og y med to endimensionale maksimeringsproblemer. Profilering er ikke et vidundermiddel og kan kun gennemføres i specielle situationer, men teknikken viser sig at være nyttigt i en række vigtige maksimeringsproblemer i statistik.

Antag at vi for en fastholdt værdi af x kan finde en y -værdi — som vi kalder $\hat{y}(x)$ for at markere at den givetvis afhænger af værdien af x — der maksimerer den endimensionale funktion $y \mapsto f(x, y)$. Vi antager altså at $\hat{y}(x)$ opfylder at

$$f(x, y) \leq f(x, \hat{y}(x)) \quad \text{for alle } y \in \mathbb{R}.$$

Vi forestiller os at denne endimensionale maksimering kan gennemføres for alle værdier af x , fx med almindelig funktionsundersøgelse for funktioner af en variabel. I så fald er det naturligt at danne den såkaldte *profilfunktion*,

$$\tilde{f}(x) = \max_{y \in \mathbb{R}} f(x, y) = f(x, \hat{y}(x)), \quad x \in \mathbb{R}.$$

Det hedder en profilmfunktion fordi grafen for \tilde{f} svarer til den profil man ser af 2D-

grafen for f hvis man stiller sig ved x og kun kigger i y -retningen.

Profilmfunktionen $\tilde{f} : \mathbb{R} \mapsto \mathbb{R}$ er kun en funktion af x og kan derfor maksimeres med metoder til maksimering af sådanne funktioner, fx almindelig funktionsundersøgelse. Antag at vi kan maksimere den, altså at vi kan finde et punkt \hat{x} så

$$\tilde{f}(x) \leq \tilde{f}(\hat{x}) \quad \text{for alle } x \in \mathbb{R}.$$

I så fald vil der gælde at $(\hat{x}, \hat{y}(\hat{x}))$ maksimerer f . For alle $(x, y) \in \mathbb{R}^2$ gælder nemlig at

$$f(x, y) \leq f(x, \hat{y}(x)) = \tilde{f}(x) \leq \tilde{f}(\hat{x}) = f(\hat{x}, \hat{y}(\hat{x})).$$

Hvis alle disse endimensionale maksimeringer er lykkedes for os, så har vi faktisk løst det todimensionale maksimeringsproblem, helt uden bekymre os om de komplikationer som metoden med stationære punkter kan give i to dimensioner. Det kræver dog at alle endimensionale maksimeringsproblemer (mht. y for fast x og mht. x) kan løses.

Når man bruger metoden ovenfor, siger man at man “profilere y ud”, men man kan naturligvis profilere x ud i stedet. Man vælger at profilere den variabel ud, der gør beregningerne nemmest.

Ovenfor antog vi at f skulle maksimeres over hele \mathbb{R}^2 . Hvis man i stedet skal maksimere f over $A \times B \subseteq \mathbb{R}^2$, så skal maksimeringen for fastholdt $x \in A$ være over $y \in B$, og man betragter kun den resulterende profilmfunktion $\tilde{f}(x)$ for $x \in A$.

Profilmetoden kan også bruges hvis man har mere end to variable. Hvis man fx har en funktion $f(x, y, z)$ af tre variable, så kan man for fastholdt værdi af x og y profilere z ud,

$$\tilde{f}(x, y) = \max_{z \in \mathbb{R}} f(x, y, z)$$

Den resulterende profilmfunktion $\tilde{f}(x, y)$ kan så angribes med enhver relevant maksimeringsmetode, fx yderligere profilering. Bortset fra at notationen hurtigt bliver uigennemskuelig — og bortset fra at regningerne kun lader sig gennemføre hvis man er heldig — så er successiv profilering en uhyre frugtbar ide.

Litteratur

- Altman, D. G. (1999). *Practical Statistics for Medical Research*. Chapman & Hall, London.
- Blæsild, P. and Granfeldt, J. (2003). *Statistics with Applications in Biology and Geology*. Chapman & Hall, London.
- Blitzstein, J. and Hwang, J. (2015). *Introduction to Probability*. CRC Press.
- Ekstrøm, C. T. and Sørensen, H. (2010). *Introduction to Statistical Data Analysis for the Life Sciences*. Chapman & Hall/CRC.
- Henningsen, I. (2006a). *En Introduktion til Statistik, bind 1*. Afdeling for Anvendt Matematik og Statistik, Københavns Universitet, 7. udgave.
- Henningsen, I. (2006b). *En Introduktion til Statistik, bind 2*. Afdeling for Anvendt Matematik og Statistik, Københavns Universitet, 4. udgave.
- Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., and Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, **317**(5384), 82.
- Ortego, J. D., Aminabhavi, T. M., Harlapur, S. F., and Balundgi, R. H. (1995). A review of polymeric geosynthetics used in hazardous waste facilities. *Journal of Hazardous Materials*, **42**, 115–156.
- Pinheiro, J. and Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.
- Samuels, M. L. and Witmer, J. A. (2003). *Statistics for the Life Sciences*. Pearson Education, Inc., New Jersey.
- Skovgaard, I. (2004). *Basal Biostatistik, Del 2*. Samfundslitteratur.

- Skovgaard, I., Stryhn, H., and Rudemo, M. (1999). *Basal Biostatistik, Del 1*. DSR Forlag.
- Sørensen, M. (2011). *En Introduktion til Sandsynlighedsregning*. Institut for Matematiske Fag, Københavns Universitet, 12. udgave.
- Venables, W. and Ripley, B. (1999). *Modern Applied Statistics with S-PLUS*. Springer, New York.

Indeks

- accept af hypotese, 46
- afvisning af hypotese, 46

- baggrundsvariabel, 111
- binomialfordelingen, 9, 10

- Capital Asset Pricing model, 134, 151
- CAPM, 134
- Carlsberg aktien, 136
- central estimator, 15, 35, 59, 85, 119

- dagligvarepriser, 66
- Danske Bank aktien, 151

- effektparameter, 112
- eksempel
 - CAPM, 134
 - dagligvarepriser, 66
 - energiforbrug, 98
 - kobbertråd, 31, 35, 38, 40, 45, 47
 - længde af kronblade, 95
 - læsetest, 47
 - malaria, 69
 - Mendelsk spaltning, 17
 - produktivitetsscore, 95
 - prothrombinindeks, 56, 60, 62, 66, 69
 - smagsforsøg, 16
 - tuberkulose, 82, 86, 89, 94, 96
 - vægt af hjerner, 69
 - vcf og blodglukose, 113, 120, 122, 126, 128, 130
 - ventetid, 20
- empirisk varians, 59
- en stikprøve, 27
 - med kendt varians, 31
 - med ukendt varians, 55
- endeligt udfaldsrum, 18
- energiforbrug, 98, 108
- estimat, 12, 14
- estimator, 9, 14
- estimeret spredning
 - for $\hat{\mu}_1$ og $\hat{\mu}_2$, 86
 - for \hat{p} , 16
 - for $\hat{\alpha}$ og $\hat{\beta}$, 120
 - for $\hat{\mu}$, ukendt varians, 59
 - for $\widehat{\mu_1 - \mu_2}$, 88
- F*-test, 95
- forklarende variabel, 111, 112

- histogram, 68, 96
- hypotese, 41
 - om α , 138
 - om β , 122
 - om μ , kendt varians, 40
 - om μ , ukendt varians, 62
 - om $\mu_1 - \mu_2$, 89
- hypotesetest, *Se* test af hypotese
- hyppighed, 9

- kobbertråd, 31, 35, 38, 40, 45, 47
- konfidensinterval, 46, 49, 120
 - for α og β , 121

- for μ , kendt varians, 37
- for μ , ukendt varians, 61
- for $\mu_1 - \mu_2$, 88
- for μ_1 og μ_2 , 87
- for regressionslinje, 127
- konklusion på test, 46
- kovariat, 111
- kritiske værdier, 41, 63
- kvotienttest, 41, 49, 123
- kvotientteststørrelse, 41, 63, 90
 - for $\alpha = 0$, 150
 - for $\beta = 0$, 123
 - for $\mu = \mu_0$, kendt varians, 43
 - for $\mu = \mu_0$, ukendt varians, 63
 - for $\mu_1 = \mu_2$, 90
- least squares method, 36
- likelihood ratio test, 41
- likelihoodfunktion, 9
 - binomialfordelingen, 12
 - endeligt udfaldsrum, 19
 - kendt varians, 33
 - lineær regression, 114
 - to stikprøver, 83
 - ukendt varians, 56
- lineær regression, 28, 111
- log-likelihoodfunktion, 14, 34
- længde af kronblade, 95, 108
- læsetest, 47
- maksimaliseringsestimat, 12
- maksimum likelihood estimat, 18, 21, 48
 - endeligt udfaldsrum, 19
 - for $(\alpha, \beta, \sigma^2)$, 115
 - for (μ, σ^2) , 57
 - for (μ_1, μ_2, σ^2) , 83
 - for $(\mu_1 = \mu_2, \sigma^2)$, 90
 - for β når $\alpha = 0$, 140
 - for μ , kendt varians, 34
 - for p , 12, 21
- malaria, 69
- markedsafkast, 135
- Mendelsk spaltning, 17
- mindste kvadraters metode, 36
- MLE, 12
- modelkontrol, 28, 49, 67, 96, 130, 133
- momentestimation, 21
- normalfordelingen, 27
- normalfordelingsantagelse
 - kontrol af, 67, 96, 134
- opdatering af estimater, 66, 94
- p -værdi, 42, 44, 46, 63, 90, 91, 123
- parameter, 10, 32, 55
- parametermængde, 10, 11, 19, 32, 55, 82
- parrede data, 71
- parret t -test, 66, 82
- prædiktion, 126, 129
- prædiktionsinterval, 129
- produktivitetsscore, 95, 108
- profilering, 58, 84, 116, 158
- profilfunktion, 158
- profillikelihoodfunktion, 58, 84
- prothrombinindeks, 56, 60, 62, 66, 69
- QQ-plot, 68, 96, 134
- R funktioner
 - abline, 145
 - confint, 145
 - dbinom, 21
 - dnorm, 51
 - fitted, 146
 - hist, 75, 103
 - lm, 143
 - mean, 49, 103
 - pbinom, 21

- plot, 145
- pnorm, 49, 51
- predict, 146
- pt, 75
- qnorm, 49, 51
- qqnorm, 76, 103
- qt, 75, 103
- rbinom, 22
- residuals, 146
- rnorm, 51
- rstandard, 146
- sd, 74
- summary, 144
- t.test, 74, 102
- var, 74, 103
- regressionslinje, 112, 126
- regressionsvariabel, 111
- relativ hyppighed, 9, 14
- residual, 130, 131
- responsvariabel, 111, 112
- risikofri aktiv, 135
- scatterplot, 113
- separate transformationer, 154
- signifikansniveau, 44
- simulation, 38, 70
- smagsforsøg, 16
- spredning, estimeret, *Se* estimeret spredning
- standard error, *Se* estimeret spredning
- standardiseret residual, 132
- statistisk analyse, 9
- statistisk model, 9–11, 48
 - binomialfordelingen, 10, 21
 - kendt varians, 32
 - lineær regression, 112, 113
 - to stikprøver, 82
 - ukendt varians, 56
- t*-fordelingen, 156
- t*-test, 65, 94, 125
- test af hypotese, 41, 46, 49, 122
 - om β , 123
 - om μ , kendt varians, 43
 - om μ , ukendt varians, 63
 - om $\mu_1 = \mu_2$, 90
- testsandsynlighed, 42, 63, 90
- to stikprøver, 28, 81
- transformation af data, 70, 96
- trunkerede data, 96
- tuberkulose, 82, 86, 89, 94, 96, 101
- type I fejl, 46
- type II fejl, 46
- u*-test, 42
- uafhængighed, 11, 27, 28, 95
- usikkerhed, 9–11, 33, 35
- varianshomogenitet, 27, 28, 81, 95
- vcf og blodglukose, 113, 120, 122, 126, 128, 130
- ventetid, 20
- vægt af hjerner, 69