STATISTIKNOTER

Simple binomialfordelingsmodeller

Jørgen Larsen

IMFUFA
Roskilde Universitetscenter

Februar 1999

IMFUFA, Roskilde Universitetscenter, Postboks 260, DK-4000 Roskilde.
Jørgen Larsen: **STATISTIKNOTER: Simple binomialfordelingsmodeller**
IMFUFA tekst nr. 304a/1995 65 sider ISSN 0106-6242

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

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

•

Om kurset og kursusmaterialet kan blandt andet siges at

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

Indhold

Indledning **5**

1 Binomialfordelingen **7**
1.1 Binomialkoefficienter 10
1.2 Egenskaber ved binomialfordelingen 13
1.3 Opgaver 16

2 Den simple binomialfordelingsmodel **21**
2.1 Estimation af parameteren p 21
2.2 En simpel statistisk hypotese 26
2.3 Kvotienttesstørrelsen 27
2.4 Opgaver 31

3 Sammenligning af binomialfordelinger **35**
3.1 Modellen 36
3.2 Hypoteseprøvning 38
3.3 Det eksakte test i en 2×2-tabel 41
3.4 Opgaver 48

4 Logistisk regression **51**
4.1 Grundmodellen 51
4.2 En dosis-respons model 52
4.3 Estimation 55
4.4 Modelkontrol 58
4.5 Hypoteser om parametrene 60
4.6 Opgaver 64

5 Stikord **65**

Indledning

Dette hæfte handler om statistisk analyse af binomialfordelte observationer, dvs. observationer der fortæller hvor mange gange et tilfældighedseksperiment giver et bestemt udfald når man gentager eksperimentet et på forhånd fastlagt antal gange. Typiske eksempler på binomialfordelte observationer er

1. antal patienter (ud af n) der bliver raske af behandlingen,
2. antal forsøgsdyr (ud af n) der dør af behandlingen,
3. antal lyskilder i et klasseværelse der stadig virker et år efter at de blev monteret,
4. antal gange man får Krone når man slår Plat-eller-Krone med en mønt n gange.

Der melder sig to forskellige slags spørgsmål i slige situationer:

modelleringsproblemet: hvordan opstiller man en matematisk-statistisk model for sådanne observationer?

inferensproblemet: hvordan kan man ud fra foreliggende observationer, der ofte er af beskedent omfang, lære noget om virkelighedens indretning (eller mere beskedent: om modellens ukendte parametre)?

Disse spørgsmål diskuteres indgående. Inferensproblemet er et generelt problem, som her behandles i forbindelse med og ved hjælp af de forskellige binomialfordelingsmodeller. Som metode til statistisk inferens benyttes *likelihood-metoden* hvis grundlæggende ideer præsenteres omhyggeligt; derimod må vi af tekniske grunde give afkald på de matematiske beviser for denne metodes fortræffeligheder.

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 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 $Y = X_1 + X_2 + \ldots + X_n$. 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* om hvilke det antages at

(a) de er stokastisk uafhængige,

(b) de alle har den samme sandsynlighed p for at antage værdien 1, altså $\mathrm{P}(X_j = 1) = p$ for ethvert j.

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*[1] for X_j som

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

eller kortere

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

5. Vi har også brug for et udtryk for den *simultane* sandsynlighedsfunktion for $X_1, X_2, \ldots, X_n$, dvs. 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 $\ldots$ 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:

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

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

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

6. Da vi nu kender den simultane sandsynlighedsfunktion for X_j-erne, kan vi bestemme sandsynlighedsfunktionen for $Y = X_1 + X_2 + \ldots + X_n$. 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:

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

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 $x_1 + x_2 + \ldots + x_n = y$ (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

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

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«.

[1] 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 $Y = X_1 + X_2 + \ldots + X_{12}$. – 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

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

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[2]

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

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

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 har vi i stedet en placering af $n-k$ 1-er og k 0-er. Heraf følger at

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

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

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

[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*[3], 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 $21 = 6 + 15$.

[3] opkaldt efter den franske videnskabsmand og tænker B. Pascal (1623-62).

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 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.

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

(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:

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

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

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

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

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

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

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

Denne sum ser ikke så rar ud, men heldigvis kan vi finde middelværdien på en anden og smartere måde. Som omtalt tidligere (side 9) kan en binomialfordelt stokastisk variabel Y fremkomme som en sum af uafhængige identisk fordelte 01-variable, så lad os sige at

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

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:

$$\begin{aligned} \mathrm{E}\, Y &= \mathrm{E}\, X_1 + \mathrm{E}\, X_2 + \ldots + \mathrm{E}\, X_n \\ &= n\, \mathrm{E}\, X_1 \end{aligned}$$

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

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

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

Dernæst ser vi på *variansen*. Variansen på Y er pr. definition tallet $\mathrm{Var}\, Y = \mathrm{E}\big((Y - \mathrm{E}Y)^2\big)$. Igen bruger vi en smart måde: 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

$$\begin{aligned} \mathrm{Var}\, Y &= \mathrm{Var}\, X_1 + \mathrm{Var}\, X_2 + \ldots + \mathrm{Var}\, X_n \\ &= n\, \mathrm{Var} X_1, \end{aligned}$$

og vi behøver nu blot at finde variansen på X_1; da X_1 kun antager værdierne 0 og 1, er $X_1^2 = X_1$, så udregningerne bliver ekstra simple:

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

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

Sammenfatning: 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 $\sqrt{np(1-p)}$.

Tabel 1.2 Eksempel 1.1: Sandsynlighedsfunktionen f for binomialfordelingen med $n = 18$ og $p = 1/6$.

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

Udregning af binomialsandsynligheder

Hvis man ønsker at udregne binomialsandsynlighederne

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

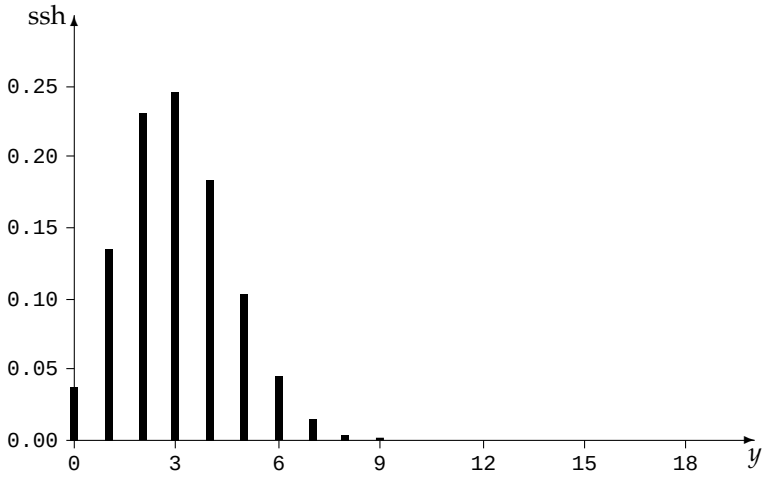
for $y = 0, 1, 2, \ldots, n$, er det som regel ikke hensigtsmæssigt bare uden videre at indsætte i formlen. Man kan med fordel benytte en rekursionsformel. Ved simple omskrivninger finder man

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

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

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

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



Figur 1.2 Eksempel 1.1: Pindediagram der viser sandsynlighedsfunktionen f for binomialfordelingen med $n = 18$ og $p = 1/6$.

$18 \cdot 1/6 = 3$ 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 Tabel 1.2. Sandsynlighedsfunktionen er vist i Figur 1.2.

1.3 Opgaver

Opgave 1.1
Tabel 1.1 (side 9) er fremstillet på den måde at man har sat et computerprogram[4] 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 9 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.

[4] et Turbo Pascal program der benytter Zaman og Marsaglia's tilfældigtalsgenerator FSU-Ultra, Version 1.05.

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 $\binom{12}{5}$ ved hjælp af Pascals trekant (og uden at bruge lommeregneren).

Udregn binomialkoefficienten $\binom{12}{5}$ 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 EY.

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 EY? Tegn grafen for EY 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 Var$Y$.

   Hvad er variansen Var$Y$ på Y? Tegn grafen for 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) $0 \le r \le n$: 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

$$\frac{\binom{R}{r} \cdot \binom{H}{n-r}}{\binom{R+H}{n}}.$$

   (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 9). 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. En sådan størrelse som p kaldes 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 retningslinie, som ikke bliver logisk-deduktivt bevist, 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 udelukkende 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.[1]

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

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

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

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

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:

- Hvis vi i modelfunktionen fixerer p og opfatter funktionen som en funktion af y alene, så har vi *sandsynlighedsfunktionen* svarende til parameterværdien p. Figur 2.1 viser en »typisk« sandsynlighedsfunktion.
- 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 $L(\,\cdot\,; y)$:

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

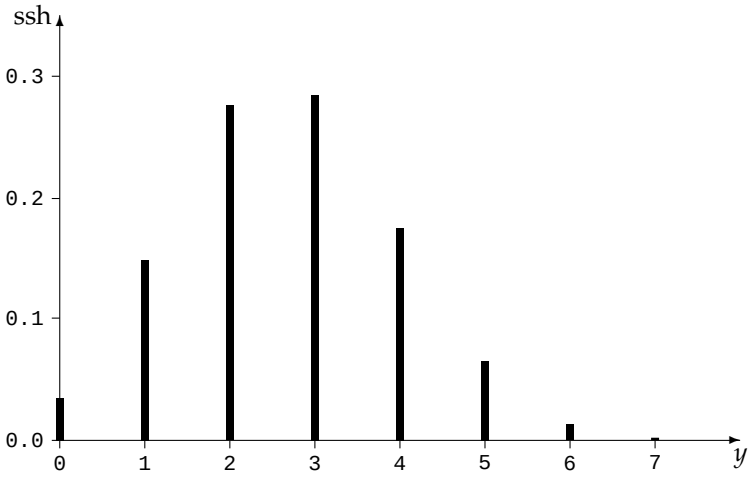
  Figur 2.2 viser en »typisk« likelihoodfunktion.

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

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



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

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 retningslinier, er da den værdi som maksimaliserer likelihoodfunktionen, og den kaldes *maksimaliseringsestimatet* (eller *maximum likelihood estimatet*) for p og betegnes $\hat{p}$ (»p hat«). Tallet $\hat{p}$ er altså bestemt ved at

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

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

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

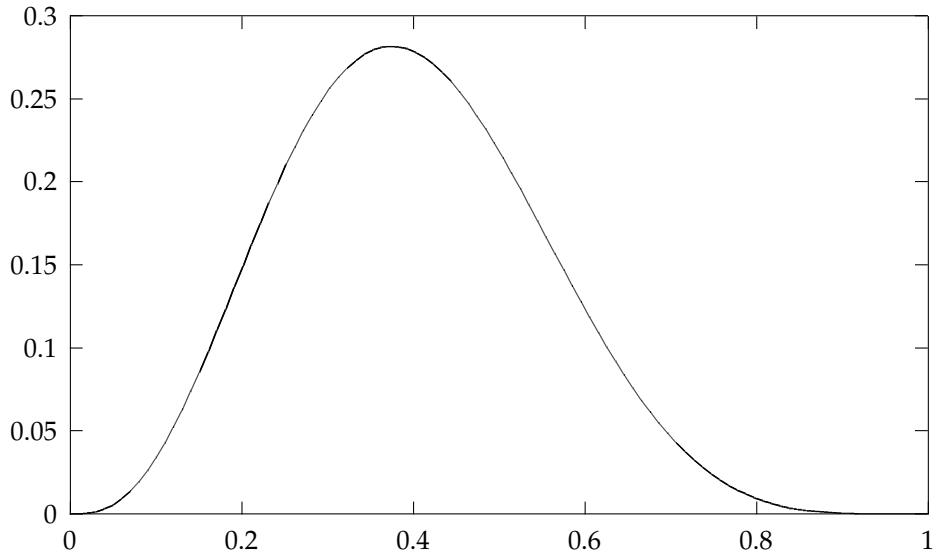
$$\ln L(p) \quad = \quad \ln \binom{144}{43} + 43 \ln p + 101 \ln(1-p).$$

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

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

Hvad er $\hat{p}$ i denne model? Svaret herpå får vi ved at løse den matematikopgave der hedder: »Bestem maksimumspunkt(er) for funktionen $p \mapsto \ln L(p)$

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



Figur 2.2 En »typisk« likelihoodfunktion $p \mapsto L(p; y) = f(y; p)$.

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:

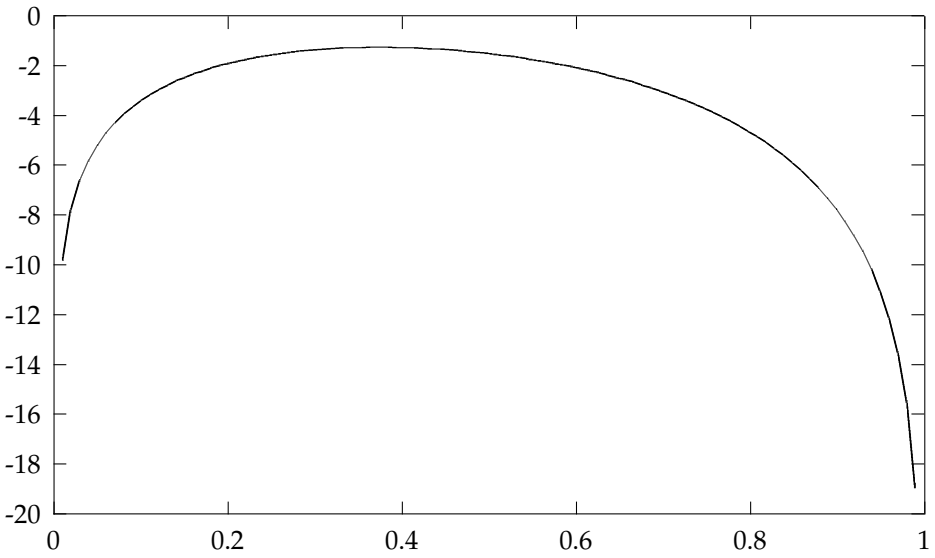
1. $0 < y < n$: Da er punktet $p = y/n$ det eneste stationære punkt for $\ln L$, og da $\ln L(0)$ og $\ln L(1)$ begge er $-\infty$, er $p = y/n$ et entydigt maksimumspunkt.
2. $y = n$: Så er $\ln L(p) = n \cdot \ln p$, hvilket er en voksende funktion af p. Den antager derfor sin største værdi når p er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

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

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

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



Figur 2.3 En log-likelihoodfunktion (svarende til likelihoodfunktionen i Figur 2.2).

også er det svar man når frem til ved at benytte den generelle fremgangsmåde der lyder

- opstil modelfunktionen,
- dan derudfra likelihoodfunktionen,
- bestem $\hat{p}$ 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 $\hat{p}$

Maksimaliseringsestimatet $\hat{p} = y/n$ 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 $\hat{p} = \hat{p}(Y) = Y/n$ 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 $\mathrm{E}\,\hat{p}(Y) = (\mathrm{E}Y)/n = p$, hvilket betyder at maksimaliseringsestimatoren $\hat{p}$ for p i middel giver det rigtige svar p – men deraf følger ikke noget om det konkrete enkelttilfælde.[2]

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

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

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

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

Sammenfattende kan vi sige at binomialparameteren p i billeeksemplet estimeres til $\hat{p} = 0.30$ 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[3] 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 $\hat{p} = {}^{43}/_{144}$ ikke er lig med 0.23, men en eksakt 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 *sandsynligheds*fordeling. Derfor kan man kun sige at

- hvis der *ikke* er stor afvigelse mellem $\hat{p}$ 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 $\hat{p}$ 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 $\hat{p}$ og p_0, og hvordan afgør man hvornår afvigelsen er stor og hvornår ikke. Afsnit 2.3 præsenterer en generel metode hvormed man kan håndtere disse spørgsmål.

[3] – men det er det ikke; denne del af eksemplet er opdigtet til lejligheden.

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 et spørgsmål 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 Kvotienttesteststørrelsen

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 $\hat{p} = \hat{p}(y)$ den bedst mulige beskrivelse af observationen y. Parameterværdier der giver en værdi af likelihoodfunktionen som ligger tæt på den maksimale værdi $L(\hat{p})$, må give en næsten lige så god beskrivelse af observationen y som $\hat{p}$ 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 $L(\hat{p})$. Hvis $L(p_0)$ er næsten lige så stor som $L(\hat{p})$, betyder det at p_0 beskriver observationen y næsten lige så godt som $\hat{p}$ 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 $L(\hat{p})$, betyder det, at p_0 giver en væsentligt dårligere beskrivelse af observationen y end $\hat{p}$ 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 $L(\hat{p})$, skal det gøres ved at dividere den mindste med den største: man danner kvotienten

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

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 $\hat{p}$, dvs. man accepterer H_0,

- en Q-værdi langt fra 1 betyder at p_0 er væsentligt dårligere end $\hat{p}$, 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

$$\begin{aligned} Q = Q(y) &= \frac{p_0^y (1-p_0)^{n-y}}{\hat{p}^y (1-\hat{p})^{n-y}} \\ &= \left(\frac{np_0}{y}\right)^y \left(\frac{n(1-p_0)}{n-y}\right)^{n-y} \end{aligned}$$

idet $\hat{p} = y/n$. 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 har 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 $Q \leq Q_{\text{obs}}$, 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 $Q \leq Q_{\text{obs}}$, 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*

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

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

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

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 værdien af $0\ln 0$ 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

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

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 kvotienttesstørrelsen $-2\ln Q$ 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.[4] 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[5] i χ^2-fordelingen, se f.eks. Tabel 2.1.

I en tabel over fraktiler i χ^2-fordelingen finder man at svarende til 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%.)

[4] Det kan nævnes at χ^2-fordelingen med 1 frihedsgrad er den kontinuerte sandsynlighedsfordeling som har tæthedsfunktion

$$f(x) \quad = \quad \frac{1}{\sqrt{2\pi}}\, x^{-1/2} \exp(-x/2),\ x > 0\,.$$

[5] 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 $\leq x$.

Tabel 2.1 Udvalgte fraktiler i χ^2-fordelingen med 1 frihedsgrad.

sandsynlighed	fraktil
0.01	0.000157
0.025	0.000982
0.05	0.00393
0.10	0.0158
0.50	0.455
0.90	2.71
0.95	3.84
0.975	5.02
0.99	6.63

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 retningslinier 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 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 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 $\hat{p}$.
2. Tegn et pindediagram over den empiriske fordeling af $\hat{p}$.
3. Tegn et pindediagram over den teoretiske fordeling af $\hat{p}$.

   TIP: Da Y er binomialfordelt, er fordelingen af $\hat{p} = Y/n$ 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 $\hat{p}$?

TIP: 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 23 står, at man kan bestemme $\hat{p}$ 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 29 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

$$-2\ln Q \approx \frac{(y-np_0)^2}{np_0(1-p_0)} = \left(\frac{y-np_0}{\sqrt{np_0(1-p_0)}}\right)^2.$$

(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, 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; og 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.

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

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.

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

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

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.				
klasse	1	2	3	...	s
1	y_1	y_2	y_3	...	y_s
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 $H_0 : p_1 = p_2 = \ldots = p_s$.

De generelle retningslinier 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

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

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

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

I bille-eksemplet bliver log-likelihoodfunktionen

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

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 $(\hat{p}_1,\hat{p}_2,\ldots,\hat{p}_s)$ 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

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

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

3.2 Hypoteseprøvning

Vi skal undersøge om det er rimeligt at antage at hypotesen $H_0 : p_1 = p_2 = \ldots = p_s$ 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 $n_\bullet = n_1 + n_2 + \ldots + n_s$ 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 $y_1,y_2,\ldots,y_s$:

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

Maksimaliseringsestimatet $\hat{p}$ for p er den værdi der maksimaliserer denne log-likelihoodfunktion, dvs. den værdi p der maksimaliserer

$$y_\bullet\ln p + (n_\bullet - y_\bullet)\ln(1-p).$$

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

Vi ved fra side 24 at løsningen er $\hat{p} = y_\cdot/n_\cdot$. Likelihoodmetoden giver altså det svar som vi formodede måtte være det rigtige. – I vort eksempel bliver $\hat{p} = {}^{188}/_{317} = 0.59$.

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

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

der bliver et tal mellem 0 og 1; en Q-værdi tæt på 1 betyder at sættet $(\hat{p}, \hat{p}, \ldots, \hat{p})$ 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 $-2\ln Q$ som bliver

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

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

Hvis vi indfører betegnelsen $\hat{y}_j = n_j\hat{p}$, kan $-2\ln Q$ omskrives til

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

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

dien af teststørrelsen til

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

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 ses at værdien

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

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 $\hat{y}_j$ og $n_j - \hat{y}_j$, $j = 1, 2, \ldots, s$ være mindst fem.* Hvis denne betingelse ikke er opfyldt, kan man eventuelt udelade de problematiske grupper eller slå nogle af grupperne sammen 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 som 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

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

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

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

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

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 $\hat{p} = {}^8\!/_{15} = 0.53$, og de »forventede« antal $\hat{y}_1 = n_1\hat{p}$, $n_1 - \hat{y}_1 = n_1(1-\hat{p})$, $\hat{y}_2 = n_2\hat{p}$ og $n_2 - \hat{y}_1 = n_2(1-\hat{p})$ 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 39)

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

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

		y_1						
		0	1	2	3	4	5	6
	0	·	·	⋆	⋆	⋆	⋆	⋆
	1	·	·	·	⋆	⋆	⋆	⋆
	2	⋆	·	·	·	⋆	⋆	⋆
	3	⋆	·	·	·	⋆	⋆	⋆
y_2	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 $\star$.

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 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 $\hat{p}$ (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 endvidere 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 $Y_1 + Y_2 = 8$

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

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

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

hvor

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

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 $\hat{\theta}$ 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 $\hat{\theta} = \hat{\theta}(2) = 0.276$. Kvotienttesstørrelsen for $\overline{H}_0$ er

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

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:

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

Bestemmelsen af ε er ukompliceret, men noget besværlig. Af Tabel 3.5 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

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

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 $\overline{H}_0$.

Tabel 3.5 Tabel over $Q(y)$ og $\overline{f}(y;1)$ til brug ved beregning af ε.

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

Vi er gået let hen over hvordan man egentlig skal finde talværdien af $\hat{\theta}$ 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.

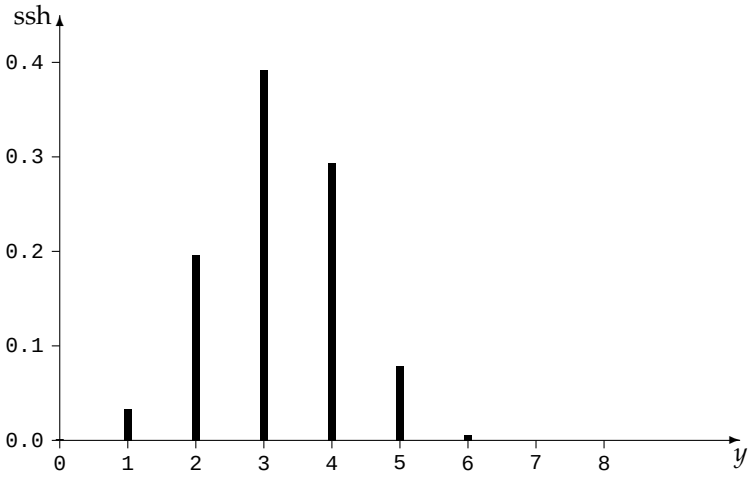
Fishers eksakte test

Det man gør når man tester en statistisk hypotese, er at man udregner værdien af en vis teststørrelse, sædvanligvis kvotienttesteststørrelsen Q eller $-2\ln Q$, der er et udtryk for hvor godt hypotesen er forenelig med de foreliggende data; dernæst bestemmer man testsandsynligheden, dvs. 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 $\overline{f}(\cdot\,;1)$ svarende til at hypotesen $\overline{H}_0$ er rigtig; det har blandt andet den fordel at man slipper for at skulle bestemme $\hat{\theta}$. Funktionen $\overline{f}(\cdot\,;1)$ er forholdsvis simpel:

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

(I øvrigt er $\overline{f}(\cdot\,;1)$ 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 $\overline{f}(y;1) \leq \overline{f}(2;1)$. For at gøre det udregner vi tælleren i højresiden af formel (3.3) for alle de mulige y-er, f.eks.



Figur 3.2 Den hypergeometriske fordeling bestemt ved sandsynlighedsfunktionen i formel (3.3).

Tabel 3.6 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

ved brug af Pascals trekant (side 11); man får da Tabel 3.6. Det ses at de y-er som er mere ekstreme end $y = 2$ (i den forstand at $\overline{f}(y;1) \leq \overline{f}(2;1) = 1260/6435$) er alle y-erne undtagen $y = 3$ og $y = 4$. Testsandsynligheden er derfor

$$\begin{aligned} \varepsilon &= 1 - \left(\overline{f}(3;1) + \overline{f}(4;1)\right) \\ &= 1 - \frac{2520 + 1890}{6435} \\ &= 31\%. \end{aligned}$$

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

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 (ubetin-

gede) 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 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 piloter 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 (*Streptococcus pyogenes*)
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 39).

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 $y_j \ln \dfrac{y_j}{\hat{y}_j}$ *og* $(n_j - y_j)\ln \dfrac{n_j - y_j}{n_j - \hat{y}_j}$ i udtrykket for $-2\ln Q$, og vis derved at man kan approksimere $-2\ln Q$ med den såkaldte *Pearsons* X^2 defineret som[1]

$$X^2 = \sum_{j=1}^{s} \frac{(y_j - \hat{y}_j)^2}{n_j\hat{p}(1-\hat{p})}.$$

[1] og opkaldt efter den engelske videnskabsmand Karl Pearson (1857-1936).

4 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, hvoraf vi nu bruger en endnu større del: I en undersøgelse[1] 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 4.1.

4.1 Grundmodellen

Første skridt i modelleringsprocessen består i at gøre sig klart at Tabel 4.1 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.[2]
- 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 4.2) og tegninger (Figur 4.1).

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

[1] Her citeret efter Pack and Morgan (1990): A mixture model for interval-censored time-to-response quantal assay data, *Biometrics* **42**, 749-757.

[2] Størrelsen »køn« er en såkaldt *faktor*, dvs. en baggrundsvariabel der kun (kan) benyttes til at inddele i grupper efter. I modsætning hertil er »dosis« en størrelse der antager rigtige talværdier på en kontinuert måleskala.

Tabel 4.1 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 4.2 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

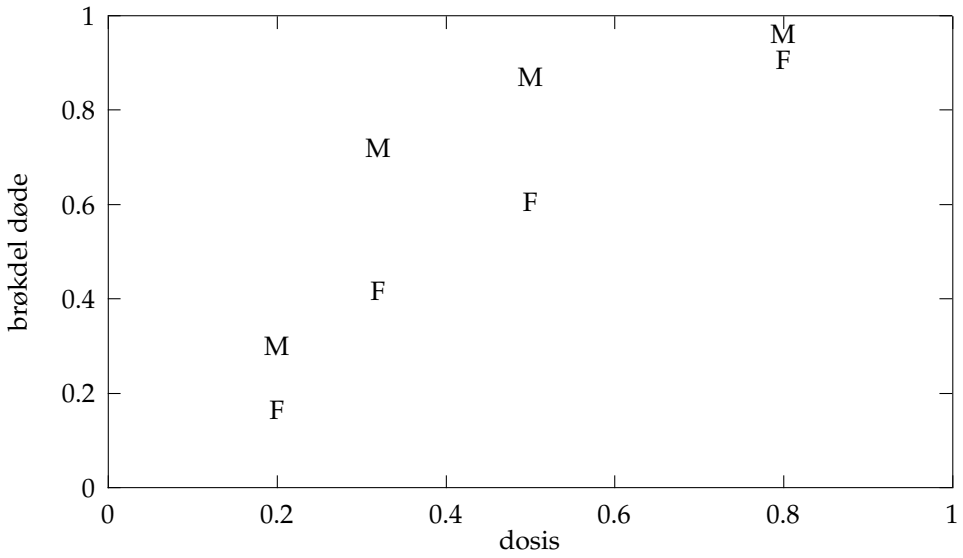
af de otte delforsøg er det nærliggende at foreslå at beskrive »antal døde« som en observation fra en binomialfordeling med en antalsparameter der er det samlede antal biller i den pågældende gruppe, og med en (ukendt) sandsynlighedsparameter der skal fortolkes som sandsynligheden for at en bille af 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 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.

4.2 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



Figur 4.1 Rismelsbillers overlevelse: Observeret dødssandsynlighed (relativ hyppighed) som funktion af dosis, for hvert køn.

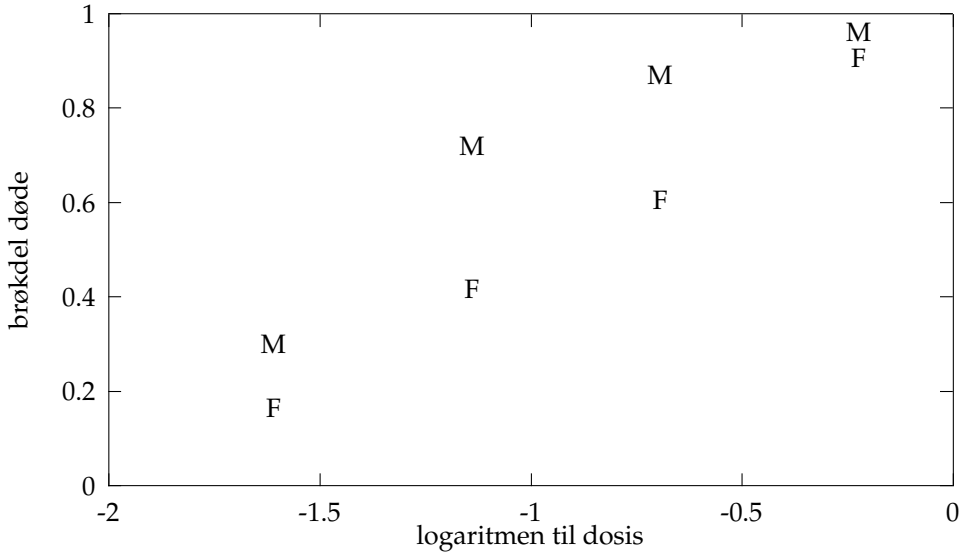
om hvordan netop dette giftstof virker i billeorganismen, kunne man formentlig give et velbegrundet forslag til hvordan sandsynligheden afhænger af dosis. Men den statistiske modelbyggers tilgang til problemet er af en langt mere jordbunden og pragmatisk karakter som vi nu skal se.

I eksemplet har eksperimentator valgt nogle tilsyneladende mærkværdige dosis-værdier (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 et 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. Derfor tegner vi Figur 4.1 en gang til idet vi nu afsætter logaritmen til dosis ud ad abscisseaksen; resultatet ses i Figur 4.2.

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:

Definition 4.1 (logit**-funktionen)**
Funktionen logit *afbilder intervallet* $]0,1[$ *på den reelle akse* $\mathbf{R}$ *og er givet ved*

$$\text{logit}(p) \quad = \quad \ln \frac{p}{1-p}.$$



Figur 4.2 Rismelsbillers overlevelse: Observeret dødssandsynlighed (relativ hyppighed) som funktion af logaritmen til dosis, for hvert køn.

Hvis $z = \text{logit}(p)$, *så er* $p = \dfrac{\exp(z)}{1+\exp(z)}$.

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 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:

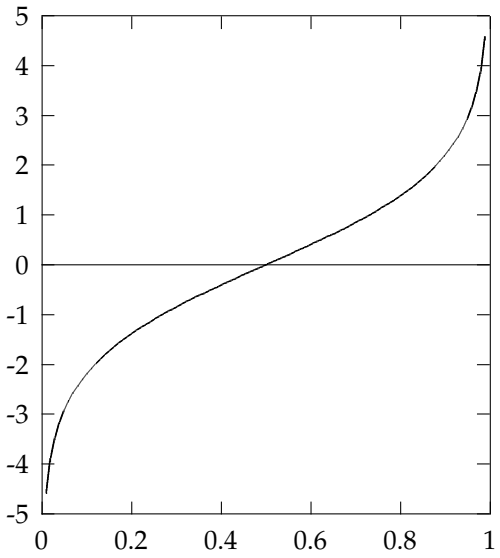
For hvert af de to køn afhænger $\text{logit}(p_d)$ lineært af $x = \ln d$,

eller mere udførligt:

Der findes konstanter α_M, β_M og α_F, β_F således at

$$\begin{aligned} \text{logit}(p_{dM}) &= \alpha_M + \beta_M \ln d \\ \text{logit}(p_{dF}) &= \alpha_F + \beta_F \ln d \end{aligned}$$

I Figur 4.4 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 linie, 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.



Figur 4.3 Del af grafen for logit-funktionen. – Der gælder at for $p \to 1$ vil $\text{logit}(p) \to +\infty$ og for $p \to 0$ vil $\text{logit}(p) \to -\infty$.

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.

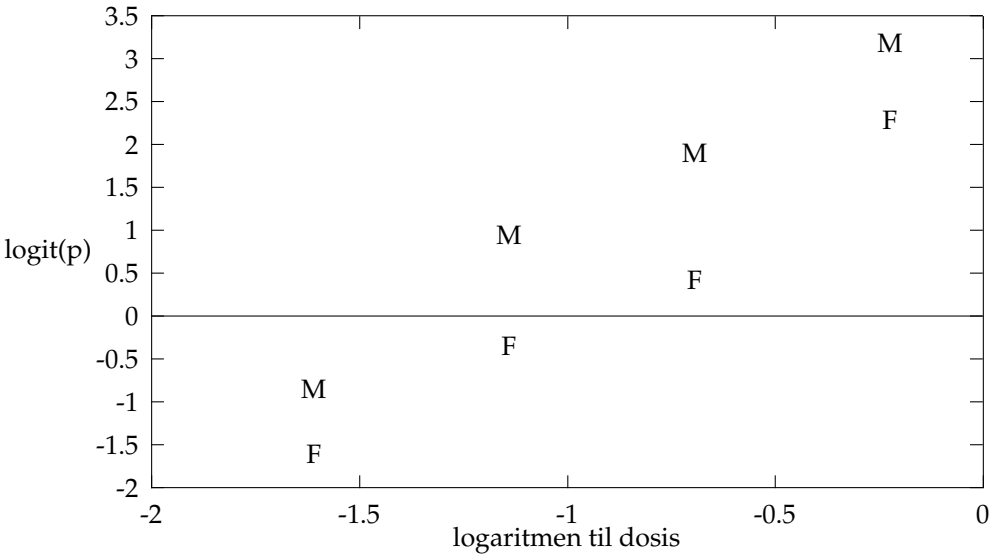
4.3 Estimation

I dette afsnit diskuteres hvordan man estimerer parametrene 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
>
> $$p_j \;=\; \text{logit}^{-1}(\alpha + \beta x_j) \;=\; \frac{\exp(\alpha + \beta x_j)}{1 + \exp(\alpha + \beta x_j)}.$$
>
> 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, og her er x_j logaritmen til koncentrationen i gruppe j.

Likelihoodfunktionen er



Figur 4.4 Rismelsbillers overlevelse: Logit til estimeret dødssandsynlighed (relativ hyppighed) som funktion af logaritmen til dosis, for hvert køn.

$$\begin{aligned} L(\alpha,\beta) &= \prod_{j=1}^{s} \binom{n_j}{y_j} p_j^{y_j} (1-p_j)^{n_j-y_j} \\ &= \prod_{j=1}^{s} \binom{n_j}{y_j} \cdot \prod_{j=1}^{s} \left(\frac{p_j}{1-p_j}\right)^{y_j} \cdot \prod_{j=1}^{s} (1-p_j)^{n_j} \\ &= \text{konstant} \cdot \prod_{j=1}^{s} \left(\frac{p_j}{1-p_j}\right)^{y_j} \cdot \prod_{j=1}^{s} (1-p_j)^{n_j}, \end{aligned}$$

og log-likelihoodfunktionen er

$$\begin{aligned} &\ln L(\alpha,\beta) \\ &= \text{konstant} + \sum_{j=1}^{s} y_j \ln \frac{p_j}{1-p_j} + \sum_{j=1}^{s} n_j \ln(1-p_j) \\ &= \text{konstant} + \sum_{j=1}^{s} y_j \operatorname{logit}(p_j) + \sum_{j=1}^{s} n_j \ln(1-p_j) \\ &= \text{konstant} + \sum_{j=1}^{s} y_j (\alpha + \beta x_j) + \sum_{j=1}^{s} n_j \ln(1-p_j) \\ &= \text{konstant} + \alpha \sum_{j=1}^{s} y_j + \beta \sum_{j=1}^{s} x_j y_j + \sum_{j=1}^{s} n_j \ln(1-p_j) \\ &= \text{konstant} + \alpha \sum_{j=1}^{s} y_j + \beta \sum_{j=1}^{s} x_j y_j - \sum_{j=1}^{s} n_j \ln\bigl(1+\exp(\alpha+\beta x_j)\bigr)\,. \end{aligned}$$

Som altid er det bedste bud på værdierne af de ukendte parametre dem der maksimaliserer likelihoodfunktionen eller log-likelihoodfunktionen. Hermed er vi stødt på 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

$$\begin{aligned}\frac{\partial}{\partial\alpha}\ln L(\alpha,\beta) &= \sum_{j=1}^{s}(y_j - n_j p_j),\\ \frac{\partial}{\partial\beta}\ln L(\alpha,\beta) &= \sum_{j=1}^{s}x_j(y_j - n_j p_j),\end{aligned}$$

og da disse som nævnt skal være 0, får vi de to ligninger

$$\sum_{j=1}^{s}(y_j - n_j p_j) = 0, \tag{4.1}$$

$$\sum_{j=1}^{s}x_j(y_j - n_j p_j) = 0, \tag{4.2}$$

med de ubekendte α og β (der indgår »skjult« i p_j).

Lad os se hvordan disse ligninger tage sig ud for hanbillernes vedkommende. Ligning (4.1) 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 (4.2)

$$\begin{aligned}&\ln(0.20)\left(43 - 144\,\frac{\exp(\alpha+\beta\ln(0.20))}{1+\exp(\alpha+\beta\ln(0.20))}\right)\\ +&\ln(0.32)\left(50 - 69\,\frac{\exp(\alpha+\beta\ln(0.32))}{1+\exp(\alpha+\beta\ln(0.32))}\right)\\ +&\ln(0.50)\left(47 - 54\,\frac{\exp(\alpha+\beta\ln(0.50))}{1+\exp(\alpha+\beta\ln(0.50))}\right)\\ +&\ln(0.80)\left(48 - 50\,\frac{\exp(\alpha+\beta\ln(0.80))}{1+\exp(\alpha+\beta\ln(0.80))}\right) = 0.\end{aligned}$$

Det ser ikke rart ud! Faktisk kan man ikke løse ligningerne hvis man med »løse ligningerne« mener at flytte rundt på symbolerne så man ender med noget af formen »α = noget kendt« og »β = noget kendt«. I stedet må man henvende sig

```
StatUnit.FitLogitLinear

Dependent relative frequency BRØKDEL, binomial totals TOTAL.

Model terms
     KØN
     KØN*LOGDOSIS

8 observations, 4 parameters estimated.
-2log(Likelihood) =               3.3637
Likelihood ratio test against full model
P[ ChiSquare(4) > -2log(Likelihood) ] =  0.498900
------------------------------------------------------------------------
StatUnit.ListParameters

                          Estimate     Std.dev.          U       P
        KØN[1]               4.270       0.5372        7.949   0.00000
        KØN[2]               2.562       0.3785        6.767   0.00000
KØN[1]*LOGDOSIS              3.138       0.3854        8.142   0.00000
KØN[2]*LOGDOSIS              2.582       0.3047        8.472   0.00000
------------------------------------------------------------------------
```

Figur 4.5 Estimation af parametrene i grundmodellen for bille-dataene: Del af udskrift fra *StatUnit* (Tue Tjurs Turbo Pascal unit til statistisk analyse). I de fire sidste linier ses at $\hat{\alpha}_M$=4.270, $\hat{\alpha}_F$=2.562, $\hat{\beta}_M$=3.138 og $\hat{\beta}_F$=2.582.

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 principielt kunne man jo også tænke sig at der var flere løsninger). Eller man kan benytte et tilpas avanceret statistikprogram til computere; det vil have indbygget nogle numerisk analyse-metoder så det kan udregne brugbare tilnærmelser til maksimaliseringsestimaterne $\hat{\alpha}$ og $\hat{\beta}$.

I Figur 4.5 er vist noget af en udskrift fra et sådant statistikprogram. Deraf fremgår det blandt andet at for hanbillerne er estimaterne $\hat{\alpha}_M = 4.270$ (med en middelfejl på 0.5372) og $\hat{\beta}_M = 3.138$ (med en middelfejl på 0.3854), og for hunbillerne er de $\hat{\alpha}_F = 2.562$ (med en middelfejl på 0.3785) og $\hat{\beta}_F = 2.582$ (med en middelfejl på 0.3047).

4.4 Modelkontrol

Vi har nu estimeret parametrene i den model der siger at

$$\text{logit}(p_{dk}) = \alpha_k + \beta_k \ln d$$

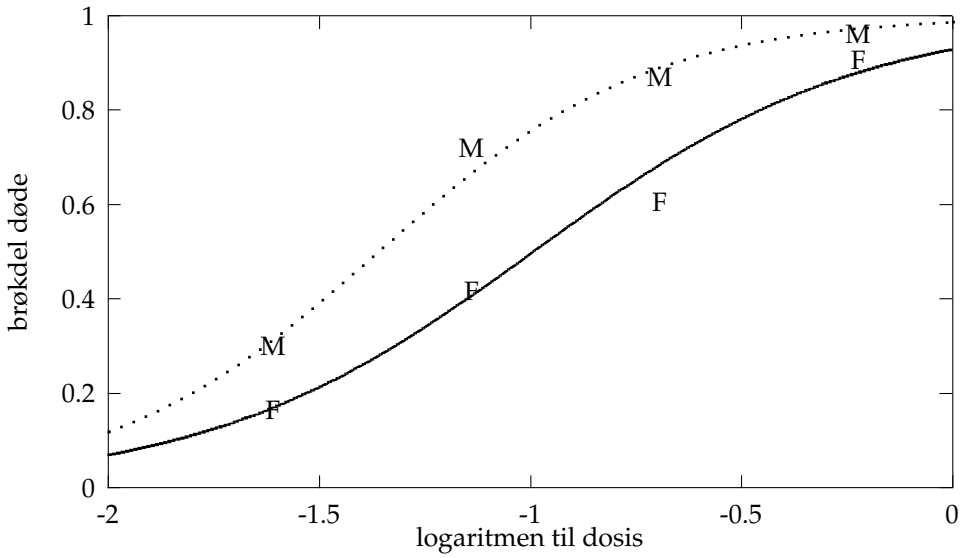
eller

$$p_{dk} = \frac{\exp(\alpha_k + \beta_k \ln d)}{1 + \exp(\alpha_k + \beta_k \ln d)}.$$

En nærliggende form for modelkontrol er derfor at indtegne graferne for de to funktioner

$$d \mapsto \frac{\exp(\alpha_M + \beta_M \ln d)}{1 + \exp(\alpha_M + \beta_M \ln d)}$$

og



Figur 4.6 Rismelsbillers overlevelse: To forskellige kurver, samt de observerede relative hyppigheder.

$$d \quad \mapsto \quad \frac{\exp(\alpha_F + \beta_F \ln d)}{1 + \exp(\alpha_F + \beta_F \ln d)}$$

i Figur 4.2 hvorved man får Figur 4.6. Den viser at modellen ikke er helt hen i vejret. Man kan desuden ved hjælp af likelihoodmetoden konstruere et numerisk test baseret på

$$Q \quad = \quad \frac{L(\hat{\alpha}_M, \hat{\alpha}_F, \hat{\beta}_M, \hat{\beta}_F)}{L_{\max}} \tag{4.3}$$

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 $\hat{p}_{dk} = \text{logit}^{-1}(\hat{\alpha}_k + \hat{\beta}_k \ln d)$ og $\hat{y}_{dk} = n_{dk}\hat{p}_{dk}$ bliver

$$\begin{aligned} Q &= \frac{\prod_k \prod_d \binom{n_{dk}}{y_{dk}} \hat{p}_{dk}^{y_{dk}} (1-\hat{p}_{dk})^{n_{dk}-y_{dk}}}{\prod_k \prod_d \binom{n_{dk}}{y_{dk}} \left(\frac{y_{dk}}{n_{dk}}\right)^{y_{dk}} \left(1-\frac{y_{dk}}{n_{dk}}\right)^{n_{dk}-y_{dk}}} \\ &= \prod_k \prod_d \left(\frac{\hat{y}_{dk}}{y_{dk}}\right)^{y_{dk}} \left(\frac{n_{dk}-\hat{y}_{dk}}{n_{dk}-y_{dk}}\right)^{n_{dk}-y_{dk}} \end{aligned}$$

og

$$-2\ln Q \quad = \quad 2\sum_k \sum_d \left(y_{dk} \ln \frac{y_{dk}}{\hat{y}_{dk}} + (n_{dk}-y_{dk}) \ln \frac{n_{dk}-y_{dk}}{n_{dk}-\hat{y}_{dk}} \right).$$

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 ($\hat{y}_{kd}$ og $n_{kd} - \hat{y}_{kd}$) 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 computerudskriften Figur 4.5 ses at $-2 \ln Q_{\text{obs}} = 3.3637$, og at den omtalte testsandsynlighed er 0.4989. – Antallet af frihedsgrader er bestemt på følgende måde: I den »fulde« model (der leverer nævneren i formel (4.3)) er der 8 parametre, én for hver gruppe; i den testede model (der leverer tælleren i formel (4.3)) 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.

4.5 Hypoteser om parametrene

Vi har opstillet en model som indeholder fire parametre, og som ser ud til at give en ganske god beskrivelse af observationerne. Næste punkt på dagsordenen er 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 4.1). I Figur 4.7 ses noget af udskriften fra et computerprogram der har løst opgaven. Parameterestimaterne er $\hat{\hat{\alpha}}_M = 3.835$ (med en middelfejl på 0.3443), $\hat{\hat{\alpha}}_F = 2.831$ (middelfejl 0.3105) og $\hat{\hat{\beta}} = 2.813$ (middelfejl 0.2386). Hypotesen 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:

$$\begin{aligned} -2 \ln Q &= -2 \ln \frac{L(\hat{\hat{\alpha}}_M, \hat{\hat{\alpha}}_F, \hat{\hat{\beta}}, \hat{\hat{\beta}})}{L(\hat{\alpha}_M, \hat{\alpha}_F, \hat{\beta}_M, \hat{\beta}_F)} \\ &= 2 \sum_k \sum_d \left(y_{dk} \ln \frac{\hat{y}_{dk}}{\hat{\hat{y}}_{dk}} + (n_{dk} - y_{dk}) \ln \frac{n_{dk} - \hat{y}_{dk}}{n_{dk} - \hat{\hat{y}}_{dk}} \right) \end{aligned}$$

```
StatUnit.FitLogitLinear

Dependent relative frequency BRØKDEL, binomial totals TOTAL.

Model terms
     KØN
     LOGDOSIS

8 observations, 3 parameters estimated.
-2log(Likelihood) =              4.6705
Likelihood ratio test against full model
P[ ChiSquare(5) > -2log(Likelihood) ] =  0.457406
-----------------------------------------------------------------
StatUnit.ListParameters

                     Estimate     Std.dev.            U          P
  KØN[1]                3.835       0.3443       11.140   0.000000
  KØN[2]                2.831       0.3105        9.118   0.000000
LOGDOSIS                2.813       0.2386       11.791   0.000000
-----------------------------------------------------------------
StatUnit.TestModelChange

1 parameters removed
-2log(Q) =               1.3067
P[ ChiSquare(1) > -2log(Q) ] =  0.252986
-----------------------------------------------------------------
```

Figur 4.7 Hypotesen om parallelle kurver for bille-dataene: Del af udskrift fra *StatUnit*. Teststørrelsen er $-2\ln Q_{\text{obs}}$=1.3067 og testsandsynligheden 0.253.

hvor $\hat{\hat{y}}_{dk} = n_{dk}\,\text{logit}^{-1}(\hat{\hat{\alpha}}_k + \hat{\hat{\beta}}\ln d)$. Man får at $-2\ln Q_{\text{obs}} = 1.3067$, 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.

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.) I Figur 4.8 ses computerudskrifter vedrørende denne hypotese. Det fremgår at når man tester H_2 i forhold til H_1, får man $-2\ln Q_{\text{obs}} = 27.5017$ der skal sammenlignes med χ^2-fordelingen med et antal frihedsgrader på $3-2=1$; sandsynligheden for at få værdier større end 27.5017 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 logit p lineært af $\ln d$; de to kurver er parallelle men ikke sammenfaldende. De estimerede kurver er

$$\begin{aligned} \text{logit}(p_{dM}) &= 3.84 + 2.81\ln d \\ \text{logit}(p_{dF}) &= 2.83 + 2.81\ln d, \end{aligned}$$

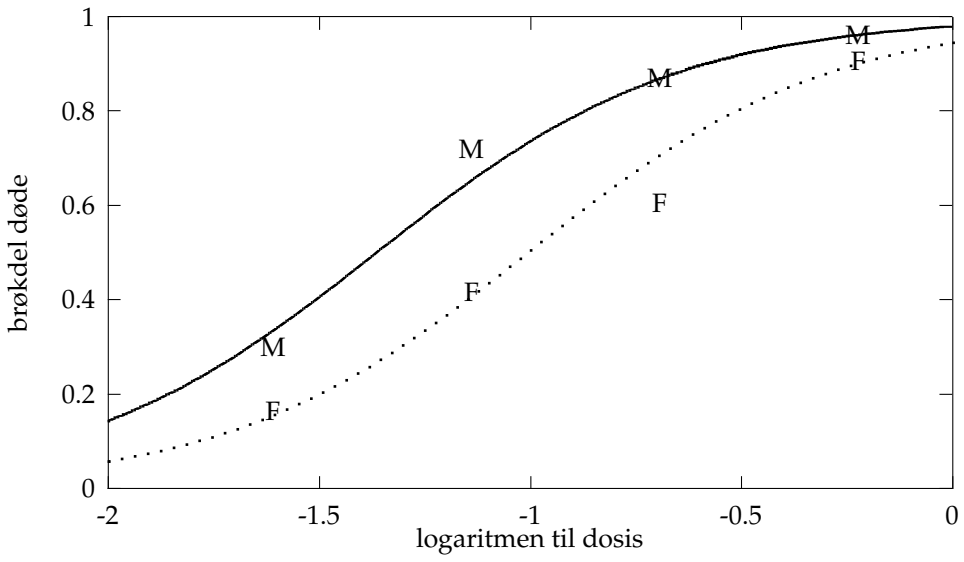
svarende til at

$$\begin{aligned} p_{dM} &= \frac{\exp(3.84 + 2.81 \ln d)}{1 + \exp(3.84 + 2.81 \ln d)} \\ p_{dF} &= \frac{\exp(2.83 + 2.81 \ln d)}{1 + \exp(2.83 + 2.81 \ln d)}. \end{aligned}$$

Figur 4.9 illustrerer situationen.

```
StatUnit.FitLogitLinear

Dependent relative frequency BRØKDEL, binomial totals TOTAL.

Model terms
     CONSTANT
     LOGDOSIS

8 observations, 2 parameters estimated.
-2log(Likelihood) =              32.1722
Likelihood ratio test against full model
P[ ChiSquare(6) > -2log(Likelihood) ] =  0.000015
-------------------------------------------------------------------------
StatUnit.ListParameters

                       Estimate        Std.dev.            U          P
CONSTANT                  3.204          0.3010       10.644   0.000000
LOGDOSIS                  2.709          0.2291       11.821   0.000000
-------------------------------------------------------------------------
StatUnit.TestModelChange

1 parameters removed
-2log(Q) =               27.5017
P[ ChiSquare(1) > -2log(Q) ] =  0.000000
-------------------------------------------------------------------------
```

Figur 4.8 Hypotesen om sammenfaldende kurver for bille-dataene: Del af udskrift fra *StatUnit*. Teststørrelsen er $-2 \ln Q_{\text{obs}}$=27.5017, og testsandsynligheden er 0.000.



Figur 4.9 Rismelsbillers overlevelse: Den endelige model.

4.6 Opgaver

Opgave 4.1
Vis at $\text{logit}(1/2) = 0$. Vis at $\text{logit}(1-p) = -\text{logit}(p)$.

Opgave 4.2
Eftervis at $\text{logit}(p) = z$ hvis og kun hvis $p = \dfrac{\exp(z)}{1+\exp(z)}$ således som det postuleres i Definition 4.1 på side 53.

Opgave 4.3
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 $p(\ln d) = 0.5$.

5 Stikord

χ^2-approksimation 30, 31, 41

01-variabel 9, 14

accept af hypotese 27, 28
antalsparameter i binomialfordeling 9

baggrundsvariabel 51
binomialfordeling 9, 13,
    definition 13
    middelværdi og varians 13
    udledning 7
    udregning af sandsynligheder 15
binomialformlen 13
binomialforsøg 7
binomialkoefficient 9, 10, 13

central estimator 25

eksakt test i en 2×2-tabel 41
eksakt test, Fishers 46
elementarforsøg 7
empirisk fordeling 17
estimat 21, 23
estimation 21
estimator 25

faktor 51
Fishers eksakte test 46
forkastelse af hypotese 27, 28
fraktil 30
frihedsgrader 30, 40
fuld model 59

hypergeometrisk fordeling 19, 46
hypotese 43, 46
    sammensat 43, 45
    simpel 43

indikatorvariabel 7

kvotienttesteststørrelse 27, 39

likelihoodfunktion 22, 23, 25
likelihoodmetoden 21
logistisk regression 51
logit 53, 55

maksimaliseringsestimat 23
maksimaliseringsestimator 25
maximum likelihood estimate 23
middelfejl 25, 26
modelfunktion 22, 37, 41
modelkontrol 58

odds 54

parameter 9, 21,
    sand værdi 21
Pascal, B. 11
Pascals trekant 11
Pearson, K. 50
Pearsons X^2 50
pindediagram 16, 17

rekursion 15

sammensat hypotese 43, 45
sandsynlighedsfunktion 8
sandsynlighedsparameter 9
signifikans 29
signifikant forskel 26
signifikant teststørrelse 29
simpel hypotese 43
simultan sandsynlighedsfunktion 8
skøn 21
standardafvigelse 14

Taylorudvikling 33, 50
testsandsynlighed 28

varians 14