TEKST NR 435 2004

BASISSTATISTIK

Jørgen Larsen

2004, 2005

TEKSTER fra

IMFUFA ROSKILDE UNIVERSITETSCENTER
INSTITUT FOR STUDIET AF MATEMATIK OG FYSIK SAMT DERES FUNKTIONER I UNDERVISNING, FORSKNING OG ANVENDELSER

IMFUFA
Roskilde Universitetscenter
Postboks 260
DK-4000 Roskilde

t +45 46 74 22 63
f +45 46 74 30 20
m imfufa@ruc.dk
w imfufa.ruc.dk

Jørgen Larsen: BASISSTATISTIK

IMFUFA tekst nr. 435/2004 231 sider issn 0106-6242
Erstatter IMFUFA tekst nr. 304/1995

Dette hæfte, som er en let revideret udgave af IMFUFA tekst 304 a-d, er undervisningsmateriale til et indledende kursus i statistik og statistiske modeller.

Om kurset og kursusmaterialet kan blandt andet siges at

- når det som et gennemgående tema påpeges 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 materialet er skrevet på dansk, er det for at bidrage til at vedligeholde traditionerne for *hvordan* og *at* man kan tale om slige emner på dansk, og også fordi dansk er det sprog som forfatteren – og vel også den forventede læser – er bedst til;
- når der foruden de sædvanlige simple modeller, metoder og eksempler også præsenteres 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;
- den væsentligste indholdsmæssige forskel fra de tidligere udgaver af kursusmaterialet er at der i den her foreliggende udgave er indføjet små afsnit der viser hvordan man kan benytte programmeringssproget R (se `http://www.r-project.org/`) til at udføre tegninger og regninger.

September 2005: rettet nogle fejl og foretaget adskillige typografiske justeringer, samt tilføjet tabeller og referencer..

Indhold

Indledning

Dette hæfte beskæftiger sig med simple eksempler på statistiske modeller. Statistiske modeller er en særlig type matematiske modeller som bruges for at beskrive talmaterialer som er behæftet med en eller anden form for tilfældig variation. De statistiske modellers force er at de kan bruges til at skille det systematiske fra det tilfældige. Der melder sig forskellige slags spørgsmål i forbindelse med statistiske modeller:

- hvordan ser modellerne ud, dvs. hvad er det for nogle matematiske ingredienser der indgår?
- hvordan finder man på en model der kan bruges i en given situation?
- hvad stiller man så op med modellen i forhold til de konkrete tal?
- hvad er det for typer af »spørgsmål« man kan stille til en statistisk model, og hvad er det for typer af »svar« man får?

Disse spørgsmål diskuteres indgående. Fremstillingen er baseret på *likelihood-metoden* hvis grundlæggende idéer præsenteres omhyggeligt; derimod må vi af tekniske grunde give afkald på de matematiske beviser for metodens fortræffeligheder.

Der præsenteres nogle af de simple og klassiske modeller for blandt andet binomialfordelte, poissonfordelte og normalfordelte observationer, men der er også eksempler på mere komplicerede modeller så som logistisk regression og multiplikative poissonmodeller.

Når man beskæftiger sig med statistik og statistiske metoder, har man brug for hensigtsmæssige regne- og tegneredskaber. Mange af de grundlæggende modeller kan uden vanskeligheder analyseres med en almindelig lommeregner som regneredskab og ternet papir til tegninger, men så snart modellerne bliver lidt mere indviklede, er det en fordel at benytte en computer med et statistikprogram.

I nærværende fremstilling er indføjet små afsnit med overskriften »Regn og tegn« der viser hvordan man kan udføre beregningerne og tegninger med programmet R. R er et 'freeware' program der kan hentes på `http://mirrors.sunsite.dk/cran/`

Teksten giver ikke nogen egentlig nærmere præsentation af R, den bedste måde at lære det på er formentlig ved en kombination af at se hvordan andre har gjort, selv prøve sig frem, og udnytte on-line hjælpen (som er relativt god). Det er dog nok en god idé med en ultrakort introduktion; en sådan gives på næste side.

Kort om statistikprogrammet R

I R opereres på *datastrukturer* så som vektorer og matricer. Det der opererer, er forskellige *funktioner*. Hvis man f.eks. har oprettet en datastruktur `tal` som indeholder en masse tal, så kan man få udregnet summen af alle disse tal ved at anvende funktionen `sum` på `tal`, dvs. man skriver `sum(tal)`. Hvis man vil have den naturlige logaritme til alle tallene, skriver man `log(tal)`.

R er primært tænkt som et interaktivt system, dvs. brugeren indtaster en kommando som R så udfører, dernæst indtaster brugeren en ny kommando osv. Kommandoer kan være instruktioner om at udføre ganske enkle regnestykker, f.eks. at lægge tallene 5 og 7 sammen (det gøres ved at skrive `5+7`), eller de kan udføre komplicerede beregninger og tegninger.

Her er nogle ting der er værd at vide fra starten:

- R skriver prompten `>` når det er parat til at modtage en kommando.
- Man afslutter R ved at kalde funktionen `q`, dvs. man skriver `q()`
- Der er forskel på store og små bogstaver.
- Decimaltal skrives med tegnet `.` (punktum) som »komma«, f.eks. `3.75`
- »Sættes lig med«-operatoren er `<-` (ɔ: tegnet `<` umiddelbart efterfulgt af tegnet `-`). Eksempel: Hvis man skriver `a <- 8.5` så bliver der oprettet en datastruktur `a` som indeholder det ene tal 8.5
- Funktionen `c` sammenkæder til en vektor.
  Eksempel: hvis man skriver `b <- c(2, 4.5, 5)` så bliver der oprettet en vektor `b` som indeholder de tre værdier 2, 4.5 og 5.
  Man får den aktuelle værdi af `b` at se ved at skrive `b` ved R-prompten.
- Funktionen `ls` skriver en liste over hvilke datastrukturer der aktuelt er defineret; man skriver `ls()`
- Der er (naturligvis) almindelige operatorer som `+` (addition), `-` (subtraktion), `*` (multiplikation), `/` (division), `^` (potensopløftning).
  En anden nyttig operator er `:` der leverer en vektor af på hinanden følgende heltal, f.eks. giver `5:12` resultatet `5 6 7 8 9 10 11 12`
- Hjælp fås med en af funktionerne `?` og `help`.
  Eksempler:
  Man kan få hjælp om funktionen `sin` ved enten at skrive `?sin` eller `help(sin)`
  Man kan få hjælp om funktionen `:` ved at skrive `?":"` eller `help(":")`
- Med R-distributionen følger udmærkede hjælpetekster i HTML-format; i Windows-udgaven af R kan man åbne disse hjælpetekster fra R-konsollens Help-menu.
- Symbolet `#` er et kommentartegn, dvs. resten af linjen (fra og med `#`) bliver ignoreret.

1 Binomialfordelingen

Binomialfordelingsmodeller kan komme på tale i situationer der har følgende grundstruktur:

- Man har et bestemt *elementarforsøg* der kan resultere i et af to mulige udfald som vi kalder 1 og 0 (eller Gunstig/Ikke-gunstig eller Succes/Fiasko).
- Det er bestemt af tilfældigheder om elementarforsøget giver det ene eller det andet udfald.
- Man udfører n gentagelser af elementarforsøget, hvor n er et på forhånd fastlagt tal, og man tæller op hvor mange af de n gentagelser der giver udfaldet 1.
- Resultatet bliver et antal y der i sagens natur er et heltal mellem 0 og n; de forskellige mulige værdier vil indtræffe med visse sandsynligheder der afhænger af tilfældighedsmekanismens nærmere indretning.
- Det samlede forsøg, altså det som består af de n elementarforsøg og som resulterer i antallet y, kaldes et *binomialforsøg*.

Her er et eksempel (hentet fra (15)) som vi vil bruge flere gange: I en undersøgelse af insekters reaktion på insektgiften pyrethrum har man udsat nogle rismelsbiller (*Tribolium castaneum*) for forskellige mængder gift og derpå set hvor mange der var døde efter 13 dages forløb. Blandt andet blev 144 han-biller udsat for en giftpåvirkning på 0.20 mg/cm^2; af disse døde de 43 i løbet af den fastsatte periode. Her kan vi sige at et elementarforsøg består i at udsætte én han-bille for påvirkningen 0.20 mg/cm^2 og så se efter om den er død eller ej efter 13 dage (dvs. »død« $\sim 1 \sim$ »Gunstigt udfald«).

Vi vil opstille en matematisk model for den beskrevne situation. Vi deler ræsonnementet op i en række punkter:

1. For hvert elementarforsøg indfører vi en såkaldt *indikatorvariabel* X der angiver om forsøget giver et 0 eller et 1. Indikatorvariablen hørende til elementarforsøg nr. j er X_j:
$$X_j = \begin{cases} 1 & \text{hvis bille nr. } j \text{ dør} \\ 0 & \text{hvis bille nr. } j \text{ ikke dør} \end{cases}$$
2. Det samlede antal døde biller kan da skrives som $Y = X_1 + X_2 + \cdots + X_n$. I eksemplet kender vi ikke de enkelte X_j-er, men kun Y; Y har værdien $y = 43$.
3. Indikatorvariablene $X_1, X_2, \ldots, X_n$ er *stokastiske variable*. (En stokastisk variabel er kort fortalt et symbol der repræsenterer det tilfældige udfald af et bestemt tilfældighedseksperiment.) Om X_j-erne antages det at

a) de alle har den samme sandsynlighed p for at antage værdien 1, altså at $\mathrm{P}(X_j = 1) = p$ for ethvert j,
b) de er stokastisk uafhængige, dvs. $\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) = \mathrm{P}(X_1 = x_1) \cdot \mathrm{P}(X_2 = x_2) \cdot \ldots \cdot \mathrm{P}(X_n = x_n)$ for vilkårlige $x_1, x_2, \ldots, x_n$.

Da X_j kun kan antage værdierne 0 og 1, og da summen af sandsynlighederne er 1, er $\mathrm{P}(X_j = 0) = 1 - p$ for ethvert j.

4. Vi kan skrive *sandsynlighedsfunktionen** for X_j som

$$\mathrm{P}(X_j = x) = \begin{cases} p & \text{hvis } x = 1 \\ 1 - p & \text{hvis } x = 0 \end{cases}$$

eller kortere

$$\mathrm{P}(X_j = x) = p^x \, (1-p)^{1-x}, \quad x \in \{0, 1\}.$$

5. Den såkaldte *simultane* sandsynlighedsfunktion for $X_1, X_2, \ldots, X_n$ er den funktion $f(x_1, x_2, \ldots, x_n)$ der angiver sandsynligheden for at der samtidigt gælder at $X_1 = x_1$ og $X_2 = x_2$ og … og $X_n = x_n$.

Da X_j-erne er stokastisk uafhængige, er den simultane sandsynlighedsfunktion for $X_1, X_2, \ldots, X_n$ produktet af de enkelte sandsynlighedsfunktioner:

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= \mathrm{P}(X_1 = x_1) \cdot \mathrm{P}(X_2 = x_2) \cdot \ldots \cdot \mathrm{P}(X_n = x_n) \\ &= p^{x_1}(1-p)^{1-x_1} \cdot p^{x_2}(1-p)^{1-x_2} \cdot \ldots \cdot p^{x_n}(1-p)^{1-x_n} \\ &= p^{x_1+x_2+\cdots+x_n} \, (1-p)^{n-(x_1+x_2+\cdots+x_n)} \end{aligned}$$

når $(x_1, x_2, \ldots, x_n)$ er et talsæt bestående af 0-er og 1-er. Det ses at hvis der i talsættet $(x_1, x_2, \ldots, x_n)$ er netop y 1-er og $n - y$ 0-er, så er

$$f(x_1, x_2, \ldots, x_n) = p^y \, (1-p)^{n-y}.$$

6. Da vi nu kender den simultane sandsynlighedsfunktion for X_j-erne, kan vi bestemme sandsynlighedsfunktionen for $Y = X_1 + X_2 + \cdots + X_n$. Sandsynligheden for at Y er lig med y, kan findes ved at summere sandsynlighederne for alle de sæt af n elementarforsøg som består af præcis y 1-udfald og $n - y$ 0-udfald:

$$\mathrm{P}(Y = y) = \sum_{x_1+x_2+\cdots+x_n=y} f(x_1, x_2, \ldots, x_n)$$

hvor meningen er at der summeres over alle talsæt $(x_1, x_2, \ldots, x_n)$ bestående af 0-er og 1-er og for hvilke $x_1 + x_2 + \cdots + x_n = y$ (dvs. hvor der er netop y 1-er og $n - y$ 0-er). Som vi netop er nået frem til, har ethvert af disse talsæt sandsynlighed $p^y(1-p)^{n-y}$, så derfor bliver

$$\mathrm{P}(Y = y) = A \cdot p^y(1-p)^{n-y}$$

hvor A står for »antallet af forskellige talsæt $(x_1, x_2, \ldots, x_n)$ bestående af y 1-er og $n - y$ 0-er«.

---

* Generelt er sandsynlighedsfunktionen for en stokastisk variabel X den funktion der til hvert tal x knytter sandsynligheden for at X antager værdien x.

Tabel 1.1 Her ses 15 eksempler på udfald af 01-variable $X_1, X_2, \ldots, X_{12}$, frembragt af en tilfældighedsmekanisme med $p = 1/3$, samt de tilsvarende værdier af $Y = X_1 + X_2 + \cdots + X_{12}$. Tallene i y-søjlen er således 15 observationer fra en binomialfordeling med $n = 12$ og $p = 1/3$.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	y
1	0	1	0	0	0	0	1	0	0	1	1	5
0	0	1	0	1	0	0	0	1	0	1	1	5
0	1	0	1	0	0	0	0	1	1	0	1	5
0	0	0	1	0	1	0	1	0	0	1	0	4
1	0	0	0	0	1	0	0	1	0	0	0	3
0	1	0	0	0	1	0	0	0	0	0	0	2
0	0	0	1	0	0	0	0	0	0	1	1	3
0	0	1	0	0	1	1	1	1	0	0	1	6
0	0	0	1	0	0	0	0	0	1	0	0	2
0	0	0	1	0	0	1	1	0	0	0	0	3
0	1	0	1	0	0	0	0	1	0	0	0	3
1	1	0	0	1	0	0	1	1	0	0	1	6
0	0	0	1	1	0	0	1	1	0	0	0	4
0	1	0	0	1	0	0	0	0	0	0	0	2
0	1	0	0	0	1	0	0	1	1	1	1	6

7. Antallet A af forskellige talsæt $(x_1, x_2, \ldots, x_n)$ bestående af y 1-er og $n - y$ 0-er afhænger af værdierne af n og y; man plejer at betegne det med symbolet $\binom{n}{y}$ (udtales »n over y«). Størrelsen $\binom{n}{y}$ kaldes en *binomialkoefficient*.

8. Alt i alt er vi dermed nået frem til at sandsynlighedsfunktionen for Y er

$$\mathrm{P}(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, 2, \ldots, n\}.$$

Den fundne sandsynlighedsfordeling for Y hedder *binomialfordelingen med sandsynlighedsparameter p og antalsparameter n*, og man siger at Y er *binomialfordelt* med parametre n og p. – Antalsparameteren n er et kendt heltal, og sandsynlighedsparameteren p, som typisk er ukendt, er et tal mellem 0 og 1.

Stokastiske variable der som X_j-erne kun kan antage værdierne 0 og 1, kaldes undertiden for *01-variable*. Der gælder altså at *hvis Y er en sum af et bestemt antal uafhængige identisk fordelte 01-variable, så er Y binomialfordelt.*

Den statistiske model for bille-forsøget kan nu kort formuleres således:

> Observationen $y = 43$ er en observeret værdi af en stokastisk variabel Y som er binomialfordelt med antalsparameter $n = 144$ og ukendt sandsynlighedsparameter $p \in [0, 1]$.

Før vi kan give os i kast med statistisk analyse af binomialfordelte observationer, er det nødvendigt at lære forskelligt om binomialfordelingen og om binomialkoefficienter.

1.1 Binomialkoefficienter

Definition 1.1: Binomialkoefficient
Binomialkoefficienten $\binom{n}{k}$ *er et symbol der betegner antallet af forskellige måder hvorpå man kan placere to symboler* 1 *og* 0 *på* n *pladser således at symbolet* 1 *kommer på* k *af pladserne og symbolet* 0 *kommer på de resterende* $n-k$ *pladser.*

Deraf følger at der er $\binom{n}{k}$ forskellige talsæt $(x_1, x_2, \ldots, x_n)$ bestående af netop k 1-er og $n-k$ 0-er.

Ud fra definitionen kan man i princippet bestemme talværdier af enhver binomialkoefficient ved simpel optælling, eksempelvis er $\binom{4}{3}$ lig med 4 fordi der er de fire placeringer $(1,1,1,0)$, $(1,1,0,1)$, $(1,0,1,1)$ og $(0,1,1,1)$ af tre 1-er og et 0 på de fire pladser.

Hvis man overhovedet skulle komme ud for i praksis at skulle udregne binomialkoefficienter, er optællingsmetoden dog ikke særlig hensigtsmæssig (prøv f.eks. at bestemme $\binom{37}{15}$ ved denne metode); vi vil på de næste par sider udlede nogle formler der kan gøre beregningsarbejdet lidt mere overkommeligt.

Hvis k er 0 eller 1 (eller n eller $n-1$), er det let at udregne $\binom{n}{k}$; fra definitionen og fra formel (1.1) får man†

$$\binom{n}{0} = 1 \quad \text{og dermed} \quad \binom{n}{n} = 1, \qquad \text{for } n = 0, 1, 2, \ldots$$

$$\binom{n}{1} = n \quad \text{og dermed} \quad \binom{n}{n-1} = n, \quad \text{for } n = 1, 2, 3, \ldots$$

I definitionen af $\binom{n}{k}$ skal man placere k 1-er og $n-k$ 0-er. Hvis man i en sådan placering kalder 1-erne for 0 og 0-erne for 1, så får vi i stedet en placering af $n-k$ 1-er og k 0-er. Heraf følger at

$$\binom{n}{k} = \binom{n}{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n \text{ og } n = 0, 1, 2, \ldots \tag{1.1}$$

De forskellige placeringer af k 1-er og $n-k$ 0-er kan opdeles i to grupper:

1. Placeringer der har et 1 på sidstepladsen. På de første $n-1$ pladser er der da netop $k-1$ 1-er, og de kan placeres på $\binom{n-1}{k-1}$ forskellige måder. Denne gruppe omfatter derfor $\binom{n-1}{k-1}$ forskellige placeringer.
2. Placeringer der har et 0 på sidstepladsen. På de første $n-1$ pladser er der da netop k 1-er, og de kan placeres på $\binom{n-1}{k}$ forskellige måder. Denne gruppe omfatter derfor $\binom{n-1}{k}$ forskellige placeringer.

Det samlede antal er lig summen af de to; dermed er vist at

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1} \quad \text{for } k = 1, 2, 3, \ldots, n \text{ og } n = 1, 2, 3, \ldots \tag{1.2}$$

† Det er dog i nogen grad en konvention at $\binom{0}{0}$ skal være 1.

n	binomialkoefficienterne $\binom{n}{k}$
0	1
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
6	1 6 15 20 15 6 1
7	1 7 21 35 35 21 7 1
⋮	⋮

Figur 1.1 Pascals trekant.

Eksempel

Som illustration bestemmes talværdien af $\binom{5}{2}$.

- Ifølge formel (1.2) er $\binom{5}{2} = \binom{4}{2} + \binom{4}{1}$, så hvis vi kender talværdierne af $\binom{4}{2}$ og $\binom{4}{1}$, kan vi løse opgaven.
- Der gælder at $\binom{4}{1} = 4$ (fordi generelt er $\binom{n}{1} = n$).
- For at udregne $\binom{4}{2}$ benytter vi formel (1.2) en gang til: $\binom{4}{2} = \binom{3}{2} + \binom{3}{1}$.
  - Der gælder at $\binom{3}{1} = 3$.
  - Der gælder også at $\binom{3}{2} = 3$ (fordi $\binom{n}{n-1} = n$).

  Dermed er $\binom{4}{2} = 3 + 3 = 6$.

Dermed er $\binom{5}{2} = \binom{4}{2} + \binom{4}{1} = 6 + 4 = 10$ – hvad man jo også kan se ved simpel optælling.

Pascals trekant

Formel (1.2) er ikke særlig velegnet når man ønsker at beregne en enkelt binomialkoefficient, men den er overordentlig praktisk hvis man ønsker at beregne alle binomialkoefficienter op til en eller anden øvre grænse for n.

Vi kender på forhånd binomialkoefficienterne med $n = 0$ og $n = 1$ (de er $\binom{0}{0} = 1$ og $\binom{1}{0} = \binom{1}{1} = 1$). Ved hjælp af formel (1.2) kan vi beregne alle koefficienter med $n = 2$, derefter alle med $n = 3$, derefter alle med $n = 4$, osv. Man plejer at stille resultaterne op i et skema der kaldes *Pascals trekant*‡, se figur 1.1. Heraf ses at f.eks. er $\binom{7}{2}$ lig 21. Hvert tal i Pascals trekant fremkommer, ifølge formel (1.2), som summen af de to nærmeste tal i rækken lige ovenover, f.eks. er $21 = 6 + 15$.

Endnu en formel

Ved brug af Pascals trekant vil det være muligt at bestemme talværdier af enhver binomialkoefficient; man skulle dog udføre en hel del additioner og have et temmelig stort ark papir for at udregne f.eks. $\binom{37}{15}$. Heldigvis findes der også en anden og mindre pladskrævende metode hvor man så til gengæld skal lave nogle multiplikationer

‡ Pascals trekant er opkaldt efter den franske videnskabsmand og tænker Blaise Pascal (1623-62).

og divisioner. Som forberedelse til denne metode skal vi bruge endnu en formel for binomialkoefficienter.

Antag igen at vi skal fordele k 1-er og $n-k$ 0-er på n pladser, men nu er et af 1-erne mærket. Vi vil bestemme antallet af synligt forskellige placeringer. Det kan regnes ud på to måder:

1. Bestem først hvilke pladser der skal have et 0: Det kan gøres på $\binom{n}{n-k} = \binom{n}{k}$ måder. Nu er der k pladser reserveret til 1-er, og der er derfor k forskellige måder at placere det mærkede 1 på. I alt er der derfor $k \cdot \binom{n}{k}$ synligt forskellige placeringer.
2. Bestem først hvilke pladser der skal have et umærket 1. Det kan gøres på $\binom{n}{k-1}$ måder. Derefter kan det mærkede 1 placeres på en af de $n-k+1$ resterende pladser. I alt er der derfor $(n-k+1) \cdot \binom{n}{k-1}$ synligt forskellige placeringer.

Da de to antal jo er ens, er $k \cdot \binom{n}{k} = (n-k+1) \cdot \binom{n}{k-1}$, og ved at flytte rundt på faktorerne fås

$$\binom{n}{k} = \frac{n-k+1}{k} \cdot \binom{n}{k-1} \quad \text{for} \quad \begin{matrix} k = 1, 2, \ldots, n \\ n = 1, 2, \ldots \end{matrix} \tag{1.3}$$

Denne formel fortæller hvordan man finder $\binom{n}{k}$ hvis man kender $\binom{n}{k-1}$.

Ved gentagne anvendelser af formel (1.3) fås i øvrigt

$$\begin{aligned} \binom{n}{k} &= \frac{n-k+1}{k} \cdot \binom{n}{k-1} \\ &= \frac{n-k+1}{k} \cdot \frac{n-k+2}{k-1} \cdot \binom{n}{k-2} \\ &= \frac{n-k+1}{k} \cdot \frac{n-k+2}{k-1} \cdot \frac{n-k+3}{k-2} \cdot \binom{n}{k-3} \\ &= \ldots \\ &= \frac{n-k+1}{k} \cdot \frac{n-k+2}{k-1} \cdot \ldots \cdot \frac{n-2}{3} \cdot \frac{n-1}{2} \cdot \frac{n}{1}, \end{aligned}$$

dvs.

$$\binom{n}{k} = \frac{n}{1} \cdot \frac{n-1}{2} \cdot \frac{n-2}{3} \cdot \ldots \cdot \frac{n-k+1}{k} \quad \text{for } k = 0, 1, 2, \ldots \tag{1.4}$$

(Hvis k er 0, er højresiden »det tomme produkt« som er 1.)

Ved hjælp af formel (1.4) kan man med papir og blyant og lommeregner let finde at $\binom{37}{15} = 9\,364\,199\,760$.

Binomialformlen

Hvorfor hedder det »binomialkoefficient«? Et *bi*-nomium er en *to*-leddet størrelse som f.eks. $a+b$. En velkendt formel fortæller hvad kvadratet på en toleddet størrelse er:

$$(a+b)^2 = a^2 + 2ab + b^2.$$

Denne formel kan generaliseres til at handle om n-te potensen af en toleddet størrelse. Hvis man i

$$(a+b)^n = \underbrace{(a+b)(a+b)\ldots(a+b)}_{n \text{ faktorer}}$$

ganger parenteserne ud, får man 2^n led der hver især er et produkt af n faktorer, en fra hvert af de n binomier. Af disse 2^n led er der netop $\binom{n}{k}$ der består af k a-er og $n-k$ b-er. Derfor er

$$\begin{aligned}(a+b)^n &= \binom{n}{0}a^0b^n + \binom{n}{1}a^1b^{n-1} + \binom{n}{2}a^2b^{n-2} + \ldots + \binom{n}{n}a^nb^0 \\ &= \sum_{k=0}^{n}\binom{n}{k}a^kb^{n-k}.\end{aligned}$$

Denne formel hedder *binomialformlen*, fordi den handler om n-te potensen af et binomium. De koefficienter der indgår i binomialformlen, kaldes naturligt nok *binomialkoefficienter*.

1.2 Egenskaber ved binomialfordelingen

Definition 1.2: Binomialfordeling
Binomialfordelingen med sandsynlighedsparameter p og antalsparameter n er den diskrete sandsynlighedsfordeling givet ved sandsynlighedsfunktionen

$$f(y) = \binom{n}{y}p^y(1-p)^{n-y}, \quad y \in \{0,1,2,\ldots,n\}.$$

Her er p et (som oftest ukendt) tal mellem 0 og 1, og n er et positivt heltal.

Middelværdi og varians

Når man har at gøre med en sandsynlighedsfordeling, kan man (som bekendt) udregne visse talstørrelser der beskriver forskellige træk ved fordelingen. Man udregner ofte fordelingens *middelværdi* (= den forventede værdi = »tyngdepunktet« i fordelingen). Hvis Y er en stokastisk variabel der har en fordeling med sandsynlighedsfunktion f, så er middelværdien pr. definition tallet $\mathrm{E}\,Y = \sum y\,f(y)$ hvor der summeres over alle de mulige y-værdier. For binomialfordelingens vedkommende er middelværdien altså tallet

$$\mathrm{E}\,Y = \sum_{y=0}^{n} y\binom{n}{y}p^y(1-p)^{n-y}.$$

Denne sum ser ikke så rar ud, men heldigvis kan vi finde middelværdien på en anden og smartere måde. Som omtalt tidligere (side 11) kan en binomialfordelt stokastisk variabel Y fremkomme som en sum af uafhængige identisk fordelte 01-variable, så lad os sige at

$$Y = X_1 + X_2 + \cdots + X_n$$

hvor $X_1, X_2, \ldots, X_n$ er uafhængige 01-variable med $\mathrm{P}(X_j = 1) = p$ for alle j. Ifølge regneregler for middelværdi er middelværdien af en sum lig summen af middelværdierne:

$$\mathrm{E}\,Y = \mathrm{E}\,X_1 + \mathrm{E}\,X_2 + \cdots + \mathrm{E}\,X_n = n\,\mathrm{E}\,X_1,$$

så problemet er nu reduceret til at bestemme $\mathrm{E}\,X_1$, og det er overkommeligt ud fra definitionen af middelværdi:

$$\mathrm{E}\,X_1 = 0 \cdot \mathrm{P}(X_1 = 0) + 1 \cdot \mathrm{P}(X_1 = 1) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Vi har dermed fundet at $\mathrm{E}\,Y = np$.

Dernæst ser vi på *variansen*. Variansen på en stokastisk variabel Y med sandsynlighedsfunktion f er pr. definition $\mathrm{Var}\,Y = \mathrm{E}\big((Y - \mathrm{E}\,Y)^2\big) = \sum (y - \mathrm{E}\,Y)^2 f(y)$ hvor der summeres over de mulige y-værdier. Når det drejer sig om vores binomialfordelte stokastiske variabel $Y = Y_1 + Y_2 + \cdots + Y_n$ kan vi benytte et smart trick: Det er en egenskab ved varians at variansen af en sum af *uafhængige* størrelser er lig summen af varianserne på de enkelte led. Derfor er

$$\mathrm{Var}\,Y = \mathrm{Var}\,X_1 + \mathrm{Var}\,X_2 + \cdots + \mathrm{Var}\,X_n = n\,\mathrm{Var}\,X_1,$$

og vi behøver nu blot at finde variansen på X_1; da X_1 kun antager værdierne 0 og 1, bliver udregningerne simple:

$$\begin{aligned}
\mathrm{Var}\,X_1 &= \mathrm{E}\big((X_1 - \mathrm{E}\,X_1)^2\big) \\
&= \mathrm{E}\big((X_1 - p)^2\big) \\
&= (0-p)^2\,\mathrm{P}(X_1 = 0) + (1-p)^2\,\mathrm{P}(X_1 = 1) \\
&= p^2(1-p) + (1-p)^2 p \\
&= p\,(1-p).
\end{aligned}$$

Vi har hermed fundet at $\mathrm{Var}\,Y = np(1-p)$.

Sammenfattende gælder at hvis den stokastiske variabel Y er binomialfordelt med parametre n og p, så er

$$\begin{aligned}
\mathrm{E}\,Y &= np, \\
\mathrm{Var}\,Y &= np(1-p).
\end{aligned}$$

Standardafvigelsen på Y er pr. definition kvadratroden af variansen, dvs. for binomialfordelingens vedkommende $\sqrt{np(1-p)}$.

Udregning af binomialsandsynligheder

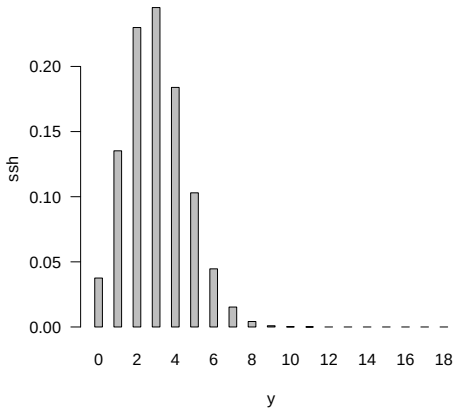
Hvis man ønsker at udregne binomialsandsynlighederne

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

for $y = 0, 1, 2, \ldots, n$, er det som regel ikke hensigtsmæssigt bare uden videre at indsætte i formlen. Man kan med fordel benytte en rekursionsformel. Ved simple omskrivninger finder man at

$$\frac{f(y)}{f(y-1)} = \frac{n-y+1}{y} \cdot \frac{p}{1-p}\ , \quad y = 1, 2, \ldots, n\,,$$

y	$f(y) = \binom{18}{y}(1/6)^y(5/6)^{18-y}$
0	0.038
1	0.135
2	0.230
3	0.245
4	0.184
5	0.103
6	0.045
7	0.015
8	0.004
9	0.001
10	0.000
11	0.000
12	0.000
13	0.000
14	0.000
15	0.000
16	0.000
17	0.000
18	0.000
	1.000



Figur 1.2 Eksempel 1.1: Tabel hhv. pindediagram over sandsynlighedsfunktionen f for binomialfordelingen med $n = 18$ og $p = 1/6$.

således at $f(y)$ let kan beregnes ud fra $f(y-1)$. Metoden bliver dermed

$$f(0) = (1-p)^n,$$
$$f(y) = f(y-1) \cdot \frac{n-y+1}{y} \cdot \frac{p}{1-p}\ , \quad y = 1, 2, \ldots, n.$$

Eksempel 1.1
Som eksempel vil vi beregne og tegne sandsynlighedsfunktionen for binomialfordelingen med $n = 18$ og $p = 1/6$. (Denne fordeling kunne f.eks. beskrive antallet af seksere ved 18 kast med en almindelig terning.) Fordelingen har i øvrigt middelværdi $18 \cdot 1/6 = 3$ og varians $18 \cdot 1/6 \cdot 5/6 = 2.5$ (svarende til standardafvigelsen 1.58). Ved at bruge den beskrevne metode udregnes fordelingens sandsynlighedsfunktion f, se figur 1.2.

1.3 Regn og tegn

Her omtales hvordan man kan foretage de forskellige beregninger med programmeringssproget R.

Binomialkoefficienter. Binomialkoefficienter udregnes med funktionen `choose`, f.eks. giver `choose(5,2)` værdien af $\binom{5}{2}$.

Binomialsandsynligheder. Binomialsandsynligheder udregnes med funktionen `dbinom`. Eksempelvis kan sandsynlighederne i binomialfordelingen med $n = 18$ og $p = 1/6$ (jf. figur 1.2) udregnes sådan:

```
n <- 18;  y <- 0:n
dbinom (y, size=n, prob=1/6)
```

Hvis man gerne ville have sandsynlighederne afrundet til tre decimaler, kunne man i stedet gøre sådan:

```
n <- 18;  y <- 0:n
round (dbinom (y, size=n, prob=1/6), digits=3)
```

Pindediagrammet kan frestilles sådan:

```
n <- 18;  y <- 0:n
barplot (dbinom (y, size=n, prob=1/6), space=1.5, names.arg=y, las=1, xlab="y", ylab="ssh")
```

Tabel 1.1. Man kan fremstille en tabel som tabel 1.1 på følgende måde. (Kaldet af `rbinom` leverer 180 tilfældige tal fra en binomialfordeling med $n = 1$ og $p = 1/3$, funktionen `matrix` putter tallene ind i en matrix med det ønskede antal rækker, funktionen `rowSums` udregner rækkesummer, og funktionen `cbind` sætter matricer sammen langs søjler (c = columns).)

```
t <- matrix (rbinom (180, size=1, prob=1/3), nrow=15)
cbind(t, rowSums(t))
```

1.4 Opgaver

Opgave 1.1
Tabel 1.1 (side 11) er fremstillet på den måde at man har sat et computerprogram til at frembringe udfald af 01-variable $X_1, X_2, \ldots, X_n$ sådan at sandsynligheden for værdien 1 hver gang er et givet tal p ($= 1/3$).

1. Udregn sandsynligheden for at få det talsæt $x_1, x_2, \ldots, x_n$ der står i række nummer 5.
2. Udregn sandsynligheden for at få det talsæt $x_1, x_2, \ldots, x_n$ der står i række nummer 7.
3. Opskriv sandsynlighedsfunktionen for $X_1, X_2, \ldots, X_n$.
4. Opskriv sandsynlighedsfunktionen for $Y = \sum_{j=1}^{n} X_j$.

Opgave 1.2
På side 11 nåede vi frem til en tilstrækkelig betingelse for at en stokastisk variabel Y er binomialfordelt. – Overvej med denne betingelse in mente om man kan benytte binomialfordelingsmodeller i nedenstående kort antydede situationer (angiv i givet fald hvad elementarforsøgene og hvad parametrene n og p er):

1. Antal toere ved fem kast med en almindelig terning.
2. Antal toere ved et kast med fem almindelige terninger.
3. Antal gange man skal kaste en almindelig terning for at få en toer.
4. Antal børn i en skoleklasse som bruger briller.

5. Antal nyregistrerede AIDS-tilfælde i Danmark i maj år 2002.
6. Antal passagerer i en HT-bus som ved forrige valg stemte på Socialdemokratiet.
7. Antal trykfejl i en bog.

Opgave 1.3
Udregn binomialkoefficienten $\binom{12}{5}$ dels ved hjælp af Pascals trekant, dels ved hjælp af formel (1.4) (og uden at bruge lommeregneren).

Opgave 1.4
I tabel 1.1 fremstilledes udfald $y_1, y_2, \ldots, y_{15}$ af en stokastisk variabel Y som er binomialfordelt med antalsparameter 12 og sandsynlighedsparameter $1/3$.

1. Udregn en tabel over fordelingen af Y (altså en tabel over sandsynlighedsfunktionen for binomialfordelingen med antalsparameter 12 og sandsynlighedsparameter $1/3$).

   Sammenlign med den empiriske fordeling af $y_1, y_2, \ldots, y_{15}$ (altså de relative hyppigheder hvormed udfaldene $0, 1, 2, \ldots, 12$ faktisk er forekommet).
2. Tegn et pindediagram over fordelingen af Y (altså en tegning i stil med figur 1.2).

   Tegn desuden et pindediagram over den empiriske fordeling. Ligner de to fordelinger hinanden?
3. Hvor mange gange ud af 15 gentagelser skulle man forvente at få observationen $Y = 5$? Hvor mange gange har man faktisk fået observationen 5?
4. Udregn middelværdien af Y. Udregn variansen og standardafvigelsen på Y.

Opgave 1.5 (Fru Hansen spiller banko)
Fru Hansen går til banko-spil de fem af ugens dage. Hun kan derfor opleve at der er 0, 1, 2, 3, 4 eller 5 dage i løbet af ugen hvor hun går hjem med en gevinst, men det er tilfældigt hvad det faktiske antal »gevinstdage« bliver. Man kan derfor (for en bestemt uge) indføre en stokastisk variabel Y som skal stå for »antal gevinstdage i den pågældende uge«. Man vil gerne vide noget om det forventede antal gevinstdage på en uge, dvs. noget om $\mathrm{E}\,Y$.

Antag at der hver dag er sandsynligheden p for at hun vinder.

1. Formulér en passende statistisk model for antallet Y af gevinstdage.
2. Hvad er det forventede antal gevinstdage $\mathrm{E}\,Y$? Tegn grafen for $\mathrm{E}\,Y$ som funktion af p.
3. For at få et indtryk af hvor meget Y kan variere fra uge til uge, vil man også gerne vide noget om $\mathrm{Var}\,Y$.

   Hvad er variansen $\mathrm{Var}\,Y$ på Y? Tegn grafen for $\mathrm{Var}\,Y$ som funktion af p; hvornår er variansen størst, og hvor stor er den da?
4. Bankospilarrangøren vil indrette det sådan at hvis man spiller hver af ugens fem »arbejdsdage«, så skal man kunne forvente netop én gevinstdag.
   a) Hvad skal han da vælge p til at være?
   b) Tegn den tilsvarende fordeling af Y.
   c) Hvor stor er variansen i fordelingen?
5. Fru Hansen vil spille i 10 uger. Hvor mange uger må hun forvente at hun ikke får en eneste gevinstdag?

Opgave 1.6 (Eksempel på simpel forsøgsplanlægning)
Ved en meningsmåling vil man spørge n personer om de er for eller mod et bestemt emne; derefter vil man udregne antallet Y af svarpersoner der er *for*.

1. Formulér en passende statistisk model for denne situation (dvs. angiv en sandsynlighedsfunktion for Y).
2. Benyt modellen til at finde standardafvigelsen på Y (for at få en idé om størrelsen af den tilfældige variation). Hvad er standardafvigelsen på den relative hyppighed Y/n?
3. Hvordan afhænger standardafvigelsen af de indgående parametre? Hvor stor skal n være for at standardafvigelsen på den relative hyppighed er 0.02 (eller mindre)?

Opgave 1.7 (Hypergeometriske sandsynligheder)
Kombinatorik er læren om at tælle. Mange kombinatoriske problemer formuleres på den måde at man taler om forskelligtfarvede *kugler* der lægges ned i og tages op af *kasser* (eller *urner*) efter bestemte regler.

Antag at man har en kasse med R røde og H hvide kugler.

1. Vis (med udgangspunkt i definition 1.1) at der er $\binom{R}{r}$ forskellige måder hvorpå man kan udtage r røde kugler uden tilbagelægning.
2. Man vil udtage n kugler i alt fra kassen, stadig uden tilbagelægning. Find antallet af forskellige måder det kan gøres på således at man får netop r røde og $n-r$ hvide kugler.
   Svaret er $\binom{R}{r} \cdot \binom{H}{n-r}$. – Det er underforstået at r et et heltal der opfylder visse betingelser:
   a) $0 \leq r \leq n$: antal udtagne røde kugler må ligge mellem 0 og det totale antal udtagne kugler (n).
   b) $r \leq R$: man kan ikke udtage flere røde kugler end der er.
   c) $n - r \leq H$: man kan ikke udtage flere hvide kugler end der er.
3. Vis at $\sum_{\text{alle } r} \binom{R}{r} \cdot \binom{H}{n-r} = \binom{R+H}{n}$.
4. Hvis man roder godt rundt i kassen inden man udtager de n kugler, kan man sige at man får udvalgt en *tilfældig delmængde* bestående af n kugler således at enhver af de $\binom{R+H}{n}$ forskellige delmængder har samme sandsynlighed for at blive udvalgt.
   Vis at sandsynligheden for at man derved får udvalgt en delmængde der indeholder netop r røde og $n-r$ hvide kugler, er $\dfrac{\binom{R}{r} \cdot \binom{H}{n-r}}{\binom{R+H}{n}}$. (Dette er et eksempel på en *hypergeometrisk* sandsynlighed.)

2 Den simple binomialfordelingsmodel

I FORRIGE KAPITEL opstillede vi en statistisk model i den simple binomialfordelingssituation (side 11). I modellen optræder to størrelser n og p der tilsammen specificerer binomialfordelingen. Størrelsen n er et kendt tal, men p er ukendt: værdien af n fastsættes ved planlægningen af forsøget, hvorimod p beskriver en egenskab ved den tilfældighedsmekanisme der frembringer observationerne; man kan også sige at p beskriver en egenskab ved naturen eller virkeligheden. Man kalder en størrelse som p for en *parameter* i den statistiske model. Man taler ofte om *den sande værdi* af parameteren p når meningen er den værdi som p »i virkeligheden« har (i modsætning til en værdi som man selv foreslår). – I dette kapitel skal vi se hvordan man kan få noget at vide om den sande værdi af p.

2.1 Estimation af parameteren p

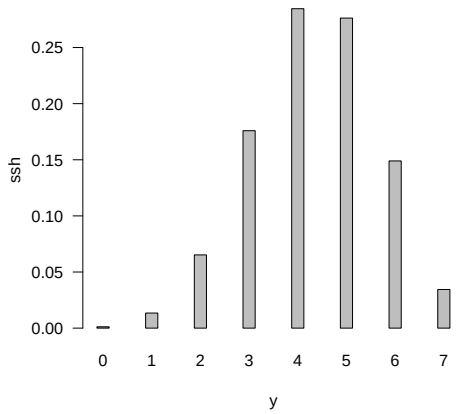
Ved hjælp af den statistiske model er det muligt at hente information ud af observationerne om hvad den sande parameterværdi sådan cirka kan være: man udregner et *skøn* eller et *estimat* over værdien af p, og selve processen hedder *estimation*.

I eksemplet med rismelsbillerne var $n = 144$ og det observerede antal gunstige udfald var $y = 43$. Da p skal fortolkes som sandsynligheden for at få et gunstigt udfald, og da man har observeret 43 gunstige ud af 144, er det nærliggende at foreslå at estimere p til den relative hyppighed $y/n = {}^{43}/_{144} = 0.30$.

I det følgende vil vi præsentere en generel estimationsmetode der kan bruges i »enhver« situation, og vi vil eftervise, at denne generelle metode fører frem til at sandsynlighedsparameteren p faktisk skal estimeres som y/n.

Likelihoodmetoden

Det er i visse simple tilfælde ret klart hvordan man »selvfølgelig« skal analysere sin statistiske model, idet der er en »umiddelbart indlysende« fremgangsmåde osv. I andre tilfælde (de fleste) er det knap så klart. Vi vil introducere et sæt overordnede principper for hvordan man bør analysere en statistisk model. Disse principper gælder (med visse tilføjelser) for »enhver« model. Indførelsen af principperne betyder *ikke*, at man slipper for overvejelser over hvad man »selvfølgelig« skal gøre og hvad der er »umiddelbart indlysende«, *men* at man i stedet for at skulle gøre overvejelserne igen og igen i hvert enkelt tilfælde så at sige overstår dem alle på en gang ved at hæve dem fra enkelttilfældene op til et overordnet niveau, hvor de udnævnes til generelle principper. – Et princip er i denne sammenhæng en norm, en retningslinje, som ikke bliver logisk-deduktivt bevist,



Figur 2.1 En »typisk« sandsynlighedsfunktion $y \mapsto f(y; p)$.

men som retfærdiggøres dels gennem generelle betragtninger og overvejelser, dels ved at levere fornuftige resultater i konkrete situationer.

Vi vil i al stilfærdighed præsentere et sådant sæt principper og vise hvordan de udmøntes i en generel metode til estimation af ukendte parametre i statistiske modeller. I dette kapitel vil vi se på hvordan den generelle metode ser ud i eksemplet »den simple binomialfordelingsmodel«, og som gennemgående eksempel på »den simple binomialfordelingsmodel« bruger vi rismelsbille-eksemplet.*

Den statistiske model i rismelsbille-eksemplet går ud på at $y = 43$ opfattes som en observation af en stokastisk variabel Y der er binomialfordelt med antalsparameter $n = 144$ og ukendt sandsynlighedsparameter $p \in [0, 1]$.

Sandsynlighedsfunktionen for Y er

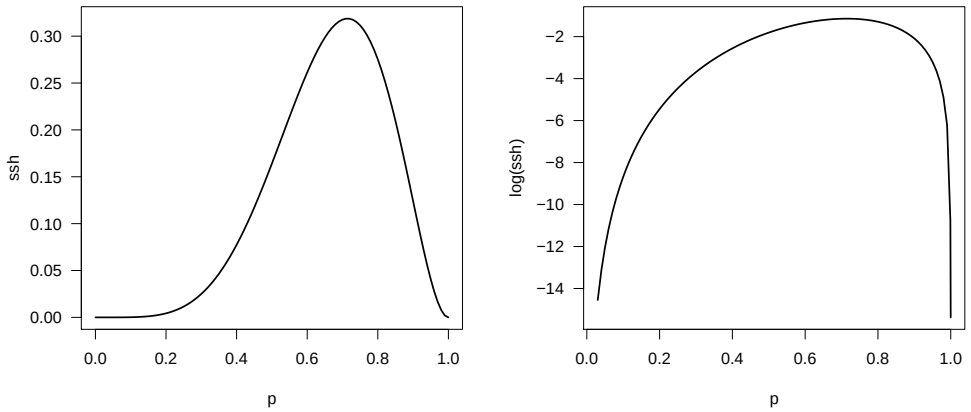
$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, 2, \ldots, n\}.$$

For at fremhæve at udtrykket afhænger af både y og p, udskifter vi betegnelsen »$f(y)$« med »$f(y; p)$«, dvs.

$$f(y; p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, 2, \ldots, n\},\ p \in [0, 1].$$

Funktionen f er nu en funktion af *to* variable, en observationsvariabel y og en parametervariabel p. Denne funktion kaldes *modelfunktionen* for den statistiske model fordi den specificerer modellen fuldstændigt: for enhver kombination af en mulig observation y og en mulig parameterværdi p angiver den sandsynligheden for at observere netop det y hvis netop det p er den rigtige parameterværdi. Modelfunktionen er flere funktioner i én:

∗ Der er altså flere niveauer af eksempler: Rismelsbille-eksemplet er et eksempel på en simpel binomialfordelingsmodel, og den simple binomialfordelingsmodel er et eksempel på en statistisk model.



Figur 2.2 Til venstre: en »typisk« likelihoodfunktion $p \mapsto L(p; y) = f(y; p)$. Til højre: den tilsvarende log-likelihoodfunktion.

- Hvis vi i modelfunktionen fixerer p og opfatter funktionen som en funktion af y alene, så har vi *sandsynlighedsfunktionen* svarende til parameterværdien p. En »typisk« sandsynlighedsfunktion er vist i figur 2.1.
- Hvis vi i modelfunktionen fixerer y og opfatter funktionen som en funktion af p alene, så har vi den såkaldte *likelihoodfunktion* svarende til observationen y. Likelihoodfunktion betegnes ofte $L(\,\cdot\,)$ eller $L(\,\cdot\,; y)$:

$$L(p) = L(p; y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad p \in [0, 1].$$

  Figur 2.2 viser en »typisk« likelihoodfunktion.

I vort eksempel er modelfunktionen

$$f(y; p) = \binom{144}{y} p^{43} (1-p)^{101}, \quad y \in \{0, 1, 2, \ldots, 144\},\ p \in [0, 1],$$

og likelihoodfunktionen svarende til observationen $y = 43$ er

$$L(p) = L(p; 43) = \binom{144}{43} p^{43} (1-p)^{101}, \quad p \in [0, 1].$$

Likelihoodfunktionsværdien $L(p; y)$ er sandsynligheden for at observere det y man faktisk har observeret, forudsat at den ukendte parameter har værdien p. Likelihoodfunktionen kan derfor anvendes til at sammenligne forskellige parameterværdiers evne til at beskrive den faktiske observation y. For hvis f.eks. $L(p_1; y) < L(p_2; y)$, så er chancen for at observere netop dette y større når p er lig p_2, end når p er lig p_1, og det må betyde at p_2 giver en bedre beskrivelse af data end p_1 gør. Den parameterværdi som giver den bedste beskrivelse efter disse retningslinjer, er da den værdi som maksimaliserer likelihoodfunktionen, og den kaldes *maksimaliseringsestimatet* (eller *maximum likelihood*

estimatet) for p og betegnes $\widehat{p}$ (»p hat«). Tallet $\widehat{p}$ er altså bestemt ved at

$$L(\widehat{p}; y) \geq L(p; y) \quad \text{for alle } p.$$

Bemærk at $\widehat{p}$ er en funktion af y.

Af bekvemmelighedsgrunde opererer man tit med *»log-likelihoodfunktionen«*, dvs. funktionen $\ln L(p)$, og man bestemmer $\widehat{p}$ som maksimumspunktet for $\ln L$ (resultatet bliver jo det samme). I vort eksempel er log-likelihoodfunktionen

$$\ln L(p) = \ln \binom{144}{43} + 43 \ln p + 101 \ln(1-p).$$

Imidlertid vil talværdierne let gøre ræsonnementerne ugennemskuelige, så vi vender tilbage til den generelle binomialfordelingsmodel hvor log-likelihoodfunktionen er

$$\ln L(p) = \ln \binom{n}{y} + y \ln p + (n-y) \ln(1-p).$$

Hvad er $\widehat{p}$ i denne model? Svaret herpå får vi ved at løse den matematikopgave der hedder: »Bestem maksimumspunkt(er) for funktionen $p \mapsto \ln L(p)$ når $p \in [0,1]$«, så det gør vi. Fra matematikken ved vi at kandidater til maksimumspunkter er dels intervalendepunkterne $p = 0$ og $p = 1$, dels de stationære punkter, dvs. de punkter hvor $\frac{d}{dp} \ln L(p) = 0$. For $0 < p < 1$ er

$$\frac{d}{dp} \ln L(p) = \frac{y}{p} - \frac{n-y}{1-p} = \frac{y - np}{p\,(1-p)}.$$

Det er hensigtsmæssigt at dele op i tre tilfælde:

1. $0 < y < n$: Så er punktet $p = y/n$ det eneste stationære punkt for $\ln L$, og da $\ln L(0)$ og $\ln L(1)$ begge er $-\infty$, er $p = y/n$ et entydigt maksimumspunkt.
2. $y = n$: Så er $\ln L(p) = n \cdot \ln p$, hvilket er en voksende funktion af p. Den antager derfor sin største værdi når p er størst mulig, dvs. når $p = 1$.
3. $y = 0$: Så er $\ln L(p) = n \cdot \ln(1-p)$, hvilket er en aftagende funktion af p. Den antager derfor sin største værdi når p er mindst mulig, dvs. når $p = 0$.

I alle tre tilfælde er der således et entydigt maksimumspunkt der kan udregnes som y/n. Vi er hermed nået frem til at *i binomialmodellen med modelfunktion*

$$f(y; p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, 2, \ldots, n\},\ p \in [0,1],$$

er maksimaliseringsestimatet $\widehat{p}$ for p givet som $\widehat{p} = y/n$.

At p skal estimeres ved den relative hyppighed y/n, kan næppe overraske nogen, det er næsten hvad man kan sige sig selv. Det interessante er at det altså også er det svar man når frem til ved at benytte den generelle fremgangsmåde der lyder

- opstil modelfunktionen,
- dan derudfra likelihoodfunktionen,
- bestem $\widehat{p}$ som maksimumspunktet for likelihoodfunktionen.

Det er vigtigt at have in mente at der tænkes at eksistere en *sand parameterværdi* som er et bestemt, ukendt tal. Vi kan principielt aldrig erfare den sande parameterværdi, men ud fra foreliggende observationer kan vi estimere den.

Middelfejlen på $\widehat{p}$

Maksimaliseringsestimatet $\widehat{p} = y/n$ er det bedste bud vi kan give på den ukendte p-værdi når vi har observeret antallet y ud af n. Den statistiske model fortæller at y er at opfatte som en observation af en stokastisk variabel Y; det medfører at vi også må opfatte estimatet y/n som en observation af en stokastisk variabel, nemlig Y/n; den stokastiske variabel $\widehat{p} = \widehat{p}(Y) = Y/n$ kaldes *maksimaliseringsestimatoren* for p. Da Y er binomialfordelt med parametre n og p, er middelværdien $\mathrm{E}\,Y$ af Y lig np, og ifølge regnereglerne for middelværdi er så $\mathrm{E}\,\widehat{p}(Y) = (\mathrm{E}\,Y)/n = p$, hvilket betyder at maksimaliseringsestimatoren $\widehat{p}$ for p i middel giver det rigtige svar p – men deraf følger ikke noget om det konkrete enkelttilfælde.†

For at få en idé om størrelsen af maksimaliseringsestimatorens tilfældige variation omkring sin middelværdi p kan man bestemme den såkaldte *middelfejl* på $\widehat{p}$, dvs. standardafvigelsen på $\widehat{p}(Y)$. Da Y har varians $np(1-p)$, er variansen på $\widehat{p}(Y) = Y/n$ lig $np(1-p)/n^2 = p(1-p)/n$, så middelfejlen på $\widehat{p}(Y)$ er

$$\sqrt{p(1-p)/n}.$$

I billeeksemplet er standardafvigelsen på $\widehat{p}$ lig $\sqrt{p(1-p)/144}$, der kan estimeres til $\sqrt{\widehat{p}(1-\widehat{p})/n} = \sqrt{0.30 \times 0.70/144} = 0.04$.

Sammenfattende kan vi sige at binomialparameteren p i billeeksemplet estimeres til $\widehat{p} = 0.30$ med en standardafvigelse på 0.04.

2.2 En simpel statistisk hypotese

Det er ikke altid at man er tilfreds med blot at *estimere* den ukendte parameter i den statistiske model, undertiden ønsker man også at opstille og teste *statistiske hypoteser* vedrørende den sande værdi af parameteren.

Antag at det i rismelsbilleeksemplet er sådan‡ at man har en referencegift hvorom man ved at når man doserer den med 0.20 mg/cm^2, så dør 23% af billerne. Den gift der er afprøvet i eksemplet, er ligeledes doseret med 0.20 mg/cm^2, og der skete som nævnt det at 43 ud af 144 biller døde. Spørgsmålet er om den afprøvede gift virker på samme måde som referencegiften.

Hvad »på samme måde« nærmere skal betyde, kan man sikkert diskutere længe og inderligt, men formuleret i den statistiske models sprog er det nemt nok: det skal betyde at $p = p_0$, altså at sandsynligheden for at en bille dør når den er blevet udsat for den afprøvede gift, er lig p_0, hvor p_0 er den kendte sandsynlighed for at dø af referencegiften (altså 0.23). Påstanden at $p = p_0$, er et eksempel på en såkaldt *statistisk hypotese*; statistiske hypoteser navngives ofte med symboler som H_0, H_1, osv., så her vil vi tale om hypotesen $H_0 : p = p_0$.

Hvordan passer den statistiske hypotese og de foreliggende observationer sammen? Man kan se at den estimerede værdi $\widehat{p} = {}^{43}/_{144}$ ikke er lig med 0.23, men en eksakt

† En estimator hvis middelværdi er lig den parameter der skal estimeres, kaldes en *central* estimator (på engelsk: an *unbiased* estimator).

‡ – men det er det ikke; denne del af eksemplet er opdigtet til lejligheden.

lighed ville også være mere end man kunne forvente, når man tager i betragtning at modellen siger at tallet $y = 43$ er en observation fra en *sandsynlighedsfordeling*. Derfor kan man kun sige at

- hvis der *ikke* er stor afvigelse mellem $\widehat{p}$ og p_0, så er der *ikke* klare tegn på at den afprøvede gift virker anderledes end referencegiften – der *er ikke* nogen *signifikant* forskel,
- og hvis der *er* stor afvigelse mellem $\widehat{p}$ og p_0, så er det tegn på at den afprøvede gift *ikke* virker på samme måde som referencegiften – der *er* en *signifikant* forskel.

Her er der to ting der behøver en nærmere præcisering: hvordan måler man *afvigelsen* mellem $\widehat{p}$ og p_0, og hvordan afgør man hvornår afvigelsen er stor og hvornår ikke. I afsnit 2.3 præsenteres en generel metode hvormed man kan håndtere disse spørgsmål.

Det faglige problem blev præsenteret på den måde at man ønskede at vide om den afprøvede gift virkede på samme måde som referencegiften, og det førte til hypotesen $H_0 : p = p_0$. Men hvis man i stedet havde stillet spørgsmålet om der var forskel på de to gifte, hvordan skulle man så have grebet sagen an? Svaret er: på nøjagtig samme måde, altså stadig ved at undersøge $H_0 : p = p_0$. Statistiske hypoteser er nemlig altid *forsimplende*, dvs. man går fra det mere omfattende til det mindre omfattende. I eksemplet begynder man derfor med den mest omfattende model, den hvor p kan være hvadsomhelst, og så opstiller man som statistisk hypotese at modellen er mindre omfattende, nemlig at p kun har lov til at have den ene værdi p_0.

2.3 Kvotienttesstørrelsen

Det blev påstået at man ved hjælp af likelihoodfunktionen kan sammenligne forskellige parameterværdiers evne til at beskrive det faktisk observerede y: hvis $L(p_1; y) < L(p_2; y)$, så giver parameterværdien p_2 en bedre beskrivelse end parameterværdien p_1 gør, inden for rammerne af den aktuelle statistiske model. I særdeleshed giver maksimaliseringsestimatet $\widehat{p} = \widehat{p}(y)$ den bedst mulige beskrivelse af observationen y. Parameterværdier der giver en værdi af likelihoodfunktionen som ligger tæt på den maksimale værdi $L(\widehat{p})$, må give en næsten lige så god beskrivelse af observationen y som $\widehat{p}$ gør. Når vi derfor skal teste en statistisk hypotese $H_0 : p = p_0$ om at den ukendte parameter p kan antages at have den kendte værdi p_0, så må det foregå ved at sammenligne likelihoodfunktionens værdi i punktet p_0 med dens maksimale værdi, altså ved at sammenligne de to tal $L(p_0)$ og $L(\widehat{p})$. Hvis $L(p_0)$ er næsten lige så stor som $L(\widehat{p})$, betyder det at p_0 beskriver observationen y næsten lige så godt som $\widehat{p}$ gør, og det betyder igen at man kan tillade sig at mene at p_0 er den sande værdi af p: man *accepterer* eller *godkender* hypotesen H_0. Hvis derimod $L(p_0)$ er væsentligt mindre end $L(\widehat{p})$, betyder det at p_0 giver en væsentligt dårligere beskrivelse af observationen y end $\widehat{p}$ gør, og det er derfor ikke rimeligt at mene at p_0 skulle være den sande værdi af p: man *forkaster* hypotesen H_0.

Når man sammenligner $L(p_0)$ og $L(\widehat{p})$, skal det gøres ved at dividere den mindste med den største: man danner kvotienten

$$Q = Q(y) = \frac{L(p_0)}{L(\widehat{p})} = \frac{L(p_0; y)}{L(\widehat{p}; y)} \ .$$

Resultatet bliver et tal mellem 0 og 1, og

- en Q-værdi nær 1 betyder at p_0 er stort set lige så god som $\widehat{p}$, dvs. man accepterer H_0,
- en Q-værdi langt fra 1 betyder at p_0 er væsentligt dårligere end $\widehat{p}$, dvs. man forkaster H_0.

Man kalder Q for *kvotienttesteststørrelsen* for den statistiske hypotese H_0.

I binomialfordelingsmodellen er $L(p) = \binom{n}{y} p^y(1-p)^{n-y}$, så

$$Q = Q(y) = \frac{p_0^y\,(1-p_0)^{n-y}}{\widehat{p}^{\,y}\,(1-\widehat{p})^{n-y}} = \left(\frac{np_0}{y}\right)^y \left(\frac{n(1-p_0)}{n-y}\right)^{n-y}$$

idet $\widehat{p} = y/n$. I eksemplet er $n = 144$, $y = 43$ og $p_0 = 0.23$, så den observerede værdi Q_{obs} af Q er

$$Q_{\text{obs}} = \left(\frac{144 \times 0.23}{43}\right)^{43} \left(\frac{144 \times 0.77}{101}\right)^{101} = 0.165\ .$$

Tallet $Q_{\text{obs}} = 0.165$ i sig selv kan vi ikke stille noget op med – det giver ingen mening at spørge om 0.165 er nær 1 eller langt fra 1 så længe vi ikke har en målestok eller et sammenligningsgrundlag. Den statistiske model fortæller at vi skal betragte y som en observation af en stokastisk variabel Y; dermed skal vi også betragte $Q_{\text{obs}} = Q(y)$ som en observation af den stokastiske variabel $Q(Y)$. Fordelingen af Y beskriver hvilke y-værdier man også kunne have fået (i stedet for den faktisk observerede) og med hvilke sandsynligheder, og den tilsvarende fordeling af $Q(Y)$ beskriver dermed hvilke Q-værdier man også kunne have fået (i stedet for 0.165) og med hvilke sandsynligheder. Takket være sandsynlighedsfordelingerne kan vi altså sammenholde den faktiske værdi $Q_{\text{obs}} = 0.165$ med alle de andre Q-værdier man også kunne have fået når p har værdien p_0.

- Hvis det er sådan at der når $p = p_0$ er en pæn chance (f.eks. over 5%) for at få Q-værdier som ligger længere væk fra 1 end Q_{obs} gør, dvs. for at få Q-værdier for hvilke $Q \le Q_{\text{obs}}$, så vil man sige at Q_{obs} ikke ligger specielt langt fra 1, og man vil *acceptere* hypotesen $H_0 : p = p_0$.
- Hvis det derimod er sådan at der når $p = p_0$ er meget lille chance (f.eks. under 5%) for at få Q-værdier som ligger længere fra 1 end Q_{obs} gør, dvs. for at få Q-værdier for hvilke $Q \le Q_{\text{obs}}$, så vil man fortolke det som at Q_{obs} i sig selv ligger usædvanligt langt fra 1, og man vil *forkaste* hypotesen $H_0 : p = p_0$.

Når man skal teste hypotesen H_0, skal man derfor bestemme *testsandsynligheden*

$$\varepsilon = \mathrm{P}_0(Q \le Q_{\text{obs}}).$$

Testsandsynligheden er sandsynligheden under H_0 for at få en værre, dvs. mindre, Q-værdi end den faktisk observerede værdi Q_{obs}. (Fodtegnet 0 på P-et angiver at sandsynligheden skal udregnes under antagelse af at hypotesen H_0 er rigtig.)

1. Hvis testsandsynligheden ε er meget lille, så forkaster man H_0 på grund af følgende ræsonnement:

a) Vi har fået en Q_{obs}-værdi der er så langt fra 1 at der, forudsat at H_0 er rigtig, kun er den meget lille sandsynlighed ε for at få en værre Q-værdi.
b) I praksis plejer man ikke at få særligt ekstreme observationer, så der må være noget galt med forudsætningerne for beregningen af ε.
c) Da vi ikke kan lave om på observationerne, må det være hypotesen H_0 det er galt med.

2. Hvis testsandsynligheden ε har en pæn størrelse, så kan man *ikke* forkaste H_0. Ræsonnementet er denne gang således:
   a) Vi har fået en Q_{obs}-værdi der ikke ligger specielt langt fra 1, thi der er nemlig, forudsat at H_0 er rigtig, en pæn chance ε for at få en værre Q-værdi.
   b) Den faktiske værdi Q_{obs} er derfor udmærket forenelig med hypotesen H_0, og der er dermed *ikke* grundlag for at forkaste H_0.

Hvis testsandsynligheden ε er så lille at man forkaster hypotesen, så siger man at teststørrelsen Q_{obs} er *signifikant* eller at der er *signifikans*.

Bestemmelse af testsandsynligheden ε

Vi vil nu for en stund holde inde med generelle betragtninger over tests og i stedet vende tilbage til den konkrete binomialfordelingsmodel hvor der viser sig et påtrængende problem, nemlig hvordan bestemmer man rent faktisk testsandsynligheden ε? Pr. definition er ε lig med sandsynligheden (når den sande parameterværdi er p_0) for at $Q(Y) \leq Q_{\text{obs}}$. Af forskellige grunde hvoraf nogle er regnetekniske og andre vil fremgå lidt senere, udregner man ofte $-2\ln Q$ i stedet for Q, og testsandsynligheden er da sandsynligheden for at $-2\ln Q(Y) \geq -2\ln Q_{\text{obs}}$. Ud fra det tidligere fundne udtryk for Q får vi at

$$-2\ln Q(y) = 2\left(y\ln\frac{y}{np_0} + (n-y)\ln\frac{n-y}{n(1-p_0)}\right), \tag{2.1}$$

så i taleksemplet er

$$-2\ln Q(y) = 2\left(y\ln\frac{y}{33.12} + (144-y)\ln\frac{144-y}{110.88}\right) \tag{2.2}$$

og dermed

$$-2\ln Q_{\text{obs}} = -2\ln Q(43) = 3.60\,.$$

Testsandsynligheden ε kan nu fås ved at summere sandsynlighederne for alle de y-er som har den egenskab at $-2\ln Q(y) \geq -2\ln Q_{\text{obs}}$ idet sandsynlighederne udregnes under antagelse af at hypotesen er rigtig, dvs. at $p = p_0$:

$$\varepsilon = \sum_{y\,:\,-2\ln Q(y) \geq -2\ln Q_{\text{obs}}} \binom{n}{y} p_0^y\,(1-p_0)^{n-y}\,.$$

Her har vi ε udtrykt ved lutter kendte størrelse.

Fremgangsmåden til bestemmelse af testsandsynligheden ε er derfor kort fortalt

1. Udregn $-2\ln Q_{\text{obs}}$.

2. Udregn $-2\ln Q(y)$ for $y = 0, 1, 2, \ldots, n$.
   (NB: Når man udregner $-2\ln Q(0)$ og $-2\ln Q(n)$, skal man sætte $0\ln 0$ til 0.)
3. Bestem de y-er for hvilke $-2\ln Q(y) \geq -2\ln Q_{\text{obs}}$.
4. Bestem binomialsandsynlighederne for de således udpegede y-er.
5. Testsandsynligheden ε er summen af disse sandsynligheder.

I taleksemplet er

$$\varepsilon = \sum_{y\,:\,-2\ln Q(y)\geq 3.60} \binom{144}{y} 0.23^y\, 0.77^{144-y}$$

hvor $-2\ln Q(y)$ er givet ved formel (2.2). Ved almindelig udregning finder man at uligheden $-2\ln Q(y) \geq 3.60$ er opfyldt for $y = 0, 1, 2, \ldots, 23$ og for $y = 43, 44, 45, \ldots, 144$. Videre finder man at $\mathrm{P}_0(Y \leq 23) = 0.0249$ og at $\mathrm{P}_0(Y \geq 43) = 0.0344$, så at den *eksakte testsandsynlighed* er $\varepsilon = 0.0249 + 0.0344 = 0.0593 \approx 5.9\%$.

χ^2-approksimationen

Ganske vist er der i afsnit 1.2 vist en udmærket algoritme til beregning af binomialsandsynligheder, men alligevel må man nok sige at ovennævnte regnestykke nok ikke er noget man lige klarer i en håndevending, medmindre man da har en datamat eller en programmerbar lommeregner til sin rådighed. Heldigvis kan den matematiske statistik komme os til hjælp idet den kan fortælle hvordan man uden større besvær kan bestemme en god tilnærmet værdi af testsandsynligheden. Man kan bevise generelt at for binomialmodellen og for en lang række andre statistiske modeller gælder at den sandsynlighedsfordeling som kvotienttesteststørrelsen $-2\ln Q$ følger når den testede hypotese er rigtig, med god tilnærmelse er af en ganske bestemt type, nemlig en såkaldt χ^2-*fordeling* (»khi-i-anden fordeling«) med et vist antal *frihedsgrader* der i vores aktuelle tilfælde er lig 1.§ Da testsandsynligheden ε jo er sandsynligheden for at få en $-2\ln Q$-værdi som er større end $-2\ln Q_{\text{obs}}$, betyder det at ε med god tilnærmelse er lig med sandsynligheden for at få en værdi større end $-2\ln Q_{\text{obs}}$ i en χ^2-fordeling med 1 frihedsgrad, og den sandsynlighed kan let bestemmes, f.eks. ved hjælp af tabeller over fraktiler¶ i χ^2-fordelingen, se f.eks. tabellen side 222.

Man finder at i χ^2-fordelingen med 1 frihedsgrad er 90%-fraktilen 2.71 og 95%-fraktilen 3.84. Den aktuelle $-2\ln Q_{\text{obs}}$-værdi 3.60 ligger mellem disse to fraktiler hvilket betyder at (det tilnærmede) ε ligger mellem 10% og 5%. (Dette harmonerer udmærket med at den eksakte testsandsynlighed er 5.9%.)

Som nævnt er χ^2-fordelingen kun en approksimation til den rigtige fordeling af $-2\ln Q$ under H_0. Man er naturligvis nødt til at have nogle retningslinjer for hvornår approksimationen er god og hvornår ikke. Man plejer at gå ud fra at hvis begge de *forventede antal* np_0 og $n(1-p_0)$ (det forventede antal døde hhv. ikke døde) er mindst

§ Det kan nævnes at χ^2-fordelingen med 1 frihedsgrad er den kontinuerte sandsynlighedsfordeling som har tæthedsfunktion $f(x) = \frac{1}{\sqrt{2\pi}}\, x^{-1/2} \exp(-x/2),\ x > 0$.

¶ En *fraktil* i en fordeling er et tal x med den egenskab at der er en vis foreskreven sandsynlighed for at få værdier $\leq x$. Eksempelvis er 90%-fraktilen et tal x således at der er sandsynlighed 90% for at få værdier $\leq x$.

fem, så kan man anvende χ^2-approksimationen. Ellers må man regne den eksakte testsandsynlighed ud efter »slavemetoden«.

De mange udregninger må følges op af en konklusion: Vi fandt en testsandsynlighed på 5.9%, dvs. hvis hypotesen H_0 er rigtig, så er der en sandsynlighed på 5.9% for at få en større værdi end den faktisk observerede værdi $-2\ln Q = 3.60$. En sådan testsandsynlighed vil almindeligvis ikke føre til at man forkaster hypotesen H_0. Vi må altså konkludere at der ikke er nogen signifikant forskel mellem den afprøvede gift og referencegiften.

2.4 Regn og tegn

Her omtales hvordan man kan foretage beregningerne med R.

Testsandsynligheden ε udregnes med funktionen `binom.test`. I det gennemgåede eksempel med $n = 144$, $y = 43$ og $p_0 = 0.23$ skriver man

```
binom.test (43, 144, 0.23)
```

som resulterer i denne udskrift:

```
        Exact binomial test

data:  43 and 144
number of successes = 43, number of trials = 144, p-value = 0.05932
alternative hypothesis: true probability of success is not equal to 0.23
95 percent confidence interval:
 0.2252674 0.3804482
sample estimates:
probability of success
             0.2986111
```

Testsandsynligheden er det der i udskriften hedder 'p-value', og $\widehat{p}$ er det der hedder 'sample estimates: probability of success'.

χ^2-fordelingen Fraktiler i χ^2-fordelingen udregnes med funktionen `qchisq`, f.eks. giver `qchisq(0.95, df=1)` 95%-fraktilen i χ^2-fordelingen med 1 frihedsgrad.

Sandsynligheder udregnes med `pchisq`, f.eks. er `1-pchisq(3.60, df=1)` sandsynligheden for at få en værdi som er større end 3.60 i χ^2-fordelingen med 1 frihedsgrad.

2.5 Opgaver

Opgave 2.1
I tabel 1.1 fremstilledes udfald $y_1, y_2, \ldots, y_{15}$ af en stokastisk variabel Y som er binomialfordelt med antalsparameter 12 og sandsynlighedsparameter $1/3$.

1. Udregn for hver af de 15 observerede y-værdier den tilsvarende værdi af $\widehat{p}$.
2. Tegn et pindediagram over den empiriske fordeling af $\widehat{p}$.
3. Tegn et pindediagram over den teoretiske fordeling af $\widehat{p}$.

   Vink: Da Y er binomialfordelt, er fordelingen af $\widehat{p} = Y/n$ en »ned-skaleret binomialfordeling« på mængden $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\}$.

4. Hvor stor er middelfejlen på $\widehat{p}$?

Vink: Tabellen var også genstand for undersøgelse i opgave 1.4.

Opgave 2.2

En haveejer går ud på en eng og indsamler frø af en plante der findes i to udgaver, en med røde blomster og en med hvide blomster. (På engen var der eksemplarer af begge slags.) Næste år sår han frøene hjemme i haven; det viser sig at der kommer 10 planter, hvoraf syv har røde og tre har hvide blomster.

1. a) Udregn sandsynligheden for at få observationen 7 i en binomialfordeling med $n = 10$ og $p = 1/2$.
   b) Udregn sandsynligheden for at få observationen 7 i en binomialfordeling med $n = 10$ og $p = 3/4$.
   c) Udregn sandsynligheden for at få observationen 7 i en binomialfordeling med $n = 10$ og $p = 1/4$.
2. Haveejerens venner og bekendte kan ved fælles hjælp finde følgende mulige forklaringer på fænomenet:
   a) Det er tilfældigt om en plante får røde eller hvide blomster, og der er samme sandsynlighed for hver af de to muligheder.
   b) Det er genetisk bestemt om en plante får røde eller hvide blomster, og »røde blomster« er dominant; i så fald er sandsynligheden $3/4$ for at en plante har røde blomster.
   c) Det er genetisk bestemt om en plante får røde eller hvide blomster, og »hvide blomster« er dominant; i så fald er sandsynligheden $1/4$ for at en plante har røde blomster.

   Hvilken af de tre forklaringer forklarer det observerede bedst?
3. En fjerde forklaring er at det simpelt hen forholder sig sådan med den eng, at den indeholder rødblomstrede og hvidblomstrede eksemplarer af planten i et ganske bestemt forhold. Hvis det er tilfældet, hvad er da det bedste bud på talværdien af dette forhold?

Opgave 2.3

Georg har slået Plat eller Krone 5 gange med en almindelig mønt og fået netop én gang Krone. Gerda siger at det da må tyde på at mønten er skæv, ellers skulle man have fået 2 eller 3 gange Krone.

For at afgøre om man på denne baggrund kan sige at mønten er skæv, kan man opstille en statistisk model og inden for rammerne af den formulere og teste en statistisk hypotese.

Gør det, dvs. opstil modellen og formulér og test hypotesen:

1. Opstil en hensigtsmæssig statistisk model og omsæt det givne problem til en statistisk hypotese.
2. Opskriv likelihoodfunktionen svarende til observationen én gang Krone. Tegn grafen for likelihoodfunktionen. Hvornår er den størst?

   Samme spørgsmål for log-likelihoodfunktionen.
3. Opskriv kvotienttesstørrelsen Q for at teste hypotesen.
4. Udregn $-2\ln Q(y)$ for alle de mulige y-værdier, og find mængden af y-er for hvilke $-2\ln Q(y) \geq -2\ln Q_{\text{obs}}$ (svarende til at $Q(y) \leq Q_{\text{obs}}$), og udregn sandsynligheden for denne mængde.

   Hvor stor er testsandsynligheden? Forkastes hypotesen?

Opgave 2.4
Formulér en hensigtsmæssig statistisk model og hypotese for at besvare følgende:

Fyns Amtsavis oplyser at bladet trykker alle indlæg om fremmede. I en tremåneders periode bragte bladet 12 læserbreve med et positivt syn på fremmede og 15 med et negativt syn. Modtager bladet stort set lige mange positive og negative indlæg?

Opgave 2.5
I en af sine forsøgsrækker med ærteplanter undersøgte Mendel om ærterne var runde eller kantede. Først dyrkede han 253 selvbestøvede heterozygote planter, og det viste sig at de ærter der kom, fordelte sig med 5474 runde og 1850 kantede. Derpå dyrkede og selvbestøvede han planter af 565 af de runde ærter fra det første forsøg. Det viste sig at 193 af disse planter udelukkende fik runde ærter, mens de resterende 372 fik både runde og kantede ærter.

Man kan nu opstille en *genetisk model* gående ud på at det er et enkelt gen der bestemmer om ærter bliver runde eller kantede, og at genet for runde ærter er dominant. En konsekvens af denne model er at efterkommerne af de 253 selvbestøvede heterozygote planter i det første forsøg skal fordele sig på runde og kantede i forholdet 3 : 1, og at ud af de 565 planter i det andet forsøg skal $1/3$ have udelukkende rundærtede efterkommere.

Hvordan stemmer Mendels observationer overens med den genetiske models forudsigelser?

Opgave 2.6
Formulér en hensigtsmæssig statistisk model og hypotese for at besvare følgende:

Kondrodystrofi er en form for dværgvækst som regnes dominant arvelig. Genet D er sygdomsgenet og d er det tilsvarende normalgen. I en undersøgelse af en række ægtepar hvor den ene ægtefælle var kondrodystrof og den anden normal (formodet genotypekombination $\mathsf{Dd}\times\mathsf{dd}$) fandt man at blandt 27 børn var 10 kondrodystrofe og 17 normale. Er dette i strid med at kondrodystrofi arves dominant?

[At kondrodystrofi arves dominant betyder i denne forbindelse at et barn med de nævnte forældre med sandsynlighed 1/2 bliver kondrodystroft.]

Opgave 2.7
På side 24 står, at man kan bestemme $\widehat{p}$ enten som maksimumspunktet for likelihoodfunktionen eller som maksimumspunktet for log-likelihoodfunktionen, for »resultatet bliver jo det samme«; det lille ord *jo* antyder, at det er en selvfølgelighed at det forholder sig sådan. Hvorfor er det det?

Opgave 2.8 (En approksimationsformel for $-2\ln Q$)
Hvis f er en to gange kontinuert differentiabel funktion af y, så kan man som bekendt approksimere $f(y)$ med følgende rækkeudvikling (Taylorudvikling) når y er tæt på y_0:

$$f(y) \approx f(y_0) + (y - y_0)\cdot f'(y_0) + \tfrac{1}{2}(y - y_0)^2 \cdot f''(y_0).$$

Man kan anvende dette på funktionen $f(y) = -2\ln Q(y)$ hvor $-2\ln Q(y)$ er som i formel (2.1) på side 28 og hvor $y_0 = np_0$.

1. Vis at den første afledede af $-2\ln Q$ er $\ln\dfrac{y}{np_0} - \ln\dfrac{n-y}{n-np_0}$.
2. Vis at den anden afledede af $-2\ln Q$ er $\dfrac{n}{y(n-y)}$.
3. Vis derved at $-2\ln Q \approx \dfrac{(y-np_0)^2}{np_0(1-p_0)} = \left(\dfrac{y-np_0}{\sqrt{np_0(1-p_0)}}\right)^2$. (Det sidste udtryk er kvadratet på en størrelse der kan fortolkes som forskellen mellem det observerede y og den forventede værdi, divideret med standardafvigelsen på Y.)

3 Sammenligning af binomialfordelinger

I KAPITEL 2 studerede vi den simple binomialfordelingsmodel, dvs. en model hvor der var én observation y fra en binomialfordeling, én sandsynlighedsparameter p der skulle estimeres, og hvor man eventuelt kunne interessere sig for en hypotese af formen $H_0 : p = p_0$.

I dette kapitel går vi et skridt videre og betragter situationer med flere binomialfordelte observationer der kan have hver sin kendte antalsparameter og hver sin ukendte sandsynlighedsparameter. Det kan være af interesse at undersøge om sandsynlighedsparametrene kan antages at være ens, eller om de er signifikant forskellige.

Som gennemgående eksempel bruger vi stadig rismelsbille-eksemplet fra (15), men nu inddrager vi en lidt større del af datamaterialet: Man har udsat nogle rismelsbiller for gift i forskellige koncentrationer, nemlig 0.20, 0.32, 0.50 og 0.80 mg/cm^2, og dernæst set hvor mange af dem der var døde efter 13 dages forløb. (Giften strøs ud på gulvet hvor billerne færdes, derfor måles koncentrationen i mængde pr. areal.) Forsøgsresultaterne er vist i tabel 3.1.

Man kan være interesseret i at undersøge om der er forskel på virkningen af de forskellige koncentrationer. Hvis der *ikke* er nogen forskel, så skulle brøkdelen af døde i hver af de fire grupper være stort set den samme, og derfor kunne det være en god idé at udregne disse brøkdele; man får dem til 0.30, 0.72, 0.87 og 0.96. Hvis der ikke er forskel på de forskellige koncentrationer, så skal forskellighederne i disse fire tal kunne forklares udelukkende ved tilfældigheder; men hvis forskellene er så store at det er urimeligt at forklare dem ved tilfældigheder alene, så er der en *signifikant* forskel mellem koncentrationerne.

Opgaven er derfor først at opstille en statistisk model for datamaterialet, og dernæst inden for rammerne af denne model at konfrontere de foreliggende observationer med hypotesen om at der ikke er forskel på koncentrationerne.

For at vi skal kunne udtale os om hvorvidt forskellene kan forklares udelukkende ved tilfældigheder, må vi have en *statistisk model* der nærmere specificerer på hvilke punkter der kommer tilfældigheder ind i billedet. Da formålet er at sammenligne sandsynlighederne for at dø ved forskellige koncentrationer, skal modellen indrettes på den måde at totalantallene 144, 69, 54 og 50 opfattes som faste tal, hvorimod antal døde 43, 50, 47 og 48 (og dermed også antal overlevende 101, 19, 7 og 2) opfattes som fremkommet via en tilfældighedsmekanisme, i modelsprog: de er observationer af stokastiske variable. Det er nærliggende at forsøge sig med en model der går ud på, at for hver koncentration har vi en situation der svarer til en simpel binomialfordelingsmodel, og at de fire situationer er uafhængige af hverandre.

Tabel 3.1 Rismelsbillers overlevelse ved forskellige giftdoser.

	koncentration			
	0.20	0.32	0.50	0.80
antal døde	43	50	47	48
antal ikke døde	101	19	7	2
i alt	144	69	54	50

De fire grupper (»situationer«) svarende til de fire koncentrationer nummereres med index j der altså kan have værdierne $1, 2, 3, 4$. Totalantallet i gruppe j er n_j hvor $n_1 = 144$, $n_2 = 69$, $n_3 = 54$ og $n_4 = 50$. Det observerede antal døde i gruppe j er y_j hvor $y_1 = 43$, $y_2 = 50$, $y_3 = 47$ og $y_4 = 48$. Totalantallene opfattes som faste tal, men de observerede antal opfattes som oberverede værdier af stokastiske variable Y_1, Y_2, Y_3 og Y_4. At gruppe nr. j modelleres med en simpel binomialfordelingsmodel betyder at Y_j er binomialfordelt med antalsparameter n_j (kendt) og en eller anden sandsynlighedsparameter p_j som er ukendt; sandsynligheden for her at observere værdien y_j er $\mathrm{P}(Y_j = y_j) = \binom{n_j}{y_j} p_j^{y_j}(1-p_j)^{n_j-y_j}$. Hvis de fire grupper er uafhængige af hverandre, er

$$\begin{aligned}\mathrm{P}(Y_1 = y_1 \text{ og } Y_2 = y_2 \text{ og } Y_3 = y_3 \text{ og } Y_4 = y_4)\\ = \mathrm{P}(Y_1 = y_1) \cdot \mathrm{P}(Y_2 = y_2) \cdot \mathrm{P}(Y_3 = y_3) \cdot \mathrm{P}(Y_4 = y_4)\end{aligned}$$

så modelfunktionen for det samlede forsøg er

$$\begin{aligned}f(y_1, y_2, y_3, y_4; p_1, p_2, p_3, p_4) = \binom{144}{y_1} p_1^{y_1}(1-p_1)^{144-y_1} \cdot \binom{69}{y_2} p_2^{y_2}(1-p_2)^{69-y_2}\\ \cdot \binom{54}{y_3} p_3^{y_3}(1-p_3)^{54-y_3} \cdot \binom{50}{y_4} p_4^{y_4}(1-p_4)^{50-y_4}.\end{aligned}$$

Det ses at modellen indeholder fire ukendte parametre p_1, p_2, p_3 og p_4, én for hver gruppe. Opgaven er nu på grundlag af denne model plus observationerne $y_1 = 43$, $y_2 = 50$, $y_3 = 47$ og $y_4 = 48$ at estimere parametrene og at vurdere om man kan tillade sig at antage at de fire parametre i virkeligheden er ens, svarende til at giftstoffet virker ens i alle fire koncentrationer.

Vi vil vise hvordan man løser denne opgave når man benytter de principper der blev lanceret i kapitel 2. Vi vil dog gøre det en anelse mere generelt ved at se på en situation med s binomialfordelinger der skal sammenlignes.

3.1 Modellen

Antag at vi har klassificeret nogle individer i to forskellige klasser »1« og »0«. Individerne er på forhånd delt op i grupper således at der er s forskellige grupper med hhv. $n_1, n_2, \ldots, n_s$ individer. Det har vist sig at i gruppe j hører y_j af individerne til klassen

»1« og de resterende $n_j - y_j$ af individerne til klassen »0«, $j = 1, 2, \ldots, s$. Skematisk ser det sådan ud:

	gruppe nr.				
	1	2	3	...	s
klasse 1	y_1	y_2	y_3	...	y_s
klasse 0	$n_1 - y_1$	$n_2 - y_2$	$n_3 - y_3$	...	$n_s - y_s$
i alt	n_1	n_2	n_3	...	n_s

Den statistiske model der benyttes til at beskrive denne situation, er at $y_1, y_2, \ldots, y_s$ betragtes som observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_s$ der er indbyrdes uafhængige og binomialfordelte således at Y_j har antalsparameter n_j og ukendt sandsynlighedsparameter p_j, $j = 1, 2, \ldots, s$. Modellen går ud fra at grupperne er forskellige idet der er en sandsynlighedsparameter for hver gruppe. Opgaven er at undersøge om grupperne kan anses for ens, dvs. den er at teste den statistiske hypotese $H_0 : p_1 = p_2 = \cdots = p_s$.

De generelle retningslinjer for hvordan man analyserer en given statistisk model siger at vi skal tage udgangspunkt i modelfunktionen og likelihoodfunktionen. *Modelfunktionen* er den simultane sandsynlighedsfunktion for Y-erne, opfattet som en funktion af både observationer og parametre, altså

$$\begin{aligned} &f(y_1, y_2, \ldots, y_s;\, p_1, p_2, \ldots, p_s) \\ &\quad = \binom{n_1}{y_1} p_1^{y_1}(1-p_1)^{n_1-y_1} \cdot \binom{n_2}{y_2} p_2^{y_2}(1-p_2)^{n_2-y_2} \cdot \ldots \cdot \binom{n_s}{y_s} p_s^{y_s}(1-p_s)^{n_s-y_s} \\ &\quad = \prod_{j=1}^{s} \binom{n_j}{y_j} p_j^{y_j}(1-p_j)^{n_j-y_j} \ . \end{aligned}$$

Ved her at holde y-erne fast og kun opfatte udtrykket som en funktion af p-erne får vi *likelihoodfunktionen* svarende til observationerne $y_1, y_2, \ldots, y_s$:

$$L(p_1, p_2, \ldots, p_s) = \prod_{j=1}^{s} \binom{n_j}{y_j} p_j^{y_j}(1-p_j)^{n_j-y_j}$$

og dermed log-likelihoodfunktionen

$$\begin{aligned} \ln L(p_1, p_2, \ldots, p_s) &= \sum_{j=1}^{s} \ln \binom{n_j}{y_j} + \sum_{j=1}^{s} \big(y_j \ln p_j + (n_j - y_j) \ln(1-p_j) \big) \\ &= \text{konstant} + \sum_{j=1}^{s} \big(y_j \ln p_j + (n_j - y_j) \ln(1-p_j) \big). \end{aligned} \tag{3.1}$$

I bille-eksemplet bliver log-likelihoodfunktionen

$$\begin{aligned}\ln L(p_1,p_2,p_3,p_4) = \text{konstant}\\ &+ 43\ln p_1 + 101\ln(1-p_1)\\ &+ 50\ln p_2 + \ 19\ln(1-p_2)\\ &+ 47\ln p_3 + \ \ 7\ln(1-p_3)\\ &+ 48\ln p_4 + \ \ 2\ln(1-p_4).\end{aligned}$$

Likelihoodfunktionen er sandsynligheden for at observere det faktisk observerede, som funktion af det ukendte sæt parametre. Det bedste estimat over de ukendte parametres værdier er det talsæt $(\widehat{p}_1,\widehat{p}_2,\ldots,\widehat{p}_s)$ som maksimaliserer likelihoodfunktionen eller log-likelihoodfunktionen. Log-likelihoodfunktionen er en funktion af s variable, men heldigvis en meget skikkelig funktion idet den (bortset fra et konstantled) er en sum af s led der hver især kun er en funktion af én variabel. Det j-te led hedder $y_j\ln p_j+(n_j-y_j)\ln(1-p_j)$, og vi ved fra tidligere (side 24) at dette udtryk antager sit maksimum når $p_j = y_j/n_j$. Vi har hermed fundet at *maksimaliseringsestimatet* for $(p_1,p_2,\ldots,p_s)$ er

$$(\widehat{p}_1,\widehat{p}_2,\ldots,\widehat{p}_s) = \left(\frac{y_1}{n_1},\frac{y_2}{n_2},\ldots,\frac{y_s}{n_s}\right).$$

I eksemplet er specielt $(\widehat{p}_1,\widehat{p}_2,\widehat{p}_3,\widehat{p}_4) = (0.30, 0.72, 0.87, 0.96)$.

3.2 Hypoteseprøvning

Vi skal undersøge om det er rimeligt at antage at hypotesen $H_0 : p_1 = p_2 = \cdots = p_s$ om ens sandsynlighedsparametre holder. Under H_0 er der ingen forskel på de s grupper, og i så fald kan vi slå dem sammen til én stor gruppe bestående af $n_\bullet = n_1+n_2+\cdots+n_s$ individer der fordeler sig med $y_\bullet = y_1, y_2, \ldots, y_s$ individer i klassen »1« og resten, dvs. $n_\bullet - y_\bullet$, i klassen »0«. Derfor må man formode at den fælles værdi p af sandsynlighedsparameteren skal estimeres ved $y_\bullet/n_\bullet$, men lad os benytte likelihoodmetoden og se hvad den siger om den sag.

Vi kalder den fælles værdi (under H_0) af $p_1, p_2, \ldots, p_s$ for p. I den oprindelige log-likelihoodfunktion (3.1) erstatter vi alle p_j-erne med p og får derved *log-likelihoodfunktionen under* H_0 svarende til observationerne $y_1, y_2, \ldots, y_s$:

$$\begin{aligned}\ln L(p,p,\ldots,p) &= \text{konstant} + \sum_{j=1}^{s}\bigl(y_j\ln p + (n_j-y_j)\ln(1-p)\bigr)\\ &= \text{konstant} + \ y_\bullet\ln p + (n_\bullet - y_\bullet)\ln(1-p).\end{aligned}$$

Maksimaliseringsestimatet $\widehat{p}$ for p er den værdi der maksimaliserer denne log-likelihoodfunktion, dvs. den værdi p der maksimaliserer $y_\bullet\ln p + (n_\bullet - y_\bullet)\ln(1-p)$. Vi ved fra side 24 at løsningen er $\widehat{p} = y_\bullet/n_\bullet$. Likelihoodmetoden giver altså det svar som vi formodede måtte være det rigtige. – I vort eksempel bliver $\widehat{p} = {}^{188}/_{317} = 0.59$.

Tabel 3.2 Rismelsbillers overlevelse ved forskellige giftdoser: forventede antal hvis giften virker på samme måde for alle fire koncentrationer.

	koncentration			
	0.20	0.32	0.50	0.80
antal døde	85.4	40.9	32.0	29.7
antal ikke døde	58.6	28.1	22.0	20.3
i alt	144	69	54	50

Likelihoodfunktionen benyttes til at vurdere et sæt parameterværdiers evne til at beskrive det faktisk observerede. Det bedste sæt parameterværdier overhovedet er $(\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_s)$. Under H_0 er det bedste sæt værdier $(\widehat{p}, \widehat{p}, \ldots, \widehat{p})$. Vi sammenligner disse to parametersæts beskrivelsesevne ved hjælp af *kvotienteststørrelsen*

$$Q = \frac{L(\widehat{p}, \widehat{p}, \ldots \widehat{p})}{L(\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_s)} ,$$

der bliver et tal mellem 0 og 1; en Q-værdi tæt på 1 betyder at sættet $(\widehat{p}, \widehat{p}, \ldots, \widehat{p})$ beskriver det observerede næsten lige så godt som $(p_1, p_2, \ldots, p_s)$ gør, dvs. vi kan godtage hypotesen H_0, hvorimod en Q-værdi langt fra 1 betyder at H_0 giver en væsentligt dårligere beskrivelse af det observerede end grundmodellen gør. Som oftest udregner man dog ikke Q, men $-2\ln Q$; den bliver

$$\begin{aligned} -2\ln Q &= -2\big(\ln L(\widehat{p}, \widehat{p}, \ldots, \widehat{p}) - \ln L(\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_s)\big) \\ &= 2\sum_{j=1}^{s}\Big(y_j \ln \frac{\widehat{p}_j}{\widehat{p}} + (n_j - y_j)\ln \frac{1-\widehat{p}_j}{1-\widehat{p}}\Big). \end{aligned}$$

Tallet $-2\ln Q$ vil altid være større end eller lig nul.

Med betegnelsen $\widehat{y}_j = n_j\widehat{p}$ kan $-2\ln Q$ omskrives til

$$-2\ln Q = 2\sum_{j=1}^{s}\Big(y_j \ln \frac{y_j}{\widehat{y}_j} + (n_j - y_j)\ln \frac{n_j - y_j}{n_j - \widehat{y}_j}\Big); \tag{3.2}$$

man kan tænke på $\widehat{y}_j$ som det »forventede« antal individer fra gruppe j der klassificeres som »1« og på $n_j - \widehat{y}_j$ som det »forventede« antal individer fra gruppe j der klassificeres som »0«.

De »forventede« antal i bille-eksemplet er vist i tabel 3.2, og man får værdien af teststørrelsen til

$$\begin{aligned} -2\ln Q_{\text{obs}} = 2\Big(&43\ln\frac{43}{85.4} + 101\ln\frac{101}{58.6} + 50\ln\frac{50}{40.9} + \;\; 19\ln\frac{19}{28.1} + \\ &47\ln\frac{47}{32.0} + \;\; 7\ln\frac{7}{22.0} + 48\ln\frac{48}{29.7} + \;\; 2\ln\frac{2}{20.3}\Big) \\ = 113.1& \end{aligned}$$

En Q-værdi tæt på 1 svarer til en $-2\ln Q$-værdi tæt på 0. Det vil sige at hvis $-2\ln Q_{\text{obs}}$ er tæt på 0, så kan vi godtage H_0, hvorimod en stor værdi af $-2\ln Q_{\text{obs}}$ tyder på en signifikant afvigelse mellem det observerede og det som H_0 foreskriver, dvs. vi må forkaste H_0. For at afgøre om tallet $-2\ln Q_{\text{obs}}$ er stort eller lille, er vi nødt til at sammenligne det med alle de andre værdier man også kunne have fået ifølge den aktuelle model når H_0 er rigtig. Derfor skal vi bestemme *testsandsynligheden* ε som er sandsynligheden for at få noget værre end det faktisk observerede, dvs. for at få en større $-2\ln Q$-værdi end den observerede, under forudsætning af at H_0 er rigtig:

$$\varepsilon = \mathrm{P}_0\big(-2\ln Q \geq -2\ln Q_{\text{obs}}\big).$$

Mere udførligt er ε defineret på følgende måde: Den statistiske model siger at observationerne $y_1, y_2, \ldots, y_s$ er observerede værdier af stokastiske variable $Y_1, Y_2, \ldots, Y_s$ der er binomialfordelte med antalsparametre $n_1, n_2, \ldots, n_s$ og – da H_0 antages rigtig – med samme sandsynlighedsparameter p. Testsandsynligheden ε er sandsynligheden for at disse stokastiske variable antager værdier som giver anledning til en $-2\ln Q$-værdi der er større end den faktisk observerede $-2\ln Q_{\text{obs}}$. Bestemmelsen af ε kan synes at være en besværlig opgave, og den kompliceres endda yderligere af at selv når H_0 er rigtig, er der en ukendt parameter inde i billedet, nemlig den fælles sandsynlighedsparameter p; hvis det skal være helt rigtigt, er vi således ikke i stand til at udregne testsandsynligheden!

Heldigvis kommer den matematiske statistik os til undsætning med et generelt resultat der fortæller at når H_0 er rigtig, så er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med et antal frihedsgrader som er $s-1$. Det betyder at testsandsynligheden ε med god tilnærmelse kan bestemmes som sandsynligheden for at få en værdi større end $-2\ln Q_{\text{obs}}$ i en χ^2-fordeling med $s-1$ frihedsgrader, kort

$$\varepsilon = \mathrm{P}\big(\chi^2_{s-1} \geq -2\ln Q_{\text{obs}}\big),$$

og den sandsynlighed er let at bestemme, f.eks. ved hjælp af tabeller over fraktiler i χ^2-fordelingen.

Antallet af *frihedsgrader* for $-2\ln Q$ findes som ændringen i antallet af frie parametre: i grundmodellen er der s frie parametre $p_1, p_2, \ldots, p_s$, under H_0 er der én fri parameter p, derfor bliver der $s-1$ frihedsgrader til teststørrelsen.

I eksemplet er $-2\ln Q_{\text{obs}} = 113.1$ og der er fire grupper, dvs. teststørrelsen har tre frihedsgrader. I en tabel over fraktiler i χ^2-fordelingen (se f.eks. side 222) ses at værdien 113.1 er langt større end 99.5%-fraktilen i χ^2-fordelingen med tre frihedsgrader, og det vil sige at testsandsynligheden ε er langt mindre end 0.5%. Værdien 113.1 er altså så stor at der, under forudsætning af at hypotesen er rigtig, kun er en helt mikroskopisk chance for at få en endnu større værdi, dvs. 113.1 er en særdeles stor værdi. Vi må derfor forkaste hypotesen H_0, eller sagt på en anden måde: Der er en signifikant forskel på de fire giftkoncentrationer.

Som nævnt er χ^2-fordelingen kun en approksimation til den rigtige fordeling af $-2\ln Q$. *For at approksimationen skal kunne bruges, skal alle de »forventede« antal* $\widehat{y}_j$ *og* $n_j - \widehat{y}_j$, $j = 1, 2, \ldots, s$ *være mindst fem.* Hvis denne betingelse ikke er opfyldt, kan man eventuelt udelade de problematiske grupper eller slå nogle af grupperne sammen

Tabel 3.3 Fordeling efter køn i to projektgrupper.

	gruppe 1	gruppe 2	i alt
drenge	2	6	8
piger	4	3	7
i alt	6	9	15

på forhånd. Hvis der kun er to grupper i det hele taget, kan man anvende det såkaldte *Fishers eksakte test* der omtales i afsnit 3.3.

3.3 Det eksakte test i en 2×2-tabel

I visse tilfælde er det ikke forsvarligt at anvende χ^2-approksimationen til fordelingen af $-2 \ln Q$, nemlig når nogle af de »forventede« antal er små. Vi skal nu omtale hvordan man kan sammenligne *to* binomialfordelinger selv om nogle af de forventede antal er under fem.

Tag som eksempel en situation hvor man på grundlag af tallene i tabel 3.3 ønsker at vurdere om der er signifikant forskel på kønsfordelingen i to projektgrupper. Ved at efterligne ræsonnementerne i begyndelsen af kapitlet kan man nå frem til følgende (forslag til den) statistiske model for disse observationer:

> De observerede antal drenge $y_1 = 2$ og $y_2 = 6$ opfattes som observationer af stokastiske variable Y_1 og Y_2 der er stokastisk uafhængige og binomialfordelte med antalsparametre $n_1 = 6$ og $n_2 = 9$ og med ukendte sandsynlighedsparametre p_1 hhv. p_2.

Den tilsvarende *modelfunktion* er

$$f(y_1, y_2; p_1, p_2) = \binom{6}{y_1} p_1^{y_1}(1-p_1)^{6-y_1} \cdot \binom{9}{y_2} p_2^{y_2}(1-p_2)^{9-y_2}.$$

Maksimaliseringsestimaterne for p_1 og p_2 er $\widehat{p}_1 = {}^2\!/_6 = {}^1\!/_3$ og $\widehat{p}_2 = {}^6\!/_9 = {}^2\!/_3$.

Lad os sætte at opgaven er at undersøge om der er en signifikant forskel på kønsfordelingen i de to grupper, eller modsat at de observerede forskelle ikke er andet end hvad man kan komme ud for på grund af tilfældigheder. Vi vil derfor teste den statistiske hypotese $H_0 : p_1 = p_2$.

Problemet

Da vi har at gøre med et specialtilfælde af det generelle problem »sammenligning af binomialfordelinger« der blev behandlet tidligere i kapitlet, kan vi nu blot gå frem efter opskriften. Maksimaliseringsestimatet for den fælles værdi under H_0 af p_1 og p_2 er $\widehat{p} = {}^8\!/_{15} = 0.53$, og de »forventede« antal $\widehat{y}_1 = n_1\widehat{p}$, $n_1 - \widehat{y}_1 = n_1(1-\widehat{p})$, $\widehat{y}_2 = n_2\widehat{p}$ og

Tabel 3.4 Forventet kønsfordeling under H_0 i de to projektgrupper.

	gruppe 1	gruppe 2	i alt
drenge	3.2	4.8	8
piger	2.8	4.2	7
i alt	6	9	15

$n_2 - \widehat{y}_1 = n_2(1-\widehat{p})$ bliver som vist i tabel 3.4. Teststørrelsen $-2\ln Q$ er dermed

$$\begin{aligned} -2\ln Q &= 2\sum\left(\text{obs. antal}\cdot\ln\frac{\text{obs. antal}}{\text{forv. antal}}\right) \\ &= 2\left(2\ln\frac{2}{3.2}+4\ln\frac{4}{2.8}+6\ln\frac{6}{4.8}+3\ln\frac{3}{4.2}\right) = 1.63. \end{aligned}$$

Store værdier af $-2\ln Q$ tyder på at hypotesen H_0 ikke holder; for at afgøre om 1.63 er en »stor« værdi, skal vi bestemme testsandsynligheden ε, dvs. sandsynligheden for at få en $-2\ln Q$-værdi som er større end 1.63 under forudsætning af at H_0 er rigtig, dvs. $\varepsilon = \mathrm{P}_0(-2\ln Q \geq 1.63)$. Der gælder at hvis de »forventede antal« alle er mindst fem, så kan ε med god tilnærmelse findes som sandsynligheden for at få en værdi på mindst 1.63 i en χ^2-fordeling med 1 frihedsgrad. *Men* i det foreliggende tilfælde er ingen af de »forventede« antal over fem, så vi kan *ikke* gå ud fra at χ^2-approksimationen er anvendelig.

Et betinget test

Derfor må man prøve at udregne ε fra 'first principles'. Hvis man udtrykker $-2\ln Q$ ved y_1 og y_2, får man (jf. (3.2) på side 37)

$$\begin{aligned} -2\ln Q(y_1,y_2) = 2\Bigg(&y_1\ln\frac{y_1}{n_1\frac{y_\bullet}{n_\bullet}} + (n_1-y_1)\ln\frac{n_1-y_1}{n_1(1-\frac{y_\bullet}{n_\bullet})} + \\ &y_2\ln\frac{y_2}{n_2\frac{y_\bullet}{n_\bullet}} + (n_2-y_2)\ln\frac{n_2-y_2}{n_2(1-\frac{y_\bullet}{n_\bullet})}\Bigg), \end{aligned}$$

hvor $y_\bullet = y_1 + y_2$ og $n_\bullet = n_1 + n_2$. Her kan talparret (y_1, y_2) antage 70 forskellige sæt værdier svarende til at $y_1 \in \{0,1,2,\ldots,6\}$ og $y_2 \in \{0,1,2,\ldots,9\}$. Man kan så udregne $-2\ln Q$ for hvert af de 70 mulige udfald og derved bestemme de udfald (y_1, y_2) for hvilke $-2\ln Q(y_1,y_2)$ er mindst 1.63. Man finder at det er de par (y_1, y_2) som er markeret med $\star$ i figur 3.1. Testsandsynligheden ε kan derefter findes som summen af sandsynlighederne $f(y_1, y_2; p, p)$ for alle udfald (y_1, y_2) for hvilke $-2\ln Q(y_1,y_2) \geq 1.63$. Denne fremgangsmåde indebærer som man hurtigt vil erfare, en hel del regnearbejde, men der er også en komplikation af mere fundamental karakter.

I kapitel 2 testede vi hypoteser gående ud på at den eneste ukendte parameter havde en bestemt på forhånd givet værdi. Når en sådan hypotese var rigtig, var der ikke flere ukendte parametre inde i billedet – den slags hypoteser kaldes *simple hypoteser*. De hypoteser vi nu har med at gøre, er af en anden slags: Der er tale om modeller med

	$y_1 =$ 0	1	2	3	4	5	6
$y_2 = 0$	·	·	⋆	⋆	⋆	⋆	⋆
1	·	·	·	⋆	⋆	⋆	⋆
2	⋆	·	·	·	⋆	⋆	⋆
3	⋆	·	·	·	⋆	⋆	⋆
4	⋆	·	·	·	·	⋆	⋆
5	⋆	⋆	·	·	·	·	⋆
6	⋆	⋆	⋆	·	·	·	⋆
7	⋆	⋆	⋆	·	·	·	⋆
8	⋆	⋆	⋆	⋆	·	·	·
9	⋆	⋆	⋆	⋆	⋆	·	·

Figur 3.1 Talpar (y_1, y_2) for hvilke $-2 \ln Q(y_1, y_2) \geq 1.63$ er markeret med $\star$.

mere end én ukendt parameter, og hypoteserne går ud på at nogle af disse parametre er ens; men selv når hypotesen er rigtig, er der stadigvæk ukendte parametre i modellen. – Denne slags hypoteser kaldes *sammensatte hypoteser*.

I det aktuelle hypoteseprøvningsproblem der altså handler om en sammensat hypotese, nåede vi ovenfor frem til at testsandsynligheden ε måtte skulle bestemmes som en sum af nogle sandsynligheder $f(y_1, y_2; p, p)$ hvor der summeres over en vis mængde (y_1, y_2)-er, og hvor der indgår den fælles men *ukendte* parameter p. For at beregne ε skal vi altså kende (den sande værdi af) den ukendte parameter p! Nu ville læseren måske nok uden at blegne indsætte værdien af $\widehat{p}$ (som er $^8/_{15}$) og så udregne ε på det grundlag (hvorved man får ε til 27%), men det ændrer ikke ved det principielle problem. Der findes imidlertid en fremgangsmåde ved hjælp af hvilken man helt kan eliminere det famøse p.

Parameteren p er sandsynligheden for at en tilfældigt valgt person er en dreng, når de to grupper er ens. Den i observationsmaterialet indeholdte information om dette p er at der ud af de i alt 15 personer viste sig at være netop 8 drenge. Man kan nu sige at det er uinteressant at der netop er 8 (og ikke 7 eller 10) drenge; det interessante er at de 8 er fordelt med 2 i gruppe 1 og 6 i gruppe 2 . Derfor skal man (sådan siger et statistisk princip) betragte *den betingede fordeling* givet at der netop var 8 drenge. I denne betingede fordeling vil det vise sig at den oprindelige sammensatte hypotese H_0 bliver til en simpel hypotese. For at se hvordan det går til, må vi oversætte det netop sagte til matematik.

Modelfunktionen i grundmodellen er som allerede nævnt

$$f(y_1, y_2; p_1, p_2) = \binom{6}{y_1} p_1^{y_1}(1-p_1)^{6-y_1} \cdot \binom{9}{y_2} p_2^{y_2}(1-p_2)^{9-y_2}.$$

Når H_0 er rigtig, har p_1 og p_2 den fælles værdi p, og modelfunktionen kommer så til at

se sådan ud:

$$\begin{aligned} f(y_1, y_2; p, p) &= \binom{6}{y_1} p^{y_1}(1-p)^{6-y_1} \cdot \binom{9}{y_2} p^{y_2}(1-p)^{9-y_2} \\ &= \binom{6}{y_1}\binom{9}{y_2} \cdot p^{y_1+y_2}(1-p)^{15-(y_1+y_2)}. \end{aligned}$$

Heraf fremgår at likelihoodfunktionen under H_0 er

$$L(p) = \text{konstant} \cdot p^{y_1+y_2}(1-p)^{15-(y_1+y_2)} \ ,$$

dvs. man kan bestemme likelihoodfunktionen (pånær en konstant faktor), blot man kender det totale antal drenge $y_1 + y_2$ – man behøver ikke kende y_1 og y_2 hver for sig.

Det snedige trick er nu at se på den *betingede fordeling* af Y_1 og Y_2 givet at $Y_1+Y_2 = 8$, altså givet at der er netop 8 drenge i alt. Påstanden er at i denne betingede fordeling bliver hypotesen H_0 til en simpel hypotese. For at indse det vil vi bestemme den betingede fordeling af Y_1 og Y_2 givet at $Y_1 + Y_2 = 8$. Ifølge de sædvanlige formler for betingede sandsynligheder er den betingede sandsynlighed for at $Y_1 = y_1$ og $Y_2 = y_2$ givet at $Y_1 + Y_2 = 8$

$$\mathrm{P}\big(Y_1 = y_1, Y_2 = y_2 \mid Y_1 + Y_2 = 8\big) = \begin{cases} \dfrac{\mathrm{P}(Y_1 = y_1) \cdot \mathrm{P}(Y_2 = 8 - y_1)}{\mathrm{P}(Y_1 + Y_2 = 8)} & \text{hvis } y_1 + y_2 = 8 \\ 0 & \text{hvis } y_1 + y_2 \neq 8, \end{cases}$$

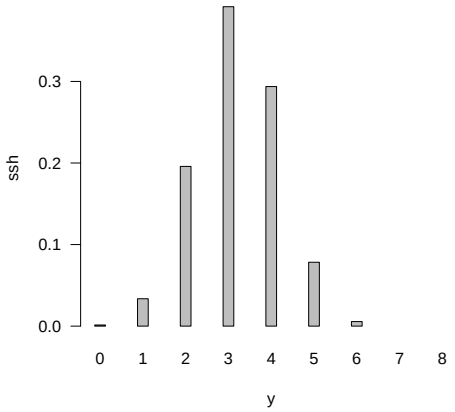
og udtrykket svarende til tilfældet $y_1 + y_2 = 8$ kan videre omskrives således (hvor y_1 erstattes af y):

$$\begin{aligned} \frac{\mathrm{P}(Y_1 = y_1) \cdot \mathrm{P}(Y_2 = 8 - y_1)}{\mathrm{P}(Y_1 + Y_2 = 8)} &= \frac{f(y, 8-y; p_1, p_2)}{\displaystyle\sum_{z=0}^{8} f(z, 8-z; p_1, p_2)} \\ &= \frac{\displaystyle\binom{6}{y} p_1^y (1-p_1)^{6-y} \cdot \binom{9}{8-y} p_2^{8-y}(1-p_2)^{9-(8-y)}}{\displaystyle\sum_{z=0}^{8} \binom{6}{z} p_1^z (1-p_1)^{6-z} \cdot \binom{9}{8-z} p_2^{8-z}(1-p_2)^{9-(8-z)}} \\ &= \frac{\displaystyle\binom{6}{y}\binom{9}{8-y}\theta^y}{\displaystyle\sum_{z=0}^{8} \binom{6}{z}\binom{9}{8-z}\theta^z}, \end{aligned}$$

hvor

$$\theta = \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2} = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

	y	$Q(y)$	$\overline{f}(y;1)$
	0	0.002	0.001
	1	0.069	0.034
$\rightarrow$	2	0.468	0.196
	3	0.979	0.392
	4	0.713	0.294
	5	0.166	0.078
	6	0.006	0.006
	7	-	0
	8	-	0
			1.001



Figur 3.2 Tabel over $Q(y)$ og $\overline{f}(y;1)$, samt pindediagram over $\overline{f}(y;1)$.

Det ses at hvor grundmodellen har to ukendte parametre p_1 og p_2, har den betingede model kun én parameter, nemlig θ. Modelfunktionen i den betingede model er

$$\overline{f}(y;\theta) = \frac{\binom{6}{y}\binom{9}{8-y}\theta^y}{\sum_{z=0}^{8}\binom{6}{z}\binom{9}{8-z}\theta^z}.$$

Af definitionen af θ følger at grundmodellens hypotese $H_0 : p_1 = p_2$ er ensbetydende med hypotesen $\overline{H}_0 : \theta = 1$ i den betingede model. Den sammensatte hypotese i grundmodellen er altså blevet til en simpel hypotese i den betingede model.

Vi kan nu teste hypotesen $\overline{H}_0$ ved brug af de sædvanlige principper, og da $\overline{H}_0$ er en simpel hypotese, er der ikke nogen principielle problemer. Der foreligger observationen $y = 2$; det tilsvarende estimat $\widehat{\theta}$ over θ er den θ-værdi der maksimaliserer den betingede likelihoodfunktion $\overline{L}(\theta) = \overline{f}(2;\theta)$, dvs. den θ-værdi som er løsning til $\frac{d}{d\theta}\overline{L}(\theta) = 0$. Man finder at $\widehat{\theta} = \widehat{\theta}(2) = 0.276$. Kvotienttesstørrelsen for $\overline{H}_0$ er

$$Q = Q(2) = \frac{\overline{L}(1)}{\overline{L}(\widehat{\theta}(2))} = \frac{\overline{L}(1)}{\overline{L}(0.276)} = 0.468.$$

Hvis Q er langt fra 1, er det tegn på at $\overline{H}_0$ skal forkastes. For at vurdere om 0.468 ligger langt fra 1, skal vi udregne testsandsynligheden ε som er sandsynligheden (under $\overline{H}_0$) for at få et y således at $Q(y)$ er mindre end eller lig med 0.468:

$$\varepsilon = \sum_{y\,:\,Q(y)\leq 0.468} \overline{f}(y;1).$$

Bestemmelsen af ε er ukompliceret, men noget besværlig. Af tabel 3.2 ses at de y-er som giver en Q-værdi der er mindre end eller lig $Q(2) = 0.468$, dvs. de y-er der

er mindst lige så uforenelige med $\overline{H}_0$ som $y = 2$ er, er y-erne $0, 1, 2, 5, 6$, således at testsandsynligheden er

$$\varepsilon = \overline{f}(0;1) + \overline{f}(1;1) + \overline{f}(2;1) + \overline{f}(5;1) + \overline{f}(6;1) = 0.315.$$

Der er altså ca. 31% chance for at få et y der er »lige så slemt eller værre« end det observerede $y = 2$ når $\overline{H}_0$ er rigtig. Man vil derfor sige at der *ikke* er nogen signifikant uoverensstemmelse mellem hypotesen $\overline{H}_0$ og det observerede $y = 2$. Sagt på en anden måde: vi kan ikke forkaste $\overline{H}_0$.

Vi er gået let hen over hvordan man egentlig skal finde talværdien af $\widehat{\theta}$ og hvordan man egentlig beregner værdier af funktionerne $\overline{L}$ og $\overline{f}$. Grunden hertil er at den just beskrevne metode, som er den principielt rigtigste, faktisk sædvanligvis ikke bruges. Den er nemlig besværlig rent regnemæssigt såfremt man skal regne med håndkraft. Det er ganske vist ingen sag at skrive et lille computerprogram der kan udføre beregningerne, men man bruger alligevel (endnu) for det meste en regnemæssigt simplere metode som vi nu vil beskrive i detaljer.

Fishers eksakte test

Det man gør når man tester en statistisk hypotese, er at man udregner værdien af en vis teststørrelse, sædvanligvis kvotienteststørrelsen Q eller $-2\ln Q$, som er et udtryk for hvor godt hypotesen er forenelig med de foreliggende data; dernæst bestemmer man testsandsynligheden, altså sandsynligheden for at få et sæt observationer som er mindst lige så »uforenelige« med hypotesen som de faktiske observationer er. I den simplere metode der nu skal omtales til løsning af det aktuelle testproblem, benytter man ikke Q som teststørrelse, men derimod sandsynlighedsfunktionen $\overline{f}(\centerdot\,;1)$ svarende til at hypotesen $\overline{H}_0$ er rigtig; det har blandt andet den fordel at man slipper for at skulle bestemme $\widehat{\theta}$. Funktionen $\overline{f}(\centerdot\,;1)$ er forholdsvis simpel:

$$\overline{f}(y;1) = \frac{\binom{6}{y}\binom{9}{8-y}}{\sum_{z=0}^{8}\binom{6}{z}\binom{9}{8-z}}. \tag{3.3}$$

(I øvrigt er funktionen $y \mapsto \overline{f}(y;1)$ sandsynlighedsfunktion for en *hypergeometrisk fordeling*, se opgave 1.7; der gælder at nævneren er lig $\binom{15}{8}$.)

Fishers eksakte test for $\overline{H}_0$ forløber nu på følgende måde: Vi har observeret $y = 2$. Vi skal bestemme de y-er for hvilke $\overline{f}(y;1) \le \overline{f}(2;1)$. For at gøre det udregner vi tælleren i højresiden af formel (3.3) for alle de mulige y-er, f.eks. ved brug af Pascals trekant (side 13); man får da tabel 3.5. Det ses at de y-er som er mere ekstreme end $y = 2$ (ekstreme i den forstand at $\overline{f}(y;1) \le \overline{f}(2;1) = 1260/6435$) er alle y-erne undtagen $y = 3$ og $y = 4$. Testsandsynligheden er derfor

$$\varepsilon = 1 - \big(\overline{f}(3;1) + \overline{f}(4;1)\big) = 1 - \frac{2520 + 1890}{6435} = 31\%.$$

Det eksakte test giver således (i dette eksempel) præcis samme resultat som det rigtige betingede test.

Tabel 3.5 Hjælpestørrelser til Fishers eksakte test.

y	$\binom{6}{y}\cdot\binom{9}{8-y}$			
0	1	· 9	=	9
1	6	· 36	=	216
2	15	· 84	=	1260
3	20	· 126	=	2520
4	15	· 126	=	1890
5	6	· 84	=	504
6	1	· 36	=	36
7	0	· 9	=	0
8	0	· 1	=	0
				6435

Hvad angår det oprindelige praktiske problem, kan vi i første omgang konkludere at $\overline{H}_0$ må accepteres, dvs. der er ikke nogen signifikant forskel på kønsfordelingen i de to grupper set fra den betingede models synspunkt. Da man kan sige at det der adskiller den betingede model og den oprindelige (ubetingede) model, er noget som er uinteressant for spørgsmålet om ens kønsfordeling i de to grupper, kan vi videre konkludere at også H_0 må accepteres, dvs. heller ikke fra grundmodellens synspunkt er der nogen signifikant forskel på kønsfordelingen i de to grupper.

3.4 Regn og tegn

Sammenligning af binomialfordelinger kan foretages med R-funktionen `prop.test`. Det gennemgåede eksempel kan behandles sådan:

```
y <- c(43, 50, 47, 48)
n <- c(144, 69, 54, 50)
prop.test (y, n)
```

hvilket resulterer i

```
  4-sample test for equality of proportions without continuity correction

data:  y out of n
X-squared = 101.783, df = 3, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4
0.2986111 0.7246377 0.8703704 0.9600000
```

De fire værdier der står under 'sample estimates', er $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \widehat{p}_4)$. Teststørrelsen X^2 (`X-squared`) er en *approksimation* (jf. opgave 3.5) til $-2\ln Q$, antallet af frihedsgrader er `df`, og testsandsynligheden er `p-value`.

Den rigtige $-2\ln Q$ med tilhørende testsandsynlighed kan udregnes sådan:

```
yhat <- n*sum(y)/sum(n)
testst <- 2*sum( y * log(y/yhat) + (n-y) * log((n-y)/(n-yhat)) )
1 - pchisq(testst, 3)    # testsandsynligheden
```

Fishers eksakte test udføres med funktionen `fisher.test`. I det gennemgåede eksempel kan det se sådan ud:

```
fisher.test (matrix (c (2, 4, 6, 3), nrow=2))
```

som giver følgende resultat:

```
Fisher's Exact Test for Count Data

data:  matrix(c(2, 4, 6, 3), nrow = 2)
p-value = 0.3147
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.01561152 3.31197442
sample estimates:
odds ratio
0.2762995
```

hvilket viser at testsandsynligheden bliver 0.3147.

3.5 Opgaver

Generelt om opgavebesvarelser

Mange statistikopgaver består af et datasæt plus en kort beskrivelse af det eksperiment eller den indsamlingsproces der frembragte dem, efterfulgt af en lakonisk besked af typen »Analysér data!« Desuden er der et eller andet (ikke altid lige klart præciseret) overordnet spørgsmål der skal besvares/belyses på baggrund af en statistisk analyse af det foreliggende datasæt.

Selv om man ikke kan (eller bør) give en generel skabelon for udformningen af besvarelsen af sådanne opgaver, kan det måske være praktisk med en »huskeliste« med punkter der ofte skal med i løsningen. Her er en sådan liste:

1. Beskriv i ord en passende statistisk model. – En »passende« model er en der dels kan tænkes at beskrive tallene, dels gør det muligt at besvare det overordnede spørgsmål.
2. Formulér modellen i matematiksprog.
3. Estimér parametrene.
4. Formulér det overordnede spørgsmål i matematiksprog, og omsæt det til en statistisk hypotese.
5. Estimér eventuelle parametre under hypotesen.
6. Udregn teststørrelsen $(-2 \ln Q)$ og find den tilsvarende testsandsynlighed.
7. Vurdér om den statistiske hypotese skal forkastes eller ej.
8. Find ud af hvad man kan konkludere om det overordnede spørgsmål.
9. Formulér konklusionen i ord.

Opgave 3.1 (Afstemning i Lejre)
Ved EF-folkeafstemningen den 2. juni 1992 om Maastricht-traktaten fordelte ja- og nejstemmerne sig på følgende måde ved de fem afstemningssteder i Lejre kommune:

	Gevninge	Herslev	Lejre	Osted	Glim
Antal gyldige ja stemmer	830	194	800	931	448
Antal gyldige nej stemmer	621	151	605	738	344

Kan man på denne baggrund sige at der er forskel på holdningen til traktaten i de fem dele af kommunen?

Opgave 3.2 (Kødkvalitet)
Ved den kødkontrol som foretages af dyrlæger på slagterier, udføres for visse dyr en bakteriologisk undersøgelse (BU) efter regler fastsatte af veterinærdirektoratet. Resultatet af undersøgelsen kan for hvert dyr noget forenklet beskrives som »godkendt« eller »kasseret«.

For bl.a. at finde ud af om der var nogen sammenhæng mellem slagteri og resultatet af BU, undersøgte man resultaterne af undersøgelserne for 672 dyr der var indsendt til et bestemt laboratorium fra forskellige slagterier. En stor del af dyrene kom fra to bestemte slagterier kaldet I og II. Man fik følgende fordeling efter BU-udfald og slagteri:

	slagteri I	slagteri II	øvrige slagterier
godkendt	134	275	146
kasseret	49	41	27

Blandt de diagnoser som kan give anledning til at der udføres BU, var halebid den hyppigst forekommende. For de 174 dyr som havde diagnose halebid, fik man følgende fordeling:

	slagteri I	slagteri II	øvrige slagterier
godkendt	30	82	25
kasseret	19	13	5

Analysér data.

Opgave 3.3 (Kampflyveres børn)
Blandt piloter i luftvåbenet siges det at pilot-børn oftere er piger end drenge. – I 1961 indsamledes data om nyfødte børn hvis fædre gjorde tjeneste som piloter i US Airforce, og man inddelte blandt andet børnene i grupper efter arten af flyvetjeneste som faderen havde haft i den måned hvor barnet blev undfanget. Det gav denne tabel:

	faderens tjeneste var		
barnets køn	i jagerfly	i transportfly	jordtjeneste
pige	51	14	38
dreng	38	16	46

Undersøg om der er hold i påstanden om at pilotfædre får flere piger end drenge.

I den samme periode var 48.7% af alle nyfødte (i USA) piger. Hvordan harmonerer pilot-dataene med dette tal?

Opgave 3.4 (Bakterier og forstørrede mandler)
Nogle mennesker er bærere af bakterien *Streptococcus pyogenes*. For at finde ud af om dette især er tilfældet for mennesker med forstørrede mandler, undersøgte man nogle børn i alderen 0-15 år. I undersøgelsen var der 497 børn hvis mandler havde normal størrelse, og af disse børn

var de 19 bærere af bakterien. Desuden var der 589 børn med noget forstørrede mandler, og heraf var de 29 bærere af bakterien. Endelig var der 293 børn med meget forstørrede mandler, og heraf var de 24 bærere af bakterien.

Tyder disse resultater på at det især er børn med forstørrede mandler der er bærere af *Streptococcus pyogenes*?

Opgave 3.5 (En approksimationsformel for $-2\ln Q$)
Denne opgave skal opfattes som en udvidelse af opgave 2.8. Formålet er at udlede en approksimation til teststørrelsen $-2\ln Q$ (formel (3.2) på side 37).

Betragt funktionen $f(y) = y\ln(y/y_0)$ hvor y_0 er en konstant.

1. Vis at $f'(y) = 1 + \ln(y/y_0)$ og at $f''(y) = 1/y$.
2. Vis at Taylorudviklingen af f omkring y_0 er
$$\begin{aligned} f(y) &\approx f(y_0) + (y-y_0)\cdot f'(y_0) + \tfrac{1}{2}(y-y_0)^2\cdot f''(y_0) \\ &= (y-y_0) + \tfrac{1}{2}\frac{(y-y_0)^2}{y_0}. \end{aligned}$$
3. Anvend ovennævnte approksimationsformel på hvert af leddene $y_j \ln\dfrac{y_j}{\widehat{y}_j}$ og $(n_j - y_j)\ln\dfrac{n_j - y_j}{n_j - \widehat{y}_j}$ i udtrykket for $-2\ln Q$, og vis derved at man kan approksimere $-2\ln Q$ med den såkaldte *Pearsons* X^2 defineret som $X^2 = \displaystyle\sum_{j=1}^{s}\frac{(y_j - \widehat{y}_j)^2}{n_j\widehat{p}(1-\widehat{p})}$ (opkaldt efter den engelske videnskabsmand Karl Pearson (1857-1936)).

4 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 tallinjen fordelingen er beliggende, og en ændring af parameterværdien skal svare til en forskydning af sandsynlighedsfordelingen hen ad tallinjen 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.*

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

  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 observationer 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

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

4.1 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 (blandt andet) 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 betegnes 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, altså
$$(\ln L)'(\overline{x}) = 0.$$
4. Af udtrykket for $\ln L$ fås
$$(\ln L)'(\mu) = \sum_{i=1}^{n} -(\ln f_0)'(x_i-\mu) = \sum_{i=1}^{n} g(x_i-\mu),$$
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
$$\sum_{i=1}^{n} g(x_i-\overline{x}) = 0. \tag{4.1}$$
5. Fidusen er nu at formel (4.1) 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 (4.1) at $g(-y) + g(y) = 0$, dvs.
   $$g(-y) = -g(y) \tag{4.2}$$
   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 $x_1 = x_2 = \cdots = x_k = -y$ og $x_{k+1} = ky$, fås at $k\,g(-y) + g(k\,y) = 0$, der ved brug af formel (4.2) kan formuleres som

$$g(k \cdot y) = k \cdot g(y) \tag{4.3}$$

gældende for vilkårligt y og $k = 1, 2, 3, \ldots$. Ved at bruge formel (4.2) endnu en gang kan man nu slutte at formel (4.3) gælder for vilkårlige reelle tal y og for vilkårlige hele tal k.

c) I formel (4.3) 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 (4.3), 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:

$$g(y) = y \cdot g(1) = g(1) \cdot y \tag{4.4}$$

for alle rationale tal y.

Medmindre g skal være en ganske overordentlig usædvanlig funktion, er det sådan at når formel (4.4) gælder for alle *rationale* tal y, så gælder den også for alle *reelle* tal y. Vi vil gå ud fra at formel (4.4) gælder for alle y, og vi er altså så nået frem til at funktionen g er en almindelig lineær funktion:

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

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

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

dvs.

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

7. Denne funktion f_0 skal være en sandsynlighedstæthed, hvilket vil sige at den skal være ikke-negativ og integrere til 1, altså $\displaystyle\int_{-\infty}^{+\infty} f_0(x)dx = 1$. 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

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

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

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

og dermed

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

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 $f(x;\mu,\sigma^2)$:

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

Der gælder at for ethvert valg af $\mu \in \mathbb{R}$ og $\sigma^2 > 0$ er dette en sandsynlighedstæthedsfunktion, nemlig for *normalfordelingen* med *positionsparameter* (eller *middelværdiparameter*) μ og *kvadratisk skalaparameter* (eller *variansparameter*) σ^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. Karl Friedrich Gauß (1777-1855) benyttede normalfordelinger til at beskrive bl.a. astronomiske målingers tilfældige afvigelser 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.

4.2 Egenskaber ved normalfordelingen

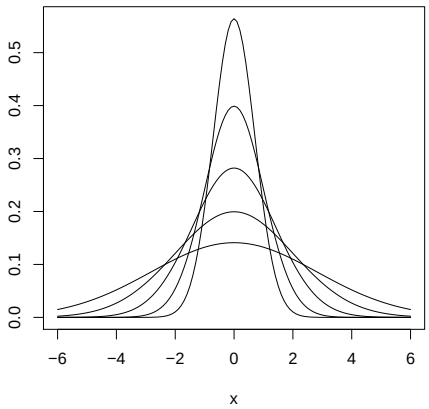
Her gives en oversigt (uden beviser) over forskellige egenskaber ved normalfordelingen:

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

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

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

b. 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 (dvs. den ene halvdel af sandsynlighedsmassen ligger til venstre for μ og den anden halvdel til højre for μ).



Figur 4.1 Tæthedsfunktioner for normalfordelinger med middelværdi 0 og varians hhv. 0.5 (den spidseste kurve), 1, 2, 4 og 8 (den fladeste kurve).

c. 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.

d. 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.

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

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

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:

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

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

$$x \longmapsto \frac{1}{\sigma}\varphi\Bigl(\frac{x-\mu}{\sigma}\Bigr)$$

og kumuleret fordelingsfunktion

$$x \longmapsto \Phi\Bigl(\frac{x-\mu}{\sigma}\Bigr).$$

g. 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 $u_\alpha + 5$.

4.3 Regn og tegn

Tæthedsfunktionen for normalfordelingen med parametre $\mu = 0$ og varians $\sigma^2 = 3$ (dvs. funktionen $x \mapsto \frac{1}{\sqrt{3}}\varphi\big(\frac{x-0}{\sqrt{3}}\big)$) kan tegnes sådan i R:

```
x <- seq(-6, 6, by=0.1)
plot (x, dnorm(x, mean=0, sd=sqrt(3)), type="l", ylab="")
```

Den kumulerede fordelingsfunktion (dvs. funktionen $x \mapsto \Phi\big(\frac{x-0}{\sqrt{3}}\big)$) kan tegnes sådan (med det samme `x` som ovenfor)

```
plot (x, pnorm(x, mean=0, sd=sqrt(3)), type="l", ylab="")
```

4.4 Opgaver

Opgave 4.1
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 4.2
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 4.3
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 4.2

Opgave 4.4
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.

5 Enstikprøveproblemet i normalfordelingen

Normalfordelingen blev i kapitel 4 udledt i forbindelse med jagten på 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 bestemt $\mathcal{N}(\mu, \sigma^2)$-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 5.1 (Lysets hastighed)
I årene 1880-82 foretog den amerikanske fysiker Albert Abraham Michelson og den amerikanske matematiker og astronom Simon Newcomb en række efter den tids forhold temmelig nøjagtige bestemmelser af lysets hastighed i luft (14). 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 5.1 (fra Stigler (20)) på side 59 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 mellem det roterende spejl der var placeret i Fort Myer på vestbredden af Potomacfloden, 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 5.1 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 5.1 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

$$
\begin{aligned}
f(y_1, y_2, \ldots, y_n; \mu, \sigma^2) &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2}\frac{(y_j-\mu)^2}{\sigma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2\right).
\end{aligned} \tag{5.1}
$$

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

$$
L(\mu, \sigma^2) = \text{konstant} \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2\right). \tag{5.2}
$$

5.1 Estimation af μ og σ^2

Vi vil bestemme maksimaliseringsestimaterne for μ og σ^2. Af udtrykket for likelihoodfunktionen ses at uanset 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:

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

altså

$$
\sum_{j=1}^{n}(y_j-\mu)^2 = \sum_{j=1}^{n}(y_j-\overline{y})^2 + n(\overline{y}-\mu)^2. \tag{5.3}
$$

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

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

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

Tabel 5.1 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	36	26	30	22	36	23
27	16	40	29	22	27	27	28	27	31	27
24	21	25	30	23	29	26	33	26	32	24
31	19	24	20	36	32	39	28	24	32	25
–2	36	28	25	21	28	29	29	27	28	29
–44	37	25	28	26	30	32	16	23	32	25

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

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

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

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2, \tag{5.4}$$

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

Eksempel 5.2 (Lysets hastighed, fortsat)
Hvis vi går ud fra at de 64 positive værdier i tabel 5.1 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

$$(27.75 \times 10^{-3} + 24.8) \times 10^{-6}\text{sek} = 24.828 \times 10^{-6}\text{sek}$$

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 (5.4), 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:

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

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.* 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:

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

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 $a+\overline{y-a}$. 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:

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

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 $s^2=6.24\times 10^{-12}$.

* Mange lommeregnere har en »statistikknap« ($\boldsymbol{\Sigma+}$) 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 $\boldsymbol{\Sigma+}$-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.

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 $\widehat{\sigma}^2$) 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

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

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. $\mathrm{E}\, s^2 = \mathrm{E}\Big(\dfrac{1}{n-1}\displaystyle\sum_{j=1}^{n}(Y_j-\overline{Y})^2\Big) = \sigma^2$. □

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 $\widehat{\mu}$, 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

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

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

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

et maximum likelihood estimat over σ^2.

5.2 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 (5.2) på side 58, og dens maksimale værdi er $L(\overline{y}, \widehat{\sigma}^2)$. Under H_0 er likelihoodfunktionen

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

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

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

Kvotienttesstørrelsen bliver derfor

$$Q = \frac{L\big(\mu_0, \widehat{\widehat{\sigma}}^2\big)}{L(\overline{y}, \widehat{\sigma}^2)} = \left(\frac{\widehat{\widehat{\sigma}}^2}{\widehat{\sigma}^2}\right)^{-n/2} \exp\left(-\left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{2\widehat{\widehat{\sigma}}^2} - \frac{\sum_{j=1}^{n}(y_j-\overline{y})^2}{2\widehat{\sigma}^2}\right)\right)$$

$$= \left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}\right)^{-n/2} \exp\Big(-\Big(\frac{n}{2}-\frac{n}{2}\Big)\Big)$$

$$= \left(\frac{\sum_{j=1}^{n}(y_j-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}\right)^{-n/2} .$$

Her omskrives kvadratsummen i tælleren ved hjælp af formel (5.3) på side 58 (med μ erstattet af μ_0), og man får

$$Q = \left(\frac{\sum_{j=1}^{n}(y_j-\overline{y})^2 + n(\overline{y}-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}\right)^{-n/2}$$

$$= \left(1 + \frac{n(\overline{y}-\mu_0)^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}\right)^{-n/2}$$

$$= \left(1 + \frac{n(\overline{y}-\mu_0)^2}{(n-1)s^2}\right)^{-n/2}$$

$$= \left(1 + \frac{1}{n-1}\Big(\frac{\overline{y}-\mu_0}{\sqrt{s^2/n}}\Big)^2\right)^{-n/2} .$$

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

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

Det er 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 William S. Gosset, der skrev den første artikel om t-testet ((21)), 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 udregnet 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.†

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.

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

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

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

og dermed

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

Eksempel 5.3 (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

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

† 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.

‡ 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.

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.

5.3 Histogrammer og fraktildiagrammer

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

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 5.1.

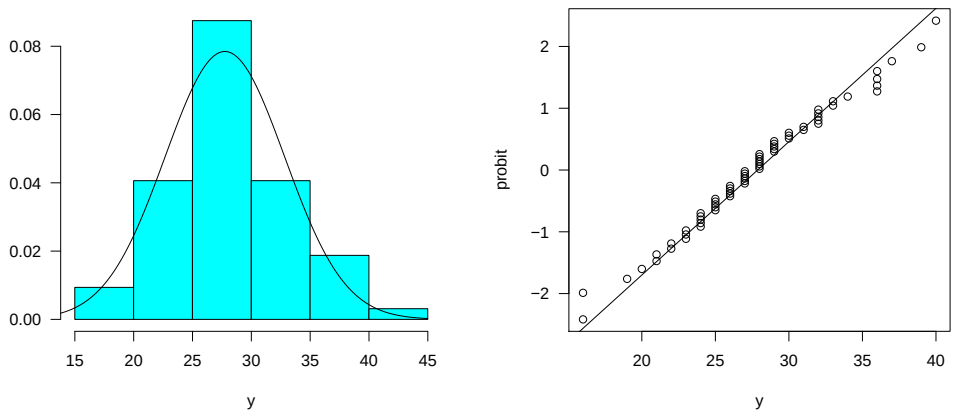
Ved udarbejdelsen af et histogram kan det være lidt af et kunststykke at vælge den rigtige intervalinddeling således at fluktuationerne bliver passende udglattet uden at tæthedens form bliver alt for udjævnet. Hvis intervallerne er for korte, bliver fluktuationerne ikke udglattet nok, er de for lange,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

Man kan godt give en lidt mere formel opskift på et histogram over et sæt observationer $y_1, y_2, \ldots, y_n$:

1. I det område hvor observationerne falder vælges delepunkter (der som regel bør være ækvidistante) $x_0 < x_1 < x_2 < \cdots < x_m$ 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

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

   Grafen for denne funktion h er histogrammet (svarende til den valgte inddeling) over observationerne $y_1, y_2, \ldots, y_n$.



Figur 5.1 Histogram (til venstre) og fraktildiagram (til højre) over de 64 målte værdier af lysets passagetid. – Den indtegnede kurve i histogrammet er tætheden for normalfordelingen med parametre $\overline{y} = 27.75$ og $s^2 = 25.8$; den rette linje i fraktildiagrammet har hældning $1/s = 0.20$ og går gennem $(\overline{y}, 0)$, altså $(27.75, 0)$.

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 $y \mapsto \Phi(\frac{y-\mu}{\sigma})$ (jf. side 53). Derfor finder man den teoretiske α-fraktil y_α ved at løse ligningen $\Phi(\frac{y_\alpha-\mu}{\sigma}) = \alpha$, hvilket giver $y_\alpha = \mu+\sigma\,\Phi^{-1}(\alpha)$. De n punkter hvis førstekoordinater er de empiriske fraktiler, og hvis andenkoordinater er de tilsvarende teoretiske fraktiler, det vil sige punkterne med koordinater

$$\Big(y_{(i)}, \mu + \sigma\,\Phi^{-1}(\tfrac{i-0.5}{n})\Big), \qquad i = 1, 2, \ldots, n,$$

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

$$\Big(y_{(i)}, \Phi^{-1}(\tfrac{i-0.5}{n})\Big), \qquad i = 1, 2, \ldots, n,$$

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

Konkret fremstiller man fraktildiagrammet ved at indtegne punkterne

$$\Big(y_{(i)}, \Phi^{-1}(\tfrac{i-0.5}{n})\Big), \qquad i = 1, 2, \ldots, n,$$

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

I figur 5.1 ses et fraktildiagram over de 64 målte værdier af lysets passagetid. Såvel histogrammet som fraktildiagrammet viser, at det ikke er ganske urimeligt at antage at måleresultaterne er normalfordelte.

5.4 Regn og tegn

Newcombs målinger af lysets passagetid

Her vises hvordan man kan analysere målingerne af lysets passagetid (tabel 5.1).

Først indlæses data (som findes i R-biblioteket `MASS`); de værdier der er større end 0, placeres i variablen `y`.

```
require (MASS)
data (newcomb)
y <- newcomb[newcomb>0]
```

Histogrammet tegnes lettest med `MASS`-funktionen `truehist` (alternativt `hist`):

```
truehist (y, las=1)
```

Normalfordelingskurven kan tegnes sådan:

```
x <- seq(10, 45, by=0.2)
lines(x, dnorm(x, mean(y), sd(y)))
```

Fraktildiagrammet plus linjen kan tegnes sådan:

```
qqnorm (y, datax=TRUE, las=1, ylab="y", xlab="probit", main="")
qqline (y, datax=TRUE)
```

Hypotesen $\mu = 23.8$ kan testes sådan:

```
t.test (y, mu=23.8)
```

Vedr. opgave 5.4

En tabel som tabel 5.3 kan fremstilles sådan:

```
t <- matrix (rnorm(200, mean=5, sd=sqrt(3)), nrow=20)
round (cbind (t, rowMeans(t), diag(var(t))), digits=2)
```

Tabel 5.2 Opgave 5.2: 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					

5.5 Opgaver

Opgave 5.1
Nedenstående 18 tal kan opfattes som en stikprøve fra en normalfordeling.

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

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 59.
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 5.2 (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 amerikanske marked har man målt kviksølvindholdet i 115 tilfældigt udvalgte sværdfisk og fået resultaterne i tabel 5.2 (fra (11)).

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.

Tabel 5.3 Data til opgave 5.4: 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

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 linje 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.

Opgave 5.3 (fortsættelse af opgave 5.2)
Løs det der er det overordnede problem i opgave 5.2, nemlig: hvordan skal man fastsætte kassationsgrænsen for de 75% af fiskene der kontrolleres, hvis man vil opnå at forbrugeren i middel højst udsættes for en kviksølvbelastning på 1 ppm.

Opgave 5.4
I tabel 5.3 er der 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 $\overline{\overline{y}} = \big(\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_{20}\big)/20$ og den empiriske varians på $\overline{y}_i$-erne, dvs. $\frac{1}{20-1}\sum_{i=1}^{20}\big(\overline{y}_i - \overline{\overline{y}}\big)^2$. Giver det cirka 5 og 0.3, som man skulle tro?
   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.

6 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.*

Hermed er problemet beskrevet i generelle vendinger. I resten af dette kapitel og i kapitel 7 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.

6.1 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 henholdsvis μ_1 og μ_2, kort

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

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 normal-

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

fordelingen) beskriver den *tilfældige variation* der altså er den samme i begge grupper (denne antagelse kan man eventuelt teste, se side 76).

Estimation af middelværdiparametrene

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

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

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, $\overline{y}_i = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} y_{ij}$. 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

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

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 variansparameteren

Maksimaliseringsestimatet $\widehat{\sigma}^2$ for σ^2 kan bestemmes som maksimumspunktet for funktionen $\sigma^2 \mapsto L(\overline{y}_1, \overline{y}_2, \sigma^2)$; man finder at

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

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*. 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 $\widehat{\sigma}^2$ 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

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

Man begrunder brugen af s_0^2 frem for $\widehat{\sigma}^2$ på lignende måde som i Enstikprøveproblemet i normalfordelingen, se side 61.

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 5 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

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

- det variansestimat man som regel benytter, er

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

  med $n-1$ frihedsgrader.

Kvotienttesstørrelsen for H_0 er

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

hvor L er defineret på side 73. 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

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

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

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

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 ($\overline{y}_1$ og $\overline{y}_2$) i forhold til den estimerede standardafvigelse på denne differens.†

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

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

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 72) 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 kvotienteteststørrelsen er en funktion af

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

† Det var et ræsonnement af denne art der førte William S. Gosset (alias 'Student') til i 1908 i (21) at foreslå en teststørrelse der næsten er vore dages *Student's t*.

‡ 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 $s_i^2 = \dfrac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$ 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 ε benyttes sandsynligheden for at få en R-værdi der ligger uden for intervallet med endepunkter R_{obs} og $1/R_{\text{obs}}$, det vil sige

- hvis $R_{\text{obs}} > 1$ så er

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

- hvis $R_{\text{obs}} < 1$ så er

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

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

Eksempel 6.1 (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.

I tabel 6.1 er et regneskema der viser hvordan man kan foretage udregningerne med »håndkraft« (se også side 59f). Hvis man blot vil opsummere resultaterne, gør man det ofte i form af en tabel som den der er vist i tabel 6.2.

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 76). 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 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.

6.2 Tostikprøveproblemet med parrede observationer

Som titlen på afsnit 6.1 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

Tabel 6.1 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
$\overline{y}_i$	$131.8/10 = 13.18$		$80.0/10 = 8.00$	
$\sum y^2 - \frac{(\sum y)^2}{n}$	$1914.36 - \frac{131.8^2}{10} = 177.236$		$708.96 - \frac{80.0^2}{10} = 68.960$	
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	$\frac{13.18 - 8.00}{\sqrt{13.68\left(\frac{1}{10} + \frac{1}{10}\right)}} = 3.13$			

Tabel 6.2 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

målinger der stammer fra hvilke individer. Vi viser situationen skematisk:

	gruppe nr. 1	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 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 5.

Eksempel 6.2 (Sovemidler)
Det kemiske stof *hyoscyamin hydrobromid* kan anvendes som sovemiddel. Stoffet findes imidlertid i to udgaver, d-hyoscyamin hydrobromid og l-hyoscyamin hydrobromid,[§] 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 6.3 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. (21)

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 6.4.

Vi vil opfatte tallene i tabel 6.4 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

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

Tabel 6.3 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

Tabel 6.4 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

dermed $\sqrt{s^2/n} = \sqrt{1.51/10}$ timer $= 0.39$ timer. Endvidere bliver 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.

Tabel 6.5 Vedr. opgave 6.1: Varmemængde (i calorier) for at smelte 1 g is med en begyndelsestemperatur på $-0.72\,^{\circ}\mathrm{C}$, 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 6.6 Vedr. opgave 6.2: Den maksimale procentdel af blodpladerne der klumper sig sammen før hhv. 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

Datamaterialet til dette eksempel er meget berømt fordi det blev benyttet til et illustrativt eksempel i den artikel fra 1908 hvor t-testet (i enstikprøveproblemet) blev introduceret (21). Artiklen er skrevet af William S. Gosset der arbejdede som biometriker ved Guinnessbryggerierne, og som benyttede 'Student' som sit *nom de plume*. (Derfor kaldes t-størrelsen ofte Student's t.)

6.3 Regn og tegn

Vi viser hvordan man kan analysere Student's data om ekstra søvntimer (tabel 6.3) med R. Datamaterialet indgår i R-distributionen, og det kan indlæses ved at skrive `data(sleep)`; derved oprettes en 'data.frame' `sleep` med de to variable `extra` og `group` (se deres indhold ved at skrive `sleep`). For at analysere data som et »tostikprøveproblem med parrede observationer« (som jo er den rigtige metode) behøver man nu blot skrive

```
t.test (extra ~ group, data=sleep, paired=TRUE)
```

Hvis man (fejlagtigt) opfattede det som et »tostikprøveproblem med uparrede observationer«, kunne man teste for ens varianser og derefter for ens middelværdier sådan:

```
var.test (extra ~ group, data=sleep)
t.test (extra ~ group, data=sleep, var.equal=TRUE)
```

6.4 Opgaver

Opgave 6.1 (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 6.5. Undersøg om der er signifikant forskel på de to metoder.

Tabel 6.7 Vedr. opgave 6.3: 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

Tip: Udregningerne bliver lettere hvis man indfører et passende beregningsnulpunkt.

Opgave 6.2 (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 6.6.

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 6.3
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 blev optaget i musen i løbet af et fastsat stykke tid. tabel 6.7 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«.

7 Ensidet variansanalyse

Sammenligning af *to* normalfordelte stikprøver er omtalt i kapitel 6. 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 71 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 7.1 (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 Greenfort, Jensen & Jeppesen (7) 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. 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 (nemlig 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 7.1. Registreringen foregår ved hjælp af et rektangulært gitternet med 416 gitterpunkter med fem centimeters afstand; gitternettet placeres på jordlodden, hvorefter

Tabel 7.1 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

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 dog 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 ud som i tabel 7.2.

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

Tabel 7.2 Den generelle k-stikprøveproblem

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}

sådan at Y_{ij} *er normalfordelt med middelværdi* μ_i *og varians* σ^2, kort

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2). \tag{7.1}$$

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; denne antagelse kan man undertiden teste, se afsnit 7.3.

7.1 Estimation af parametrene

Middelværdiparametrene

De ukendte middelværdiparametre $\mu_1, \mu_2, \ldots, \mu_k$ i grundmodellen (7.1) estimeres ved hjælp af maximum likelihood metoden, altså som de værdier der maksimaliserer likelihoodfunktionen

$$\begin{aligned} L(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2) &= \prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2} \frac{(y_{ij} - \mu_i)^2}{\sigma^2} \right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 \right) \end{aligned} \tag{7.2}$$

hvor $n = n_1 + n_2 + \cdots + n_k$ 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, $\overline{y}_i = \dfrac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$. Ved at benytte formlen for kvadratet på en toleddet størrelse fås

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

Tabel 7.3 Fuglegræseksemplet: nogle beregnede størrelser.

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

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

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

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i-\mu_i)^2\ . \tag{7.3}$$

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, kort $\widehat{\mu}_i = \overline{y}_i$.

Variansparameteren

Maksimaliseringsestimatet $\widehat{\sigma}^2$ 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

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

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 $\widehat{\sigma}^2$ 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

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

Man begrunder brugen af s_0^2 frem for $\widehat{\sigma}^2$ på lignende måde som i enstikprøveproblemet i normalfordelingen, se side 61.

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,

$$s_0^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 \tag{7.4}$$

  med $n - k$ frihedsgrader.

I tabel 7.3 er vist de værdier man finder i Fuglegræs-eksemplet.

7.2 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*:

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

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 om 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 under H_0 og under grundmodellen. Vi ved fra forrige afsnit at i grundmodellen maksimaliseres likelihoodfunktionen af værdierne $\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k$ og $\widehat{\sigma}^2$, hvor $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$. 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 5 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.

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

- 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.

Kvotienttesteststørrelsen for H_0 er

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

hvor L er defineret i formel (7.2) side 87. 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

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

For at kunne omforme Q yderligere skal vi bruge følgende omskrivning der fås af formel (7.3) på side 88 hvis man erstatter μ_i med $\overline{y}$:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{y}_i - \overline{y})^2, \tag{7.5}$$

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 (7.4) på side 89).
- *Variationen mellem grupper* (dvs. gruppegennemsnittenes variation omkring det totale gennemsnit) beskrives af

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

Men vi skal videre med omskrivningen af udtrykket for Q. Ved hjælp af formel (7.5) kan vi omskrive Q til

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

hvoraf det fremgår 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. I praksis benytter man altid F som teststørrelse for H_0. 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 $(k-1, n-k)$ når hypotesen H_0 er rigtig. Derfor kan testsandsynligheden $\varepsilon = \mathrm{P}_0\big(F > F_{\mathrm{obs}}\big)$ bestemmes som

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

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

Tabel 7.4 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	

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 variansanalysen i et såkaldt *variansanalyseskema*. Tabel 7.4 er et variansanalyseskema for Fuglegræs-eksemplet.

Eksempel 7.2 (Fuglegræs, konklusion)
Af variansanalyseskemaet (tabel 7.4) fremgår at F-teststørrelsen for hypotesen om ens grupper 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.

7.3 Bartletts test for varianshomogenitet

I normalfordelingsmodeller er det en forudsætning for en meningsfuld sammenligning af forskellige gruppers middelværdiparametre at grupperne har samme varians.* 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*.

Den generelle situation er stadig den der blev præsenteret på side 86, og vi ønsker nu 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 (7.1) fra side 87 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 (7.1) som en hypotese i forhold til den nye grundmodel.

* 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).

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.† Hvis man omskriver kvotienttesstørrelsen i den nævnte betingede fordeling, når man frem til at man kan benytte følgende størrelse (*Bartletts teststørrelse*) som teststørrelse for hypotesen om varianshomogenitet:

$$B = -\sum_{i=1}^{k} f_i \ln \frac{s_i^2}{s_0^2} \; ; \tag{7.6}$$

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.

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

og s_0^2 er det sædvanlige estimat over den fælles varians σ^2 (formel (7.4) på side 89). Bemærk i øvrigt at s_0^2 er et vægtet gennemsnit af s_i^2-erne med frihedsgraderne som vægte, $s_0^2 = \dfrac{1}{f} \displaystyle\sum_{i=1}^{k} f_i s_i^2$, hvor $f = f_1 + f_2 + \cdots + f_k$ 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

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

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 76. (Dette tostikprøvetest er ikke baseret på nogen χ^2-approksimationer, så det har ingen restriktioner på antallene af frihedsgrader.)

Eksempel 7.3 (Fuglegræs: test for varianshomogenitet)
Som illustration udregnes Bartletts teststørrelse i Fuglegræs-eksemplet. Vi udvider det tidligere regneskema i tabel 7.3 og får tabel 7.5. Derefter kan vi udregne B_{obs}:

$$B_{\mathrm{obs}} = -\Big(3 \ln \frac{78.00}{34.46} + 3 \ln \frac{4.25}{34.46} + 3 \ln \frac{14.67}{34.46} + 3 \ln \frac{40.92}{34.46}\Big) = 5.87\,.$$

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

† Dette hænger sammen med at man måske også bør *estimere* variansparametrene i denne betingede fordeling, se side 61.

Tabel 7.5 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

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.

7.4 Regn og tegn

Her vises hvordan man kan udføre ensidet variansanalyse og Bartletts test med R.

Fuglegræseksemplet

Data findes i en fil hvis første linjer ser sådan ud (`dg`-søjlen indeholder dækningsgraderne og `gr`-søjlen gruppenumre):

```
gr dg
1 17
1 38
1 23
1 26
2 19
```

Her følger (dele af) en R-session; brugeren skriver de linjer der begynder med `>`.

Først indlæses data, og man fortæller at værdierne i `gr` ikke skal opfattes som tal, men som navne på grupper, dvs. faktorer:

```
> fuglegrs <- read.table("h:/bog/txt304ny/fuglegrs.dat", header=TRUE, nrows=20)
> fuglegrs$gr <- factor(fuglegrs$gr)
```

Fit grundmodellen `G`:

```
> G <- lm (dg ~ gr - 1, data = fuglegrs)
> summary(G)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
gr1   26.000      2.935   8.858 1.30e-06 ***
gr2   16.250      2.935   5.537 0.000129 ***
```

Tabel 7.6 Opgave 7.1: Tørstofindhold (i g) i planter under forskellige dyrkningsbetingelser.

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

```
gr3   30.000      2.935  10.221 2.83e-07 ***
gr4   23.250      2.935   7.921 4.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.87 on 12 degrees of freedom
F-statistic: 69.09 on 4 and 12 DF,  p-value: 3.507e-08
```

Bartletts test for varianshomogenitet:

```
> bartlett.test(dg ~ gr - 1, data = fuglegrs)

Bartlett test for homogeneity of variances

data:  dg by gr
Bartlett's K-squared = 5.1583, df = 3, p-value = 0.1606
```

Hypotesen H_0 testes i forhold til grundmodellen:

```
> H0 <- update(G, dg ~ 1)
> anova(H0, G)
Analysis of Variance Table

Model 1: dg ~ 1
Model 2: dg ~ gr - 1
Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     15 815.75
2     12 413.50  3    402.25 3.8912 0.03736 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Det ses at F-teststørrelsen bliver 3.8912 og den tilsvarende testsandsynlighed 0.03736.

Data til opgave 7.1

Talmaterialet indlæses med `data(PlantGrowth)` hvorved der oprettes en 'data.frame' `PlantGrowth` indeholdende de to variable `weight` og `group`.

Tabel 7.7 Opgave 7.2: 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

7.5 Opgaver

Opgave 7.1 (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 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 7.6.

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

Opgave 7.2 (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. (10)

De fire grupper af børn er 1) sorte børn i blandede skoler, 2) hvide børn i blandede skoler, 3) sorte børn i adskilte skoler, og 4) hvide børn i adskilte skoler. Der er undersøgt 50 børn fra hver gruppe. Resultaterne fremgår af tabel 7.7 (fra (18)).

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

Tabel 7.8 Vedr. opgave 7.3: 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 7.9 Vedr. opgave 7.3: 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

Opgave 7.3 (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),
  - 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 7.8.

Undersøg ved hjælp af ensidet variansanalyse om man kan sige at det tilsatte hormon faktisk virker vækstfremmende.
Tip: Benyt eventuelt tabel 7.9.

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?

8 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 en 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 skala), 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 siger 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*.

Under alle omstændigheder er der tale om *lineær* regressionsanalyse, hvilket betyder at baggrundsvariablene indgår *lineært* i udtrykket for middelværdien. Det giver selvfølgelig en vis begrænsning i, hvor generelle man kan lave denne type modeller, men på den anden side kan man vælge baggrundsvariablene helt frit, og det er også tilladt at danne nye baggrundsvariable ud fra gamle. Hvis man f.eks. har én »naturligt givet« baggrundsvariabel t som er en nærmere fastlagt tidsstørrelse, kan man evt. indføre en ny baggrundsvariabel t^2, således at man alt i alt får den lineære regressionsmodel $\mathrm{E}\, Y_i = \beta_0 + \beta_1 t + \beta_2 t^2$.

Regressionsanalyse går kort fortalt ud på at finde en statistisk model hvormed man kan beskrive en y-variabel ved hjælp af en kendt simpel funktion af nogle baggrundsvariable og nogle såkaldte *parametre*. Parametrene er de samme for alle observationssæt, hvorimod baggrundsvariablene typisk ikke er det. Parametrenes værdier bestemmes ud fra data således at man får det bedste *fit*.

Man må naturligvis ikke forvente at den statistiske model leverer en perfekt beskrivelse, et perfekt fit, dels fordi den model man måtte finde frem til næppe er fuldstændig rigtig, dels fordi en af pointerne med statistiske modeller netop er, at de kun beskriver hovedtrækkene i datamaterialet og ser stort på de finere detaljer. Der vil derfor være en vis forskel mellem den observerede værdi y og den såkaldt *fittede* værdi $\widehat{y}$, dvs. den værdi som man ifølge regressionsmodellen skulle få med de givne værdier af baggrundsvariablene. Denne forskel kaldes *residualet* og betegnes ofte e. Vi har så opspaltningen

$$\begin{array}{ccccc} y & = & \widehat{y} & + & e \\ \text{observeret værdi} & = & \text{fittet værdi} & + & \text{residual}\,. \end{array}$$

Residualerne er det som modellen *ikke* beskriver, og derfor er det naturligt at man (eller rettere modellen) anser dem for *tilfældige*, ɔ: tilfældige tal fra en vis sandsynlighedsfordeling.

To væsentlige forudsætninger for at kunne benytte regressionsanalyse er

1. at det ikke er x-erne, men kun y-erne og residualerne, der er behæftede med *tilfældig variation* (»usikkerhed«),
2. at de enkelte målinger er *stokastisk uafhængige* af hinanden, hvilket vil sige at de tilfældigheder der indvirker på én bestemt y-værdi (efter at man har taget højde for baggrundsvariablene) ikke har nogen sammenhæng med de tilfældigheder der spiller ind på de øvrige y-værdier.

De simpleste eksempler på regressionsanalyse er dem hvor der kun er én enkelt baggrundsvariabel, som vi så kan betegne x. Opgaven bliver da at beskrive y-værdierne ved hjælp af en kendt simpel funktion af x. Det simpleste ikke-trivielle bud på en sådan funktion må vel være en funktion af typen $x \mapsto \beta_0 + x\beta_1$ hvor β_0 og β_1 er to parametre, dvs. man formoder at y afhænger lineært af x. Derved får man den såkaldte simple lineære regressionsmodel.

De følgende kapitler beskæftiger sig med forskellige væsentlige aspekter af regressionsmodeller og regressionsanalyse: Hvordan vælger man værdierne af β-erne så man får det bedste fit? Hvordan afgør man om en bestemt model er god *nok*? Hvis man har flere forskellige baggrundsvariable til sin rådighed, hvordan afgør man så hvilke af dem der skal med i modellen og hvilke ikke?

8.1 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 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:

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

Denne model kan kort skrives som

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2). \tag{8.1}$$

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 8.1 (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 og registrerede blandt andet øjenfarve, gemyt, kunstneriske evner, sygdomme, valg af ægtefælle, frugtbarhed, og altså højde.

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 nærmen sig det gennemsnitlige så Galton som en tilbagegang og kaldte det derfor en *regression*.* †

I tabel 8.1 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

∗ *regression* betyder *tilbagegang*.

† 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 8.1 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	-	-

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 8.2 (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.

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 8.2.

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

Tabel 8.2 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

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 (8.1) giver en fornuftig beskrivelse af datamaterialet?
3. Hvordan tester man hypoteser om parametrene?

8.2 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 (8.1) hørende likelihoodfunktion

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

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

$$\sum_{i=1}^{n}\big(y_i-(\alpha+\beta x_i)\big)^2. \tag{8.2}$$

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 58), i tostikprøveproblemet (side 73) og i ensidet variansanalyse (side 87). 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 (8.2) således:

$$\begin{aligned}\sum_{i=1}^{n}\big(y_i-(\alpha+\beta x_i)\big)^2 &= \sum_{i=1}^{n}\big((y_i-\overline{y})+(\overline{y}-(\alpha+\beta\overline{x}))-\beta(x_i-\overline{x})\big)^2\\ &= \sum_{i=1}^{n}(y_i-\overline{y})^2\\ &\quad+n\big(\overline{y}-(\alpha+\beta\overline{x})\big)^2\\ &\quad+\beta^2\sum_{i=1}^{n}(x_i-\overline{x})^2\\ &\quad-2\beta\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\ ,\end{aligned}\tag{8.3}$$

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.‡

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

$$\widehat{\beta}=\frac{SP_{xy}}{SS_x}.$$

Ifølge betragtningerne ovenfor er det dertil svarende bedste valg af α

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

Hermed har vi løst estimationsproblemet for så vidt angår α og β.

Den *estimerede regressionslinje* er (den linje hvis ligning er) $y=\widehat{\alpha}+\widehat{\beta}x$.

Undertiden, især når man skal udføre beregningerne mere eller mindre med håndkraft, kan man have fornøjelse af et andet udtryk for $\widehat{\beta}$ eller måske snarere for SP_{xy} og SS_x.

‡ SS = Sum of Squared deviations, SP = Sum of Products.

Ved almindelige og lette formelmanipulationer finder man følgende formler, hvor hver gang det første lighedstegn er definitionslighedstegnet og det andet viser det alternative udtryk:

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

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) $\widehat{\alpha}$ og $\widehat{\beta}$ i kvadratsummen (8.2), så den er

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

   Hvis man i stedet indsætter i udtrykket (8.3) og reducerer, får man et alternativt udtryk for residualkvadratsummen, nemlig

$$\begin{aligned} \sum_{i=1}^{n}\Big(y_i - (\widehat{\alpha}+\widehat{\beta}x_i)\Big)^2 &= \sum_{i=1}^{n}(y_i-\overline{y})^2 - \widehat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2 \\ &= SS_y - \widehat{\beta}^2 SS_x \\ &= SS_y - \frac{SP_{xy}^2}{SS_x}. \end{aligned}$$

2. *Antallet af frihedsgrader* er $n-2$ fordi der er n observationer og der er estimeret 2 middelværdiparametre. Variansen σ^2 estimeres derfor ved

$$\begin{aligned} s_{02}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\Big(y_i - (\widehat{\alpha}+\widehat{\beta}x_i)\Big)^2 \\ &= \frac{1}{n-2}\Big(SS_y - \frac{SP_{xy}^2}{SS_x}\Big). \end{aligned} \tag{8.4}$$

Eksempel 8.3 (Fædre og sønner, fortsat fra side 103)
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.

Tabel 8.3 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

På grundlag af tallene i tabel 8.1 udregnes først nogle hjælpestørrelser, se tabel 8.3, og ved hjælp af disse udregnes

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

Den estimerede regressionskoefficient er

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

og den estimerede skæring med ordinataksen er

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

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.

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

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

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 $\widehat{\alpha} + \widehat{\beta}x_i$, 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.

8.3 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[§], og i den forstand er middelfejlen et direkte mål for hvor unøjagtigt estimatet er.[¶]

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 $\widehat{\beta}$ er $\sqrt{\sigma^2/SS_x}$.
2. a) Middelfejlen på $\widehat{\alpha}$ er $\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{SS_x}\right)}$.
   b) Estimatorerne $\widehat{\alpha}$ og $\widehat{\beta}$ er korrelerede, og korrelation mellem dem er $-1\Big/\sqrt{1+\frac{SS_x}{n\overline{x}^2}}$.
3. a) Middelfejlen på $\widehat{\alpha}+\widehat{\beta}\overline{x}$ er $\sqrt{\sigma^2/n}$.
   b) Estimatorerne $\widehat{\alpha}+\widehat{\beta}\overline{x}$ og $\widehat{\beta}$ 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_{02}^2, og derved få de *estimerede middelfejl*.

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

8.4 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

§ Det er især rigtigt for normalfordelte estimatorer – såsom $\widehat{\alpha}$ og $\widehat{\beta}$.

¶ Eksempel: Hvis man har udregnet $\widehat{\beta}$ til 1.534 og middelfejlen på $\widehat{\beta}$ til 0.3, så véd man at ca. 95% af alle de andre $\widehat{\beta}$-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 $\widehat{\beta}$ *ikke* skal angives med tre decimaler, men kun med én.

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 $n = n_1 + n_2 + \cdots + n_k$. Regressionsmodellen (8.1) 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 106) er i den nye notation

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

hvor der er benyttet følgende betegnelser:

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

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

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

Parameterestimaterne er stadig

$$\widehat{\beta} = \frac{SP_{xy}}{SS_x} \ , \qquad \widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \ ,$$

$$s_{02}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\Big(y_i - (\widehat{\alpha} + \widehat{\beta}x_i)\Big)^2 = \frac{1}{n-2}\left(SS_y - \frac{SP_{xy}^2}{SS_x}\right). \tag{8.5}$$

Eksempel 8.4 (Kvælning af hunde, fortsat fra side 104)
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 114.)

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 $\widehat{\alpha}$, $\widehat{\beta}$ 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 8.4. Heraf fås den estimerede regressionskoefficient til

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

og det estimerede skæringspunkt med ordinataksen til

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

Variansen estimeres ved

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

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

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

Tabel 8.4 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

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 $\widehat{\beta}$ er (jf. side 109)

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

og middelfejlen på $\widehat{\alpha}$ er

$$\sqrt{\Big(\frac{1}{n} + \frac{\overline{x}^2}{SS_x}\Big)s_{02}^2} = \sqrt{\Big(\frac{1}{25} + \frac{(222/25)^2}{1232.64}\Big)4.85}\ \mu\text{mol}\,\text{l}^{-1} = 0.7\ \mu\text{mol}\,\text{l}^{-1}.$$

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

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

8.5 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 regressionslinje* og ser efter om punkterne fordeler sig passende tilfældigt omkring linjen. 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.

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:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(y_{ij} - (\widehat{\alpha} + \widehat{\beta}x_i)\Big)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k} n_i\Big(\overline{y}_i - (\widehat{\alpha} + \widehat{\beta}x_i)\Big)^2 \qquad (8.6)$$

(denne opspaltning følger af formel (7.3) på side 88 ved at erstatte μ_i med $\widehat{\alpha} + \widehat{\beta}x_i$); den tilsvarende opspaltning af frihedsgraderne er

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

Formel (8.6) viser hvordan residualkvadratsummen, der kan siges at beskrive *den samlede variation omkring regressionslinjen*, deles op i en sum af en kvadratsum vedrørende *variationen inden for grupper* og en kvadratsum vedrørende *gruppernes variation omkring regressionslinjen*, se også figur 8.1.

Ved at dividere kvadratsummerne med deres frihedsgrader fås s^2-størrelserne: dels de tidligere indførte s_{02}^2 med $n-2$ frihedsgrader (formel (8.5) side 111) og s_0^2 med $n-k$ frihedsgrader (side 89), dels

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

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

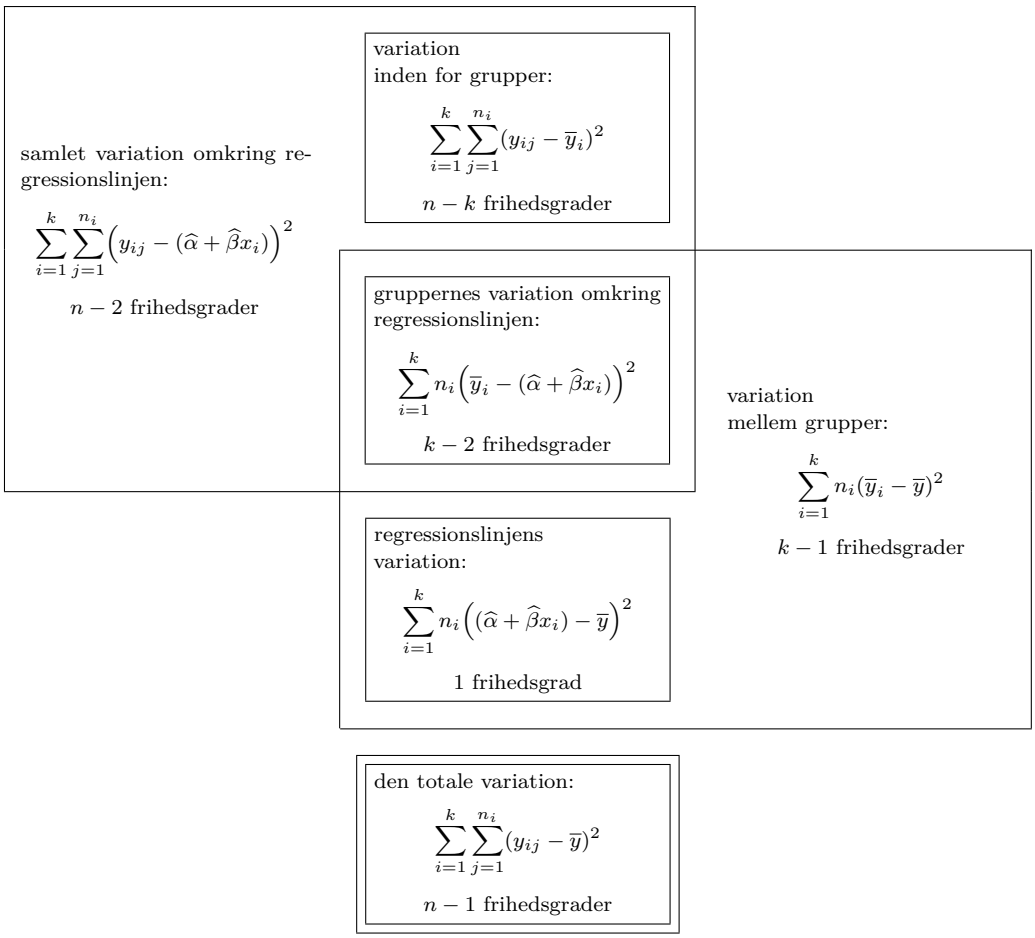
$$F = \frac{s_2^2}{s_0^2}$$

det vil sige gruppernes variation omkring linjen 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 $n-k$, så testsandsynligheden ε er givet som

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

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.



Figur 8.1 Skematisk oversigt over nogle af de i kapitlet forekommende kvadratsummer med tilhørende frihedsgrader.

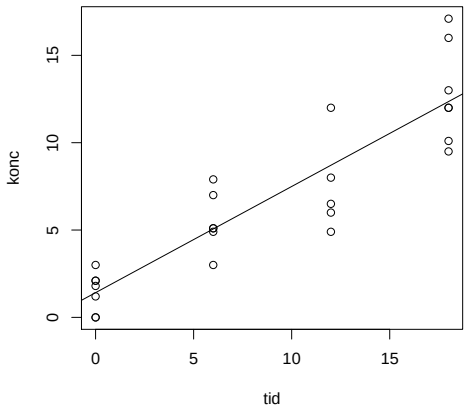
- Hvis ε ikke er meget lille (og F dermed ikke er signifikant stor), så kan vi godtage den lineære regressionsmodel $Y_{ij} \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$ hvor parametrene α og β estimeres ved $\widehat{\alpha}$ og $\widehat{\beta}$ og hvor parameteren σ^2 estimeres ved s_{02}^2 med $n-2$ frihedsgrader (jf. side 111).

Bemærk i øvrigt, at kvadratsummen vedrørende gruppernes variation omkring linjen ifølge formel (8.6) kan skrives som

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

det kan være praktisk ved udregninger.

Eksempel 8.5 (Kvælning af hunde, fortsat)
Vi vil undersøge om det kan antages at hypoxantinkoncentrationen afhænger lineært af hypoxiens



Figur 8.2 Kvælning af hunde: Sammenhørende værdier af hypoxantinkoncentration og hypoxivarighed, samt den estimerede regressionslinje.

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 111 ff) bestemt de talværdier der i givet fald er de bedste estimater over parametrene, og derved fået den estimerede regressionslinje til

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

I figur 8.2 er indtegnet dels sammenhørende værdier af varighed og koncentration, dels den estimerede regressionslinje. 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 8.5. Det fremgår blandt andet at den kvadratsum der beskriver *variationen mellem grupper*, er 101.32 med 21 frihedsgrader. På side 112 fandt vi regressionsmodellens residualkvadratsum til 111.50 med 23 frihedsgrader, så kvadratsummen hørende til gruppernes variation omkring regressionslinjen er $111.50 - 101.32 = 10.18$ med $23 - 21 = 2$ frihedsgrader. Teststørrelsen for hypotesen om at gruppemiddelværdierne ligger på en ret linje, er da

$$F = \frac{10.18/2}{101.32/21} = \frac{5.09}{4.82} = 1.06$$

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 8.6.

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 8.5 i Bartletts teststørrelse og får

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

Tabel 8.5 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

Tabel 8.6 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.linjen	2	10.18	5.09	5.09/4.82=1.06
samlet variation omkring regr.linjen	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.

8.6 Test af hypoteser om linjens parametre

Man kan naturligvis teste hypoteser om regressionslinjens 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 106, og hvor

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

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

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

er estimatet $\widehat{\beta}$ over β divideret med den estimerede standardafvigelse (dvs. den estimerede middelfejl) på $\widehat{\beta}$, jf. side 109. Man kan sige at t-størrelsen måler hvor langt $\widehat{\beta}$ 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 linjen *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 linjen går gennem $(0,0)$, kan det gøres med F-teststørrelsen $F = s_4^2/s_{02}^2$, hvor s_4^2 er »kvadratsummen« $\dfrac{n\,SS_x\,\widehat{\alpha}^2}{SS_x + n\overline{x}^2}$ 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

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

er forholdet mellem estimatet $\widehat{\alpha}$ over α og den estimerede middelfejl på $\widehat{\alpha}$. 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

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

og estimatet over σ^2 bliver

$$s_{04}^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\Big(y_{ij} - \widehat{\widehat{\beta}} x_i\Big)^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \widehat{\widehat{\beta}}^2 \sum_{i=1}^{k} n_i x_i^2\Big).$$

8.7 Regn og tegn

Hypoxi-eksemplet

Data findes i en separat fil som i nedenstående kode-eksempel hedder `h:/bog/txt304ny/hypoxi.dat`, og hvis første linjer ser sådan ud:

```
tid konc
 0 0
 0 0
 0 1.2
 0 1.8
 0 2.1
 0 2.1
 0 3
 6 3
 6 4.9
```

Her er noget R-kode der viser hvordan man kan få udført tegninger (figur 8.2) og beregninger:

```
hypoxi <- read.table ("h:/bog/txt304ny/hypoxi.dat", header=TRUE)
plot (hypoxi)                      # plot værdierne
M <- lm (konc ~ tid, data=hypoxi) # estimér modellen y=a+bx
abline (coef(M))                   # og indtegn linjen.
# sammenlign med modellen hvor de fire grupper (defineret ved x) har samme middelværdi:
M0 <- update (M, . ~ as.factor(tid) - 1)
# og test M i forhold til M0:
anova (M, M0)
```

Tabel 8.7 Anscombe's data (opgave 8.1).

datasæt 1		datasæt 2		datasæt 3		datasæt 4	
x	y	x	y	x	y	x	y
10	8.04	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

Vedr. opgaverne

En del af datasættene indgår i R-distributionen:

- Forbes' data (opgave 8.2):
  Skriv `require(MASS)` og derefter `data(forbes)`. Derved indlæses en 'data.frame' `forbes` med de to variable `bp` ('boiling point') og `pres` ('pressure'). Herefter kan man skrive f.eks. `plot(forbes)` og `lm (pres ~ bp, data=forbes)`
- Anscombe's data (opgave 8.1):
  Skriv `data(anscombe)`. Derved indlæses en 'data.frame' `anscombe` med de variable `x1`, `x2`, `x3`, `x4`, `y1`, `y2`, `y3` og `y4`, og f.eks. datasæt 1 består så af `anscombe$x1` og `anscombe$y1`.
- Legemsvægt og hjernevægt (opgave 8.3):
  Der findes et datasæt af samme art. Skriv `require(MASS)` og derefter `data(Animals)`, hvorved der indlæses en 'data.frame' `Animals`. Herefter kan man f.eks. afprøve R's interaktive grafik: skriv `plot(log(Animals))` for at få et plot af logaritmen til hjernevægt mod logaritmen til legemsvægt. Skriv så
  `identify(log(Animals), labels=row.names(Animals))`
  og gå ind på grafikvinduet og venstreklik på et af datapunkterne.

8.8 Opgaver

Opgave 8.1 (Anscombe's data)
I tabel 8.7 er vist fire forskellige sæt af (til formålet konstruerede) talpar (x, y) der kan underkastes en lineær regressionsanalyse (3).

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 $\widehat{\alpha}$, $\widehat{\beta}$ og s_{02}^2 over parametrene α, β og σ^2.

Tabel 8.8 Forbes' barometriske målinger (opgave 8.2). – Kogepunktet er angivet i °F, 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. Lav for hvert datasæt et såkaldt *scatterplot*, dvs. en tegning med punkterne (x_i, y_i), og indtegn den estimerede regressionslinje.
3. Hvad kan man lære heraf?

Opgave 8.2 (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 (6) 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 8.8 (fra (22)).

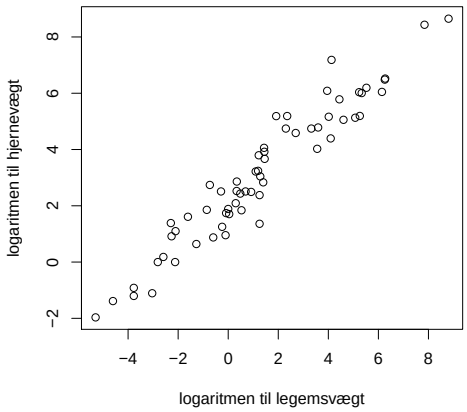
1. Lufttrykket er angivet i 'inches Hg'. Nutildags måles lufttryk i hPa (hektopascal = millibar). Hvordan omregner man lufttrykkene til hPa?

   Kogepunkterne er angivet i °F. 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°F og 100 °C til 212°F.
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 linje der fitter punkterne bedst.

   Indtegn den estimerede linje i figuren.

   Hvordan passer linjen til punkterne?



Figur 8.3 Opgave 8.3: Logaritmen til hjernevægt afsat mod logaritmen til legemsvægt.

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

   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 8.3 (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? I tabel 8.9 på næste side vises den gennemsnitlige legemsvægt og den gennemsnitlige hjernevægt for et antal pattedyr (1) (jf. også (22)). Dyrene er ordnet efter legemsvægt. Opgaven er nu 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 8.3.

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

$$\text{hjernevægt} = \text{konstant} \cdot \text{legemsvægt}^{2/3}. \tag{8.7}$$

   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 (8.7).

Tabel 8.9 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
genette	1.410	17.50
plump-lori	1.400	12.50

(fortsættes)

a Procavia habessinica
b Dendrohyrax

(tabel 8.9 fortsat)

art	legemsvægt (kg)	hjernevægt (g)
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

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?

b) Hvis formel (8.7) gælder, hvilken sammenhæng er der da mellem logaritmen til hjernevægt og logaritmen til legemsvægt?
Hvordan harmonerer formodningen (8.7) 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 generelt finde den bedste rette linje 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 linje (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.

Tabel 8.10 Opgave 8.4: 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 8.11 Opgave 8.4: 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 8.12 Opgave8.4: 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 regressionslinjen	2	2.261966	1.13
omkring regressionslinjen	18	45.925301	2.55
regressionslinjen	1	54.183790	54.18
total	19	100.109091	5.27

Opgave 8.4 (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 8.10 hvor $^{14}CO_2$-aktiviteten angives i dpm (disintegrations per minute) pr. μl porevandsprøve.

I tabel 8.11 er opgivet forskellige hjælpestørrelser, og tabel 8.12 viser (dele af) et variansana-

lyseskema; 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 forløb.
3. Man har en formodning om at den specifikke aktivitet afhænger lineært af tiden. – Estimér regressionslinjen 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?

9 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 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 $x_{i1}, x_{i2}, \ldots, x_{ip}$:

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

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 kort formuleres som $\mathrm{E}\, Y_i = \sum_{j=0}^{p} x_{ij}\beta_j$ eller bedre

$$Y_i \sim \mathcal{N}(\sum_{j=0}^{p} x_{ij}\beta_j, \sigma^2). \tag{9.1}$$

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.

9.1 Estimation af parametrene

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

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

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

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

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 $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.* 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

$$a_{jk} = \sum_{i=1}^{n} x_{ij}x_{ik}, \quad j = 0, 1, 2, \ldots, p;\ k = 0, 1, 2, \ldots, p.$$

Ved at løse de $p+1$ ligninger får man estimaterne $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p$. – 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 forstand at den ikke indeholder anden information end hvad der allerede er indeholdt i de øvrige. (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.) I sådanne situationer plejer man at fjerne den eller de overflødige variable.

Sluttelig kan man udregne residualkvadratsummen

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

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

og variansestimatet

$$s_0^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \Big(y_i - \sum_{j=0}^{p} x_{ij} \widehat{\beta}_j \Big)^2 \tag{9.2}$$

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

9.2 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*

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

og se hvordan de fordeler sig. Hvis modellen (9.1) 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 – jo flere frihedsgrader der er, jo mere uafhængige er de. Man kan derfor se efter om residualerne ser ud til at være nogenlunde uafhængige og normalfordelte.

I afsnit 8.5 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

Størrelsen af variansskønnet s_0^2 fortæller *ikke* noget om hvor godt modellen passer, kun noget om hvor meget punkterne varierer omkring regressionsfladen; en stor 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 $\widehat{y}_i = \sum_{j=0}^{p} x_{ij}\widehat{\beta}_j$. Man kan udregne R^2 efter en af følgende to formler:

$$R^2 = \frac{\Big(\sum_{i=1}^{n}(y_i - \overline{y})(\widehat{y}_i - \overline{y})\Big)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2 \cdot \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}\ , \tag{9.3}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}\ . \tag{9.4}$$

Formel (9.3) fortæller at R^2 er kvadratet på korrelationskoefficienten mellem de observerede og de fittede værdier. Formel (9.4) 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 for de fire datasæt i opgave 8.1!

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

9.3 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 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 *Ockhams ragekniv*, og som siger at man ikke skal antage eksistensen af flere ting end nødvendigt.

Det kan forekomme at man har mange flere baggrundsvariable end man med rimelighed kan have 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 (9.2)).

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

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

og sammenholder resultatet med t-fordelingen med $n - (p + 1)$ frihedsgrader (ɔ: 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 $\widehat{\beta}_k$ signifikant forskellig fra 0, dvs. baggrundsvariabel nr. k har en signifikant virkning og skal derfor forblive i modellen.

Eksempel 9.1 (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 ((5), her citeret efter (18)). 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 9.1 (fra (18)).

Antropologerne mener at x_1 (brøkdel levet i de nye omgivelser) er et godt mål for hvor længe personerne har levet i de civiliserede omgivelser, og at det derfor må være interessant at se hvor godt x_1 kan forklare blodtrykket y. Første skridt er derfor at fitte en simpel lineær regressionsmodel med x_1 som forklarende variabel. Man finder den estimerede regressionslinje 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 9.1, 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 128)

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

Man finder at $\widehat{\beta}_0 = 60.8775$, $\widehat{\beta}_1 = -26.78738$ og $\widehat{\beta}_2 = 1.21726$, så den estimerede regressionsligning er $y = 61 - 27x_1 + 1.2x_2$, og variansestimatet bliver denne gang 96 med 36 frihedsgrader.

Tabel 9.1 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			

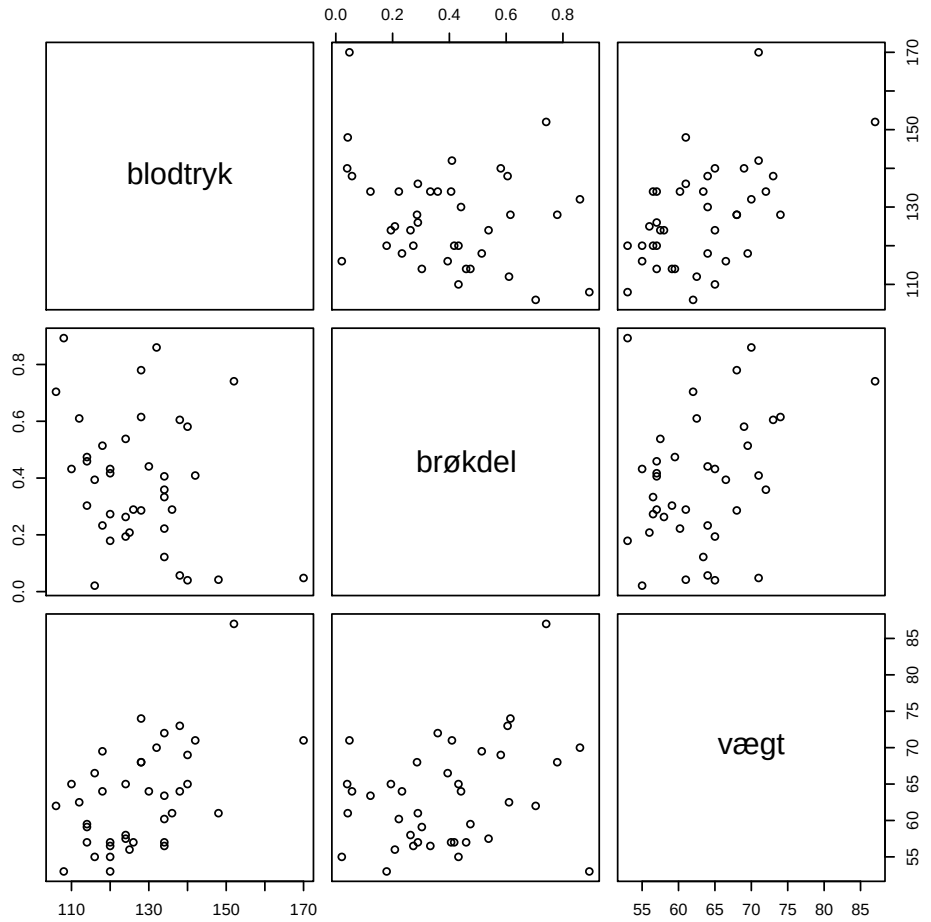
Det ses at ved at inddrage x_2 er variansen gået drastisk ned, fra 163 til 96. Deraf kan man dog ikke slutte at den nye regressionsligning giver en *god* beskrivelse af data, kun at den er bedre end den forrige. Man bør undersøge residualerne for at kunne vurdere modellens kvalitet – det vil vi dog ikke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

Når man lader et statistikprogram foretage udregningerne, vil man også få oplyst parameterestimaternes middelfejl (eng.: 'standard error') og få at vide om parametrene hver især er signifikant forskellige fra 0. I det konkrete tilfælde får man 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 man har, jo højere blodtryk; *men* da det nok også er sådan at jo længere tid man har boet i »civilisationen«, desto mere vejer man, så vil de to virkninger udjævne hinanden hvis man ikke sørger for at inddrage begge forklarende variable.

9.4 Regn og tegn

Peru-eksemplet

Her er en udskrift af en R-session hvor Peru-tallene (eksempel 9.1) analyseres. Datamaterialet indlæses fra en fil der i udskriften hedder `h:/bog/txt304ny/peru.dat`; de første linjer i denne fil ser sådan ud:



Figur 9.1 Indianere i Peru: *Scatterplot-matrix* over de tre variable systolisk blodtryk, brøkdel af livet i de nye omgivelser, og vægt.

```
blodtryk  del  vaegt
170 0.048 71.0
120 0.273 56.5
125 0.208 56.0
148 0.042 61.0
```

Her er R-kørslen (brugeren har skrevet de linjer der begynder med >, scatterplot-matricen fremstilles med funktionen `pairs`, regressionsmodellerne fittes med `lm`):

```
> Peru <- read.table("h:/bog/txt304ny/peru.dat", nrows = 50, header = TRUE)

> pairs (Peru, labels = c("blodtryk", "brøkdel", "vægt"))

> M1 <- lm(blodtryk ~ del, data = Peru)

> summary(M1)
```

```
Call:
lm(formula = blodtryk ~ del, data = Peru)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  133.497      4.038  33.059   <2e-16 ***
del          -15.756      9.014  -1.748   0.0888 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.77 on 37 degrees of freedom
Multiple R-Squared: 0.07628,    Adjusted R-squared: 0.05131
F-statistic: 3.055 on 1 and 37 DF,  p-value: 0.08877


> M2 <- update(M1, . ~ del + vaegt)

> summary(M2)

Call:
lm(formula = blodtryk ~ del + vaegt, data = Peru)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.8775    14.2785   4.264 0.000139 ***
del         -26.7874     7.2180  -3.711 0.000694 ***
vaegt         1.2173     0.2337   5.209 7.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.775 on 36 degrees of freedom
Multiple R-Squared: 0.4733,     Adjusted R-squared: 0.4441
F-statistic: 16.18 on 2 and 36 DF,  p-value: 9.725e-06


> anova(M1, M2)
Analysis of Variance Table

Model 1: blodtryk ~ del
Model 2: blodtryk ~ del + vaegt
Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     37 6033.2
2     36 3440.0  1    2593.3 27.139 7.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vedr. opgave 9.1

Data til opgave 9.1 følger med R-programmet. Skriv `data(trees)` for at hente dem ind i arbejdsområdet, og skriv `trees` for at se tallene. Man kan få hjælp med `?trees`.

Tabel 9.2 Opgave 9.1: 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			

9.5 Opgaver

Opgave 9.1 (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 9.2 ((13), her taget fra (18)).

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 9.2 (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). Måleresultaterne ses i tabel 9.3.

Opgaven er at give en simpel beskrivelse af sammenhængen mellem flowrate og vanddybde. (Hydrologer kan sikkert opstille fornemme differentialligningsmodeller der beskriver denne sammenhæng, forudsat at flodens sider og bund ikke er alt for uregelmæssige. Det er slet ikke det vi er ude efter her. Statistikeren søger blot efter en simpel beskrivelse af de empiriske data.)

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

Tabel 9.3 Opgave 9.2: 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

2. Beregn den bedste rette linje 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 linje? Hvorfor?
4. Hvad er konklusionen mht. sammenhængen mellem flowrate og vanddybde?

10 Logistisk regression

I KAPITEL 3 har vi beskæftiget os med sammenligning af binomialfordelinger og set hvordan man vurderer om der er en signifikant forskel på dem. I nogle situationer er man imidlertid ikke udelukkende interesseret i at vurdere om der er en forskel eller ej, man vil også gerne kunne give en nærmere beskrivelse af forskellen. Vi skal i det følgende vise hvordan man kan indbygge såkaldte *baggrundsvariable* i modellen for (måske) at nå frem til at kunne beskrive forskellen mellem de pågældende binomialfordelinger. – Indeværende kapitel kan desuden ses som et lidt større eksempel på statistisk modelbygningsarbejde.

Som gennemgående eksempel benytter vi endnu engang rismelsbille-eksemplet, nu en større del: I en undersøgelse (jf. (15)) af insekters reaktion over for insektgiften pyrethrum har man udsat nogle rismelsbiller (*Tribolium castaneum*) for forskellige mængder gift og derpå set hvor mange der var døde efter 13 dages forløb. Der er fire forskellige giftkoncentrationer, og forsøget er udført dels på han-biller, dels på hun-biller. Resultaterne (i reduceret form) ses i tabel 10.1.

10.1 Grundmodellen

Første skridt i modelleringsprocessen består i at gøre sig klart at tabel 10.1 giver oplysninger om flere forskellige slags størrelser der skal have hver deres status i modellen:

- Størrelserne »dosis« og »køn« er *baggrundsvariable* der benyttes til at inddele de $144 + 69 + 54 + \ldots + 47 = 641$ elementarforsøg i grupper, idet man forestiller sig at »dosis« og »køn« kan have betydning for udfaldene af de enkelte delforsøg; det kan endda tænkes at selve talværdierne af »dosis« har betydning.
- Totalantallene $(144, 69, 54, \ldots, 47)$ er kendte konstanter, nemlig antal »identiske« gentagelser af de enkelte elementarforsøg.
- Antal døde $(43, 50, 47, \ldots, 43)$ er *observerede værdier af stokastiske variable.*

For overhovedet at få en idé om talmaterialets beskaffenhed kan man lave nogle simple udregninger (tabel 10.2) og tegninger (figur 10.1).

Da der er lavet forsøg med fire forskellige doser og to forskellige køn, er der otte delforsøg med hver sin binomialfordeling, eller sagt mere præcist: i hvert af de otte delforsøg er det nærliggende at foreslå at beskrive »antal døde« som en observation fra en binomialfordeling med en antalsparameter der er det samlede antal biller i den pågældende gruppe, og med en (ukendt) sandsynlighedsparameter der skal fortolkes som sandsynligheden for at en bille af det pågældende køn dør af giften doseret i den pågældende koncentration. Her er det ikke så interessant blot at få at vide om der er en signifikant forskel på grupperne eller ej, det ville være langt mere spændende hvis

Tabel 10.1 Rismelsbillers overlevelse: Tabellen viser *antal døde / totalantal* for hvert køn og for fire forskellige doser (mg/cm^2).

dosis	M	F
0.20	43/144	26/152
0.32	50/ 69	34/ 81
0.50	47/ 54	27/ 44
0.80	48/ 50	43/ 47

Tabel 10.2 Rismelsbillers overlevelse: Observeret dødssandsynlighed (relativ hyppighed) i hver af de otte grupper.

dosis	M	F
0.20	0.30	0.17
0.32	0.72	0.42
0.50	0.87	0.61
0.80	0.96	0.91

man kunne give en nærmere beskrivelse af hvordan sandsynligheden for at dø afhænger af giftkoncentrationen, og hvis man kunne udtale sig om hvorvidt giften virker ens på hanner og hunner. Vi indfører noget notation og præciserer modellen:

1. I den gruppe der svarer til dosis d (hvor $d \in \{0.20, 0.32, 0.50, 0.80\}$) og køn k (hvor $k \in \{M, F\}$), er der n_{dk} biller hvoraf y_{dk} døde.
2. Det antages at y_{dk} er en observation af en stokastisk variabel Y_{dk} som er binomialfordelt med kendt antalsparameter n_{dk} og med sandsynlighedsparameter p_{dk}.
3. Det antages desuden at de enkelte Y_{dk}-er er stokastisk uafhængige.

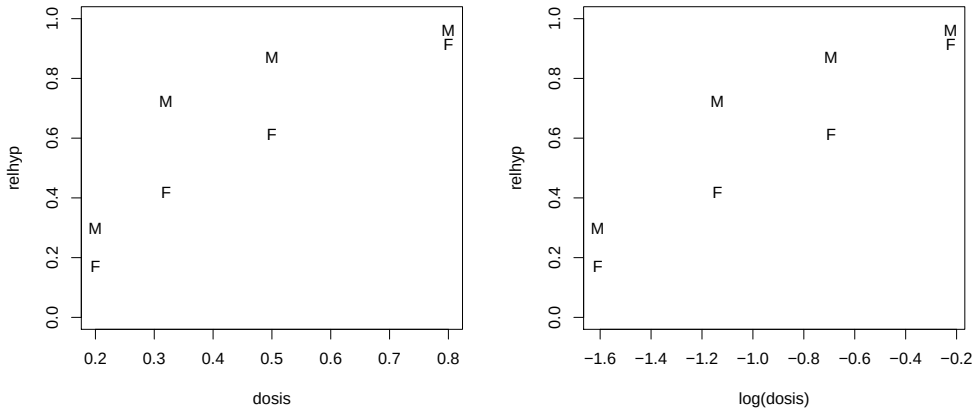
Opgaven er at finde en model der fortæller hvordan p_{dk} afhænger af d og k. Først vil vi se på hvordan man modellerer dosisafhængigheden.

10.2 En dosis-respons model

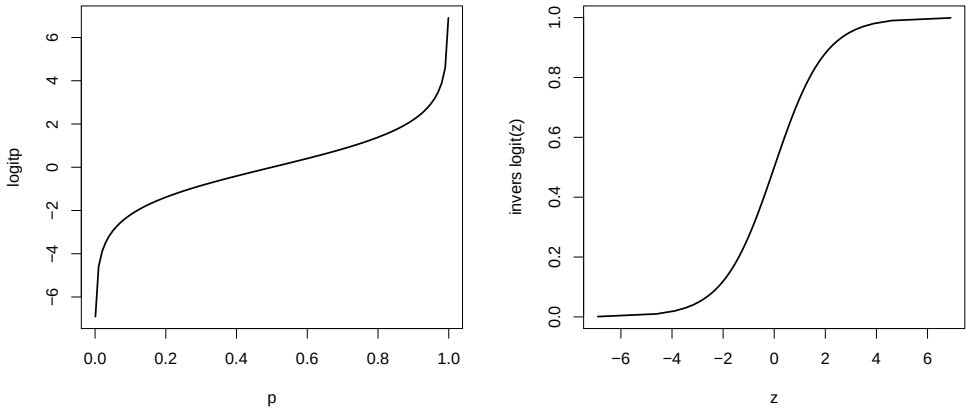
Hvordan er sammenhængen mellem giftkoncentrationen (dosis) d og sandsynligheden p_d for at en bille dør ved denne dosis? Hvis man vidste en hel masse om hvordan netop dette giftstof virker i billeorganismen, kunne man formentlig give et velbegrundet forslag til hvordan sandsynligheden afhænger af dosis. Men den statistiske modelbyggers tilgang til problemet er af en langt mere jordbunden og pragmatisk karakter, som vi nu skal se.

I eksemplet har eksperimentator valgt nogle tilsyneladende mærkværdige dosisværdier (0.20, 0.32, 0.50 og 0.80). Hvis man ser nærmere efter, opdager man dog at der (næsten) er tale om en kvotientrække, idet kvotienten mellem hvert tal og det næste er (næsten) den samme, nemlig 1.6. Det tager den statistiske modelbygger som et fingerpeg om at dosis antagelig skal måles på en *logaritmisk* skala, dvs. man skal interessere sig for hvordan sandsynligheden for at dø afhænger af *logaritmen* til dosis. Dette er grunden til at figur 10.1 også viser den relative hyppighed af døde biller afsat mod *logaritmen* til dosis.

Vi skal modellere sandsynlighedernes afhængighed af baggrundsvariablen $\ln d$. En af de simpleste former for afhængighed er *lineær* afhængighed. Imidlertid ville det være en dårlig idé at foreslå at p_d skulle afhænge lineært af $\ln d$ (altså at $p_d = \alpha + \beta \ln d$ for passende valgte konstanter α og β) fordi dette ville være uforeneligt med kravet om at sandsynlighederne skal ligge mellem 0 og 1. Ofte gør man så det at man omregner p_d til en ny skala og postulerer at »p_d på den ny skala« afhænger lineært af $\ln d$. Omregningen foregår ved hjælp af en særlig funktion ved navn logit:



Figur 10.1 Rismelsbillers overlevelse: Observeret dødssandsynlighed (relativ hyppighed) som funktion af dosis (venstre delfigur) og logaritmen til dosis (højre delfigur), for hvert køn.



Figur 10.2 Venstre del: grafen for logit-funktionen. Højre del: grafen for den omvendte funktion til logit-funktionen.
Der gælder at for $p \to 1$ vil $\text{logit}(p) \to +\infty$, og for $p \to 0$ vil $\text{logit}(p) \to -\infty$.

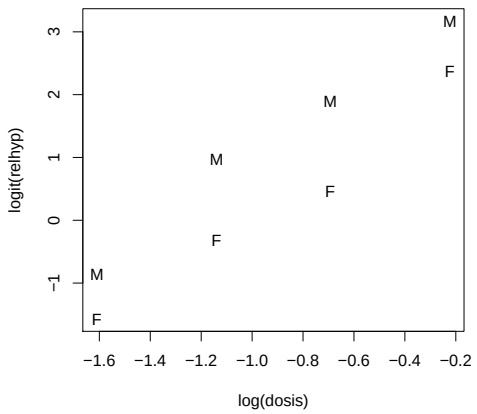
Definition 10.1: logit-funktionen
Funktionen logit *afbilder intervallet* $]0,1[$ *på den reelle akse* $\mathbb{R}$ *og er givet ved*

$$\text{logit}(p) = \ln \frac{p}{1-p}.$$

Hvis $z = \text{logit}(p)$*, så er* $p = \dfrac{\exp(z)}{1+\exp(z)}$.

Figur 10.2 viser grafen for logit-funktionen og dens omvendte funktion.

Bemærkning: Når p er sandsynligheden for en bestemt hændelse (f.eks. at dø), så er $p/(1-p)$ forholdet mellem sandsynligheden for hændelsen og sandsynligheden for den modsatte hændelse; dette tal kaldes med et udtryk hentet fra spillebranchen for *odds* for



Figur 10.3 Rismelsbillers overlevelse: Logit til estimeret dødssandsynlighed (relativ hyppighed) som funktion af logaritmen til dosis, for hvert køn.

den pågældende hændelse. Vi kan dermed sige at logit-funktionen udregner logaritmen til odds.

Vi vil nu foreslå/postulere følgende ofte anvendte model for sammenhængen mellem dosis og sandsynligheden for at dø:

> For hvert af de to køn afhænger $\text{logit}(p_d)$ lineært af $x = \ln d$,

eller mere udførligt:

> Der findes konstanter α_M, β_M og α_F, β_F således at
>
> $$\text{logit}(p_{dM}) = \alpha_M + \beta_M \ln d$$
> $$\text{logit}(p_{dF}) = \alpha_F + \beta_F \ln d \ .$$

I figur 10.3 er logit til de relative hyppigheder afsat mod logaritmen til dosis; hvis modellen er rigtig, skal hvert af de to punktsæt fordele sig tilfældigt omkring en ret linje, og det ser jo ikke helt urimeligt ud; det kræver dog en nærmere undersøgelse for at afgøre om modellen giver en tilstrækkeligt god beskrivelse af datamaterialet.

I de følgende afsnit skal vi se hvordan man estimerer de ukendte parametre (α-erne og β-erne), hvordan man undersøger om modellen er god nok, og hvordan man sammenligner giftens virkningen på han- og hunbiller.

10.3 Estimation

I dette afsnit diskuteres hvordan man estimerer de ukendte parametre α og β i modellen $\text{logit}(p) = \alpha + \beta x$, eller rettere i følgende model:

> Observationerne $y_1, y_2, \ldots, y_s$ er observationer af uafhængige binomialfordelte stokastiske variable $Y_1, Y_2, \ldots, Y_s$, hvor Y_j er binomialfordelt med

antalsparameter n_j (kendt) og sandsynlighedsparameter p_j, og hvor

$$\text{logit}(p_j) = \alpha + \beta x_j,$$

svarende til at

$$p_j = \frac{\exp(\alpha + \beta x_j)}{1 + \exp(\alpha + \beta x_j)}.$$

Her er $x_1, x_2, \ldots, x_s$ kendte tal, og α og β er ukendte parametre.

I bille-eksemplet har vi en sådan model for hvert af de to køn; som x_j bruges logaritmen til koncentrationen i gruppe j.

Likelihoodfunktionen er

$$\begin{aligned} L(\alpha,\beta) &= \prod_{j=1}^{s} \binom{n_j}{y_j} p_j^{y_j} (1-p_j)^{n_j - y_j} \\ &= \prod_{j=1}^{s} \binom{n_j}{y_j} \cdot \prod_{j=1}^{s} \Big(\frac{p_j}{1-p_j} \Big)^{y_j} \cdot \prod_{j=1}^{s} (1-p_j)^{n_j} \\ &= \text{konstant} \cdot \prod_{j=1}^{s} \Big(\frac{p_j}{1-p_j} \Big)^{y_j} \cdot \prod_{j=1}^{s} (1-p_j)^{n_j}, \end{aligned}$$

og log-likelihoodfunktionen er

$$\begin{aligned} \ln L(\alpha,\beta) &= \text{konstant} + \sum_{j=1}^{s} y_j \ln \frac{p_j}{1-p_j} + \sum_{j=1}^{s} n_j \ln(1-p_j) \\ &= \text{konstant} + \sum_{j=1}^{s} y_j \,\text{logit}(p_j) + \sum_{j=1}^{s} n_j \ln(1-p_j) \\ &= \text{konstant} + \sum_{j=1}^{s} y_j(\alpha + \beta x_j) + \sum_{j=1}^{s} n_j \ln(1-p_j) \\ &= \text{konstant} + \alpha \sum_{j=1}^{s} y_j + \beta \sum_{j=1}^{s} x_j y_j + \sum_{j=1}^{s} n_j \ln(1-p_j) \\ &= \text{konstant} + \alpha \sum_{j=1}^{s} y_j + \beta \sum_{j=1}^{s} x_j y_j - \sum_{j=1}^{s} n_j \ln\big(1 + \exp(\alpha + \beta x_j)\big). \end{aligned}$$

Som altid er det bedste bud på værdierne af de ukendte parametre dem der maksimaliserer likelihoodfunktionen eller log-likelihoodfunktionen. Hermed er vi nået til det delproblem der består i at finde maksimumspunkt(er) for funktionen $\ln L$ af de to variable α og β. Den generelle fremgangsmåde går ud på at man søger maksimumspunkterne blandt de stationære punkter for funktionen, dvs. punkter hvor de partielle afledede $\frac{\partial}{\partial \alpha} \ln L$ og

$\frac{\partial}{\partial\beta}\ln L$ er nul. Man finder at

$$\frac{\partial}{\partial\alpha}\ln L(\alpha,\beta)=\sum_{j=1}^{s}(y_j-n_jp_j),$$
$$\frac{\partial}{\partial\beta}\ln L(\alpha,\beta)=\sum_{j=1}^{s}x_j(y_j-n_jp_j),$$

og da disse som nævnt skal være 0, får vi de to ligninger

$$\sum_{j=1}^{s}(y_j-n_jp_j)=0, \tag{10.1}$$
$$\sum_{j=1}^{s}x_j(y_j-n_jp_j)=0, \tag{10.2}$$

med de to ubekendte α og β (der indgår »skjult« i p_j).

Lad os se hvordan disse ligninger tage sig ud for hanbillernes vedkommende. Ligning (10.1) er

$$\begin{aligned}&\left(43-144\,\frac{\exp(\alpha+\beta\ln(0.20))}{1+\exp(\alpha+\beta\ln(0.20))}\right)\\+&\left(50\ -69\,\frac{\exp(\alpha+\beta\ln(0.32))}{1+\exp(\alpha+\beta\ln(0.32))}\right)\\+&\left(47\ -54\,\frac{\exp(\alpha+\beta\ln(0.50))}{1+\exp(\alpha+\beta\ln(0.50))}\right)\\+&\left(48\ -50\,\frac{\exp(\alpha+\beta\ln(0.80))}{1+\exp(\alpha+\beta\ln(0.80))}\right)=0\end{aligned}$$

og ligning (10.2)

$$\begin{aligned}&\ln(0.20)\left(43-144\,\frac{\exp(\alpha+\beta\ln(0.20))}{1+\exp(\alpha+\beta\ln(0.20))}\right)\\+&\ln(0.32)\left(50\ -69\,\frac{\exp(\alpha+\beta\ln(0.32))}{1+\exp(\alpha+\beta\ln(0.32))}\right)\\+&\ln(0.50)\left(47\ -54\,\frac{\exp(\alpha+\beta\ln(0.50))}{1+\exp(\alpha+\beta\ln(0.50))}\right)\\+&\ln(0.80)\left(48\ -50\,\frac{\exp(\alpha+\beta\ln(0.80))}{1+\exp(\alpha+\beta\ln(0.80))}\right)=0.\end{aligned}$$

Det ser ikke rart ud! Faktisk kan man ikke løse disse ligninger, hvis man med »løse ligningerne« mener at flytte rundt på symbolerne så man ender med et resultat af formen »α = noget kendt« og »β = noget kendt«. I stedet må man henvende sig i den afdeling af matematikken der hedder *numerisk analyse*, for at få at vide hvordan man finder en numerisk approksimation til en løsning, hvis der altså overhovedet er en løsning (og man kunne jo også frygte at der var flere løsninger). Eller man kan benytte et passende

statistikprogram på computeren; det vil have indbygget nogle numerisk analyse-metoder så det kan udregne værdierne af maksimaliseringsestimaterne $\widehat{\alpha}$ og $\widehat{\beta}$.

Her er noget af en udskrift fra statistikprogrammet R (koden kan ses i afsnit 10.6 side 147ff):

```
Call:
glm(formula = Ymat ~ sex/(1 + log(dosis)) - 1, family = binomial, data = biller)

Deviance Residuals:
[1]  -0.40748   1.03188  -0.46056  -0.51234   0.06967   0.24861  -0.98679
[8]   0.78606

Coefficients:
Estimate Std. Error z value Pr(>|z|)
sexM              4.2704     0.5372   7.949 1.88e-15 ***
sexF              2.5617     0.3785   6.767 1.31e-11 ***
sexM:log(dosis)   3.1381     0.3854   8.142 3.89e-16 ***
sexF:log(dosis)   2.5816     0.3047   8.472  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 238.0077  on 8  degrees of freedom
Residual deviance:   3.3637  on 4  degrees of freedom
AIC: 44.438
```

Det fremgår blandt andet at for hanbillerne er estimaterne $\widehat{\alpha}_M = 4.270$ (med en middelfejl på 0.5372) og $\widehat{\beta}_M = 3.138$ (med en middelfejl på 0.3854), og for hunbillerne er de $\widehat{\alpha}_F = 2.562$ (med en middelfejl på 0.3785) og $\widehat{\beta}_F = 2.582$ (med en middelfejl på 0.3047).

10.4 Modelkontrol

Vi har nu estimeret parametrene i den model der siger at

$$\operatorname{logit}(p_{dk}) = \alpha_k + \beta_k x \qquad \text{eller} \qquad p_{dk} = \frac{\exp(\alpha_k + \beta_k x)}{1 + \exp(\alpha_k + \beta_k x)}$$

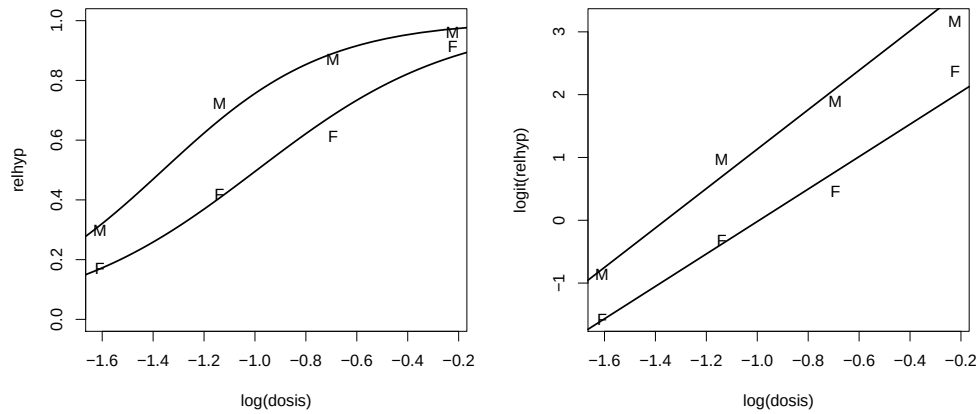
hvor $x = \ln d$. En nærliggende form for modelkontrol er derfor at indtegne graferne for de to funktioner

$$x \mapsto \widehat{\alpha}_M + \widehat{\beta}_M x \qquad \text{og} \qquad x \mapsto \widehat{\alpha}_F + \widehat{\beta}_F x$$

i figur 10.3 og at indtegne graferne for de to funktioner

$$x \mapsto \frac{\exp(\widehat{\alpha}_M + \widehat{\beta}_M x)}{1 + \exp(\widehat{\alpha}_M + \widehat{\beta}_M x)} \qquad \text{og} \qquad x \mapsto \frac{\exp(\widehat{\alpha}_F + \widehat{\beta}_F x)}{1 + \exp(\widehat{\alpha}_F + \widehat{\beta}_F x)}$$

i den højre delfigur af figur 10.1; derved får man henholdsvis højre og venstre del af figur 10.4. Den viser at modellen ikke er helt hen i vejret. Man kan desuden ved hjælp



Figur 10.4 Rismelsbillers overlevelse: To forskellige kurver, samt de observerede relative hyppigheder.

af likelihoodmetoden konstruere et numerisk test baseret på

$$Q = \frac{L(\widehat{\alpha}_M, \widehat{\alpha}_F, \widehat{\beta}_M, \widehat{\beta}_F)}{L_{\max}} \tag{10.3}$$

hvor $L_{\max}$ er likelihoodfunktionens maksimale værdi i den »fulde« model (grundmodellen) hvor p_{dk} estimeres ved den relative hyppighed y_{dk}/n_{dk}.

Med betegnelserne $\widehat{p}_{dk} = \text{logit}^{-1}(\widehat{\alpha}_k + \widehat{\beta}_k \ln d)$ og $\widehat{y}_{dk} = n_{dk}\widehat{p}_{dk}$ bliver

$$\begin{aligned} Q &= \frac{\displaystyle\prod_k \prod_d \binom{n_{dk}}{y_{dk}} \widehat{p}_{dk}^{\,y_{dk}} (1-\widehat{p}_{dk})^{n_{dk}-y_{dk}}}{\displaystyle\prod_k \prod_d \binom{n_{dk}}{y_{dk}} \left(\frac{y_{dk}}{n_{dk}}\right)^{y_{dk}} \left(1-\frac{y_{dk}}{n_{dk}}\right)^{n_{dk}-y_{dk}}} \\ &= \prod_k \prod_d \left(\frac{\widehat{y}_{dk}}{y_{dk}}\right)^{y_{dk}} \left(\frac{n_{dk}-\widehat{y}_{dk}}{n_{dk}-y_{dk}}\right)^{n_{dk}-y_{dk}} \end{aligned}$$

og

$$-2\ln Q = 2\sum_k \sum_d \left(y_{dk} \ln \frac{y_{dk}}{\widehat{y}_{dk}} + (n_{dk}-y_{dk}) \ln \frac{n_{dk}-y_{dk}}{n_{dk}-\widehat{y}_{dk}} \right).$$

Store værdier af $-2\ln Q$ (svarende til små værdier af Q) er tegn på at der er for stor uoverensstemmelse mellem de observerede antal (y_{kd} og $n_{kd} - y_{kd}$) og de forudsagte antal ($\widehat{y}_{kd}$ og $n_{kd} - \widehat{y}_{kd}$) til at modellen kan siges at være god nok. En observeret værdi $-2\ln Q_{\text{obs}}$ er »stor« hvis der kun er lille sandsynlighed for at få en større værdi; denne sandsynlighed (testsandsynligheden) kan bestemmes omtrentligt som sandsynligheden for i χ^2-fordelingen med 4 frihedsgrader at få en værdi større end $-2\ln Q_{\text{obs}}$. I R-udskriften side 143 ses at $-2\ln Q_{\text{obs}} = 3.3637$; den tilhørende testsandsynlighed er 0.4989. – Antallet af frihedsgrader er bestemt på følgende måde: I den »fulde« model (der leverer nævneren i formel (10.3)) er der 8 parametre, én for hver gruppe; i den

testede model (der leverer tælleren i formel (10.3)) er der 4 parametre, nemlig α_M, β_M, α_F og β_F; antal frihedsgrader er ændringen i antal parametre, dvs. $8 - 4 = 4$.

Da der er henved 50% chance for at få et sæt observationer der harmonerer dårligere med den postulerede model, må vi konkludere at modellen ser ud til at være anvendelig.

10.5 Hypoteser om parametrene

Efter at vi har fået opstillet en model som indeholder fire parametre, og som ser ud til at give en ganske god beskrivelse af observationerne, er næste punkt på dagsordenen at undersøge om modellen kan forsimples. Eksempelvis kan man undersøge om de to kurver er parallelle, og hvis det kan accepteres, kan man derefter undersøge om kurverne er sammenfaldende. Vi formulerer derfor to statistiske hypoteser:

1. Hypotesen om parallelle kurver: $H_1 : \beta_M = \beta_F$, eller mere udførligt: Der findes konstanter α_M, α_F og β således at

$$\begin{aligned} \text{logit}(p_{dM}) &= \alpha_M + \beta \ln d \\ \text{logit}(p_{dF}) &= \alpha_F + \beta \ln d. \end{aligned}$$

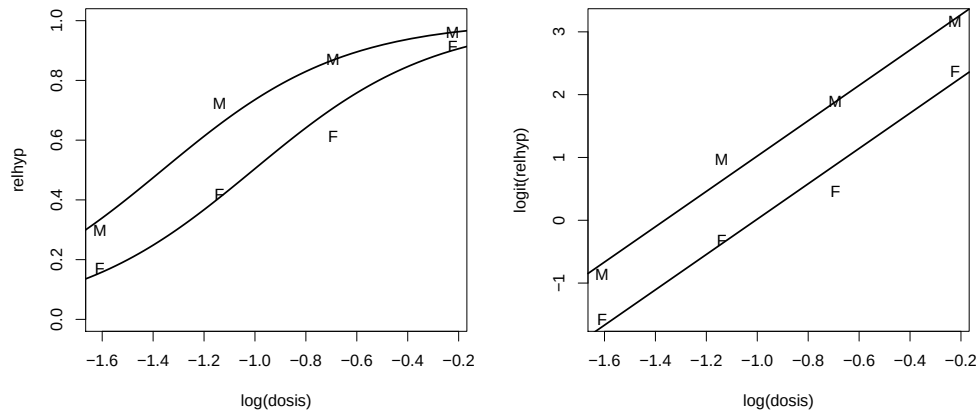
2. Hypotesen om sammenfaldende kurver: $H_2 : \alpha_M = \alpha_F$ og $\beta_M = \beta_F$, eller mere udførligt: Der findes konstanter α og β således at

$$\begin{aligned} \text{logit}(p_{dM}) &= \alpha + \beta \ln d \\ \text{logit}(p_{dF}) &= \alpha + \beta \ln d. \end{aligned}$$

Vi undersøger først hypotesen H_1 om parallelle kurver. De tre parametre estimeres ved maximum likelihood metoden, og det er et problem af samme sværhedsgrad som i grundmodellen (afsnit 10.1). Her ses dele af R-programmets svar:

```
Call:  glm(formula = Ymat ~ sex + log(dosis) - 1, family = binomial, data = biller)

Coefficients:
sexM        sexF  log(dosis)
3.835       2.831       2.813

Degrees of Freedom: 8 Total (i.e. Null);  5 Residual
Null Deviance:      238
Residual Deviance: 4.67         AIC: 43.74

Analysis of Deviance Table

Model 1: Ymat ~ sex + log(dosis) - 1
Model 2: Ymat ~ sex/(1 + log(dosis)) - 1
Resid. Df Resid. Dev Df Deviance
1         5     4.6705
2         4     3.3637  1   1.3067
```

Det ses at parameterestimaterne er $\widehat{\widehat{\alpha}}_M = 3.835$ (med en middelfejl på 0.3443), $\widehat{\widehat{\alpha}}_F = 2.831$ (middelfejl 0.3105) og $\widehat{\widehat{\beta}} = 2.813$ (middelfejl 0.2386). Hypotesen om parallelle



Figur 10.5 Rismelsbillers overlevelse: To parallelle kurver, samt de observerede relative hyppigheder.

kurver testes med det sædvanlige kvotienttest hvor man sammenligner den maksimale likelihoodfunktion under antagelse af H_1 med den maksimale likelihoodfunktion i den senest accepterede model:

$$\begin{aligned} -2\ln Q &= -2\ln\frac{L(\widehat{\widehat{\alpha}}_M, \widehat{\widehat{\alpha}}_F, \widehat{\widehat{\beta}}, \widehat{\widehat{\beta}})}{L(\widehat{\alpha}_M, \widehat{\alpha}_F, \widehat{\beta}_M, \widehat{\beta}_F)} \\ &= 2\sum_k \sum_d \left(y_{dk} \ln \frac{\widehat{y}_{dk}}{\widehat{\widehat{y}}_{dk}} + (n_{dk} - y_{dk}) \ln \frac{n_{dk} - \widehat{y}_{dk}}{n_{dk} - \widehat{\widehat{y}}_{dk}} \right) \end{aligned}$$

hvor $\widehat{\widehat{y}}_{dk} = n_{dk}\,\mathrm{logit}^{-1}(\widehat{\widehat{\alpha}}_k + \widehat{\widehat{\beta}}\ln d)$. Man får at $-2\ln Q_{\mathrm{obs}} = 1.3067$, der kan sammenlignes med χ^2-fordelingen med $4-3=1$ frihedsgrader (ændring i antal parametre). Testsandsynligheden (dvs. sandsynligheden for at få værdier større end 1.3067) er ca. 25%, dvs. værdien 1.3067 er ikke usædvanligt stor. Modellen med parallelle kurver giver således ikke en signifikant dårligere beskrivelse af observationerne end den hidtidige model gør, se også figur 10.5.

Efter således at have accepteret hypotesen H_1 kan vi gå videre med hypotesen H_2 om sammenfaldende kurver. (Hvis H_1 var blevet forkastet, ville man ikke gå videre til H_2.) Man får følgende resultat med R:

```
Call:  glm(formula = Ymat ~ log(dosis), family = binomial, data = biller)

Coefficients:
(Intercept)   log(dosis)
3.204        2.709

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      238
Residual Deviance: 32.17        AIC: 69.25


Analysis of Deviance Table
```

```
Model 1: Ymat ~ log(dosis)
Model 2: Ymat ~ sex + log(dosis) - 1
Resid. Df Resid. Dev Df Deviance
1         6     32.172
2         5      4.670  1   27.502
```

Det fremgår at når man tester H_2 i forhold til H_1, får man $-2\ln Q_{\text{obs}} = 27.502$ der skal sammenlignes med χ^2-fordelingen med et antal frihedsgrader på $3-2=1$; sandsynligheden for at få værdier større end 27.502 er lig nul med adskillige betydende cifre, hvilket viser at modellen med sammenfaldende kurver giver en væsentligt dårligere beskrivelse af observationerne end den forrige model gør. Vi må derfor forkaste hypotesen om sammenfaldende kurver.

Konklusionen på det hele er således at vi kan beskrive sammenhængen mellem dosis d og sandsynligheden p for at dø på den måde at for hvert køn afhænger $\operatorname{logit} p$ lineært af $\ln d$; de to kurver er parallelle, men ikke sammenfaldende. De estimerede kurver er

$$\begin{aligned}\operatorname{logit}(p_{dM}) &= 3.84 + 2.81\ln d\\ \operatorname{logit}(p_{dF}) &= 2.83 + 2.81\ln d,\end{aligned}$$

svarende til at

$$\begin{aligned}p_{dM} &= \frac{\exp(3.84+2.81\ln d)}{1+\exp(3.84+2.81\ln d)}\\ p_{dF} &= \frac{\exp(2.83+2.81\ln d)}{1+\exp(2.83+2.81\ln d)}.\end{aligned}$$

10.6 Regn og tegn

Her er den R-kode der foretager udregningerne og tegningerne til dette kapitel.

Data er på forhånd lagt i en fil som her hedder `h:/bog/txt304ny/biller.dat`, og som har følgende indhold:

```
sex dosis dead total
 M   0.2   43   144
 M   0.32  50    69
 M   0.5   47    54
 M   0.8   48    50
 F   0.2   26   152
 F   0.32  34    81
 F   0.5   27    44
 F   0.8   43    47
```

Tallene indlæses i en 'data.frame' `biller`, og der udregnes nogle ekstra størrelser:

```
biller <- read.table ("h:/bog/txt304ny/biller.dat", nrows=10, header=TRUE)

biller$overlev <- biller$total - biller$dead
biller$relhyp <- biller$dead / biller$total
```

```
# glm-funktionen (se senere) har brug for en 'Ymat':
biller$Ymat <- cbind(biller$dead, biller$overlev)

biller # skriv indholdet af biller
```

Fremstilling af delfigurerne i figur 10.1:

```
# relhyp som funktion af dosis, for hvert køn.
# ('ylim=c(0,1)' tvinger y-aksen til at gå fra 0 til 1)
plot(relhyp ~ dosis, pch=as.character(sex), ylim=c(0,1), data=biller, ask=FALSE)
# samme, men med log(dosis) ud ad x-aksen:
plot(relhyp ~ log(dosis), pch=as.character(sex), ylim=c(0,1), data=biller, ask=FALSE)
```

Vi definerer funktionen $\text{logit}(p) = \ln(p/(1-p))$ og tegner graferne for funktionen selv og den omvendte (figur 10.2):

```
# Definér en funktion logit
logit <- function (p){ log(p/(1-p)) }

# tegn logit-funktionen:
p <- c(0.001, 1:99/100, 0.999)
logitp <- logit(p)
plot (p, logitp, type="l", lwd=2)

# den omvendte funktion får vi ved at bytte om på argumenterne:
plot (logitp, p, type="l", lwd=2, xlab="z", ylab="invers logit(z)")
```

Så tegnes figur 10.3:

```
# logit(relhyp) som funktion af log(dosis), for hvert køn:
plot(logit(relhyp) ~ log(dosis), pch=as.character(sex), data=biller, ask=FALSE)
```

Estimation i grundmodellen

Grundmodellen specificeres ved formlen `Ymat ~ sex/(1+log(dosis))-1`. Det der står til venstre for `~` er den afhængige variabel. På højresiden betyder `1+log(dosis)` regression med `log(dosis)` som uafhængig variabel og med eksplicit konstantled; `sex/(1+log(dosis))` betyder så at dette gøres for hvert niveau af `sex`, og `sex/(1+log(dosis))-1` betyder at den samlede regression ikke skal have et konstantled:

```
grundmodel <- glm (Ymat ~ sex/(1+log(dosis))-1, family=binomial, data=biller)
summary(grundmodel)

grundmodel$coef # koefficienterne

# grundmodellens 'deviance' er -2lnQ-størrelsen:
grundmodel$deviance

grundmodel$df.residual # dens antal frihedsgrader er
# testsandsynligheden er
1-pchisq(grundmodel$deviance, df=grundmodel$df.res)
# -- eller
1-pchisq(3.3637, df=4)
```

Vi kan så lave den forrige tegning med de estimerede regressionslinjer indtegnet (figur 10.4) (`abline` indtegner linjen $y = a + bx$ i det aktuelle plot):

```
# logit(relhyp) som funktion af log(dosis) samt estimeret linje, for hvert køn:
plot (logit(relhyp) ~ log(dosis), pch=as.character(sex), data=biller, ask=FALSE)
abline (grundmodel$coef[c(1,3)], lwd=2) # koeff. nr. 1 og 3 hører til hanbillerne
abline (grundmodel$coef[c(2,4)], lwd=2) # koeff. nr. 2 og 4 hører til hunbillerne
```

Det er lidt sværere at få indtegnet de estimerede kurver i figuren med relative hyppigheder ud ad ordinataksen:

```
plot(relhyp ~ log(dosis), pch=as.character(sex), ylim=c(0,1), data=biller, ask=FALSE)
p <- seq(0.1, 0.98, by=0.01)
logitp <- logit(p)
lines ((logitp-grundmodel$coef[1])/grundmodel$coef[3], p, lwd=2)
lines ((logitp-grundmodel$coef[2])/grundmodel$coef[4], p, lwd=2)
```

To parallelle linjer

Vi opdaterer modellen så den giver to parallelle linjer; derefter testes den i forhold til grundmodellen:

```
model1 <- update (grundmodel, . ~ sex + log(dosis) -1)
anova (model1, grundmodel) # test af model 1 i forhold til grundmodellen
```

Så tegnes figur 10.5:

```
plot (logit(relhyp) ~ log(dosis), pch=as.character(sex), data=biller, ask=FALSE)
abline (model1$coef[c(1,3)], lwd=2)
abline (model1$coef[c(2,3)], lwd=2)

plot(relhyp ~ log(dosis), pch=as.character(sex), ylim=c(0,1), data=biller, ask=FALSE)
p <- seq(0.1, 0.98, by=0.01)
logitp <- logit(p)
lines ((logitp-model1$coef[1])/model1$coef[3], p, lwd=2)
lines ((logitp-model1$coef[2])/model1$coef[3], p, lwd=2)
```

Ingen forskel på de to køn

```
model2 <- update (model1, . ~ log(dosis))
anova (model2, model1) # test af model2 i forhold til model1
```

10.7 Opgaver

Opgave 10.1
Vis at $\text{logit}(1/2) = 0$. Vis at $\text{logit}(1-p) = -\text{logit}(p)$.

Opgave 10.2
Eftervis at $\text{logit}(p) = z$ hvis og kun hvis $p = \dfrac{\exp(z)}{1+\exp(z)}$ således som det postuleres i definition 10.1 på side 138.

Opgave 10.3
Indfør en funktion $p(x)$ ved $p(x) = \dfrac{\exp(a+bx)}{1+\exp(a+bx)}$, dvs. $\text{logit}(p(x)) = a + bx$. (Her er a og b to konstanter.)

1. Skitsér grafen for $p(x)$ når $a = 3$ og $b = 0.5$.
2. Skitsér grafen for $p(x)$ når $a = 3$ og $b = -0.5$.

3. Løs ligningen $p(x) = 0.5$ (for generelle a og b).
   Lad os sige at $x = \ln d$ hvor d er en dosis af et giftstof, og at $p(x) = p(\ln d)$ betyder sandsynligheden for at dø når giften gives i dosis d. Man er undertiden interesseret i at finde den dosis for hvilken sandsynligheden for at dø netop er 50% (den såkaldte LD50), dvs. finde det d for hvilket $p(\ln d) = 0.5$.

11 Poissonfordelingen

Dette kapitel introducerer poissonfordelingen der er opkaldt efter den franske matematiker og fysiker Siméon-Denis Poisson (1781-1840). Poissonfordelingsmodeller kan blandt andet komme på tale når man har at gøre med *antalsobservationer* der angiver hvor mange gange et bestemt fænomen optræder i et vist tidsrum og/eller et vist geografisk område eller lignende (det kunne for eksempel være trafikulykker på et år på en bestemt vejstrækning).

Det gennemgående eksempel i dette kapitel hidrører fra (4) og kan måske i første omgang forekomme lidt kuriøst, men optræder i næsten alle lærebøger i statistik:

Eksempel 11.1 (Hestespark)
For hvert af de 20 år fra 1875 til 1894 har man for hvert af den prøjsiske armés 10 regimenter registreret hvor mange soldater der døde fordi de blev sparket af en hest. Det vil sige at man for hvert af de 200 »regiment-år« kender antal dødsfald som følge af hestespark.

Man kan give en oversigt over disse tal ved at angive i hvor mange regiment-år der var 0 dødsfald, i hvor mange der var 1 dødsfald, i hvor mange der var 2, osv., dvs. man klassificerer regiment-årene efter antal dødsfald. Det viste sig at det største antal dødsfald pr. regiment-år var fire. Ved klassificeringen bliver der derfor fem klasser svarende til 0, 1, 2, 3 og 4 døde pr. år. Tabel 11.1 viser hvordan de faktiske tal blev.

Man må formode at det i høj grad var tilfældigheder der bestemte om en given soldat blev sparket til døde af en hest eller ej. Derfor er det også i høj grad tilfældigheder der har afgjort om et givet regiment i et givet år nu fik 0 eller 1 eller 2 osv. døde som følge af hestespark. Der kan således være fornuft i at beskæftige sig med denne modelbygningsopgave: *Find et forslag til en matematisk model der kan levere sandsynligheder for at have netop y døde i et bestemt regiment, $y = 0, 1, 2, \ldots$.*

11.1 Udledning

En væsentlig del af problemløsningsprocessen består i at oversætte problemet til matematik i en passende generel formulering. Vi går frem i en række punkter der dels leder frem til en sådan passende formulering, dels leverer en løsning.

1. Hestesparkeksemplet handler om at man 200 gange har foretaget sig noget bestemt, nemlig fulgt et regiment igennem et år og set hvor mange dødsfald der var som følge af hestespark.
2. Der er et »grundeksperiment« der består i at man i et vist tidsinterval (af længde 1 år) holder øje med hvor mange gange en bestemt type begivenhed (dødsfald ved hestespark) indtræffer.

Tabel 11.1 Antal dødsfald som følge af hestespark i den prøjsiske armé.

antal dødsfald y	antal regiment-år med y dødsfald
0	109
1	65
2	22
3	3
4	1
	200

3. Grundeksperimentet består i at der i tidsintervallet fra t_0 til t_1 registreres antal forekomster af en bestemt art begivenhed.
4. Vi kan dele tidsintervallet fra t_0 til t_1 op i et antal lige store delintervaller som hver især har længden Δt. På den måde bliver der

$$n = n(\Delta t) = \frac{t_1 - t_0}{\Delta t}$$

delintervaller. (I hestesparkeksemplet kan man for eksempel dele intervallet $]t_0, t_1]$ af længde 1 år op i 365 delintervaller af længde $\Delta t = 1\,\text{dag}$.)
Antallet af begivenheder i det store interval er (selvfølgelig) lig med summen af antal begivenheder i de enkelte delintervaller.
5. Fidusen ved at dele op i delintervaller er at hvis Δt er tilstrækkelig lille, så er det meget usandsynligt at der indtræffer to eller flere begivenheder i *samme* delinterval. Sagt på en anden måde, hvis Δt er meget lille, så er det samlede antal begivenheder i intervallet $]t_0, t_1]$ stort set lig med antallet af de delintervaller hvori der forekommer mindst én begivenhed.
6. Vi har nu fået lavet problemet om til at handle om 01-variable, nemlig om

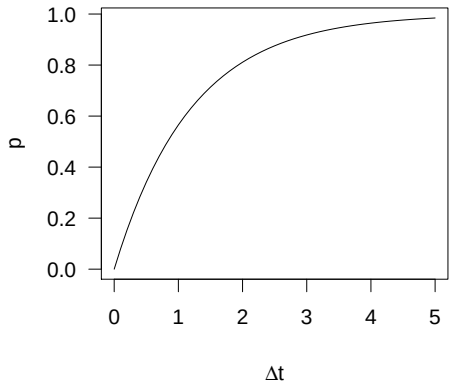
$$I_j = \begin{cases} 1 & \text{hvis der er mindst én begivenhed i delinterval nr. } j \\ 0 & \text{hvis der ingen begivenhed er i delinterval nr. } j \end{cases}$$

$j = 1, 2, \ldots, n$.
Hvis Δt er meget lille, så er det samlede antal Y af begivenheder i intervallet $]t_0, t_1]$ ifølge betragtningerne i punkt 5 cirka lig med $I_1 + I_2 + \cdots + I_n$.
7. Antag at der i alle $n = n(\Delta t)$ delintervaller er *den samme* sandsynlighed $p = p(\Delta t)$ for at der sker en begivenhed. (Der bliver altså ikke i løbet af perioden indført nye sikkerhedsforanstaltninger der nedsætter chancen for at blive sparket til døde af en hest. Og antallet af soldater og af heste i regimentet er stort set konstant året igennem.)
Antag også at det der sker i ét interval, er *stokastisk uafhængigt* af det der sker i andre intervaller. (Selv om der tilfældigvis er to soldater der i begyndelsen af året blev sparket til døde af heste, så tager de øvrige soldater i regimentet ikke ekstra forholdsregler i resten af året i den anledning.)

8. Da $I_1, I_2, \ldots, I_n$ således er uafhængige og identisk fordelte 01-variable, er $\sum_{j=1}^{n} I_j$ binomialfordelt med parametre $n = n(\Delta t)$ og $p = p(\Delta t)$, og da totalantallet Y af begivenheder i $]t_0, t_1]$ cirka er lig med $\sum_{j=1}^{n} I_j$, er Y således cirka binomialfordelt med parametre n og p.
   Forbeholdet »cirka« bortfalder når Δt bliver tilstrækkelig lille, dvs. vi skal på et senere stadium lade Δt gå mod nul.
9. Den måde hvorpå n afhænger af Δt, er simpel, idet som tidligere anført

$$n = n(\Delta t) = \frac{t_1 - t_0}{\Delta t}.$$

   Derimod mangler vi at overveje hvordan p afhænger af Δt.
10. Det må være rimeligt at formode at p er en forholdsvis pæn funktion af Δt, bl.a. med den egenskab at $p(\Delta t) \to 0$ når $\Delta t \to 0$, og at $p(\Delta t) \to 1$ når $\Delta t \to +\infty$, så $p(\Delta t)$ må have et udseende i retning af



   Vi vil gå ud fra at $p(\Delta t)$ er differentiabel fra højre i $\Delta t = 0$, mere præcist at der eksisterer et tal $\lambda > 0$ således at

$$\lim_{\Delta t \to 0} \frac{p(\Delta t)}{\Delta t} = \lambda.$$

   Der gælder altså at $p(\Delta t) \approx \lambda \Delta t$ for små værdier af Δt.
11. I punkt 8 nåede vi frem til at Y er cirka binomialfordelt, dvs. at

$$\mathrm{P}(Y = y) \approx \binom{n}{y} p^y (1-p)^{n-y} \tag{11.1}$$

   hvor »≈« bliver til »=« når $\Delta t \to 0$. Derfor må det næste skridt være at bestemme grænseværdien $\lim \binom{n}{y} p^y (1-p)^{n-y}$ under den grænseovergang hvor $\Delta t \to 0$ og dermed $n = \dfrac{t_1 - t_0}{\Delta t} \to \infty$.

I punkt 10 vedtog vi at der under denne grænseovergang skal gælde at $\dfrac{p}{\Delta t} = \dfrac{p(\Delta t)}{\Delta t} \to \lambda$, og derfor vil

$$np = \frac{(t_1 - t_0) \cdot p}{\Delta t} \longrightarrow \lambda \cdot (t_1 - t_0). \tag{11.2}$$

12. Vi omskriver binomialsandsynligheden på følgende måde:

$$\begin{aligned}\binom{n}{y} p^y (1-p)^{n-y} &= \frac{n}{1}\,\frac{n-1}{2}\,\ldots\,\frac{n-y+1}{y} \cdot p^y \cdot (1-p)^{-y} \cdot (1-p)^n \\ &= \underbrace{1\,(1-\tfrac{1}{n})\,\ldots\,(1-\tfrac{y-1}{n})}_{\text{(a)}} \cdot \underbrace{\frac{(np)^y}{y!}}_{\text{(b)}} \cdot \underbrace{(1-p)^{-y}}_{\text{(c)}} \cdot \underbrace{(1-p)^n}_{\text{(d)}}.\end{aligned}$$

13. Under grænseovergangen vil de forskellige faktorer opføre sig på forskellige måder:

    a) $\underbrace{1\,(1-\frac{1}{n})\,\ldots\,(1-\frac{y-1}{n})}_{y \text{ faktorer}} \to 1^y = 1.$

    b) $\dfrac{(np)^y}{y!} \to \dfrac{\big(\lambda \cdot (t_1 - t_0)\big)^y}{y!}.$

    c) $(1-p)^{-y} \to (1-0)^{-y} = 1\ .$

    d) $(1-p)^n \to \exp\big(-\lambda \cdot (t_1 - t_0)\big)$, hvilket indses således:

        i. Da funktionen $x \mapsto \ln x$ er differentiabel i $x = 1$ med differentialkvotient 1, vil for $h \to 0$

$$\frac{\ln(1+h)}{h} = \frac{\ln(1+h) - \ln 1}{h} \to 1\ .$$

        ii. Ved at benytte dette samt formel (11.2) fås

$$n \ln(1-p) = -np \cdot \frac{\ln(1-p)}{-p} \to -\lambda \cdot (t_1 - t_0).$$

        iii. Ved at tage exp på begge sider heraf fås at

$$(1-p)^n \to \exp\big(-\lambda \cdot (t_1 - t_0)\big)$$

        som ønsket.

Alt i alt vil binomialsandsynligheden i formel (11.1) konvergere mod

$$\frac{\big(\lambda \cdot (t_1 - t_0)\big)^y}{y!}\ \exp\big(-\lambda \cdot (t_1 - t_0)\big).$$

Vi er hermed nået frem til følgende forslag til en statistisk model: Sandsynligheden for at der i et bestemt regiment er netop y dødsfald i perioden $]t_0, t_1]$ må være

$$\mathrm{P}(Y = y) = \frac{\big(\lambda \cdot (t_1 - t_0)\big)^y}{y!}\ \exp\big(-\lambda \cdot (t_1 - t_0)\big), \tag{11.3}$$

hvor λ er en positiv konstant og $y = 0, 1, 2, 3, \ldots$ – Bemærk at de hjælpestørrelser n og Δt som vi indførte i punkt 4, helt er forsvundet.

I formel (11.3) optræder den ukendte parameter λ der i punkt 10 blev indført som værende cirka sandsynligheden for en begivenhed i et meget kort tidsinterval divideret med tidsintervallets længde. Størrelsen λ har derfor dimensionen tid^{-1}, dvs. λ angives i f.eks. dag^{-1} eller år^{-1}.

Jo større λ er, jo tilbøjeligere er begivenhederne til at indtræffe; λ er en såkaldt *intensitet* (der i hestesparkeksemplet specielt kunne kaldes for en *ulykkesintensitet* eller en *dødsintensitet*).

11.2 Definition og egenskaber

Man definerer poissonfordelingen således:

Definition 11.1: Poissonfordeling
Poissonfordelingen med parameter $\mu \geq 0$ er den sandsynlighedsfordeling på udfaldsrummet $\mathscr{X} = \{0, 1, 2, \ldots\}$ som har sandsynlighedsfunktion

$$f(y; \mu) = \frac{\mu^y}{y!} \exp(-\mu).$$

I figur 11.1 ses nogle poissonfordelinger.

Det resultat vi fandt i forrige afsnit, kan derefter udtrykkes på den måde at antallet af dødsfald i et bestemt regiment i perioden fra t_0 til t_1 er poissonfordelt med parameter $\mu = \lambda\,(t_1 - t_0)$ hvor λ betegner dødsintensiteten.

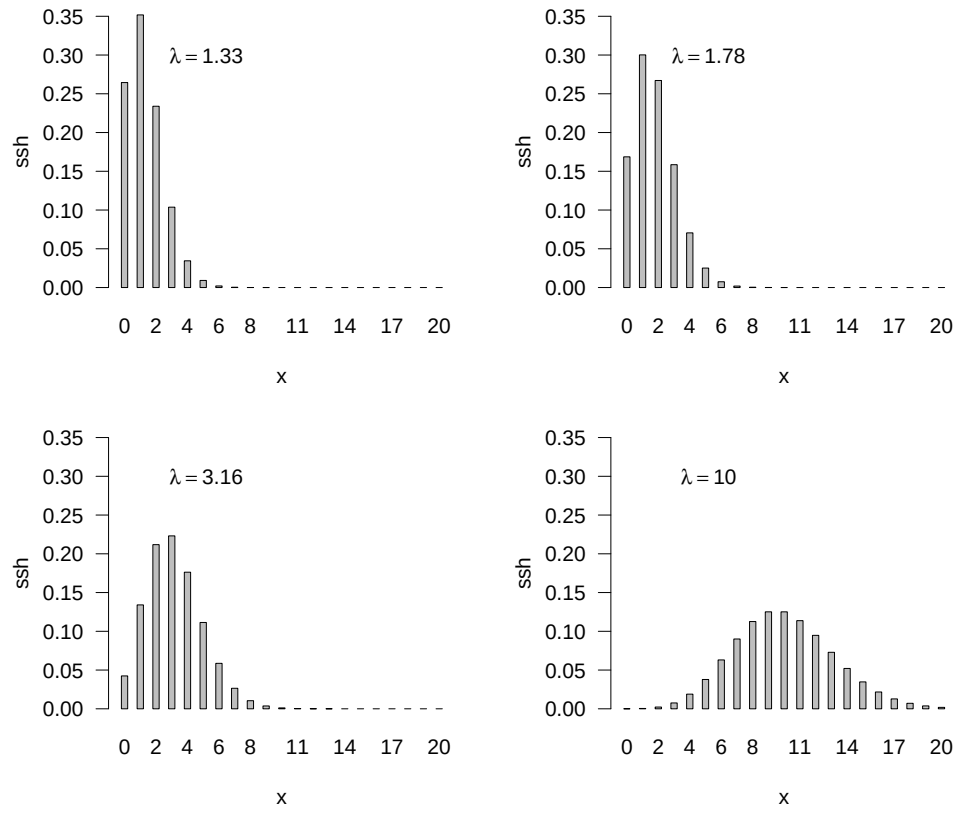
Bemærkning
Strengt taget bør definitionen af poissonfordelingen følges op af en redegørelse for at $f(y; \mu)$ faktisk *er* en sandsynlighedsfunktion, dvs. at der er tale om ikke-negative tal der summerer til 1. Det er klart at f-erne er ikke-negative; at de summerer til 1 følger af eksponentialfunktionens rækkeudvikling. □

En egenskab ved poissonfordelingen er at dens middelværdi er lig dens varians: *Hvis den stokastiske variabel Y er poissonfordelt med parameter μ, så er* $\mathrm{E}\,Y = \mu$ *og* $\mathrm{Var}\,Y = \mu$.

Bevis
Det vises på følgende måde der kræver lidt kendskab til uendelige rækker: Middelværdien af Y er pr. definition $\mathrm{E}\,Y = \sum_{y=0}^{\infty} y \cdot f(y)$, og der gælder:

$$\begin{aligned}\mathrm{E}\,Y &= \sum_{y=0}^{\infty} y\,\frac{\mu^y}{y!}\exp(-\mu) = \sum_{y=1}^{\infty} y\,\frac{\mu^y}{y!}\exp(-\mu)\\ &= \mu\Big(\sum_{y=1}^{\infty}\frac{\mu^{y-1}}{(y-1)!}\Big)\exp(-\mu) = \mu\Big(\sum_{n=0}^{\infty}\frac{\mu^n}{n!}\Big)\exp(-\mu) = \mu.\end{aligned}$$

Variansen af Y er $\mathrm{Var}(Y) = \mathrm{E}\big((Y - \mathrm{E}\,Y)^2\big) = \mathrm{E}(Y^2) - (\mathrm{E}\,Y)^2$. Vi kender $\mathrm{E}\,Y$, men $\mathrm{E}(Y^2)$ er besværlig at regne ud; det er smart at benytte omskrivningen $\mathrm{E}(Y^2) = \mathrm{E}\big(Y(Y-1) + Y\big) = \mathrm{E}\big(Y(Y-1)\big) + \mathrm{E}\,Y$.



Figur 11.1 Poissonfordelinger med middelværdier 1.33, 1.78, 3.16 og 10.

Nu er

$$\begin{aligned}\mathrm{E}(Y(Y-1)) &= \sum_{y=0}^{\infty} y(y-1)\,\frac{\mu^y}{y!}\exp(-\mu) = \sum_{y=2}^{\infty} y(y-1)\,\frac{\mu^y}{y!}\exp(-\mu)\\ &= \mu^2\Big(\sum_{y=2}^{\infty}\frac{\mu^{y-2}}{(y-2)!}\Big)\exp(-\mu) = \mu^2\Big(\sum_{n=0}^{\infty}\frac{\mu^n}{n!}\Big)\exp(-\mu) = \mu^2,\end{aligned}$$

så $\mathrm{Var}(Y) = \mathrm{E}\big(Y(Y-1)\big) + \mathrm{E}\,Y - (\mathrm{E}\,Y)^2 = \mu^2 + \mu - \mu^2 = \mu$. □

11.3 Afrunding

Her er nogle flere eksempler på situationer der kan give poissonfordelte antal:

- Antal tilfælde af en bestemt (ikke-smittende) sygdom i et bestemt tidsrum.
- Antal ulykkestilfælde af en bestemt art i et bestemt tidsrum.
- Antal omdannelser af atomer i et radioaktivt stof i et bestemt tidsrum (der er forsvindende i forhold til stoffets halveringstid).
- Antal trykfejl i en bog. – Her er »tidsaksen« simpelthen teksten forstået som en følge af tegn. Der er altså tale om en diskret tidsakse, og ræsonnementerne der

førte frem til poissonfordelingen, beror i høj grad på at tidsaksen er kontinuert. Men hvis der kun er få trykfejl i forhold til antallet af bogstaver og tegn, så kan man »næsten ikke« se at tidsaksen faktisk er diskret. Derfor finder man på alligevel at anvende poissonfordelingen.

- Antal bombenedfald i London under det tyske bombardement under Anden Verdenskrig – her er »tidsaksen« det (todimensionale) geografiske område London.

11.4 Opgaver

Opgave 11.1

I næste kapitel vil det vise sig at det i hestesparkeksemplet er fornuftigt at estimere λ ved $\hat{\lambda} = 0.61$ dødsfald pr. år.

Lav en tegning (f.eks. i stil med dem der er fire af i figur 11.1) der viser hvordan poissonfordelingen med parameter 0.61 ser ud.

Tip: Når man skal udregne $f(y;\mu) = \dfrac{\mu^y}{y!}\exp(-\mu)$ for en hel masse y-værdier, kan det være smart at gøre det rekursivt:

$$f(0;\mu) = \exp(-\mu)$$
$$f(y;\mu) = \frac{\mu}{y} \cdot f(y-1;\mu), \quad y = 1, 2, 3, \ldots$$

Opgave 11.2 (Raunkiær-cirklinger)

Inden for planteøkologi bestemmer man (i Danmark) ofte planters skudtæthed ved hjælp af en metode der kaldes Raunkiær-cirklinger. I sin simpleste form er metoden som følger (tænk på at det handler om at undersøge planter på en mark): Man anbringer et tilfældigt sted på prøvearealet en cirkel med areal a og ser efter om den planteart man undersøger, findes inden for cirklen eller ej; dette gentages n gange (idet man sørger for at de n cirkler ikke overlapper). Typisk er $a = 0.1\text{m}^2$ og $n = 10$.

Antag for eksempel at man i 10 cirklinger med en 0.1m^2 cirkel fik netop 7 tilfælde hvor planten blev fundet inden for cirklen.

Man ønsker som nævnt at bestemme skudtætheden λ (der måles i antal/m^2). Man må derfor gøre en antagelse om at en bestemt slags sandsynlighedsmodel har placeret skuddene ud over marken. Den simpleste antagelse er at skuddene er placeret efter en poisson-proces, hvilket betyder at antal skud i et delområde med areal a er poissonfordelt med parameter λa, og at antal skud i disjunkte delområder er stokastisk uafhængige.

1. Hvad er sandsynligheden for at man ved én cirkling oplever at der er netop k skud inde i cirklen?
2. Hvad er sandsynligheden for at man ved én cirkling oplever at der er mindst et skud inde i cirklen?

   Tip: »mindst et« er det modsatte af »ingen«.
3. Hvis man udfører $n = 10$ cirklinger, hvad er da sandsynligheden for at der i netop $y = 7$ tilfælde findes mindst et skud inde i cirklen?

(Raunkiær-cirklinger genoptages i opgave 12.3.)

12 En- og flerstikprøveproblemer i poissonfordelingen

I KAPITEL 11 NÅEDE VI ved teoretiske overvejelser frem til at antallet af dødsfald pr. regiment pr. år måtte være poissonfordelt med parameter $\mu = \lambda \cdot 1\,\text{år}$, men stemmer det overhovedet med virkeligheden, og hvordan estimerer man intensiteten λ?

Vi skal i dette kapitel beskæftige os med estimation af parametre og test af hypoteser om parametre i poissonfordelinger, og med spørgsmålet om kontrol af modellen.

12.1 Enstikprøveproblemet

I hestespark-eksemplet fra kapitel 11 er situationen den at der er $n = 200$ uafhængige observationer $y_1, y_2, \ldots, y_n$ fra poissonfordelingen med parameter $\mu = \lambda \cdot 1\,\text{år}$. Det er et eksempel på et *enstikprøveproblem* fordi der er tale om et antal observationer, en *stikprøve*, fra en og samme fordeling.

Generelt har man at gøre med uafhængige observationer $y_1, y_2, \ldots, y_n$ fra en poissonfordeling med parameter μ, svarende til at *modelfunktionen* er

$$f(y_1, y_2, \ldots, y_n; \mu) = \prod_{j=1}^{n} \frac{\mu^{y_j}}{y_j!} \exp(-\mu) = \frac{\mu^{y_\bullet}}{\prod_{j=1}^{n} y_j!} \exp(-n\mu).$$

Her er $y_\bullet$ statistikerens sædvanlige korte skrivemåde for $y_1 + y_2 + \cdots + y_n$.

Estimation af parameteren

Poissonparameteren μ estimeres ved likelihoodmetoden. Likelihoodfunktionen svarende til observationerne $y_1, y_2, \ldots, y_n$ er

$$L(\mu) = \frac{\mu^{y_\bullet}}{\text{konstant}} \exp(-n\mu),$$

så at

$$\ln L(\mu) = \text{konstant} + y_\bullet \ln \mu - n\mu \ .$$

Ifølge de sædvanlige principper er det bedste estimat over μ den værdi $\widehat{\mu}$ der maksimaliserer L eller $\ln L$. For at bestemme denne værdi løser vi ligningen $\frac{d}{d\mu} \ln L = 0$. Man finder at

$$\frac{d}{d\mu} \ln L(\mu) = \frac{y_\bullet}{\mu} - n$$

som er lig 0 netop når μ er lig $\overline{y} = y_\bullet/n$. Funktionen $\ln L$ har altså stationært punkt i $\mu = \overline{y}$, og da dens anden afledede

$$\frac{d^2}{d\mu^2} \ln L(\mu) = -\frac{y_\bullet}{\mu^2}$$

altid er negativ, er $\overline{y}$ et maksimumspunkt. Dermed er vist at maksimaliseringsestimatet for μ er gennemsnittet af observationerne:*

$$\widehat{\mu} = \overline{y} = \frac{1}{n} \sum_{j=1}^{n} y_j \ .$$

I taleksemplet får man

$$\sum_{j=1}^{200} y_j = 0 \cdot 109 + 1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot 1 = 122,$$

så at $\widehat{\mu} = 122/200 = 0.61$ og dermed $\widehat{\lambda} = \widehat{\mu}/1\text{år} = 0.61\text{år}^{-1}$, dvs. dødsintensiteten er 0.61 dødsfald pr. år for hvert regiment. – Det ses at $\widehat{\lambda}$ fremkommer som antal dødsfald divideret med antal regiment-år.

Modelkontrol

Når vi holder os inden for klassen af poissonfordelingsmodeller, får vi den bedste beskrivelse af hestespark-observationerne ved at bruge intensiteten $\widehat{\lambda} = 0.61$ dødsfald pr. år for hvert regiment.

For at få et fingerpeg om hvor god denne »bedste beskrivelse« er, udregner vi nogle »forventede« antal under forudsætning af at modellen er rigtig: Ifølge modellen er sandsynligheden for at der i et bestemt regiment-år er netop y dødsfald,

$$f(y; \widehat{\lambda}) = \frac{(\widehat{\lambda} \cdot 1\text{år})^y}{y!} \exp(-\widehat{\lambda} \cdot 1\text{år}).$$

Ud af de 200 regiment-år skulle man derfor forvente ca. $200 \cdot f(0; \widehat{\lambda})$ tilfælde med 0 dødsfald, ca. $200 \cdot f(1; \widehat{\lambda})$ tilfælde med 1 dødsfald, ca. $200 \cdot f(2; \widehat{\lambda})$ tilfælde med 2 dødsfald, osv. Disse forventede tal udregnes, og man får tabel 12.1. Det ses at de »forventede« antal stemmer fint overens med de observerede, og det må vi tage som tegn på at poissonmodellen ikke er helt hen i vejret.

Dispersionstestet

Undertiden vil man gerne udføre et numerisk test for rimeligheden af at antage at et sæt observationer $y_1, y_2, \ldots, y_n$ er en stikprøve fra en poissonfordeling. Vi vil omtale en nem metode hertil.

∗ Det er i udledningen forudsat at $y_\bullet > 0$. Hvis $y_\bullet = 0$, så er log-likelihoodfunktionen lig $-n\mu$+konstant, og den antager sit maksimum når $\mu = 0$.

Tabel 12.1 Hestespark-eksemplet: De observerede antal år med y dødsfald sammenlignet med de »forventede« antal år med y dødsfald beregnet ud fra poissonmodellen.

antal dødsfald y	observeret antal år	»forventet« antal år
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6
5+	0	0.1
	200	200.0

Som nævnt side 155 har poissonfordelingen den egenskab at middelværdi og varians er ens. Man kunne derfor udregne den empiriske middelværdi

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

og den empiriske varians

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

og så se efter om de to er nogenlunde ens. Det viser sig hensigtsmæssigt at gøre dette på den måde at man udregner størrelsen

$$d = \frac{s^2}{\overline{y}}.$$

Hvis modelantagelsen er rigtig, skal d være tæt på 1, så man vil forkaste modellen hvis *enten* d_{obs} er så meget større end 1 at der kun er lille sandsynlighed (for eksempel 0.025) for at få en større værdi, *eller* d_{obs} er så meget mindre end 1 at der kun er lille sandsynlighed (for eksempel 0.025) for at få en mindre værdi. – Man kan bevise at når modellen er rigtig (og poissonparameteren ikke er alt for lille), så vil d med god tilnærmelse følge en såkaldt χ^2/f-fordeling med $f = n - 1$ frihedsgrader.

I hestespark-eksemplet fandt vi tidligere at $\overline{y} = 0.61$. Videre er

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \overline{y})^2 &= 109 \cdot (0 - 0.61)^2 + 65 \cdot (1 - 0.61)^2 \\ &\quad + 22 \cdot (2 - 0.61)^2 + 3 \cdot (3 - 0.61)^2 + 1 \cdot (4 - 0.61)^2 \\ &= 121.58,\end{aligned}$$

så $s^2 = 121.58/199 = 0.611$ og dermed $d_{\text{obs}} = 0.611/0.61 = 1.002$. Den fundne d_{obs}-værdi ligger meget tæt på 1, også målt i forhold til χ^2/f-fordelingen med 199 frihedsgrader, og det numeriske test bekræfter dermed indtrykket af at poissonfordelingen giver en god beskrivelse af tallene.

Tabel 12.2 Resultater af mikrokernetællinger.

Behandlingsgruppen			*Kontrolgruppen*		
Mus nr.	Antal optalte celler r	Antal mikrokerne-celler y	Mus nr.	Antal optalte celler r	Antal mikrokerne-celler y
1	2096	1	1	2077	2
2	2138	10	2	2181	6
3	2086	7	3	2030	2
	6320	18		6288	10

12.2 Sammenligning af to poissonfordelinger

Vi vil diskutere spørgsmålet om sammenligning af to poissonfordelinger ud fra følgende eksempel.

Eksempel 12.1 (Ultralydsscanning)
Det er meget udbredt at foretage ultralydsscanning af gravide kvinder. Det menes/frygtes imidlertid at fostrene kan lide skade derved, idet der måske sker kromosomforandringer. For at undersøge dette nærmere har man udført en række laboratorieforsøg med mus (12).

Et antal drægtige mus udsættes for ultralydsbestråling i et vist stykke tid, hvorefter man undersøger leverceller fra fostrene for at se om der er dannet såkaldte mikrokerneceller. Mikrokerner i en celle opstår som følge af kromosomforandringer og/eller -ødelæggelser.

I dette eksempel (der kun behandler en del af forsøgets talmateriale) optræder to grupper à tre mus: en behandlingsgruppe og en kontrolgruppe. Behandlingsgruppen har fået ultralyd, hvorefter man har ladet gå 18 timer inden musen blev dræbt og prøverne udtaget. Kontrolgruppen er behandlet på samme måde, på nær at der denne gang ikke blev tændt for ultralydapparatet. Fra hver mus udtog man otte prøver; i alt undersøgte man for hver mus ca. 2000 celler og afgjorde om det var en mikrokernecelle eller ej. Derved fremkom resultaterne i tabel 12.2. Spørgsmålet er om disse tal tyder på at ultralyd har en skadelig virkning.

Modelopstilling

For hver mus er der øjensynlig *to* størrelser der er uforudsigelige, nemlig antal optalte celler r og antal mikrokerneceller y. Når vi skal formulere den statistiske model, skal vi tage stilling til om både r og y eller kun den ene af dem skal opfattes som observation af en stokastisk variabel. De størrelser der opfattes som udfald af stokastiske variable, er de størrelser for hvilke den statistiske model påtager sig at beskrive hvilke andre udfald man også kunne have fået.

I den foreliggende problemstilling er det der er genstand for den grundlæggende interesse, formentlig chancen for at en celle omdannes til en mikrokernecelle. I den forbindelse er det uinteressant at søge at opstille en model der kan påtage sig at beskrive variationen i antal optalte celler pr. mus. De indgående tider er de samme for alle forsøgsdyr; derfor behøver vi ikke indbygge tidsafhængigheder i modellen. Derimod skal vi søge at formulere en model der kan beskrive variationen i antallet af mikrokerneceller

i en prøve af en given størrelse, udtaget fra en mus der har fået en given behandling. I modellen skal r-erne derfor indgå som givne konstanter og y-erne som udfald af stokastiske variable.

Da der for en enkelt mus optælles et meget stort antal celler der hver især har en meget lille chance for at være blevet omdannet til en mikrokernecelle, kan vi antage (jf. trykfejlseksemplet side 156) at antal mikrokerneceller i en prøve med r celler er poissonfordelt med parameter $\mu = \lambda r$, hvor λ er en »omdannelsesintensitet«, nemlig sandsynligheden for at en optalt celle er en mikrokernecelle. Den systematiske forskel mellem behandlingsgrupperne skal beskrives ved hjælp af modellens parametre, så derfor skal mus med samme behandling have samme intensitet λ, hvorimod behandlingsgruppen og kontrolgruppen skal have hver sit λ.

Vi indfører lidt notation for at kunne formulere modellen præcist:

$$\begin{aligned} r_{ij} &= \text{ antal optalte celler fra mus nr. } j \text{ i gruppe } i, \\ y_{ij} &= \text{ antal mikrokerneceller fra mus nr. } j \text{ i gruppe } i, \end{aligned}$$

hvor $i = 1$ svarer til behandlingsgruppen og $i = 2$ til kontrolgruppen. Det vil sige at tabel 12.2 skematisk ser således ud:

	gruppe 1	
1	r_{11}	y_{11}
2	r_{12}	y_{12}
3	r_{13}	y_{13}
	$r_{1\bullet}$	$y_{1\bullet}$

	gruppe 2	
1	r_{21}	y_{21}
2	r_{22}	y_{22}
3	r_{23}	y_{23}
	$r_{2\bullet}$	$y_{2\bullet}$

Modellen er da at tallene $y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}$ skal opfattes som observerede værdier af stokastisk uafhængige poissonfordelte stokastiske variable $Y_{11}, Y_{12}, Y_{13}, Y_{21}, Y_{22}, Y_{23}$ hvor Y_{ij} har parameter $\mu_{ij} = \lambda_i r_{ij}$. Her er λ_1 og λ_2 ukendte parametre der beskriver den systematiske forskel mellem behandlingsgruppen og kontrolgruppen, og r_{ij}-erne er kendte konstanter. *Modelfunktionen* bliver

$$\prod_{i=1}^{2}\prod_{j=1}^{3} \frac{(\lambda_i r_{ij})^{y_{ij}}}{y_{ij}!} \exp(-\lambda_i r_{ij}) \ . \tag{12.1}$$

Det oprindelige spørgsmål om observationerne tyder på at ultralyd er skadeligt, kan nu oversættes til modellens sprog. Da den systematiske forskel mellem grupperne beskrives ved hjælp af parametrene λ_1 og λ_2, bliver det relevante spørgsmål om observationerne tyder på at λ_1 og λ_2 er signifikant forskellige; med andre ord skal vi teste den statistiske hypotese $H_0 : \lambda_1 = \lambda_2$.

Estimation af parametre

Maksimaliseringsestimaterne over λ_1 og λ_2 skal bestemmes på grundlag af *likelihood-funktionen*. Ud fra modelfunktionen (12.1) får vi

$$\begin{aligned} L(\lambda_1,\lambda_2) &= \prod_{i=1}^{2}\prod_{j=1}^{3}\frac{(\lambda_i r_{ij})^{y_{ij}}}{y_{ij}!}\exp(-\lambda_i r_{ij}) \\ &= \text{konstant}\cdot\prod_{i=1}^{2}\prod_{j=1}^{3}\lambda_i^{y_{ij}}\exp(-\lambda_i r_{ij}) \\ &= \text{konstant}\cdot\prod_{i=1}^{2}\lambda_i^{y_{i\bullet}}\exp(-\lambda_i r_{i\bullet}) \end{aligned}$$

hvor konstanten afhænger af r-erne og y-erne, men ikke af λ_1 og λ_2. Vi ser at likelihoodfunktionen er den samme som man ville have fået, hvis man udelukkende havde set på totalantallene $y_{1\bullet}$ og $y_{2\bullet}$ for hver mus og havde sagt at det var $Y_{1\bullet}$ og $Y_{2\bullet}$ der var poissonfordelte med parametre $\lambda_1 r_{1\bullet}$ hhv. $\lambda_2 r_{2\bullet}$. Derfor bliver estimatet over λ_i

$$\widehat{\lambda}_i = \frac{y_{i\bullet}}{r_{i\bullet}},$$

nemlig det totale antal observerede mikrokerneceller i gruppe i divideret med det totale antal optalte celler i gruppe i, hvilket også er et estimat der virker umiddelbart rimeligt.

For at estimere det fælles λ under H_0 betragtes likelihoodfunktionen $L_0(\lambda) = L(\lambda,\lambda)$:

$$L_0(\lambda) = \text{konstant}\cdot\prod_{i=1}^{2}\lambda^{y_{i\bullet}}\exp(-\lambda r_{i\bullet}) = \text{konstant}\cdot\lambda^{y_{\bullet\bullet}}\exp(-\lambda r_{\bullet\bullet})$$

som har maksimum i $\widehat{\lambda} = y_{\bullet\bullet}/r_{\bullet\bullet}$. Det er også hvad man umiddelbart skulle vente, thi når H_0 er rigtig, er der ingen forskel på de to grupper, dvs. der er i realiteten kun tale om én enkelt gruppe bestående af $r_{\bullet\bullet}$ celler hvoraf $y_{\bullet\bullet}$ er mikrokerneceller.

I eksemplet bliver estimaterne

$$\begin{array}{lrcl} \widehat{\lambda}_{\text{behandling}} & 18/6320 & = & 2.8\cdot 10^{-3} \text{ mikrokerneceller pr. 1000 celler} \\ \widehat{\lambda}_{\text{kontrol}} & 10/6288 & = & 1.6\cdot 10^{-3} \text{ mikrokerneceller pr. 1000 celler} \\ \widehat{\lambda}_{\text{fælles}} & 28/12608 & = & 2.2\cdot 10^{-3} \text{ mikrokerneceller pr. 1000 celler.} \end{array}$$

Man kan spørge hvor stor tiltro man nu kan have til disse tal. Det er ikke i statistikerens magt at udtale noget fornuftigt om diverse eksterne fejlkilder der eventuelt måtte have været i spil (det véd eksperimentator bedre). Statistikeren kan udtale sig om dén tilfældige variation der beskrives af den statistiske model, for eksempel konkretiseret til *middelfejlene på estimatorerne*. Lad os derfor bestemme middelfejlen (ɔ: standardafvigelsen) på $\widehat{\lambda}_i$ i det foreliggende eksempel: Da $\widehat{\lambda}_i = Y_{i\bullet}/r_{i\bullet}$, er den søgte størrelse $\sqrt{\operatorname{Var}\widehat{\lambda}_i} = \sqrt{\operatorname{Var}\left(\dfrac{Y_{i\bullet}}{r_{i\bullet}}\right)}$. Ifølge regnereglerne for varianser og kvadratrødder er

$$\sqrt{\operatorname{Var}\left(\frac{Y_{i\bullet}}{r_{i\bullet}}\right)} = \sqrt{\frac{\operatorname{Var}(Y_{i\bullet})}{r_{i\bullet}^2}} = \frac{\sqrt{\operatorname{Var}(Y_{i\bullet})}}{r_{i\bullet}}.$$

Da $Y_{i\bullet}$ er poissonfordelt med parameter $\lambda_i r_{i\bullet}$, er $\mathrm{Var}(Y_{i\bullet}) = \lambda_i r_{i\bullet}$ (se side 155), så middelfejlen på $\widehat{\lambda}_i$ er

$$\frac{\sqrt{\mathrm{Var}(Y_{i\bullet})}}{r_{i\bullet}} = \sqrt{\frac{\lambda_i}{r_{i\bullet}}}.$$

Da vi ikke kender λ_i men kun et estimat $\widehat{\lambda}_i = y_{i\bullet}/r_{i\bullet}$, kan vi kun udregne en talværdi for *den estimerede middelfejl* på λ_i, og den bliver

$$\sqrt{\frac{\widehat{\lambda}_i}{r_{i\bullet}}} = \sqrt{\frac{y_{i\bullet}/r_{i\bullet}}{r_{i\bullet}}} = \frac{\sqrt{y_{i\bullet}}}{r_{i\bullet}}.$$

Man finder de estimerede middelfejl på $\widehat{\lambda}_{\text{behandling}}$, $\widehat{\lambda}_{\text{kontrol}}$ og $\widehat{\lambda}_{\text{fælles}}$ til $0.67 \cdot 10^{-3}$, $0.50 \cdot 10^{-3}$ og $0.42 \cdot 10^{-3}$.

Som læseren nok vil have bemærket, benytter vi ved beregningen af de forskellige estimater slet ikke de individuelle værdier af r og y for de enkelte mus, vi benytter kun totalerne for hver gruppe. Er det da lige meget hvad værdierne for de enkelte mus er? Ja, det er det faktisk, *sålænge der ikke er tvivl om poissonmodellens brugbarhed*. Men hvis vi er på udkig efter indicier for (eller imod) anvendeligheden af poissonmodellen, så er det i høj grad påkrævet at kende de enkelte værdier. For den statistiske model skal jo beskrive enkeltobservationernes tilfældige variation omkring et bestemt niveau, og hvis man vil vurdere antagelsen om at den tilfældige variation kan beskrives ved netop en poissonfordeling, så skal man undersøge enkeltobservationernes faktiske variation og vurdere om den ligner den fittede poissonfordeling.

Hypoteseprøvning

Som nævnt skal vi teste den statistiske hypotese $H_0 : \lambda_1 = \lambda_2$. Det gøres på traditionel vis med et kvotienttest. Vi udregner kvotienttesteststørrelsen

$$\begin{aligned} Q &= \frac{L(\widehat{\lambda}, \widehat{\lambda})}{L(\widehat{\lambda}_1, \widehat{\lambda}_2)} = \frac{\widehat{\lambda}^{y_{1\bullet}} \widehat{\lambda}^{y_{2\bullet}} \cdot \exp(-\widehat{\lambda} r_{1\bullet} - \widehat{\lambda} r_{2\bullet})}{\widehat{\lambda}_1^{y_{1\bullet}} \widehat{\lambda}_2^{y_{2\bullet}} \cdot \exp(-\widehat{\lambda}_1 r_{1\bullet} - \widehat{\lambda}_2 r_{2\bullet})} \\ &= \left(\frac{\widehat{\lambda}}{\widehat{\lambda}_1}\right)^{y_{1\bullet}} \left(\frac{\widehat{\lambda}}{\widehat{\lambda}_2}\right)^{y_{2\bullet}} \frac{\exp(-y_{1\bullet} - y_{2\bullet})}{\exp(-y_{1\bullet} - y_{2\bullet})} = \left(\frac{\widehat{\lambda} r_{1\bullet}}{y_{1\bullet}}\right)^{y_{1\bullet}} \left(\frac{\widehat{\lambda} r_{2\bullet}}{y_{2\bullet}}\right)^{y_{2\bullet}} \\ &= \left(\frac{\widehat{y}_{1\bullet}}{y_{1\bullet}}\right)^{y_{1\bullet}} \left(\frac{\widehat{y}_{2\bullet}}{y_{2\bullet}}\right)^{y_{2\bullet}}, \end{aligned}$$

hvor $\widehat{y}_{i\bullet} = \widehat{\lambda} r_{i\bullet}$ er det »forventede« antal mikrokerneceller i gruppe i, forudsat at H_0 er rigtig. Derfor er

$$-2 \ln Q = 2\Big(y_{1\bullet} \ln \frac{y_{1\bullet}}{\widehat{y}_{1\bullet}} + y_{2\bullet} \ln \frac{y_{2\bullet}}{\widehat{y}_{2\bullet}}\Big).$$

Små værdier af Q, dvs. store værdier af $-2 \ln Q$, er *signifikante*, dvs. de er tegn på at hypotesen H_0 *ikke* er forenelig med de foreliggende data. For at vurdere om $-2 \ln Q_{\text{obs}}$ er signifikant stor, skal man bestemme testsandsynligheden

$$\varepsilon = \mathrm{P}_0\big(-2 \ln Q \geq -2 \ln Q_{\text{obs}}\big),$$

Tabel 12.3 Fordelingen af $n = 647$ kvinder efter antallet y af ulykkestilfælde i en fem ugers periode.

y	f_y = antal kvinder med y ulykker
0	447
1	132
2	42
3	21
4	3
5	2
6+	0
	647

altså sandsynligheden under H_0 for at få et observationssæt der er mindst lige så afvigende som det foreliggende. Ved beregningen af ε kan man udnytte at når H_0 er rigtig, så er $-2\ln Q$ med god tilnærmelse† χ^2-fordelt med $f = 2-1$ frihedsgrader (nemlig antal parametre i grundmodellen minus antal parametre under H_0), således at ε med god tilnærmelse kan udregnes som sandsynligheden for at få en værdi større end eller lig med $-2\ln Q_{\text{obs}}$ i χ^2-fordelingen med 1 frihedsgrad:

$$\varepsilon = \mathrm{P}\big(\chi^2_1 \geq -2\ln Q_{\text{obs}}\big).$$

I taleksemplet er $\widehat{y}_{1\bullet} = 14.0$ og $\widehat{y}_{2\bullet} = 14.0$, så

$$-2\ln Q = 2\Big(18\,\ln\frac{18}{14.0} + 10\,\ln\frac{10}{14.0}\Big) = 2.32\ .$$

I χ^2-fordelingen med 1 frihedsgrad er 80%-fraktilen 1.64 og 90%-fraktilen 2.71, så den fundne $-2\ln Q$-værdi svarer til et ε på mellem 10% og 20%. Man vil almindeligvis sige at en sådan ε-værdi ikke er lille nok til at man vil forkaste H_0. Vi kan dermed konkludere at de foreliggende tal *ikke* giver statistisk belæg for at mene at ultralyd er skadeligt. (På den anden side giver de næppe heller belæg for at mene at ultralyd *ikke* er skadeligt.)

12.3 Et sværere eksempel

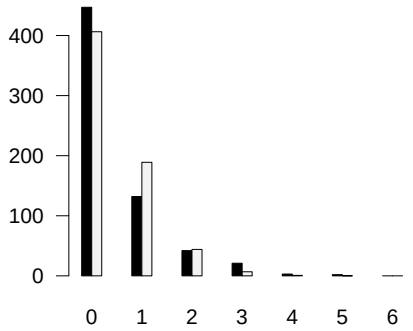
I dette afsnit gennemgås et eksempel hvor poissonfordelingen søges anvendt; det viser sig imidlertid at den model der foreslås i første omgang, ikke passer særlig godt; derfor må man finde på en anden model.

Præsentation af eksemplet

Man har undersøgt hvor mange ulykkestilfælde hver enkelt arbejder på en granatfabrik i England kom ud for i løbet af en fem ugers periode. Det hele foregik under første verdenskrig, så de pågældende arbejdere var kvinder (mens mændene var soldater).

† χ^2-approksimationen kan anvendes når de forventede antal $\widehat{y}_{i\bullet}$ er mindst fem.

y	f_y	$\widehat{f}_y$
0	447	406.3
1	132	189.0
2	42	44.0
3	21	6.8
4	3	0.8
5	2	0.1
6+	0	0.0
	647	647.0



Figur 12.1 Model 1: Observerede antal f_y (sorte søjler) og forventede antal $\widehat{f}_y$ (lyse søjler).

I tabel 12.3 ses fordelingen af $n = 647$ kvinder efter antallet y af ulykkestilfælde i en fem ugers periode. Man søger en statistisk model der kan beskrive dette talmateriale. (Eksemplet stammer fra (8), og er her i landet især kendt via sin forekomst i (9) der i mere end en menneskealder har været en toneangivende dansk lærebog i statistik.)

Lad y_i betegne antal ulykker som kvinde nr. i kommer ud for; y_i tænkes at være en observation af en stokastisk variabel Y_i, $i = 1, 2, \ldots, n$. Vi går ud fra at de stokastiske variable $Y_1, Y_2, \ldots, Y_n$ er indbyrdes uafhængige (men det er måske en lidt diskutabel antagelse). Vi benytter betegnelsen f_y for antallet af kvinder der har været ude for netop y ulykker, dvs. i det foreliggende tilfælde er $f_0 = 447$, $f_1 = 132$ osv. Det samlede antal ulykker er da lig $0f_0 + 1f + 2f_2 + \ldots = \sum_{y=0}^{\infty} y f_y = 301$.

Model 1

I første omgang kan man forsøge sig med en model hvor $Y_1, Y_2, \ldots, Y_n$ er uafhængige og identisk poissonfordelte med parameter μ, dvs.

$$\mathrm{P}(Y_i = y) = \frac{\mu^y}{y!} \exp(-\mu).$$

Poissonfordelingen kommer ind i billedet ud fra en forestilling om at ulykkerne sker »helt tilfældigt«, og man kan sige at parameteren μ beskriver kvindernes »ulykkestilbøjelighed«.

I denne model estimeres μ ved $\widehat{\mu} = \overline{y} = 301/647 = 0.465$ (ɔ: der sker 0.465 ulykker pr. kvinde pr. fem uger). Det forventede antal kvinder med y ulykker er $\widehat{f}_y = n \dfrac{\widehat{\mu}^y}{y!} \exp(-\widehat{\mu})$; værdierne heraf ses i figur 12.1. Det ses at der ikke er nogen særlig god overensstemmelse mellem de observerede og de forventede antal. Man kan udregne variansen til $s^2 = 0.692$, og det er næsten halvanden gange middelværdien, hvilket er endnu et tegn på at poissonmodellen er dårlig. Man kan derfor give sig til at overveje en anden model.

Model 2

Man kan udvide model 1 på følgende måde

- Det antages stadig at $Y_1, Y_2, \ldots, Y_n$ er uafhængige og poissonfordelte, men nu tillader vi at de har hver sin middelværdi, dvs. Y_i er poissonfordelt med parameter μ_i, $i = 1, 2, \ldots, n$. Hvis modelopstillingen gjorde holdt her, ville der være en parameter for hver person, og man ville få et perfekt fit (ɔ: $\widehat{\mu}_i = y_i$, $i = 1, 2, \ldots, n$). Men der endnu et trin i modelopbygningen:
- Det antages endvidere at $\mu_1, \mu_2, \ldots, \mu_n$ er uafhængige observationer fra en og samme sandsynlighedsfordeling. Denne sandsynlighedsfordeling skal være en kontinuert fordeling på den positive halvakse, og det viser sig bekvemt at benytte en fordeling med en tæthedsfunktion af formen

$$g(\mu) = \frac{1}{\Gamma(\kappa)\beta^\kappa}\mu^{\kappa-1}\exp(-\mu/\beta), \quad \mu > 0.$$

(Symbolet $\Gamma(\kappa)$ betegner den såkaldte Gammafunktion, udregnet i κ. Pr. definition er $\Gamma(\kappa) = \int_0^{+\infty} t^{\kappa-1}\exp(-t)\,dt$. Hvis m er et naturligt tal, så er $\Gamma(m+1) = m!$. Gammafunktionen kommer ind i billedet fordi tæthedsfunktionen g skal integrere til 1, og det gør den da også, hvilket ses ved at foretage substitutionen $t = \mu/\beta$.) Fordelingen med denne tæthedsfunktion g er en *gammafordeling* med formparameter $\kappa > 0$ og skalaparameter $\beta > 0$.
- Sandsynligheden for at en kvinde kommer ud for y ulykker ville nu være lig med $\dfrac{\mu^y}{y!}\exp(-\mu)$, hvis vi altså kendte værdien af μ for den pågældende kvinde. Men da vi kun véd at μ følger fordelingen med tæthedsfunktion g, bliver den faktiske sandsynlighed for y ulykker et vægtet middeltal af værdierne $\dfrac{\mu^y}{y!}\exp(-\mu)$ med $g(\mu)$-værdierne som vægte, og det betyder at sandsynligheden for at en kvinde kommer ud for netop y ulykker alt i alt bliver

$$\begin{aligned}
\mathrm{P}(Y = y) &= \int_0^{+\infty} \frac{\mu^y}{y!}\exp(-\mu)\cdot g(\mu)\,d\mu \\
&= \int_0^{+\infty} \frac{\mu^y}{y!}\exp(-\mu)\,\frac{1}{\Gamma(\kappa)\beta^\kappa}\,\mu^{\kappa-1}\exp(-\mu/\beta)\,d\mu \\
&= \frac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}\Big(\frac{1}{\beta+1}\Big)^\kappa\Big(\frac{\beta}{\beta+1}\Big)^y \\
&= \frac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}p^\kappa(1-p)^y,
\end{aligned}$$

hvor $p = 1/(\beta+1)$. Med betegnelsen $\binom{y+\kappa-1}{y} = \dfrac{\Gamma(y+\kappa)}{y!\,\Gamma(\kappa)}$ (som hvis κ er et naturligt tal, blot er den sædvanlige definition af binomialkoefficient) er sandsynligheden for y ulykker

$$\mathrm{P}(Y = y) = \binom{y+\kappa-1}{y}p^\kappa(1-p)^y, \qquad y = 0, 1, 2, \ldots.$$

Denne fordeling af Y er den såkaldte *negative binomialfordeling* med formparameter κ og sandsynlighedsparameter p. – κ kan være et vilkårligt positivt tal, og p et vilkårligt tal mellem 0 og 1 (fordi $p = 1/(1+\beta)$ hvor $\beta > 0$).

Den negative binomialfordeling har *to* parametre man kan »skrue på«, og man kan håbe at det derved er muligt at få denne model til at passe bedre til observationerne end Model 1 gjorde.

I den nye model kan middelværdien vises at være $\mathrm{E}(Y) = \kappa(1-p)/p$ som vi kalder μ, og variansen kan vises at være $\mathrm{Var}(Y) = \kappa(1-p)/p^2 = \mu/p = \mu + \mu^2/\kappa$; heraf ses at variansen altid er større end middelværdien. – I det foreliggende talmateriale fandt vi netop at variansen var større end middelværdien, så foreløbig kan det ikke udelukkes at den negative binomialfordelingsmodel er brugbar.

Undertiden bruger man en anden parametrisering af fordelingen: i stedet for κ og p bruger man κ og μ.

Estimation af parametrene i Model 2

Vi benytter som altid likelihoodmetoden til estimation af de ukendte parametre. Likelihoodfunktionen er

$$\begin{aligned} L(\kappa, p) &= \prod_{i=1}^{n} \binom{y_i + \kappa - 1}{y_i} p^{\kappa} (1-p)^{y_i} \\ &= p^{n\kappa} (1-p)^{y_1 + y_2 + \cdots + y_n} \prod_{i=1}^{n} \binom{y_i + \kappa - 1}{y_i} \\ &= \text{konstant} \cdot p^{n\kappa} (1-p)^{y_\bullet} \prod_{k=1}^{\infty} (\kappa + k - 1)^{\sum_{j=k}^{\infty} f_j}, \end{aligned}$$

hvor f_k stadig betegner antal observationer som har værdien k. Logaritmen til likelihoodfunktionen bliver derfor (pånær en konstant)

$$\ln L(\kappa, p) = n\kappa \ln p + y_\bullet \ln(1-p) + \sum_{k=1}^{\infty} \left(\sum_{j=k}^{\infty} f_j \right) \ln(\kappa + k - 1)$$

der i det konkrete eksempel antager det mere uskyldige udseende

$$\begin{aligned} \ln L(\kappa, p) = {} & 647\kappa \ln p + 301 \ln(1-p) \\ & + 200 \ln \kappa + 68 \ln(\kappa + 1) + 26 \ln(\kappa + 2) \\ & + 5 \ln(\kappa + 3) + 2 \ln(\kappa + 4). \end{aligned}$$

Vi overlader til computeren at bestemme maksimumspunktet for denne funktion. Man får at $\widehat{\kappa} = 0.865$ og $\widehat{\mu} = 0.465$ (så $\widehat{p} = \widehat{\kappa}/(\widehat{\mu} + \widehat{\kappa}) = 0.650$). De tilsvarende forventede antal

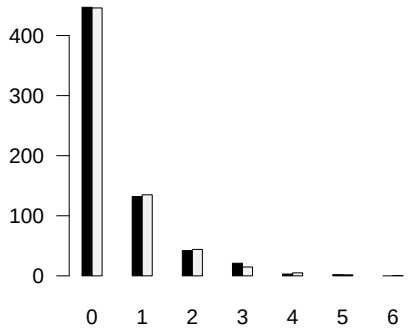
$$\widehat{\widehat{f}}_y = n \binom{y + \widehat{\kappa} - 1}{y} \widehat{p}^{\widehat{\kappa}} (1 - \widehat{p})^y$$

ses i figur 12.2. På baggrund heraf tillader vi os at konkludere at den negative binomialfordelingsmodel beskriver observationerne godt nok.

12.4 Regn og tegn

Her vises hvordan man kan udføre de forskellige beregninger og tegninger med R.

y	f_y	$\widehat{\widehat{f}}_y$
0	447	445.9
1	132	134.9
2	42	44.0
3	21	14.7
4	3	5.0
5	2	1.7
6+	0	0.9
	647	647.1



Figur 12.2 Model 2: Observerede antal f_y (sorte søjler) og forventede antal $\widehat{\widehat{f}}_y$ (hvide søjler).

Hestespark-eksemplet

Her er det lettest at »tilbageregne« de fem observerede antal (jf. tabel 12.1) til 200 y-værdier; det gøres med funktionen `rep`. Poissonsandsynlighederne udregnes med funktionen `dpois`

```
# Vi genskaber de oprindelige 200 observationer med "gentagelsesfunktionen" rep
y <- rep (  0:4, c (109, 65, 22, 3, 1))
y              # skriv de 200 værdier
mean(y)        # gennemsnittet
round (200 * dpois(0:4, mean(y)), digits=1) # de forventedede antal m. 1 decimal
var(y)         # den estimerede varians
d <- var(y)/mean(y) # dispersionsteststørrelsen
f <- 199 # antal frihedsgrader
pchisq (f*min(d,1/d), f) + 1 - pchisq (f*max(d,1/d), f) # testsandsynligheden
```

Ultralyd-eksemplet

Data (tabel 12.2) indlæses fra en datafil der i den viste udskrift hedder `h:/bog/txt304ny/ultralyd.dat`, og hvis indhold er

```
Gruppe     Optalte Antal
Behandling 2096     1
Behandling 2138    10
Behandling 2086     7
Kontrol    2077     2
Kontrol    2181     6
Kontrol    2030     2
```

Selve R-koden kommer her.

Funktionen `glm` fitter en såkaldt generaliseret lineær model; den udregner logaritmen til parametrene, så derfor må man anvende `exp` for at få de rigtige parametre.

```
# indlæs data til data.frame'n  Ulyd
Ulyd <- read.table("h:/bog/txt304ny/ultralyd.dat", nrows=10, header=TRUE)
# fit grundmodellen
G <- glm (Antal ~ 0 + Gruppe, family=poisson, data=Ulyd, offset=log(Optalte))
round (exp (G$coef), digits=4) # de to lambda.hat'er
```

```
# Hypotesen om ens parametre
H0 <- update (G, . ~ 1)
round (exp (H0$coef), digits=4) # det fælles lambda.hat

# test af hypotesen H0 i forhold til grundmodellen G
anova (H0, G) # giver -2lnQ (Deviance) på 2.2774 med 1 frihedsgrad
1 - pchisq (2.2774, 1) # testsandsynligheden
```

Granat-eksemplet

Den første model (jf. side 167) behandles på samme måde som hestesparkeksemplet. Til den negative binomialfordelingsmodel (jf. side 167f) benyttes en særlig `glm`-variant `glm.nb` der findes i biblioteket `MASS`.

```
f <- c (447, 132, 42, 21, 3, 2, 0) # frekvenser
y <- rep (0:6, f)
mean (y) # my.hat i poisson-modellen
var (y)
# de forventedede antal i poissonmodellen:
f.hat <- round (647 * dpois(0:6, mean(y)), digits=1)

# plot observerede og forventede i model 1
barplot (rbind (f, f.hat), beside=TRUE, las=1, names.arg=0:6, space=c(0,3),
  col= gray ( c(0, 0.95)))

# negativ biomial-model
require (MASS)
NB <- glm.nb (y ~ 1, link=identity)
mu.hat <- NB$coef
kappa.hat <- NB$theta                # fordi glm.nb kalder  kappa  for  theta
p.hat <- kappa.hat/(kappa.hat+mu.hat)
mu.hat ; kappa.hat ; p.hat

# de forventedede antal:
f.hathat <- round (647 * dnbinom (0:6, mu = mu.hat, size=kappa.hat), digits=1)
f.hathat

# plot observerede og forventede i model 2
barplot (rbind (f, f.hathat), beside=TRUE, las=1, names.arg=0:6, space=c(0,3),
col= gray ( c(0, 0.95)))
```

Vedr. opgave 12.2

Man kan fremstille indholdet af en tabel svarende til tabel 12.6 sådan her (`rpois` laver tilfældige poissonfordelte tal):

```
y <- matrix (rpois(200, 3.14), nrow=20)
y # de 20 rækker med hver 10 tilfældige poissonfordelte tal
apply (y, 1, sum)    # de 20 rækkesummer
apply (y, 1, mean)   # de 20 rækkegennemsnit
apply (y, 1, var)    # de 20 rækkevarianser
apply (y, 1, (function(x){ var(x) / mean(x) }))   # d^2 værdierne
```

Tabel 12.4 Opgave 12.1: Antal tidsintervaller f_y hvor der udsendes netop y α-partikler.

y	f_y	y	f_y
0	57	8	45
1	203	9	27
2	383	10	10
3	525	11	4
4	532	12	0
5	408	13	1
6	273	14	1
7	139		

Tabel 12.5 Opgave 12.4: Fordelingen af drenge fra to vandværksdistrikter efter antal DMF-tænder.

y	antal med y DMF-tænder	
	gamle vv.	hjælpe-vv.
0	1	
1	1	
2	8	6
3	3	5
4	13	6
5	7	3
6	8	7
7	2	4
8		3
9	2	4
10		1
11		
12		
13		1

12.5 Opgaver

Opgave 12.1 (Udsendelse af α-partikler)
I et berømt eksperiment har Rutherford og Geiger talt op hvor mange α-partikler der udsendes fra en bestemt portion af det radioaktive stof Polonium i et tidsinterval af længde 7.5 sekund; man har foretaget optællingen for i alt 2608 sådanne tidsintervaller. Resultaterne fremgår af tabel 12.4 (fra (16), genoptrykt med mindre rettelser i (17)).

Det formodes at antal α-partikler udsendt i et tidsinterval af længde t (som er meget mindre end stoffets halveringstid) kan opfattes som en observation af en poissonfordelt stokastisk variabel med parameter $\lambda \cdot t$, hvor λ er en slags strålingsintensitet.

1. Gør rede for rimeligheden af poissonfordelingsantagelsen, og præcisér den statistiske model.
2. Estimér λ ud fra de givne observationer.
3. Hvad kan dispersionstestet fortælle om rimeligheden af den foreslåede model?

Opgave 12.2
Tabel 12.6 indeholder 20 stikprøver $y_1, y_2, \ldots, y_{10}$ fra en poissonfordeling med $\mu = 3.14$.

1. Udregn $\widehat{\mu}$ for hver stikprøve. Hvordan fordeler $\widehat{\mu}$ sig omkring μ?
2. Udregn dispersionsteststørrelsen d for hver stikprøve. Hvordan ligger værdierne i forhold til χ^2/f-fordelingen?
3. Man kan bevise at en sum af uafhængige poissonfordelte størrelser er poissonfordelt med en parameter der er lig summen af parametrene. Derfor kan man opfatte de 20 værdier i $y_{\bullet}$-søjlen som 20 observationer fra en poissonfordeling med parameter 10μ ($= 31.4$).
   Udregn parameterestimatet og dispersionsteststørrelsen for disse 20 observationer.

Tabel 12.6 20 eksempler på udfald af stokastiske variable $Y_1, Y_2, \ldots, Y_{10}$ frembragt af en poissonfordelings-tilfældighedsmekanisme med $\mu = 3.14$.

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	$y_\bullet$	$\overline{y}$	s^2
2	4	2	2	0	5	2	3	4	2	26	2.60	2.04
3	3	3	0	4	3	3	3	3	5	30	3.00	1.56
4	2	5	1	0	6	5	6	1	2	32	3.20	5.07
3	1	4	3	3	2	0	2	3	3	24	2.40	1.38
3	4	5	4	4	2	3	6	2	1	34	3.40	2.27
2	5	6	6	5	3	4	2	4	2	39	3.90	2.54
2	2	6	2	4	2	4	1	1	6	30	3.00	3.56
4	1	2	0	2	3	7	4	4	2	29	2.90	3.88
1	3	4	2	1	2	2	2	3	3	23	2.30	0.90
3	3	2	2	5	4	2	3	6	4	34	3.40	1.82
6	5	5	3	3	2	3	3	3	2	35	3.50	1.83
3	4	1	4	3	4	4	3	3	4	33	3.30	0.90
1	2	3	2	1	2	1	2	5	5	24	2.40	2.27
4	2	3	1	5	8	5	5	2	1	36	3.60	4.93
3	4	3	4	3	1	5	2	3	5	33	3.30	1.57
2	2	2	6	4	5	3	2	2	0	28	2.80	3.07
3	4	3	2	3	2	3	2	0	1	23	2.30	1.34
2	0	1	2	2	5	6	4	2	2	26	2.60	3.38
1	2	4	2	3	3	4	0	3	4	26	2.60	1.82
6	0	7	3	0	6	3	4	3	4	36	3.60	5.60

Opgave 12.3
Fortsættelse af opgave 11.2.

1. Antag at der er udført n cirklinger, og at i netop y tilfælde fandtes der et skud inde i cirklen. Hvordan skal man på denne baggrund estimere λ?
2. Hvis man skal kunne opdage sjældne plantearter med denne metode, skal man nok bruge mere end 10 cirklinger. Antag at man stadig bruger cirkler med areal $a = 0.1\,\mathrm{m}^2$. Hvis en art vokser med en tæthed på ca. en pr. $5\,\mathrm{m}^2$ (dvs. $\lambda = 0.2\,\mathrm{m}^{-2}$), hvor mange cirklinger skal man da foretage for at være 90% sikker på at opdage planten?

   Tip: Opskriv først sandsynligheden for at man i n cirklinger *ikke* opdager planten.

Opgave 12.4 (Fluor i drikkevandet)
Det menes at fluor i drikkevandet kan modvirke huller i tænderne. I 1960-erne foretog man en undersøgelse af børns »tandstatus« og sammenholdt den med koncentrationen af fluor-ioner i drikkevandet fra det lokale vandværk. Tabel 12.5 viser data fra to vandværksdistrikter i Næstved. Man har bestemt antal DMF-tænder, dvs. tænder med huller efter caries samt udtrukne og plomberede tænder, hos de 12-årige drenge i de to distrikter. (Det kan i øvrigt nævnes at F^--koncentrationen ved det gamle vandværk var 1.9 ppm og ved hjælpevandværket 1.2 ppm.) – Undersøg ved hjælp af en poissonfordelingsmodel om der er en signifikant forskel på forekomsten af DMF-tænder i de to vandværksdistrikter.

13 Multiplikative poissonmodeller

I DETTE KAPITEL gennemgås et eksempel på en såkaldt multiplikativ poissonmodel. Modellen er ganske vist en smule mere indviklet end hvad der hidtil er blevet præsenteret, men på den anden side er det en type modeller der benyttes en del. Derudover er eksemplet interessant på den måde at man tilsyneladende kan nå frem til modstridende konklusioner blot ved at ændre en smule på fremgangsmåden ved analysen af modellen.

13.1 Det gennemgående eksempel: Lungekræft i Fredericia

I midten af 1970-erne var der en større debat om hvorvidt der var særlig stor risiko for at få lungekræft når man boede i byen Fredericia. Grunden til at der kunne være en større risiko, var at der i Fredericia var en betydelig mængde luftforurenende industri som tilmed lå midt inde i byen. For at kunne afgøre spørgsmålet indsamlede man data om lungekræfthyppigheden i perioden 1968-71, dels i Fredericia, dels i byerne Horsens, Kolding og Vejle. De tre sidste byer skulle tjene som sammenligningsgrundlag, idet det var byer af nogenlunde samme art som Fredericia, pånær den mistænkte industri.

Lungekræft opstår tit som et resultat af daglige påvirkninger af skadelige stoffer gennem mange år. En eventuel større risiko i Fredericia kunne måske derfor vise sig ved at lungekræftpatienterne fra Fredericia var yngre end dem fra kontrolbyerne, og det er under alle omstændigheder tilfældet at lungekræft optræder med meget forskellig hyppighed i forskellige aldersklasser. Det er derfor ikke nok at se på totalantallene af lungekræfttilfælde, man skal se på antallene af tilfælde i forskellige aldersklasser. De foreliggende tal er vist i tabel 13.1 (fra (2)). Da antallene af lungekræfttilfælde i sig selv ikke siger noget sålænge man ikke kender risikogruppernes størrelse, må man også rapportere antal indbyggere i de forskellige aldersklasser og byer, se tabel 13.2 (fra (2)).

Det der nu er statistikerens opgave, er at beskrive tallene i tabel 13.1 ved hjælp af en statistisk model hvori der indgår nogle parametre der i en passende forstand beskriver risikoen for at få lungekræft når man tilhører en bestemt aldersgruppe og bor i en bestemt by. Endvidere ville det være formålstjenligt hvis man kunne udskille nogle parametre der beskrev »byvirkninger« (dvs. forskelle mellem byer) efter at man på en eller anden måde havde taget højde for forskellene mellem aldersgrupperne.

13.2 Modelopstilling

Den statistiske model skal ikke modellere variationen i antallet af indbyggere i de forskellige byer og aldersklasser, så derfor vil vi anse disse antal for givne konstanter. Det

Tabel 13.1 Lungekræfttilfælde i fire byer fordelt på aldersklasser.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	11	13	4	5	33
55-59	11	6	8	7	32
60-64	11	15	7	10	43
65-69	10	10	11	14	45
70-74	11	12	9	8	40
75+	10	2	12	7	31
i alt	64	58	51	51	224

er antallene af lungekræfttilfælde der skal opfattes som observerede værdier af stokastiske variable, og det er fordelingen af disse stokastiske variabel der skal specificeres af den statistiske model. Vi indfører noget notation:

$$y_{ij} = \text{antal tilfælde i aldersgruppe } i \text{ i by } j,$$
$$r_{ij} = \text{antal personer i aldersgruppe } i \text{ i by } j,$$

hvor $i = 1, 2, 3, 4, 5, 6$ nummererer aldersgrupperne, og $j = 1, 2, 3, 4$ nummererer byerne. Observationerne y_{ij} opfattes som observerede værdier af stokastiske variable Y_{ij}.

Inspireret af kapitel 11 kunne man foreslå at Y_{ij} skulle være poissonfordelt med en parameter μ_{ij} der afhænger af aldersgruppe og by (modellen skal ikke indeholde observationsperiodens længde da denne er konstant lig 4 år). Hvis vi skriver μ_{ij} som $\mu_{ij} = \lambda_{ij} \cdot r_{ij}$, så kan intensiteten λ_{ij} fortolkes som antal lungekræfttilfælde pr. person i aldersgruppe i i by j i den betragtede fireårsperiode, dvs. λ er den *alders- og byspecifikke cancer-incidens*. Endvidere vil vi gå ud fra at de enkelte Y_{ij}-er er stokastisk uafhængige. *Grundmodellen* er altså at

> de stokastiske variable Y_{ij} er stokastisk uafhængige og poissonfordelte således at Y_{ij} har parameter $\lambda_{ij} r_{ij}$ hvor λ_{ij}-erne er ukendte positive parametre.

Det er let nok at estimere parametrene i grundmodellen. Eksempelvis estimeres intensiteten λ_{21} for 55-59–årige i Fredericia til $11/800 = 0.014$ (dvs. 0.014 tilfælde pr. person pr. 4 år). Den generelle opskrift er $\widehat{\lambda}_{ij} = y_{ij}/r_{ij}$.

Nu var det jo tanken at vi gerne ville kunne komme til at sammenligne byerne efter at vi havde taget højde for deres forskellige aldersfordelinger, og det kan ikke uden videre lade sig gøre i grundmodellen. Derfor vil vi undersøge om det lader sig gøre at beskrive data med en anden model hvor λ_{ij} er spaltet op i et produkt $\alpha_i \, \beta_j$ af en *aldersvirkning* α_i og en *byvirkning* β_j. Hvis dette lader sig gøre, er vi heldigt stillede, for så kan vi sammenligne byerne ved at sammenligne byparametrene β_j.

Vi vil derfor i første omgang teste den statistiske hypotese

$$H_0 : \lambda_{ij} = \alpha_i \, \beta_j$$

hvor α_1, α_2, α_3, α_4, α_5, α_6, β_1, β_2, β_3, β_4 er ukendte parametre. (Mere udførligt lyder hypotesen: Der findes parametre α_1, α_2, α_3, α_4, α_5, α_6, β_1, β_2, β_3, β_4 således at der for

Tabel 13.2 Antal indbyggere i de forskellige aldersklasser i de fire byer.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	3059	2879	3142	2520	11600
55-59	800	1083	1050	878	3811
60-64	710	923	895	839	3367
65-69	581	834	702	631	2748
70-74	509	634	535	539	2217
75+	605	782	659	619	2665
i alt	6264	7135	6983	6026	26408

by j og aldersgruppe i gælder at lungekræftrisikoen λ_{ij} fås som $\lambda_{ij} = \alpha_i\,\beta_j$.) – Hypotesen H_0 specificerer en såkaldt *multiplikativ* model fordi aldersparametre og byparametre indgår multiplikativt.

En detalje vedrørende parametriseringen

Der er det særlige ved parametriseringen af modellen under H_0 at den ikke er injektiv. At en parametrisering er *injektiv* betyder at forskellige parametersæt giver forskellige udgaver af modellen.

De 10 parametre $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4$ indgår udelukkende i modellen via produkterne $\alpha_i\beta_j$ $(= \lambda_{ij})$. Antag nu at to parametersæt

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4)$$

og

$$(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)$$

giver anledning til de samme produkter, dvs. antag at

$$\alpha_i\beta_j = \alpha_i^*\beta_j^* \tag{13.1}$$

for alle i og j. Så gælder også

$$\alpha_i/\alpha_i^* = \beta_j^*/\beta_j \tag{13.2}$$

for alle i og j. Da højresiden af formel (13.2) ikke involverer i, så kan venstresiden heller ikke afhænge af i, det vil sige der findes en konstant c således at $\alpha_i/\alpha_i^* = c$ og dermed $\alpha_i^* = \alpha_i/c$ for alle i. Videre er $\beta_j^*/\beta_j = \alpha_i/\alpha_i^* = c$, det vil sige $\beta_j^* = c\beta_j$ for alle j. Parametersættet $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)$ må altså nødvendigvis være af formen

$$\begin{aligned}(\alpha_1^*,&\alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_5^*, \alpha_6^*, \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)\\ &= (\frac{\alpha_1}{c}, \frac{\alpha_2}{c}, \frac{\alpha_3}{c}, \frac{\alpha_4}{c}, \frac{\alpha_5}{c}, \frac{\alpha_6}{c}, c\beta_1, c\beta_2, c\beta_3, c\beta_4)\end{aligned} \tag{13.3}$$

hvor c er en positiv konstant. Omvendt gælder også at hvis det stjernede parametersæt er defineret ved formel (13.3), så vil formel (13.1) være opfyldt. Hermed har vi fået

klarlagt dels at parametriseringen ikke er injektiv, dels hvilke parametersæt der giver den samme model.

De 10 parametre skal pålægges ét bånd for at få en injektiv parametrisering. Et sådant bånd kan være at $\alpha_1 = 1$, eller at $\alpha_1 + \alpha_2 + \cdots + \alpha_6 = 1$, eller at $\alpha_1\alpha_2 \ldots \alpha_6 = 1$, eller det tilsvarende for β, osv.

I det aktuelle eksempel vil vi benytte betingelsen $\beta_1 = 1$, dvs. vi definerer at parameteren for Fredericia skal være lig 1. Med denne betingelse er parametriseringen injektiv, for hvis både β_1 og $\beta_1^* = c\beta_1$ skal være 1, så må c nødvendigvis være lig 1. Samtidig noterer vi at der er $10 - 1 = 9$ forskellige parametre at estimere.

13.3 Den multiplikative model

I den multiplikative model lader det sig ikke gøre at opskrive simple udtryk for estimaterne, man er henvist til at benytte numeriske metoder for at bestemme talværdierne i de konkrete tilfælde. En computer med noget ordentligt statistikprogrammel vil uden videre kunne levere de ønskede værdier.

Parametersættet $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_1, \beta_2, \beta_3, \beta_4)$ (hvor $\beta_1 = 1$) skal ifølge de sædvanlige principper bestemmes så det maksimaliserer likelihoodfunktionen. I grundmodellen er likelihoodfunktionen

$$\begin{aligned} L &= \prod_{i=1}^{6}\prod_{j=1}^{4} \frac{(\lambda_{ij} r_{ij})^{y_{ij}}}{y_{ij}!} \exp(-\lambda_{ij} r_{ij}) \\ &= \text{konstant} \cdot \prod_{i=1}^{6}\prod_{j=1}^{4} \lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij} r_{ij}). \end{aligned}$$

Når vi her erstatter λ_{ij} med $\alpha_i\beta_j$, får vi likelihoodfunktionen under H_0:

$$\begin{aligned} L_0 &= \text{konstant} \cdot \prod_{i=1}^{6}\prod_{j=1}^{4} \alpha_i^{y_{ij}} \beta_j^{y_{ij}} \exp(-\alpha_i\beta_j r_{ij}) \\ &= \text{konstant} \cdot \Big(\prod_{i=1}^{6} \alpha_i^{y_{i\bullet}}\Big)\Big(\prod_{j=1}^{4} \beta_j^{y_{\bullet j}}\Big) \exp\Big(-\sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j r_{ij}\Big). \end{aligned}$$

Den tilsvarende *log-likelihoodfunktion* $\ln L_0$ er

$$\ln L_0 = \text{konstant} + \Big(\sum_{i=1}^{6} y_{i\bullet} \ln \alpha_i\Big) + \Big(\sum_{j=1}^{4} y_{\bullet j} \ln \beta_j\Big) - \sum_{i=1}^{6}\sum_{j=1}^{4} \alpha_i\beta_j r_{ij}.$$

Opgaven er nu at bestemme det parametersæt $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)$ der maksimaliserer L_0. Denne opgave lader sig ikke løse sådan lige uden videre; i praksis vil statistikeren benytte sig af noget programmel til analyse af *generaliserede lineære modeller*, idet den multiplikative poissonmodel er et specialtilfælde heraf.

Med programmet R får man (jf. afsnit 13.8)

$$\begin{aligned}
\widehat{\alpha}_1 &= 0.004 & \beta_1 &= 1\\
\widehat{\alpha}_2 &= 0.011 & \widehat{\beta}_2 &= 0.719\\
\widehat{\alpha}_3 &= 0.016 & \widehat{\beta}_3 &= 0.690\\
\widehat{\alpha}_4 &= 0.021 & \widehat{\beta}_4 &= 0.762\\
\widehat{\alpha}_5 &= 0.023 & &\\
\widehat{\alpha}_6 &= 0.015. & &
\end{aligned}$$

Efter at have bestemt de bedste estimater over α-erne og β-erne skal vi beskæftige os med hvor god en beskrivelse de faktisk giver af datamaterialet.

Formelt består opgaven i at *teste* multiplikativitetshypotesen H_0, og dette gøres som sædvanlig med et kvotienttest: Man udregner $-2\ln Q$ hvor

$$Q = \frac{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}{L(\widehat{\lambda}_{11}, \widehat{\lambda}_{12}, \ldots, \widehat{\lambda}_{63}, \widehat{\lambda}_{64})}.$$

Små værdier af Q eller store værdier af $-2\ln Q$ er signifikante, dvs. de tyder på at H_0 *ikke* giver en tilstrækkelig god beskrivelse af data. For at afgøre om $-2\ln Q_{\text{obs}}$ er signifikant stor, skal vi se på testsandsynligheden ε, altså sandsynligheden for at få en værre $-2\ln Q$-værdi forudsat at H_0 er rigtig:

$$\varepsilon = \mathrm{P}_0\big(-2\ln Q \geq -2\ln Q_{\text{obs}}\big).$$

Når H_0 er rigtig, er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $f = 24 - 9 = 15$ frihedsgrader (forudsat at de forventede antal alle er mindst fem). Det betyder at testsandsynligheden kan bestemmes som $\varepsilon = \mathrm{P}\big(\chi^2_{15} \geq -2\ln Q_{\text{obs}}\big)$. Efter en del omskrivninger finder man at

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}\ln\frac{y_{ij}}{\widehat{y}_{ij}}.$$

Man kan udregne de forventede antal lungekræfttilfælde $\widehat{y}_{ij} = \widehat{\alpha}_i\widehat{\beta}_j r_{ij}$ i hver enkelt by og aldersklasse (tabel 13.3). Det kan måske også være interessant at finde de estimerede alders- og by-specifikke lungekræftintensiteter $\widehat{\alpha}_i\widehat{\beta}_j$ (tabel 13.4).

Indsættes tallene fra tabel 13.1 og tabel 13.3 i udtrykket for $-2\ln Q$, får man $-2\ln Q_{\text{obs}} = 23.45$. I χ^2-fordelingen med $f = 24-9 = 15$ frihedsgrader er 90%-fraktilen 22.3 og 95%-fraktilen 25.0. Den opnåede værdi $-2\ln Q_{\text{obs}} = 23.45$ svarer altså til en testsandsynlighed ε på mellem 5% og 10%, og der er dermed ikke alvorlig evidens imod modellens brugbarhed. Vi tillader os at gå ud fra at modellen faktisk *er* anvendelig, dvs. at *lungekræftrisikoen afhænger multiplikativt af by og alder.*

Hermed er vi nået frem til en statistisk model der beskriver data ved hjælp af nogle by-parametre og nogle alders-parametre, men uden parametre svarende til en vekselvirkning mellem by og alder. Det betyder at den forskel der er mellem byerne,

Tabel 13.3 De forventede antal $\widehat{y}_{ij}$ af lungekræfttilfælde under den multiplikative poissonmodel.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	11.01	7.45	7.80	6.91	33.17
55-59	8.64	8.41	7.82	7.23	32.10
60-64	11.64	10.88	10.13	10.48	43.13
65-69	12.20	12.59	10.17	10.10	45.06
70-74	11.66	10.44	8.45	9.41	39.96
75+	8.95	8.32	6.73	6.98	30.98
i alt	64.10	58.09	51.10	51.11	224.40

Tabel 13.4 Estimerede alders- og byspecifikke lungekræftintensiteter i perioden 1986-71 under den multiplikative poissonmodel. Værdierne er antal pr. 1000 indbyggere pr. 4 år.

aldersklasse	Fredericia	Horsens	Kolding	Vejle
40-54	3.6	2.6	2.5	2.7
55-59	10.8	7.8	7.5	8.2
60-64	16.4	11.8	11.3	12.5
65-69	21.0	15.1	14.5	16.0
70-74	22.9	16.5	15.8	17.4
75+	14.8	10.6	10.2	11.3

er den samme for alle aldersklasser, og at den forskel der er mellem aldersklasserne, er den samme i alle byer. Når vi skal sammenligne byerne kan vi derfor gøre det ved udelukkende at betragte β-erne.

13.4 Ens byer?

Det hele går ud på at vurdere om der er nogen signifikant forskel på byerne. Hvis der ikke er nogen forskel, så må byparametrene være ens, dvs. $\beta_1 = \beta_2 = \beta_3 = \beta_4$, og da $\beta_1 = 1$, må den fælles værdi være 1. Derfor skall vi teste den statistiske hypotese

$$H_1 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 1.$$

Hypotesen skal testes i forhold til den aktuelle grundmodel H_0, så teststørrelsen bliver

$$Q = \frac{L_1(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \widehat{\widehat{\alpha}}_3, \widehat{\widehat{\alpha}}_4, \widehat{\widehat{\alpha}}_5, \widehat{\widehat{\alpha}}_6)}{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4, \widehat{\alpha}_5, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}$$

hvor

$$L_1(\alpha_1, \alpha_2, \ldots, \alpha_6) = L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, 1, 1, 1)$$

er likelihoodfunktionen under H_1, og $\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6$ er estimaterne over $\alpha_1, \alpha_2, \ldots, \alpha_6$ under H_1, dvs. $\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6$ maksimaliserer L_1.

Likelihoodfunktionen L_1 kan omskrives til et produkt af seks funktioner, hver med sit α:

$$L_1(\alpha_1,\alpha_2,\ldots,\alpha_6) = \text{konstant}\cdot\prod_{i=1}^{6}\prod_{j=1}^{4}\alpha_i^{y_{ij}}\exp(-\alpha_i r_{ij})$$
$$= \text{konstant}\cdot\prod_{i=1}^{6}\alpha_i^{y_{i\bullet}}\exp(-\alpha_i r_{i\bullet})\ .$$

Maksimaliseringsestimatet findes derfor til $\widehat{\widehat{\alpha}}_i = \dfrac{y_{i\bullet}}{r_{i\bullet}}$. Talværdierne bliver

$$\begin{array}{lcccl}
\widehat{\widehat{\alpha}}_1 & = & 33/11600 & = & 0.0028\\
\widehat{\widehat{\alpha}}_2 & = & 32/3811 & = & 0.0084\\
\widehat{\widehat{\alpha}}_3 & = & 43/3367 & = & 0.0128\\
\widehat{\widehat{\alpha}}_4 & = & 45/2748 & = & 0.0164\\
\widehat{\widehat{\alpha}}_5 & = & 40/2217 & = & 0.0180\\
\widehat{\widehat{\alpha}}_6 & = & 31/2665 & = & 0.0116.
\end{array}$$

Kvotienttesstørrelsen er

$$Q = \frac{L_1(\widehat{\widehat{\alpha}}_1,\widehat{\widehat{\alpha}}_2,\widehat{\widehat{\alpha}}_3,\widehat{\widehat{\alpha}}_4,\widehat{\widehat{\alpha}}_5,\widehat{\widehat{\alpha}}_6)}{L_0(\widehat{\alpha}_1,\widehat{\alpha}_2,\widehat{\alpha}_3,\widehat{\alpha}_4,\widehat{\alpha}_5,\widehat{\alpha}_6,1,\widehat{\beta}_2,\widehat{\beta}_3,\widehat{\beta}_4)}$$
$$= \frac{\displaystyle\prod_{i=1}^{6}\prod_{j=1}^{4}\widehat{\widehat{\alpha}}_i^{y_{ij}}\exp(-\widehat{\widehat{\alpha}}_i r_{ij})}{\displaystyle\prod_{i=1}^{6}\prod_{j=1}^{4}(\widehat{\alpha}_i\widehat{\beta}_j)^{y_{ij}}\exp(-\widehat{\alpha}_i\widehat{\beta}_j r_{ij})}$$
$$= \prod_{i=1}^{6}\prod_{j=1}^{4}\Big(\frac{\widehat{\widehat{y}}_{ij}}{\widehat{y}_{ij}}\Big)^{y_{ij}}\cdot\exp\Big(-\sum_{i=1}^{6}\sum_{j=1}^{4}\widehat{\widehat{y}}_{ij}+\sum_{i=1}^{6}\sum_{j=1}^{4}\widehat{y}_{ij}\Big)$$

hvor $\widehat{y}_{ij} = \widehat{\alpha}_i\widehat{\beta}_j r_{ij}$ (som hidtil), og $\widehat{\widehat{y}}_{ij} = \widehat{\widehat{\alpha}}_i r_{ij}$.

Da $\displaystyle\sum_{i=1}^{6}\sum_{j=1}^{4}\widehat{\widehat{y}}_{ij} = \sum_{i=1}^{6}\sum_{j=1}^{4}\widehat{y}_{ij} = \sum_{i=1}^{6}\sum_{j=1}^{4}y_{ij}$, er $\displaystyle Q = \prod_{i=1}^{6}\prod_{j=1}^{4}\Big(\frac{\widehat{\widehat{y}}_{ij}}{\widehat{y}_{ij}}\Big)^{y_{ij}}$ og dermed

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4}y_{ij}\ln\frac{\widehat{y}_{ij}}{\widehat{\widehat{y}}_{ij}}\ .$$

Store værdier af $-2\ln Q$ er signifikante. Man skal sammenholde $-2\ln Q$ med χ^2-fordelingen med $f = 9-6 = 3$ frihedsgrader.

De forventede tal er vist i tabel 13.5. Indsættes værdierne fra tabel 13.1, tabel 13.3 og tabel 13.5 i udtrykket for $-2\ln Q$, fås $-2\ln Q_{\text{obs}} = 4.86$. I χ^2-fordelingen med $f = 9-6 = 3$ frihedsgrader er 80%-fraktilen 4.64 og 90%-fraktilen 6.25, således at testsandsynligheden ε er næsten 20%. De foreliggende observationer er altså fint forenelige med hypotesen H_1 om at der ikke er nogen forskel på byerne. Sagt på en anden måde, *der er ikke nogen signifikant forskel på byerne.*

Tabel 13.5 De forventede antal $\widehat{\widehat{y}}_{ij}$ af lungekræfttilfælde under antagelsen om at der ikke er forskel på byerne.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	8.70	8.19	8.94	7.17	33.00
55-59	6.72	9.10	8.82	7.38	32.02
60-64	9.09	11.81	11.46	10.74	43.10
65-69	9.53	13.68	11.51	10.35	45.07
70-74	9.16	11.41	9.63	9.70	39.90
75+	7.02	9.07	7.64	7.18	30.91
i alt	50.22	63.26	58.00	52.52	224.00

13.5 En anden mulighed

Det er sjældent tilfældet at der er én bestemt måde at undersøge en praktisk problemstilling på ved hjælp af en statistisk model og en statistisk hypotese. Det aktuelle spørgsmål om der er en øget risiko for lungekræft ved at bo i Fredericia, blev i forrige afsnit belyst ved at vi testede hypotesen H_1 om ens byparametre. Det viste sig at H_1 kunne accepteres, og man kan således sige at der ikke er nogen signifikant forskel på de fire byer.

Nu kan man imidlertid angribe problemet på en anden måde. Man kan sige at det hele drejer sig om at vurdere om det er farligere at bo i Fredericia end i en af de tre øvrige byer. Dermed er det indirekte forudsat at de tre øvrige byer er stort set ens, hvilket man måake burde teste. Man kunne derfor anlægge følgende strategi for formulering og test af hypoteser:

1. Vi går stadig ud fra den multiplikative poissonmodel H_0 som grundmodel.
2. Først undersøges om det kan antages at de tre byer Horsens, Kolding og Vejle er ens, dvs. vi vil teste hypotesen

$$H_2 : \beta_2 = \beta_3 = \beta_4$$

3. Hvis H_2 bliver accepteret, er der et fælles niveau β for de tre »kontrolbyer«. Vi kan derefter sammenligne Fredericia med dette fælles niveau ved at teste om $\beta_1 = \beta$. Da β_1 pr. definition er lig 1, er den hypotese der skal testes,

$$H_3 : \beta = 1.$$

Vi skal således teste hypotesen $H_2 : \beta_2 = \beta_3 = \beta_4$ om ens kontrolbyer i forhold til den multiplikative model H_0. Det gøres med et kvotienttest

$$Q = \frac{L_2(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta})}{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_6, 1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4)}$$

hvor

$$L_2(\alpha_1, \alpha_2, \ldots, \alpha_6, \beta) = L_0(\alpha_1, \alpha_2, \ldots, \alpha_6, 1, \beta, \beta, \beta)$$

Tabel 13.6 De forventede antal $\widetilde{y}_{ij}$ af lungekræfttilfælde under H_2.

aldersklasse	Fredericia	Horsens	Kolding	Vejle	i alt
40-54	10.95	7.44	8.12	6.51	33.02
55-59	8.64	8.44	8.19	6.85	32.12
60-64	11.64	10.93	10.60	9.93	43.10
65-69	12.20	12.65	10.64	9.57	45.06
70-74	11.71	10.53	8.88	8.95	40.07
75+	8.95	8.36	7.04	6.61	30.96
i alt	64.09	58.35	53.47	48.42	224.33

er likelihoodfunktionen under H_2, og $\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta}$ er maksimaliseringsestimaterne under H_2. Når H_2 er rigtig, er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $f = 9-7 = 2$ frihedsgrader.

Modellen H_2 svarer til en multiplikativ poissonmodel med to byer (nemlig Fredericia og resten) og seks aldersklasser, og der er derfor ingen principielt nye problemer forbundet med at estimere parametrene under H_2. Man finder

$$\begin{aligned}
\widetilde{\alpha}_1 &= 0.004\\
\widetilde{\alpha}_2 &= 0.011\\
\widetilde{\alpha}_3 &= 0.016\\
\widetilde{\alpha}_4 &= 0.021\\
\widetilde{\alpha}_5 &= 0.023\\
\widetilde{\alpha}_6 &= 0.015\\
\widetilde{\beta}_1 &= 1\\
\widetilde{\beta} &= 0.722.
\end{aligned}$$

Endvidere bliver

$$-2\ln Q = 2\sum_{i=1}^{6}\sum_{j=1}^{4} y_{ij}\ln\frac{\widehat{y}_{ij}}{\widetilde{y}_{ij}}$$

hvor $\widehat{y}_{ij} = \widehat{\alpha}_i\widehat{\beta}_j r_{ij}$, se tabel 13.3, og

$$\begin{aligned}
\widetilde{y}_{i1} &= \widetilde{\alpha}_i\, r_{i1}\\
\widetilde{y}_{ij} &= \widetilde{\alpha}_i\widetilde{\beta}\, r_{ij}, \quad j = 2,3,4.
\end{aligned}$$

De forventede antal $\widetilde{y}_{ij}$ ses i tabel 13.6. Når man indsætter værdierne fra tabel 13.1, tabel 13.3 og tabel 13.6 i det netop fundne udtryk for $-2\ln Q$, fås $-2\ln Q_{\text{obs}} = 0.253$ der skal sammenholdes med χ^2-fordelingen med $f = 9-7 = 2$ frihedsgrader. I χ^2-fordelingen med $f = 2$ frihedsgrader er 20%-fraktilen 0.446, så testsandsynligheden er altså godt 80%, og det betyder at H_2 er udmærket forenelig med de foreliggende data. Vi kan altså udmærket tillade os at gå ud fra at der ikke er nogen signifikant forskel mellem de tre byer.

Herefter kan vi gå over til at teste H_3, der går ud på at alle fire byer er ens, og at der er de seks forskellige aldersgrupper med hver sin parameter α_i. Under forudsætning af H_2 er H_3 identisk med hypotesen H_1 fra tidligere, så estimaterne over aldersparametrene er $\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6$ fra side 181.

I denne omgang skal vi teste H_3 $(= H_1)$ i forhold til den nu gældende grundmodel H_2. Teststørrelsen er $-2\ln Q$ hvor

$$Q = \frac{L_1(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6)}{L_2(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, \widetilde{\beta})} = \frac{L_0(\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \ldots, \widehat{\widehat{\alpha}}_6, 1, 1, 1, 1)}{L_0(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_6, 1, \widetilde{\beta}, \widetilde{\beta}, \widetilde{\beta})}$$

der let omformes til

$$Q = \prod_{i=1}^{6} \prod_{j=1}^{4} \Big(\frac{\widehat{\widehat{y}}_{ij}}{\widetilde{y}_{ij}} \Big)^{y_{ij}}$$

så at

$$-2\ln Q = 2 \sum_{i=1}^{6} \sum_{j=1}^{4} y_{ij} \ln \frac{\widetilde{y}_{ij}}{\widehat{\widehat{y}}_{ij}}.$$

Store værdier af $-2\ln Q$ er signifikante. Når H_3 er rigtig, er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $f = 7 - 6 = 1$ frihedsgrad (forudsat at de indgående forventede antal er mindst fem).

Ved at indsætte værdierne fra tabel 13.1, tabel 13.5 og tabel 13.6 i det seneste udtryk for $-2\ln Q$ fås $-2\ln Q_{\text{obs}} = 4.61$ der med 1 frihedsgrad svarer til en testsandsynlighed på lidt over 3%. På det grundlag vil man almindeligvis *forkaste* hypotesen H_3 $(= H_1)$. Konklusionen bliver altså at *der ikke er signifikant forskel på lungekræfthyppigheden i de tre byer Horsens, Kolding og Vejle, hvorimod Fredericia har en signifikant anderledes lungekræfthyppighed.*

Den relative lungekræfthyppighed i de tre ens byer i forhold til Fredericia estimeres til $\widetilde{\beta} = 0.7$, så lungekræfthyppigheden i Fredericia er altså signifikant *større.*

Se det var jo en pæn og klar konklusion, der blot er stik modsat den vi nåede frem til på side 181!

13.6 Sammenligning af de to fremgangsmåder

Vi har benyttet to forskellige fremgangsmåder der kun var en smule forskellige, men gav helt forskellige resultater. De to fremgangsmåder er begge opbygget over følgende skema:

1. Find en passende grundmodel.
2. Formuler en hypotese der giver en forsimpling af den aktuelle grundmodel.
3. Test hypotesen i forhold til den aktuelle grundmodel.
4. a) Hvis hypotesen accepteres, så har vi derved fået en ny aktuel grundmodel (nemlig den gamle med de simplifikationer som den accepterede hypotese giver).
   Fortsæt da med punkt 2

Tabel 13.7 Oversigt over de to fremgangsmåder.

Første fremgangsmåde

Model/Hypotese		$-2\ln Q$	f	ε
M:	vilkårlige parametre	22.65	$24-9=15$	godt 5%
H:	multiplikativitet			
M:	multiplikativitet	5.67	$9-6=3$	ca. 20
H:	fire ens byer			

Anden fremgangsmåde

Model/Hypotese		$-2\ln Q$	f	ε
M:	vilkårlige parametre	22.65	$24-9=15$	godt 5%
H:	multiplikativitet			
M:	multiplikativitet	0.40	$9-7=2$	godt 80%
H:	de tre byer ens			
M:	de tre byer ens	5.27	$7-6=1$	ca. 2
H:	de fire byer ens			

b) Hvis hypotesen forkastes, så slut. Data beskrives da ved den senest anvendte grundmodel.

Begge fremgangsmåder tog udgangspunkt i den samme poissonmodel, de adskiller sig udelukkende ved valgene af hypoteser i punkt 2; tabel 13.7 giver en oversigt over de to fremgangsmåder.

I den første fremgangsmåde tages skridtet fra den multiplikative model til »fire ens« på én gang, hvilket giver en teststørrelse på 4.86, som, da den kan fordeles på 3 frihedsgrader, ikke er signifikant. I den anden fremgangsmåde spalter vi op i

1. multiplikativitet → »tre ens«, og
2. »tre ens« → »fire ens«,

og det viser sig så at de 4.86 med 3 frihedsgrader spaltes op i 0.25 med 2 frihedsgrader og 4.61 med 1 frihedsgrad, og her er det sidste bidrag signifikant. – Det kan undertiden være hensigtsmæssigt at foretage en sådan trinvis testning. Man bør dog ikke stræbe efter at spalte op i så mange tests som muligt, men kun teste hypoteser der er *rimelige* i den foreliggende faglige sammenhæng.

13.7 Om teststørrelser

Læseren vil måske have bemærket visse fælles træk ved de $-2\ln Q$-udtryk der forekommer i dette kapitel. De er alle af formen

$$-2\ln Q = 2\sum \text{obs.antal} \cdot \ln \frac{\text{Modellens forventede antal}}{\text{Hypotesens forventede antal}}$$

og er (tilnærmelsesvis) χ^2-fordelt med et antal frihedsgrader som er »det reelle antal parametre under modellen« minus »det reelle antal parametre under hypotesen«. Dette

gælder faktisk helt generelt når man tester hypoteser om poissonfordelte observationer (forudsat at summen af de forventede antal er lig summen af de observerede antal, og forudsat at de forventede antal alle er mindst 5).

13.8 Regn og tegn

Her vises hvordan man kan foretage de forskellige beregninger med computerprogrammet R, især funktionen `glm` der benyttes til generaliserede lineære modeller. Af forskellige grunde er det nemlig i visse sammenhænge praktisk at omparametrisere den multiplikative poissonmodel til en såkaldt log-lineær model idet man skriver

$$\ln \mathrm{E}\, Y_{ij} = \ln(\alpha_i \beta_j r_{ij}) = \ln \alpha_i + \ln \beta_j + \ln r_{ij}$$

hvor man så benytter $\ln \alpha_i$-erne og $\ln \beta_j$-erne som parametre. Funktionen `glm` udregner $\ln \widehat{\alpha}_i$-erne og $\ln \widehat{\beta}_j$-erne.

Data indlæses fra en fil der i dette tilfælde hedder `h:/bog/txt304ny/frcia.dat`. De første 10 linjer af filen ser sådan ud:

```
 y   r   Alder   By
11 3059   40-54  Fredericia
11  800   55-59  Fredericia
11  710   60-64  Fredericia
10  581   65-69  Fredericia
11  509   70-74  Fredericia
10  605   75+    Fredericia
13 2879   40-54  Horsens
 6 1083   55-59  Horsens
15  923   60-64  Horsens
```

Her kommer selve R-kommandoerne:

```
Frcia <- read.table ("h:/bog/txt304ny/frcia.dat", nrows=25, header = TRUE)

require (stats) # indlæser pakken stats der definerer xtabs
```

Tallene til tabel 13.1:

```
obs <- xtabs ( y ~ Alder + By, data=Frcia)
obs           # skriv indmaden af tabellen
rowSums (obs) #       rækkesummerne
colSums (obs) #       søjlesummerne
sum (obs)     #       totalsummen
```

og det samme med tabel 13.2

```
antal <- xtabs ( r ~ Alder + By, data=Frcia)
antal; rowSums (antal); colSums (antal); sum (antal)
```

Vedr. afsnit 13.3, Den multiplikative model:

```
# estimation:
H0 <- glm( y ~ 0 + Alder + By, offset=log(r), family=poisson, data=Frcia)
```

```
H0$deviance                          # -2lnQ
H0$df.res                            # antal frihedsgrader
1 - pchisq (H0$deviance, H0$df.res) # den tilsv. testsandsynlighed

# Koefficienterne (dvs. de estimerede alfa'er og beta'er):
round (exp(H0$coef), digits=3) # skrives med 3 cifre efter kommaet

# de forventede værdier(med én decimal)
T0 <- round (xtabs (H0$fitted ~ Alder+By , data=Frcia), digits=1)
T0; rowSums (T0); colSums (T0); sum (T0)

# de estimerede intensiteter
round (xtabs (H0$fitted*1000/r ~ Alder+By , data=Frcia), digits=1)
```

Vedr. afsnit 13.4, Ens byer?:

```
# vi opdaterer modellen H0 så den kun indeholder variablen Alder:
H1 <- update(H0, . ~ 0 + Alder)

# de estimerede alders-parametre:
round (exp(H1$coef), digits=3)

# Udregner -2lnQ for H1 mod H0:
anova (H1, H0) # giver en -2lnQ (Deviance) på 4.859 med 3 frihedsgr.
1 - pchisq(4.859, 3) # testsandsynligheden

# de forventede værdier
T1 <- round (xtabs (H1$fitted ~ Alder+By , data=Frcia), digits=1)
T1; rowSums (T1); colSums (T1); sum (T1)
```

Vedr. afsnit 13.5, En anden mulighed: Her er tale om at de tre byer slås sammen til én kontrolby. Vi definerer derfor en ny faktor `isFr` der fortæller om byen er Fredericia eller Kontrol:

```
Frcia$isFr <- factor(ifelse(Frcia$By == "Fredericia", "Fredericia", "Kontrol"))

# så fitter vi den nye model (som er en delmodel af H0)
H2 <- update (H0, . ~ 0 + Alder + isFr)

round (exp(H2$coef), digits=3) # de estimerede koefficienter

# udregn -2lnQ for H2 mod H0:
anova (H2, H0) # giver en -2lnQ (Deviance) på 0.2526 med 2 frihedsgr.
1 - pchisq (0.2526, 2)  # testsandsynligheden

# de forventede værdier
T2 <- round (xtabs (H2$fitted ~ Alder+By , data=Frcia), digits=1)
T2; rowSums (T2); colSums (T2); sum (T2)
```

Endelig testes hypotesen om ens byer i forhold til H_2:

```
H3 <- update (H2, . ~ 0 + Alder)
round (exp(H3$coef), digits=3) # estimaterne
anova (H3, H2) # giver en -2lnQ (Deviance) på 4.6065 med 1 frihedsgr.
1 - pchisq (4.6065, 1)
```

14 Multinomialfordelingen

Multinomialfordelingen kan ses som en naturlig generalisation af binomialfordelingen:

- I situationer hvor man har at gøre med n gentagelser af et elementarforsøg der kan resultere i et af *to* mulige udfald, vil antallet af gange man får den ene slags udfald, blive *binomialfordelt*.
- I situationer hvor man har at gøre med n gentagelser af et elementarforsøg der kan resultere i et af r mulige udfald, vil man et vist antal gange, y_1, få det første udfald, et vist antal gange, y_2, det andet udfald, ..., og et vist antal gange, y_r, det r-te udfald; talsættet $(y_1, y_2, \ldots, y_r)$ bliver *multinomialfordelt*.

Eksempel 14.1
En simpel form for politisk meningsmålingsundersøgelse kunne bestå i at man tilfældigt udvælger n personer og spørger dem hvilket af de r politiske partier de ville stemme på hvis der var folketingsvalg i morgen.

Her består elementarforsøget i at spørge én person og notere den pågældendes svar ned. Den samlede undersøgelse resulterer i at et vist antal y_1 svarer det første parti, et vist antal y_2 svarer det andet parti, ..., og et vist antal y_r svarer det r-te parti. Da der i alt er spurgt n personer, vil der gælde at $y_1 + y_2 + \cdots + y_r = n$, forudsat at alle de adspurgte faktisk svarer.

Den multinomialfordelingsmodel vi i det følgende vil diskutere, svarer til at når man vælger en tilfældig person, så vil denne med en vis sandsynlighed p_1 svare parti nr. 1, med en vis sandsynlighed p_2 svare parti nr. 2, ..., og med en vis sandsynlighed p_r svare parti nr. r. Da vi forudsætter at alle adspurgte giver et af de r mulige svar, er $p_1 + p_2 + \cdots + p_r = 1$.

14.1 Den grundlæggende multinomialfordelingsmodel

Antag at vi har klassificeret n individer i r klasser; i den generelle diskussion kaldes klasserne $A_1, A_2, \ldots, A_r$, i en konkret modelsituation har de ofte nogle mere sigende betegnelser. Skematisk er situationen som vist i figur 14.1.

Vi går ud fra at de n individer stammer fra en og samme »population«, således at hver gang man tilfældigt udvælger et individ, er der sandsynligheden p_1 for at individet tilhører klassen A_1, sandsynligheden p_2 for at individet tilhører klassen A_2, osv. Sandsynlighederne $p_1, p_2, \ldots, p_r$ (der summerer til 1) er *ukendte parametre* der er karakteristiske for populationen.

Hermed har vi sådan set beskrevet den statistiske model for ét individ. Når der er et større antal individer, plejer man ikke at angive hvilken klasse hvert enkelt individ viser sig at tilhøre, man nøjes med at angive hvor mange individer der er i hver klasse, dvs. man angiver de observerede værdier af de stokastiske variable $Y_1, Y_2, \ldots, Y_r$ defineret som Y_i = antal individer der viser sig at tilhøre klassen A_i, $(i = 1, 2, \ldots, r)$.

klasse-nummer	klasse-navn	observeret antal
1	A_1	y_1
2	A_2	y_2
3	A_3	y_3
$\vdots$	$\vdots$	$\vdots$
r	A_r	y_r
i alt		n

Figur 14.1 Multinomialfordelingssituationen, skematisk.

Den statistiske model vi skal nå frem til, skal specificere sandsynlighedsfordelingen for sættet $(Y_1, Y_2, \ldots, Y_r)$ af stokastiske variable, eller sagt på en anden måde, vi skal fastlægge $\mathrm{P}(Y_1 = y_1,\ Y_2 = y_2,\ \ldots,\ Y_r = y_r)$ som funktion af $(y_1, y_2, \ldots, y_r)$.

Hvis der kun er *to* klasser, så er der tale om et binomialfordelingsproblem. For at løse problemet med r klasser går vi frem på en måde der er stærkt inspireret af udledningen af binomialfordelingen i begyndelsen af kapitel 1.

Vi indfører nogle hjælpevariable $X_1, X_2, \ldots, X_n$ således at X_d betegner navnet på den klasse som individ nr. d tilhører, dvs. $X_d = A_i$ hvis og kun hvis individ nr. d tilhører klassen A_i. Der gælder så at $\mathrm{P}(X_d = A_i) = p_i$. Da individerne tænkes valgt uafhængigt af hverandre, må de forskellige X_d-er være stokastisk uafhængige, således at f.eks.

$$\mathrm{P}(X_{d_1} = A_{i_1},\ X_{d_2} = A_{i_2}) = p_{i_1} p_{i_2} \quad \text{hvis } d_1 \neq d_2.$$

Hvis vi har n klassenavne $x_1, x_2, \ldots, x_n$, og hvis det er sådan at netop y_i af x-erne er et A_i, $i = 1, 2, \ldots, r$, så er

$$\begin{aligned}
&\mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\
&\quad = \mathrm{P}(X_1 = x_1) \cdot \mathrm{P}(X_2 = x_2) \cdot \ldots \cdot \mathrm{P}(X_n = x_n) \\
&\quad = p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r}.
\end{aligned}$$

Den søgte sandsynlighed $\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r)$ fås nu ved at summere disse sandsynligheder over alle mulige n-tupler $(x_1, x_2, \ldots, x_n)$ bestående af y_1 A_1-er, y_2 A_2-er, $\ldots$, y_r A_r-er:

$$\begin{aligned}
&\mathrm{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r) \\
&\quad = \sum \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\
&\quad = \sum p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r} \\
&\quad = \Big(\sum 1\Big) \cdot p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r}
\end{aligned}$$

hvor summationstegnet hver gang betyder summation over de n-tupler $(x_1, x_2, \ldots, x_n)$ som består af y_1 A_1-er, y_2 A_2-er osv. Symbolet $\sum 1$ kommer på den måde til at betyde

antallet af forskellige sådanne n-tupler $(x_1, x_2, \ldots, x_n)$; dette antal plejer man at betegne med symbolet

$$\binom{n}{y_1\ y_2\ \ldots\ y_r}$$

der kaldes en *multinomialkoefficient* (eller *polynomialkoefficient*). Den fundne sandsynlighedsfunktion

$$\mathrm{P}\big(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r\big) = \binom{n}{y_1\ y_2\ \ldots\ y_r} p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r}$$

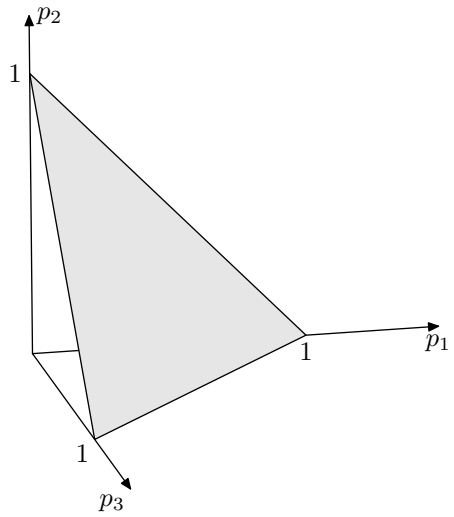
er sandsynlighedsfunktionen for en *multinomialfordeling* (eller *polynomialfordeling*) *med parametre* n *og* $\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{pmatrix}$.

Multinomialkoefficienter

Definition 14.1: Multinomialkoefficient
Multinomialkoefficienten $\binom{n}{y_1\ y_2\ \ldots\ y_r}$ *betegner antallet af forskellige måder hvorpå man kan placere* r *symboler* $A_1, A_2, \ldots, A_r$ *på* n *pladser således at symbolet* A_1 *kommer på* y_1 *af pladserne, symbolet* A_2 *kommer på* y_2 *af pladserne, ..., symbolet* A_r *kommer på* y_r *af pladserne.*

Man kan let udlede formler der gør det muligt at udregne multinomialkoefficienter. Vi illustrerer fremgangsmåden med et eksempel, hvor vi vil bestemme talværdien af $\binom{7}{2\ 3\ 2}$:

1. Det søgte tal er pr. definition antallet af placeringer af symbolerne A_1, A_2 og A_3 på syv pladser således at to af pladserne får et A_1, tre af pladserne et A_2 og to af pladserne et A_3. – En mulig placering er $A_1, A_3, A_1, A_2, A_2, A_2, A_3$.
2. Vi kan bestemme en placering ved først at bestemme hvilke to pladser der skal have et A_1, dernæst hvilke tre pladser der skal have et A_2, og så endelig placere et A_3 på de to tiloversblevne pladser.
   a) Der er $\binom{7}{2} = 21$ forskellige placeringer af de to A_1-er (jf. definitionen af binomialkoefficienter på side 12).
   b) Hver gang vi har placeret de to A_1-er, er der fem pladser tilbage, og på de fem pladser skal vi fordele tre A_2-er og to A_3-er; dette kan gøres på $\binom{5}{3} = 10$ forskellige måder. Hver gang vi har en af de $\binom{7}{2}$ placeringer af A_1, er der altså $\binom{5}{3}$ placeringer af A_2 og A_3.
3. I alt er der derfor $\binom{7}{2} \cdot \binom{5}{3}$ forskellige placeringer af A-erne så
$$\binom{7}{2\ 3\ 2} = \binom{7}{2} \cdot \binom{5}{3} = 21 \cdot 10 = 210.$$
4. Vi kan benytte formlen $\binom{n}{k} = \dfrac{n!}{k!\,(n-k)!}$ og få
$$\binom{7}{2\ 3\ 2} = \binom{7}{2} \cdot \binom{5}{3} = \frac{7!}{2!\,5!} \cdot \frac{5!}{3!\,2!} = \frac{7!}{2!\,3!\,2!}.$$



Figur 14.2 Sandsynlighedssimplexet i det tredimensionale rum.

Et generelt udtryk for multinomialkoefficienter fås på ganske tilsvarende måde. Man skal placere y_1 A_1-er, y_2 A_2-er, ..., og y_r A_r-er på n pladser ($n = y_1 + y_2 + \cdots + y_r$). Først kan A_1-erne placeres på $\binom{n}{y_1}$ forskellige måder; dernæst kan A_2-erne placeres på de de resterende $n - y_1$ pladser på $\binom{n-y_1}{y_2}$ forskellige måder; dernæst kan A_3-erne placeres på de resterende $n - y_1 - y_2$ pladser på $\binom{n-y_1-y_2}{y_3}$ måder, osv. Slutresultatet bliver at

$$\binom{n}{y_1\ y_2\ \ldots\ y_r} = \frac{n!}{y_1!\, y_2!\, \ldots\, y_r!}$$

når $y_1 + y_2 + \cdots + y_r = n$.

Definition 14.2: Multinomialfordeling
At den r-dimensionale stokastiske variabel $(Y_1, Y_2, \ldots, Y_r)$ er multinomialfordelt med antalsparameter n og sandsynlighedsparameter $\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{pmatrix}$ betyder at

$$\mathrm{P}\big(Y_1 = y_1, Y_2 = y_2, \ldots, Y_r = y_r\big) = \binom{n}{y_1\ y_2\ \ldots\ y_r} p_1^{y_1} p_2^{y_2} \ldots p_r^{y_r} \tag{14.1}$$

når $y_1, y_2, \ldots, y_r$ er ikke-negative heltal der summerer til n.

Estimation af parametrene

I den generelle situation er modelfunktionen givet ved formel (14.1), og likelihoodfunktionen er dermed $L(\boldsymbol{p}) = \text{konstant} \cdot p_1^{y_1} p_2^{y_2} \dots p_r^{y_r}$. Spørgsmålet er nu hvordan man estimerer parameteren $\boldsymbol{p}$.

De almene principper for analyse af statistiske modeller påbyder at estimere $\boldsymbol{p}$ ved det r-dimensionale talsæt $\widehat{\boldsymbol{p}}$ der maksimaliserer likelihoodfunktionen. Likelihoodfunktionen er en funktion af $\boldsymbol{p}$, dvs. af de r variable $p_1, p_2, \dots, p_r$; disse kan ikke variere frit, men opfylder bibetingelserne $p_1 \geq 0$, $p_2 \geq 0$, $\dots$, $p_r \geq 0$, og $p_1 + p_2 + \cdots + p_r = 1$.

I specialtilfældet $r = 3$ kan vi anskueliggøre mulighedsområdet, dvs. mængden af $\boldsymbol{p}$-er der opfylder bibetingelserne, som et trekantet område, det såkaldte sandsynlighedssimplex, i det tredimensionale rum, se figur 14.2.

Opgaven er at bestemme det punkt $\widehat{\boldsymbol{p}} = \begin{pmatrix} \widehat{p}_1 \\ \widehat{p}_2 \\ \vdots \\ \widehat{p}_r \end{pmatrix}$ som ligger i mulighedsområdet, og hvor likelihoodfunktionen L antager sin største værdi. I matematikken diskuteres generelle metode til bestemmelse af maksimumspunkter for funktioner af mange variable, men disse metoder skal vi ikke komme ind på her. Derimod vil vi løse det specielle problem der vedrører multinomialfordelingen. Dertil skal vi bruge følgende

Sætning 14.1
Antag at $a_1, a_2, \dots, a_r$ er givne ikke-negative tal, og betragt funktionen

$$f : (p_1, p_2, \dots, p_r) \longmapsto p_1^{a_1} p_2^{a_2} \dots p_r^{a_r}$$

defineret på mængden af ikke-negative talsæt $(p_1, p_2, \dots, p_r)$ der summerer til 1. Vi sætter $a_\bullet = a_1 + a_2 + \cdots + a_r$ og $\widehat{p}_i = a_i / a_\bullet$, $i = 1, 2, \dots, r$.

Da har f et entydigt maksimumspunkt, nemlig $(\widehat{p}_1, \widehat{p}_2, \dots, \widehat{p}_r)$.

Bevis
Vi vil sammenligne funktionsværdierne $f(p_1, p_2, \dots, p_r)$ og $f(\widehat{p}_1, \widehat{p}_2, \dots, \widehat{p}_r)$ ved at se på størrelsen $\ln \dfrac{f(p_1, p_2, \dots, p_r)}{f(\widehat{p}_1, \widehat{p}_2, \dots, \widehat{p}_r)}$ som er negativ hvis og kun hvis $f(p_1, p_2, \dots, p_r) < f(\widehat{p}_1, \widehat{p}_2, \dots, \widehat{p}_r)$.

Der gælder først at $\ln \dfrac{f(p_1, p_2, \dots, p_r)}{f(\widehat{p}_1, \widehat{p}_2, \dots, \widehat{p}_r)} = \displaystyle\sum_{i=1}^{r} a_i \ln \frac{p_i}{\widehat{p}_i}$. Nu benyttes en egenskab ved logaritmefunktionen, nemlig at $\ln t \leq t - 1$ for alle $t > 0$, og med lighedstegn hvis og kun hvis $t = 1$. Derfor er

$$\begin{aligned} \sum_{i=1}^{r} a_i \ln \frac{p_i}{\widehat{p}_i} &\leq \sum_{i=1}^{r} a_i \Big(\frac{p_i}{\widehat{p}_i} - 1 \Big) = \sum_{i=1}^{r} \Big(\frac{a_i p_i}{a_i / a_\bullet} - a_i \Big) \\ &= \sum_{i=1}^{r} p_i a_\bullet - \sum_{i=1}^{r} a_i = a_\bullet - a_\bullet = 0, \end{aligned}$$

hvor »mindre end eller lig med« bliver »lig med« hvis og kun hvis alle tallene $p_i / \widehat{p}_i$ er lig 1, dvs. hvis og kun hvis $p_i = \widehat{p}_i$ for alle i. □

Tabel 14.1 Genotypefordeling af torsk fra tre lokaliteter i Østersøen.

	lokalitet		
genotype	Lolland	Bornholm	Ålandsøerne
AA	27	14	0
Aa	30	20	5
aa	12	52	75
i alt	69	86	80

Anvendt på funktionen $(p_1, p_2, \ldots, p_r) \mapsto p_1^{y_1}\, p_2^{y_2} \ldots p_r^{y_r}$ fortæller sætningen at likelihoodfunktionen L antager sit maksimum i det entydigt bestemte punkt $(\frac{y_1}{n}, \frac{y_2}{n}, \ldots, \frac{y_r}{n})$. Derfor er maksimaliseringsestimatet $\widehat{\boldsymbol{p}}$ for $\boldsymbol{p}$ givet ved

$$\widehat{\boldsymbol{p}} = \begin{pmatrix} \widehat{p}_1 \\ \widehat{p}_2 \\ \vdots \\ \widehat{p}_r \end{pmatrix} = \begin{pmatrix} y_1/n \\ y_2/n \\ \vdots \\ y_r/n \end{pmatrix}.$$

Parameteren p_i, der jo er sandsynligheden for at et individ tilhører klassen A_i, skal altså estimeres ved den relative hyppighed y_i/n af A_i-individer i stikprøven.

14.2 Sammenligning af multinomialfordelinger

Man har undertiden brug for at kunne sammenligne forskellige multinomialfordelinger for at afgøre om de har samme sandsynlighedsparameter. Her er et eksempel; det vil blive analyseret mere indgående i kapitel 16:

Eksempel 14.2 (Torsk i Østersøen)
Den 6. marts 1961 fangede nogle havbiologer 69 torsk ved Lolland og undersøgte arten af blodets hæmoglobin i hver enkelt torsk. Senere på året fangede man også nogle torsk ved Bornholm og ved Ålandsøerne og bestemte deres genotype. (19)

Man mener at hæmoglobin-arten bestemmes af ét enkelt gen, og det som biologerne bestemte, var torskenes genotype for så vidt angår dette gen. Genet kan optræde i to udgaver som traditionen tro kaldes for A og a, og de mulige genotyper er da AA, Aa og aa. Den fundne genotypefordeling på hver lokalitet ses i tabel 14.1. I dette afsnit vil vi udelukkende opfatte symbolerne AA, Aa og aa som *navne* på klasser man klassificerer torskene i. I kapitel 16 vil vi smugle lidt genetik ind i en mere udbygget statistisk model for tallene.

På hver geografisk lokalitet er der sket det at man har klassificeret et antal torsk i tre mulige klasser, så derfor kan man sige at der på hver lokalitet er tale om en multinomialfordelingssituation (når der er tre klasser, taler man også om en *tri*nomialfordeling). Det kunne måske være af interesse at undersøge om genotypefordelingen er den samme på de tre lokaliteter, altså om sandsynligheden for at en torsk har en bestemt genotype, er den samme for alle tre lokaliteters vedkommende. (Skønt når man ser på tallene virker denne formodning lidet plausibel.)

	gruppe nr.				
klasse	1	2	3	...	s
A_1	y_{11}	y_{12}	y_{13}	...	y_{1s}
A_2	y_{21}	y_{22}	y_{23}	...	y_{2s}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_r	y_{r1}	y_{r2}	y_{r3}	...	y_{rs}
i alt	n_1	n_2	n_3	...	n_s

Figur 14.3 Sammenligning af multinomialfordelinger, generelt. y_{ij} betegner antallet af individer fra gruppe j der tilhører klassen A_i.

Den generelle model

I den generelle model antages det at vi har klassificeret nogle individer i r forskellige klasser $A_1, A_2, \ldots, A_r$. Individerne er på forhånd delt op i grupper, og der er s forskellige grupper med hhv. $n_1, n_2, \ldots, n_s$ individer. Det har vist sig at i gruppe j hører y_{1j} af individerne til gruppen A_1, y_{2j} af individerne til gruppen A_2, $y_{3,j}$ af individerne til gruppen A_3, osv. Skematisk ser situationen ud som vist i figur 14.3.

I torskeeksemplet er der $s = 3$ grupper svarende til de tre geografiske lokaliteter og $r = 3$ klasser svarende til de tre forskellige genotyper.

Den statistiske model der benyttes til at beskrive denne situation er:

- for hvert j (dvs. for hver gruppe) opfattes det r-dimensionale talsæt

$$\boldsymbol{y}_j = \begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{rj} \end{pmatrix}$$

  som en observeret værdi af en r-dimensional stokastisk variabel

$$\boldsymbol{Y}_j = \begin{pmatrix} Y_{1j} \\ Y_{2j} \\ \vdots \\ Y_{rj} \end{pmatrix};$$

- de stokastiske variable $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_s$ er stokastisk uafhængige (dvs. de forskellige grupper er stokastisk uafhængige);
- den stokastiske variabel $\boldsymbol{Y}_j$ er multinomialfordelt med antalsparameter n_j og med ukendt sandsynlighedsparameter

$$\boldsymbol{p}_j = \begin{pmatrix} p_{1j} \\ p_{2j} \\ \vdots \\ p_{rj} \end{pmatrix}$$

hvor p_{ij}-erne er ikke-negative tal med $p_{1j} + p_{2j} + \ldots + p_{rj} = 1$ for hvert j.
Modellen tager altså udgangspunkt i at grupperne er systematisk forskellige (mht. den foretagne klassificering), og den beskriver den såkaldte *systematiske variation mellem grupperne* ved hjælp af de s sandsynlighedsparametre $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s$. Den såkaldte *tilfældige variation inden for grupper* beskrives ved sandsynlighedsfordelingerne (multinomialfordelingerne).

Opgaven er nu at undersøge om grupperne kan anses for ens, dvs. den er at teste den statistiske hypotese

$$H_0 : \ \boldsymbol{p}_1 = \boldsymbol{p}_2 = \cdots = \boldsymbol{p}_s$$

eller mere udførligt

$$H_0 : \ \begin{pmatrix} p_{11} \\ p_{21} \\ \vdots \\ p_{r1} \end{pmatrix} = \begin{pmatrix} p_{12} \\ p_{22} \\ \vdots \\ p_{r2} \end{pmatrix} = \ldots = \begin{pmatrix} p_{1s} \\ p_{2s} \\ \vdots \\ p_{rs} \end{pmatrix} .$$

De generelle retningslinjer for hvordan man analyserer en given statistisk model, siger at vi skal begynde med at opskrive modelfunktionen og derudaf få likelihoodfunktionen. Da de enkelte grupper er stokastisk uafhængige, er den samlede modelfunktion lig med et produkt af del-modelfunktionerne for de enkelte grupper, dvs. *den samlede modelfunktion* er

$$f(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_s; \boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s) = \prod_{j=1}^{s} \binom{n_j}{y_{1j}\ y_{2j}\ \ldots\ y_{rj}} p_{1j}^{y_{1j}} p_{2j}^{y_{2j}} \ldots p_{rj}^{y_{rj}} .$$

Likelihoodfunktionen er dermed

$$L(\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s) = \text{konstant} \cdot \prod_{j=1}^{s} p_{1j}^{y_{1j}} p_{2j}^{y_{2j}} \ldots p_{rj}^{y_{rj}} \tag{14.2}$$

hvor konstanten er produktet af de s multinomialkoefficienter. I torskeeksemplet er likelihoodfunktionen

$$L(\boldsymbol{p}_{\mathrm{L}}, \boldsymbol{p}_{\mathrm{B}}, \boldsymbol{p}_{\mathrm{Å}}) = \text{konstant} \cdot p_{1\mathrm{L}}^{27} p_{2\mathrm{L}}^{30} p_{3\mathrm{L}}^{12} \cdot p_{1\mathrm{B}}^{14} p_{2\mathrm{B}}^{20} p_{3\mathrm{B}}^{52} \cdot p_{1\mathrm{Å}}^{0} p_{2\mathrm{Å}}^{5} p_{3\mathrm{Å}}^{75} .$$

Likelihoodfunktionen er sandsynligheden for at observere det faktisk observerede, betragtet som funktion af det ukendte sæt parametre. Som sædvanlig udnævner vi de værdier der maksimaliserer likelihoodfunktionen (eller log-likelihoodfunktionen) til at være de bedste estimater over de ukendte parametre. I den foreliggende model er likelihoodfunktionen et produkt af s del-likelihoodfunktioner der hver især vedrører én enkelt gruppe og ét enkelt $\boldsymbol{p}_j$. Når vi skal maksimalisere L mht. $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s$, kan det derfor ske ved at maksimalisere hver del-likelihoodfunktion for sig. Det j-te delproblem er en simpel multinomialfordelingsmodel, så derfor følger det uden videre af resultatet på

side 194 at $\widehat{p}_{ij} = y_{ij}/n_j$ for alle i og j. I taleksemplet er specielt

$$\widehat{\boldsymbol{p}}_{\mathrm{L}} = \begin{pmatrix} \widehat{p}_{1\mathrm{L}} \\ \widehat{p}_{2\mathrm{L}} \\ \widehat{p}_{3\mathrm{L}} \end{pmatrix} = \begin{pmatrix} 27/69 \\ 30/69 \\ 12/69 \end{pmatrix} = \begin{pmatrix} 0.39 \\ 0.43 \\ 0.17 \end{pmatrix},$$

$$\widehat{\boldsymbol{p}}_{\mathrm{B}} = \begin{pmatrix} \widehat{p}_{1\mathrm{B}} \\ \widehat{p}_{2\mathrm{B}} \\ \widehat{p}_{3\mathrm{B}} \end{pmatrix} = \begin{pmatrix} 14/86 \\ 20/86 \\ 52/86 \end{pmatrix} = \begin{pmatrix} 0.16 \\ 0.23 \\ 0.60 \end{pmatrix},$$

$$\widehat{\boldsymbol{p}}_{\mathrm{Å}} = \begin{pmatrix} \widehat{p}_{1\mathrm{Å}} \\ \widehat{p}_{2\mathrm{Å}} \\ \widehat{p}_{3\mathrm{Å}} \end{pmatrix} = \begin{pmatrix} 0/80 \\ 5/80 \\ 75/80 \end{pmatrix} = \begin{pmatrix} 0.00 \\ 0.06 \\ 0.94 \end{pmatrix}.$$

Hypoteseprøvning

Vi skal herefter undersøge om det er rimeligt at antage at hypotesen

$$H_0 : \boldsymbol{p}_1 = \boldsymbol{p}_2 = \cdots = \boldsymbol{p}_s$$

om ens sandsynlighedsparametre holder. Under H_0 er der ingen forskel på de s grupper, så da kan vi lige så godt slå dem sammen til én stor gruppe bestående af $n_\bullet = n_1 + n_2 + \cdots + n_s$ individer der fordeler sig med

$$\begin{array}{rcccl}
y_{1\bullet} & = & y_{11} + y_{12} + \ldots + y_{1s} & = & \displaystyle\sum_{j=1}^{s} y_{ij} \quad \text{i klassen } A_1 \\
y_{2\bullet} & = & y_{21} + y_{22} + \ldots + y_{2s} & = & \displaystyle\sum_{j=1}^{s} y_{2j} \quad \text{i klassen } A_2 \\
\vdots & & \vdots & & \vdots \\
y_{i\bullet} & = & y_{i1} + y_{i2} + \ldots + y_{is} & = & \displaystyle\sum_{j=1}^{s} y_{ij} \quad \text{i klassen } A_i \\
\vdots & & \vdots & & \vdots \\
y_{r\bullet} & = & y_{r1} + y_{r2} + \ldots + y_{rs} & = & \displaystyle\sum_{j=1}^{s} y_{rj} \quad \text{i klassen } A_r
\end{array}$$

Man må derfor formode at den fælles værdi p_i af sandsynligheden for at tilhøre klassen A_i skal estimeres ved $y_{i\bullet}/n_\bullet$, men lad os prøve at gå frem efter likelihoodmetoden.

Vi kalder den fælles værdi (under H_0) af $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_s$ for $\boldsymbol{p}$,

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_r \end{pmatrix}.$$

Tabel 14.2 Genotypefordeling hos torsk fra tre lokaliteter i Østersøen: forventede antal under antagelse af ens fordelinger på de tre lokaliteter.

	lokalitet		
genotype	Lolland	Bornholm	Ålandsøerne
AA	12.0	15.0	14.0
Aa	16.1	20.1	18.7
aa	40.8	50.9	47.3
i alt	68.9	86.0	80.0

I likelihoodfunktionen (14.2) erstatter vi alle $\boldsymbol{p}_j$-erne med $\boldsymbol{p}$ og får derved *likelihoodfunktionen under* H_0:

$$\begin{aligned} L(\boldsymbol{p}, \boldsymbol{p}, \ldots, \boldsymbol{p}) &= \text{konstant} \cdot \prod_{j=1}^{s} p_1^{y_{1j}} p_2^{y_{2j}} \ldots p_r^{y_{rj}} \\ &= \text{konstant} \cdot p_1^{y_{1\bullet}} p_2^{y_{2\bullet}} \ldots p_r^{y_{r\bullet}}. \end{aligned}$$

Det valg af $p_1, p_2, \ldots, p_r$ der maksimaliserer denne likelihoodfunktion, er ifølge sætningen på side 193 netop $\widehat{p}_i = y_{i\bullet}/n_\bullet$ som formodet.

I taleksemplet bliver $\widehat{\boldsymbol{p}} = \begin{pmatrix} 41/235 \\ 55/235 \\ 139/235 \end{pmatrix} = \begin{pmatrix} 0.17 \\ 0.23 \\ 0.59 \end{pmatrix}$.

Når man vil vurdere hvor godt det faktisk observerede beskrives under H_0 i forhold til den aktuelle grundmodels beskrivelse, skal man udregne *kvotienttesstørrelsen*

$$Q = \frac{L(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{p}}, \ldots, \widehat{\boldsymbol{p}})}{L(\widehat{\boldsymbol{p}}_1, \widehat{\boldsymbol{p}}_2, \ldots, \widehat{\boldsymbol{p}}_s)}$$

eller værdien $-2 \ln Q$. En Q-værdi tæt på 1, dvs. en $-2 \ln Q$-værdi tæt på 0, betyder at H_0 beskriver data næsten lige så godt som grundmodellen gør, hvorimod en Q-værdi nær 0, dvs. en stor $-2 \ln Q$-værdi, betyder at H_0 giver en væsentligt dårligere beskrivelse end grundmodellen gør. Man plejer at udregne $-2 \ln Q$ (og ikke Q).

Når man indsætter udtrykkene for L i Q, får man let at

$$\begin{aligned} -2 \ln Q &= 2 \sum_{j=1}^{s} \Big(y_{1j} \ln \frac{y_{1j}}{\widehat{y}_{1j}} + y_{2j} \ln \frac{y_{2j}}{\widehat{y}_{2j}} + \ldots + y_{rj} \ln \frac{y_{rj}}{\widehat{y}_{rj}} \Big) \\ &= 2 \sum_{j=1}^{s} \sum_{i=1}^{r} y_{ij} \ln \frac{y_{ij}}{\widehat{y}_{ij}}, \end{aligned}$$

hvor $\widehat{y}_{ij} = \widehat{p}_i n_j = y_{i\bullet} n_j / n_\bullet$ er det »forventede« antal individer fra gruppe j der klassificeres som A_i.

For at bestemme $-2\ln Q$ i taleksemplet udregnes først de forventede antal, se tabel 14.2. Dermed er

$$\begin{aligned}-2\ln Q_{\text{obs}} = 2\Big(&27\ln\frac{27}{12.0} + 14\ln\frac{14}{15.0} + 0\ln\frac{0}{14.0}\\ &+ 30\ln\frac{30}{16.1} + 20\ln\frac{20}{20.1} + 5\ln\frac{5}{18.7}\\ &+ 12\ln\frac{12}{40.8} + 52\ln\frac{52}{50.9} + 75\ln\frac{75}{47.3}\Big)\\ = 107.8\end{aligned}$$

For at afgøre om en opnået $-2\ln Q_{\text{obs}}$-værdi (som f.eks. 107.8) nu er tæt på 0 eller ej, skal man sammenligne den med alle de andre $-2\ln Q$-værdier man også kunne have fået ifølge den aktuelle model, når H_0 er rigtig. Vi skal derfor finde *testsandsynligheden* ε, dvs. sandsynligheden for at få en værre (større) $-2\ln Q$-værdi end den observerede, under forudsætning af at H_0 er rigtig: $\varepsilon = \mathrm{P}_0\big(-2\ln Q \geq -2\ln Q_{\text{obs}}\big)$. Når man skal bestemme ε, kan man udnytte en generel matematisk sætning der fortæller at når H_0 er rigtig, så er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $(r-1)(s-1)$ frihedsgrader således at ε med god tilnærmelse kan bestemmes som sandsynligheden for at få en værdi større end $-2\ln Q_{\text{obs}}$ i en χ^2-fordeling med $(r-1)(s-1)$ frihedsgrader, kort

$$\varepsilon = \mathrm{P}\big(\chi^2_{(r-1)(s-1)} \geq -2\ln Q_{\text{obs}}\big),$$

og denne sandsynlighed er let at bestemme ved hjælp af tabeller over fraktiler i χ^2-fordelingen.

Antallet af frihedsgrader for $-2\ln Q$ findes som *ændringen i antallet af frie parametre*: i grundmodellen er der for hver af de s grupper $r-1$ parametre (fordi der er r klasser og dermed r sandsynligheder der skal summere til 1), altså i alt $s(r-1)$ parametre; under H_0 er der i realiteten kun én gruppe og dermed $r-1$ frie parametre; antallet af frihedsgrader for teststørrelsen er derfor $s(r-1)-(r-1) = (r-1)(s-1)$.

Bemærk at χ^2-fordelingen kun er en approksimation; for at man skal kunne bruge den, *skal alle de »forventede« antal* $\widehat{y}_{ij} = \widehat{p}_i n_j = y_{i\bullet} n_j / n_\bullet$ *være mindst fem*. Hvis denne betingelse ikke er opfyldt, kan man måske opnå at den bliver opfyldt ved at man udelader nogle grupper eller klasser eller slår nogle grupper eller klasser sammen.

I det gennemgående taleksempel er der ingen problemer med at de »forventede« antal er for små. Vi kan derfor uden videre sammenligne $-2\ln Q_{\text{obs}} = 107.8$ med χ^2-fordelingen med $(3-1)(3-1) = 4$ frihedsgrader. Da 99.9%-fraktilen i denne fordeling er 18.47, er testsandsynligheden ε mindre end 0.1%. Da det således er temmelig usandsynligt at få en værre værdi af teststørrelsen $-2\ln Q$ end 107.8, er teststørrelsen *signifikant* stor, og vi forkaster H_0. Man må altså sige at der er en signifikant forskel på genotypen af torsk på de tre geografiske lokaliteter. – Denne konklusion er ikke overraskende hvis man man sammenligner tabel 14.1 og 14.2.

14.3 Regn og tegn

Her demonstreres ved at gennemregne torskeeksemplet hvordan man med R kan foretage beregningerne til sammenligning af multinomialfordelinger.

Data indlæses fra filen `h:/bog/txt304ny/r-torsk1.r` der har følgende indhold:

```
Lokalitet Genotype Antal
Lolland AA 27
Lolland Aa 30
Lolland aa 12
Bornholm AA 14
Bornholm Aa 20
Bornholm aa 52
Ålandsøerne AA 0
Ålandsøerne Aa 5
Ålandsøerne aa 75
```

Man kan nu klare sig med nogle få R-kommandoer (det er ikke nogen fejl at der i kaldet af `glm` blandt andet står `family=poisson`):

```
Torsk <- read.table ("h:/bog/txt304ny/torsk.dat", nrows=10, header = TRUE)
require (stats) # indlæs pakken stats der definerer funktionen  xtabs
andet
```

tabel 14.1:

```
obs <- xtabs (Antal ~ Genotype + Lokalitet, data=Torsk)
rbind (obs, "i alt" = colSums (obs))
```

Så ser vi på hypotesen H_0 om ingen forskel på lokaliteter; i udskriften fra `glm` svarer `Likelihood Ratio` til $-2\ln Q$:

```
H0 <- glm (Antal ~ Genotype + Lokalitet, family=poisson, data=Torsk)
H0$deviance                          # -2lnQ
H0$df.res                            # antal frihedsgrader
1 - pchisq (H0$deviance, H0$df.res) # testsandsynligheden
```

Vi putter de fittede værdier ind i data.frame'n `Torsk`; så er det let at lave tabel 14.2 (de forventede værdier):

```
Torsk$Forventet <- H0$fitted
forv <- xtabs (Forventet ~ Genotype + Lokalitet, data=Torsk)
round ( forv, digits=1)
```

De estimerede sandsynligheder kan f.eks. udregnes sådan:

```
round (rowSums(forv)/sum(forv), digits=2)
```

14.4 Opgaver

Opgave 14.1 (Medarbejderaktier)
Det er blevet almindeligt at firmaer indfører ordninger med medarbejderaktier; derved skulle medarbejderne komme til at føle større medansvar og forpligtelse over for deres arbejdsplads. Det er dog ikke altid at firmaets opfordring til medarbejderne om at blive aktionærer opfattes

Tabel 14.3 Opgave 14.1: respondenternes fordeling på motiv og medarbejderkategori.

	arbejdere	funktionærer	mellemledere	topledere
for at bevare jobbet	77	25	11	8
som en investering	37	13	8	4
tror på idéen	35	14	7	11

på samme måde af alle medarbejdergrupper. For at danne sig et indtryk af medarbejderes motiver til at erhverve sig aktier har man foretaget et rundspørge blandt medarbejderne på en bestemt virksomhed (som har en medarbejderaktie-ordning) og bedt dem nævne deres motiver for at gå med i aktieordningen. Svarmulighederne var »for at bevare jobbet«, »som en investering« og »tror på idéen med medarbejderaktier«.

Hvad kan man på baggrund af svarfordelingen i tabel 14.1 sige om en eventuel sammenhæng mellem medarbejdernes motiver for at deltage i ordningen og arten af deres arbejde?

Opgave 14.2 (Test af simpel hypotese)
Antag at $(Y_1, Y_2, \ldots, Y_r)$ er multinomialfordelt med parametre n og $\boldsymbol{p}$, og lad

$$\boldsymbol{p}_0 = \begin{pmatrix} p_{01} \\ p_{02} \\ \vdots \\ p_{0r} \end{pmatrix}$$

være et sæt *kendte* ikke-negative tal der summerer til 1. Man ønsker at teste hypotesen $H_0 : \boldsymbol{p} = \boldsymbol{p}_0$ (eller altså $p_i = p_{0i}$ for alle i).

1. Udled $-2 \ln Q$-størrelsen for denne hypotese.
2. Der gælder at når H_0 er rigtig, så er $-2 \ln Q$ asymptotisk χ^2-fordelt med et antal frihedsgrader der kan udregnes som ændringen i antal frie parametre.
   Hvad er antallet af frihedsgrader for $-2 \ln Q$?

15 Tosidede kontingenstabeller

En af pointerne i kapitel 14 er at når man klassificerer et antal individer (fra en bestemt population) efter ét kriterium med r klasser $A_1, A_2, \ldots, A_r$, så kan det være fornuftigt at forsøge sig med en model der siger at hvis Y_i betegner antallet af A_i-individer i stikprøven, $i = 1, 2, \ldots, r$, så er den r-dimensionale stokastiske variabel $(Y_1, Y_2, \ldots, Y_r)$ multinomialfordelt. I dette kapitel skal vi se hvorledes en bestemt art struktur i inddelingskriteriet kan afspejle sig i den statistiske model. Den pågældende struktur består i at der rent faktisk inddeles efter *to* kriterier på en gang.

Her er først en præsentation af det talmateriale der benyttes som gennemgående eksempel i dette kapitel.

Eksempel 15.1 (Hjernesvulstpatienter)
Man har klassificeret 141 hjernesvulstpatienter efter svulstens *art* (»godartet«, »ondartet« og »andet«) og *placering* i hjernevævet (»ved panden«, »ved tindingen« og »andre steder«). Resultaterne heraf fremgår af tabel 15.1. Man er interesseret i at finde ud af om disse tal tyder på at der er en sammenhæng mellem svulstens art og dens placering.

Man kan sige at man har klassificeret $n = 141$ patienter som hørende til én af ni forskellige klasser, og at man derfor ifølge overvejelserne i kapitel 14 kan betragte det observerede talsæt (23, 21, ..., 17) som en observation af en multinomialfordelt stokastisk variabel. Imidlertid kan man også tænke på situationen på den måde at patienterne er klassificeret efter to kriterier på én gang, hvor hvert kriterium har tre niveauer.

15.1 Grundmodellen

Antag at vi har klassificeret n individer efter *to* kriterier. Det første kriterium har r niveauer og klasserne $A_1, A_2, \ldots, A_r$, og det andet kriterium har s niveauer og klasserne $B_1, B_2, \ldots, B_s$. Skematisk ser det sådan ud:

		kriterium 2				
	klasse	B_1	B_2	...	B_s	sum
kriterium 1	A_1	y_{11}	y_{12}	...	y_{1s}	$y_{1\bullet}$
	A_2	y_{21}	y_{22}	...	y_{2s}	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	y_{r1}	y_{r2}	...	y_{rs}	$y_{r\bullet}$
	sum	$y_{\bullet 1}$	$y_{\bullet 2}$	...	$y_{\bullet s}$	n

Tabel 15.1 141 hjernesvulstpatienter fordelt efter svulstens art og placering.

		placering			
		pande	tinding	andet	sum
art	godartet	23	21	34	78
	ondartet	9	4	24	37
	andet	6	3	17	26
	sum	38	28	75	141

hvor

$$
\begin{aligned}
y_{ij} &= \text{ antal individer i klassen } A_iB_j \ (= A_i \cap B_j),\\
y_{i\bullet} &= \sum_{j=1}^{s} y_{ij} = \text{ antal individer i klassen } A_i,\\
y_{\bullet j} &= \sum_{i=1}^{r} y_{ij} = \text{ antal individer i klassen } B_j.
\end{aligned}
$$

Da der er tale om at et antal individer er klassificeret i et antal klasser, benytter vi som grundmodel en *multinomialfordelingsmodel*: Den rs-dimensionale observation $\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{rs} \end{pmatrix}$ er en observeret værdi af en rs-dimensional stokastisk variabel $\boldsymbol{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{rs} \end{pmatrix}$ som er multinomialfordelt med antalsparameter n og sandsynlighedsparameter $\boldsymbol{p} = \begin{pmatrix} p_{11} \\ p_{12} \\ \vdots \\ p_{rs} \end{pmatrix}$.

Størrelsen p_{ij} er sandsynligheden for at et individ udvalgt tilfældigt fra »populationen« vil tilhøre klassen A_iB_j, og den estimeres ved

$$\widehat{p}_{ij} = y_{ij}/n \ . \tag{15.1}$$

15.2 Uafhængighedshypotesen

Den struktur der er i inddelingskriteriet (nemlig at der inddeles efter to kriterier på en gang), har foreløbig kun givet sig udslag i den måde de variable og parametrene er navngivet på (med index ij). Vi vil nu formulere en model der svarer til at der ikke er nogen sammenhæng mellem de to inddelingskriterier.

Den »sammenhæng« der kan være tale om, er ikke en årsagssammenhæng, men en statistisk sammenhæng. At der ikke er nogen sammenhæng mellem kriterium A og kriterium B, skal betyde at A og B i en vis forstand »virker« uafhængigt af hinanden, således at forstå at en oplysning om hvilken B-klasse et individ tilhører, ikke indeholder nogen information om hvilken A-klasse individet tilhører, og omvendt. Det skal nu formaliseres i en matematisk model.

Vi indfører nogle hjælpevariable $\boldsymbol{X}_d = (X_{dA}, X_{dB})$, således at X_{dA} er navnet på den A-klasse som individ nr. d tilhører, og X_{dB} er navnet på den B-klasse som individ nr. d tilhører, det vil sige at

$$\boldsymbol{X}_d = (A_i, B_j) \text{ betyder: individ nr. } d \text{ tilhører } A\text{-klassen } A_i \text{ og } B\text{-klassen } B_j.$$

At der ikke er nogen sammenhæng mellem A og B betyder hermed at en oplysning om værdien af X_{dB} ikke indeholder nogen information om værdien af X_{dA} (og omvendt), og det betyder at *de stokastiske variable X_{dA} og X_{dB} er stokastisk uafhængige*, dvs.

$$\mathrm{P}(X_{dA} = A_i \text{ og } X_{dB} = B_j) = \mathrm{P}(X_{dA} = A_i) \cdot \mathrm{P}(X_{dB} = B_j).$$

Nu er pr. definition $\mathrm{P}(X_{dA} = A_i,\ X_{dB} = B_j) = p_{ij}$, så at der ikke er nogen sammenhæng mellem A og B betyder altså at $p_{ij} = \alpha_i\,\beta_j$ hvor vi har sat $\alpha_i = \mathrm{P}(X_{dA} = A_i)$ og $\beta_j = \mathrm{P}(X_{dB} = B_j)$. Sammenfattende kan vi derfor sige at den matematiske formulering af antagelsen om at der ikke er nogen (statistisk) sammenhæng mellem kriterierne A og B, bliver at

$$p_{ij} = \alpha_i\,\beta_j$$

for alle i og j, hvor $\alpha_1, \alpha_2, \ldots, \alpha_r$ er ikke-negative tal der summerer til 1, og $\beta_1, \beta_2, \ldots, \beta_s$ er ikke-negative tal der summerer til 1. Udtrykt i ord går antagelsen ud på at sandsynligheden p_{ij} for på én gang at tilhøre både A_i og B_j er lig produktet af sandsynligheden α_i for at tilhøre A_i og sandsynligheden β_j for at tilhøre B_j.

I stedet for at tale om at der ikke er nogen sammenhæng mellem A og B, taler man ofte om at der er *uafhængighed* mellem A og B, og den statistiske hypotese

$$H_0 : p_{ij} = \alpha_i\beta_j \quad \text{for alle } i \text{ og } j,$$

hvor de ukendte parametre $(\alpha_1, \alpha_2, \ldots, \alpha_r)$ og $(\beta_1, \beta_2, \ldots, \beta_s)$ er ikke-negative talsæt der hver især summerer til 1, hedder da *uafhængighedshypotesen*.

At der er uafhængighed mellem A og B, udtrykker man undertiden på den måde at der ikke er nogen (signifikant) *vekselvirkning* mellem A og B. Når der ikke er nogen vekselvirkning mellem A og B, beskrives hele den *systematiske variation* i talmaterialet af de såkaldte *rækkevirkninger* (A-virkninger) $\alpha_1, \alpha_2, \ldots, \alpha_r$ der beskriver den systematiske forskel mellem rækker, og af de såkaldte *søjlevirkninger* (B-virkninger) $\beta_1, \beta_2, \ldots, \beta_s$ der beskriver den systematiske forskel mellem søjler.

Estimation af parametrene

Likelihoodfunktionen i grundmodellen er en almindelig multinomial-likelihoodfunktion:

$$L(\boldsymbol{p}) = \text{konstant} \cdot \prod_{i=1}^{r} \prod_{j=1}^{s} p_{ij}^{y_{ij}}$$

Tabel 15.2 Estimaterne over grundmodellens parametre p_{ij} og uafhængighedsmodellens parametre α_i og β_j i hjernesvulsteksemplet. Tallene er sandsynligheder i procent.

		placering			sum =
		pande	tinding	andet	$\widehat{\alpha}_i$
art	godartet	14.9	11.0	29.4	55.3
	ondartet	7.1	5.2	14.0	26.2
	andet	5.0	3.7	9.8	18.4
sum = $\widehat{\beta}_j$		27.0	19.9	53.2	100.0

hvor konstanten er en multinomialkoefficient.

Estimaterne over parametrene $\alpha_1, \alpha_2, \ldots, \alpha_r$ og $\beta_1, \beta_2, \ldots, \beta_s$ i uafhængighedsmodellen er de værdier der maksimaliserer $L(\boldsymbol{p})$ hvor man for p_{ij} indsætter $p_{ij} = \alpha_i\beta_j$, dvs. de værdier der maksimaliserer

$$\begin{aligned} &L_0(\alpha_1, \alpha_2, \ldots, \alpha_r, \beta_1, \beta_2, \ldots, \beta_s) \\ &\quad = \text{konstant} \cdot \prod_{i=1}^{r}\prod_{j=1}^{s}(\alpha_i\beta_j)^{y_{ij}} \\ &\quad = \text{konstant} \cdot \prod_{i=1}^{r}\prod_{j=1}^{s}\alpha_i^{y_{ij}} \cdot \prod_{i=1}^{r}\prod_{j=1}^{s}\beta_j^{y_{ij}} \\ &\quad = \text{konstant} \cdot \prod_{i=1}^{r}\alpha_i^{y_{i\bullet}} \cdot \prod_{j=1}^{s}\beta_j^{y_{\bullet j}} \ . \end{aligned}$$

Det ses at L_0 er et produkt af en funktion af α-erne og en funktion af β-erne. Ifølge sætning 14.1 antager disse to funktioner deres maksimumsværdier i hhv.

$$(\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_r) = \left(\frac{y_{1\bullet}}{n}, \frac{y_{2\bullet}}{n}, \ldots, \frac{y_{r\bullet}}{n}\right) \tag{15.2}$$

og

$$(\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_s) = \left(\frac{y_{\bullet 1}}{n}, \frac{y_{\bullet 2}}{n}, \ldots, \frac{y_{\bullet s}}{n}\right). \tag{15.3}$$

Dette er så maksimaliseringsestimaterne for parametrene. Resultatet er i øvrigt hvad man umiddelbart skulle forvente, idet f.eks. sandsynligheden α_i for at tilhøre A-klassen A_i estimeres ved den observerede relative hyppighed $y_{i\bullet}/n$ af A_i.

I taleksemplet bliver $L = \text{konstant} \cdot p_{11}^{23}\, p_{12}^{21}\, p_{13}^{34}\, p_{21}^{9}\, p_{22}^{4}\, p_{23}^{24}\, p_{31}^{6}\, p_{32}^{3}\, p_{33}^{17}$. Ved at indsætte de aktuelle talværdier i (15.1), (15.2) og (15.3) fås estimaterne over de ukendte parametre, se tabel 15.2.

Test for uafhængighed

Teststørrelsen for uafhængighedshypotesen H_0 er likelihoodkvotientstørrelsen Q eller $-2\ln Q$. Når man indsætter de fundne estimater i udtrykket for Q, får man

$$\begin{aligned} Q &= \frac{L_0(\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_r, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_s)}{L(\widehat{p}_{11}, \widehat{p}_{12}, \ldots, \widehat{p}_{rs})} \\ &= \frac{\prod_{i=1}^{r}\prod_{j=1}^{s}(\widehat{\alpha}_i\widehat{\beta}_j)^{y_{ij}}}{\prod_{i=1}^{r}\prod_{j=1}^{s}(\widehat{p}_{ij})^{y_{ij}}} = \prod_{i=1}^{r}\prod_{j=1}^{s}\left(\frac{\widehat{\alpha}_i\widehat{\beta}_j}{\widehat{p}_{ij}}\right)^{y_{ij}} \\ &= \prod_{i=1}^{r}\prod_{j=1}^{s}\left(\frac{\widehat{y}_{ij}}{y_{ij}}\right)^{y_{ij}}, \end{aligned}$$

hvor $\widehat{y}_{ij} = n\widehat{\alpha}_i\widehat{\beta}_j = y_{i\cdot}y_{\cdot j}/n$ er det »forventede« antal individer i klassen A_iB_j under uafhængighedshypotesen. Dermed bliver

$$-2\ln Q = 2\sum_{i=1}^{r}\sum_{j=1}^{s} y_{ij}\ln\frac{y_{ij}}{\widehat{y}_{ij}}\ .$$

Værdier af $-2\ln Q$ tæt på 0 tyder på at H_0 giver en næsten lige så god beskrivelse af data som grundmodellen gør, hvorimod store $-2\ln Q$-værdier betyder at H_0 giver en væsentlig dårligere beskrivelse end grundmodellen gør, og i så fald vil man forkaste hypotesen om uafhængighed mellem rækker og søjler.

De »forventede« antal i hjernesvulsteksemplet er vist i tabel 15.3; herudfra fås

$$\begin{aligned} -2\ln Q_{\text{obs}} = 2\Bigg(&23\ln\frac{23}{21.0} + 21\ln\frac{21}{15.5} + 34\ln\frac{34}{41.5} \\ &+ 9\ln\frac{9}{10.0} + 4\ln\frac{4}{7.3} + 24\ln\frac{24}{19.7} \\ &+ 6\ln\frac{6}{7.0} + 3\ln\frac{3}{5.2} + 17\ln\frac{17}{13.8}\Bigg) \\ = 8.1& \end{aligned}$$

Når vi skal afgøre om en opnået $-2\ln Q_{\text{obs}}$-værdi (som f.eks. 8.1) er signifikant stor, skal vi sammenligne den med alle de andre $-2\ln Q$-værdier man også kunne have fået såfremt uafhængighedshypotesen H_0 var rigtig. Vi skal derfor bestemme *testsandsynligheden* ε, dvs. sandsynligheden for at få en større $-2\ln Q$-værdi end den observerede, under forudsætning af at H_0 er rigtig:

$$\varepsilon = \mathrm{P}_0\big(-2\ln Q \geq -2\ln Q_{\text{obs}}\big).$$

Når man skal bestemme ε, kan man udnytte en generel matematisk sætning der fortæller at når H_0 er rigtig, så er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med $(r-1)(s-1)$ frihedsgrader, således at ε med god tilnærmelse kan bestemmes som sandsynligheden for

Tabel 15.3 Den »forventede« fordeling af 141 hjernesvulstpatienter under forudsætning af uafhængighed mellem svulstens art og placering.

		placering			
		pande	tinding	andet	sum
art	godartet	21.0	15.5	41.5	78
	ondartet	10.0	7.3	19.7	37
	andet	7.0	5.2	13.8	26
	sum	38.0	28.0	75.0	141

at få en værdi større end $-2\ln Q_{\text{obs}}$ i en χ^2-fordeling med $(r-1)(s-1)$ frihedsgrader, kort

$$\varepsilon = \mathrm{P}\big(\chi^2_{(r-1)(s-1)} \geq -2\ln Q_{\text{obs}}\big).$$

Denne sandsynlighed er let at bestemme ved hjælp af tabeller over fraktiler i χ^2-fordelingen.

Antallet af frihedsgrader for $-2\ln Q$ findes som *ændringen i antallet af frie parametre*: i grundmodellen er der rs sandsynlighedsparametre der summerer til 1, dvs. der er $rs-1$ frie parametre; under H_0 er der r rækkeparametre der summerer til 1, samt s søjleparametre der summerer til 1, dvs. $(r-1)+(s-1)$ frie parametre; antallet af frihedsgrader for teststørrelsen er dermed

$$\big(rs-1\big) - \big((r-1)+(s-1)\big) = (r-1)(s-1).$$

Bemærk at χ^2-fordelingen kun er en approksimation; for at den skal kunne anvendes, kræves det at *alle de »forventede« antal være mindst fem*. – Hvis denne betingelse ikke er opfyldt, kan man eventuelt slå nogle rækker eller nogle søjler sammen.

I hjernesvulsteksemplet er de »forventede« antal over fem, så vi kan roligt anvende χ^2-approksimationen. Tabelopslag viser at i χ^2-fordelingen med $(3-1)(3-1)=4$ frihedsgrader er 90%-fraktilen 7.78 og 95%-fraktilen 9.49, således at teststørrelsen $-2\ln Q_{\text{obs}} = 8.1$ svarer til en testsandsynlighed på mellem 5% og 10%. På det grundlag vil man sædvanligvis *ikke* forkaste H_0. Det kan altså konkluderes at der tilsyneladende ikke er nogen sammenhæng mellem svulstens art og dens placering. Det vil blandt andet sige at man ikke ud fra kendskab til *placeringen* af en svulst kan sige noget om, hvorvidt den vil være godartet eller ej.

15.3 Jævnføring med andre tilsvarende modeller

Den læser der har studeret afsnit 14.2 om sammenligning af multinomialfordelinger, vil måske have bemærket, at de dér præsenterede metoder har store ligheder med dem i indeværende kapitel. Vi kan opregne nogle af lighederne:

1. Der foreligger nogle observerede *antal* y_{ij} anbragt i et tosidet skema.
2. Man udregner nogle »forventede« antal $\widehat{y}_{ij}$ efter opskriften *rækkesum* gange *søjlesum* divideret med *totalsum*.

3. Man udregner en teststørrelse $-2\ln Q_{\text{obs}} = \sum y \ln(y/\widehat{y})$.
4. Man sammenligner $-2\ln Q_{\text{obs}}$ med χ^2-fordelingen med $(r-1)(s-1)$ frihedsgrader.

Selv om man *foretager sig* det samme i de to tilfælde, er det imidlertid på grundlag af to forskellige modeller:

- I det ene tilfælde (dette kapitel) klassificerer man nogle individer efter *to* kriterier, og opgaven er da at undersøge om der er en sammenhæng mellem disse to kriterier.
- I det andet tilfælde (afsnit 14.2) er individerne på forhånd delt ind i nogle grupper inden de klassificeres efter *et* kriterium. Opgaven er da at undersøge om der er forskel på grupperne (med hensyn til hvordan gruppernes individer fordeles på klasserne).

Om man skal benytte den ene eller den anden model, er således et spørgsmål om hvorledes man har designet det forsøg der har leveret talmaterialet. I eksemplet i dette kapitel sagde vi at det handlede om at man havde taget 141 hjernesvulstpatienter og klassificeret dem efter *to* kriterier; derved blev det et eksempel der illustrerede dette kapitels model og metoder. Hvis det derimod havde handlet om at man havde taget 38 patienter med svulst i panden, 28 med svulst i tindingen og 75 hvor svulsten ikke var lokaliseret til pande eller tinding, og dernæst klassificeret disse patienter efter svulstens art, så havde det været et afsnit 14.2-eksempel.

De to modeller er nært beslægtede; hvis man i dette kapitels model betinger med søjlesummerne, dvs. betinger med at $Y_{\bullet 1} = n_1$, $Y_{\bullet 2} = n_2$, …, $Y_{\bullet s} = n_s$, så får man modellen i afsnit 14.2, og uafhængighedshypotesen overføres til afsnit 14.2's H_0.

15.4 Regn og tegn

Som nævnt i afsnit 15.3 er udregningerne i forbindelse med test for uafhængighed i en kontingenstabel de samme som udregningerne i forbindelse med sammenligning af multinomialfordelinger, så vi kunne her nøjes med at henvise til afsnit 14.3.

For god ordens skyld viser vi dog også et eksempel her, nemlig hjernesvulsteksemplet. Datafilen `h:/bog/txt304ny/svulst.dat` har følgende indhold:

```
art      placering antal
godartet pande      23
ondartet pande       9
andet    pande       6
godartet tinding    21
ondartet tinding     4
andet    tinding     3
godartet andet      34
ondartet andet      24
andet    andet      17
```

Selve R-koden kommer her:

```
Hjsvulst <- read.table ("h:/bog/txt304ny/svulst.dat", nrows=15, header = TRUE)
require (stats) # indlæs pakken stats der definerer funktionen  xtabs
```

```
obs <- xtabs (antal ~ art + placering , data=Hjsvulst)
obs  # dette er tabellen over observerede antal
```

Som teststørrelse kan man bruge Pearsons X^2 der er en approksimation til $-2\ln Q$ (se evt. opgave 3.5 side 48)

```
chisq.test (obs)   # Pearsons X^2 :
```

eller man kan få udregnet den rigtige $-2\ln Q$ f.eks. sådan her:

```
H0 <- glm (antal ~ art + placering - 1, family=poisson, data=Hjsvulst)
H0$deviance   # dette er -2lnQ
1 - pchisq (H0$deviance, H0$df.res) # dette er testsandsynligheden

Hjsvulst$forventet <- H0$fitted  # de forventede antal

forv <- xtabs (forventet ~ art + placering , data=Hjsvulst)
round (forv, digits=1)  # dette er tabellen over forventede antal
```

15.5 Opgaver

Opgave 15.1 (Hår- og øjenfarve)
Ved en sundhedsundersøgelse af 283 piger i St. Clement Street skole i Aberdeen blev hår- og øjenfarve observeret med et resultat som vist i nedenstående tabel. Viser dette materiale en sammenhæng mellem hårfarve og øjenfarve?

		Hårfarve			
		lys	rød	neutral	mørk
Øjenfarve	blå	30	4	27	6
	lys	30	5	28	11
	neutral	21	7	40	22
	mørk	6	3	23	20

16 Et større eksempel: Torsk i Østersøen

I DETTE KAPITEL vil vi tage et tidligere omtalt eksempel op til nærmere behandling. Eksemplet er blandt andet et eksempel på at man kan indbygge noget teori i den statistiske model, og et eksempel der viser nytten af maximum likelihood metoden til parameterestimation.

16.1 Præsentation af eksemplet

Den 6. marts 1961 fangede nogle havbiologer 69 torsk ved Lolland og undersøgte arten af blodets hæmoglobin i hver enkelt torsk. Senere på året fangede man desuden nogle torsk ved Bornholm og ved Ålandsøerne og bestemte deres genotype. (19)

Man mener at hæmoglobin-arten bestemmes af ét enkelt gen, og det som biologerne bestemte, var torskenes genotype for så vidt angår dette gen. Genet optræder i to udgaver som traditionelt kaldes A og a, og de mulige genotyper er da AA, Aa og aa. I tabel 16.1 på næste side ses den fundne genotypefordeling for hver af de tre lokaliteter.

På hver geografisk lokalitet er der sket det at man har klassificeret et antal torsk i tre mulige klasser, så på hver lokalitet er der tale om en multinomialfordelingssituation (når der er tre klasser, taler man også om en *tri*nomialfordeling). Som grundmodel benytter vi derfor den model der siger at de tre observations»vektorer«

$$\boldsymbol{y}_{\mathrm{L}} = \begin{pmatrix} y_{1\mathrm{L}} \\ y_{2\mathrm{L}} \\ y_{3\mathrm{L}} \end{pmatrix} = \begin{pmatrix} 27 \\ 30 \\ 12 \end{pmatrix}, \quad \boldsymbol{y}_{\mathrm{B}} = \begin{pmatrix} y_{1\mathrm{B}} \\ y_{2\mathrm{B}} \\ y_{3\mathrm{B}} \end{pmatrix} = \begin{pmatrix} 14 \\ 20 \\ 52 \end{pmatrix} \text{ og } \quad \boldsymbol{y}_{\mathrm{Å}} = \begin{pmatrix} y_{1\mathrm{Å}} \\ y_{2\mathrm{Å}} \\ y_{3\mathrm{Å}} \end{pmatrix} = \begin{pmatrix} 0 \\ 5 \\ 75 \end{pmatrix}$$

stammer fra hver sin multinomialfordeling med antalsparametre henholdsvis $n_{\mathrm{L}} = 69$, $n_{\mathrm{B}} = 86$ og $n_{\mathrm{Å}} = 80$ og med sandsynlighedsparametre henholdsvis

$$\boldsymbol{p}_{\mathrm{L}} = \begin{pmatrix} p_{1\mathrm{L}} \\ p_{2\mathrm{L}} \\ p_{3\mathrm{L}} \end{pmatrix}, \quad \boldsymbol{p}_{\mathrm{B}} = \begin{pmatrix} p_{1\mathrm{B}} \\ p_{2\mathrm{B}} \\ p_{3\mathrm{B}} \end{pmatrix}, \text{ og } \quad \boldsymbol{p}_{\mathrm{Å}} = \begin{pmatrix} p_{1\mathrm{Å}} \\ p_{2\mathrm{Å}} \\ p_{3\mathrm{Å}} \end{pmatrix}.$$

16.2 Hardy-Weinberg ligevægt

Grundmodellen er at hver geografisk lokalitet har sin egen multinomialfordeling, og at hver multinomialfordeling har en sandsynlighedsparameter

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

Tabel 16.1 (= tabel 14.1) Genotypefordeling af torsk fra tre lokaliteter i Østersøen.

genotype	Lolland	Bornholm	Ålandsøerne
AA	27	14	0
Aa	30	20	5
aa	12	52	75
i alt	69	86	80

hvor p_1, p_2 og p_3 kan være hvilkesomhelst tre ikke-negative tal der summerer til 1. Imidlertid kan man argumentere for at der under visse omstændigheder må være en bestemt sammenhæng mellem de tre p-er.

Lad os antage at i en bestemt torskegeneration optræder de tre genotyper AA, Aa og aa med hyppighederne p_1, p_2 og p_3 (hvor $p_1 + p_2 + p_3 = 1$). Lad os desuden antage at næste generation fremstilles ved »tilfældig parring« således at hvert af en torskeunges to hæmoglobin-gener vælges uafhængigt af hinanden på følgende måde: først vælges et tilfældigt forældre-individ, dernæst vælges et tilfældigt af dette individs hæmoglobin-gener. Sandsynligheden for at vælge A er da $p_1 + {}^1\!/_2\, p_2$ hvilket vi kalder β, og sandsynligheden for at vælge a er ${}^1\!/_2\, p_2 + p_3 = 1 - \beta$. I den nye generation bliver genotypefordelingen derfor

$$\begin{aligned} \text{AA:} &\quad \beta^2 \\ \text{Aa:} &\quad 2\beta(1-\beta) \\ \text{aa:} &\quad (1-\beta)^2. \end{aligned}$$

(Man kan notere at de tre sandsynligheder summerer til 1: $\beta^2 + 2\beta(1-\beta) + (1-\beta)^2 = \big(\beta + (1-\beta)\big)^2 = 1$). Det ses at genotypefordelingen i den nye generation ikke kan være hvadsomhelst, men at der er en vis sammenhæng mellem de tre sandsynligheder, styret af størrelsen β. Vi kan prøve at se hvad der sker hvis der er en tilsvarende sammenhæng mellem sandsynlighederne i forældregenerationen. Lad os sige at i forældregenerationen er fordelingen

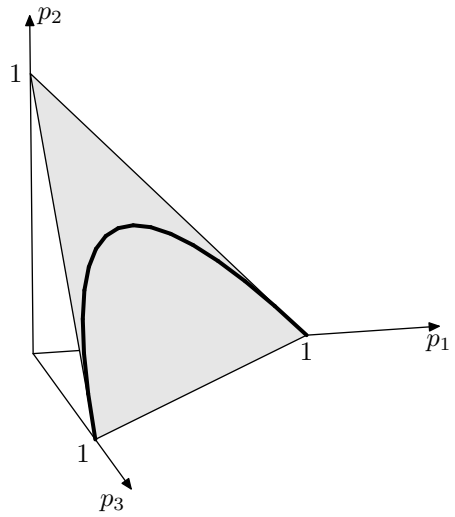
$$\begin{aligned} \text{AA:} &\quad p_1 = \alpha^2 \\ \text{Aa:} &\quad p_2 = 2\alpha(1-\alpha) \\ \text{aa:} &\quad p_3 = (1-\alpha)^2. \end{aligned}$$

Så bliver $\beta = p_1 + {}^1\!/_2\, p_2 = \alpha^2 + {}^1\!/_2\, 2\alpha(1-\alpha) = \alpha$, dvs. sandsynlighederne er uforandrede fra den ene generation til den anden.

Man siger at populationen er i *Hardy-Weinberg ligevægt* hvis det er sådan at de tre genotyper optræder i forholdet

$$\begin{aligned} \text{AA:} &\quad p_1 = \beta^2 \\ \text{Aa:} &\quad p_2 = 2\beta(1-\beta) \\ \text{aa:} &\quad p_3 = (1-\beta)^2 \end{aligned}$$

for en eller anden værdi af $\beta \in [0,1]$. Hvis der er Hardy-Weinberg ligevægt, er det altså kun visse sandsynlighedstripler (p_1, p_2, p_3) der kan komme på tale, se figur 16.1.



Figur 16.1 Det tonede område er sandsynlighedssimplexet, dvs. mængden af tripler $\boldsymbol{p} = (p_1, p_2, p_3)$ af ikke-negative tal der summerer til 1. Kurven består af de $\boldsymbol{p}$-er der kan optræde hvis der er Hardy-Weinberg ligevægt.

16.3 Hypotesen om Hardy-Weinberg ligevægt

Vi vil undersøge om der er Hardy-Weinberg ligevægt på hver af de tre lokaliteter. Vi begynder med Lolland.

At der er Hardy-Weinberg ligevægt ved Lolland kan formuleres som den statistiske hypotese

$$H_{\mathrm{L}} : \begin{pmatrix} p_{1\mathrm{L}} \\ p_{2\mathrm{L}} \\ p_{3\mathrm{L}} \end{pmatrix} = \begin{pmatrix} \beta_{\mathrm{L}}^2 \\ 2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}) \\ (1-\beta_{\mathrm{L}})^2 \end{pmatrix}.$$

I grundmodellen er likelihoodfunktionen $L(p_{1\mathrm{L}}, p_{2\mathrm{L}}, p_{3\mathrm{L}}) = \text{konstant} \cdot p_{1\mathrm{L}}^{27}\, p_{2\mathrm{L}}^{30}\, p_{3\mathrm{L}}^{12}$, der har maksimum i $\widehat{\boldsymbol{p}}_L = \begin{pmatrix} 27/69 \\ 30/69 \\ 12/69 \end{pmatrix}$. Under H_{L} er likelihoodfunktionen

$$\begin{aligned} L_{\mathrm{L}}(\beta_L) &= L\big(\beta_{\mathrm{L}}^2, 2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}), (1-\beta_{\mathrm{L}})^2\big) \\ &= \text{konstant} \cdot \big(\beta_{\mathrm{L}}^2\big)^{27} \big(2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}})\big)^{30} \big((1-\beta_{\mathrm{L}})^2\big)^{12} \\ &= \text{konstant} \cdot \beta_{\mathrm{L}}^{2\cdot 27+30} (1-\beta_{\mathrm{L}})^{30+2\cdot 12}, \end{aligned}$$

som har maksimum i $\widehat{\beta}_{\mathrm{L}} = \dfrac{2\cdot 27+30}{2\cdot 69} = \dfrac{84}{138} = 0.609$, dvs. $\widehat{\beta}_{\mathrm{L}}$ er det observerede antal A divideret med det samlede antal gener.

Man tester hypotesen ved brug af den sædvanlige kvotientteststørrelse Q som er $L\big(\widehat{\beta}_{\mathrm{L}}^2, 2\widehat{\beta}_{\mathrm{L}}(1-\widehat{\beta}_{\mathrm{L}}), (1-\widehat{\beta}_{\mathrm{L}})^2\big) / L\big(\widehat{p}_{1\mathrm{L}}, \widehat{p}_{2\mathrm{L}}, \widehat{p}_{3\mathrm{L}}\big)$ eller $-2\ln Q$; sidstnævnte kan udtrykkes

Tabel 16.2 Forventede antal $\widehat{y}$ under forudsætning af Hardy-Weinberg ligevægt på hver lokalitet.

genotype	Lolland	Bornholm	Ålandsøerne
AA	25.6	6.7	0.1
Aa	32.9	34.6	4.8
aa	10.6	44.7	75.1
i alt	69	86	80

som

$$-2\ln Q = 2\sum_{i=1}^{r} y_i \ln \frac{y_i}{\widehat{y}_i}$$

hvor $(\widehat{y}_1, \widehat{y}_2, \widehat{y}_3) = \left(n_{\mathrm{L}}\widehat{\beta}_{\mathrm{L}}^2, n_{\mathrm{L}}2\widehat{\beta}_{\mathrm{L}}(1-\widehat{\beta}_{\mathrm{L}}), n_{\mathrm{L}}(1-\widehat{\beta}_{\mathrm{L}})^2\right)$ er de »forventede« antal under H_{L}.

Man finder at $-2\ln Q = 0.52$ med $(3-1)-1 = 1$ frihedgrader, svarende til en testsandsynlighed på ca. 47%, så man kan sagtens antage at torskebestanden ved Lolland er i Hardy-Weinberg ligevægt.

Noget tilsvarende kan gøres med de to andre lokaliteter. Man får maksimaliseringsestimaterne $\widehat{\beta}_{\mathrm{B}} = 0.279$ og $\widehat{\beta}_{\mathrm{Å}} = 0.031$. De forventede antal $\widehat{y}$ ses i tabel 16.2. Ved Ålandsøerne kan man oplagt antage Hardy-Weinberg ligevægt. (Man kan ikke benytte χ^2-approksimationen til $-2\ln Q$ fordi et af de forventede antal er alt for lille. Til gengæld reproducerer modellen jo observationerne særdeles fint.) Ved Bornholm er der større uoverensstemmelse mellem de observerede og de forventede antal, og teststørrelsen er her $-2\ln Q = 14.4$, svarende til en testsandsynlighed af størrelsesorden 10^{-4}.

16.4 En samlet model

Man kan sige at hypotesen om Hardy-Weinberg ligevægt er sådan en »pæn« hypotese fordi man kan »forstå« (dvs. levere en simpel forklaring på) den. Derfor er det ærgerligt at Bornholm tilsyneladende falder uden for det pæne billede. For at reparere på tingene kunne man forsøge sig med en modificeret hypotese H_1 gående ud på at

- ved Lolland er der Hardy-Weinberg ligevægt med parameter β_{L},
- ved Ålandsøerne er der Hardy-Weinberg ligevægt med parameter $\beta_{\mathrm{Å}}$,
- ved Bornholm er populationen en blanding af Lollandstorsk og Ålandstorsk i forholdet $\alpha : (1-\alpha)$ hvor $\alpha \in]0,1[$ er en ukendt parameter.

Mere præcist går H_1 altså ud på at der findes værdier af β_{L}, $\beta_{\mathrm{Å}}$ og α så

Tabel 16.3 Forventede antal $\widehat{\widehat{y}}$ i blandingsmodellen.

genotype	Lolland	Bornholm	Ålandsøerne
AA	25.7	13.7	0.1
Aa	32.8	20.3	4.8
aa	10.4	52.0	75.1
i alt	69	86	80

$$\boldsymbol{p}_{\mathrm{L}} = \begin{pmatrix} \beta_{\mathrm{L}}^2 \\ 2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}) \\ (1-\beta_{\mathrm{L}})^2 \end{pmatrix},$$

$$\boldsymbol{p}_{\mathrm{Å}} = \begin{pmatrix} \beta_{\mathrm{Å}}^2 \\ 2\beta_{\mathrm{Å}}(1-\beta_{\mathrm{Å}}) \\ (1-\beta_{\mathrm{Å}})^2 \end{pmatrix},$$

$$\boldsymbol{p}_{\mathrm{B}} = \alpha\boldsymbol{p}_{\mathrm{L}} + (1-\alpha)\boldsymbol{p}_{\mathrm{Å}} = \begin{pmatrix} \alpha\beta_{\mathrm{L}}^2 + (1-\alpha)\beta_{\mathrm{Å}}^2 \\ \alpha 2\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}) + (1-\alpha)2\beta_{\mathrm{Å}}(1-\beta_{\mathrm{Å}}) \\ \alpha(1-\beta_{\mathrm{L}})^2 + (1-\alpha)(1-\beta_{\mathrm{Å}})^2 \end{pmatrix}.$$

Bemærk at der nu er tale om én samlet model for alle tre lokaliteter.

Den samlede likelihoodfunktion bliver produktet af de tre del-likelihoodfunktioner for de tre lokaliteter. Det er bekvemt at operere med *logaritmen* til likelihoodfunktionen, så den skriver vi op:

$$\begin{aligned}
\ln L(\beta_{\mathrm{L}}, \beta_{\mathrm{Å}}, \alpha) &= 27\ln p_{1\mathrm{L}} + 30\ln p_{2\mathrm{L}} + 12\ln p_{3\mathrm{L}} \\
&\quad + 14\ln p_{1\mathrm{B}} + 20\ln p_{2\mathrm{B}} + 52\ln p_{3\mathrm{B}} \\
&\quad + 0\ln p_{1\mathrm{Å}} + 5\ln p_{2\mathrm{Å}} + 75\ln p_{3\mathrm{Å}} \\
&= \text{konstant} + 84\ln\beta_{\mathrm{L}} + 54\ln(1-\beta_{\mathrm{L}}) \\
&\quad + 14\ln\bigl(\alpha\beta_{\mathrm{L}}^2 + (1-\alpha)\beta_{\mathrm{Å}}^2\bigr) \\
&\quad + 20\ln\bigl(\alpha\beta_{\mathrm{L}}(1-\beta_{\mathrm{L}}) + (1-\alpha)\beta_{\mathrm{Å}}(1-\beta_{\mathrm{Å}})\bigr) \\
&\quad + 52\ln\bigl(\alpha(1-\beta_{\mathrm{L}})^2 + (1-\alpha)(1-\beta_{\mathrm{Å}})^2\bigr) \\
&\quad + 5\ln\beta_{\mathrm{Å}} + 155\ln(1-\beta_{\mathrm{Å}}).
\end{aligned}$$

Der synes ikke at være nogen praktisk anvendelig analytisk måde at maksimalisere denne funktion på, så man må benytte en iterationsmetode. Som startværdier til en sådan kan vi benytte de tidligere fundne estimater $\widehat{\beta}_{\mathrm{L}} = 0.609$ og $\widehat{\beta}_{\mathrm{Å}} = 0.031$ og vælge α så det forventede antal Aa ved Bornholm er lig det observerede, dvs. ved at løse ligningen $\alpha \cdot 2\widehat{\beta}_{\mathrm{L}}(1-\widehat{\beta}_{\mathrm{L}}) + (1-\alpha)\cdot 2\widehat{\beta}_{\mathrm{Å}}(1-\widehat{\beta}_{\mathrm{Å}}) = 20/86$, hvilket giver $\alpha \approx 0.414$.

Man finder at $\ln L$ antager sit maksimum i $(\widehat{\widehat{\beta}}_{\mathrm{L}}, \widehat{\widehat{\beta}}_{\mathrm{Å}}, \widehat{\widehat{\alpha}}) = (0.611, 0.031, 0.425)$. Herefter kan vi udregne den forventede genotypefordeling de tre steder, se tabel 16.3. Det ses at der er langt bedre overensstemmelse mellem de observerede og de »forventede«

værdier i denne model. Hvis man tester modellen i forhold til grundmodellen med en vilkårlig trinomialfordeling hvert sted, får man en $-2\ln Q$-størrelse på 0.7, og selv om de forventede antal ikke alle er mindst 5, kan man jo alligevel godt skæve til χ^2-fordelingen med $3 \cdot (3-1) - 3 = 3$ frihedsgrader.

Alt i alt må man konkludere, at modellen med Hardy-Weinberg ligevægt ved Lolland og ved Ålandsøerne og med en blandingspopulation ved Bornholm giver en god beskrivelse af de foreliggende observationer.

16.5 Regn og tegn

Her ses hvordan man kan udføre beregningerne til dette kapitel.

De simple modeller

```
# de observerede antal
yL <- c(27, 30, 12)
yB <- c(14, 20, 52)
yA <- c(0, 5, 75)
y <- cbind(yL, yB, yA)
```

Vi definerer nu en funktion `betahat` der ud fra en observation $\boldsymbol{y}$ udregner $\widehat{\beta} = \frac{2y_1+y_2}{2y_\bullet}$ (det forudsættes at `y` er en vektor af længde 3). Deruden definerer vi en funktion `p.beta` der udregner $\boldsymbol{p}$ som funktion af β:

```
betahat <- function(y){ (2*y[1] + y[2])/(2*sum(y)) }

p.beta <- function(b) { c(b^2, 2*b*(1-b), (1-b)^2) }
```

Ved hjælp af disse funktioner udregnes de enkelte $\widehat{\beta}$'er og de tilsvarende vektorer $(\widehat{y}_1, \widehat{y}_2, \widehat{y}_3)$ af observerede antal (jf. tabel 16.2):

```
betaL <- betahat (yL)
betaB <- betahat (yB)
betaA <- betahat (yA)

yhatL <- p.beta(betaL) * sum(yL)
yhatB <- p.beta(betaB) * sum(yB)
yhatA <- p.beta(betaA) * sum(yA)
```

Så defineres en funktion `minus2lnQ` der udregner $-2\ln Q$ (i summen skal man se bort fra led hvor det observerede antal er 0 idet $0\ln 0$ er lig med 0). Funktionen bruges til at udregne $-2\ln Q$ på hver af de tre lokaliteter:

```
minus2lnQ <- function (obs.antal, hyp.antal) {
  2*sum (ifelse (obs.antal>0, obs.antal*log(obs.antal/hyp.antal), 0))
}

minus2lnQ (yL, yhatL)
minus2lnQ (yB, yhatB)
minus2lnQ (yA, yhatA)
```

Den store model

Vi definerer en funktion `fkt` som (pånær et konstantled) er $-\ln L(\beta_{\mathrm{L}}, \beta_{\mathrm{Å}}, \alpha)$. Denne funktion skal minimaliseres. – Bemærk at funktionen skal defineres som en funktion af én tredimensional variabel (som er kaldt `b`).

```
fkt <- function (b){
  -( 84*log( b[1]) + 54*log(1-b[1])
    +14*log( b[3]*b[1]^2 + (1-b[3])*b[2]^2     )
    +20*log( b[3]*b[1]*(1-b[1]) + (1-b[3])*b[2]*(1-b[2]))
    +52*log( b[3]*(1-b[1])^2  + (1-b[3])*(1-b[2])^2 )
    + 5*log( b[2]) + 155*log(1-b[2]))
}
```

Vi prøver nu to forskellige R-funktioner til optimering, nemlig `nlm` og `optim`; de skulle gerne give det samme!

(For at øge forvirringen er det sådan at `nlm` skal have funktionen som første argument og vektoren af startværdier som andet argument, og `optim` skal have dem i modsat rækkefølge...)

```
metode1 <- nlm (fkt, c(0.609, 0.031, 0.414))
metode2 <- optim (c(0.609, 0.031, 0.414), fkt)
```

Vi vælger at bruge estimaterne fra metode 2.

```
# De estimerede sandsynlighedsvektorer er
phathatL <- p.beta(metode2$par[1])
phathatA <- p.beta(metode2$par[2])
phathatB <- metode2$par[3]*phathatL + (1-metode2$par[3])*phathatA

# De forventede antal er
yhathatL <- phathatL*sum(yL)
yhathatB <- phathatB*sum(yB)
yhathatA <- phathatA*sum(yA)

# de forventede antal
yhathat <- cbind (yhathatL, yhathatB, yhathatA)
yhathat

# En -2lnQ-størrelse for den samlede model (df=3*(3-1)-3=3)
minus2lnQ (y, yhathat)
```

Referencer

[1] Allison, T. & Cicchetti, D. V. (1976). Sleep in mammals: ecological and constitutional correlates, *Science* **194**: 732–34.

[2] Andersen, E. B. (1977). Multiplicative poisson models with unequal cell rates, *Scandinavian Journal of Statistics* **4**: 153–8.

[3] Anscombe, F. J. (1973). Graphs in statistical analysis, *The American Statistician* **27**: 17–21.

[4] Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen*, Teubner, Leipzig.

[5] Davin, E. P. (1975). *Blood pressure among residents of the tambo valley*, Master's thesis, The Pennsylvania State University.

[6] Forbes, J. D. (1857). Further experiments and remarks on the measurement of heights by the boiling point of water, *Transactions of the Royal Society of Edingburgh* **21**: 135–43.

[7] Greenfort, A., Jensen, C. S. F. & Jeppesen, S. (1987). *Planter og planter imellem*, Biologispeciale, Roskilde Universitetscenter.

[8] Greenwood, M. & Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents, *Journal of the Royal Statistical Society* **83**: 255–79.

[9] Hald, A. (1948, 1968). *Statistiske Metoder*, Akademisk Forlag, København.

[10] Hinton, M. (1968). Doctoral thesis, Teachers College, Columbia University.

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

[12] Meillier, L. & Toldbod, I. (1985). *På skærmen står et lille hjerte og banker ... ultralyd og biologiske skadevirkninger – afprøvet for kromosombrud i mikrokernetesten*, Biologispeciale, Roskilde Universitetscenter.

[13] Meyer, H. A. (1953). *Forest Mensuration*, Penns Valley Publishers, Inc., State College, Pennsylvania.

[14] Newcomb, S. (1891). Measures of the velocity of light made under the direction of the Secretary of the Navy during the years 1880-1882, *Astronomical Papers* **2**: 107–230.

[15] Pack, S. E. & Morgan, B. J. T. (1990). A mixture model for interval-censored time-to-response quantal assay data, *Biometrics* **42**: 749–57.

[16] Rutherford, E. & Geiger, H. (1910). The probability variations in the distribution of α particles, *Philosophical Magazine* **xx**: 698–707.

[17] Rutherford, E. & Geiger, H. (1963). The probability variations in the distribution of α particles, *in* J. Chadwick (ed.), *The Collected Papers of Lord Rutherford of Nelson*, Vol. 2, George Allen and Unwin, London, pp. 203–11.

[18] Ryan, T. A., Joiner, B. L. & Ryan, B. F. (1976). *MINITAB Student Handbook*, Duxbury Press, North Scituate, Massachusetts.

[19] Sick, K. (1965). Haemoglobin polymorphism of cod in the baltic and the danish belt sea, *Hereditas* **54**: 19–48.

[20] Stigler, S. M. (1977). Do robust estimators work with *real* data?, *The Annals of Statistics* **5**: 1055–98.

[21] 'Student' (1908). The probable error of a mean, *Biometrika* **6**: 1–25.

[22] Weisberg, S. (1980). *Applied Linear Regression*, Wiley series in probability and mathematical statistics, John Wiley & Sons.

Tabeller

Det gælder for de allerfleste af de fordelinger som den praktisk arbejdende statistiker benytter, at hverken fordelingsfunktionen eller den inverse fordelingsfunktion (der leverer fraktilerne i fordelingen) er lette at beregne når hjælpemidlerne er papir og blyant og »almindelige« matematiske funktioner (så som addition, multiplikation, division, kvadratrod, kvadrering, logaritmefunktion, eksponentialfunktion osv.). I tidligere tider var det et betydeligt regnearbejde at udregne pålidelige numeriske approksimationer til de almindelige fordelinger og deres fraktiler, og man nøjedes derfor med at udregne funktionsværdierne for udvalgte værdier af argumenterne (her er en af forklaringerne på de magiske fem procent!), og statistiske tabeller var noget meget dyrebart (og copyright-belagt). – I vore dage er det anderledes. Enhver kan nu på en almindelig pc'er på ingen tid udregne de almindeligt brugte fordelingsfunktioner og fraktiler med stor præcision.

Tabellerne på de følgende sider er udregnet ved brug af Tue Tjurs unit `distr` til Turbo Pascal (`http://www.mes.cbs.dk/~sttt/`).

Fraktiler i χ^2-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	50	90	95	97.5	99	99.5	99.9
1	0.45	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	13.36	15.51	17.53	20.09	21.95	26.12
9	8.34	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	17.28	19.68	21.92	24.72	26.76	31.26
12	11.34	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	28.41	31.41	34.17	37.57	40.00	45.31
21	20.34	29.62	32.67	35.48	38.93	41.40	46.80
22	21.34	30.81	33.92	36.78	40.29	42.80	48.27
23	22.34	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	40.26	43.77	46.98	50.89	53.67	59.70
31	30.34	41.42	44.99	48.23	52.19	55.00	61.10
32	31.34	42.58	46.19	49.48	53.49	56.33	62.49
33	32.34	43.75	47.40	50.73	54.78	57.65	63.87
34	33.34	44.90	48.60	51.97	56.06	58.96	65.25
35	34.34	46.06	49.80	53.20	57.34	60.27	66.62
36	35.34	47.21	51.00	54.44	58.62	61.58	67.99
37	36.34	48.36	52.19	55.67	59.89	62.88	69.35
38	37.34	49.51	53.38	56.90	61.16	64.18	70.70
39	38.34	50.66	54.57	58.12	62.43	65.48	72.05
40	39.34	51.81	55.76	59.34	63.69	66.77	73.40

Fraktiler i χ^2-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	50	90	95	97.5	99	99.5	99.9
41	40.34	52.95	56.94	60.56	64.95	68.05	74.74
42	41.34	54.09	58.12	61.78	66.21	69.34	76.08
43	42.34	55.23	59.30	62.99	67.46	70.62	77.42
44	43.34	56.37	60.48	64.20	68.71	71.89	78.75
45	44.34	57.51	61.66	65.41	69.96	73.17	80.08
46	45.34	58.64	62.83	66.62	71.20	74.44	81.40
47	46.34	59.77	64.00	67.82	72.44	75.70	82.72
48	47.34	60.91	65.17	69.02	73.68	76.97	84.04
49	48.33	62.04	66.34	70.22	74.92	78.23	85.35
50	49.33	63.17	67.50	71.42	76.15	79.49	86.66
51	50.33	64.30	68.67	72.62	77.39	80.75	87.97
52	51.33	65.42	69.83	73.81	78.62	82.00	89.27
53	52.33	66.55	70.99	75.00	79.84	83.25	90.57
54	53.33	67.67	72.15	76.19	81.07	84.50	91.87
55	54.33	68.80	73.31	77.38	82.29	85.75	93.17
56	55.33	69.92	74.47	78.57	83.51	86.99	94.46
57	56.33	71.04	75.62	79.75	84.73	88.24	95.75
58	57.33	72.16	76.78	80.94	85.95	89.48	97.04
59	58.33	73.28	77.93	82.12	87.17	90.72	98.32
60	59.33	74.40	79.08	83.30	88.38	91.95	99.61
61	60.33	75.51	80.23	84.48	89.59	93.19	100.89
62	61.33	76.63	81.38	85.65	90.80	94.42	102.17
63	62.33	77.75	82.53	86.83	92.01	95.65	103.44
64	63.33	78.86	83.68	88.00	93.22	96.88	104.72
65	64.33	79.97	84.82	89.18	94.42	98.11	105.99
66	65.33	81.09	85.96	90.35	95.63	99.33	107.26
67	66.33	82.20	87.11	91.52	96.83	100.55	108.53
68	67.33	83.31	88.25	92.69	98.03	101.78	109.79
69	68.33	84.42	89.39	93.86	99.23	103.00	111.06
70	69.33	85.53	90.53	95.02	100.43	104.21	112.32
71	70.33	86.64	91.67	96.19	101.62	105.43	113.58
72	71.33	87.74	92.81	97.35	102.82	106.65	114.84
73	72.33	88.85	93.95	98.52	104.01	107.86	116.09
74	73.33	89.96	95.08	99.68	105.20	109.07	117.35
75	74.33	91.06	96.22	100.84	106.39	110.29	118.60
76	75.33	92.17	97.35	102.00	107.58	111.50	119.85
77	76.33	93.27	98.48	103.16	108.77	112.70	121.10
78	77.33	94.37	99.62	104.32	109.96	113.91	122.35
79	78.33	95.48	100.75	105.47	111.14	115.12	123.59
80	79.33	96.58	101.88	106.63	112.33	116.32	124.84

90% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.22
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.91
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63
75	2.77	2.37	2.16	2.02	1.93	1.85	1.80	1.75	1.72	1.69	1.58
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56
150	2.74	2.34	2.12	1.98	1.89	1.81	1.76	1.71	1.67	1.64	1.53
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.52
300	2.72	2.32	2.10	1.96	1.87	1.79	1.74	1.69	1.65	1.62	1.51
400	2.72	2.32	2.10	1.96	1.86	1.79	1.73	1.69	1.65	1.61	1.50
500	2.72	2.31	2.09	1.96	1.86	1.79	1.73	1.68	1.64	1.61	1.50

95% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96	1.80
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.73
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.70
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.69
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.69

97.5% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	984.87
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.50
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.44
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.39
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.34
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11
75	5.23	3.88	3.30	2.96	2.74	2.58	2.46	2.37	2.29	2.22	2.01
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97
150	5.13	3.78	3.20	2.87	2.65	2.49	2.37	2.28	2.20	2.13	1.92
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	1.90
300	5.07	3.73	3.16	2.83	2.61	2.45	2.33	2.23	2.16	2.09	1.88
400	5.06	3.72	3.15	2.82	2.60	2.44	2.32	2.22	2.15	2.08	1.87
500	5.05	3.72	3.14	2.81	2.59	2.43	2.31	2.22	2.14	2.07	1.86

99% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.28
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57	2.29
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.16
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.13
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.10
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.08
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.07

Fraktiler i t-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	90	95	97.5	99	99.5	99.9	f
1	3.078	6.314	12.706	31.821	63.657	318.309	1
2	1.886	2.920	4.303	6.965	9.925	22.327	2
3	1.638	2.353	3.182	4.541	5.841	10.215	3
4	1.533	2.132	2.776	3.747	4.604	7.173	4
5	1.476	2.015	2.571	3.365	4.032	5.893	5
6	1.440	1.943	2.447	3.143	3.707	5.208	6
7	1.415	1.895	2.365	2.998	3.499	4.785	7
8	1.397	1.860	2.306	2.896	3.355	4.501	8
9	1.383	1.833	2.262	2.821	3.250	4.297	9
10	1.372	1.812	2.228	2.764	3.169	4.144	10
11	1.363	1.796	2.201	2.718	3.106	4.025	11
12	1.356	1.782	2.179	2.681	3.055	3.930	12
13	1.350	1.771	2.160	2.650	3.012	3.852	13
14	1.345	1.761	2.145	2.624	2.977	3.787	14
15	1.341	1.753	2.131	2.602	2.947	3.733	15
16	1.337	1.746	2.120	2.583	2.921	3.686	16
17	1.333	1.740	2.110	2.567	2.898	3.646	17
18	1.330	1.734	2.101	2.552	2.878	3.610	18
19	1.328	1.729	2.093	2.539	2.861	3.579	19
20	1.325	1.725	2.086	2.528	2.845	3.552	20
21	1.323	1.721	2.080	2.518	2.831	3.527	21
22	1.321	1.717	2.074	2.508	2.819	3.505	22
23	1.319	1.714	2.069	2.500	2.807	3.485	23
24	1.318	1.711	2.064	2.492	2.797	3.467	24
25	1.316	1.708	2.060	2.485	2.787	3.450	25
30	1.310	1.697	2.042	2.457	2.750	3.385	30
50	1.299	1.676	2.009	2.403	2.678	3.261	50
75	1.293	1.665	1.992	2.377	2.643	3.202	75
100	1.290	1.660	1.984	2.364	2.626	3.174	100
150	1.287	1.655	1.976	2.351	2.609	3.145	150
200	1.286	1.653	1.972	2.345	2.601	3.131	200
400	1.284	1.649	1.966	2.336	2.588	3.111	400

Stikord