STATISTIKNOTER

Simple normalfordelingsmodeller

Jørgen Larsen

IMFUFA
Roskilde Universitetscenter

Februar 1999

IMFUFA, Roskilde Universitetscenter, Postboks 260, DK-4000 Roskilde.
Jørgen Larsen: **STATISTIKNOTER: Simple normalfordelingsmodeller**
IMFUFA tekst nr. 304b/1995 98 sider ISSN 0106-6242

Dette hæfte er en del af undervisningsmaterialet til et kursus i statistik og statistiske modeller. Undervisningsmaterialet omfatter blandt andet følgende titler:

a. Simple binomialfordelingsmodeller
b. Simple normalfordelingsmodeller
c. Simple Poissonfordelingsmodeller
d. Simple multinomialfordelingsmodeller
e. Mindre matematisk-statistisk opslagsværk, indeholdende bl.a. ordforklaringer, resuméer og tabeller

•

Om kurset og kursusmaterialet kan blandt andet siges at

- når det er et gennemgående tema at påpege at likelihoodmetoden kan benyttes som et overordnet princip for valg af estimatorer og teststørrelser, er det blandt andet begrundet i *at* likelihoodmetoden har mange egenskaber der fra et matematisk-statistisk synspunkt anses for ønskelige, *at* likelihoodmetoden er meget udbredt og nyder stor anerkendelse (ikke mindst i Danmark), og *at* det i al almindelighed er værd at gøre opmærksom på at man også inden for faget statistik har overordnede og strukturerende begreber og metoder;
- når kursusmaterialet er skrevet på dansk (og ikke for eksempel på 'scientific English'), er det for at bidrage til at vedligeholde traditionerne for *hvordan* og *at* man kan tale om slige emner på dansk, og så sandelig også fordi dansk er det sprog som forfatteren – og vel også den forventede læser – er bedst til;
- når hæfterne foruden de sædvanlige simple modeller, metoder og eksempler også indeholder eksempler der er væsentligt sværere, er det for at antyde nogle af de retninger man kan arbejde videre i, og for at der kan være lidt udfordringer til den krævende læser.

Indhold

1 Normalfordelingen

Man har meget ofte brug for en type sandsynlighedsfordelinger der kan beskrive hvordan målinger varierer tilfældigt omkring et bestemt niveau, når det skal være sådan at de faktisk observerede værdier lige så godt tilfældigvis kan være lidt *over* som lidt *under* det teoretisk rigtige niveau. For at kunne finde frem til sådanne fordelinger må vi præcisere lidt nøjere hvad det er der søges.

Fordelingerne skal benyttes til at beskrive den tilfældige variation af målinger af længder, masser, koncentrationer osv., altsammen størrelser der måles på en *kontinuert* skala. Første punkt i problempræciseringen er derfor:

> *Der søges en type kontinuerte fordelinger.*

Fordelingerne skal beskrive den tilfældige variation omkring et vist niveau. Dette niveau skal indgå som en parameter μ, så modelfunktionen skal derfor være en funktion af både en observationsvariabel x og en parametervariabel μ:

> *Modelfunktionen er* $f(x;\mu)$.

Parameteren μ skal beskrive *hvor* på tallinien fordelingen er beliggende, og en ændring af parameterværdien skal svare til en forskydning af sandsynlighedsfordelingen hen ad tallinien uden at fordelingens form i øvrigt ændres. Mere præcist vil vi antage at

> *Fordelingen svarende til parameterværdien* μ *fås ved at forskyde fordelingen svarende til parameterværdien* 0 *stykket* μ*, dvs.*
>
> $$f(x;\mu) \quad = \quad f(x-\mu;0)$$
>
> *hvor* μ *i princippet kan antage alle mulige værdier.*

Denne betingelse udtrykker man også på den måde at μ skal være en *positionsparameter*.

Disse tre betingelser er ikke nok til at fastlægge fordelingen, så man er nødt til at stille nogle flere krav. Vi vil stille en *statistisk betingelse*, en betingelse der handler om hvordan man skal analysere data fra den søgte fordeling: Da parameteren μ skal beskrive det niveau hvoromkring observationerne fordeler sig, kan man mene at det må være rimeligt at den ukendte parameter μ skal estimeres ved *gennemsnittet af observationerne*. Da det tillige er et gennemgående princip at man altid skal benytte *maksimaliseringsestimater*, vil vi stille følgende krav:

Maksimaliseringsestimatet for μ skal være gennemsnittet af observationerne.

I næste afsnit viser vi at disse betingelser fører frem til den såkaldte *normalfordeling* med middelværdiparameter μ og variansparameter σ^2, det vil sige fordelingen med tæthedsfunktion

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

1.1 Udledning af normalfordelingen

Vi vil i dette afsnit gøre rede for at normalfordelingen faktisk er svar*et* på ønsket om en type kontinuerte fordelinger på den reelle akse således at fordelingerne er parametriseret med en positionsparameter, og således at maksimaliseringsestimatet for positionsparameteren er gennemsnittet af observationerne. Det går for sig på denne måde:

1. Modelfunktionen hørende til et forsøg med én observation kaldes som nævnt $f(x;\mu)$. Modelfunktionen svarende til et forsøg med n observationer $x_1, x_2, \ldots, x_n$ er da $\prod_{i=1}^{n} f(x_i;\mu)$, så likelihoodfunktionen er

   $$L(\mu) \;=\; \prod_{i=1}^{n} f(x_i;\mu).$$

2. Da der skal være tale om en positionsparameter må der gælde at

   $$f(x;\mu) \;=\; f(x-\mu;0) \;=\; f_0(x-\mu),$$

   hvor f_0 er brugt som en kort betegnelse for $f(\,\cdot\,;0)$. Likelihoodfunktionen kan derfor skrives som

   $$L(\mu) \;=\; \prod_{i=1}^{n} f_0(x_i-\mu),$$

   og log-likelihoodfunktionen er tilsvarende

   $$\ln L(\mu) \;=\; \sum_{i=1}^{n} \ln f_0(x_i-\mu).$$

3. Vi har stillet som krav at $\ln L$ skal antage sin maksimale værdi i punktet $\mu=\overline{x}$. Hvis vi desuden går ud fra at f_0 og dermed også $\ln L$ er en pæn differentiabel funktion, så er den afledede $(\ln L)'$ lig 0 i dette maksimumspunkt:

   $$(\ln L)'(\overline{x}) \;=\; 0.$$

4. Af udtrykket for $\ln L$ fås

$$\begin{aligned}(\ln L)'(\mu) &= \sum_{i=1}^{n} -(\ln f_0)'(x_i - \mu) \\ &= \sum_{i=1}^{n} g(x_i - \mu),\end{aligned}$$

hvor g er en kort betegnelse for $-(\ln f_0)'$. Kravet om at maksimaliseringsestimatet skal være lig gennemsnittet $\overline{x}$, betyder derfor at funktionen g skal opfylde betingelsen

$$\sum_{i=1}^{n} g(x_i - \overline{x}) = 0. \tag{1.1}$$

5. Fidusen er nu at formel (1.1) skal gælde for *alle* valg af $x_1, x_2, \ldots, x_n$, og ved at indsætte nogle tilpas snedigt valgte x-er kan man få at vide hvordan funktionen g nødvendigvis må se ud.

   (a) Ved at vælge $n = 2$ og $x_2 = -x_1 = y$ (hvorved $\overline{x} = 0$) fås af formel (1.1) at $g(-y) + g(y) = 0$, dvs.

   $$g(-y) = -g(y) \tag{1.2}$$

   for vilkårligt y. Specielt er $g(0) = 0$.

   (b) Ved at vælge $n = k + 1$ og lade de k første x-er være ens og lade gennemsnittet være 0, mere præcist ved at vælge $x_1 = x_2 = \ldots = x_k = -y$ og $x_{k+1} = ky$, fås at $k\,g(-y) + g(k\,y) = 0$, der ved brug af formel (1.2) kan formuleres som

   $$g(k \cdot y) = k \cdot g(y) \tag{1.3}$$

   gældende for vilkårligt y og $k = 1, 2, 3, \ldots$. Ved at bruge formel (1.2) endnu en gang kan man nu slutte at formel (1.3) gælder for vilkårlige reelle tal y og for vilkårlige hele tal k.

   (c) I formel (1.3) kan vi vælge $y = {}^{j}/_{k}$ hvor j og k er heltal. Derved fås at $g(j) = k\,g({}^{j}/_{k})$, dvs. at $g({}^{j}/_{k}) = {}^{1}/_{k}\,g(j)$.
   Men vi kan også vælge $y = 1$ og $k = j$ i formel (1.3), og derved får vi $g(j) = j\,g(1)$. Alt i alt er dermed $g({}^{j}/_{k}) = {}^{j}/_{k}\,g(1)$, hvilket vi formulerer sådan:

   $$\begin{aligned}g(y) &= y \cdot g(1) \\ &= g(1) \cdot y\end{aligned} \tag{1.4}$$

   for alle rationale tal y.

Medmindre g skal være en ganske overordentlig usædvanlig funktion, er det sådan at når formel (1.4) gælder for alle *rationale* tal y, så gælder den også for alle *reelle* tal y. Vi vil gå ud fra at formel (1.4) gælder for alle

y, og vi er altså så nået frem til at funktionen g er en almindelig lineær funktion:

$$g(x) \quad = \quad c\,x$$

for en passende valgt konstant c.

6. Da g blot var en kort betegnelse for funktionen $-(\ln f_0)'$, kan vi dernæst finde f_0: Hvis $-(\ln f_0)'(x) = c\,x$, så er

$$\ln f_0(x) \quad = \quad -\tfrac{1}{2}c\,x^2 + \text{konstant},$$

dvs.

$$f_0(x) \quad = \quad \text{konstant} \cdot \exp\left(-\tfrac{1}{2}c\,x^2\right).$$

7. Denne funktion f_0 skal være en sandsynlighedstæthed hvilket vil sige at den skal være ikke-negativ og integrere til 1, altså $\int_{-\infty}^{+\infty} f_0(x)dx = 1$. For at dette sidste skal kunne lade sig gøre må konstanten c nødvendigvis være positiv; traditionen tro omdøber vi c til $1/\sigma^2$ hvorved tæthedsfunktionen får udseendet

$$f_0(x) \quad = \quad \text{konstant} \cdot \exp\left(-\tfrac{1}{2}\frac{x^2}{\sigma^2}\right).$$

Den betingelse at f_0 skal integrere til 1 fastlægger konstanten; man kan vise at den skal være $1/\sqrt{2\pi\sigma^2}$. Dermed har vi fundet at

$$f_0(x) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{x^2}{\sigma^2}\right)$$

og dermed

$$\begin{aligned} f(x;\mu) \quad &= \quad f_0(x-\mu) \\ &= \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right). \end{aligned}$$

8. Det oprindelige problem bestod i at finde en type fordelinger hvor der indgik en *positionsparameter* μ. I den fundne løsning optræder imidlertid også en størrelse σ^2 der er kommet ind i billedet som en integrationskonstant. Denne størrelse udnævner vi til en *parameter*, og samtidig omdøbes $f(x;\mu)$ til $f(x;\mu,\sigma^2)$:

$$f(x;\mu,\sigma^2) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right).$$

Der gælder at for ethvert valg af $\mu \in \mathbf{R}$ og $\sigma^2 > 0$ er dette en sandsynlighedstæthedsfunktion, nemlig for *normalfordelingen* med *positionsparameter* (eller *middelværdiparameter*) μ og *kvadratisk skalaparameter* (eller *variansparameter*) σ^2, kort $\mathcal{N}(\mu,\sigma^2)$.

Resultatet af ovenstående udledninger er således at *hvis* vi er på jagt efter en type kontinuerte sandsynlighedsfordelinger hvor der optræder en positionsparameter, og *hvis* vi forlanger at denne positionsparameter skal estimeres ved gennemsnittet af observationerne, *så* er normalfordelinger den eneste type fordelinger der kan komme på tale. (Strengt taget har vi ikke vist at normalfordelingerne faktisk har den ønskede egenskab, men det kommer i det følgende.)

Normalfordelinger kaldes også Gauß-fordelinger. K.F.Gauß (1777-1855) benyttede normalfordelinger til at beskrive bl.a. astronomiske målingers tilfældige fra den sande værdi. I værket *Theoria Motus Corporum Coelestium in Sectionibus Conicus Arbientium* (dvs. Teori om de himmelske legemers bevægelser i keglesnit omkring solen) argumenterede han for normalfordelingen på en måde der meget ligner den der er benyttet her.

1.2 Egenskaber ved normalfordelingen

Her gives en oversigt (uden beviser) over forskellige egenskaber ved normalfordelingen:

- Normalfordelingen med parametre μ og σ^2, kort $\mathcal{N}(\mu, \sigma^2)$-fordelingen, er den sandsynlighedsfordeling på den reelle talakse $\mathbf{R}$ som har tæthedsfunktionen

$$f(x; \mu, \sigma^2) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right).$$

  Her kan parameteren μ være et vilkårligt reelt tal og parameteren σ^2 et vilkårligt positivt tal.

- Parameteren μ er en *positionsparameter*, dvs. hvis X er $\mathcal{N}(\mu, \sigma^2)$-fordelt og a en konstant, så vil $a + X$ være $\mathcal{N}(a + \mu, \sigma^2)$-fordelt.

  Desuden er μ *middelværdien* i $\mathcal{N}(\mu, \sigma^2)$-fordelingen.

  Endvidere er μ *medianen* i $\mathcal{N}(\mu, \sigma^2)$-fordelingen.

- Parameteren σ^2 er en *kvadratisk skalaparameter*, hvilket vil sige at hvis X er $\mathcal{N}(0, \sigma^2)$-fordelt og b en konstant, så vil bX være $\mathcal{N}(0, b^2\sigma^2)$-fordelt.

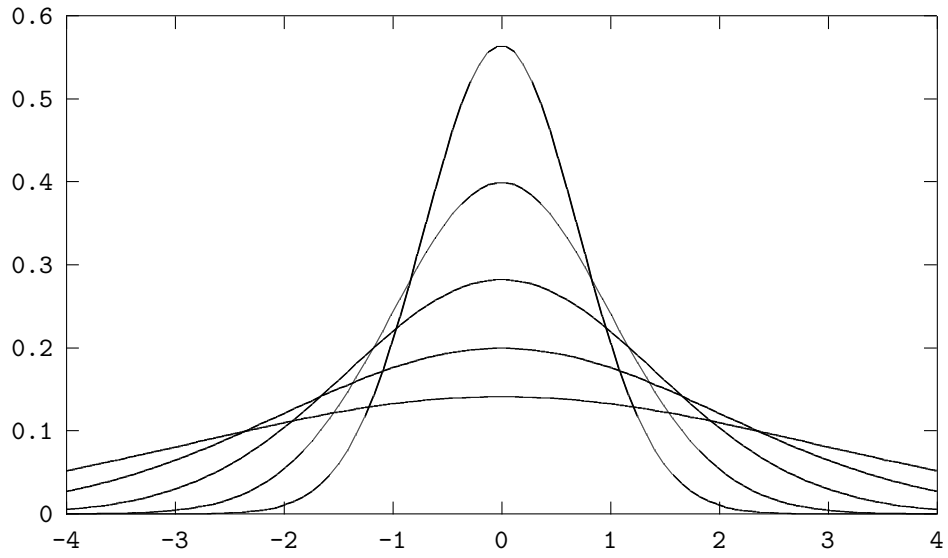
  Desuden er σ^2 *variansen* i $\mathcal{N}(\mu, \sigma^2)$-fordelingen, og dermed er σ *standardafvigelsen* i $\mathcal{N}(\mu, \sigma^2)$-fordelingen.

  Undertiden kaldes $1/\sigma^2$ for *præcisionen* i fordelingen, fordi $1/\sigma^2$ er et udtryk for hvor snævert fordelingen er koncentreret om sin middelværdi.

- Hvis X er $\mathcal{N}(\mu, \sigma^2)$-fordelt, så vil $a + bX$ være $\mathcal{N}(a + b\mu, b^2\sigma^2)$-fordelt; her betegner a og b konstanter.

- Den *normerede normale fordeling* er $\mathcal{N}(0, 1)$-fordelingen. Dens tæthed betegnes ofte φ:

$$\varphi(x) \quad = \quad \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}x^2), \quad x \in \mathbf{R}.$$



Figur 1.1 Tæthedsfunktioner for normalfordelinger med middelværdi 0 og varians hhv. 0.5, 1, 2, 4 og 8.

Dens kumulerede fordelingsfunktion betegnes tilsvarende Φ, dvs. $\Phi(u)$ er sandsynligheden for at en $\mathcal{N}(0,1)$-variabel er mindre end eller lig u:

$$\begin{aligned}\Phi(u) &= \int_{-\infty}^{u} \varphi(x)\,dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp(-\tfrac{1}{2}x^2)\,dx.\end{aligned}$$

- En $\mathcal{N}(\mu,\sigma^2)$-variabel har tæthedsfunktion

$$x \;\mapsto\; \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$$

og kumuleret fordelingsfunktion

$$x \;\mapsto\; \Phi\left(\frac{x-\mu}{\sigma}\right).$$

- Hvis α er et tal mellem 0 og 1 så har ligningen $\Phi(u) = \alpha$ netop én løsning, nemlig *α-fraktilen* u_α i den normerede normale fordeling.

  Ved at lægge fem til fraktilerne fås de såkaldte *probits* (dvs. probability units):

$$\text{probit}(\alpha) \;=\; u_\alpha + 5.$$

  I statistiske tabelværker findes tabeller over $\Phi(u)$ og over fraktilerne u_α eller $u_\alpha + 5$.

1.3 Opgaver

Opgave 1.1
Diskutér om det vil være rimeligt at benytte normalfordelingsmodeller (med uafhængige observationer) i de situationer der kort antydes her:

1. Bredden af kraniet på 20 toårige grønlandske sneharer fanget ved Søndre Strømfjord en bestemt sommer.
2. Vindstyrken kl. 12 på en bestemt lokalitet på 50 på hinanden følgende dage.
3. Vægten af 100 tilfældigt udvalgte sild landet i Gilleleje en bestemt dag.
4. Koncentrationen af NO_x kl. 16.30 ved Nørreport Station hver dag i november måned.
5. Høstudbyttet på hver af 10 forsøgsparceller (à 500 m^2) med en ny sort vinterbyg.
6. Vægten af leveren i 27 fem uger gamle forsøgsmus.
7. Antal nyregistrerede AIDS-tilfælde i Danmark i hver af 12 på hinanden følgende måneder.
8. Antal nyregistrerede leukæmi-tilfælde i Danmark i hver af 12 på hinanden følgende måneder.
9. Levetiden af 50 elektriske 40W pærer af samme fabrikat.
10. Det årlige antal trafikulykker i København og Frederiksberg kommuner hvor cyklister er indblandet, for hvert af årene 1980-1990.

Opgave 1.2
Løs ved hjælp af passende tabeller følgende delopgaver:

1. Find 25%-fraktilen i den normerede normalfordeling $\mathcal{N}(0,1)$.
2. Find 75%-fraktilen i den normerede normalfordeling $\mathcal{N}(0,1)$.
3. Find et interval af formen $[-x,x]$ som indeholder 50% af sandsynlighedsmassen i den normerede normalfordeling $\mathcal{N}(0,1)$.
4. Find et interval af formen $[-x,x]$ som indeholder 95% af sandsynlighedsmassen i den normerede normalfordeling $\mathcal{N}(0,1)$.
5. Hvor stor en del af sandsynlighedsmassen i den normerede normalfordeling $\mathcal{N}(0,1)$ er indeholdt i intervallet $[-1,1]$?

Opgave 1.3
Løs ved hjælp af passende tabeller følgende delopgaver:

1. Udtryk 25%-fraktilen i normalfordelingen $\mathcal{N}(\mu, \sigma^2)$ ved μ og σ^2.
2. Udtryk 75%-fraktilen i normalfordelingen $\mathcal{N}(\mu, \sigma^2)$ ved μ og σ^2.
3. Angiv et interval af formen $[\mu - x, \mu + x]$ som indeholder 50% af sandsynlighedsmassen i normalfordelingen $\mathcal{N}(\mu, \sigma^2)$.
4. Angiv et interval af formen $[\mu - x, \mu + x]$ som indeholder 95% af sandsynlighedsmassen i normalfordelingen $\mathcal{N}(\mu, \sigma^2)$.
5. Hvor stor en del af sandsynlighedsmassen i normalfordelingen $\mathcal{N}(\mu, \sigma^2)$ er indeholdt i intervallet $[\mu - \sigma, \mu + \sigma]$?

TIP: Udnyt eventuelt Opgave 1.2

Opgave 1.4
Generelt er en α-fraktil i en fordeling et tal x_α med den egenskab at brøkdelen α af fordelingen ligger til venstre for x_α.

Find α-fraktilen x_α i $\mathcal{N}(\mu, \sigma^2)$-fordelingen udtrykt ved μ, σ^2 og ved α-fraktilen u_α i den normerede normalfordeling.
TIP: Værdien af den kumulerede fordelingsfunktion (for $\mathcal{N}(\mu, \sigma^2)$) udregnet i x_α skal være lig α. Den kumulerede fordelingsfunktion kan udtrykkes ved Φ.

2 Enstikprøveproblemet i normalfordelingen

Normalfordelingen blev i Kapitel 1 udledt i forbindelse med søgningen efter en fordeling hvor positionsparameteren estimeres ved gennemsnittet af observationerne. Vi mangler imidlertid at gøre rede for at normalfordelingen faktisk *har* denne eftertragtede egenskab, men det vil ske i indeværende Kapitel som led i behandlingen af »enstikprøveproblemet i normalfordelingen«.

Enstikprøveproblemet i normalfordelingen handler om en enkelt *stikprøve*, altså et antal uafhængige observationer, $y_1, y_2, \ldots, y_n$, fra en $\mathcal{N}(\mu, \sigma^2)$-fordeling. Parametrene μ og σ^2 er ukendte, og problemet er at bestemme estimater over dem og måske teste hypoteser om dem. En anden side af sagen er *modelkontrolproblemet*, dvs. spørgsmålet om hvordan man vurderer om observationerne nu også med rimelighed kan beskrives som værende normalfordelte.

Eksempel 2.1 (Lysets hastighed)
I årene 1879-82 foretog den amerikanske fysiker A.A.Michelson og den amerikanske matematiker og astronom S.Newcomb en række efter den tids forhold temmelig nøjagtige bestemmelser af lysets hastighed i luft. Deres metoder var baseret på Foucaults idé med at sende en lysstråle fra et hurtigt roterende spejl hen på et fjernt fast spejl som returnerer lysstrålen til det roterende hvor man måler dens vinkelforskydning i forhold til den oprindelige lysstråle. Hvis man kender rotationshastigheden samt afstanden mellem spejlene, kan man derved bestemme lyshastigheden.

I Tabel 2.1 er vist resultaterne af de 66 målinger som Newcomb foretog i perioden 24. juli til 5. september 1882 i Washington, D.C. I Newcombs opstilling var der 3721 m

Tabel 2.1 Newcombs bestemmelser af lysets passagetid af en strækning på 7442 m. Tabelværdierne $\times 10^{-3} + 24.8$ er passagetiden i 10^{-6} sek.

28	26	33	24	34	–44
27	16	40	–2	29	22
24	21	25	30	23	29
31	19	24	20	36	32
36	28	25	21	28	29
37	25	28	26	30	32
36	26	30	22	36	23
27	27	28	27	31	27
26	33	26	32	32	24
39	28	24	25	32	25
29	27	28	29	16	23

mellem det roterende spejl der var placeret i Fort Myer på vestbredden af Potomac-floden, og det faste spejl der var anbragt på George Washington-monumentets fundament. Den størrelse som Newcomb rapporterer, er lysets *passagetid*, altså den tid som det er om at tilbagelægge den pågældende distance.

Af de 66 værdier i Tabel 2.1 skiller to sig ud, nemlig -44 og -2, der synes at være »outliers«, altså tal der tilsyneladende ligger *for* langt væk fra flertallet af observationerne. Det er altid et vanskeligt spørgsmål at afgøre om det er forsvarligt at se bort fra »outliere«.

I analysen af tallene i Tabel 2.1 vil vi vælge at se bort fra de to nævnte observationer således at vi kun har at gøre med 64 observationer.

I den generelle situation foreligger der størrelser $y_1, y_2, \ldots, y_n$ der antages at være observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_n$ som er uafhængige identisk $\mathcal{N}(\mu, \sigma^2)$-fordelte; her er μ og σ^2 ukendte parametre. *Modelfunktionen* er

$$\begin{aligned} f(y_1, y_2, \ldots, y_n; \mu, \sigma^2) &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2} \frac{(y_j - \mu)^2}{\sigma^2} \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - \mu)^2 \right). \end{aligned} \tag{2.1}$$

Likelihoodfunktionen svarende til observationerne $y_1, y_2, \ldots, y_n$ er derfor

$$L(\mu, \sigma^2) = \text{konstant} \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - \mu)^2 \right). \tag{2.2}$$

2.1 Estimation af μ og σ^2

Vi vil bestemme maksimaliseringsestimaterne for μ og σ^2. Af udtrykket for likelihoodfunktionen ses at ligegyldigt hvilken værdi σ^2 måtte have, så er den bedste μ-værdi, altså den μ-værdi som maksimaliserer $\mu \mapsto L(\mu, \sigma^2)$, den værdi som *minimaliserer* kvadratsummen $\sum_{j=1}^{n} (y_j - \mu)^2$. Ved at benytte formlen for kvadratet på en toleddet størrelse kan kvadratsummen omskrives på

følgende måde hvor $\overline{y}$ betegner gennemsnittet af y-erne:

$$\begin{aligned}
&\sum_{j=1}^{n}(y_j-\mu)^2 \\
&= \sum_{j=1}^{n}\left((y_j-\overline{y})+(\overline{y}-\mu)\right)^2 \\
&= \sum_{j=1}^{n}\left((y_j-\overline{y})^2+2(y_j-\overline{y})(\overline{y}-\mu)+(\overline{y}-\mu)^2\right) \\
&= \sum_{j=1}^{n}(y_j-\overline{y})^2+\sum_{j=1}^{n}2(y_j-\overline{y})(\overline{y}-\mu)+\sum_{j=1}^{n}(\overline{y}-\mu)^2 \\
&= \sum_{j=1}^{n}(y_j-\overline{y})^2+2(\overline{y}-\mu)\sum_{j=1}^{n}(y_j-\overline{y})+n\,(\overline{y}-\mu)^2 \\
&= \sum_{j=1}^{n}(y_j-\overline{y})^2+n(\overline{y}-\mu)^2,
\end{aligned}$$

altså

$$\sum_{j=1}^{n}(y_j-\mu)^2 = \sum_{j=1}^{n}(y_j-\overline{y})^2+n(\overline{y}-\mu)^2. \tag{2.3}$$

Heraf ses at kvadratsummen er mindst netop når μ er lig med $\overline{y}$. Derfor *er* maksimaliseringsestimatet for μ faktisk gennemsnittet af observationerne,

$$\hat{\mu} = \overline{y},$$

således som det jo også var tanken at det skulle være.

Herefter kan man bestemme maksimaliseringsestimatet for σ^2 som maksimumspunktet for funktionen

$$\sigma^2 \mapsto L(\overline{y},\sigma^2),$$

og man finder at den antager sit maksimum når σ^2 har værdien

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j-\overline{y})^2.$$

Imidlertid benytter man som regel *ikke* dette estimat over σ^2, men derimod

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\overline{y})^2, \tag{2.4}$$

hvor divisoren $n-1$ i denne forbindelse kaldes for *antallet af frihedsgrader* for variansestimatet s^2.

Eksempel 2.2 (Lysets hastighed, fortsat)
Hvis vi går ud fra at de 64 positive værdier i Tabel 2.1 kan betragtes som observationer fra en og samme normalfordeling, så skal denne normalfordelings middelværdi estimeres til $\overline{y}=27.75$ og dens varians til $s^2=25.8$ med 63 frihedsgrader. Det betyder at passagetidens middelværdi estimeres til

$$(27.75\times10^{-3}+24.8)\times10^{-6}\text{sek}=24.828\times10^{-6}\text{sek}$$

og passagetidens varians estimeres til

$$25.8 \times (10^{-3} \times 10^{-6}\text{sek})^2 = 25.8 \times 10^{-6}(10^{-6}\text{sek})^2$$

med 63 frihedsgrader, dvs. standardafvigelsen estimeres til

$$\sqrt{25.8 \times 10^{-6}}10^{-6}\text{sek} = 0.005 \times 10^{-6}\text{sek}.$$

Beregningstips og -tricks

Når man skal udregne en konkret s^2-værdi, kan man naturligvis bare indsætte talværdierne i formel (2.4), det vil sige først udregne gennemsnittet $\overline{y}$, så trække det fra alle y_j-erne og kvadrere og summere, og til sidst dividere med $n-1$. Hvis man regner med håndkraft/lommeregner, er det imidlertid ofte en fordel at udnytte at summen af de kvadratiske afvigelser kan omskrives på følgende måde:

$$\begin{aligned}\sum_{j=1}^{n}(y_j-\overline{y})^2 &= \sum_{j=1}^{n}\left(y_j^2-2y_i\overline{y}+\overline{y}^2\right)\\ &= \sum_{j=1}^{n}y_j^2-n\overline{y}^2\\ &= \sum_{j=1}^{n}y_j^2-\frac{1}{n}\left(\sum_{j=1}^{n}y_j\right)^2.\end{aligned}$$

Summen af de kvadratiske afvigelser kan altså udregnes ved at man først finder *summen* og *summen af kvadraterne* af observationerne, og så indsætter dem i ovenstående forholdsvis simple formel.[1] Bemærk dog at metoden er temmelig følsom overfor afrundingsfejl (fordi den ender med at man skal trække to ofte meget store positive tal fra hinanden).

Metoden illustreres med et eksempel der samtidig omtaler endnu et par smarte tricks. Betragt følgende (konstruerede!) talmateriale:

$$\begin{aligned}y_1 &= 59837021\\ y_2 &= 59837023\\ y_3 &= 59837022\\ y_4 &= 59837021\\ y_5 &= 59837028\\ y_6 &= 59837023\\ y_7 &= 59837025\end{aligned}$$

Når vi her skal udregne gennemsnittet $\overline{y}$ af y_j-erne, er det smart at indføre et såkaldt beregningsnulpunkt a, f.eks. $a = 59837020$, og så udregne $\overline{y}$ som

[1] Mange lommeregnere har en »statistikknap« ($\Sigma+$) der gør det let at udregne $\overline{y}$ og s^2. Lommeregneren benytter tre hukommelsesregistre hvor den gemmer henholdsvis n, $\sum y$ og $\sum y^2$. Når man indtaster et tal og trykker på $\Sigma+$-tasten, opdateres de tre registre. Til sidst trykker man på nogle passende taster, og lommeregneren udregner $\overline{y}$ på den oplagte måde og s^2 ved hjælp af den her præsenterede formel.

$a + \overline{y-a}$. Med det omtalte valg af a bliver $\overline{y-a} = (1+3+2+1+8+3+5)/7 = \frac{23}{7} = 3\frac{2}{7} \approx 3.29$, og dermed $\overline{y} = 59837023.29$.

Summen af de kvadratiske afvigelser ændres ikke når man trækker det samme tal a fra alle y_j-erne (fordi det netop drejer sig om *afvigelser*). Ved beregningen kan vi derfor lade som om observationerne er tallene $y_j - a$, altså $1, 3, 2, 1, 8, 3, 5$; summen af disse tal fandt vi ovenfor til 23, og summen af deres kvadrater er $1+9+4+1+64+9+25 = 113$ så summen af de kvadratiske afvigelser (af y_j-erne eller af $y_j - a$-erne) er $113 - \frac{1}{7} \cdot 23^2 = 37\frac{3}{7} \approx 37.43$; endelig er så $s^2 = \frac{1}{7-1} \cdot 37\frac{3}{7} = 6\frac{5}{21} \approx 6.24$.

Men hvad nu hvis observationerne havde været f.eks. 10^6 gange mindre:

$$\begin{aligned} y_1 &= 59.837021 \\ y_2 &= 59.837023 \\ y_3 &= 59.837022 \\ y_4 &= 59.837021 \\ y_5 &= 59.837028 \\ y_6 &= 59.837023 \\ y_7 &= 59.837025 \end{aligned}$$

Så ville gennemsnittet ligeledes være blevet 10^6 gange mindre, og s^2 ville være blevet $10^6 \cdot 10^6 = 10^{12}$ gange mindre, altså $\overline{y} = 59.83702329$ og $s^2 = 6.24 \times 10^{-12}$.

Hvorfor benyttes s^2?

Det kan der argumenteres for på forskellige måder. Det lettest håndterlige og forståelige argument er at s^2 (i modsætning til $\hat{\sigma}^2$) er en *central* estimator over σ^2, hvilket vil sige at middelværdien af den stokastiske variabel s^2 er lig σ^2, altså $\mathrm{E}\, s^2 = \sigma^2$, således at estimatoren »i middel« rammer den rigtige værdi.

Bevis for at s^2 er central:
Antag at $Y_1, Y_2, \ldots, Y_n$ er uafhængige $\mathcal{N}(\mu, \sigma^2)$-variable. Der gælder at

$$\begin{aligned} \sum_{j=1}^{n} (Y_j - \overline{Y})^2 &= \sum_{j=1}^{n} \left((Y_j - \mu)^2 + 2(Y_j - \mu)(\mu - \overline{Y}) + (\mu - \overline{Y})^2 \right) \\ &= \sum_{j=1}^{n} (Y_j - \mu)^2 - n(\overline{Y} - \mu)^2. \end{aligned}$$

Ved at tage middelværdi fås (idet vi undervejs benytter at $\mathrm{E}(\overline{Y}) = \mu$ og $\mathrm{Var}(\overline{Y}) = \sigma^2/n$):

$$\begin{aligned} \mathrm{E} \sum_{j=1}^{n} (Y_j - \overline{Y})^2 &= \sum_{j=1}^{n} \mathrm{E}(Y_j - \mu)^2 - n\mathrm{E}(\overline{Y} - \mu)^2 \\ &= n\mathrm{Var}(Y) - n\mathrm{Var}(\overline{Y}) \\ &= (n-1)\mathrm{Var}(Y) \\ &= (n-1)\sigma^2, \end{aligned}$$

dvs.

$$\mathrm{E}\left(\frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \overline{Y})^2 \right) = \sigma^2.$$

□

Mod dette argument kan man indvende at det er baseret på et nyt princip (princippet om centrale estimatorer) der tilsyneladende blot er hentet ind på scenen til denne lejlighed. Hvis likelihoodmetoden virkelig skal være noget der er værd at beskæftige sig med, så burde man kunne basere sin argumentation udelukkende på den. Det kan man også til en vis grad, og det skal nu antydes hvordan.

De to parametre μ og σ^2 i normalfordelingen opfattes sædvanligvis ikke som værende ligestillede. Man plejer at tænke på middelværdiparameteren μ som den primære da den jo beskriver den *systematiske* variation, nemlig det *niveau* hvorom observationerne fordeler sig, hvorimod variansparameteren σ^2 der »kun« beskriver den tilfældige variation, kommer i anden række. Som en konsekvens heraf kan man mene at man ikke skal estimere de to parametre samtidigt, men at man *først* skal estimere μ og *dernæst* σ^2. Man skal derfor til estimationen af σ^2 kun benytte det der er tilbage af (informationen i) talmaterialet efter at man først har estimeret μ.

Hvis der f.eks. foreligger de fem observationer $3.2, 5.7, 2.1, 7.4, 3.1$ som tænkes at stamme fra en $\mathcal{N}(\mu, \sigma^2)$-fordeling, så estimeres først den »væsentlige« parameter μ ved gennemsnittet $(3.2 + 5.7 + 2.1 + 7.4 + 3.1)/5 = 21.5/5 = 4.3$. Dernæst skal man estimere σ^2 der skal beskrive den tilfældige variation omkring niveauet 4.3. Da det nu kan siges at være *givet* at de fem værdier skal have gennemsnit 4.3, dvs. at de fem afvigelser fra gennemsnittet skal summere til 0, så er der på sin vis kun fire *forskellige* afvigelser. Når man skal estimere variansen (der jo er den forventede kvadratiske afvigelse af en observation fra middelværdien), bliver det derfor som summen af de kvadratiske afvigelser divideret med *fire*:

$$\begin{aligned}
&\Big((3.2-4.3)^2 + (5.7-4.3)^2 + (2.1-4.3)^2 \\
&+ (7.4-4.3)^2 + (3.1-4.3)^2\Big)/4 \\
&= \quad \Big((-1.1)^2 + 1.4^2 + (-2.2)^2 + 3.1^2 + (-1.2)^2\Big)/4 \\
&= \quad 19.08/4 \\
&= \quad 4.77
\end{aligned}$$

Man siger at der er fire *frihedsgrader* fordi når det er fixeret at de fem observationer skal have et bestemt gennemsnit (f.eks. 4.3), så kan man vælge fire af de fem afvigelser fra gennemsnittet frit.

Ovenstående argument for at dividere summen af de kvadratiske afvigelser med $n-1$ i stedet for med n kan jo roligt siges at være noget løst og upræcist, men det kan faktisk godt præciseres. Det forhold at variansparameteren σ^2 tænkes at spille en underordnet rolle i forhold til middelværdiparameteren μ, og at dette skal afspejles i den måde parametrene skal estimeres på, kan formaliseres på følgende måde: Man skal først estimere μ på sædvanlig måde, men dernæst skal man estimere σ^2 i den *betingede model* hvor man betinger med $\hat{\mu}$, altså med $\overline{y}$. Estimatet over σ^2 skal være maximum likelihood estimatet, men man skal vel at mærke benytte likelihoodfunktionen svarende til *den betingede fordeling af* $Y_1, Y_2, \ldots, Y_n$ *givet at* $\overline{Y}$ *er lig med* $\overline{y}$. Hvis det skal gå bare nogenlunde matematisk korrekt til, er det ikke noget simpelt problem at bestemme denne betingede fordeling – det skyldes at der er tale om kontinuerte fordelinger. Men hvis man i al naivitet regner med at der gælder nogenlunde det samme som for diskrete fordelinger, blot med tæthedsfunktioner i stedet for sandsynlighedsfunktioner, så skulle den betingede tæthedsfunktion være

$$\frac{\text{tæthedsfunktionen for } Y_1, Y_2, \ldots, Y_n}{\text{tæthedsfunktionen for } \overline{Y}}\,.$$

Da $Y_1, Y_2, \ldots, Y_n$ er uafhængige $\mathcal{N}(\mu, \sigma^2)$-variable, vil gennemsnittet $\overline{Y}$ være $\mathcal{N}(\mu, \sigma^2/n)$-fordelt. Derfor bliver den betingede tæthedsfunktion

$$\begin{aligned}
&\frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2\right)}{\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left(-\frac{1}{2}\frac{(\overline{y}-\mu)^2}{\sigma^2/n}\right)} \\
&= \quad \text{konstant} \cdot (\sigma^2)^{-\frac{n-1}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\overline{y})^2\right).
\end{aligned}$$

Opfattet som funktion af σ^2 skulle dette så være den betingede likelihoodfunktion (hvor i øvrigt μ meget bekvemt er forsvundet ud af billedet), altså den likelihoodfunktion der skal benyttes ved estimation af σ^2. Den betingede likelihoodfunktion er en funktion af én variabel σ^2, og man finder

at den antager sit maksimum i ét punkt, nemlig når σ^2 har værdien s^2. Der gælder altså at i den betingede model er størrelsen

$$s^2 \quad = \quad \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\overline{y})^2$$

et maximum likelihood estimat over σ^2.

2.2 Test af hypotese om middelværdien

Man er undertiden interesseret i at undersøge om de foreliggende data er forenelige med en antagelse om at den teoretiske middelværdi μ har en bestemt værdi (f.eks. 0). Mere formelt ønsker man at teste den statistiske hypotese $H_0 : \mu = \mu_0$ hvor μ_0 er et kendt tal.

Hypoteser om parametre i normalfordelinger testes principielt på samme måde som alle andre statistiske hypoteser, nemlig ved brug af et kvotienttest der sammenligner likelihoodfunktionens maksimale værdi under hypotesen med den maksimale værdi overhovedet under den givne model. Likelihoodfunktionen er givet i formel (2.2) på side 14, og dens maksimale værdi er $L(\overline{y}, \hat{\sigma}^2)$. Under H_0 er likelihoodfunktionen

$$L_0(\sigma^2) \quad = \quad L(\mu_0, \sigma^2)$$

og den antager sin maksimumsværdi når σ^2 er lig med

$$\hat{\hat{\sigma}}^2 \quad = \quad \frac{1}{n}\sum_{j=1}^{n}(y_j-\mu_0)^2.$$

Kvotienttesstørrelsen bliver derfor

$$\begin{aligned}
Q &= \frac{L(\mu_0, \hat{\hat{\sigma}}^2)}{L(\overline{y}, \hat{\sigma}^2)} \\
&= \left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right)^{-n/2} \exp\left(-\left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{2\hat{\hat{\sigma}}^2} - \frac{\sum_{j=1}^{n}(y_j-\overline{y})^2}{2\hat{\sigma}^2}\right)\right) \\
&= \left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}\right)^{-n/2} \exp\left(-(\frac{n}{2}-\frac{n}{2})\right) \\
&= \left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}\right)^{-n/2}.
\end{aligned}$$

Her omskrives kvadratsummen i tælleren ved hjælp af formel (2.3) på side 15 (med μ erstattet af μ_0), og man får

$$\begin{aligned} Q &= \left(\frac{\sum_{j=1}^{n}(y_j-\overline{y})^2 + n(\overline{y}-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2} \right)^{-n/2} \\ &= \left(1 + \frac{n(\overline{y}-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2} \right)^{-n/2} \\ &= \left(1 + \frac{n(\overline{y}-\mu_0)^2}{(n-1)s^2} \right)^{-n/2} \\ &= \left(1 + \frac{1}{n-1} \left(\frac{\overline{y}-\mu_0}{\sqrt{s^2/n}} \right)^2 \right)^{-n/2} . \end{aligned}$$

Størrelsen $(\overline{y}-\mu_0)/\sqrt{s^2/n}$ plejer man at betegne t:

$$t = \frac{\overline{y}-\mu_0}{\sqrt{s^2/n}} ,$$

og med denne betegnelse har vi at

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

Nu er det jo sådan at *små* værdier af Q tyder på at hypotesen H_0 *ikke* er forenelig med data, og det ses at små Q-værdier er ensbetydende med t-værdier langt fra 0, dvs. med store $|t|$-værdier. Man kan derfor benytte t som teststørrelse i stedet for Q, hvilket er praktisk da t er lettere at beregne end Q. – Undertiden kaldes t-teststørrelsen for *Student's t* fordi W.S.Gosset der skrev den første artikel om t-testet (i 1908), skrev under pseudonymet 'Student'.

Bemærk at t-teststørrelsen også ud fra en umiddelbar betragtning forekommer at være en fornuftig teststørrelse idet den måler afvigelsen $\overline{y}-\mu_0$ mellem den observerede og den teoretiske middelværdi i forhold til $\sqrt{s^2/n}$ som er den estimerede middelfejl på på $\overline{y}$ (dvs. standardafvigelsen på $\overline{y}$).

Når man har fundet værdien af teststørrelsen t, er næste skridt i testproceduren at bestemme *testsandsynligheden*, altså sandsynligheden for at få en mere ekstrem værdi af teststørrelsen end den faktisk opnåede, forudsat at hypotesen H_0 er rigtig. En matematisk sætning fortæller at når H_0 er rigtig, så følger t-størrelsen en bestemt fordeling, nemlig en såkaldt *t-fordeling med $f = n-1$ frihedsgrader*; frihedsgradsantallet i t-fordelingen arves fra frihedsgradsantallet for variansestimatet s^2 i nævneren.[2]

[2]Når H_0 er er rigtig, afhænger fordelingen af t hverken af μ_0 eller af σ^2, hvilket er bekvemt da vi jo ikke kender de nøjagtige værdier heraf.

I statistiske tabelværker kan man finde tabeller over fraktiler i t-fordelingen, og ved hjælp af sådanne tabeller er det let at bestemme testsandsynligheder i t-testet. Man skal dog være opmærksom på at en »mere ekstrem t-værdi« som oftest vil sige en t-værdi således at $|t| > |t_{\text{obs}}|$, dvs.

$$t > |t_{\text{obs}}| \quad \text{eller} \quad t < -|t_{\text{obs}}| \,.$$

Man vil altså forkaste hypotesen både hvis t_{obs} er meget stor og hvis den er meget lille.[3] Der gælder at t-fordelingen er symmetrisk omkring 0, hvilket medfører at

$$\mathrm{P}_0\big(t > |t_{\text{obs}}|\big) \quad = \quad \mathrm{P}_0\big(t < -|t_{\text{obs}}|\big)$$

og dermed

$$\mathrm{P}_0\big(|t| > |t_{\text{obs}}|\big) \quad = \quad 2\,\mathrm{P}_0\big(t > |t_{\text{obs}}|\big) \,.$$

Eksempel 2.3 (Lysets hastighed, fortsat)
I vore dage er en meter pr. definition den strækning som lyset i vacuum gennemløber på $1/299\,792\,458$ sekund, hvoraf følger at lysets hastighed er 299 792 458 meter pr. sekund. Med denne hastighed vil lyset være $\tau_0 = 2.48238 \times 10^{-5}$ sekunder om at tilbagelægge strækningen på de 7442 meter. Størrelsen τ_0 svarer til en tabelværdi på $((\tau_0 \times 10^6) - 24.8) \times 10^3 = 23.8$, så det ville være interessant at undersøge om de foreliggende data er forenelige med hypotesen om at den ukendte middelværdi μ har værdien $\mu_0 = 23.8$. Derfor vil vi teste den statistiske hypotese $H_0 : \mu = 23.8$.

Vi har tidligere fundet at $\overline{y} = 27.75$ og $s^2 = 25.8$, så t-teststørrelsen er

$$t \;=\; \frac{27.75 - 23.8}{\sqrt{25.8/64}} \;=\; 6.2 \,.$$

Da der ikke er nogen grund til at tro at der kun skulle kunne forekomme afvigelser i én retning, skal testet være tosidet. Testsandsynligheden er derfor sandsynligheden for at få t-værdier som enten er større end 6.2 eller mindre end -6.2. Ved tabelopslag kan man finde at i t-fordelingen med 63 frihedsgrader er 99.95%-fraktilen lidt over 3.4, dvs. der mindre end 0.05% sandsynlighed for at få en værdi som er større end 6.2, og testsandsynligheden er dermed mindre $2 \times 0.05\% = 0.1\%$. En så lille testsandsynlighed betyder at man må *forkaste* hypotesen. Newcombs målinger af lysets passagetid stemmer altså *ikke* overens med hvad vi i dag ved om lysets hastighed.

Vi ser at Newcombs passagetider er en smule for store, og da den lyshastighed vi her har benyttet, er lysets hastighed i vacuum, kan noget af forklaringen være at lyset bevæger sig en smule langsommere i luft end i vacuum.

2.3 Histogrammer og fraktildiagrammer

For at få en idé om modellens rimelighed vil man ofte i et »enstikprøveproblem i normalfordelingen« tegne histogrammer og fraktildiagrammer.

[3] Et sådant test kaldes et *tosidet test*, i modsætning til et *ensidet test* der regner med at de »ekstreme« afvigelser kun kan være til den ene side, f.eks. den positive, så at man kun forkaster hvis den observerede t-værdi er meget stor.

Figur 2.1 Histogram over 64 målte værdier af lysets passagetid. – Den indtegnede kurve er tætheden for normalfordelingen med parametre $\overline{y} = 27.75$ og $s^2 = 25.8$.

Histogrammer

Et histogram over et sæt observationer $y_1, y_2, \ldots, y_n$ fås på følgende måde:

1. Inddel observationsaksen i et antal delintervaller, gerne lige store, sådan at der ikke er nogen observationer i intervalendepunkterne.

2. Tæl op hvor mange observationer der er i hvert interval.

3. Tegn rektangler hvis grundflader er delintervallerne, og hvis arealer er lig med den brøkdel af observationerne som ligger inden for det pågældende delinterval. (Hvis der er a observationer i et interval af længde l, skal rektanglets højde være a/nl.)

4. Histogrammet skal ligne tæthedsfunktionen for den formodede sandsynlighedsfordeling (her en normalfordeling). Det er derfor en god idé at indtegne den estimerede fordelings tæthedsfunktion i samme figur som histogrammet, se Figur 2.1.

Ved udarbejdelsen af et histogram kan det være lidt af et kunststykke at vælge den rigtige intervalinddeling således at fluktuationerne bliver passende udglattet uden at tæthedens form bliver alt for udjævnet. Hvis intervallerne er for korte, bliver fluktuationerne ikke udglattet nok, er de for lange, sker der en for stor udjævning af tæthedens form.

Man kan godt opskrive definitionen på et histogram over et sæt observationer $y_1, y_2, \ldots, y_n$ lidt mere formelt:

1. I det område hvor observationerne falder vælges delepunkter (der som regel bør være ækvidistante) $x_0 < x_1 < x_2 < \ldots < x_m$ hvor x_0 er mindre end den mindste og x_m større end den største af y-observationerne.

2. Bestem antallet n_j af y-er i det j-te interval (som er $]x_{j-1}, x_j]$).

3. Definer den stykkevis konstante funktion h som

$$h(y) \quad = \quad \begin{cases} \dfrac{n_j/n}{x_j - x_{j-1}} & \text{når} \quad y \in]x_{j-1}, x_j] \,, \\ 0 & \text{når} \quad y \leq x_0 \text{ eller } y > x_m \,. \end{cases}$$

   Så er histogrammet (svarende til den valgte inddeling) over observationerne $y_1, y_2, \ldots, y_n$ ganske simpelt grafen for h.

Figur 2.2 Fraktildiagram over 64 målte værdier af lysets passagetid.

Fraktildiagrammer

Når man har et sæt observationer $y_1, y_2, \ldots, y_n$, benytter man traditionelt betegnelsen $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ for de *ordnede observationer*, dvs. y-erne stillet op i voksende rækkefølge.

Nu er det sådan at hvis alle de observerede y-er er forskellige, så er brøkdelen $(i-1)/n$ af observationerne strengt mindre end tallet $y_{(i)}$, og brøkdelen i/n af dem er mindre end eller lig med tallet $y_{(i)}$. Som et kompromis kan man da sige at brøkdelen $(i-0.5)/n$ af dem er mindre end tallet $y_{(i)}$, med andre ord er $y_{(i)}$ en $\frac{i-0.5}{n}$-fraktil i den empiriske fordeling. – Generelt defineres en α-fraktil i en fordeling som et tal y_α med den egenskab at brøkdelen α af fordelingen ligger til venstre for y_α.

Et fraktildiagram er kort fortalt en tegning hvor man afsætter teoretiske fraktiler mod empiriske fraktiler. Hvis y-erne er observationer fra $\mathcal{N}(\mu, \sigma^2)$-fordelingen, så er den teoretiske fordelingsfunktion funktionen $y \mapsto \Phi(\frac{y-\mu}{\sigma})$ (side 10). Derfor finder man den teoretiske α-fraktil y_α ved at løse ligningen $\Phi(\frac{y_\alpha-\mu}{\sigma}) = \alpha$, hvilket giver $y_\alpha = \mu + \sigma \cdot \Phi^{-1}(\alpha)$. De n punkter hvis førstekoordinater er de empiriske fraktiler, og hvis andenkoordinater er de tilsvarende teoretiske fraktiler, det vil sige punkterne med koordinater

$$\left(y_{(i)}, \mu + \sigma \cdot \Phi^{-1}(\tfrac{i-0.5}{n})\right), \quad i = 1, 2, \ldots, n,$$

bør da ligge nogenlunde omkring en ret linie gennem $(0, 0)$ med hældning 1. Dette er ensbetydende med at punkterne med koordinater

$$\left(y_{(i)}, \Phi^{-1}(\tfrac{i-0.5}{n})\right), \quad i = 1, 2, \ldots, n,$$

ligger nogenlunde omkring den rette linie gennem $(\mu, 0)$ med hældning $1/\sigma$.

Konkret fremstiller man fraktildiagrammet ved at indtegne punkterne

$$\left(y_{(i)}, \Phi^{-1}(\tfrac{i-0.5}{n})\right), \quad i = 1, 2, \ldots, n$$

i et koordinatsystem hvor man desuden indtegner den rette linie gennem $(\overline{y}, 0)$ med hældning $1/s$; funktionen Φ^{-1} findes tabelleret i statistiske tabelværker og er en standardfunktion i statistikprogrammer til computere.

Med *sandsynlighedspapir* kan man fremstille fraktildiagrammer med håndkraft uhyre let og uden at skulle bekymre sig om funktionen Φ^{-1} (den er nemlig indbygget i sandsynlighedspapiret). Sandsynlighedspapiret er indrettet på den måde at ordinataksen har to skalaer: en *probit-skala* som er ækvidistant og går fra knap 2 til godt 8, og en (ikke-ækvidistant) *sandsynlighedsskala* med sandsynligheder i procent, gående fra 0.05 til 99.95. Man afsætter nu punkterne

Tabel 2.2 Data til Opgave 2.1.

0.606	0.619	0.645
0.693	0.740	0.761
0.768	0.798	0.843
0.849	0.891	0.965
0.970	0.996	1.129
1.265	1.378	1.421

$(y_{(i)}, \frac{i-0.5}{n})$ idet man benytter sandsynlighedsskalaen på ordinataksen; hvis tallene er normalfordelte, skal punkterne fordele sig omkring den rette linie der kan indtegnes ved at benytte probit-skalaen på ordinataksen og lade linien gå gennem punkterne $(\overline{y} - s, 4)$, $(\overline{y}, 5)$, $(\overline{y} + s, 6)$ osv.

Figur 2.2 viser et fraktildiagram over de 64 målte værdier af lysets passagetid. Såvel histogrammet (side 22) som fraktildiagrammet viser, at det ikke er ganske urimeligt at antage at måleresultaterne er normalfordelte.

2.4 Opgaver

Opgave 2.1
Tallene i Tabel 2.2 kan opfattes som et »enstikprøveproblem« i normalfordelingen. Vi betegner tallene $y_1, y_2, \ldots, y_n$ $(n = 18)$.

1. Udregn gennemsnittet $\overline{y}$ af observationerne.
2. Udregn summen af kvadratiske afvigelser $\sum_{j=1}^{n} (y_j - \overline{y})^2$ på to måder,
   (a) dels på den »umiddelbare« måde, dvs. udregn de 18 differenser $y_j - \overline{y}$, kvadrér differenserne og summér dem,
   (b) dels ved at benytte det snedige trick fra side 16.
3. Udregn variansskønnet og skønnet over standardafvigelsen.
4. Standardafvigelsen på gennemsnittet $\overline{y}$ er $1/\sqrt{n}$ gange standardafvigelsen på y-erne. Udregn den estimerede standardafvigelse på gennemsnittet.

   (Standardafvigelsen på gennemsnittet kaldes ofte *middelfejlen* på $\overline{y}$.)
5. Med hvor mange cifre bør man angive værdien af $\overline{y}$?

Opgave 2.2 (Kviksølv i sværdfisk)
Sværdfisk kan være en kulinarisk oplevelse, men de er sundest når de ikke indeholder alt for mange tungmetaller. I en undersøgelse af sværdfisk på det

Tabel 2.3 Opgave 2.2: Kviksølvindhold (ppm) i 115 sværdfisk, de ordnede observationer.

0.05	0.07	0.07	0.13	0.13	0.19	0.24	0.25	0.28	0.32
0.39	0.45	0.46	0.53	0.54	0.56	0.60	0.60	0.61	0.62
0.65	0.71	0.72	0.75	0.76	0.79	0.81	0.81	0.82	0.82
0.82	0.83	0.83	0.83	0.84	0.85	0.89	0.90	0.91	0.92
0.92	0.93	0.95	0.95	0.97	0.97	0.98	1.00	1.00	1.01
1.02	1.04	1.05	1.05	1.08	1.10	1.12	1.12	1.14	1.14
1.15	1.16	1.20	1.20	1.20	1.20	1.20	1.21	1.22	1.25
1.25	1.26	1.27	1.27	1.29	1.29	1.29	1.29	1.30	1.31
1.32	1.32	1.37	1.37	1.39	1.39	1.40	1.40	1.41	1.42
1.43	1.44	1.45	1.54	1.54	1.58	1.58	1.60	1.60	1.62
1.62	1.66	1.66	1.68	1.69	1.72	1.74	1.85	1.89	1.96
2.06	2.10	2.23	2.25	2.72					

amerikanske marked har man målt kviksølvindholdet i 115 tilfældigt udvalgte sværdfisk og fået resultaterne i Tabel 2.3.[4]

Ifølge de amerikanske sundhedsmyndigheder bør konsumfisk ikke indeholde over 1 ppm kviksølv. Den fisk der sælges via de autoriserede salgskanaler, kan man kontrollere (med stikprøvekontroller), og man kan så kassere de partier der indeholder for meget kviksølv. Imidlertid sælges der også en del fisk uden om kontrolmyndighederne – i USA regner man med ca. 25%. Man er interesseret i at vide hvordan man skal vælge kassationsgrænsen for de 75% kontrollerede fisk for at opnå at gennemsnitsindholdet af kviksølv i de fisk der når frem til forbrugeren, bliver 1 ppm (eller derunder). Hvis man skal kunne beregne denne grænse, er man nødt til at kende fordelingen af kviksølvindhold i sværdfisk.

1. Det ville være bekvemt hvis observationerne kunne beskrives ved en normalfordeling, så det ønsker man at undersøge.
   (a) Udregn estimaterne $\overline{y}$ og s^2 over μ og σ^2.
   (b) Tegn et histogram over kviksølvindholdet i de 115 sværdfisk. Indtegn (skitsemæssigt) den fittede normalfordelingstæthed (dvs. tætheden for normalfordelingen med parametre $\overline{y}$ og s^2).
   (c) Tegn et fraktildiagram (f.eks. på sandsynlighedspapir). Indtegn den rette linie der svarer til den fittede normalfordeling.
2. I den oprindelige analyse af tallene gik man ud fra at kviksølvkoncentrationen i sværdfisk var *logaritmisk normalfordelt*, hvilket betyder at *logaritmen* til koncentrationerne er normalfordelt. Diskutér denne formodning.

TIP: Summen af observationerne er 126.70, og summen af kvadraterne er 168.0858. For *logaritmen* (den naturlige logaritme) til observationerne er de tilsvarende tal -7.9070 og 56.8102.

[4] Lee & Krutchkoff (1980): Mean and variance of partially-truncated distributions. *Biometrics* **36,** 531-6.

Tabel 2.4 Data til Opgave 2.4: 20 eksempler på udfald af stokastiske variable $Y_1, Y_2, \ldots, Y_{10}$ frembragt af en normalfordelings-tilfældighedsmekanisme med middelværdi 5 og varians 3.

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	$\overline{y}$	s^2
5.80	5.06	7.69	4.10	5.13	5.49	1.91	6.90	4.34	5.61	5.20	2.50
7.43	7.03	4.92	4.69	8.43	5.24	5.86	2.26	4.17	4.81	5.48	3.18
6.45	4.06	7.48	6.57	4.87	6.42	5.00	6.05	4.11	1.25	5.23	3.23
3.63	5.55	6.90	5.80	3.90	6.79	4.71	3.97	5.49	8.99	5.57	2.76
4.79	6.01	2.71	7.31	5.18	4.48	6.50	4.21	7.98	5.05	5.42	2.44
3.67	4.92	1.72	6.80	5.14	5.08	5.85	5.03	3.49	5.21	4.69	1.99
5.32	7.16	5.63	4.70	4.38	7.18	5.53	5.25	4.99	5.09	5.52	0.89
2.34	3.57	5.10	4.03	3.17	7.48	5.37	4.05	5.47	5.43	4.60	2.16
5.56	5.32	6.25	7.43	1.16	2.62	6.87	5.31	1.70	6.42	4.86	4.95
7.79	6.08	8.13	4.98	3.27	6.01	3.26	1.82	3.28	5.79	5.04	4.39
5.54	3.58	5.26	4.79	3.97	6.01	4.47	4.98	2.47	3.44	4.45	1.18
3.87	5.79	5.56	5.36	8.25	7.48	3.21	4.37	1.81	5.26	5.10	3.64
4.87	8.49	5.54	7.83	3.91	3.61	3.10	5.15	6.80	3.92	5.32	3.40
3.95	5.24	7.46	6.46	3.36	3.21	7.58	3.26	4.83	8.06	5.34	3.69
8.75	4.19	3.41	8.17	3.46	3.89	4.62	7.08	6.18	4.16	5.39	4.00
5.32	6.49	6.13	4.28	5.52	4.37	5.37	6.00	4.51	2.98	5.10	1.13
7.10	5.57	3.76	5.31	4.15	4.53	3.99	5.09	4.25	6.53	5.03	1.25
3.61	4.80	3.44	6.29	3.72	0.19	6.48	5.90	6.30	5.92	4.66	3.92
1.45	4.01	7.06	6.61	0.47	2.20	3.07	4.88	6.15	5.15	4.11	5.10
4.21	6.90	5.06	5.60	7.80	4.12	6.22	5.91	6.42	4.30	5.65	1.53

Opgave 2.3 (fortsættelse af Opgave 2.2; svær)
Løs det der er det overordnede problem i Opgave 2.2, nemlig hvordan skal man fastsætte kassationsgrænsen for de 75% af fiskene der kontrolleres hvis man vil opnå at forbrugeren i gennemsnit højst udsættes for en kviksølvbelastning på 1 ppm.

Opgave 2.4
Tabel 2.4 indeholder 20 stikprøver $y_1, y_2, \ldots, y_{10}$ fra en normalfordeling med parametre $\mu = 5$ og $\sigma^2 = 3$.

1. Hvordan fordeler de enkelte stikprøvers estimerede middelværdier $\overline{y}$ sig omkring den teoretiske middelværdi $\mu = 5$?

2. Man kan bevise at gennemsnittet af n $\mathcal{N}(\mu, \sigma^2)$-fordelte størrelser kan opfattes som en observation fra $\mathcal{N}(\mu, \sigma^2/n)$-fordelingen. De 20 gennemsnit $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_{20}$ skulle altså være observationer fra en normalfordeling med middelværdi 5 og varians $3/10$. Ser det ud til at passe?

   (a) Udregn gennemsnittet $\overline{\overline{y}} = (\overline{y}_1 + \overline{y}_2 + \ldots + \overline{y}_{20})/20$ og den empiriske varians på $\overline{y}_i$-erne, dvs. $\dfrac{1}{20-1}\sum_{i=1}^{20}(\overline{y}_i - \overline{\overline{y}})^2$.

   Giver det cirka 5 og 0.3 ?

(b) Tegn et fraktildiagram over $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_{20}$.

3. Udregn for hver af de 20 stikprøver t-teststørrelsen for hypotesen $\mu = 5$.

   Hvordan fordeler t-værdierne sig?

   Udregn de 20 testsandsynligheder. Hvor mange af dem er under 5% ?

   Er tingene som man skulle forvente – og hvad skulle man egentlig forvente?

4. I realiteten foreligger der jo 200 observationer fra en og samme normalfordeling. Skitsér hvordan man ud fra disse 200 observationer kunne teste hypotesen om at den teoretiske middelværdi er lig 5.

3 Tostikprøveproblemer i normalfordelingen

En ofte forekommende situation er at der foreligger målinger af en bestemt egenskab hos et antal individer der på forhånd vides at tilhøre forskellige grupper. Alt afhængigt af karakteren af målingerne kan man så benytte den ene eller anden eller tredje statistiske model/metode for dels at beskrive, dels at sammenligne de pågældende grupper. I dette kapitel skal vi diskutere metoder der kan benyttes, når

- der på hvert individ er målt én enkelt talværdi,
- talværdien opfattes som værende en værdi på en kontinuert måleskala,
- man vælger at beskrive den tilfældige variation med en normalfordeling.

Når betingelserne er formuleret i vendinger som »opfattes som værende« og »vælger at beskrive«, skyldes det at normalfordelingen ofte benyttes også i situationer hvor man kunne pege på andre mere rigtige fordelinger. Tit er der en eller to forholdsvis gode grunde til alligevel at benytte normalfordelingen. Den ene grund er *Den Centrale Grænseværdisætning* der siger at *summer* af et større antal stokastiske variable under visse milde omstændigheder med god tilnærmelse er normalfordelt, og de størrelser man laver statistiske modeller for, er netop tit sådanne summer. Den anden grund er rent pragmatisk: Normalfordelingsmodeller er fra et matematisk-statistisk synspunkt særdeles »pæne« i den forstand at når man i normalfordelingsmodeller benytter de generelle statistiske principper, så bliver resultatet næsten altid pæne og simple metoder der ofte er lette at forstå og giver nemme og forståelige udregninger osv. Som følge heraf er normalfordelingsmodeller studeret og beskrevet i alle detaljer, og man kan for det meste finde en teoretisk gennemregnet model der passer til ens behov.

Hvori består problemet?

Antag at der er tale om en situation hvor man på hvert af et antal »individer« har målt værdien af en bestemt variabel Y. Individer skal her forstås i meget bred forstand: det kan bl.a. være personer, forsøgsdyr, jordlodder eller f.eks. de enkelte realisationer af forsøget »måling af lysets hastighed«. Individerne er opdelt i grupper ud fra nogle kriterier som er kendt på forhånd (inden forsøget starter), og som ikke afhænger af hvilken værdi Y nu måtte have. I den statistiske model for Y-erne vil man regne med at den forskel der er mellem

(Y-værdierne hos) individerne *inden for* en bestemt gruppe, er *tilfældig*, og at den forskel der er *mellem* forskellige grupper, er *systematisk*. En normalfordelingsmodel til denne situation er da indrettet på den måde at

- den *systematiske* forskel mellem grupper beskrives ved hjælp af middelværdiparametre, og
- den *tilfældige* forskel inden for grupper beskrives ved hjælp af dels normalfordelingen, dels variansparametre i normalfordelingen.

Det statistiske problem består tit i at man ønsker at sammenligne grupperne for at vurdere om den systematiske forskel mellem dem er signifikant, dvs. om den forskel der er mellem grupperne, er stor målt i forhold til den tilfældige variation inden for de enkelte grupper. Man ønsker derfor at kunne måle forskellen mellem grupperne med en målestok der er kalibreret efter størrelsen af den tilfældige variation inden for grupperne.

Det man egentlig er interesseret i, er altså information om middelværdiparametrene. Men for at der kan være en veldefineret målestok at måle dem med, må man først sikre sig at det har mening at tale om *den* tilfældige variation inden for grupper. Derfor man i må modellen gøre den antagelse (som undertiden kan testes) at der er *varianshomogenitet*, dvs. at de forskellige grupper har samme variansparameter.[1]

Hermed er problemet beskrevet i generelle vendinger. I resten af dette kapitel og i Kapitel 4 skal vi se hvordan det kan løses.

Der er tradition for at man giver en særlig omtale af den situation hvor der er *to* grupper der skal sammenlignes, så det gør vi også her.

3.1 Tostikprøveproblemet med uparrede observationer

Man har to grupper af »individer«, og på hvert individ har man målt værdien af en bestemt variabel Y. Individerne i den ene gruppe hører ikke sammen med dem i den anden gruppe på nogen måde, de er *uparrede*. Der behøver heller ikke være lige mange observationer i de to grupper. Skematisk ser situationen sådan ud:

gruppe	observationer					
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}

Her betegner y_{ij} observation nr. j i gruppe nr. i, $i = 1, 2$. Grupperne har henholdsvis n_1 og n_2 observationer. Vi vil gå ud fra at forskellen mellem observationer inden for en gruppe er tilfældig, hvorimod der er en systematisk forskel på to de grupper – det er derfor at observationerne er inddelt i grupper! Endelig antages at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij} som er normalfordelte med samme varians σ^2 og med middelværdier

[1] Man kan dog klare sig med en antagelse om at gruppernes variansparametre er kendte pånær en konstant faktor.

henholdsvis μ_1 og μ_2, kort

$$\begin{aligned} Y_{1j} &\sim \mathcal{N}(\mu_1, \sigma^2) \\ Y_{2j} &\sim \mathcal{N}(\mu_2, \sigma^2). \end{aligned}$$

På denne måde beskriver de to middelværdiparametre μ_1 og μ_2 den *systematiske variation*, dvs. de to gruppers niveauer, medens variansparameteren σ^2 (samt normalfordelingen) beskriver den *tilfældige variation* der altså er den samme i begge grupper (denne antagelse kan man eventuelt teste, se side 35).

Estimation af μ_1 og μ_2

Estimater over de ukendte middelværdiparametre μ_1 og μ_2 findes ved maximum likelihood metoden, altså som de værdier der maksimaliserer likelihoodfunktionen

$$\begin{aligned} &L(\mu_1, \mu_2, \sigma^2) \\ &= \prod_{j=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(y_{1j}-\mu_1)^2}{\sigma^2}\right) \times \prod_{j=1}^{n_2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(y_{2j}-\mu_2)^2}{\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\mu_2)^2\right)\right), \end{aligned}$$

hvor $n = n_1 + n_2$ er det samlede antal observationer. Det ses at hvis σ^2 er fast, så er det at *maksimalisere* likelihoodfunktionen L med hensyn til μ_1 og μ_2 det samme som det at *minimalisere* kvadratsummen

$$\sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\mu_2)^2,$$

og den opgave er som vi skal se, let at løse.

Vi lader $\overline{y}_i$ betegne gennemsnittet i gruppe i, $\overline{y}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$. Det snedige trick er nu følgende omskrivning af det j-te led fra gruppe 1 (vi benytter formlen for kvadratet på en toleddet størrelse):

$$\begin{aligned} (y_{1j}-\mu_1)^2 &= \big((y_{1j}-\overline{y}_1) + (\overline{y}_1-\mu_1)\big)^2 \\ &= (y_{1j}-\overline{y}_1)^2 + 2(y_{1j}-\overline{y}_1)(\overline{y}_1-\mu_1) + (\overline{y}_1-\mu_1)^2. \end{aligned}$$

Når vi summerer over j, bliver summen af de dobbelte produkter 0 fordi summen af afvigelserne fra $\overline{y}_1$ er 0, så

$$\begin{aligned} \sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2 &= \sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2 + \sum_{j=1}^{n_1}(\overline{y}_1-\mu_1)^2 \\ &= \sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2 + n_1(\overline{y}_1-\mu_1)^2. \end{aligned}$$

Fra gruppe 2 kommer der et tilsvarende bidrag, så alt i alt kan den kvadratsum der skal minimaliseres, skrives som

$$\begin{aligned}&\sum_{j=1}^{n_1}(y_{1j}-\mu_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\mu_2)^2\\ &= \sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_2)^2+n_1(\overline{y}_1-\mu_1)^2+n_2(\overline{y}_2-\mu_2)^2.\end{aligned}$$

Det ses at de værdier af μ_1 og μ_2 der gør kvadratsummen mindst, er $\mu_1 = \overline{y}_1$ og $\mu_2 = \overline{y}_2$. Vi har dermed fundet at maksimaliseringsestimaterne for gruppemiddelværdierne μ_1 og μ_2 er gruppegennemsnittene $\overline{y}_1$ og $\overline{y}_2$.

Estimation af σ^2

Maksimaliseringsestimatet $\hat{\sigma}^2$ for σ^2 kan bestemmes som maksimumspunktet for funktionen

$$\sigma^2 \quad \mapsto \quad L(\overline{y}_1, \overline{y}_2, \sigma^2)\,;$$

man finder at

$$\hat{\sigma}^2 \quad = \quad \frac{1}{n}\left(\sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_2)^2\right).$$

En størrelse som $y_{ij}-\overline{y}_i$ der er forskellen mellem den faktiske observation og det bedst mulige *fit* under den aktuelle model, kaldes undertiden for et *residual*.[2] Derfor kaldes en størrelse som

$$\sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_2)^2$$

for en *residualkvadratsum*, og man kan sige at maksimaliseringsestimatet $\hat{\sigma}^2$ for σ^2 er lig med residualkvadratsummen divideret med antallet af observationer. Som regel benytter man imidlertid et andet estimat over σ^2, nemlig residualkvadratsummen divideret med *antallet af frihedsgrader* $n-2$ (antal observationer minus antal estimerede middelværdiparametre), dvs. man estimerer variansen ved

$$s_0^2 \quad = \quad \frac{1}{n-2}\left(\sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\overline{y}_2)^2\right).$$

Man begrunder brugen af s_0^2 frem for $\hat{\sigma}^2$ på lignende måde som i Enstikprøveproblemet i normalfordelingen, se side 17.

[2] Et *residual* betyder: noget der er til rest.

Hypotesen $\mu_1 = \mu_2$

For at vurdere om der er en signifikant forskel på de to gruppers middelværdier, testes den statistiske hypotese

$$H_0 \ : \ \mu_1 = \mu_2 \ .$$

Når hypotesen H_0 er rigtig, er der tale om et »enstikprøveproblem« med $n = n_1 + n_2$ observationer, så vi ved fra Kapitel 2 at

- den fælles værdi af middelværdiparameteren estimeres ved det totale gennemsnit

$$\overline{y} \ = \ \frac{1}{n}\left(\sum_{j=1}^{n_1} y_{1j} + \sum_{j=1}^{n_2} y_{2j}\right),$$

- maksimaliseringsestimatet over variansparameteren σ^2 er

$$\hat{\sigma}^2 \ = \ \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2,$$

- det variansestimat man som regel benytter, er

$$\begin{aligned} s_{01}^2 \ &= \ \frac{1}{n-1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 \\ &= \ \frac{1}{n-1}\left(\sum_{j=1}^{n_1}(y_{1j} - \overline{y})^2 + \sum_{j=1}^{n_2}(y_{2j} - \overline{y})^2\right), \end{aligned}$$

  med $n - 1$ frihedsgrader.

Kvotienttesstørrelsen for H_0 er

$$Q \ = \ \frac{L(\overline{y}, \overline{y}, \hat{\hat{\sigma}}^2)}{L(\overline{y}_1, \overline{y}_2, \hat{\sigma}^2)}$$

hvor L er defineret på side 31. Når man indsætter udtrykkene for estimaterne i Q, bliver det udtryk som exp skal anvendes på, simpelthen $-n/2$, både i tæller og nævner; udtrykket for Q kan derfor reduceres til

$$\begin{aligned} Q \ &= \ \left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right)^{-n/2} \\ &= \ \left(\frac{\sum_{j=1}^{n_1}(y_{1j} - \overline{y})^2 + \sum_{j=1}^{n_2}(y_{2j} - \overline{y})^2}{\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \overline{y}_2)^2}\right)^{-n/2} . \end{aligned}$$

Nævnerkvadratsummen er lig $(n-2)s_0^2$. Tællerkvadratsummen kan omskrives på følgende måde hvor vi undervejs benytter at $\overline{y} = (n_1\overline{y}_1 + n_2\overline{y}_2)/(n_1 + n_2)$:

$$\begin{aligned}
&\sum_{j=1}^{n_1}(y_{1j}-\overline{y})^2 + \sum_{j=1}^{n_2}(y_{2j}-\overline{y})^2 \\
&= \sum_{j=1}^{n_1}\left((y_{1j}-\overline{y}_1)+(\overline{y}_1-\overline{y})\right)^2 + \sum_{j=1}^{n_2}\left((y_{2j}-\overline{y}_2)+(\overline{y}_2-\overline{y})\right)^2 \\
&= \sum_{j=1}^{n_1}(y_{1j}-\overline{y}_1)^2 + n_1(\overline{y}_1-\overline{y})^2 + \sum_{j=1}^{n_2}(y_{2j}-\overline{y}_2)^2 + n_2(\overline{y}_2-\overline{y})^2 \\
&= (n-2)s_0^2 + n_1(\overline{y}_1-\overline{y})^2 + n_2(\overline{y}_2-\overline{y})^2 \\
&= (n-2)s_0^2 + n_1\left(\frac{n_2(\overline{y}_1-\overline{y}_2)}{n_1+n_2}\right)^2 + n_2\left(\frac{n_1(\overline{y}_1-\overline{y}_2)}{n_1+n_2}\right)^2 \\
&= (n-2)s_0^2 + \frac{(\overline{y}_1-\overline{y}_2)^2}{\frac{1}{n_1}+\frac{1}{n_2}}.
\end{aligned}$$

Med betegnelsen

$$t = \frac{\overline{y}_1-\overline{y}_2}{\sqrt{s_0^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

kan Q derfor udtrykkes som

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

Det ses at Q er en aftagende funktion af $|t|$, dvs. små Q-værdier er ensbetydende med store $|t|$-værdier, så man skal forkaste H_0 hvis $|t|$ er stor.

Man plejer at benytte t (*Student's t*) som teststørrelse fordi den har en umiddelbart forståelig fortolkning: den måler differensen mellem de to middelværdiestimater $(\overline{y}_1 - \overline{y}_2)$ i forhold til den estimerede standardafvigelse på denne differens.[3]

Testsandsynligheden, dvs. sandsynligheden for at få et sæt observationer der harmonerer *dårligere* med H_0 end de foreliggende observationer, bestemmes som[4]

$$\begin{aligned}
\varepsilon &= \mathrm{P}_0\big(|t| > |t_{\mathrm{obs}}|\big) \\
&= \mathrm{P}_0\big(t > |t_{\mathrm{obs}}| \text{ eller } t < -|t_{\mathrm{obs}}|\big) \\
&= 2\cdot\mathrm{P}_0\big(t > |t_{\mathrm{obs}}|\big),
\end{aligned}$$

[3] Det var et ræsonnement af denne art der førte Gosset (alias 'Student') til i 1908 at foreslå en teststørrelse der næsten er vore dages *Student's t*.

[4] Dette test er *tosidet* fordi de ekstreme t-værdier er på begge sider af 0, og det er det man som oftest bruger. Men en sjælden gang er man i en situation hvor man er aldeles sikker på at hvis ikke $\mu_1 = \mu_2$, så er (lad os sige) $\mu_1 < \mu_2$, den modsatte ulighed er utænkelig, og i så fald vil man kun forkaste H_0 hvis t er langt fra 0 *og negativ*. Man foretager da et *ensidet* test og udregner testsandsynligheden som $\mathrm{P}_0(t < t_{\mathrm{obs}})$.

hvor det sidste lighedstegn er en konsekvens af at t-fordelingen er symmetrisk om 0.

Hvis H_0 er rigtig, så følger t en såkaldt t-fordeling med $n-2$ frihedsgrader (frihedsgradsantallet arves fra variansskønnet s_0^2), og denne fordeling findes i statistiske tabeller. Hvis t_f betegner en stokastisk variabel som er t-fordelt med f frihedsgrader, så kan testsandsynligheden altså findes som

$$\varepsilon = 2\,\mathrm{P}\big(t_{n-2} > |t_{\mathrm{obs}}|\big)\,.$$

Test for varianshomogenitet

I det foregående er vi gået ud fra at observationerne i den ene gruppe har samme varians som observationerne i den anden gruppe. Denne antagelse kan man imidlertid godt teste. Det foregår på den måde at man opstiller den lidt generellere model der tillader varianserne at være forskellige, og i den model tester man så om varianserne kan antages at være ens.

Den lidt generellere model (generellere end på side 31) er

$$\begin{aligned} Y_{1j} &\sim \mathcal{N}(\mu_1, \sigma_1^2) \\ Y_{2j} &\sim \mathcal{N}(\mu_2, \sigma_2^2). \end{aligned}$$

Nu kan man opstille sine likelihoodfunktioner og estimere parametrene og teste hypotesen $H : \sigma_1^2 = \sigma_2^2$. Det viser sig at kvotienttesteststørrelsen er en funktion af

$$R = \frac{s_1^2}{s_2^2},$$

hvor $s_i^2 = \dfrac{1}{n_i - 1}\displaystyle\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$ er variansskønnet (med $n_i - 1$ frihedsgrader) i gruppe i, $i = 1, 2$.

Man plejer at benytte R som teststørrelse, og man skal forkaste hypotesen om ens varianser hvis R enten er meget større end 1 eller meget mindre end 1, dvs. der er tale om et tosidet test. Som testsandsynlighed benyttes sandsynligheden for at få en R-værdi der ligger uden for intervallet med endepunkter R_{obs} og $1/R_{\mathrm{obs}}$, så testsandsynligheden er

- hvis $R_{\mathrm{obs}} > 1$ så

$$\begin{aligned} \varepsilon &= \mathrm{P}_0\big(R > R_{\mathrm{obs}}\big) + \mathrm{P}_0\left(R < \frac{1}{R_{\mathrm{obs}}}\right) \\ &= \mathrm{P}_0\big(R > R_{\mathrm{obs}}\big) + \mathrm{P}_0\left(\frac{1}{R} > R_{\mathrm{obs}}\right), \end{aligned}$$

- hvis $R_{\mathrm{obs}} < 1$ så

$$\begin{aligned} \varepsilon &= \mathrm{P}_0\big(R < R_{\mathrm{obs}}\big) + \mathrm{P}_0\left(R > \frac{1}{R_{\mathrm{obs}}}\right) \\ &= \mathrm{P}_0\left(\frac{1}{R} > \frac{1}{R_{\mathrm{obs}}}\right) + \mathrm{P}_0\left(R > \frac{1}{R_{\mathrm{obs}}}\right). \end{aligned}$$

Der gælder at når hypotesen om varianshomogenitet er rigtig, så vil R følge den såkaldte F-fordeling med (f_1, f_2) frihedsgrader hvor f_1 og f_2 er antal frihedsgrader for s_1^2 og s_2^2. Da man har tabeller over fraktiler i F-fordelingen, er det let at bestemme testsandsynligheden ε. Hvis man yderligere udnytter en særlig egenskab ved F-fordelinger, nemlig at hvis R følger F_{f_1,f_2}-fordelingen, så vil $1/R$ følge F_{f_2,f_1}-fordelingen, så kan fremgangsmåden forsimples til

1. Lad $s^2_{\max}$ og $s^2_{\min}$ betegne henholdsvis det største og det mindste af tallene s_1^2 og s_2^2.
2. Lad $R^\star = s^2_{\max}/s^2_{\min}$.
3. Så er testsandsynligheden lig

   sandsynligheden for værdier større end $R^\star_{\text{obs}}$ i F-fordelingen med (f_1, f_2) frihedsgrader

   $+$ sandsynligheden for værdier større end $R^\star_{\text{obs}}$ i F-fordelingen med (f_2, f_1) frihedsgrader,

   altså

$$\varepsilon \;=\; \mathrm{P}\Big(F_{f_1,f_2} \geq R^\star_{\text{obs}}\Big) + \mathrm{P}\Big(F_{f_2,f_1} \geq R^\star_{\text{obs}}\Big)\,.$$

Et eksempel

Eksempel 3.1 (C-vitamin)
C-vitamin (ascorbinsyre) er et veldefineret kemisk stof som man sagtens kan fremstille i laboratoriet (og i industrien), og man kan jo i sin naivitet forestille sig at virkningen i den menneskelige organisme af det »kunstige« C-vitamin er præcis lige så god som virkningen af det i naturen forekommende. For at undersøge om det nu også forholder sig sådan har man foretaget et eksperiment, ikke med mennesker men med marsvin (små gnavere).

Man delte 20 nogenlunde ens marsvin op i to grupper hvoraf den ene fik appelsinsaft, og den anden fik en tilsvarende mængde »kunstigt« C-vitamin. Efter seks ugers behandling målte man længden af fortændernes odontoblaster (det tandbensdannende væv). Man fik da disse resultater (i hver gruppe er observationerne ordnet efter størrelse):

appelsinsaft:	8.2	9.4	9.6	9.7	10.0	14.5	15.2	16.1	17.6	21.5
kunstigt C-vitamin:	4.2	5.2	5.8	6.4	7.0	7.3	10.1	11.2	11.3	11.5

Man kan fastslå at der må være tale om en art tostikprøveproblem. Karakteren af observationerne gør at det ikke er urimeligt at forsøge sig med en normalfordelingsmodel af en slags, og det er alt i alt nærliggende at sige at der er tale om et »tostikprøveproblem med uparrede normalfordelte observationer«. Vi vil analysere observationerne ved brug af denne model, mere nøjagtigt vil vi undersøge om odontoblasternes middelvækst er den samme i de to grupper.

Tabel 3.1 er et regneskema der viser hvordan man kan foretage udregningerne med »håndkraft« (se også side 16f). Hvis man blot vil opsummere resultaterne, gør man det ofte i form af en tabel som den der er vist i Tabel 3.2.

Tabel 3.1 C-vitamin-eksemplet: regneskema.

	Appelsinsaft		Kunstigt C-vitamin	
	y	y^2	y	y^2
	8.2	67.24	4.2	17.64
	9.4	88.36	5.2	27.04
	9.6	92.16	5.8	33.64
	9.7	94.09	6.4	40.96
	10.0	100.00	7.0	49.00
	14.5	210.25	7.3	53.29
	15.2	231.04	10.1	102.01
	16.1	259.21	11.2	125.44
	17.6	309.76	11.3	127.69
	21.5	462.25	11.5	132.25
sum	131.8	1914.36	80.0	708.96
$\overline{y}_i$	$131.8/10 = 13.18$		$80.0/10 = 8.00$	
$\sum y^2 - \frac{(\sum y)^2}{n}$	$1914.36 - \frac{131.8^2}{10} = 177.236$		$709.96 - \frac{80.0^2}{10} = 68.960$	
s_i^2	$\frac{177.236}{10-1} = 19.69$		$\frac{68.960}{10-1} = 7.66$	
s_0^2	$\frac{177.236 + 68.960}{(10-1)+(10-1)} = 13.68$			
t	$\frac{13.18 - 8.00}{\sqrt{13.68\left(\frac{1}{10}+\frac{1}{10}\right)}} = 3.13$			

Da metoden til sammenligning af middelværdierne i de to grupper forudsætter at de to grupper har samme varians, kan man eventuelt også teste hypotesen om varianshomogenitet (se side 35). Testet er baseret på varianskvotienten

$$R = \frac{s^2_{\text{appelsinsaft}}}{s^2_{\text{kunstigt}}} = \frac{19.69}{7.66} = 2.57\,.$$

Denne værdi skal sammenholdes med F-fordelingen med $(9, 9)$ frihedsgrader i et tosidet test. Tabelopslag viser at 95%-fraktilen er 3.18 og 90%-fraktilen 2.44; der er derfor mellem 10 og 20 procents chance for at få en værre R-værdi selv om hypotesen er rigtig, og på dette grundlag vil vi ikke afvise antagelsen om varianshomogenitet. Den fælles varians estimeres til $s_0^2 = 13.68$ med 18 frihedsgrader.

Vi kan nu gå over til det egentlige, nemlig at teste om der er signifikant forskel på to gruppers niveauer. Til det formål udregnes t-teststørrelsen

$$t = \frac{13.18 - 8.00}{\sqrt{13.68\left(\frac{1}{10}+\frac{1}{10}\right)}} = \frac{5.18}{1.65} = 3.13\,.$$

Den fundne værdi skal sammenholdes med t-fordelingen med 18 frihedsgrader. I denne fordeling er 99.5%-fraktilen 2.878, hvoraf vi kan slutte at der er mindre end 1% chance

Tabel 3.2 C-vitamin-eksemplet: nogle beregnede størrelser.
n står for antal observationer y, S for Sum af y-er, $\overline{y}$ for gennemsnit af y-er, f for antal frihedsgrader, SS for Sum af kvadratiske afvigelser ('Sum of Squared deviations'), og s^2 for variansestimater (SS/f).

gruppe	n	S	$\overline{y}$	f	SS	s^2
appelsinsaft	10	131.8	13.18	9	177.236	19.69
kunstigt C-vit.	10	80.0	8.00	9	68.960	7.66
sum	20	211.8		18	246.196	
gennemsnit			10.59			13.68

for at få en værdi numerisk større end 3.13. Konklusionen bliver derfor at der er en klart signifikant forskel mellem de to grupper. Som det ses af tallene, består forskellen i at den »kunstige« gruppe har *mindre* odontoblastvækst end appelsingruppen. Kunstigt C-vitamin synes altså ikke at virke så godt som det naturlige.

3.2 Tostikprøveproblemet med parrede observationer

Som titlen på Afsnit 3.1 lader ane, er der også et tostikprøveproblem med *parrede* observationer. Situationen er her at observationerne hører sammen på to ledder: dels hører hver observation til en af to mulige grupper, dels hører observationerne sammen to og to, de er parrede. Typiske eksempler er målinger på nogle forsøgsdyr (eller -personer) af en bestemt variabel *før* og *efter* en behandling; de to grupper består da af henholdsvis målingerne før og målingerne efter, og observationerne er parrede idet man véd hvilke målinger der stammer fra hvilke individer.

Vi viser situationen skematisk:

	gruppe nr. 1	gruppe nr. 2
par nr. 1	y_{11}	y_{12}
par nr. 2	y_{21}	y_{22}
$\vdots$	$\vdots$	$\vdots$
par nr. i	y_{i1}	y_{i2}
$\vdots$	$\vdots$	$\vdots$
par nr. r	y_{r1}	y_{r2}

Der er r observationspar, og det i-te par består af y_{i1} og y_{i2}.

Ved opbygningen af en statistisk model bør man naturligvis udnytte den information der ligger i at vi véd hvilke observationer der hører sammen. Man kunne forestille sig at det forholdt sig på den enkle måde at forskellen mellem den »sande« værdi af en gruppe 2-måling og den »sande« værdi af den tilsvarende gruppe 1-måling havde den samme værdi δ for alle parrene. Der er

Tabel 3.3 Antal ekstra søvntimer ved behandling med hyoscyamin hydrobromid.

person	dextro-	laevo-
1	0.7	1.9
2	–1.6	0.8
3	–0.2	1.1
4	–1.2	0.1
5	–0.1	–0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

altså ikke noget i vejen for at de enkelte par kan være voldsomt forskellige, blot *forskellen* mellem de to medlemmer af et par er den samme (pånær tilfældige afvigelser) for alle par.

Hvis det forholder sig på denne måde, er der en uhyre simpel måde at analysere tallene på: man udregner differenserne $d_i = y_{i2} - y_{i1}$ og undersøger om de fordeler sig tilfældigt omkring 0. Hvis man er parat til at antage at differenserne $d_1, d_2, \ldots, d_n$ er observationer fra en normalfordeling med middelværdi δ og varians σ^2, så er vi tilbage ved et *enstikprøveproblem* i normalfordelingen, og så er det bare at slå tilbage til Kapitel 2.

Eksempel 3.2 (Sovemidler)
Det kemiske stof *hyoscyamin hydrobromid* kan anvendes som sovemiddel. Stoffet findes imidlertid i to udgaver, d-hyoscyamin hydrobromid og l-hyoscyamin hydrobromid,[5] og man er interesseret i at finde ud af om de to udgaver er lige gode. Derfor har man udført en forsøgsrække hvor man på 10 forsøgspersoner har bestemt stoffernes søvnforlængende virkning. I Tabel 3.3 er vist det gennemsnitlige antal ekstra søvntimer pr. nat for hver person, dels ved behandling med d-udgaven, dels ved behandling med l-udgaven af stoffet.

Da der er tale om at man på nogle forsøgspersoner har målt effekten af først en, så en anden behandling, vil det være nærliggende at søge at analysere talmaterialet ved hjælp af en model af typen »tostikprøveproblem med parrede observationer«. Derfor bestemmes differenserne mellem virkningerne af laevo- og dextroudgaven af stoffet, se Tabel 3.4.

Vi vil opfatte tallene i Tabel 3.4 som et »enstikprøveproblem i normalfordelingen«, og spørgsmålet om de to stoffer virker lige godt kan da præciseres til spørgsmålet om tallenes middelværdi er signifikant forskellig fra 0. Dette kan testes som en statistisk hypotese.

Gennemsnittet af differenserne i tabellen er $\overline{d} = 1.58$ timer, og estimatet over variansen på differenserne er $s^2 = 1.51$ timer^2 (med 9 frihedsgrader), svarende til at den estimerede standardafvigelse er $s = 1.23$ timer. Den estimerede standardafvigelse på gennemsnittet er dermed $\sqrt{s^2/n} = \sqrt{1.51/10}$ timer $= 0.39$ timer. Endvidere bliver

[5] l = laevo = venstre, d = dextro = højre (angiver til hvilken side stoffet afbøjer polariseret lys).

Tabel 3.4 Differenser mellem l- og d-hyoscyamin hydrobromids søvnforlængende virkning.

person	differens (timer)
1	1.2
2	2.4
3	1.3
4	1.3
5	0.0
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

t-teststørrelsen

$$t = \frac{\overline{d} - 0}{\sqrt{s^2/n}} = \frac{1.58 \text{ timer}}{0.39 \text{ timer}} = 4.06 .$$

I t-fordelingen med 9 frihedsgrader er 99.5%-fraktilen 3.25 og 99.9%-fraktilen 4.29, så testsandsynligheden ligger et sted mellem 0.2% og 1%. Der er således ganske klart signifikans, dvs. de to stoffer virker signifikant forskelligt (og som man ser er l-stoffet det mest virksomme).

Dette var så et eksempel på et tostikprøveproblem med parrede observationer, men hvad var der sket hvis man af vanvare var kommet til at analysere det som om der var tale om uparrede observationer?

Den t-størrelse man så ville udregne, var en anden. Tælleren ville være den samme fordi differensen mellem gennemsnittene er lig gennemsnittet af differenserne. Det variansestimat der skulle benyttes i nævneren, er estimatet over den fælles varians i de to grupper, og det udregnes til $s_0^2 = 3.605 \text{ timer}^2$ med 18 frihedsgrader, og teststørrelsen ville derfor blive

$$t = \frac{1.58 \text{ timer}}{\sqrt{3.605 \text{ timer}^2 \left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{1.58}{0.85} = 1.86 .$$

Denne gang ville vi få 18 frihedsgrader i t-fordelingen, og det vil sige at 95%-fraktilen er 1.73 og 97.5%-fraktilen 2.10. Der ville altså være et sted mellem 5% og 10% chance for at få en mere ekstrem t-størrelse end 1.86, og man vil derfor almindeligvis sige at $t_{\text{obs}} = 1.86$ *ikke* er signifikant stor. Dette test ville således ikke vise nogen signifikant forskel på de to stoffer.

Grunden til at de to analyser giver forskellige resultater, er at der er en temmelig stor forskel på forsøgspersonerne:

- I den først benyttede model (parrede observationer) elimineres en stor del af personforskellene ved at man går over til at analyserer differenserne. Til gengæld får variansestimatet kun 9 frihedsgrader.
- I den anden model (uparrede observationer) skal al forskellen mellem personer beskrives af variansparameteren (fordi forskellen mellem personer i denne omgang udelukkende anses for tilfældig), og til gengæld får variansestimatet hele 18 frihedsgrader. – På den anden side indebærer det at hvis der *er* stor forskel mellem personer, så bliver variansestimatet også stort.

Datamaterialet til dette eksempel er meget berømt fordi det blev benyttet til et illustrativt eksempel i dén artikel hvor *t*-testet (i enstikprøveproblemet) blev introduceret ('Student' (1908): The Probable Error of a Mean. *Biometrika* **6**, 1-25). Artiklen er skrevet af W.S.Gosset der arbejdede som biometriker ved Guinnessbryggerierne og benyttede 'Student' som sit *nom de plume*.

Tabel 3.5 Opgave 3.1: Varmemængde (i calorier) for at smelte 1 g is med en begyndelsestemperatur på $-0.72\,℃$, bestemt ved to forskellige metoder.

Metode A	Metode B
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

Tabel 3.6 Opgave 3.2: Den maksimale procentdel af blodpladerne der klumper sig sammen efter en given påvirkning.

før	efter
25	27
25	29
27	37
44	56
30	46
67	82
53	57
53	80
52	61
60	59
28	43

3.3 Opgaver

Opgave 3.1 (Is's smeltevarme)
Man ønsker at sammenligne to forskellige metoder (A og B) til bestemmelse af is's smeltevarme. Eksperimenter har givet resultaterne i Tabel 3.5. Undersøg om der er signifikant forskel på de to metoder.
TIP: Udregningerne bliver lettere hvis man indfører et passende beregningsnulpunkt.

Opgave 3.2 (Rygning og blodpropper)
På 11 forsøgspersoner har man taget blodprøver før og efter de røg en cigaret, og man har så undersøgt blodpladernes tendens til at klumpe sig sammen (sådanne klumper kan udvikle sig til regulære blodpropper). Resultaterne ses i Tabel 3.6.

Undersøg om resultaterne tyder på at rygning påvirker blodpladernes tendens til at klumpe sig sammen. (Der er øjensynligt tale om et tostikprøveproblem af en slags; der kan så være tale om *parrede* eller *uparrede* observationer. Det kan være illustrativt at forsøge sig med begge slags modeller. Hvad er forskellen? Argumentér for at den ene af dem er mere rigtig end den anden.)

Opgave 3.3
Man har foretaget nogle forsøg med mus for at finde ud af om de to forskellige former for jernioner Fe^{2+} og Fe^{3+} optages med forskellig hastighed i organismen. Dette er af betydning når man skal sammensætte kosttilskud (eksempelvis vitaminpiller) til mennesker.

Som led i et større forsøg har man givet 18 mus Fe^{2+} og 18 andre mus Fe^{3+}, i begge tilfælde i 1.2 millimolar opløsninger indgivet oralt. Jernatomerne var radioaktivt mærkede således at det var muligt at måle hvor meget jern der

Tabel 3.7 Opgave 3.3: Procentdel optaget jern samt titalslogaritmen til procentdel optaget jern, for 18 mus der har fået Fe^{2+} og 18 mus der har fået Fe^{3+}. – De enkelte søjler indeholder de *ordnede* observationer.

	y		$\log_{10} y$	
	Fe^{2+}	Fe^{3+}	Fe^{2+}	Fe^{3+}
	2.20	4.04	0.342	0.606
	2.93	4.16	0.467	0.619
	3.08	4.42	0.489	0.645
	3.49	4.93	0.543	0.693
	4.11	5.49	0.614	0.740
	4.95	5.77	0.695	0.761
	5.16	5.86	0.713	0.768
	5.54	6.28	0.744	0.798
	5.68	6.97	0.754	0.843
	6.25	7.06	0.796	0.849
	7.25	7.78	0.860	0.891
	7.90	9.23	0.898	0.965
	8.85	9.34	0.947	0.970
	11.96	9.91	1.078	0.996
	15.54	13.46	1.191	1.129
	15.89	18.40	1.201	1.265
	18.30	23.89	1.262	1.378
	18.59	26.39	1.269	1.421
sum	147.67	173.38	14.862	16.339
sum af kvadrater	1715.9265	2431.1648	13.662209	15.885527

blev optaget i musen i løbet af et fastsat stykke tid. Tabel 3.7 viser hvor stor en procentdel af den tilførte mængde jern der blev optaget af musen.

1. Ved data af denne type kan man erfaringsmæssigt ofte beskrive *logaritmen* til observationerne med en normalfordeling. Undersøg om det er rimeligt at gøre det i dette tilfælde.

2. Undersøg om data tyder på at Fe^{2+} og Fe^{3+} optages på samme måde (sammenlign for eksempel de to stikprøver af logaritmerede målinger).

3. Man vil planlægge et nyt forsøg af samme slags, blot med et andet antal mus. Det nye forsøg skal kunne afgøre om der er en reel forskel på 0.1 (på den logaritmiske skala) mellem Fe^{2+}- og Fe^{3+}-optagelsen. I den forbindelse kan man vælge at sige at »en reel forskel på 0.1« skal betyde at hvis tælleren i t-teststørrelsen, altså differensen mellem middeltallene, er større end eller lig 0.1 (eller mindre end eller lig -0.1), så vil testsandsynligheden blive mindre end eller lig 5% (»der er signifikans på niveau 5%«).

   Spørgsmålet er hvor mange mus der skal benyttes: Omsæt ovenstående præcisering af »en reel forskel på 0.1« til matematik, og få derved en ulighed der kan løses med hensyn til den ubekendte »antal mus«.

Det således planlagte forsøg skulle angiveligt kunne afgøre om der er en reel forskel på 0.1. Diskutér hvilken status man skal tillægge en sådan »afgørelse«.

4 Ensidet variansanalyse

Sammenligning af *to* normalfordelte stikprøver er omtalt i Kapitel 3. Man kommer dog ofte ud for at skulle sammenligne mere end to stikprøver, og derfor er man nødt til også at have metoder til det såkaldte k-stikprøveproblem, dvs. den situation hvor der foreligger k grupper af normalfordelte observationer, og hvor man ønsker at vurdere om der er en signifikant forskel på disse k grupper (se side 29 for en generel formulering af problemet). Den metode der benyttes for at sammenligne *middelværdierne* i k grupper af normalfordelte observationer, kaldes (måske lidt overraskende) for *ensidet variansanalyse*.

Eksempel 4.1 (Dækningsgrad for Fuglegræs)
På dyrkede marker er ukrudt jo pr. definition en uting, og landmanden kan overveje om han skal sprøjte mod den slags ukrudt han anser for værst. Men når man fjerner én slags ukrudt, kan det være at det ikke bare er afgrøden der derved får forbedrede vækstforhold, men også de resterende ukrudtsarter! Måske er det en ligefrem fordel at have så mange forskellige ukrudtsarter som muligt fordi de så kan holde hinanden i skak.

For at undersøge ukrudtsplanters indbyrdes konkurrence på en kornmark har man udført et større forsøg der består i at på forskellige dele af en stor mark luger man på et bestemt tidspunkt forskellige ukrudtsarter bort, og derefter ser man hvorledes resten af arterne så trives.[1] Mere præcist er marken delt op i 16 jordlodder som er delt ind i fire grupper med hver fire lodder. Den første gruppe er en kontrolgruppe hvor intet luges bort, men i hver af grupperne to, tre og fire luges én bestemt ukrudtsart bort (henholdsvis Snerle pileurt, Fuglegræs og Hvidmelet gåsefod). Én gang før og tre gange efter bortlugningen registrerer man hvilke planter der er på de forskellige lodder og i hvor stor udstrækning. Den første registrering skal tjene til at fastlægge det niveau som den senere udvikling skal måles ud fra.

De fire grupper er fordelt på marken i et romersk kvadrat:

3	4	1	2
2	1	4	3
4	3	2	1
1	2	3	4

De fire lodder der udgør en gruppe, er altså placeret fire helt forskellige steder på marken; derved har man en chance for at kunne tage højde for eventuelle variationer i jordbund og mikroklima henover marken.

Forsøget har givet et stort talmateriale som kan analyseres på mange måder. Her skal vi kun se på en enkelt detalje i forbindelse med fastlæggelsen af et udgangsniveau på grundlag af den første registrering. Vi vil studere forekomsten af Fuglegræs, *Stellaria media*, ved den første registrering, se Tabel 4.1. Registreringen foregår ved hjælp af et rektangulært gitternet med 416 gitterpunkter med fem centimeters afstand; gitternettet

[1] A. Greenfort, C.S.F. Jensen & S. Jeppesen (1987): *Planter og planter imellem*. Biologispeciale, RUC.

Tabel 4.1 Dækningsgrader for Fuglegræs ved første registrering.

gruppe	dækningsgrader			
1	17	38	23	26
2	19	16	16	14
3	25	33	29	33
4	27	16	30	20

placeres på jordlodden hvorefter man i hvert gitterpunkt ser efter om der findes noget af en Fuglegræs-plante eller ej. Som mål for *dækningsgraden* for arten benyttes antallet af gitterpunkter hvor arten blev registreret. Dækningsgraden bliver på denne måde et helt tal mellem 0 og 416.

Da den første registrering udførtes inden der blev foretaget nogen bortlugning, kan der ikke på dette tidspunkt være tale om nogen behandlingseffekt (lugningseffekt). De forskelle der er på lodderne og på grupperne, må alene skyldes »startbetingelserne«, dvs. de lokale variationer i jordbund og klima og de forskellige antal planter af den pågældende art som der nu tilfældigvis var på de enkelte områder af marken. Da man ønsker at vurdere hvordan behandlingerne påvirker grupperne, kan det være af interesse at få en idé om hvor forskellige (eller hvor ens) grupperne egentlig er ved forsøgets start. Hvis grupperne nemlig er stort set ens, kan man bestemme et fælles startniveau hvorudfra den senere udvikling kan vurderes, men hvis der er en signifikant forskel mellem grupperne, så er man nødt til at vurdere hver gruppes udvikling ud fra dens eget startniveau. Derfor vil vi gerne sammenligne de fire grupper og vurdere om forskellen mellem grupperne er stor i forhold til den tilfældige variation inden for grupperne.

Den statistiske model: Da observationerne er fremkommet som en sum af et vist antal 01-størrelser svarende til om planten er fraværende eller til stede i det pågældende gitterpunkt, kunne man mene at det smager lidt af en binomialfordelingssituation (eller eventuelt en Poissonfordelingssituation da n er temmelig stor). Hertil kan man indvende at ikke alle binomialfordelingsbetingelserne er opfyldt idet de enkelte 01-størrelser næppe er uafhængige med samme sandsynlighed for »1«, og det kan medføre en større tilfældig variation inden for de enkelte grupper end hvad binomialfordelingen kan forklare. Man kan derfor idet man går let hen over at der er tale om diskrete observationer, forsøge sig med en normalfordelingsmodel, hvor man jo ved hjælp af variansparameteren kan modellere den tilfældige variation særskilt. Vi vil benytte en statistisk model der går ud på at observationer i samme gruppe opfattes som observationer fra en og samme normalfordeling, og at de fire grupper har hver deres normalfordeling. Det statistiske problem er da at undersøge om de fire normalfordelinger kan tænkes at være ens.

Det generelle k-stikprøveproblem i normalfordelingen kan formuleres på følgende måde:

Der foreligger nogle observationer y som er ordnet i k grupper med n_i observationer i gruppe nr. i, $i = 1, 2, \ldots, k$; observation nr. j fra gruppe nr. i betegnes y_{ij}. Skematisk ser det sådan ud:

gruppe	observationer					
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in_i}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kn_k}

Vi går ud fra at forskellen mellem observationerne inden for en gruppe er tilfældig, hvorimod der er en systematisk forskel mellem grupperne. Vi går endvidere ud fra at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij}. Den tilfældige variation vil vi beskrive ved hjælp af en normalfordeling, og det skal derfor alt i alt være sådan at Y_{ij} *er normalfordelt med middelværdi* μ_i *og varians* σ^2, kort

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2). \tag{4.1}$$

Herved beskriver middelværdiparametrene $\mu_1, \mu_2, \ldots, \mu_k$ *den systematiske variation*, nemlig de enkelte gruppers niveauer, medens variansparameteren σ^2 (samt normalfordelingen) beskriver *den tilfældige variation* inden for grupperne. Den tilfældige variation antages at være den samme i alle grupperne, og denne antagelse kan man undertiden teste, se Afsnit 4.3.

4.1 Estimation af parametrene

Estimation af $\mu_1, \mu_2, \ldots, \mu_k$

De ukendte middelværdiparametre $\mu_1, \mu_2, \ldots, \mu_k$ i grundmodellen (4.1) estimeres ved hjælp af maximum likelihood metoden, altså som de værdier der maksimaliserer likelihoodfunktionen

$$\begin{aligned} & L(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) \\ &= \prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2} \frac{(y_{ij} - \mu_i)^2}{\sigma^2} \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \right) \end{aligned} \tag{4.2}$$

hvor $n = n_1 + n_2 + \ldots + n_k$ er det samlede antal observationer. Det ses at hvis σ^2 er fast, så er det at *maksimalisere* likelihoodfunktionen L med hensyn til

$\mu_1, \mu_2, \ldots, \mu_k$ det samme som det at *minimalisere* kvadratsummen

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2,$$

og den opgave er let at løse:

Vi lader $\overline{y}_i$ betegne gennemsnittet i gruppe i, $\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$. Ved at benytte formlen for kvadratet på en toleddet størrelse fås

$$\begin{aligned} (y_{ij} - \mu_i)^2 &= \big((y_{ij} - \overline{y}_i) + (\overline{y}_i - \mu_i)\big)^2 \\ &= (y_{ij} - \overline{y}_i)^2 + 2(y_{ij} - \overline{y}_i)(\overline{y}_i - \mu_i) + (\overline{y}_i - \mu_i)^2 \, ; \end{aligned}$$

når vi her holder i fast og summerer over j, så bliver summen af de dobbelte produkter 0 fordi $\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)$ er lig med 0 ifølge definitionen af $\overline{y}_i$; hvis vi endelig også summerer over i, får vi

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{y}_i - \mu_i)^2 \, . \tag{4.3}$$

Opgaven er at minimalisere venstresiden; men de μ-er der minimaliserer venstresiden, er de samme som dem der minimaliserer den anden kvadratsum på højresiden, og den bliver mindst mulig, nemlig 0, netop når μ_i er lig $\overline{y}_i$, $i = 1, 2, \ldots, k$. Vi har dermed fundet at maksimaliseringsestimatet for den i-te gruppes middelværdi er lig med gennemsnittet af observationerne i gruppen, $\hat{\mu}_i = \overline{y}_i$.

Estimation af σ^2

Maksimaliseringsestimatet $\hat{\sigma}^2$ for σ^2 kan bestemmes som maksimumspunktet for funktionen

$$\sigma^2 \mapsto L(\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k, \sigma^2).$$

Man finder at

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2.$$

En størrelse som $y_{ij} - \overline{y}_i$ der er forskellen mellem den faktiske observation og det bedst mulige *fit* under den aktuelle model, kaldes for et *residual*, og $\hat{\sigma}^2$ kan derfor beskrives som værende residualkvadratsummen divideret med antallet af observationer. Som regel benytter man imidlertid et andet estimat over σ^2, nemlig residualkvadratsummen divideret med *antallet af frihedsgrader* $n - k$ (antal observationer minus antal estimerede parametre), dvs. man benytter variansestimatet

$$s_0^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2.$$

Tabel 4.2 Fuglegræseksemplet: nogle beregnede størrelser.

i	n_i	$y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$	$\overline{y}_i$	$\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$
1	4	104	26.00	234.00
2	4	65	16.25	12.75
3	4	120	30.00	44.00
4	4	93	23.25	122.75
sum	16	382		413.50
gennemsnit			23.88	

$$s_0^2 = \frac{1}{16-4}\,413.50 = 34.46$$

Man begrunder brugen af s_0^2 frem for $\hat{\sigma}^2$ på lignende måde som i Enstikprøveproblemet i normalfordelingen, se side 17.

Sammenfattende har vi altså at

- middelværdiparameteren μ_i i den i-te gruppe estimeres ved gennemsnittet $\overline{y}_i$ af observationerne i gruppen,
- gruppernes fælles varians σ^2 estimeres ved residualkvadratsummen divideret med antallet af frihedsgrader,

$$s_0^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 \tag{4.4}$$

  med $n - k$ frihedsgrader.

I Tabel 4.2 er vist de værdier man finder i Fuglegræs-eksemplet.

4.2 Hypotesen om ens grupper

I dette afsnit skal vi beskæftige os med spørgsmålet om hvordan man undersøger om de k grupper kan antages at have samme middelværdi. Opgaven er således at teste hypotesen H_0 om at der ikke er nogen signifikant forskel mellem grupperne, også kaldet hypotesen om *homogenitet mellem grupper*:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \,.$$

Ofte er det ikke H_0 man er interesseret i, men dens negation: at der *er* en signifikant forskel mellem grupperne, fordi man har et ønske eller et håb om at kunne vise at grupperne *ikke* er ens. Når det alligevel er H_0 man tester og ikke dens negation, så hænger det sammen med to generelle træk ved formulering og test af statistiske hypoteser:

1. De hypoteser man kan teste, er altid hypoteser der består i en *forsimpling* af den aktuelle grundmodel – typisk tester man at nogle parametre er ens, mens grundmodellen tillader dem at være forskellige.

2. Det er informativt at *forkaste* en hypotese: Vi får at vide at der er en signifikant uoverensstemmelse mellem hypotese og observationer.

   Derimod viser det ofte ingenting at få accepteret en hypotese: Det kan være at man simpelt hen bare har for få observationer til at kunne afsløre noget som helst.

Vi skal nu se hvordan man tester hypotesen H_0 om ens middelværdier. Man kan gå frem efter den sædvanlige opskrift, dvs. opstille en kvotienttestørrelse der sammenligner likelihoodfunktionens maksimale værdier hhv. under H_0 og under grundmodellen. Vi ved fra side 49 at likelihoodfunktionen maksimaliseres af værdierne $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k, \hat{\sigma}^2$ hvor $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$. Dernæst skal vi finde ud af hvilke værdier der maksimaliserer likelihoodfunktionen under H_0:

Når H_0 er rigtig, er der tale om et enstikprøveproblem, og fra Kapitel 2 ved vi at

- den fælles middelværdi μ estimeres ved det totale gennemsnit

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij},$$

- maksimaliseringsestimatet over den fælles varians σ^2 er kvadratafvigelsessummen omkring $\overline{y}$ divideret med n, dvs.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2,$$

- det variansskøn man som regel bruger, er

$$s_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$$

  med $n-1$ frihedsgrader.

Kvotienttestørrelsen for H_0 er

$$Q = \frac{L(\overline{y}, \overline{y}, \ldots, \overline{y}, \hat{\sigma}^2)}{L(\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k, \hat{\sigma}^2)},$$

hvor L er defineret på side 47. Når man indsætter udtrykkene for estimaterne i Q, så bliver det udtryk som exp skal anvendes på ganske enkelt $-n/2$, både i

tæller og nævner, så udtrykket for Q kan reduceres til

$$\begin{aligned} Q &= \left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right)^{-n/2} \\ &= \left(\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2}\right)^{-n/2}. \end{aligned}$$

For at kunne omforme Q yderligere skal vi bruge følgende omskrivning der fås af formel (4.3) på side 48 hvis man erstatter μ_i med $\overline{y}$:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i-\overline{y})^2, \tag{4.5}$$

dvs. den totale kvadratsum der beskriver y_{ij}-ernes variation om det totale gennemsnit $\overline{y}$, spaltes op i en sum af et bidrag der beskriver »variationen inden for grupperne« og et bidrag der beskriver »variationen mellem grupperne«.

Parallelt med opspaltningen af kvadratsummen har vi opspaltningen

$$n-1 = (n-k)+(k-1)$$

af frihedsgraderne, og ved at dividere kvadratsummerne med de tilsvarende antal frihedsgrader får vi variansestimater der beskriver forskellige variationer:

- *Variationen omkring totalgennemsnittet* (dvs. enkeltobservationernes variation omkring totalgennemsnittet) beskrives af

  $$s_{01}^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2$$

  som er variansestimatet under H_0.

- *Variationen inden for grupper* (dvs. enkeltobservationernes variation omkring deres respektive gruppegennemsnit) beskrives af

  $$s_0^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2$$

  som er variansestimatet i grundmodellen (formel (4.4) på side 49).

- *Variationen mellem grupper* (dvs. gruppegennemsnittenes variation omkring det totale gennemsnit) beskrives af

  $$\begin{aligned} s_1^2 &= \frac{1}{k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i-\overline{y})^2 \\ &= \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{y}_i-\overline{y})^2. \end{aligned}$$

Men vi skal videre med omskrivningen af udtrykket for Q. Ved hjælp af formel (4.5) kan vi omskrive Q til

$$\begin{aligned} Q &= \left(1 + \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{y}_i - \overline{y})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2} \right)^{-n/2} \\ &= \left(1 + \frac{(k-1)s_1^2}{(n-k)s_0^2} \right)^{-n/2}, \end{aligned}$$

hvilket viser at Q er en monotont aftagende funktion af størrelsen

$$F = \frac{s_1^2}{s_0^2},$$

således at store værdier af F svarer til små værdier af Q og dermed er tegn på at H_0 bør forkastes.

Som teststørrelse for H_0 benytter man i praksis altid F. Man kan forstå F som *forholdet mellem variationen mellem grupper og variationen inden for grupper*.

Man forkaster hypotesen om homogenitet mellem grupper når variationen *mellem* grupper er væsentligt større end variationen *inden for* grupper. Man kan bevise at F-teststørrelsen følger den såkaldte F-fordeling med frihedsgrader $(k-1, n-k)$, forudsat at hypotesen H_0 er rigtig. Derfor kan testsandsynligheden $\varepsilon = \mathrm{P}_0(F > F_{\mathrm{obs}})$ bestemmes som

$$\varepsilon = \mathrm{P}(F_{k-1,n-k} > F_{\mathrm{obs}})$$

der let findes ved hjælp af en tabel over fraktiler i F-fordelingen.

Vi har hermed løst den opgave der gik ud på at sammenligne k grupper af normalfordelte observationer. Man kan sige at F-teststørrelsen sammenligner to variansestimater, og derfor kaldes analysemetoden for en *variansanalyse*; da observationerne er inddelt efter ét kriterium (nemlig hvilken gruppe de tilhører), kaldes analysen for *ensidet variansanalyse*. Det er kutyme at give en oversigt over en variansanalyse i et såkaldt *variansanalyseskema*. Tabel 4.3 er et variansanalyseskema for Fuglegræs-eksemplet.

Eksempel 4.2 (Fuglegræs, konklusion)
Tabel 4.3 viser variansanalyseskemaet for ensidet variansanalyse i Fuglegræs-eksemplet. Det ses at F-teststørrelsen bliver 3.9, og denne værdi skal sammenholdes med fraktilerne i F-fordelingen med frihedsgrader 3 og 12; i denne fordeling er 95%-fraktilen 3.49 og 97.5%-fraktilen 4.47, så testsandsynligheden er knap 4%. På den baggrund vil man sædvanligvis være stemt for at *forkaste* hypotesen om ens middelværdier i grupperne. Man må altså konstatere at de fire grupper synes at være forskellige allerede inden man begynder at give dem hver deres behandling. Det kan virke overraskende, men det må hænge sammen med at der på forhånd er betydelige forskelle på de enkelte dele af marken. Når man sidenhen skal undersøge hvordan behandlingerne virker, er man nødt til at tage hensyn til denne forskellighed.

Tabel 4.3 Fuglegræs-eksemplet: *Variansanalyseskema*.
f står for antal frihedsgrader, SS for Sum af kvadratiske afvigelser,
$s^2 = SS/f$.

variation	f	SS	s^2	test
inden for grupper	12	413.50	34.46	
mellem grupper	3	402.25	134.08	134.08/34.46=3.9
total	15	815.75	54.38	

4.3 Bartletts test for varianshomogenitet

I normalfordelingsmodeller er det en forudsætning for en meningsfuld sammenligning af forskellige gruppers middelværdiparametre at grupperne har samme varians.[2] I dette afsnit skal vi omtale et test der kan anvendes når man ønsker at vurdere om et antal grupper af normalfordelte observationer kan antages at have samme varians, det vil sige om der er *varianshomogenitet*. Testet kan ikke benyttes hvis en af grupperne kun indeholder en enkelt observation, og for at man skal kunne anvende den tilnærmede fordeling af teststørrelsen, skal hver gruppe indeholde mindst seks observationer, eller rettere: i hver enkelt gruppe skal variansestimatet have mindst fem frihedsgrader.

Den generelle situation er stadig den der blev præsenteret på side 47, og vi ønsker i denne omgang at teste antagelsen om at grupperne har samme variansparameter σ^2. Den måde man kan gribe et sådant problem an på, er at man indlejrer den statistiske model i en større model, og så tester man på helt sædvanlig vis om man kan reducere den store model til den oprindelige model. I det aktuelle tilfælde indlejrer vi den oprindelige model (4.1) fra side 47 i en større model der tillader grupperne at have hver deres egen varians, nemlig modellen

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

Dernæst tester vi (4.1) som en hypotese i forhold til den nye grundmodel.

Den hypotese der skal testes, handler kun om en del af modellens parametre, og for så at sige at slippe af med de parametre der ikke har noget med hypotesen at gøre (altså med μ_i-erne), kan man teste hypotesen i den betingede fordeling givet de estimerede middelværdiparametre.[3] Hvis man omskriver kvotienteststørrelsen i den nævnte betingede fordeling, når man frem til at man kan benytte følgende størrelse (*Bartletts teststørrelse*) som teststørrelse for hypotesen om varianshomogenitet:

$$B = -\sum_{i=1}^{k} f_i \ln \frac{s_i^2}{s_0^2}\,; \tag{4.6}$$

[2] Man kan eventuelt klare sig med en antagelse om at gruppernes varianser er af formen: en ukendt fælles parameter ganget med en kendt konstant (der kan afhænge af gruppen).

[3] Dette hænger sammen med at man måske også bør *estimere* variansparametrene i denne betingede fordeling, se side 17.

her betegner s_i^2 estimatet over variansen σ_i^2 i den i-te gruppe, og f_i er antallet af frihedsgrader for s_i^2, dvs.

$$\begin{aligned} s_i^2 &= \frac{1}{f_i}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2, \\ f_i &= n_i - 1, \end{aligned}$$

og s_0^2 er det sædvanlige estimat over den fælles varians σ^2 (formel (4.4) på side 49). Bemærk i øvrigt at s_0^2 er et vægtet gennemsnit af s_i^2-erne med frihedsgraderne som vægte, $s_0^2 = \frac{1}{f}\sum_{i=1}^{k} f_i s_i^2$, hvor $f = f_1 + f_2 + \ldots + f_k$ er antallet af frihedsgrader for s_0^2.

Teststørrelsen B (som i virkeligheden er en $-2\ln Q$-størrelse) er altid et positivt tal, og store værdier af B er signifikante, dvs. tyder på at hypotesen om varianshomogenitet er forkert. Hvis hypotesen er rigtig, er B nogenlunde χ^2-fordelt med $k-1$ frihedsgrader, således at det er let at bestemme den omtrentlige testsandsynlighed som

$$\varepsilon = \mathrm{P}\left(\chi^2_{k-1} \geq B_{\mathrm{obs}}\right).$$

Denne χ^2-approksimation er god når alle f_i-erne er store; som tommelfingerregel siger man at de alle skal være mindst 5.

Hvis der kun er to grupper (dvs. $k = 2$), kan man alternativt teste hypotesen om varianshomogenitet med et test baseret på forholdet mellem de to variansestimater; dette er omtalt i forbindelse med tostikprøveproblemet i normalfordelingen, se side 35. (Dette tostikprøvetest er ikke baseret på nogen χ^2-approksimationer, så det har ingen restriktioner på antallene af frihedsgrader.)

Eksempel 4.3 (Fuglegræs: test for varianshomogenitet)
Som illustration udregnes Bartletts teststørrelse i Fuglegræs-eksemplet. Vi udvider det tidligere regneskema i Tabel 4.2 og får Tabel 4.4. Derefter kan vi udregne B_{obs}:

$$\begin{aligned} B_{\mathrm{obs}} &= -\left(3\ln\frac{78.00}{34.46} + 3\ln\frac{4.25}{34.46} + 3\ln\frac{14.67}{34.46} + 3\ln\frac{40.92}{34.46}\right) \\ &= 5.87\,. \end{aligned}$$

Betingelsen om at alle f_i-erne skal være mindst fem er ikke opfyldt (idet de alle er tre), så det er begrænset hvor χ^2-fordelt B kan forventes at være; men hvis vi ser lidt stort på det, så skulle B altså være ca. χ^2-fordelt med $4-1=3$ frihedsgrader når hypotesen om varianshomogenitet er rigtig. I χ^2_3-fordelingen er 80%-fraktilen 4.64 og 90%-fraktilen 6.25, således at under forudsætning af at hypotesen er rigtig, er der i størrelsesordenen 10% sandsynlighed for at få en værre B-værdi end den opnåede; på dette grundlag kan vi ikke forkaste hypotesen om varianshomogenitet.

4.4 Opgaver

Opgave 4.1 (Sammenligning af gødskningsmetoder)
I et dyrkningsforsøg vil man undersøge hvordan to gødskningsmetoder virker. Man har dyrket 10 planter med den ene metode, 10 med den anden, og 10

Tabel 4.4 Fuglegræseksemplet: nogle beregnede størrelser.
n står for antal observationer y, S for Sum af y-er, $\overline{y}$ for gennemsnit af y-er, f for antal frihedsgrader, SS for Sum af kvadratiske afvigelser ('Sum of Squared deviations'), og s^2 for variansestimat (SS/f).

gruppe	n	S	$\overline{y}$	f	SS	s^2
1	4	104	26.00	3	234.00	78.00
2	4	65	16.25	3	12.75	4.25
3	4	120	30.00	3	44.00	14.67
4	4	93	23.25	3	122.75	40.92
sum	16	382		12	413.50	
gennemsnit			23.88			34.46

Tabel 4.5 Opgave 4.1: Tørstofindhold (i g) i planter under forskellige dyrkningsbetingelser, samt visse hjælpestørrelser til beregningerne.

	kontrol	metode A	metode B
	4.17	4.81	6.31
	5.58	4.17	5.12
	5.18	4.41	5.54
	6.11	3.59	5.50
	4.50	5.87	5.37
	4.61	3.83	5.29
	5.17	6.03	4.92
	4.53	4.89	6.15
	5.33	4.32	5.80
	5.14	4.69	5.26
sum	50.32	46.61	55.26
sum af kvadrater	256.2702	222.9185	307.1296

planter som en kontrolgruppe under »sædvanlige« omstændigheder. Efter en bestemt vækstperiode er planterne høstet, og man har målt tørstofindhold i hver af dem. De opnåede resultater fremgår af Tabel 4.5.

Analysér talmaterialet. (Opstil en passende statistisk model, estimér parametrene, test relevante hypoteser; kan man foretage modelkontrol?)

Opgave 4.2 (Etnocentricisme)
En forsker ved Columbia University ville undersøge om det amerikanske skolesystems integration af børn af forskellig race gav sig udslag i at børnene fik forskellige holdninger til deres egen og til andre racer. Han udsatte derfor fire grupper af børn for en *etnocentricisme-test* der måler i hvilken grad det enkelte barn foretrækker at omgås og respektere børn af samme etniske gruppe som det selv frem for børn af andre etniske grupper. (Et barn får altså et højt etnocentricisme-tal hvis det i høj grad foretrækker kammerater af sin egen race.)

Tabel 4.6 Opgave 4.2: Etnocentricisme-tal for fire grupper af børn.

1. sorte børn i blandede skoler:
   15 12 14 15 22 21 18 18 23 22 14 22 15 7 17 12 18 17 19 18 14 19 13 21 21 12 16 14 22 16 17 20 18 22 20 23 20 14 20 17 13 17 14 16 15 15 12 17 13 24
2. hvide børn i blandede skoler:
   12 12 13 11 16 12 19 12 5 8 20 7 12 24 13 18 14 18 8 16 9 19 9 1 9 11 9 17 16 16 12 7 9 24 13 15 20 14 17 8 15 16 6 5 14 7 12 22 14 11
3. sorte børn i adskilte skoler:
   11 11 13 13 9 21 21 9 13 13 11 10 12 18 19 18 12 18 17 19 21 22 22 17 12 13 21 14 20 19 15 19 12 12 16 14 16 11 15 12 9 15 11 11 10 10 14 12 11 13
4. hvide børn i adskilte skoler:
   23 17 14 18 16 18 15 21 22 20 10 18 16 13 10 19 10 15 22 15 12 11 9 14 21 10 15 14 7 14 21 10 14 10 24 24 12 9 14 13 14 16 12 14 22 21 15 9 9 9

De fire grupper af børn er

1. sorte børn i blandede skoler,
2. hvide børn i blandede skoler,
3. sorte børn i adskilte skoler,
4. hvide børn i adskilte skoler.

Der er undersøgt 50 børn fra hver gruppe. Resultaterne fremgår af Tabel 4.6.

Analysér talmaterialet.

(Datamaterialets størrelse gør det muligt også at vurdere rimeligheden af en antagelse om at observationerne i de enkelte grupper er uafhængige normalfordelte observationer.)

TIP: Hjælpestørrelser til beregningerne:

	sum	sum af kvadrater
sorte børn i blandede skoler	854	15254
hvide børn i blandede skoler	647	9607
sorte børn i adskilte skoler	727	11313
hvide børn i adskilte skoler	751	12325

Opgave 4.3 (Kyllingers vækst)

Man har foretaget en forsøgsrække med kyllinger for at bedømme virkningen af et formodet vækstfremmende hormon. Forsøget kan tænkes opbygget på følgende måde:

- Den *eksperimentelle enhed* består af et antal kyllinger der lever i samme hønsehus og får samme kost; *måleresultatet* er den gennemsnitlige vægt af de fuldvoksne fugle.
- De eksperimentelle enheder er inddelt i *tre grupper*:
  - én gruppe får normal kost (kontrolgruppen),

Tabel 4.7 Opgave 4.3: Gennemsnitlig vægt (i pund) af de fuldvoksne fugle i hver af de i alt 24 eksperimentelle enheder.

kontrol	lav dosis	høj dosis
3.93	3.99	3.96
3.78	3.96	3.94
3.88	3.96	4.02
3.93	4.03	4.06
3.84	4.10	3.94
3.75	4.02	4.09
3.98	4.06	4.17
3.84	3.92	4.12

Tabel 4.8 Opgave 4.3: Nogle hjælpestørrelser til beregningerne.

gruppe	antal	sum	sum af kvadrater
kontrol	8	30.93	119.6267
lav dosis	8	32.04	128.3446
høj dosis	8	32.30	130.4642
sum	24	95.27	378.4355

- én gruppe får normal kost plus hormonet i lav dosis,
- én gruppe får normal kost plus hormonet i høj dosis.

Hver gruppe indeholder otte eksperimentelle enheder.

Resultatet af forsøget ses i Tabel 4.7.

Undersøg ved hjælp af ensidet variansanalyse om man kan sige at det tilsatte hormon faktisk virker vækstfremmende.

TIP: Benyt eventuelt Tabel 4.8.

Undersøgelsen bør suppleres med forskellige former for modelkontrol. Man kan således kontrollere antagelsen om varianshomogenitet ved hjælp af Bartletts test. – Hvilke muligheder er der for grafiske tests af normalfordelingsantagelsen?

5 Simpel lineær regressionsanalyse

Regressionsanalyse handler om at undersøge, hvordan en målt størrelse afhænger af en eller flere såkaldte baggrundsvariable.

Antag at der foreligger et statistisk datamateriale som er fremkommet ved at man på hvert af et antal »individer« (f.eks. forsøgspersoner eller forsøgsdyr eller enkelt-laboratorieforsøg osv.) har målt værdien af et antal størrelser (variable). En af disse størrelser indtager en særstilling idet man nemlig gerne vil »beskrive« eller »forklare« denne størrelse ved hjælp af de øvrige. Tit kalder man den variabel der skal beskrives, for y, og de variable ved hjælp af hvilke man vil beskrive, for $x_1, x_2, \ldots, x_p$. Andre betegnelser fremgår af følgende oversigt:

$x_1, x_2, \ldots, x_p$	y
baggrundsvariable	modelleret variabel
uafhængige variable	afhængig variabel
forklarende variable	forklaret variabel
	responsvariabel

Her skitseres et par eksempler:

1. Lægen observerer den tid y som patienten overlever efter at være blevet behandlet for sygdommen, men lægen har også registreret en mængde baggrundsoplysninger om patienten, så som køn, alder, vægt, detaljer om sygdommen osv. Nogle af baggrundsoplysningerne kan måske indeholde information om hvor længe patienten kan forventes at overleve.

2. I en række nogenlunde ens i-lande har man bestemt mål for lungekræftforekomst, cigaretforbrug og forbrug af fossilt brændstof, altsammen pr. indbygger. Man kan da udnævne lungekræftforekomst til y-variabel og søge at »forklare« den ved hjælp af de to andre variable der så får rollen som forklarende variable.

3. Man ønsker at undersøge et bestemt stofs giftighed. Derfor giver man det i forskellige koncentrationer til nogle grupper af forsøgsdyr og ser hvor mange af dyrene der dør. Her er koncentrationen x en uafhængig variabel hvis værdi eksperimentator bestemmer, og antallet y af døde er den afhængige variabel.

En *statistisk model* i den slags situationer skal blandt andet

- udtrykke middelværdien af y-variablen som en simpel og »pæn« funktion af de forklarende variable, og
- angive hvilken sandsynlighedsfordeling der skal beskrive y-ernes tilfældige variation.

I det følgende skal vi beskæftige os med modeller hvor den tilfældige variation beskrives af en *normalfordeling* og hvor middelværdien kan skrives som en linearkombination af (to eller flere) ukendte parametre med de forklarende variable som koefficienter. Den slags modeller kan generelt formuleres på følgende måde: For hvert individ i ($= 1, 2 \ldots, n$) foreligger der dels en måling af en størrelse y (på en kontinuert måleskala), dels værdier af p baggrundsvariable $x_1, x_2, \ldots, x_p$. For hvert i har man altså de $p + 1$ tal

$$x_{i1}, x_{i2}, \ldots, x_{ip},\ y_i,$$

hvor y_i betegner den værdi af y der er målt på det i-te individ, og hvor x_{ij} betegner værdien af den j-te baggrundsvariabel hos individ nr. i. Modellen er da at tallene $y_1, y_2, \ldots, y_n$ opfattes som observerede værdier af uafhængige normalfordelte stokastiske variable $Y_1, Y_2, \ldots, Y_n$ hvor

$$\begin{aligned} Y_i &\sim \mathcal{N}(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2) \\ &= \mathcal{N}(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p, \sigma^2). \end{aligned}$$

Her er koefficienterne $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ ukendte parametre der fastlægger hvordan middelværdien bestemmes kvantitativt ud fra de forklarende variable, og variansparameteren σ^2 beskriver den tilfældige variation omkring middelværdien.

Den generelle model omtales nærmere i kapitlet om *multipel lineær regressionsanalyse*, men vi skal først og fremmest undersøge det vigtige specialtilfælde *simpel lineær regressionsanalyse*.

5.1 Præsentation af modellen

Resten af dette kapitel handler om den situation hvor der foreligger et antal talpar (x, y), og hvor man ønsker at opstille en statistisk model for y-erne; x-erne skal indgå i modellen på den måde at middelværdien af Y kan skrives som $\alpha + \beta x$ for passende valg af parametrene α og β. Skematisk ser det sådan ud hvis der er n talpar og par nr. i betegnes (x_i, y_i):

baggrundsvariabel	observation
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

Vi formulerer en statistisk model for y-erne på følgende måde:

- tallene $y_1, y_2, \ldots, y_n$ er observerede værdier af nogle stokastiske variable $Y_1, Y_2, \ldots, Y_n$;
- de stokastiske variable $Y_1, Y_2, \ldots, Y_n$ er uafhængige og normalfordelte med samme varians σ^2;
- tallene $x_1, x_2, \ldots, x_n$ betragtes som faste tal – de er altså *ikke* (i denne model) observerede værdier af stokastiske variable;
- middelværdien af den i-te måling kan skrives som $\alpha + \beta x_i$, dvs. som en linearkombination af to ukendte parametre α og β og med koefficienterne 1 og x_i:

$$\mathrm{E}\, Y_i \quad = \quad \alpha + \beta x_i, \quad i = 1, 2, \ldots, n.$$

Denne model kan kort skrives som

$$Y_i \quad \sim \quad \mathcal{N}(\alpha + \beta x_i, \sigma^2). \tag{5.1}$$

Modellen beskriver y-ernes *systematiske* variation ved hjælp af parametrene α og β og de kendte konstanter $x_1, x_2, \ldots, x_n$; den beskriver den *tilfældige* variation ved hjælp af normalfordelingen og den ukendte variansparameter σ^2. Modellen kaldes en *simpel lineær regressionsanalyse*-model, og β kaldes *regressionskoefficienten*.

De to størrelser x og y indgår på vidt forskellig måde i modellen, og det er derfor ikke ligegyldigt hvad man lader være x og hvad y. I nogle tilfælde er det ganske klart hvad der er »observation«, og hvad der er »baggrundsvariabel«, men i andre tilfælde er det i høj grad et valg man træffer. Her kommer to eksempler der illustrerer de to muligheder.

Eksempel 5.1 (Fædre og sønner)
I slutningen af 1800-tallet opstod i England faget *biometri*, et fag i grænseområdet mellem (hvad vi i vore dage forstår ved) statistik og biologi. De emner biometrikerne tog op, var i høj grad emner med forbindelse til den nye og kontroversielle arvelighedslære idet de håbede at kunne finde bekræftelser på og numeriske beskrivelser af evolutionsteorien. Desuden var nogle af biometrikerne meget optaget af den almindelige debat om de sociale problemer i samfundet (og de var store), og de måtte derfor gøre sig overvejelser over hvad arvelighedslæren kunne fortælle om samfundets udvikling.

Biometrikeren F. Galton (1822-1911) spekulerede over det tilsyneladende almindelige forfald: hvordan kunne det være at fremragende fædre ikke fik tilsvarende fremragende sønner (– eller var det bare noget man syntes?). Nu er det vanskeligt at finde et mål for »fremragende-hed«, så Galton gav sig til at undersøge *højde* i stedet. Han foranstaltede en større indsamling af data om medlemmer af britiske familier.[1]

Galton foretog det vi nutildags kalder en regressionsanalyse, og han fandt at høje fædre gennemsnitligt fik sønner der ikke var så høje som de selv, men dog lå over gennemsnittet i befolkningen. Omvendt fik små fædre gennemsnitligt sønner der var højere end dem selv, men dog lå under gennemsnittet i befolkningen. Denne tilsyneladende

[1] Han indsamlede data om blandt andet øjenfarve, gemyt, kunstneriske evner, sygdomme, valg af ægtefælle, frugtbarhed, og altså højde.

Tabel 5.1 Fædre og sønner: Fordelingen af 1078 par af far og søn efter faderens højde og sønnens højde. Højderne er angivet i inches.

		Faderens højde																
		59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
	60	-	-	-	-	1	-	1	-	-	-	-	-	-	-	-	-	-
	61	-	-	-	-	1	-	-	-	1	-	-	-	-	-	-	-	-
	62	-	1	-	-	-	1	-	-	1	-	-	-	-	-	-	-	-
S	63	-	-	-	2	2	2	4	5	3	1	-	-	1	-	-	-	-
ø	64	1	-	2	4	3	4	8	9	3	1	2	1	1	-	-	-	-
n	65	2	1	-	2	3	10	13	11	7	6	4	2	-	-	-	-	-
n	66	-	-	1	2	5	9	10	17	18	16	5	2	3	1	-	-	-
e	67	-	2	2	5	3	14	20	26	26	19	13	14	3	-	1	-	-
n	68	-	-	2	2	8	10	10	24	31	24	30	13	8	10	2	-	-
s	69	-	-	1	-	5	5	13	18	16	24	29	22	10	4	2	-	1
	70	-	-	-	-	1	3	6	19	12	20	22	19	14	6	3	2	1
h	71	-	-	-	-	-	3	5	9	10	19	15	21	11	8	5	1	1
ø	72	-	-	-	-	-	-	3	1	7	8	11	11	10	9	3	-	-
j	73	-	-	-	-	-	-	1	1	2	8	6	6	8	6	3	-	1
d	74	-	-	-	-	1	-	2	2	-	5	2	3	6	3	3	-	2
e	75	-	-	-	-	-	-	-	-	-	1	2	-	2	1	2	1	-
	76	-	-	-	-	-	-	-	-	-	1	-	-	1	1	1	-	-
	77	-	-	-	-	-	-	-	-	-	1	-	1	-	-	2	-	-
	78	-	-	-	-	-	-	-	-	-	-	1	1	-	-	1	-	-

nærmen sig det gennemsnitlige så Galton som en tilbagegang og kaldte det derfor en *regression*.[2] [3]

I Tabel 5.1 er gengivet et talmateriale som to andre biometrikere indsamlede, idet de for 1078 par af far og søn registrerede faderens højde og sønnens højde. Tabellen skal læses på den måde at der f.eks. var syv tilfælde ud af de 1078 hvor faderen var 67 inches og sønnen 65 inches.

Der er tale om en situation med $n = 1078$ talpar (x, y), men det er ikke uden videre klart at den ene af de to højder er en »baggrundsvariabel« og den anden en »observation«, faktisk må man vel sige at de er »observationer« begge to. Alligevel kan man *vælge* at opfatte f.eks. faderens højde som »baggrundsvariabel« og sønnens højde som »observation« og så foretage en såkaldt »regression af sønnens højde på faderens højde«; det kan man *vælge* at gøre hvis man er interesseret i at undersøge hvordan man kan forudsige, *prædiktere*, sønnens højde ud fra faderens.

Eksempel 5.2 (Kvælning af hunde)
Man ved at *hypoxi* (nedsat ilttilførsel til hjernen) kan bevirke at der dannes forskellige skadelige stoffer i hjernen, og det kan i værste fald medføre alvorlige hjerneskader. (Hypoxi kan blandt andet forekomme ved fødsler.) Man er derfor interesseret i at udvikle en simpel metode til at afgøre *om* der har være hypoxi og i givet fald hvor længe. Man har udført en række forsøg for at undersøge om koncentrationen af *hypoxantin* i cerebrospinalvæsken kan benyttes som hypoxiindikator.

[2] *regression* betyder *tilbagegang*.

[3] Vi kan altså takke Galton for betegnelsen regressionsanalyse. Det er vistnok også ham der skal have æren for at have udbredt betegnelsen *normalfordelingen* om normalfordelingen.

Tabel 5.2 Kvælning af hunde: Målinger af hypoxantinkoncentration til de fire forskellige tidspunkter. I hver gruppe er observationerne ordnet efter størrelse.

varighed (min)	koncentration (μmol/l)						
0	0.0	0.0	1.2	1.8	2.1	2.1	3.0
6	3.0	4.9	5.1	5.1	7.0	7.9	
12	4.9	6.0	6.5	8.0	12.0		
18	9.5	10.1	12.0	12.0	13.0	16.0	17.1

Syv hunde er (under bedøvelse) blevet udsat for iltmangel ved sammenpresning af luftrøret, og hypoxantinkoncentrationen måltes efter 0, 6, 12 og 18 minutters forløb. Det var af forskellige grunde ikke muligt at foretage målinger på alle syv hunde til alle fire tidspunkter, og det kan heller ikke afgøres hvordan målinger og hunde hører sammen. Resultaterne af forsøget er vist i Tabel 5.2.

Man kan anskue situationen på den måde at der foreligger $n = 25$ par sammenhørende værdier af koncentration og varighed. Varighederne er kendte størrelser – de indgår i forsøgsplanen – hvorimod koncentrationerne kan betragtes som observerede værdier af stokastiske variable: tallene er ikke ens fordi der er en vis biologisk variation og en vis forsøgsusikkerhed, og det kan passende modelleres som tilfældig variation. Det er derfor nærliggende at søge at modellere tallene ved hjælp af en regressionsmodel med koncentration som y-variabel og varighed som x-variabel. Man kan naturligvis ikke på forhånd vide om varigheden i sig selv er en hensigtsmæssig forklarende variabel. Måske viser det sig at man bedre kan beskrive koncentrationen som en lineær funktion af *logaritmen* til varigheden end som en lineær funktion af selve varigheden, men det betyder blot at der er tale om en lineær regressionsmodel med logaritmen til varigheden som forklarende variabel.

Der melder sig nu forskellige spørgsmål:

1. Hvordan estimerer man de indgående parametre α, β og σ^2?

2. Hvordan vurderer man om en model af formen (5.1) giver en fornuftig beskrivelse af datamaterialet?

3. Hvordan tester man hypoteser om parametrene?

5.2 Estimation af parametrene

Vi estimerer α og β ved maximum likelihood metoden der på grund af normalfordelingsantagelsen er det samme som en *mindste kvadraters metode*, og vi estimerer σ^2 som residualkvadratsummen divideret med antallet af frihedsgrader.

Estimation af α og β

Parametrene α og β estimeres ved at maksimalisere den til grundmodellen (5.1) hørende likelihoodfunktion

$$
\begin{aligned}
& L(\alpha, \beta, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2} \frac{\left(y_i - (\alpha + \beta x_i)\right)^2}{\sigma^2} \right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2 \right).
\end{aligned}
$$

Det fremgår heraf at de bedste estimater over α og β er de værdier der *minimaliserer* kvadratsummen

$$
\sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2 . \tag{5.2}
$$

Disse værdier kan man enten bestemme ved hjælp af standardmetoder til bestemmelse af ekstremumspunkter for funktioner af to variable, eller man kan søge at slippe lettere om ved det ved at foretage snedige omskrivninger af kvadratsummen på lignende måde som ved estimation i enstikprøveproblemet (side 15), i tostikprøveproblemet (side 31) og i ensidet variansanalyse (side 48). Vi prøver med den snedige omskrivning:

Det er hensigtsmæssigt at operere med x-ernes og y-ernes afvigelser fra deres gennemsnit $\overline{x}$ og $\overline{y}$. Derfor omskrives kvadratsummen (5.2) således:

$$
\begin{aligned}
& \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2 \\
&= \sum_{i=1}^{n} \left((y_i - \overline{y}) + (\overline{y} - (\alpha + \beta\overline{x})) - \beta(x_i - \overline{x}) \right)^2 \\
&= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\
&\quad + n\left(\overline{y} - (\alpha + \beta\overline{x})\right)^2 \\
&\quad + \beta^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&\quad - 2\beta \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) ,
\end{aligned} \tag{5.3}
$$

idet de øvrige to dobbelte produkter fra kvadreringen af den treleddede størrelse bliver 0. Omskrivningen har ført til et udtryk hvor α kun optræder i ét led, nemlig $n(\overline{y} - (\alpha + \beta\overline{x}))^2$, og dette led antager sin mindsteværdi 0 netop når α er lig $\overline{y} - \beta\overline{x}$. Dernæst skal vi bestemme β så det minimaliserer summen af de tre øvrige led, dvs. minimaliserer udtrykket

$$
\beta^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 - 2\beta \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) + \sum_{i=1}^{n} (y_i - \overline{y})^2
$$

eller kort

$$
\beta^2 SS_x - 2\beta SP_{xy} + SS_y
$$

hvor vi har benyttet de ofte anvendte betegnelser SS_x hhv. SS_y for sum af kvadratiske afvigelser af x-er hhv. y-er, og SP_{xy} for sum af produkter af afvigelser af x-er og y-er.[4]

Udtrykket $\beta^2 SS_x - 2\beta SP_{xy} + SS_y$ er en andengradsfunktion af β, og da koefficienten til β^2 er positiv, så har funktionen ét minimumspunkt, og det findes ved at differentiere og sætte den afledede lig 0; man får da at β estimeres ved

$$\hat{\beta} \;=\; \frac{SP_{xy}}{SS_x}.$$

Ifølge betragtningerne ovenfor er det dertil svarende bedste valg af α

$$\hat{\alpha} \;=\; \overline{y} - \hat{\beta}\,\overline{x}.$$

Hermed har vi løst estimationsproblemet for så vidt angår α og β. Den *estimerede regressionslinie* er (linien hvis ligning er)

$$y \;=\; \hat{\alpha} + \hat{\beta}x.$$

Undertiden, især når man skal udføre beregningerne mere eller mindre med håndkraft, kan man have fornøjelse af et andet udtryk for $\hat{\beta}$ eller måske snarere for SP_{xy} og SS_x. Ved almindelige og lette formelmanipulationer finder man følgende formler, hvor hver gang det første lighedstegn er definitionslighedstegnet og det andet viser det alternative udtryk:

$$\begin{aligned}
SP_{xy} &= \sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) &&= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right),\\
SS_x &= \sum_{i=1}^{n}(x_i-\overline{x})^2 &&= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2,\\
SS_y &= \sum_{i=1}^{n}(y_i-\overline{y})^2 &&= \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2.
\end{aligned}$$

Estimation af σ^2

Variansestimatet er som altid residualkvadratsummen divideret med antallet af frihedsgrader:

1. *Residualkvadratsummen* får vi ved at erstatte α og β med (udtrykkene for) $\hat{\alpha}$ og $\hat{\beta}$ i kvadratsummen (5.2), så den er

   $$\sum_{i=1}^{n}\left(y_i - (\hat{\alpha}+\hat{\beta}x_i)\right)^2.$$

   Hvis man i stedet indsætter i udtrykket (5.3) og reducerer, får man et alternativt udtryk for residualkvadratsummen, nemlig

   $$\sum_{i=1}^{n}(y_i-\overline{y})^2 - \hat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2 \;=\; SS_y - \hat{\beta}^2 SS_x \;=\; SS_y - \frac{SP_{xy}^2}{SS_x}.$$

[4] SS = Sum of Squared deviations, SP = Sum of Products.

2. *Antallet af frihedsgrader* er $n-2$ fordi der er n observationer og der er estimeret 2 middelværdiparametre.

Variansen σ^2 estimeres derfor ved

$$\begin{aligned} s_{02}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2 \\ &= \frac{1}{n-2}\left(SS_y-\frac{SP_{xy}^2}{SS_x}\right). \end{aligned} \tag{5.4}$$

Eksempel 5.3 (Fædre og sønner, fortsat fra side 62)
Vi vil udregne »regressionen af sønnens højde på faderens højde«, dvs. vi vil bruge sønnens højde som y og faderens højde som x i en lineær regression.

På grundlag af tallene i Tabel 5.1 udregnes først nogle hjælpestørrelser, se Tabel 5.3, og ved hjælp af disse udregnes

$$\begin{aligned} SP_{xy} &= \sum_{i=1}^{n}x_iy_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right) \\ &= 5015024-\frac{72979\times 74018}{1078} = 4114.260, \\ SS_x &= \sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 \\ &= 4948575-\frac{72979^2}{1078} = 8005.018, \\ SS_y &= \sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2 \\ &= 5090344-\frac{74018^2}{1078} = 8095.091. \end{aligned}$$

Den estimerede regressionskoefficient er

$$\hat{\beta} = SP_{xy}/SS_x = 4114.260/8005.018 = 0.514,$$

og den estimerede skæring med ordinataksen er

$$\hat{\alpha} = \overline{y}-\hat{\beta}\overline{x} = \frac{74018}{1078}-0.514\times\frac{72979}{1078} = 33.86.$$

Regressionsmodellen anviser altså følgende relation:

$$\text{søns højde} = 33.86+0.514\times\text{fars højde}.$$

Residualkvadratsummen er

$$SS_y-SP_{xy}^2/SS_x = 8095.091-4114.260^2/8005.018 = 5980.525$$

så den estimerede varians er $s_{02}^2 = 5980.525/(1078-2) = 5.558$ med 1076 frihedsgrader.

Der er naturligvis også den mulighed at udregne regressionen af faderens højde på sønnens højde. Man vil da få

$$\text{fars højde} = 32.79+0.508\times\text{søns højde}$$

og en estimeret varians på $s_{02}^2 = 5.495$, ligeledes med 1076 frihedsgrader. Som det ses, er det ikke ligegyldigt hvilken af de to højder man benytter som x og hvilken som y.

Tabel 5.3 Fædre og sønner: Hjælpestørrelser til beregningerne.

Sum af	
1	1078
faders højde	72979
søns højde	74018
faders højde × faders højde	4948575
søns højde × søns højde	5090344
faders højde × søns højde	5015024

Afrundingsfejl

De forskellige formeludtryk for SP_{xy}, SS_x, SS_y og s_{02}^2 er allesammen lige rigtige set fra et matematisk synspunkt. Men hvis man tænker på dem som forskrifter for hvordan man skal regne tingene ud, så har de hver deres fordele og ulemper. Hvis man f.eks. skal udregne s_{02}^2, så er formlen

$$s_{02}^2 = \frac{1}{n-2}\left(SS_y - \frac{SP_{xy}^2}{SS_x}\right)$$

praktisk fordi den viser hvordan man finder s_{02}^2 ud fra tre tal som man formentlig allerede har regnet ud i anden forbindelse; men formlen er upraktisk fordi den indebærer at man skal trække to ofte næsten lige store positive tal (SS_y og SP_{xy}^2/SS_x) fra hinanden, og det betyder at det hele let kan ende i afrundingsfejl såfremt man ikke har regnet med tilstrækkeligt mange cifre i mellemregningerne. Omvendt er formlen

$$s_{02}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\hat{\alpha} + \hat{\beta}x_i)\right)^2$$

ikke nær så følsom over for afrundingsfejl, men den er til gengæld besværlig at regne ud fordi der skal man udregne de n prædikterede værdier $\hat{\alpha} + \hat{\beta}x_i$, derpå de tilsvarende residualer, og endelig summen af de kvadrerede residualer.

Moralen må derfor være at enten skal man være doven men tænke sig om, eller også skal man regne meget, men så behøver man ikke tænke sig om.

5.3 Parameterestimaternes middelfejl

Regressionsanalyse er i udpræget grad et forsøg på at modellere *kvantitative sammenhænge*, og derfor er det ikke tilstrækkeligt blot at udregne parameterestimaterne, man skal også skaffe sig en idé om hvor præcise de er.

Når man *tester hypoteser*, foregår det ved at man udregner værdien af en passende valgt teststørrelse der fungerer som et mål for hvor godt de foreliggende observationer stemmer overens med hypotesen. Derefter bestemmer man den såkaldte testsandsynlighed der er sandsynligheden for at få et sæt observationer der stemmer dårligere overens med hypotesen end de faktiske

observationer gør. Når man overhovedet kan tale om en sådan sandsynlighed, er det takket være den statistiske model; den statistiske model fortæller nemlig at observationerne kan opfattes som observerede værdier af stokastiske variable der følger en nærmere angivet sandsynlighedsfordeling, og man kan derfor sige at den statistiske model sætter os i stand til at sammenligne de faktiske observationer med alle de andre sæt observationer man også kunne have fået idet man tager hensyn til, med hvilke sandsynligheder de forekommer.

En anden side af dette at »sammenligne med hvad man ellers kunne have fået« er bestemmelse af estimatorernes *middelfejl*. Et estimat er jo regnet ud på grundlag af de faktiske observationer, men ved hjælp af den statistiske model kan man få svar på spørgsmålet: hvilke andre talværdier af estimatet kunne man også have fået og med hvilke sandsynligheder. For da estimatet er en funktion af observationerne, og da observationerne opfattes som observerede værdier af stokastiske variable, så kan estimatet også opfattes som en observeret værdi af en vis stokastisk variabel, *estimatoren*, hvis sandsynlighedsfordeling man i princippet kan finde. Ofte er man endda kun interesseret i at vide inden for hvilke grænser størstedelen af sandsynlighedsmassen er beliggende, og til det brug udregner man den såkaldte *middelfejl*, dvs. *estimatorens standardafvigelse*. Som en tommelfingerregel gælder nemlig at intervallet *middelværdien plus/minus to gange standardafvigelsen* afgrænser ca. 95% af sandsynlighedsmassen[5], og i den forstand er middelfejlen et direkte mål for hvor unøjagtigt estimatet er.[6]

Vi skal ikke komme nærmere ind på *hvordan* man når frem til formeludtryk for middelfejl, men her er nogle resultater for den lineære regressionsmodel:

1. Middelfejlen på $\hat{\beta}$ er $\sqrt{\sigma^2/SS_x}$.
2. (a) Middelfejlen på $\hat{\alpha}$ er $\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{SS_x}\right)}$.
   (b) Estimatorerne $\hat{\alpha}$ og $\hat{\beta}$ er korrelerede; korrelation mellem dem er $-1\Big/\sqrt{1+\frac{SS_x}{n\overline{x}^2}}$.
3. (a) Middelfejlen på $\hat{\alpha}+\hat{\beta}\overline{x}$ er $\sqrt{\sigma^2/n}$.
   (b) Estimatorerne $\hat{\alpha}+\hat{\beta}\overline{x}$ og $\hat{\beta}$ er ukorrelerede.

Disse udtryk er de teoretiske middelfejl hvori optræder den teoretiske varians σ^2 på Y. Da vi ikke kender parameteren σ^2, må vi i stedet indsætte et estimat over den, f.eks. s^2_{02}, og derved få de estimerede middelfejl.

Af udtrykket for middelfejlen på $\hat{\beta}$ ses at det er en fordel at x-værdierne ligger spredt over et stort interval for så bliver SS_x stor og middelfejlen derved lille.

À propos middelfejl kan det være værd at nævne at middelfejlen på en estimator s^2 over variansparameteren σ^2 i en normalfordelingsmodel er lig

[5] Det er især rigtigt hvis estimatoren er normalfordelt.

[6] Eksempel: Hvis man har udregnet $\hat{\beta}$ til 1.534 og middelfejlen på $\hat{\beta}$ til 0.3, så véd man at ca. 95% af alle de andre $\hat{\beta}$-værdier man også kunne have fået, ligger i et interval af længde 1.2 (nemlig intervallet $\beta \pm 2 \times 0.3$), og deraf bør man bl.a. drage den konsekvens at $\hat{\beta}$ *ikke* skal angives med tre decimaler, men kun med én.

$\sigma^2\sqrt{2/f}$, hvor f er antallet af frihedsgrader for s^2. Deraf ses hvordan variansestimatet bliver bedre jo flere frihedsgrader det har.

5.4 En anden formulering af modellen

I den oprindelige formulering af den lineære regressionsmodel var der tale om et antal »uspecificerede« talpar (x, y). Ofte er det sådan at der foreligger flere målinger af y for hvert x (det er for eksempel tilfældet i eksemplet med kvælning af hunde). Det gør ikke spor at der er flere talpar med det samme x, men undertiden er det hensigtsmæssigt at notationen kan indfange dette forhold, bl.a. når man vil lave regneopskrifter der er overkommelige at benytte med »håndkraft«. Vi præsenterer derfor nu en anden formulering af den lineære regressionsmodel. Skematisk ser situationen sådan ud:

baggrundsvariabel	observationer			
x_1	y_{11}	y_{12}	$\cdots$	y_{1n_1}
x_2	y_{21}	y_{22}	$\cdots$	y_{2n_2}
x_3	y_{31}	y_{32}	$\cdots$	y_{3n_3}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_k	y_{k1}	y_{k2}	$\cdots$	y_{kn_k}

hvor det nu er sådan at værdierne $x_1, x_2, \ldots, x_k$ af baggrundsvariablen x er *forskellige*; hørende til den i-te x-værdi er der de n_i observationer $y_{i1}, y_{i2}, \ldots, y_{in_i}$; det samlede antal observationer er $n = n_1 + n_2 + \ldots + n_k$. Regressionsmodellen (5.1) skrives nu som

$$Y_{ij} \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2).$$

De tidligere indførte hjælpestørrelser SP_{xy}, SS_x og SS_y (side 65) er i den nye notation

$$\begin{aligned}
SP_{xy} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_i-\overline{x})(y_{ij}-\overline{y}) = \sum_{i=1}^{k} n_i(x_i-\overline{x})(\overline{y}_i-\overline{y}) \\
&= \sum_{i=1}^{k} n_i x_i \overline{y}_i - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)\left(\sum_{i=1}^{k} n_i \overline{y}_i\right), \\
SS_x &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_i-\overline{x})^2 = \sum_{i=1}^{k} n_i(x_i-\overline{x})^2 \\
&= \sum_{i=1}^{k} n_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)^2, \\
SS_y &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i \overline{y}_i\right)^2
\end{aligned}$$

Tabel 5.4 Kvælning af hunde: beregningsskema. x-værdierne er varighed i minutter, y-værdierne er koncentration i μmol/l.

i	n_i	x_i	$\overline{y}_i$	$n_i x_i$	$n_i \overline{y}_i$	$n_i x_i \overline{y}_i$	$n_i x_i^2$	$\sum_{j=1}^{n_i} y_{ij}^2$
1	7	0	1.46	0	10.2	0.0	0	22.50
2	6	6	5.50	36	33.0	198.0	216	196.44
3	5	12	7.48	60	37.4	448.8	720	310.26
4	7	18	12.81	126	89.7	1614.6	2268	1197.67
sum	25			222	170.3	2261.4	3204	1726.87

hvor der er benyttet følgende betegnelser:

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \text{ er gennemsnittet af } y\text{-erne hørende til } x_i,$$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{y}_i \text{ er totalgennemsnittet af } y\text{-erne, og}$$

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_i = \frac{1}{n} \sum_{i=1}^{k} n_i x_i \text{ er gennemsnittet af } x\text{-erne.}$$

Parameterestimaterne er stadig

$$\begin{aligned} \hat{\beta} &= \frac{SP_{xy}}{SS_x}, \\ \hat{\alpha} &= \overline{y} - \hat{\beta}\overline{x}, \text{ og} \\ s_{02}^2 &= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - (\hat{\alpha} + \hat{\beta} x_i)\right)^2 \\ &= \frac{1}{n-2} \left(SS_y - \frac{SP_{xy}^2}{SS_x} \right). \end{aligned}$$

Bemærkningerne om afrundingsfejl (side 67) er stadig aktuelle.

Eksempel 5.4 (Kvælning af hunde, fortsat fra side 63)
Vi vil antage at hypoxantinkoncentrationen kan beskrives ved en lineær regressionsmodel med hypoxivarigheden som uafhængig variabel. (Denne antagelse vil blive undersøgt nærmere i en senere fortsættelse af eksemplet, se side 74.)

Vi lader x_1, x_2, x_3 og x_4 betegne de fire tidspunkter 0, 6, 12 og 18 min, og vi lader y_{ij} betegne den j-te koncentrationsværdi til tid x_i. Med de indførte betegnelser kan den tidligere foreslåede statistiske model for talmaterialet formuleres som

$$Y_{ij} \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2).$$

Vi vil udregne værdierne af estimaterne $\hat{\alpha}$, $\hat{\beta}$ og s_{02}^2 over modellens parametre. Man kan selvfølgelig overlade regnearbejdet til en datamat, men det er på den anden side

ikke uoverkommeligt at gøre det med håndkraft. Indledningsvis udregnes forskellige hjælpestørrelser mm., se Tabel 5.4. Heraf fås den estimerede regressionskoefficient til

$$\begin{aligned}
\hat{\beta} &= \frac{SP_{xy}}{SS_x} \\
&= \frac{\sum_{i=1}^{k} n_i x_i \overline{y}_i - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)\left(\sum_{i=1}^{k} n_i \overline{y}_i\right)}{\sum_{i=1}^{k} n_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i x_i\right)^2} \\
&= \frac{2261.4 - \frac{222 \times 170.3}{25}}{3204 - \frac{222^2}{25}}\ \mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1} \\
&= 0.61\ \mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1},
\end{aligned}$$

og det estimerede skæringspunkt med ordinataksen til

$$\begin{aligned}
\hat{\alpha} &= \overline{y} - \hat{\beta}\,\overline{x} \\
&= \frac{170.3\ \mu\text{mol}\,\text{l}^{-1}}{25} - \frac{0.61\ \mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1} \times 222\ \text{min}}{25} \\
&= 1.4\ \mu\text{mol}\,\text{l}^{-1}.
\end{aligned}$$

Variansen estimeres ved

$$s_{02}^2 = \frac{1}{n-2}\left(SS_y - \frac{SP_{xy}^2}{SS_x}\right).$$

Her er

$$\begin{aligned}
SS_y &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i \overline{y}_i\right)^2 \\
&= (1726.87 - 170.3^2/25)\ \mu\text{mol}^2\,\text{l}^{-2} \\
&= 566.79\ \mu\text{mol}^2\,\text{l}^{-2},
\end{aligned}$$

og

$$\frac{SP_{xy}^2}{SS_x} = \frac{749.14^2}{1232.64}\ \mu\text{mol}^2\,\text{l}^{-2} = 455.29\ \mu\text{mol}^2\,\text{l}^{-2},$$

så residualkvadratsummen er $(566.79 - 455.29)\ \mu\text{mol}^2\,\text{l}^{-2} = 111.50\ \mu\text{mol}^2\,\text{l}^{-2}$ og

$$s_{02}^2 = \frac{111.50}{23}\ \mu\text{mol}^2\,\text{l}^{-2} = 4.85\ \mu\text{mol}^2\,\text{l}^{-2},$$

svarende til en estimeret standardafvigelse på 2.2 μmol/l.

Middelfejlen på $\hat{\beta}$ er (jf. side 68)

$$\begin{aligned}
\sqrt{\frac{s_{02}^2}{SS_x}} &= \sqrt{\frac{4.85}{1232.64}}\ \mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1} \\
&= 0.06\ \mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1},
\end{aligned}$$

og middelfejlen på $\hat{\alpha}$ er

$$\begin{aligned}\sqrt{\left(\frac{1}{n}+\frac{\bar{x}^2}{SS_x}\right)s_{02}^2} &= \sqrt{\left(\frac{1}{25}+\frac{(222/25)^2}{1232.64}\right)4.85}\ \mu\text{mol}\,\text{l}^{-1}\\ &= 0.7\,\mu\text{mol}\,\text{l}^{-1}.\end{aligned}$$

Størrelsen af de to middelfejl viser at det er passende at angive $\hat{\beta}$ med to og $\hat{\alpha}$ med én decimal, så vi må konkludere at den *estimerede regressionslinie* er

$$y = 1.4\,\mu\text{mol}\,\text{l}^{-1} + 0.61\,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1}\cdot x\,.$$

5.5 Modelkontrol

Ved simpel lineær regressionsanalyse er den første og vigtigste form for modelkontrol den uhyre simple: at lave en tegning. I et koordinatsystem afsætter man punkterne (x_i, y_i), man indtegner den *estimerede regressionslinie* og ser efter om punkterne fordeler sig passende tilfældigt omkring linien. En tegning kan som regel også afsløre hvad der i givet fald måtte være galt med den lineære regressionsmodel.

Tit kan man også foretage et numerisk test for om den lineære regressionsmodel er brugbar. Det foregår ved at man indlejrer regressionsmodellen i en større model, og derefter tester man på helt sædvanlig vis regressionsmodellen som en hypotese i forhold til den større model. En nødvendig forudsætning for at dette kan lade sig gøre er at der er flere y-er til det samme x; for man bærer sig nemlig ad på den måde at man inddeler y-erne i *grupper* bestående af y-er med samme x, og som den »større model« benytter man en ensidet variansanalysemodel. Vi skal nu se hvordan det nærmere går for sig. – Det følgende skal læses som en fortsættelse af Afsnit 5.4.

Regressionsmodellen

$$Y_{ij} \sim \mathcal{N}(\alpha+\beta x_i, \sigma^2)$$

indlejres i en større model, nemlig i den ensidede variansanalysemodel med k grupper svarende til de k niveauer af x:

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2).$$

Vi benytter så denne model som grundmodel og tester den førstnævnte model som en hypotese i forhold hertil, det vil sige vi tester hypotesen

$$H_2 : \mu_i = \alpha+\beta x_i\,.$$

Teststørrelsen for at teste H_2 er i princippet en kvotient Q mellem to likelihoodfunktionsværdier, men på samme måde som i forbindelse med ensidet variansanalyse kan Q omskrives til en kvotient F mellem to s^2-størrelser. Før

vi specificerer disse størrelser nærmere er det hensigtsmæssigt at opskrive følgende spaltning af regressionsmodellens residualkvadratsum:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-(\hat{\alpha}+\hat{\beta}x_i)\right)^2 \tag{5.5}$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2+\sum_{i=1}^{k}n_i\left(\overline{y}_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2$$

(denne opspaltning følger af formel (4.3) på side 48 ved at erstatte μ_i med $\hat{\alpha}+\hat{\beta}x_i$); den tilsvarende opspaltning af frihedsgraderne er

$$n-2 \quad = \quad (n-k)+(k-2).$$

Formel (5.5) viser hvordan residualkvadratsummen der kan siges at beskrive *den samlede variation omkring regressionslinien*, deles op i en sum af en kvadratsum vedrørende *variationen inden for grupper* og en kvadratsum vedrørende *gruppernes variation omkring regressionslinien*, se også Figur 5.1.

Ved at dividere kvadratsummerne med deres frihedsgrader fås s^2-størrelserne: dels de tidligere indførte s_{02}^2 med $n-2$ frihedsgrader (side 70) og s_0^2 med $n-k$ frihedsgrader (side 49), dels

$$s_2^2 \quad = \quad \frac{1}{k-2}\sum_{i=1}^{k}n_i\left(\overline{y}_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2.$$

Teststørrelsen for hypotesen H_2 om at gruppemiddelværdierne faktisk ligger på en ret linie, er

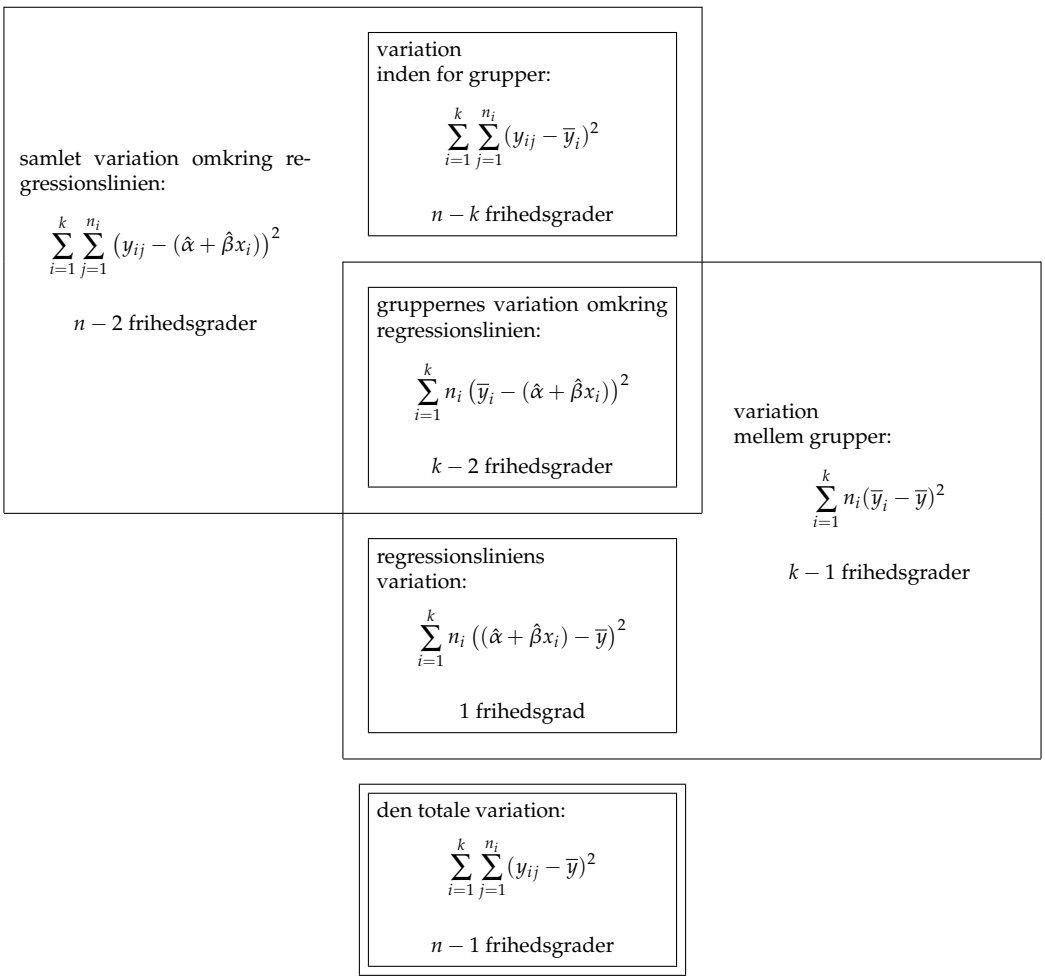
$$F \quad = \quad \frac{s_2^2}{s_0^2}$$

det vil sige gruppernes variation omkring linien målt i forhold til variationen inden for grupperne. Store værdier af F er signifikante, og hvis H_2 er rigtig, vil F følge F-fordelingen med frihedsgrader $k-2$ og $n-k$ således at testsandsynligheden ε er givet som

$$\varepsilon \quad = \quad \mathrm{P}\left(F_{k-2,\,n-k} > F_{\mathrm{obs}}\right)$$

der bestemmes ved hjælp af en tabel over F-fordelingen.

- Hvis ε er meget lille (og F dermed er signifikant stor), så må vi forkaste linearitetshypotesen H_2. Så står vi tilbage med den ensidede variansanalysemodel $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ hvor parametrene $\mu_1, \mu_2, \ldots, \mu_k$ estimeres ved $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k$ og hvor parameteren σ^2 estimeres ved s_0^2 med $n-k$ frihedsgrader.
- Hvis ε ikke er meget lille (og F dermed ikke er signifikant stor), så kan vi godtage den lineære regressionsmodel $Y_{ij} \sim \mathcal{N}(\alpha+\beta x_i, \sigma^2)$ hvor parametrene α og β estimeres ved $\hat{\alpha}$ og $\hat{\beta}$ og hvor parameteren σ^2 estimeres ved s_{02}^2 med $n-2$ frihedsgrader (jf. side 70).



Figur 5.1 Skematisk oversigt over nogle af de i kapitlet forekommende kvadratsummer med tilhørende frihedsgrader.

Bemærk i øvrigt, at kvadratsummen vedrørende gruppernes variation omkring linien ifølge formel (5.5) kan skrives som

$$\sum_{i=1}^{k} n_i\left(\overline{y}_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2 \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-(\hat{\alpha}+\hat{\beta}x_i)\right)^2-\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_i})^2;$$

det kan være praktisk ved udregninger.

Eksempel 5.5 (Kvælning af hunde, fortsat)
Vi vil undersøge om det kan antages at hypoxantinkoncentrationen afhænger lineært af hypoxiens varighed. Da vi er i en situation hvor der er en del y-er til hvert x, er det muligt at udføre det numeriske test for modellen.

Vi har tidligere (side 70 ff) bestemt de talværdier der i givet fald er de bedste esti-

Figur 5.2 Kvælning af hunde: Sammenhørende værdier af hypoxantinkoncentration og hypoxivarighed, samt den estimerede regressionslinie.

Tabel 5.5 Kvælning af hunde: Nogle hjælpestørrelser til beregningerne.

i	n_i	$\sum_{j=1}^{n_i} y_{ij}$	$\overline{y}_i$	f_i	$\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$	s_i^2
1	7	10.2	1.46	6	7.64	1.27
2	6	33.0	5.50	5	14.94	2.99
3	5	37.4	7.48	4	30.51	7.63
4	7	89.7	12.81	6	48.23	8.04
sum	25	170.3		21	101.32	
gennemsnit			6.81			4.82

mater over parametrene, og derved fået den estimerede regressionslinie til

$$y \quad = \quad 1.4\,\mu\text{mol}\,\text{l}^{-1} + 0.61\,\mu\text{mol}\,\text{l}^{-1}\,\text{min}^{-1} \cdot x\,.$$

I Figur 5.2 er indtegnet dels sammenhørende værdier af varighed og koncentration, dels den estimerede regressionslinie. Efter tegningen at dømme er den lineære regressionsmodel ikke helt hen i vejret. I håbet om at kunne bestyrke troen på modellen vil vi udføre det numeriske test for den lineære model.

Som midlertidig grundmodel vil vi benytte en ensidet variansanalysemodel baseret på de fire grupper bestemt af x-erne. Indledningsvis udregnes forskellige hjælpestørrelser mm., se Tabel 5.5. Det fremgår blandt andet at den kvadratsum der beskriver *variationen mellem grupper*, er 101.32 med 21 frihedsgrader. På side 71 fandt vi regressionsmodellens residualkvadratsum til 111.50 med 23 frihedsgrader, så kvadratsummen hørende til gruppernes variation omkring regressionslinien er $111.50 - 101.32 = 10.18$ med $23 - 21 = 2$ frihedsgrader. Teststørrelsen for hypotesen om at gruppemiddelværdierne ligger på en ret linie, er da

$$F \;=\; \frac{10.18/2}{101.32/23} \;=\; \frac{5.09}{4.82} \;=\; 1.06$$

der skal sammenlignes med F-fordelingen med 2 og 21 frihedsgrader, og i denne fordeling er der mere end 30% sandsynlighed for at få en værdi som er større end den observerede der altså på ingen måde er signifikant. Vi har således fået bekræftet linearitetshypotesen.

Traditionelt opsummerer man udregninger og testresultater i et variansanalyseskema, se Tabel 5.6.

Variansanalysemodellen såvel som den lineære regressionsmodel forudsætter at der er varianshomogenitet, så det kan man jo også teste. Vi indsætter s^2-værdierne fra Tabel 5.5 i Bartletts teststørrelse og får

$$\begin{aligned} B \quad &= \quad -\left(6 \ln \frac{1.27}{4.82} + 5 \ln \frac{2.99}{4.82} + 4 \ln \frac{7.63}{4.82} + 6 \ln \frac{8.04}{4.82}\right) \\ &= \quad 5.5 \end{aligned}$$

Tabel 5.6 Kvælning af hunde: Variansanalyseskema vedrørende test af linearitetshypotesen. I skemaet står f for antal frihedsgrader, SS for sum af kvadratiske afvigelser, og $s^2 = SS/f$.

variation	f	SS	s^2	test
inden for grupper	21	101.32	4.82	
gruppernes var. omkring regr.linien	2	10.18	5.09	5.09/4.82=1.06
samlet variation omkring regr.linien	23	111.50	4.85	

der skal sammenlignes med χ^2-fordelingen med $k - 1 = 3$ frihedsgrader. Tabelopslag viser at der er over 10% chance for at få en større B-værdi end værdien 5.5 som derfor ikke er signifikant stor. Med andre ord kan vi opretholde antagelsen om varianshomogenitet.

Alt i alt er der således ikke noget der taler imod at vi beskriver hypoxidataene med en lineær regressionsmodel med hypoxivarighed som uafhængig variabel og hypoxantinkoncentration som afhængig variabel.

5.6 Test af hypoteser om liniens parametre

Man kan naturligvis teste hypoteser om regressionsliniens parametre. Fremgangsmåden er den samme som altid: først estimeres parametrene under hypotesen, dernæst udregnes kvotienten Q mellem de to maksimale likelihoodfunktionsværdier, og endelig bestemmes sandsynligheden for at få et værre sæt observationer, dvs. et sæt observationer der giver et mindre Q. Som ved alle andre tests af hypoteser der har med middelværdier i normalfordelingen at gøre, kan Q omskrives til en F-størrelse der er mere praktisk at have med at gøre, og når der er tale om hypoteser om en enkelt parameter, kan man som en yderligere forsimpling benytte en t-teststørrelse der måske er mere forståelig.

Vi skal ikke her komme ind på de nærmere detaljer, men blot forklare hvordan teststørrelserne kommer til at se ud i disse specielle tilfælde.

Hypotesen $\beta = 0$

Hvis man vil teste hypotesen $H_3 : \beta = 0$ om at regressionskoefficienten er 0, dvs. y afhænger ikke (lineært) af x, så bliver F-teststørrelsen $F = s_3^2/s_{02}^2$ hvor s_{02}^2 er det bedste variansestimat under den aktuelle model, se side 66, og hvor

$$s_3^2 = \frac{1}{1}\sum_{i=1}^{k} n_i \left((\hat{\alpha} + \hat{\beta}x_i) - \overline{y}\right)^2 = \hat{\beta}^2 SS_x = SP_{xy}^2/SS_x$$

er den såkaldte *regressionsliniens variation*. Store værdier af F er signifikante. Der gælder at $F = t^2$ hvor

$$t = \frac{\hat{\beta}}{\sqrt{s_{02}^2/SS_x}}$$

er estimatet $\hat{\beta}$ over β divideret med den estimerede standardafvigelse (dvs. den estimerede middelfejl) på $\hat{\beta}$, jf. side 68. Man kan sige at t-størrelsen måler hvor langt $\hat{\beta}$ ligger fra den formodede værdi 0 når man benytter middelfejlen som målestok. Store værdier af $|t|$ er signifikante.

Man kan bevise at under H_3 vil t være t-fordelt med det antal frihedsgrader som s_{02}^2 har, dvs. med $n-2$ frihedsgrader. Det betyder at testsandsynligheden kan findes ved hjælp af tabeller over t-fordelingen som

$$\varepsilon \;=\; \mathrm{P}\big(|t_{n-2}| > |t_{\mathrm{obs}}|\big) \;=\; 2\,\mathrm{P}\big(t_{n-2} > |t_{\mathrm{obs}}|\big)\,.$$

(Hvis man vil benytte F som teststørrelse, er $\varepsilon = \mathrm{P}(F_{1,n-2} > F_{\mathrm{obs}})$.)

Hvis hypotesen H_3 kan godkendes, skal man udregne et revideret estimat over α og et forbedret estimat over variansen σ^2. Hypotesen H_3 betyder at den forklarende variabel x ikke er nødvendig, men at alle Y-er har samme middelværdi α, dvs. der er tale om et enstikprøveproblem. Under H_3 er estimatet over α derfor totalgennemsnittet $\overline{y}$, og estimatet over σ^2 er

$$s_{03}^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2.$$

Hypotesen $\alpha = 0$

Undertiden følger det af den faglige problemstilling at linien *skal* gå gennem $(0,0)$, dvs. at $\alpha = 0$, i andre situationer kan man være interesseret i at teste hypoteser om α blot for at nå til en så simpel beskrivelse af data som muligt. Hvis man ønsker at teste hypotesen $H_4 : \alpha = 0$ om at linien går gennem $(0,0)$, kan det gøres med F-teststørrelsen $F = s_4^2/s_{02}^2$, hvor s_4^2 er »kvadratsummen« $\dfrac{n\,SS_x\,\hat{\alpha}^2}{SS_x + n\overline{x}^2}$ divideret med sit frihedsgradsantal 1, og s_{02}^2 er variansestimatet under linearitetshypotesen. Store værdier af F er signifikante. Der gælder at $F = t^2$ hvor

$$t \;=\; \frac{\hat{\alpha}}{\sqrt{s_{02}^2\left(\frac{1}{n}+\frac{\overline{x}^2}{SS_x}\right)}}$$

er forholdet mellem estimatet $\hat{\alpha}$ over α og den estimerede middelfejl på $\hat{\alpha}$. Store værdier af $|t|$ er signifikante.

Man kan bevise at under H_4 vil t følge t-fordelingen med samme antal frihedsgrader som variansestimatet i nævneren, dvs. $n-2$ frihedsgrader. Det betyder at testsandsynligheden kan findes ved hjælp af tabeller over t-fordelingen som

$$\varepsilon \;=\; \mathrm{P}\big(|t_{n-2}| > |t_{\mathrm{obs}}|\big) \;=\; 2\,\mathrm{P}\big(t_{n-2} > |t_{\mathrm{obs}}|\big)\,.$$

Hvis hypotesen H_4 kan godkendes, skal man udregne et revideret estimat over regressionskoefficienten β og et forbedret estimat over σ^2. Det nye estimat

over β bliver

$$\hat{\hat{\beta}} = \frac{\sum_{i=1}^{k} n_i x_i \overline{y}_i}{\sum_{i=1}^{k} n_i x_i^2}$$

og estimatet over σ^2 bliver

$$\begin{aligned} s_{04}^2 &= \frac{1}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \hat{\hat{\beta}} x_i \right)^2 \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \hat{\hat{\beta}}^2 \sum_{i=1}^{k} n_i x_i^2 \right). \end{aligned}$$

5.7 Opgaver

Opgave 5.1 (Anscombe's data)
I Tabel 5.7 er vist fire forskellige sæt af (til formålet konstruerede) talpar (x, y) der kan underkastes en lineær regressionsanalyse.

1. Hvis man ikke tænkte nærmere over det, kunne man måske finde på at bære sig ad som om tallene $y_1, y_2, \ldots, y_{11}$ i et givet datasæt var observerede værdier af uafhængige stokastiske variable $Y_1, Y_2, \ldots, Y_{11}$ hvor Y_i var normalfordelt med middelværdi $\alpha + \beta x_i$ og varians σ^2.

   Udregn for hvert datasæt estimaterne $\hat{\alpha}$, $\hat{\beta}$ og s_{02}^2 over parametrene α, β og σ^2.

2. Lav for hvert datasæt et såkaldt *scatterplot*, dvs. en tegning med punkterne (x_i, y_i), og indtegn den estimerede regressionslinie.

3. Hvad kan man lære heraf?

Opgave 5.2 (Forbes' barometriske målinger)
Som bekendt aftager lufttrykket med højden over havets overflade, og derfor kan et barometer benyttes som højdemåler. Imidlertid kan man også bestemme højden ved at koge vand fordi vands kogepunkt aftager med lufttrykket. I 1840erne og 1850erne foretog den skotske fysiker James D. Forbes på 17 forskellige lokaliteter i Alperne og i Skotland en række målinger hvor han bestemte dels vands kogepunkt, dels luftens tryk (omregnet til lufttrykket ved en standardlufttemperatur). Resultaterne er vist i Tabel 5.8.

1. Lufttrykket er angivet i 'inches Hg'. Nutildags måles lufttryk i hPa (hektopascal = millibar). Hvordan omregner man lufttrykkene til hPa?

   Kogepunkterne er angivet i ℉. Hvordan omregner man dem til °C?

   TIP: Der gælder at 1 inch = 2.54 cm og 760 mm Hg = 1013.250 hPa. Endvidere svarer 0 °C til 32℉ og 100 °C til 212℉.

Tabel 5.7 Anscombe's data (opgave 5.1).

datasæt 1		datasæt 2		datasæt 3		datasæt 4	
x	y	x	y	x	y	x	y
10	8.03	7	7.26	11	7.81	8	6.58
8	6.95	4	3.10	4	5.39	8	5.76
13	7.58	14	8.10	5	5.73	8	7.71
9	8.81	9	8.77	13	12.74	8	8.84
11	8.33	8	8.14	14	8.84	8	8.47
14	9.96	10	9.14	12	8.15	8	7.04
6	7.24	13	8.74	10	7.46	8	5.25
4	4.26	11	9.26	9	7.11	19	12.50
12	10.84	6	6.13	6	6.08	8	5.56
7	4.82	12	9.13	7	6.42	8	7.91
5	5.68	5	4.74	8	6.77	8	6.89

Tabel 5.8 Forbes' barometriske målinger (opgave 5.2). – Kogepunktet er angivet i ℉, lufttrykket i 'inches Kviksølv'.

Kogepunkt	Lufttryk
194.5	20.79
194.3	20.79
197.9	22.40
198.4	22.67
199.4	23.15
199.9	23.35
200.9	23.89
201.1	23.99
201.4	24.02
201.3	24.01
203.6	25.14
204.6	26.57
209.5	28.49
208.6	27.76
210.7	29.04
211.9	29.88
212.2	30.06

2. Meningen med eksperimentet er at undersøge *om* og *hvordan* man kan forudsige lufttrykket (og dermed højden over havet) på grundlag af en bestemmelse af vands kogepunkt. Lav et *scatterplot* for at se *om* det skulle være muligt.

3. Bestem den rette linie der fitter punkterne bedst.

   Indtegn den estimerede linie i figuren.

   Hvordan passer linien til punkterne?

4. Fysikerne kan fortælle os at det næppe er lufttrykket selv der afhænger lineært af kogepunktet, men snarere logaritmen til lufttrykket.[7]

   Derfor kan man forsøge sig med *logaritmen* til lufttrykkene i stedet for. Bliver det bedre af det?

Hvis man skal have nogen praktisk fornøjelse af sådanne kogepunktsbestemmelser, er man nødt til at kende den rigtige sammenhæng mellem højde og lufttryk. Sålænge vi holder os til bjerghøjder, aftager lufttrykket eksponentielt med højden, og der gælder at hvis lufttrykket ved havets overflade er p_0 (f.eks. 1013.25 hPa) og lufttrykket i højden h er p_h, så er

$$h \approx 8150\,\mathrm{m} \cdot \ln \frac{p_0}{p_h}\,.$$

Opgave 5.3 (Pattedyrs legemsvægt og hjernevægt)
Man kunne umiddelbart forestille sig at store dyr har en større hjerne end små dyr – eller er det måske de mere intelligente dyr der har de store hjerner? Tabel 5.9 viser den gennemsnitlige legemsvægt og den gennemsnitlige hjernevægt for et antal pattedyr. Dyrene er ordnet efter legemsvægt.

Opgaven går ud på at undersøge hvordan hjernens vægt afhænger af legemsvægten.

1. Hvordan vil et scatterplot af hjernevægt mod legemsvægt (dvs. med legemsvægt som x og hjernevægt som y) se ud?

   Man vil få en mere overskuelig fremstilling af tallene ved at afsætte *logaritmen* til hjernevægt mod *logaritmen* til legemsvægt, se Figur 5.3.

2. Nogle biologer mener at der kunne tænkes at gælde en relation af typen

   $$\text{hjernevægt} \quad = \quad \text{konstant} \cdot \text{legemsvægt}^{2/3}. \tag{5.6}$$

   Begrundelsen skulle være at *hjernens* størrelse og dermed vægt er proportional med dyrets overflade (der skal være nerveforbindelser ud til alle punkter på overfladen), hvorimod *legemets* vægt er proportional med dyrets rumfang. Da overflade er proportional med rumfang$^{2/3}$, når man alt i alt til formel (5.6).

[7] Der er med god tilnærmelse en lineær sammenhæng mellem logaritmen til trykket og den reciprokke af den absolutte temperatur T. For tal i den størrelsesorden som vi her har med at gøre, er T^{-1} imidlertid stort set en lineær funktion af T.

Tabel 5.9 Legemsvægt og hjernevægt for 62 pattedyrearter.

art	legemsvægt (kg)	hjernevægt (g)
afrikansk elefant	6654.000	5712.00
asiatisk elefant	2547.000	4603.00
giraf	529.000	680.00
hest	521.000	655.00
ko	465.000	423.00
okapi	250.000	490.00
gorilla	207.000	406.00
svin	192.000	180.00
æsel	187.100	419.00
brasiliansk tapir	160.000	169.00
jaguar	100.000	157.00
gråsæl	85.000	325.00
menneske	62.000	1320.00
kæmpebæltedyr	60.000	81.00
får	55.500	175.00
chimpanse	52.160	440.00
gråulv	36.330	119.50
kænguro	35.000	56.00
ged	27.660	115.00
rådyr	14.830	98.20
bavian	10.550	179.50
husarabe	10.000	115.00
rhesusabe	6.800	179.00
vaskebjørn	4.288	39.20
rød ræv	4.235	50.40
grøn marekat	4.190	58.00
gulbuget murmeldyr	4.050	17.00
klippegrævling[a]	3.600	21.00
nibæltet bæltedyr	3.500	10.80
pungodder	3.500	3.90
polarræv	3.385	44.50
kat	3.300	25.60
myrepindsvin	3.000	25.00
kanin	2.500	12.10
trægrævling[b]	2.000	12.30
nordamerikansk opossum	1.700	6.30
kuskus	1.620	11.40

(fortsættes)

[a] *Procavia habessinica*
[b] *Dendrohyrax*

(a) Præcisér dette argument.

TIP: Hvis man havde et matematisk model-dyr som var kugleformet eller terningeformet, så kunne man let finde både dets overflade og dets rumfang.

Hvad med »rigtige« dyr?

(fortsat)

art	legemsvægt (kg)	hjernevægt (g)
genette	1.410	17.50
plump-lori	1.400	12.50
bæveregern	1.350	8.10
marsvin	1.040	5.50
afrikansk kæmpepungrotte	1.000	6.60
arktisk jordegern[a]	0.920	5.70
børstesvin	0.900	2.60
pindsvin	0.785	3.50
klippegrævling[b]	0.750	12.30
ørkenpindsvin	0.550	2.40
natabe	0.480	15.50
chinchilla	0.425	6.40
rotte	0.280	1.90
galago	0.200	5.00
muldvarpegnaver	0.122	3.00
guldhamster	0.120	1.00
træspidsmus	0.104	2.50
egern	0.101	4.00
østamerikansk muldvarp	0.075	1.20
stjernemuldvarp	0.060	1.00
bisamrotte	0.048	0.33
stor brun flagermus	0.023	0.30
mus	0.023	0.40
lille brun flagermus	0.010	0.25
lille korthalet spidsmus	0.005	0.14

[a] *Citellus (Spermophilus) undulatus ablusus*
[b] *Heterohyrax brucci*

(b) Hvis formel (5.6) gælder, hvilken sammenhæng er der da mellem logaritmen til hjernevægt og logaritmen til legemsvægt?

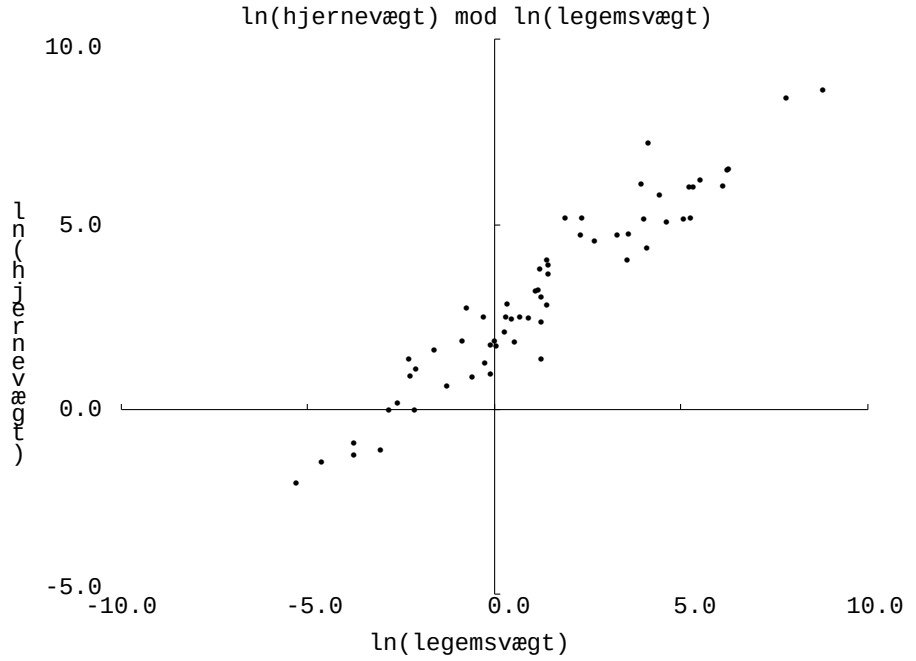
Hvordan harmonerer formodningen (5.6) med de observerede data? (Det har næppe mening at udregne en teststørrelse – for hvad er den statistiske model? Men til orientering kan det oplyses at hypotesen $H : \beta = \beta_0$ i den sædvanlige regressionsmodel testes med $t = \frac{\beta - \beta_0}{\sqrt{s_{02}^2/SS_x}}$ der er t-fordelt med $n-2$ frihedsgrader.)

3. Hvordan kan man i almindelighed finde den bedste rette linie med en given hældning β_0?

   TIP: $y = \alpha + \beta_0 x \Leftrightarrow y - \beta_0 x = \alpha$.

4. Find i det konkrete eksempel den bedste rette linie (i log-log figuren) med hældning $2/3$ og indtegn den.

   TIP: Gennemsnittet af værdierne af ln(legemsvægt) er 1.338 og gennemsnittet af værdierne af ln(hjernevægt) er 3.140.



Figur 5.3 Logaritmen til hjernevægt afsat mod logaritmen til legemsvægt.

Opgave 5.4 (Hydrolysering af urea i sedimenter)
Talmaterialet til denne opgave stammer fra en undersøgelse af sedimenter fra Norsminde Fjord, foretaget af Bente Lómstein, Institut for genetik og økologi, Århus Universitet.

Formålet med undersøgelsen var at bestemme den rate hvormed *urea* ($CO(NH_2)_2$) hydrolyseres til NH_4^+ og CO_2 i sedimentet fra fjorden. En del af undersøgelsen bestod i at man indsprøjtede en spormængde af radioaktivt mærket urea, $^{14}CO(NH_2)_2$, i et antal sedimentkerner, og derefter målte man til forskellige tidspunkter hvor meget $^{14}CO_2$ der udskiltes på det pågældende tidspunkt.

Der blev indsprøjtet $^{14}CO(NH_2)_2$ i 20 sedimentkerner, og efter henholdsvis 28.5, 72, 109 og 141 minutters forløb udtog man fem af disse kerner og målte den specifikke aktivitet af $^{14}CO_2$. Måleresultaterne ses i Tabel 5.10 hvor $^{14}CO_2$-aktiviteten angives i dpm (disintegrations per minute) pr. μl porevandsprøve.

I Tabel 5.11 er opgivet forskellige hjælpestørrelser, og Tabel 5.12 viser (dele af) et variansanalyseskema; indholdet af disse tabeller kan måske være til hjælp ved besvarelsen af nedenstående spørgsmål.

1. Lav en tegning der viser de faktiske måleresultater efter de forskellige antal minutters forløb.

2. Undersøg ved hjælp af en ensidet variansanalyse om der er signifikant forskel på den specifikke aktivitet efter de forskellige antal minutters for-

løb.

3. Man har en formodning om at den specifikke aktivitet afhænger lineært af tiden. – Estimér regressionslinien og indtegn den i figuren.

4. Da der er flere målinger til hvert tidspunkt, kan man udføre et numerisk test for om den specifikke aktivitet afhænger lineært af tiden. Gør det.

5. Hvor stor er middelfejlen på de estimerede parametre?

Tabel 5.10 Opgave 5.4: Den specifikke aktivitet i sedimentprøver efter forskellige antal minutters forløb.

tid x	specifik aktivitet y				
28.5	3.115	3.775	7.583	5.318	4.301
72.0	7.683	6.642	9.525	6.239	6.117
109.0	9.161	10.234	6.640	7.468	9.322
141.0	7.856	11.987	6.986	9.773	9.419

Tabel 5.11 Opgave 5.4: Nogle hjælpestørrelser.

i	$n_i x_i$	$n_i \overline{y}_i$	$n_i x_i \overline{y}_i$	$n_i x_i^2$	$\sum_j y_{ij}^2$	$\sum_j (y_{ij} - \overline{y}_i)^2$
1	142.5	24.092	686.622	4061.25	128.235464	12.150571
2	360.0	36.206	2606.832	25920.00	270.213088	8.038201
3	545.0	42.825	4667.925	59405.00	375.418985	8.622860
4	705.0	46.021	6488.961	99405.00	438.438191	14.851703
sum	1752.5	149.144	14450.340	188791.25	1212.305728	43.663335

Tabel 5.12 Opgave 5.4: Dele af et variansanalyseskema.

Variation	f	SS	s^2
inden for grupper	16	43.663335	2.73
mellem grupper	3	56.445756	18.82
total	19	100.109091	5.27
inden for grupper	16	43.663335	2.73
gruppernes variation omkring regressionslinien	2	2.261966	1.13
omkring regressionslinien	18	45.925301	2.55
regressionslinien	1	54.183790	54.18
total	19	100.109091	5.27

6 Multipel lineær regressionsanalyse

Ofte ønsker man at opbygge en regressionsmodel der inddrager mere end én forklarende variabel. Vi vil derfor nu betragte den situation hvor der for hvert af et antal »individer« foreligger dels en observation y, dels værdier $x_1, x_2, \ldots, x_p$ af p baggrundsvariable: Til individ nr. i hører observationen y_i og værdierne $x_{i1}, x_{i2}, \ldots, x_{ip}$ af de forklarende variable. Skematisk ser det sådan ud:

baggrundsvariable				observation
x_{11}	x_{12}	$\ldots$	x_{1p}	y_1
x_{21}	x_{22}	$\ldots$	x_{2p}	y_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\ldots$	x_{np}	y_n

Den statistiske model for y-erne indrettes på følgende måde:

- Tallene $y_1, y_2, \ldots, y_n$ er observerede værdier af de stokastiske variable $Y_1, Y_2, \ldots, Y_n$.
- De stokastiske variable $Y_1, Y_2, \ldots, Y_n$ er uafhængige og normalfordelte med samme varians σ^2.
- x-erne betragtes som faste tal – de er altså ikke observerede værdier af stokastiske variable.
- Middelværdien af den i-te måling kan skrives som $\alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p$, dvs. som en linearkombination af $p+1$ ukendte parametre $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ med koefficienterne $x_{i1}, x_{i2}, \ldots, x_{ip}$:

$$\mathrm{E}\, Y_i \quad = \quad \alpha + \sum_{j=1}^{p} x_{ij}\beta_j, \quad i = 1, 2, \ldots, n.$$

Af æstetiske grunde indfører man gerne en ekstra baggrundsvariabel x_0 der er lig med 1 for alle i, og samtidig kalder man α for β_0. Så kan man nemlig skrive $\alpha + \sum_{j=1}^{p} x_{ij}\beta_j$ som $\sum_{j=0}^{p} x_{ij}\beta_j$, og modellen kan kortfattet formuleres som

$$\mathrm{E}\, Y_i \quad = \quad \sum_{j=0}^{p} x_{ij}\beta_j$$

eller bedre

$$Y_i \sim \mathcal{N}(\sum_{j=0}^{p} x_{ij}\beta_j, \sigma^2). \tag{6.1}$$

Denne model er en såkaldt *multipel lineær regressionsmodel*. Den beskriver y-ernes *systematiske variation* ved hjælp af de $p+1$ parametre $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ plus de kendte konstanter x_{ij}, og den beskriver den tilfældige variation ved hjælp af normalfordelingen og variansparameteren σ^2.

6.1 Estimation af parametrene

Som altid estimeres modellens parametre ved at maksimalisere likelihoodfunktionen der i dette tilfælde er

$$\begin{aligned} & L(\beta_0, \beta_1, \beta_2, \ldots, \beta_p, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2 \right). \end{aligned}$$

Heraf ses at de bedste estimater over β-erne er dem der minimaliserer kvadratsummen

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij}\beta_j \right)^2.$$

De generelle metoder til minimalisering af funktioner af flere variable fortæller at minimumspunktet findes som det punkt hvor alle de $p+1$ partielle afledede (mht. de $p+1$ β-er) er lig 0. Hvis man skriver op hvad det betyder og omskriver en smule, når man frem til $p+1$ ligninger med de $p+1$ ubekendte $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.[1] Den j-te af disse såkaldte *estimationsligninger* er

$$a_{j0}\beta_0 + a_{j1}\beta_1 + a_{j2}\beta_2 + \ldots + a_{jp}\beta_p = \sum_{i=1}^{n} x_{ij} y_i$$

hvor

$$a_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik}, \qquad \begin{matrix} j = 0, 1, 2, \ldots, p \\ k = 0, 1, 2, \ldots, p \end{matrix}$$

Ved at løse de $p+1$ ligninger får man estimaterne $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$. – Ligningerne har »som oftest« netop én løsning. Undertiden er der uendelig mange løsninger; det er tilfældet hvis en af de forklarende variable er overflødig i den

[1] I matrix-notation kan disse ligninger skrives kort som $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$.

forstand at den ikke indeholder anden information end hvad der allerede er indeholdt i de øvrige.[2] I sådanne situationer plejer man at fjerne den eller de overflødige variable.

Sluttelig kan man udregne residualkvadratsummen

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j\right)^2$$

og variansestimatet

$$s_0^2 \quad = \quad \frac{1}{n-(p+1)}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j\right)^2 \tag{6.2}$$

der har $n-(p+1)$ frihedsgrader.

6.2 Modelkontrol

I tilfældet $p = 1$, dvs. simpel lineær regression, kan man kontrollere sin model ved hjælp af enkle tegninger. Det lader sig ikke gøre når p er større end 1, så der må man finde på andre metoder. Én ting der er fornuftig at foretage sig, er at udregne *residualerne*

$$e_i \quad = \quad y_i - \sum_{j=0}^{p} x_{ij}\hat{\beta}_j$$

og se hvordan de fordeler sig. Hvis modellen (6.1) er rigtig, er de teoretiske residualer $y_i - \sum_{j=0}^{p} x_{ij}\beta_j$ uafhængige $\mathcal{N}(0,\sigma^2)$-fordelte. Vi kender kun de empiriske residualer $e_1, e_2, \ldots, e_n$; det kan vises at hvis modellen er rigtig, så vil de empiriske residualer være $\mathcal{N}(0,\sigma^2)$-fordelte og næsten uafhængige[3]. Man kan derfor se efter om residualerne ser ud til at være nogenlunde uafhængige og normalfordelte.

I Afsnit 5.5 omtaltes et *numerisk test* for linearitetshypotesen. Dette test kunne udføres når der var flere y-værdier til hvert enkelt x således at man kunne indføre nogle grupper og bestemme en variation inden for grupper. Når der er tale om *multipel* regressionsanalyse kan man gøre noget tilsvarende, forudsat at der er flere y-værdier for hvert enkelt sæt værdier $(x_1, x_2, \ldots, x_p)$ af de forklarende variable. Denne forudsætning er sædvanligvis kun opfyldt hvis man har sørget for det ved planlægningen af forsøget.

Variansskønnet s_0^2

Variansskønnet s_0^2 fortæller ikke noget om hvor godt modellen passer, kun noget om hvor meget punkterne varierer omkring regressionsfladen; en stor

[2] Mere præcist gælder at ligningerne har en entydig løsning hvis og kun hvis det ikke er muligt at udtrykke nogen af de forklarende variable som en linearkombination af de øvrige.

[3] Jo flere frihedsgrader der er, jo mere uafhængige er de.

værdi af s_0^2 kan meget vel skyldes at der simpelthen er stor tilfældig variation på den slags y-målinger som man nu har med at gøre, modellens øvrige kvaliteter ufortalte.

Derimod kan det undertiden være fornuftigt at benytte størrelsen af s_0^2 som kriterium når man skal udvælge baggrundsvariable. Hvis der eksempelvis er 20 baggrundsvariable at vælge imellem, og man har besluttet sig for højst at ville have tre med i sin model, så kan det være fornuftigt at vælge de tre der giver den mindste s_0^2. Man bør dog også skele til om de tre der derved bliver udvalgt, virker som fornuftige baggrundsvariable i den givne sammenhæng.

Determinationskoefficienten R^2

Nogle brugere af regressionsanalyse er meget begejstrede for den såkaldte *determinationskoefficient* R^2 eller *kvadratet på den multiple korrelationskoefficient* der i en vis forstand udtaler sig om graden af overensstemmelse mellem de observerede værdier $y_1, y_2, \ldots, y_n$ og de fittede værdier $\hat{y}_i = \sum_{j=0}^{p} x_{ij}\hat{\beta}_j$.

Man kan udregne R^2 efter en af følgende to formler:[4]

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{y})\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2 \cdot \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}, \tag{6.3}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}. \tag{6.4}$$

Formel (6.3) fortæller at R^2 er kvadratet på korrelationskoefficienten mellem de observerede og de fittede værdier. Formel (6.4) fortæller at R^2 er et udtryk for hvor stor en del af den samlede variation omkring totalgennemsnittet der beskrives af modellen. Der er dem der mener at R^2 derfor også er et udtryk for hvor godt modellen passer, men prøv så at udregne R^2 i Opgave 5.1!

Bemærk at R^2 kun kan benyttes når der er et konstantled med i regressionen.

6.3 Udvælgelse af baggrundsvariable

Undertiden foreligger der et større sortiment af baggrundsvariable, og i første omgang kunne man måske fristes til at tro at jo flere baggrundsvariable man inddrager, jo bedre. Det er selvfølgelig rigtigt at jo flere baggrundsvariable man medtager, jo nøjagtigere et fit kan man få, men det er ikke nødvendigvis det der

[4] Det er ikke umiddelbart indlysende, men dog rigtigt, at de to udtryk giver samme resultat.

er meningen med at benytte en statistisk model. Formålet med at benytte statistiske modeller er at få en *reduktion* af data, og det vil blandt andet sige at man skal stræbe efter en statistisk model med væsentligt færre baggrundsvariable (og dermed parametre) end antallet af observationer. I det hele taget skal man holde sig det princip efterretteligt som går under navnet *Occams ragekniv*, og som siger at man ikke skal antage eksistensen af flere ting end nødvendigt.

Undertiden har man mange flere baggrundsvariable end man med rimelighed kunne tage med i modellen, og så er man stillet over for den opgave at udvælge en passende delmængde af dem. Det første kriterium må da være at man kun bør medtage variable der kan tænkes at have noget at gøre med den y-variabel der er tale om. Derudover skal man have fat i et sæt baggrundsvariable der gør s_0^2 forholdsvis lille. Bemærk i denne forbindelse at man i udtrykket for s_0^2 tager hensyn til antallet af baggrundsvariable (formel (6.2)).

Når man skal afgøre hvilke baggrundsvariable der måske kan undværes, kan man benytte sig af at man med et t-test for hver enkelt variabel kan vurdere om den tilsvarende parameter er signifikant forskellig fra 0, dvs. om variablen har en signifikant virkning. Antag f.eks. at man har en model med p baggrundsvariable plus en konstant, og at man ønsker at undersøge om variabel nr. k behøver være med i modellen. Så udregner man

$$t \quad = \quad \frac{\hat{\beta}_k}{\text{estimeret middelfejl på } \hat{\beta}_k}$$

og sammenholder resultatet med t-fordelingen med $n - (p + 1)$ frihedsgrader (= antal frihedsgrader for s_0^2). Hvis t er tæt på nul, vil man acceptere hypotesen om at β_k er nul, og det betyder at man kan se bort fra baggrundsvariabel nr. k og altså gå videre med en reduceret model med kun $p - 1$ baggrundsvariable; hvis t er langt fra nul er $\hat{\beta}_k$ signifikant forskellig fra 0, dvs. baggrundsvariabel nr. k har en signifikant virkning og skal derfor forblive i modellen.

Eksempel 6.1 (Indianere i Peru)
Ændringer i menneskers livsbetingelser kan give sig udslag i fysiologiske ændringer, eksempelvis i ændret blodtryk.

En gruppe antropologer har undersøgt hvordan blodtrykket ændrer sig hos peruvianske indianere der flyttes fra deres oprindelige primitive samfund i de høje Andesbjerge til den såkaldte civilisation, dvs. storbyen, der i øvrigt ligger i langt mindre højde over havets overflade end deres oprindelig bopæl. Antropologerne udvalgte en stikprøve på 39 mænd over 21 år der havde undergået en sådan flytning. På hver af disse måltes blodtrykket (både det systoliske og det diastoliske) samt en række baggrundsvariable, heriblandt alder, antal år siden flytningen, højde, vægt og puls. Som om det ikke kunne være nok har man udregnet endnu en baggrundsvariabel, nemlig »brøkdel af livet levet i de nye omgivelser«, dvs. antal år siden flytning divideret med nuværende alder. Man forestillede sig at denne baggrundsvariabel kunne have stor »forklaringsevne«.

Her vil vi ikke se på hele talmaterialet, men kun på *blodtrykket* (det systoliske) der skal optræde som y-variabel, og på de to x-variable *brøkdel af livet i de nye omgivelser* og *vægt*. Disse er angivet i Tabel 6.1.

Antropologerne mente at x_1 (brøkdel levet i de nye omgivelser) var et godt mål for hvor længe personerne havde levet i de civiliserede omgivelser, og at det derfor måtte være interessant at se hvor godt x_1 kunne forklare blodtrykket y. Første skridt er derfor

Tabel 6.1 Indianere i Peru: Sammenhørende værdier af y: systolisk blodtryk (mm Hg), x_1: brøkdel af livet i de nye omgivelser, og x_2: vægt (kg).

y	x_1	x_2	y	x_1	x_2
170	0.048	71.0	114	0.474	59.5
120	0.273	56.5	136	0.289	61.0
125	0.208	56.0	126	0.289	57.0
148	0.042	61.0	124	0.538	57.5
140	0.040	65.0	128	0.615	74.0
106	0.704	62.0	134	0.359	72.0
120	0.179	53.0	112	0.610	62.5
108	0.893	53.0	128	0.780	68.0
124	0.194	65.0	134	0.122	63.4
134	0.406	57.0	128	0.286	68.0
116	0.394	66.5	140	0.581	69.0
114	0.303	59.1	138	0.605	73.0
130	0.441	64.0	118	0.233	64.0
118	0.514	69.5	110	0.432	65.0
138	0.057	64.0	142	0.409	71.0
134	0.333	56.5	134	0.222	60.2
120	0.417	57.0	116	0.021	55.0
120	0.432	55.0	132	0.860	70.0
114	0.459	57.0	152	0.741	87.0
124	0.263	58.0			

at fitte en simpel lineær regressionsmodel med x_1 som forklarende variabel. Man finder den estimerede regressionslinie til

$$y = 134 - 16x_1$$

og det tilhørende variansestimat er 163 med 37 frihedsgrader.

Hvis man i et koordinatsystem afsætter y mod x_1, viser det sig imidlertid, se Figur 6.1, at det bestemt ikke virker særlig rimeligt at hævde at (middelværdien af) y afhænger lineært af x_1. Derfor må man give sig til at overveje om andre af de målte baggrundsvariable med fordel kan inddrages.

Nu ved man at en persons *vægt* har betydning for den pågældendes blodtryk, så næste modelforslag er en multipel regressionsmodel med både x_1 og x_2 som forklarende variable. Estimaterne over parametrene β_0, β_1 og β_2 i regressionsligningen

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2$$

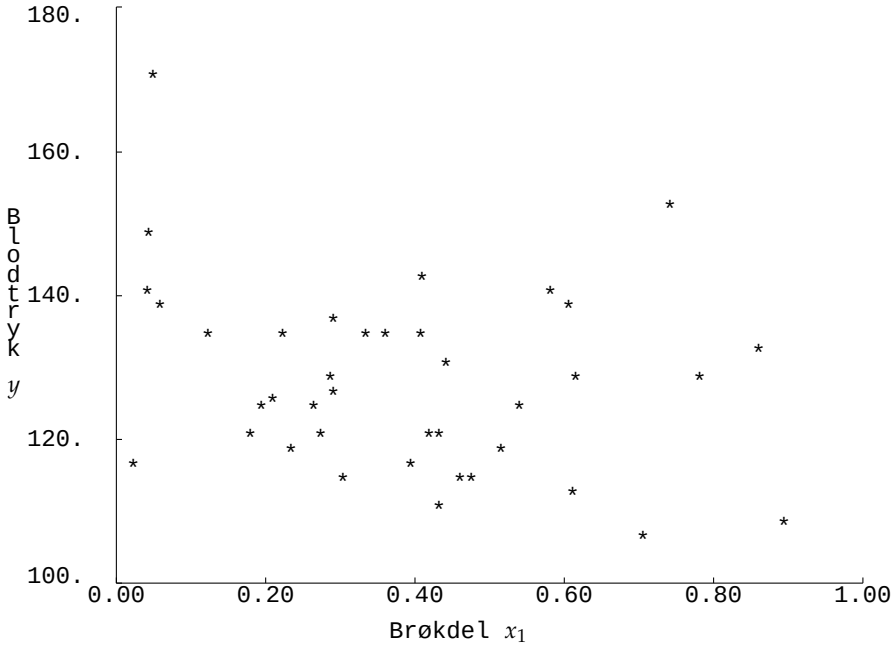
bestemmes som løsning til estimationsligningerne (jf. side 88)

$$\begin{array}{rcrcrcr} 39\,\beta_0 & + & 15.066\,\beta_1 & + & 2463.20\,\beta_2 & = & 4969 \\ 15.066\,\beta_0 & + & 7.826896\,\beta_1 & + & 969.7395\,\beta_2 & = & 1887.944 \\ 2463.20\,\beta_0 & + & 969.7395\,\beta_1 & + & 157488.16\,\beta_2 & = & 315680.8 \end{array}$$

Man finder at $\hat{\beta}_0 = 60.8775$, $\hat{\beta}_1 = -26.78738$ og $\hat{\beta}_2 = 1.21726$, så den estimerede regressionsligning er

$$y = 61 - 27x_1 + 1.2x_2$$

og variansestimatet bliver denne gang 96 med 36 frihedsgrader.



Figur 6.1 Indianere i Peru: Blodtrykket y afsat mod brøkdel af livet siden flytning x_1.

Det ses at ved at inddrage x_2 er variansen gået drastisk ned, fra 163 til 96. Deraf kan man dog ikke slutte at den nye regressionsligning giver en *god* beskrivelse af data, kun at den er bedre end den forrige. Man bør undersøge residualerne for at kunne vurdere modellens kvalitet – det vil vi dog ikke gøre her.

Hvis man lader en computer foretage udregningerne, vil man sandsynligvis også få oplyst parameterestimaternes middelfejl og få at vide om parametrene hver især er signifikant forskellige fra 0. I det konkrete tilfælde vil man da få at vide at når man kun bruger x_1, så er koefficienten til x_1 *ikke* signifikant forskellig fra 0, men når man benytter både x_1 og x_2, så er alle koefficienter signifikant forskellige fra 0. Det kan man fortolke på den måde at blodtrykket afhænger signifikant af både x_1 og x_2 således at jo længere man har levet i de nye omgivelser jo lavere blodtryk, og jo større vægt jo højere blodtryk; *men* da det nok også er sådan at jo længere tid man har boet i »civilisationen«, jo mere vejer man, så vil de to virkninger udjævne hinanden hvis man ikke sørger for at inddrage begge forklarende variable.

Tabel 6.2 Opgave 6.1: Diameter d (i *inches*), højde h (i *feet*) og rumfang v (i kubik*feet*) for 31 sortkirsebærtræer.

d	h	v	d	h	v
8.3	70	10.3	12.9	85	33.8
8.6	65	10.3	13.3	86	27.4
8.8	63	10.2	13.7	71	25.7
10.5	72	16.4	13.8	64	24.9
10.7	81	18.8	14.0	78	34.5
10.8	83	19.7	14.2	80	31.7
11.0	66	15.6	14.5	74	36.3
11.0	75	18.2	16.0	72	38.3
11.1	80	22.6	16.3	77	42.6
11.2	75	19.9	17.3	81	55.4
11.3	79	24.2	17.5	82	55.7
11.4	76	21.0	17.9	80	57.3
11.4	76	21.4	18.0	80	51.5
11.7	69	21.3	18.0	80	51.0
12.0	75	19.1	20.6	87	77.0
12.9	74	22.2			

6.4 Opgaver

Opgave 6.1 (Træers rumfang)
Inden for skovbruget er man interesseret i at kunne vurdere et træs indhold af tømmer, dvs. dets *rumfang*, uden alt for stort besvær. Nogle størrelser der er nemme at bestemme, er *diameter* og *højde*, og det ville være praktisk hvis man kunne forudsige et træs rumfang så nogenlunde ud fra disse to størrelser.

Man har derfor målt diameteren d (i en højde af 4.5 feet over jorden), højden h og rumfanget (volumenet) v for 31 træer af en bestemt slags (sortkirsebærtræer i Allegheny National Forest, Pennsylvania). Resultaterne er vist i Tabel 6.2.

Opgaven er nu at undersøge, *om* man med en simpel statistisk model kan bestemme v ud fra kendskab til d og h, og i givet fald *hvordan* og *hvor godt*.
TIP: Der er mulighed for forskellige regressionsanalyser. Man kan også prøve at udnytte at rumfang er noget med højde gange tværsnitsareal.

Opgave 6.2 (Vands strømningsforhold i en flod)
I forbindelse med en undersøgelse af vands strømningsforhold i en flod har man på et bestemt sted målt flowraten i forskellige dybder. Flowraten er den mængde vand der passerer et givet tværsnit af floden i et givet tidsrum (så den måles altså i f.eks. m^3 pr. m^2 pr. sekund). Tabel 6.3 viser sammenhørende værdier af vanddybde og flowrate.

Opgaven er at give en simpel beskrivelse af sammenhængen mellem flowrate og vanddybde.[¶]

[¶]Hydrologer kan sikkert opstille fornemme differentialligningsmodeller der beskriver denne

Tabel 6.3 Opgave 6.2: Flowraten i forskellige vanddybder.

dybde	flowrate
0.34	0.636
0.29	0.319
0.28	0.734
0.42	1.327
0.29	0.487
0.41	0.924
0.76	7.350
0.73	5.890
0.46	1.979
0.40	1.124

1. Lav et scatterplot af flowrate mod dybde. Ser punkterne ud til at ligge på en ret linie?

2. Beregn den bedste rette linie og indtegn den (det er altid lettere at vurdere om punkter ligger omkring en bestemt kurve når man har kurve og punkter i samme tegning).

3. Man kunne forestille sig at en *andengradskurve* ville give en bedre beskrivelse af punkterne. Opstil og løs de estimationsligninger der bestemmer den bedste andengradskurve.

   TIP: Dvs. foretag en multipel regression med de to forklarende variable $x_1 =$ dybde og $x_2 =$ dybde^2.

   Er andengradskurven bedre end den rette linie? Hvorfor?

4. Hvad er konklusionen mht. sammenhængen mellem flowrate og vanddybde?

sammenhæng, forudsat at flodens sider og bund ikke er alt for uregelmæssige. Det er slet ikke det vi er ude efter her. Statistikeren vil blot søge efter en simpel beskrivelse af de empiriske data.

7 Stikord

8 De »manglende« figurer

På grund af tekniske vanskeligheder kan visse figurer ikke gengives i den rigtige størrelse inde i teksten (og de gengives så slet ikke). Her kan man se figurerne uskaleret (klik på »Her er«):

Her er Figur 2.1 der hører til på side 22. Figurteksten er: Histogram over 64 målte værdier af lysets passagetid. – Den indtegnede kurve er tætheden for normalfordelingen med parametre $\overline{y} = 27.75$ og $s^2 = 25.8$.

Her er Figur 2.2 der hører til på side 23. Figurteksten er: Fraktildiagram over 64 målte værdier af lysets passagetid.

Her er Figur 5.2 der hører til på side 75. Figurteksten er: Kvælning af hunde: Sammenhørende værdier af hypoxantinkoncentration og hypoxivarighed, samt den estimerede regressionslinie.