

# **IDRÆTSSTATISTIK**

## **BIND 2**

Det Naturvidenskabelige Fakultet  
Aarhus Universitet  
Reprocenter

© Preben Blæsild og Jørgen Granfeldt 2001

# Forord

Denne bog er skrevet til brug i et statistikkursus for bachelorstuderende ved Center for Idræt, Aarhus Universitet.

Bag bogen ligger samme holdninger både til statistisk analyse og til begynderundervisning i statistik, der primært retter sig mod brugere, som i Blæsild og Granfeldt (2000) *Statistik for biologer og geologer*.

Et vigtigt holdepunkt i statistisk analyse er modelbegrebet. Man vælger en statistisk model, som kan belyse den faglige problemstilling. Det vil sige, at parametrene i modellen kan fortolkes i den faglige problemstilling, og at interessante faglige hypoteser svarer til restriktioner på parametrene. En faglig hypotese afprøves ved at undersøge (teste), om man kan acceptere en reduktion af modellen til en ny model, som er enklere ved at have færre parametre.

Gør man sig det klart, kan man hurtigt lære at analysere temmelig komplicerede problemstillinger korrekt. Ydermere bliver analysen til at følge også for folk, som hverken er specialister på det faglige område eller er professionelle statistikere.

Et tidsvarende brugerkursus i statistik må benytte EDB og en statistisk programpakke. Ved dette kursus er valgt regnearket *Excel* og den statistik pakke der under navnet *Dataanalyse* optræder som et ”tilføjelsesprogram” til *Excel*, men der er ikke benyttet faciliteter, som er specielle for denne statistik pakke, og bogen kan uden vanskelighed anvendes sammen med andre statistiske programpakker. Argumentet for at benytte *Excel* er, at regnearket er tilgængeligt på de fleste PC-er imodsatning til mere kostbare og specialiserede statistiske programpakker såsom for eksempel *SAS*, *Genstat* og *BMDP*. Disse programpakker er designet specielt til brug i forbindelse med statistisk analyse og kan derfor udføre beregningerne i meget mere avancerede statistiske modeller end regnearket *Excel* kan. Disse noter demonstrerer forhåbenligt at i forbindelse med et elementært kursus i statistik er *Excel* et brugbart alternativ.

Når man bruger statistiske programpakker i undervisningen bliver modellerne, som beskrevet ovenfor, det faste holdepunkt når man skal orientere sig i udskrifterne. Man kan bruge en programpakke til statistisk analyse, når man har lært dels at specificere modeller i programpakken og dels at teste reduktionen fra én model til en simplere ved at hente relevante oplysninger ud fra udskrifterne fra estimationen i de to modeller.

Kun få kan lære statistik uden at få metoderne ind gennem fingrene. Vi har derfor valgt både at præsentere, hvordan de enkleste modeller kan regnes på lommeregner, og hvordan de kan regnes ved at orientere sig i udskrifter fra en programpakke. For normalfordelte data vises både for én, to og  $k$  observationsrækker, samt én regressionslinje, hvordan modellerne regnes igennem på lommegner, mens en mere kompliceret model som tosidet variansanalyse kun skal kunne klares med henvisning til programudskrifter.

Et statistikkursus for studerende, der ikke har et vist kendskab til de mest basale begreber i sandsynlighedsteorien, fremstår for os som en umulighed. I Kapitel 2 introduceres og/eller repeteres disse begreber, der illustreres ved en række eksempler, som er valgt ud fra det princip, at de matematisk skulle være lette at håndtere. Kapitel 3 er at betragte som et katalog vedrørende definition af og egenskaber ved de fordelinger som anvendes i forbindelse med de statistiske modeller i de senere kapitler. Kapitel 2 gennemgås efter diskussionen i Kapitel 1 af grafiske og numeriske metoder i forbindelse med beskrivende statistik. Herefter fortsættes med modellerne for normalfordelte data i Kapitel 4 idet de hertil relaterede fordelinger fra Kapitel 3 omtales undervejs. Efter adskillige eksempler på statistisk analyse i forbindelse med normalfordelingen i Kapitel 4 diskuteres hovedtrækkene i en analyse af en parametrisk statistisk model i generelle termer i Kapitel 5. Derefter gennemgås Kapitel 6 om multinomialfordelte data og Kapitel 7 om Poissonfordelte data. Bogen slutter med omtale af nogle simple ikke-parametriske test i Kapitel 8. Som nævnt ovenfor foretrækker vi at betragte parametriske statistiske modeller. Formålet med Kapitel 8 er at orientere læserne om at ikke alle deler denne holdning og for at give et kort indblik i de alternative metoder.

Det vil være muligt at læse kapitlerne i en anden rækkefølge, men man skal være opmærksom på, at de statistiske grundbegreber som nulhypotese, test, testsandsynlighed, signifikansniveau og så videre gennemgås i forbindelse med Afsnit 4.2.

Uden dataeksempler, som udspringer af en faglig problemstilling, bliver en lærebog til et brugerkursus i statistik temmelig uinteressant. En del af eksemplerne er taget fra Andersen (1998) *Statistik for Idrætsstuderende* med forfatterens tilladelse, hvilket vi er taknemmelige for. Vi vil også gerne takke medarbejdere og studerende ved Center for Idræt, Aarhus Universitet og ved Institut for Idræt, Københavns Universitet, som har stillet data og deres historie til rådighed for bogens eksempler og opgaver.

Bogen er blevet brugt ved Idrætsstatistik i efteråret 2000 og bygger på erfaringer fra et lignede kursus i efteråret 1999 og en særlig tak går til Jakob Krabbe Pedersen og Lars Bo Kristensen for deres store indstuds som instruktører på disse to kurser og for deres påvisning af trykfejl.

Bogen er skrevet  $\text{\LaTeX}$ , og Jacob Goldbach har skrevet de stylefiler i  $\text{\LaTeX}$ , som definerer

udsendet af bogen, men derudover har Jacob Goldbach tålmodigt besvaret utallige spørgsmål om  $\text{\LaTeX}$  ligesom Frank Allan Hansen, Niels Væver Hartvig og Michael Kjærgård Sørensen velvilligt har assisteret os.

I forhold til versionen af bogen fra maj 2001 er der rettet en del trykfejl og nogle få figurer er blevet tilføjet. Vi vil gerne takke Lars Madsen for meget kompetent bistand med  $\text{\LaTeX}$  spørgsmål i forbindelse med revisionen og Michael Kjærgård Sørensen for at have produceret de nye figurer.

Århus, august 2005

Preben Blæsild og Jørgen Granfeldt



# Indhold

<b>1</b>	<b>Data og beskrivende statistik</b>	<b>1.1</b>
1.1	Prik- og pindediagrammer . . . . .	1.4
1.2	Histogrammer . . . . .	1.5
1.3	Empiriske størrelser . . . . .	1.7
1.4	Grupperede data . . . . .	1.18
1.5	Kvalitative data . . . . .	1.23
1.6	Flerdimensionale data . . . . .	1.27
	Anneks til Kapitel 1 . . . . .	1.31
	Opgaver til Kapitel 1 . . . . .	1.41
<b>2</b>	<b>Begreber fra sandsynlighedsteorien</b>	<b>2.1</b>
2.1	Sandsynlighedsrum . . . . .	2.1
2.1.1	Definition af sandsynlighedsmål . . . . .	2.1
2.1.2	Regneregler for sandsynligheder . . . . .	2.3
2.1.3	Betingede sandsynligheder og uafhængighed . . . . .	2.6
2.2	Stokastiske variable . . . . .	2.9
2.2.1	Diskrete stokastiske variable . . . . .	2.12
2.2.2	Kontinuerte stokastiske variable . . . . .	2.16
2.3	Stokastiske vektorer . . . . .	2.19
2.3.1	Diskrete stokastiske vektorer . . . . .	2.19
2.3.2	Kontinuerte stokastiske vektorer . . . . .	2.20
2.3.3	Marginale fordelinger . . . . .	2.22
2.3.4	Uafhængighed . . . . .	2.24
2.3.5	Betingede fordelinger . . . . .	2.25
2.4	Middelværdi og varians . . . . .	2.26
	Opgaver til Kapitel 2 . . . . .	2.32

<b>3</b>	<b>Specielle fordelinger</b>	<b>3.1</b>
3.1	Normalfordelingen og relaterede fordelinger . . . . .	3.1
3.1.1	Normalfordelingen . . . . .	3.1
3.1.2	Den todimensionale normalfordeling . . . . .	3.4
3.1.3	$\chi^2$ -fordelingen . . . . .	3.5
3.1.4	$t$ -fordelingen . . . . .	3.8
3.1.5	$F$ -fordelingen . . . . .	3.10
3.2	Diskrete fordelinger . . . . .	3.12
3.2.1	Binomialfordelingen . . . . .	3.12
3.2.2	Multinomialfordelingen . . . . .	3.15
3.2.3	Poissonfordelingen . . . . .	3.16
3.2.4	Den hypergeometriske fordeling . . . . .	3.17
3.2.5	Den negative binomialfordeling . . . . .	3.19
	Opgaver til Kapitel 3 . . . . .	3.22
<b>4</b>	<b>Normalfordelte data</b>	<b>4.1</b>
4.1	Fraktilsammenligning . . . . .	4.2
4.1.1	Ugrupperede observationer . . . . .	4.2
4.1.2	Grupperede data . . . . .	4.7
4.1.3	Transformation . . . . .	4.8
	Anneks til Afsnit 4.1 . . . . .	4.10
4.2	Én observationsrække med kendt varians . . . . .	4.13
	Anneks til Afsnit 4.2 . . . . .	4.19
	Hovedpunkter til Afsnit 4.2 . . . . .	4.20
4.3	Én observationsrække med ukendt varians . . . . .	4.21
	Anneks til Afsnit 4.3 . . . . .	4.28
	Hovedpunkter til Afsnit 4.3 . . . . .	4.30
4.4	To observationsrækker . . . . .	4.32
4.4.1	Test for varianshomogenitet . . . . .	4.35
4.4.2	Ens varians . . . . .	4.38
4.4.3	Forskellig varians . . . . .	4.42
4.4.4	Parrede observationer . . . . .	4.45
	Anneks til Afsnit 4.4 . . . . .	4.50
	Hovedpunkter til Afsnit 4.4 . . . . .	4.55
4.5	$k$ observationsrækker . . . . .	4.59
4.5.1	Test for varianshomogenitet . . . . .	4.61

4.5.2	Test for ens middelværdier . . . . .	4.64
4.5.3	Forskelle og ligheder i behandlingen af to og $k$ observationsrækker . . .	4.68
4.5.4	Notation og test i forbindelse med en følge af modeller . . . . .	4.69
	Anneks til Afsnit 4.5 . . . . .	4.70
	Hovedpunkter til Afsnit 4.5 . . . . .	4.73
4.6	Lineær regression . . . . .	4.76
4.6.1	Lineær regression uden gentagelser . . . . .	4.76
4.6.2	Lineær regression med gentagelser . . . . .	4.83
4.6.3	Hypoteser om regressionsparametrene . . . . .	4.90
4.6.4	Korrelation og/eller regression . . . . .	4.94
	Anneks til Afsnit 4.6 . . . . .	4.101
	Hovedpunkter til Afsnit 4.6 . . . . .	4.106
4.7	Tosidet variansanalyse . . . . .	4.111
	Anneks til Afsnit 4.7 . . . . .	4.135
	Hovedpunkter til Afsnit 4.7 . . . . .	4.137
	Opgaver til Kapitel 4 . . . . .	4.142
<b>Indeks</b>		<b>I.1</b>
<b>5</b>	<b>Statistisk analyse</b>	<b>5.1</b>
5.1	Data . . . . .	5.2
5.2	Modelopstilling . . . . .	5.2
5.3	Modelkontrol . . . . .	5.4
5.4	Statistisk inferens . . . . .	5.5
5.5	Likelihood inferens . . . . .	5.7
5.6	Begreber fra generel testteori . . . . .	5.14
5.7	Approksimativ likelihood teori . . . . .	5.17
5.8	Afsluttende bemærkninger . . . . .	5.22
	Opgaver til Kapitel 5 . . . . .	5.23
<b>6</b>	<b>Multinomialfordelte data</b>	<b>6.1</b>
6.1	Eksempler . . . . .	6.2
6.2	Inferens i én multinomialfordeling. . . . .	6.4
6.2.1	Test af simpel hypotese . . . . .	6.10
6.2.2	Uafhængighed af inddelingskriterier . . . . .	6.11
6.3	Inferens i flere multinomialfordelinger . . . . .	6.15

6.3.1	Homogenitet af flere multinomialfordelinger . . . . .	6.15
6.4	Fishers eksakte test . . . . .	6.19
6.5	Test for goodness of fit . . . . .	6.24
	Anneks til Kapitel 6 . . . . .	6.28
	Hovedpunkter til Kapitel 6 . . . . .	6.31
	Opgaver til Kapitel 6 . . . . .	6.35
<b>7</b>	<b>Poissonfordelte data</b>	<b>7.1</b>
7.1	Eksempler . . . . .	7.2
7.2	Sandsynlighedsteoretiske resultater vedrørende Poissonfordelingen . . . . .	7.3
7.3	Én observationsrække . . . . .	7.7
7.4	Inferens i flere fordelinger . . . . .	7.11
7.4.1	Poissonmodellen med proportionale parametre . . . . .	7.12
7.4.2	Den multiplikative Poissonmodel . . . . .	7.18
	Anneks til Kapitel 7 . . . . .	7.29
	Hovedpunkter til Kapitel 7 . . . . .	7.33
	Opgaver til Kapitel 7 . . . . .	7.39
<b>8</b>	<b>Ikke-parametriske test</b>	<b>8.1</b>
8.1	Fortegnstestet . . . . .	8.2
8.2	Rangtest . . . . .	8.4
8.2.1	Wilcoxons test for én observationsrække . . . . .	8.5
8.2.2	Wilcoxons test for to observationsrækker . . . . .	8.7
8.2.3	Kruskal-Wallis test . . . . .	8.11
	Anneks til Kapitel 8 . . . . .	8.15
	Hovedpunkter til Kapitel 8 . . . . .	8.18
	Opgaver til Kapitel 8 . . . . .	8.21
<b>A</b>	<b>Forskellige matematiske begreber</b>	<b>A.1</b>
A.1	Notation fra mængdelæren . . . . .	A.1
A.2	Rækker . . . . .	A.3
A.3	Dobbeltintegraler og partiel differentiation . . . . .	A.4
A.3.1	Dobbeltintegraler . . . . .	A.5
A.3.2	Partiel differentiation . . . . .	A.5
<b>B</b>	<b>Simulerede fraktildiagrammer</b>	<b>B.1</b>

**C Matematiske symboler**

**C.1**

**D Det græske alfabet**

**D.1**

**Indeks**

**I.1**



## 5 Statistisk analyse

Vi har i Kapitel 4 set adskillige eksempler på statistiske analyser og i disse eksempler er estimer og teststørrelser valgt ud fra heuristiske argumenter. Disse valg er dog baseret på en generel metode, der omtales i dette kapitel. Denne metode kan benyttes i andre situationer, hvor valg af estimatore og teststørrelser ikke kan baseres på heuristiske argumenter.

Kapitlet indeholder en beskrivelse af de vigtigste ingredienser i en statistisk analyse samt en præsentation af de basale matematiske og/eller filosofiske begreber, der ligger til grund for de statistiske metoder, vi betragter i disse noter. Næsten alle de statistiske metoder, der er blevet eller vil blive omtalt i noterne, kan faktisk opfattes som specialtilfælde - eller illustrationer - af den generelle metodik, som diskuteres i dette kapitel. Eneste undtagelse er metoderne i Kapitel 8. Formålet med kapitlet er at fremstille de grundliggende begreber og ideer så overskueligt som muligt, og vi har valgt at gøre dette med reference til teorien for én normalfordelt observationsrække med kendt varians i Afsnit 4.2.

En nybegynder i statistisk analyse kan betragte kapitlet som udstilling af fundamentale begreber i statistisk analyse, som er blevet og også senere vil blive anvendt og illustreret igen og igen. En mere erfaren læser kan derimod betragte kapitlet som et lille opslagsværk vedrørende begreber og terminologi i statistisk analyse.

Afsnit 5.1 vedrører videnskabelige eksperimenter og data. Vi har valgt at fokusere på tre hovedingredienser eller aktiviteter i en statistisk analyse

- i) modelopstilling
- ii) modelkontrol
- iii) statistisk inferens

som omtales i Afsnit 5.2 - 5.4. Statistik inferens baseret på begrebet *likelihood* diskuteres i Afsnit 5.5 og i Afsnit 5.6 omtales nogle få begreber fra den generelle testteori. Approksimative statistiske metoder omtales i Afsnit 5.7 og endelig indeholder Afsnit 5.8 nogle afsluttende bemærkninger.

## 5.1 Data

Udgangspunktet for en statistisk analyse er et *datasæt*  $\mathbf{x}$ , der er resultatet af et *eksperiment*, udført med det formål at få indblik i en speciel *faglig sammenhæng*. Betegnelsen eksperiment skal her forstås i en bred forstand. Data fra idræt kan for eksempel være bestemmelser af kondital, hæmatokritværdier eller andre fysiologiske målinger. Data er ofte indsamlet for at få indblik i, hvorledes træning eller konkurrence påvirker målingerne. En anden form for data er resultater fra konkurrencer, der studeres for at få indsigt i, hvordan forskellige personer eller hold klarer sig i forhold til hinanden eller for at sammenligne præstationer udført under forskellige omstændigheder.

## 5.2 Modelopstilling

Karakteristisk for et datasæt  $\mathbf{x}$  i et eksperiment er, at det er *stokastisk*; det vil sige, at hvis man gentager eksperimentet eller målingerne under lignende omstændigheder, bliver resultatet ikke nødvendigvis  $\mathbf{x}$ . Dette er i modsætning til en deterministisk situation, hvor udfald på forhånd kan bestemmes med sikkerhed. Men selv om udfaldene af eksperimentet ikke kan angives på forhånd er der ofte en regelmæssighed på et højere niveau, som man netop kan erkende, hvis forsøget gentages mange gange. En byggesten i beskrivelsen af et eksperiment er derfor en *sandsynlighedsteoretisk model*.

En sandsynlighedsteoretisk model består af tre komponenter: 1) *udfaldsrummet*,  $\mathcal{X}$ , som er samtlige værdier (udfald), som eksperimentet kan få; 2) *hændelsessystemet*,  $\mathcal{A}$ , som omfatter alle de hændelser vi vil betragte; og 3) *sandsynlighedsmålet*,  $P$ , som angiver sandsynligheden af alle hændelser i  $\mathcal{A}$ .

Det stokastiske element i et eksperiment beskrives af hændelsessystemet og sandsynlighedsmålet, som beskriver alle hændelser vi er interesserede i og deres sandsynligheder. Vi beskriver ofte det stokastiske ved et datasæt ved at opfatte data  $\mathbf{x}$  som en realisation af en stokastisk vektor  $\mathbf{X}$ . Denne stokastiske vektor kan man tænke på som identitetsafbildningen på udfaldsrummet  $\mathcal{X}$  og dens fordeling som givet ved sandsynlighedsmålet  $P$ .

Vi indskrænker os til kun betragte *diskrete* og *kontinuerte* stokastiske vektorer. Hændelsessystemet vil omfatte alle etpunktsmængder, alle intervaller og alle mængder, der kan dannes ud fra dem med de sædvanlige mængdeoperationer, som foreningsmængde, fællesmængde og komplementærmængde. Sandsynlighedsmålene på disse hændelsessystemer kan repræsenteres enten ved deres *fordelingsfunktionen*  $F$  eller deres *tæthedsfunktion*  $f$ .

En *statistisk model* er en parametriseret mængde af sandsynlighedsteoretiske modeller. Sædvanligvis er udfaldsrummene og hændelsessystemerne identiske for alle de sandsynlighedsteo-

retiske modeller, og i det tilfælde kan man tænke på en statistisk model som en sandsynlighedsteoretisk model, hvor sandsynlighedsmålet er blevet erstattet med en *parametriseret* klasse af sandsynlighedsmål,  $\mathcal{P} = \{P_{\boldsymbol{\omega}} | \boldsymbol{\omega} \in \Omega\}$ . Alternativt kan klassen af sandsynlighedsmål repræsenteres med en parametriseret klasse af fordelinger,  $\mathcal{F} = \{F_{\boldsymbol{\omega}} | \boldsymbol{\omega} \in \Omega\}$ , eller en parametriseret klasse af tætheder  $\{f(\cdot; \boldsymbol{\omega}) | \boldsymbol{\omega} \in \Omega\}$ . Her er *parameteren*  $\boldsymbol{\omega} = (\omega_1, \dots, \omega_k)$ , og vi antager altså, at  $\Omega$ , *parameterrummet* (*parametermængden*), er en delmængde af  $R^k$ . Parameteren  $\boldsymbol{\omega}$  bør vælges, således at den er relevant for det faglige problem, der ligger til grund for eksperimentet. Det vil sige, at parameteren skal vælges, således at udsagn vedrørende det faglige problem kan formuleres ved hjælp af  $\boldsymbol{\omega}$ .

Med undtagelse af modellerne i Kapitel 8 er alle de statistiske modeller, der betragtes i disse noter, på formen

$$(\mathcal{X}, \mathcal{A}; \mathcal{P}) = (\mathcal{X}, \mathcal{A}; \{P_{\boldsymbol{\omega}} | \boldsymbol{\omega} \in \Omega\}).$$

Vores foretrukne repræsentation af sandsynlighedsmålene er via tætheder, og vi kalder funktionen

$$\begin{aligned} \mathcal{X} \times \Omega &\rightarrow R \\ (\mathbf{x}, \boldsymbol{\omega}) &\rightarrow f(\mathbf{x}; \boldsymbol{\omega}) \end{aligned} \tag{5.1}$$

for *modelfunktionen*. Modelfunktionen er tætheden som funktion af både udfaldet  $\mathbf{x}$  og parameteren  $\boldsymbol{\omega}$ .

For at gøre de matematiske overvejelser lettere vil vi antage, at parametermængden  $\Omega$  kan vælges som et *område* i  $R^k$ ; det vil sige, at  $\Omega$  er en *åben*<sup>1</sup> og *sammenhængende*<sup>2</sup> delmængde af  $R^k$ .

Vi har nu fået fastlagt de termer og den notation vi vil bruge i omtalen af statistiske modeller. *Modelopstilling* opfatter vi som den proces, hvor man identificerer komponenterne i den statistiske model: udfaldsrum, hændelsessystem og klassen af fordelinger. Det er sædvanligvis uproblematisk at bestemme sig for udfaldsrummet, og dermed er hændelsessystemet også givet. Det væsentligste arbejde er i forbindelse med identifikation af den parametriserede klasse af fordelinger, som man vil betragte. Det betyder også, at man i omtalen af modellerne ofte undlader at nævne hele triplet  $(\mathcal{X}, \mathcal{A}; \{P_{\boldsymbol{\omega}} | \boldsymbol{\omega} \in \Omega\})$ , men fokuserer på fordelingerne  $\{P_{\boldsymbol{\omega}} | \boldsymbol{\omega} \in \Omega\}$ . Endda går man ofte så vidt, at man nøjes med at specificere parametermængden  $\Omega$ , idet både udfaldsrum, hændelsessystem og fordelingsklasse er underforstået.

I arbejdet med at identificere en klasse af fordelinger inddrager man almindelig og specifik viden om forsøgsomstændighederne og undertiden erfaringer fra statistiske analyser af lignende forsøg. Sædvanligvis er de indledende grafiske procedurer, der omtales i Kapitel 1, særdeles

<sup>1</sup> $\Omega$  er åben, hvis et vilkårligt punkt  $\boldsymbol{\omega} \in \Omega$  er centrum for en kugle, der helt er indeholdt i  $\Omega$ .

<sup>2</sup> $\Omega$  er sammenhængende, hvis to vilkårlige punkter  $\boldsymbol{\omega}$  og  $\boldsymbol{\omega}'$  i  $\Omega$  kan forbindes med hinanden ved hjælp af linjestykker, der alle er indeholdt i  $\Omega$ .

nyttige i forbindelse med modelopstilling. Dette trin i en statistisk analyse kræver ofte en så betydelig indsigt i den faglige sammenhæng, at et samarbejde mellem fagmanden fra idræt og statistikeren er påkrævet.

### 5.3 Modelkontrol

Dette punkt i en statistisk analyse vedrører vurdering af rimeligheden af den opstillede statistiske model. Det undersøges, om data  $\mathbf{x}$  strider mod en eller flere væsentlige konsekvenser af modellen. Hvis dette er tilfældet, forkastes modellen og en ny opstilles; hvis ikke, er man klar til at gå videre til næste punkt i analysen, statistisk inferens. Bemærk, at man ved den skitserede procedure på ingen måde opnår sikkerhed for, at modellen er korrekt. Det er vanskeligt at give en generel beskrivelse af dette punkt i en statistisk analyse, idet metoderne dels afhænger af modellen og dels af de betragtede aspekter ved modellen.

Desuden skal det understreges, at modelkontrol ikke er begrænset til de indledende faser af en statistisk undersøgelse. I mange modeller, for eksempel i regressionsmodeller, sker den væsentligste del af modelkontrollen efter, at man har estimeret i modellen.

Som det fremgår af næsten alle de følgende kapitler, indgår såvel grafiske som numeriske undersøgelser i kontrollen af en model.

#### Eksempel 4.1 (Fortsat)

Ved opstillingen af en model for data  $\mathbf{x}$  som består af de 15 målinger  $x_1, \dots, x_{15}$  af laktat koncentrationen i den samme blodprøve med en kendt koncentration på 80 mg/l benytter vi oplysningen om, at erfaringsmæssigt kan sådanne målinger betragtes som normalfordelte med en spredning på 5 mg/l. Vi opfatter derfor de 15 målinger som realisationer af uafhængige og identisk fordelte stokastiske variable  $X_1, \dots, X_{15}$ . Vi betragter altså modellen

$$X_i \sim N(\mu, \sigma_0^2), \quad i = 1, \dots, n,$$

hvor  $n = 15$  og  $\sigma_0^2 = 25$ . Parameteren  $\mu$  varierer i  $R$ , og da de stokastiske variable er uafhængige er modelfunktionen

$$\begin{aligned} f(\mathbf{x}; \mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2\sigma_0^2}(x_i-\mu)^2} \\ &= \left( \frac{1}{2\pi\sigma_0^2} \right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i-\mu)^2}. \end{aligned} \quad (5.2)$$

Modellen kontrolleres ved hjælp af en fraktilsammenligning, som beskrevet i Afsnit 4.1. □

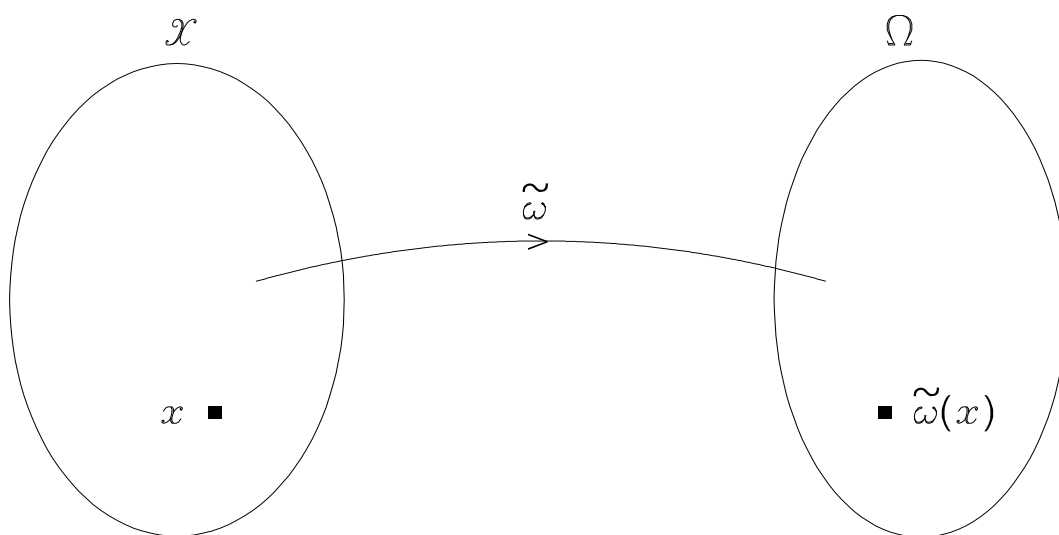
## 5.4 Statistisk inferens

Formålet med en statistisk analyse er at opnå indsigt i den faglige problemstilling, der gav anledning til eksperimentet. Ved modelopstillingen blev parameteren  $\omega$  valgt, således at den repræsenterer de aspekter ved det faglige problem, som er af speciel interesse. Statistisk inferens vedrører spørgsmålet om at formulere udsagn om parameteren  $\omega$  - og dermed om det faglige problem - på baggrund af data  $\mathbf{x}$ , udfaldet af eksperimentet. Disse udsagn har som formål at angive, i hvilken grad de forskellige parameterverdier  $\omega$ , eller rettere de tilsvarende fordelingsfunktioner  $F_\omega$  (eller tæthedsfunktioner  $f(\cdot; \omega)$ ), kan anses for at give en rimelig beskrivelse af data  $\mathbf{x}$ . *Estimationsteori* og *testteori* anses traditionelt som de vigtigste discipliner i statistisk inferens.

I estimationsteorien søges en afbildning

$$\begin{aligned} \tilde{\omega} : \mathcal{X} &\rightarrow \Omega \\ \mathbf{x} &\rightarrow \tilde{\omega}(\mathbf{x}), \end{aligned} \tag{5.3}$$

der til data  $\mathbf{x}$  tilordner en bestemt parameterverdi  $\tilde{\omega}(\mathbf{x})$ , se Figur 5.1. Denne værdi omtales som *estimatet* for (skønnet over) parameteren  $\omega$ . Den tilsvarende stokastiske vektor  $\tilde{\omega}(\mathbf{X})$  omtales som en *estimator* for  $\omega$ . Vi vil ofte bruge notationen  $\tilde{\omega} \rightarrow \omega$  eller  $\omega \leftarrow \tilde{\omega}$  til at antyde, at  $\tilde{\omega}$  er et estimat for  $\omega$ .



**Figur 5.1** Illustration af en estimator  $\tilde{\omega}$ .

Det er ofte en del af en statistisk analyse at undersøge, om en enklere statistiske model end den, der som udgangspunkt blev opstillet, giver en tilfredsstillende beskrivelse af data. Det kan netop være på den måde, man formulerer og besvarer et relevant fagligt spørgsmål. Lad  $\Omega_0$  betegne en delmængde af parameterrummet  $\Omega$ . Hypotesen

$$H_0 : \boldsymbol{\omega} \in \Omega_0 \quad (5.4)$$

repræsenterer da en *reduktion* af den statistiske model. Hvis  $\Omega_0$  kun har ét element  $\boldsymbol{\omega}_0$ , omtales hypotesen som en *simpel hypotese* eller som en *punkthypotese*. I modsat fald betegnes hypotesen som *sammensat*. *Testteorien* angiver metoder til at vurdere, om hypotesen  $H_0$  er rimelig eller ej på grundlag af data  $\mathbf{x}$ . Matematisk set er et test blot en opdeling af værdimængden  $\mathcal{X}$  i to disjunkte mængder

$$R = \{\mathbf{x} \in \mathcal{X} : H_0 \text{ forkastes på grundlag af } \mathbf{x}\} \quad (5.5)$$

$$A = \{\mathbf{x} \in \mathcal{X} : H_0 \text{ forkastes ikke på grundlag af } \mathbf{x}\},$$

der betegnes som henholdsvis *forkastelses-* og *acceptområdet* for  $H_0$ . Mængden  $R$  (ikke at forveksle med de reelle tal  $R$ ) omtales undertiden også som det *kritiske område* for  $H_0$ .

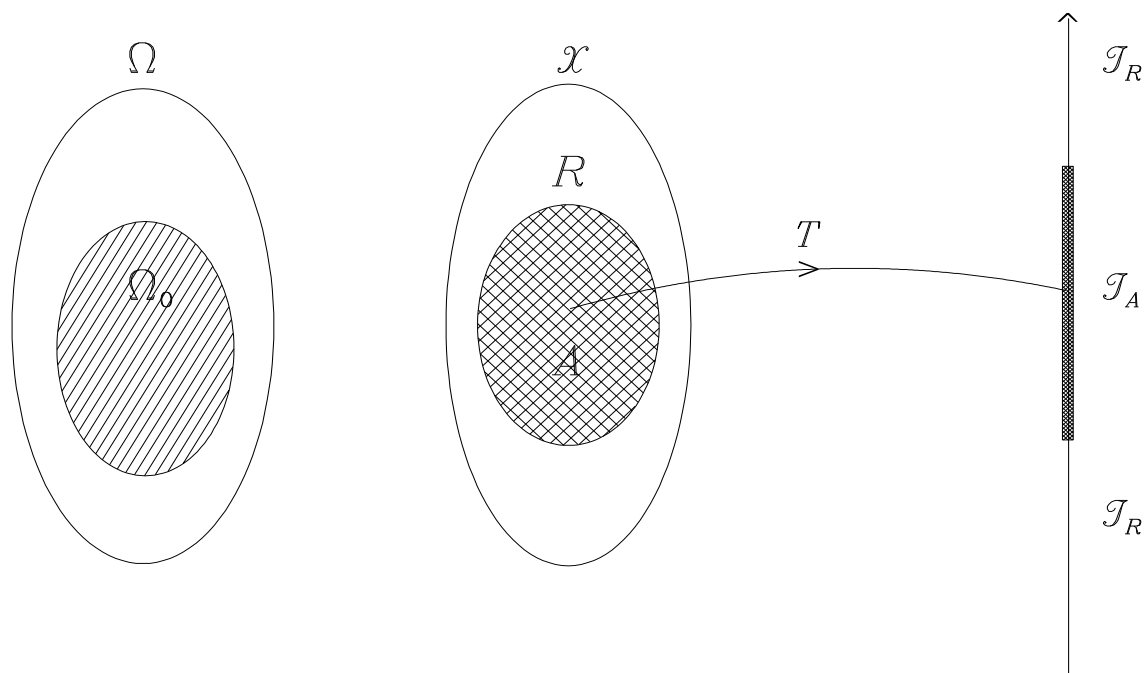
Ofte fås den betragtede opdeling af værdimængden  $\mathcal{X}$  som beskrevet på følgende måde, se også Figur 5.2: Lad  $T$  være en afbildning af  $\mathcal{X}$  ind i de reelle tal og lad  $\mathcal{T}_R$  og  $\mathcal{T}_A$  være en opdeling af værdimængden  $\mathcal{T} = T(\mathcal{X})$  i to disjunkte mængder. Hvis

$$R = T^{-1}(\mathcal{T}_R) = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \in \mathcal{T}_R\} \quad (5.6)$$

$$A = T^{-1}(\mathcal{T}_A) = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \in \mathcal{T}_A\},$$

omtales  $T$  som en *testor* af hypotesen  $H_0$ . Værdien  $T(\mathbf{x})$  af  $T$  svarende til data  $\mathbf{x}$  omtales som *teststørrelsen*.

Ud fra heuristiske argumenter er det ofte muligt at angive estimatorer og testorer i simple, konkrete situationer. Imidlertid er det naturligvis af værdi at have en general metodik, baseret på simple principper, der anviser estimatorer og testorer også i mere komplicerede situationer. Den metodik, vi skal omtale i det følgende, baserer sig på *likelihood funktionen*, som introduceres i det næste afsnit. De hertil hørende størrelser omtales som henholdsvis *maksimum likelihood estimatoren* og *likelihood ratio testoren*.



Figur 5.2 Illustration af en testor  $T$  for hypotesen  $H_0$ .

## 5.5 Likelihood inferens

Ideerne bag likelihood inferens og de første grundlæggende udviklinger af dette begreb skyldes den engelske genetiker R. A. Fisher. Likelihood inferens er baseret på *likelihood funktionen*, som vi nu introducerer og diskuterer.

Fra formuleringen af den statistiske model i Afsnit 5.2 ses det, at for fast værdi af parameteren  $\boldsymbol{\omega}$  er modelfunktionen  $f(\mathbf{x}; \boldsymbol{\omega})$  tæthedsfunktionen for den stokastiske vektor  $\mathbf{X}$ . Hvis  $P_{\boldsymbol{\omega}}$  betegner sandsynlighedsmålet svarende til tæthedsfunktionen  $f(\mathbf{x}; \boldsymbol{\omega})$  har vi derfor, at

$$f(\mathbf{x}; \boldsymbol{\omega}) = P_{\boldsymbol{\omega}}(\mathbf{X} = \mathbf{x}), \quad (5.7)$$

hvis  $\mathbf{X}$  er diskret. Hvis  $\mathbf{X}$  er kontinuert er relationen mellem  $f(\mathbf{x}; \boldsymbol{\omega})$  og  $P_{\boldsymbol{\omega}}$  givet ved

$$f(\mathbf{x}; \boldsymbol{\omega}) d\mathbf{x} \approx P_{\boldsymbol{\omega}}(\mathbf{X} \in I_{\mathbf{x}}), \quad (5.8)$$

hvor  $I_{\mathbf{x}}$  er en lille mængde omkring  $\mathbf{x}$ , hvis indhold er  $d\mathbf{x}$ .

For fast værdi af  $\boldsymbol{\omega}$  beskriver modelfunktionen altså sandsynlighederne knyttet til alle mulige realisationer af  $\mathbf{X}$ . Data  $\mathbf{x}$  er imidlertid en bestemt og fast realisation af  $\mathbf{X}$ , og da vi ønsker at udtale os om forskellige værdier af  $\boldsymbol{\omega}$  i lys af data  $\mathbf{x}$ , kunne vi prøve at betragte modelfunktionen som funktion af  $\boldsymbol{\omega}$  for fastholdt  $\mathbf{x}$ . Vi har da stadig fortolkningen, at  $f(\mathbf{x}; \boldsymbol{\omega})$  er sandsynligheden af observationen  $\mathbf{x}$ , hvis parameteren er  $\boldsymbol{\omega}$ . Det har vi direkte via (5.7), hvis  $\mathbf{X}$  er diskret, eller

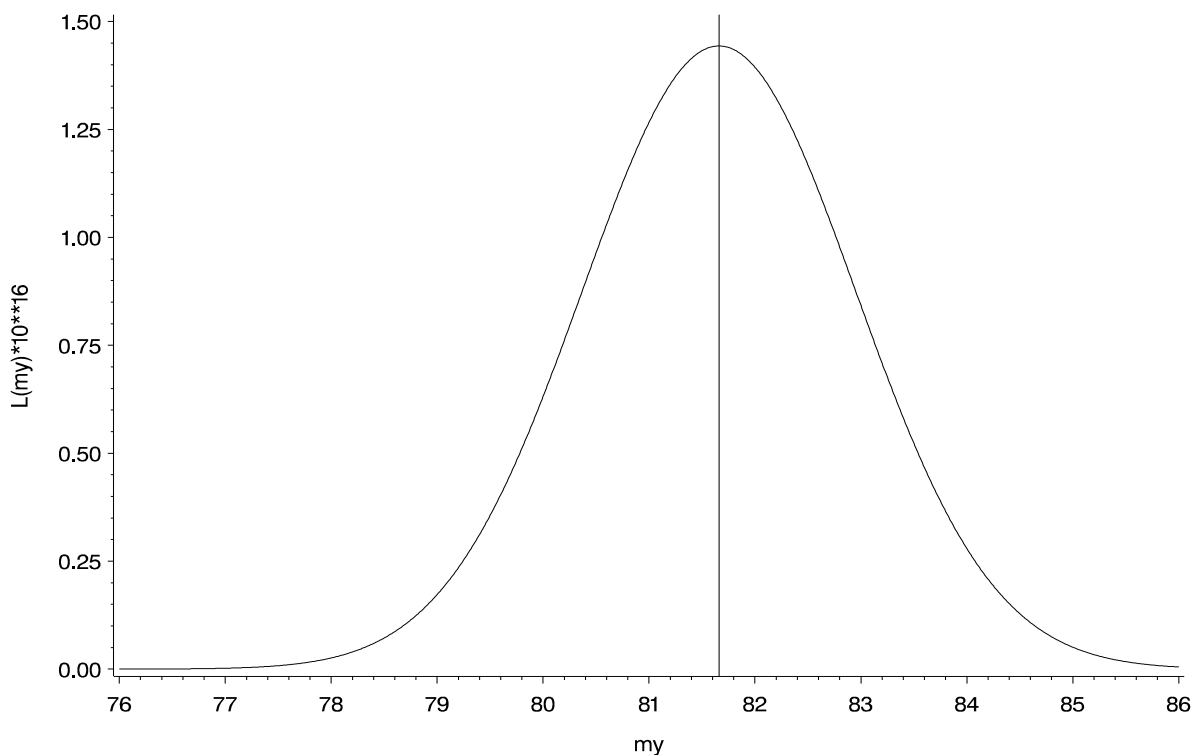
via fortolkningen i (5.8), hvis  $\mathbf{X}$  er kontinuert. I den forstand er  $f(\mathbf{x}; \boldsymbol{\omega})$  et udtryk for troligheden eller rimeligheden af  $\boldsymbol{\omega}$  i lys af data  $\mathbf{x}$ . R.A. Fisher valgte termen *likelihood*, fordi likelihood i lighed med probability i engelsk daglig tale bruges til at udtrykke grader af tiltro. Ved at vælge en anden term end probability understregede Fisher, at vi ikke har at gøre med sandsynligheder på parametrene.

Termen likelihood er ikke oversat til dansk, og vi kalder  $f(\mathbf{x}; \boldsymbol{\omega})$  som funktion af  $\boldsymbol{\omega}$  for *likelihood funktionen* og betegner den

$$L(\boldsymbol{\omega}) = f(\mathbf{x}; \boldsymbol{\omega}) \quad \boldsymbol{\omega} \in \Omega, \quad (5.9)$$

idet vi underforstår afhængigheden af de observerede data. Men hvis vi ønsker at understrege, at vi betragter funktionen svarende til data  $\mathbf{x}$ , skriver vi  $L(\boldsymbol{\omega}; \mathbf{x})$  i stedet for  $L(\boldsymbol{\omega})$ .

Et eksempel på en likelihood funktion kan ses i Figur 5.3.



**Figur 5.3** Likelihood funktionen  $L(\mu)$  (gandet med  $10^{16}$ ) for middelværdien  $\mu$  i en normalfordelt observationsrække med kendt varians ( $\sigma_0^2 = 25$ ) for data i Eksempel 4.1.

Likelihood funktionen laver en ordening i parametermængden. Hvis vi et øjeblik betragter kun to parameterverdier  $\boldsymbol{\omega}_1$  og  $\boldsymbol{\omega}_2$ , og på baggrund af data  $\mathbf{x}$  ønsker at vælge, hvilken af de to parameterverdier, der bedst forklarer data, må det blive den, som har den største værdi af likelihood funktionen  $L(\boldsymbol{\omega})$ , fordi det er den som gør data mest sandsynlig. Vi siger, at værdien  $\boldsymbol{\omega}_1$  er mere *likely* end  $\boldsymbol{\omega}_2$  i lys af data  $\mathbf{x}$ , hvis  $L(\boldsymbol{\omega}_1) > L(\boldsymbol{\omega}_2)$ . På dansk vil vi undertiden bruge

ordet *trolig* i denne tekniske betydning, og altså sige, at  $\boldsymbol{\omega}_1$  er mere *trolig* end  $\boldsymbol{\omega}_2$  i lys af data  $\mathbf{x}$ , hvis  $L(\boldsymbol{\omega}_1) > L(\boldsymbol{\omega}_2)$ .

Likelihood funktionens ordning af parametermængden leder umiddelbart til, at hvis vi vil angive én parameterværdi, som er i bedst overensstemmelse med data  $\mathbf{x}$ , må det blive den værdi, som gør de observerede data mest sandsynlige, det vil sige den værdi, hvor likelihood funktionen antager sit maksimum. Vi har hermed introduceret begrebet *maksimum likelihood estimation*. Hvis der eksisterer en entydigt bestemt værdi  $\hat{\boldsymbol{\omega}}$ , for hvilken likelihood funktionen  $L(\cdot)$  antager sit maksimum, det vil sige

$$L(\hat{\boldsymbol{\omega}}) > L(\boldsymbol{\omega}) \quad \text{for alle } \boldsymbol{\omega} \in \Omega \text{ således at } \boldsymbol{\omega} \neq \hat{\boldsymbol{\omega}},$$

kaldes denne værdi  $\hat{\boldsymbol{\omega}}$  af parameteren for *maksimum likelihood estimatet* for  $\boldsymbol{\omega}$ . Med andre ord er maksimum likelihood estimatet  $\hat{\boldsymbol{\omega}} = (\hat{\boldsymbol{\omega}}(\mathbf{x}))$  den mest trolige værdi af parameteren  $\boldsymbol{\omega}$  i lys af data  $\mathbf{x}$ . Den tilsvarende stokastiske vektor  $\hat{\boldsymbol{\omega}}(\mathbf{X})$  omtales som *maksimum likelihood estimatoren*.

Undertiden er det lettere at maksimere *log likelihood funktionen*

$$l(\boldsymbol{\omega}) = \ln L(\boldsymbol{\omega}) \quad \boldsymbol{\omega} \in \Omega, \quad (5.10)$$

end selve likelihood funktionen  $L(\cdot)$ . I de modeller, vi betragter, er likelihood funktionen (mindst) to gange differentiabel med kontinuerte (partielle) afledede, og det letter arbejdet med at finde den værdi, hvor likelihood funktionen antager sit maksimum. Da parametermængden er antaget at være et område, kan  $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \dots, \hat{\omega}_k)$  findes som en løsning til ligningerne

$$\frac{\partial l}{\partial \omega_j}(\boldsymbol{\omega}) = 0, \quad j = 1, 2, \dots, k. \quad (5.11)$$

Disse ligninger, der kaldes *likelihood ligningerne*, kan undertiden løses eksplicit, men i nogle tilfælde må man benytte numeriske procedurer for at finde  $\hat{\boldsymbol{\omega}}$ . Desuden må man også vurdere om en løsning til likelihood ligningerne er et punkt, hvor likelihood funktionen antager sit maksimum.

Ofte består data  $\mathbf{x}$  af  $n$  enkeltmålinger  $x_1, \dots, x_n$ , det vil sige  $\mathbf{x} = (x_1, \dots, x_n)$ . Hvis vi som model kan benytte, at  $x_1, \dots, x_n$  er udfald af uafhængige og identisk fordelte stokastiske variable  $X_1, \dots, X_n$ , hvor tæthedsfunktionen for  $X_i$  er  $f(x_i; \boldsymbol{\omega}), i = 1, \dots, n$ , vil vi omtale data som én *observationsrække fra fordelingen*  $F_{\boldsymbol{\omega}}$ . Antagelsen om uafhængighed af de stokastiske variable medfører - som bekendt fra sandsynlighedsteorien - at tæthedsfunktionen for  $\mathbf{X}$  er produktet af tæthedsfunktionerne for  $X_i, i = 1, \dots, n$ . Likelihood funktionen  $L(\cdot)$  og log likelihood funktionen  $l(\cdot)$  bliver derfor i denne situation henholdsvis

$$L(\boldsymbol{\omega}) = \prod_{i=1}^n f(x_i; \boldsymbol{\omega}) \quad (5.12)$$

og

$$l(\boldsymbol{\omega}) = \sum_{i=1}^n \ln f(x_i; \boldsymbol{\omega}). \quad (5.13)$$

### Eksempel 4.1 (Fortsat)

Af (5.2) ses, at likelihoodfunktionen for  $\mu$  er

$$L(\mu) = \left( \frac{1}{2\pi\sigma_0^2} \right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2}, \quad (5.14)$$

se Figur 5.3, og dermed at log likelihood funktionen for  $\mu$  er

$$l(\mu) = -\frac{n}{2} \ln(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2. \quad (5.15)$$

Differentieres log likelihood funktionen  $l$  i (5.15) én gang med hensyn til  $\mu$  og sættes lig med 0, fås likelihood ligningen

$$0 = \frac{dl}{d\mu}(\mu) = \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \mu).$$

Løses ligningen med hensyn til  $\mu$  fås løsningen

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

som maksimerer  $l$ . Maksimum likelihood estimatet for middelværdien  $\mu$  er gennemsnittet af observationerne. Som nævnt i forbindelse med Eksempel 4.1 er dette et intuitivt rimeligt estimat. Det er en realisation af den stokastiske variabel

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma_0^2}{n}\right),$$

som har den rigtige middelværdi  $\mu$  og en varians  $\sigma_0^2/n$ , som aftager med antallet af observationer. □

Vi giver nu en diskussion af testteori baseret på likelihood funktionen. Indledningsvis bemærker vi to ting. For det første er værdien af likelihood funktionen beregnet i maksimum likelihood estimatet,  $L(\hat{\boldsymbol{\omega}}(\mathbf{x}))$ , en funktion af data  $\mathbf{x}$ , og dermed en stokastisk variabel. For det andet er fortolkningen af  $L(\hat{\boldsymbol{\omega}}(\mathbf{x}))$ , at det er den maksimale sandsynlighed for data  $\mathbf{x}$  i den givne statistiske model.

Antag, at vi i en statistisk model med parametermængde  $\Omega$  og på grundlag af data  $\mathbf{x}$  ønsker at undersøge, om data kan beskrives med delmodellen  $\Omega_0$ , hvor  $\Omega_0$  betegner en delmængde af  $\Omega$ , det vil sige  $\Omega_0 \subseteq \Omega$ .

Vi bruger sprogbroen, at vi ønsker at teste hypotesen

$$H_0 : \boldsymbol{\omega} \in \Omega_0.$$

Lad  $\hat{\boldsymbol{\omega}}(\mathbf{x})$  betegne maksimum likelihood estimatet for  $\boldsymbol{\omega}$  i den oprindelige model, og lad  $\hat{\boldsymbol{\omega}}_0(\mathbf{x})$  betegne maksimum likelihood estimatet for  $\boldsymbol{\omega}$  under  $H_0$ , det vil sige i den statistiske model med parametermængde  $\Omega_0$ . *Likelihood ratio teststørrelsen*  $Q(\mathbf{x})$  defineres da som:

$$Q(\mathbf{x}) = \frac{\max_{\boldsymbol{\omega} \in \Omega_0} L(\boldsymbol{\omega})}{\max_{\boldsymbol{\omega} \in \Omega} L(\boldsymbol{\omega})} = \frac{L(\hat{\boldsymbol{\omega}}_0(\mathbf{x}))}{L(\hat{\boldsymbol{\omega}}(\mathbf{x}))}. \quad (5.16)$$

Man bemærker, at  $Q(\mathbf{x}) \leq 1$  fordi det er samme funktion, der maksimeres i tæller og nævner og at der maksimeres over en mindre mængde i tælleren. Desuden er  $0 < Q(\mathbf{x})$ , da  $Q(\mathbf{x})$  er et forhold mellem to sandsynligheder. Alt i alt er altså  $0 < Q(\mathbf{x}) \leq 1$ .

Vi ser dernæst på fortolkningen af likelihood ratio teststørrelsen.  $Q(\mathbf{x})$  er troværdighedsforholdet mellem den mest trolige værdi  $\hat{\boldsymbol{\omega}}_0$  af  $\boldsymbol{\omega}$  under  $H_0$  og den mest trolige værdi  $\hat{\boldsymbol{\omega}}$  af  $\boldsymbol{\omega}$  overhovedet. Hvis  $Q(\mathbf{x}) \approx 1$  er  $L(\hat{\boldsymbol{\omega}}_0) \approx L(\hat{\boldsymbol{\omega}})$ ; der eksisterer altså en værdi af parameteren under hypotesen, der er næsten ligeså trolig som den mest trolige værdi overhovedet, og vi har derfor ingen grund til at betvivle  $H_0$  i denne situation. Hvis derimod  $Q(\mathbf{x}) \approx 0$  er  $L(\hat{\boldsymbol{\omega}}_0) \ll L(\hat{\boldsymbol{\omega}})$ ; den mest trolige værdi under hypotesen er altså meget mindre trolig end den mest trolige værdi overhovedet, og derfor må vi betvivle  $H_0$ . Med andre ord, *observationen  $\mathbf{x}$  er kritisk for  $H_0$  hvis  $Q(\mathbf{x})$  er lille.*

Helt på samme måde som likelihood funktionen lavede en ordning i parametermængden  $\Omega$ , laver likelihood ratio teststørrelsen en ordning i udfaldsrummet  $\mathcal{X}$ . Vi siger, at  $\mathbf{x}_1$  er *mere (eller sige så) kritisk* for  $H_0$  som  $\mathbf{x}_2$ , hvis

$$Q(\mathbf{x}_1) \leq Q(\mathbf{x}_2).$$

Begrundelsen er, at  $Q(\mathbf{x})$  er forholdet mellem den maksimale sandsynlighed for data under hypotesen relativt til den maksimale sandsynlighed under modellen.

For at få et indtryk af hvor lille  $Q(\mathbf{x})$  skal være, før vi forkaster  $H_0$ , betragtes mængden af alle mulige udfald  $\mathbf{y}$  af eksperimentet, som er mindst lige så kritiske for  $H_0$  som det observerede udfald  $\mathbf{x}$ , det vil sige mængden

$$\{\mathbf{y} \in \mathcal{X} : Q(\mathbf{y}) \leq Q(\mathbf{x})\}. \quad (5.17)$$

For at vurdere størrelsen af mængden i (5.17) relativt til størrelsen af  $\mathcal{X}$  benytter vi sandsynlighedsteorien.

Hvis hypotesen  $H_0$  er simpel, det vil sige hvis  $\Omega_0 = \{\omega_0\}$ , omtales sandsynligheden for mængden i (5.17),

$$\varepsilon(\mathbf{x}) = P_{\omega_0}(\{\mathbf{y} \in \mathcal{X} : Q(\mathbf{y}) \leq Q(\mathbf{x})\}), \quad (5.18)$$

som *testsandsynligheden* for likelihood ratio testet. (Synonymt bruges betegnelserne *det observerede signifikansniveau* eller *p-værdien*.) Det ses af (5.18), at testsandsynligheden er *sandsynligheden - beregnet under  $H_0$  - for de mulige udfald  $\mathbf{y}$ , der er mindst lige så kritiske for  $H_0$  som det observerede udfald  $\mathbf{x}$* . Er  $\varepsilon(\mathbf{x})$  lille, er der således ikke stor sandsynlighed for at få udfald, der er mindst lige så kritiske for  $H_0$  som det observerede udfald  $\mathbf{x}$ , og derfor forkaster vi  $H_0$ . Altså *hvis  $\varepsilon(\mathbf{x})$  er lille forkastes  $H_0$* . Er  $\varepsilon(\mathbf{x})$  stor, er der stor sandsynlighed for udfald, der er mindst lige så kritiske som  $\mathbf{x}$ , og følgelig er der ingen grund til at forkaste  $H_0$ ; vi siger da, at  $H_0$  accepteres. Altså *hvis  $\varepsilon(\mathbf{x})$  er stor accepteres  $H_0$* . Bemærk, at accept af  $H_0$  på ingen måde betyder, at vi har bevist (i matematisk forstand) rigtigheden af  $H_0$ , men blot at vi i det nærværende forsøg ikke har kunnet konstatere *signifikante* (betydningsfulde) afvigelser fra  $H_0$ .

Lad os for en sikkerheds skyld fremhæve logikken bag det argument, der implicerer, at hypotesen  $H_0$  forkastes, hvis testsandsynligheden  $\varepsilon(\mathbf{x})$  er lille. Statistikerens betragter to præmisser: 1) 'enten er hypotesen  $H_0$  falsk eller også er en hændelse med lille sandsynlighed indtruffet' og 2) 'en hændelse med lille sandsynlighed indtræffer ikke'. Ud fra disse to præmisser drages konklusionen 'hypotesen  $H_0$  er falsk'.

Et spørgsmål er stadig ubesvaret. Hvor lille skal testsandsynligheden  $\varepsilon(\mathbf{x})$  være, før vi forkaster  $H_0$ ? Principielt afhænger svaret af hypotesens natur. For eksperimentelt at afvise velrenommerede videnskabelige hypoteser, som for eksempel Newtons 2. lov, kræves, at der konstateres stærkt signifikante afvigelser fra hypotesen, det vil sige at testsandsynligheden skal være meget lille, for eksempel 0.1%. Mindre velbegrundede hypoteser, såsom at koncentrationen af laktat i en blodprøve er 80mg/l, forkastes for langt større testsandsynligheder (1 eller 5%).

Likelihood ratio testet med *signifikansniveau*  $\alpha$  forkaster hypotesen  $H_0$ , hvis

$$\varepsilon(\mathbf{x}) \leq \alpha, \quad (5.19)$$

hvilket medfører, at det tilsvarende forkastelsesområde (eller kritiske område)  $R_\alpha$  er

$$R_\alpha = \{\mathbf{x} \in \mathcal{X} : \varepsilon(\mathbf{x}) \leq \alpha\} \quad (5.20)$$

samt at det tilsvarende acceptområde er

$$A_\alpha = \{\mathbf{x} \in \mathcal{X} : \varepsilon(\mathbf{x}) > \alpha\}. \quad (5.21)$$

I den statistiske litteratur er det foretrukne signifikansniveau traditionelt 5%, men også niveauet 1% benyttes i forbindelse med mere velbegrundede hypoteser. I dette kursus vil vi i forbindelse med eksempler og opgaver benytte test på 5%-niveau, medmindre andet er nævnt.

Vi afslutter dette afsnit med nogle bemærkninger vedrørende likelihood ratio testet i det tilfælde, hvor hypotesen  $H_0$  er sammensat, det vil sige hvor delmængden  $\Omega_0$ , der specificerer hypotesen, har mere end ét element. I dette tilfælde defineres testsandsynligheden for likelihood ratio testet som

$$\varepsilon(\mathbf{x}) = \sup_{\boldsymbol{\omega} \in \Omega_0} P_{\boldsymbol{\omega}}(\{\mathbf{y} \in \mathcal{X} : Q(\mathbf{y}) \leq Q(\mathbf{x})\}), \quad (5.22)$$

altså som den største af sandsynlighederne - under  $H_0$  - for mængden i (5.17). Forkastelses- og acceptområdet defineres også i dette tilfælde som i formlerne (5.20) og (5.21).

Temmelig ofte er det vanskeligt (eller umuligt) at beregne de eksakte værdier af testsandsynligheden for likelihood ratio testet som defineret i (5.18) eller (5.22). I Afsnit 5.7 diskuterer vi, hvorledes man beregner approksimationer for testsandsynlighederne i sådanne situationer.

#### Eksempel 4.1 (Fortsat)

Vi betragter nu test af hypotesen  $H_0 : \mu = \mu_0 = 80$ .

Likelihood ratio teststørrelsen  $Q(\mathbf{x})$  er forholdet mellem maksimum af likelihood funktionen under  $H_0$  og maksimum af likelihood funktionen uden  $H_0$ 's begrænsning.

$$Q(\mathbf{x}) = \frac{\max_{\mu \in H_0} L(\mu)}{\max_{\mu \in R} L(\mu)} = \frac{L(\mu_0)}{L(\bar{x})}.$$

Hvis  $Q(\mathbf{x})$  er meget lille, forklares observationen meget dårligere under  $H_0$  end under den oprindelige model uden restriktioner på  $\mu$ . Så de værdier, der er mere kritiske for  $H_0$  end observationen  $\mathbf{x}$ , er  $\{\mathbf{y} \mid Q(\mathbf{y}) \leq Q(\mathbf{x})\}$ . Igen ser man af tekniske grunde på  $\ln Q$ .

$$\begin{aligned} \ln Q(\mathbf{x}) &= l(\mu_0) - l(\bar{x}) \\ &= -\frac{1}{2\sigma_0^2} \left[ \sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= -\frac{n(\bar{x} - \mu_0)^2}{2\sigma_0^2} \\ &= -\frac{1}{2} u^2(\mathbf{x}), \end{aligned} \quad (5.23)$$

hvor  $u(\mathbf{x})$  netop er teststørrelsen (4.5), som blev udledt i Eksempel 4.1. De observationer, som er mere kritiske for  $H_0$  end observationen  $\mathbf{x}$  er

$$\begin{aligned} \{\mathbf{y} \mid Q(\mathbf{y}) \leq Q(\mathbf{x})\} &= \{\mathbf{y} \mid -2\ln Q(\mathbf{y}) \geq -2\ln Q(\mathbf{x})\} \\ &= \{\mathbf{y} \mid u^2(\mathbf{y}) \geq u^2(\mathbf{x})\} \\ &= \{\mathbf{y} \mid |u(\mathbf{y})| \geq |u(\mathbf{x})|\}. \end{aligned}$$

og man ser, at likelihood ratio testet for  $H_0$  er det samme som testet baseret på (4.5)

$$u(\mathbf{x}) = u(x_1, \dots, x_n) = \frac{\bar{x} - \mu_0}{\sqrt{\sigma_0^2/n}}.$$

□

## 5.6 Begreber fra generel testteori

I Afsnit 5.5 har vi diskuteret et specielt signifikanstest, nemlig likelihood ratio testet. Som nævnt er det undertiden vanskeligt at finde testsandsynligheden for dette test og derfor også de tilsvarende forkastelses- og acceptområder. I sådanne situationer betragter man sommetider alternative teststørrelser, der findes ved hjælp af heuristiske argumenter og/eller sandsynlighedsteoretiske overvejelser. I dette afsnit giver vi en kortfattet omtale af egenskaber ved en generel teststørrelse  $T$  som defineret i Afsnit 5.4. Bemærkningerne er derfor gyldige for såvel likelihood ratio testet som for de alternative test.

Signifikanstestet af hypotesen  $H_0 : \boldsymbol{\omega} \in \Omega_0$  svarende til testoren  $T$  siges at have *signifikansniveau*  $\alpha$  (eller kort,  $T$  er et *test på niveau*  $\alpha$ ), hvis

$$\sup_{\boldsymbol{\omega} \in \Omega_0} P_{\boldsymbol{\omega}}(\mathbf{X} \in R) = \alpha, \quad (5.24)$$

altså hvis den største sandsynlighed for at forkaste  $H_0$ , det vil sige den største sandsynlighed for at  $\mathbf{X}$  tilhører det kritiske område  $R$  - beregnet under  $H_0$  - er  $\alpha$ . Med andre ord, signifikansniveauet  $\alpha$  for et test er mål for risikoen for at forkaste en sand hypotese. Det er indlysende, at det ville være ønskeligt, at  $\alpha$  var 0, men sådanne signifikanstest findes ikke.

Det er karakteristisk for statistisk inferens, at det *ikke med sikkerhed er muligt* at udtale sig, om hypotesen  $H_0$  er sand eller falsk. På dette punkt adskiller statistisk inferens sig fra matematik og logik. I de to sidstnævnte discipliner drager man konklusioner på grundlag af faste præmisser. I statistisk inferens drager man konklusioner på grundlag af data, der betragtes som en realisation af en stokastisk vektor, hvis variation beskrives ved hjælp af en sandsynlighedsteoretisk model. Konklusionerne i statistisk inferens formuleres derfor - naturligvis - ved hjælp af sandsynlighedsteorien.

En anden vigtig forskel mellem de tre discipliner består i, at matematik og logik er *deduktive*, det vil sige, at de slutter fra det generelle til det specielle. I modsætning hertil er statistisk inferens *induktiv*, idet man her slutter fra det specielle (data) til det generelle (en videnskabelig model).

I forbindelse med testteori taler man undertiden om fejl af type I og type II. Disse fremgår af Tabel 5.1. Bemærk, at sandsynligheden for at begå en fejl af type I præcis er signifikansniveauet  $\alpha$ .

Kvaliteten af et statistisk test afhænger blandt andet af dets evne til at afsløre signifikante afvigelser fra hypotesen  $H_0$ , hvilket kan udtrykkes ved *styrkefunktionen* for testet. Med betegnelserne fra Afsnit 5.4 er styrkefunktionen for testoren  $T$  af hypotesen  $H_0 : \boldsymbol{\omega} \in \Omega_0$  defineret som

$$pow(\boldsymbol{\omega}) = P_{\boldsymbol{\omega}}(T \in \mathcal{T}_R),$$

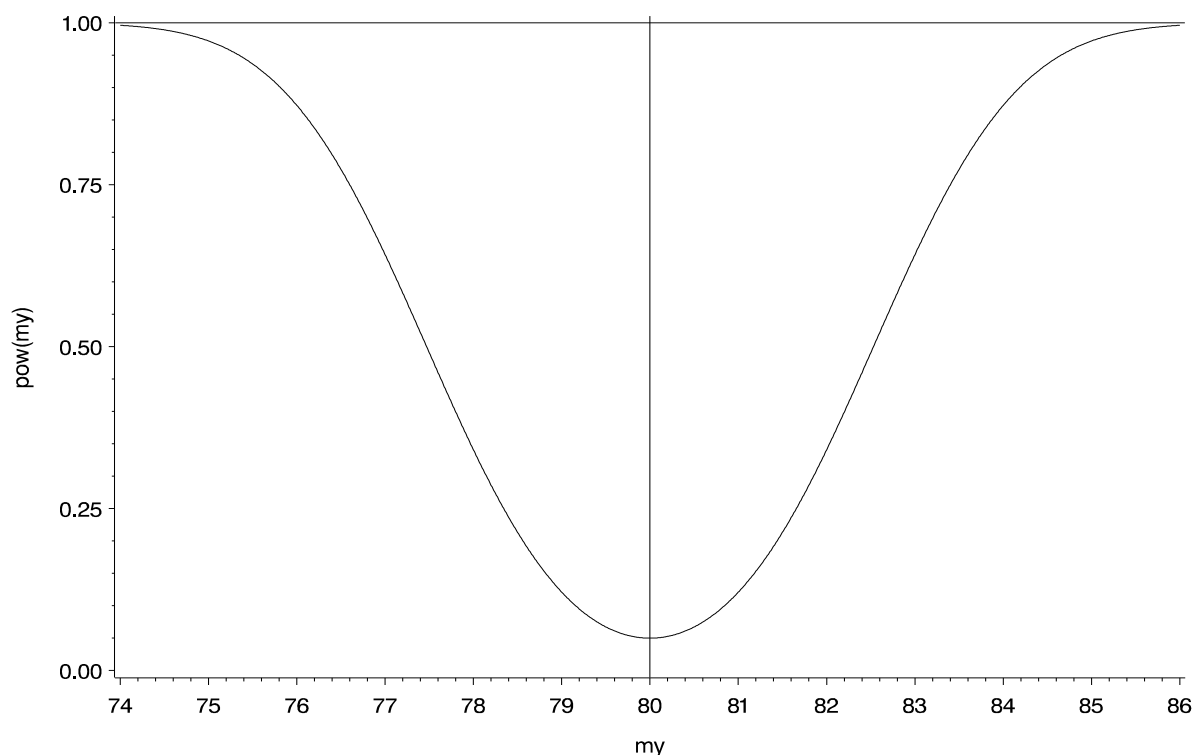
	$H_0$ forkastes	$H_0$ accepteres
$H_0$ sand	type I	ingen
$H_0$ falsk	ingen	type II

**Tabel 5.1** De forskellige typer af fejl i testteorien.

det vil sige, at for enhver værdi af parameteren  $\omega$  er styrken  $pow(\omega)$  sandsynligheden - beregnet ved hjælp af sandsynlighedsmålet svarende til  $\omega$  - for at forkaste hypotesen  $H_0$ . Bemærk, at hvis hypotesen er simpel,  $\Omega_0 = \{\omega_0\}$ , så er  $pow(\omega_0)$  netop lig med signifikansniveauet  $\alpha$ , samt at hvis vi for  $\omega \neq \omega_0$  lader  $\beta(\omega)$  betegne sandsynlighed for fejl af type II - svarende til parameterværdien  $\omega$  - så er

$$\beta(\omega) = 1 - pow(\omega).$$

Ideelt set burde værdien af styrkefunktionen for en simpel hypotese  $H_0 : \omega = \omega_0$  derfor være konstant lig med 1 med undtagelse af værdien i  $\omega_0$ , som burde være 0. Som nævnt ovenfor findes der imidlertid ikke testorer med en sådan styrkefunktion. Et eksempel på en styrkefunktion er vist i Figur 5.4.



**Figur 5.4** Styrkefunktionen for  $u$ -testet på niveau 5% for hypotesen  $H_0 : \mu = 80$ . Standardafvigelsen  $\sigma$  er 5, svarende til problemstillingen i Eksempel 4.1.

Vi afslutter dette afsnit med at omtale konfidensområder, som er et begreb, hvis definition

er relateret til testteorien og som ofte benyttes i anvendelser. I lys af data  $\mathbf{x}$  er  $(1 - \alpha)$  konfidensområdet for parameteren  $\boldsymbol{\omega}$  defineret som

$$C_{1-\alpha}(\mathbf{x}) = \{\boldsymbol{\omega}_0 \mid \text{hypotesen } H_0 : \boldsymbol{\omega} = \boldsymbol{\omega}_0 \text{ accepteres ved et signifikantest på niveau } \alpha \text{ på grundlag af data } \mathbf{x}\}. \quad (5.25)$$

Hvis parameteren er en-dimensional er området typisk et interval,  $(1 - \alpha)$  konfidensintervallet.

Der er indlysende, at konfidensområdet afhænger af det betragtede test samt det valgte signifikansniveau. Test udføres sædvanligvis på niveau 5%, og de tilsvarende områder er i så tilfælde 95% konfidensområder.

En fortolkning af 95% konfidensområder baserer sig på fortolkningen af sandsynligheder som grænseværdier af relative hyppigheder. Antag, at eksperimentet, der resulterede i data  $\mathbf{x}$ , blev gentaget et uendeligt antal gange og antag, at man for resultatet  $\mathbf{y}$  af hver gentagelse af eksperimentet beregnede området  $C_{1-\alpha}(\mathbf{y})$ . Den sande værdi af parameteren  $\boldsymbol{\omega}$  ville da være indeholdt i det beregnede område i 95% af gentagelserne.

Denne fortolkning er naturligvis ikke så gavnlige, når man står med sit interval  $C_{1-\alpha}(\mathbf{x})$  beregnet på grundlag af data  $\mathbf{x}$ . Men det er samme fortolkning, som vi har mødt i forbindelse med test. Enten omfatter intervallet  $C_{1-\alpha}(\mathbf{x})$  den sande parameter eller også er der indtruffet en hændelse med en sandsynlighed mindre end  $\alpha$ .

Undertiden omtales konfidensintervallet for  $\boldsymbol{\omega}$  som *intervalestimatet* for  $\boldsymbol{\omega}$ . Et sædvanligt estimat, for eksempel maksimum likelihood estimatet  $\hat{\boldsymbol{\omega}}(\mathbf{x})$ , udpeger kun én værdi af parameteren i lys af data  $\mathbf{x}$ . Konfidensintervallet eller intervalestimatet  $C_{1-\alpha}(\mathbf{x})$  er i praksis værdifuldt, fordi det ikke blot udpeger en enkelt værdi af  $\boldsymbol{\omega}$  men er et udtryk for, hvor meget information data  $\mathbf{x}$  indeholder vedrørende den ukendte parameter  $\boldsymbol{\omega}$ . Hvis konfidensintervallet er stort, er der mange værdier af parameteren  $\boldsymbol{\omega}$ , der giver en rimelig beskrivelse af data  $\mathbf{x}$ , og i så tilfælde indeholder  $\mathbf{x}$  begrænset information om  $\boldsymbol{\omega}$ . Hvis derimod konfidensintervallet er lille, er der relativt få værdier af parameteren, der giver en fornuftig beskrivelse af data  $\mathbf{x}$ , og  $\mathbf{x}$  indeholder derfor megen information om værdien af  $\boldsymbol{\omega}$ .

#### Eksempel 4.1 (Fortsat)

For  $u$ -testet for hypotesen  $H_0 : \mu = \mu_0$  er acceptområdet ved et test på niveau  $\alpha$

$$-u_{1-\alpha/2} \leq u = \frac{\bar{x} - \mu_0}{\sqrt{\sigma_0^2/n}} \leq u_{1-\alpha/2} \quad (5.26)$$

og dermed er værdien af styrkefunktionen  $pow(\mu)$  for  $u$ -testet på niveau  $\alpha$  beregnet i punktet

$\mu$  lig med

$$\begin{aligned} \text{pow}(\mu) &= 1 - P_{\mu}(-u_{1-\alpha/2} \leq \frac{\bar{X} - \mu_0}{\sqrt{\sigma_0^2/n}} \leq u_{1-\alpha/2}) \\ &= 1 - P_{\mu}(-u_{1-\alpha/2} \sqrt{\sigma_0^2/n} + \mu_0 \leq \bar{X} \leq u_{1-\alpha/2} \sqrt{\sigma_0^2/n} + \mu_0). \end{aligned}$$

Under sandsynlighedsmålet  $P_{\mu}$  er  $\bar{X} \sim N(\mu, \sigma_0^2/n)$  så

$$\begin{aligned} \text{pow}(\mu) &= 1 - \left( \Phi\left(\frac{u_{1-\alpha/2} \sqrt{\sigma_0^2/n} + \mu_0 - \mu}{\sqrt{\sigma_0^2/n}}\right) - \Phi\left(\frac{-u_{1-\alpha/2} \sqrt{\sigma_0^2/n} + \mu_0 - \mu}{\sqrt{\sigma_0^2/n}}\right) \right) \\ &= 1 - \Phi\left(u_{1-\alpha/2} + \frac{\mu_0 - \mu}{\sqrt{\sigma_0^2/n}}\right) + \Phi\left(-u_{1-\alpha/2} + \frac{\mu_0 - \mu}{\sqrt{\sigma_0^2/n}}\right), \end{aligned}$$

se Figur 5.4.

Af (5.26) fås, at  $(1 - \alpha)$  konfidensintervallet for  $\mu$  er

$$\bar{x} - u_{1-\alpha/2} \sqrt{\sigma_0^2/n} \leq \mu \leq \bar{x} + u_{1-\alpha/2} \sqrt{\sigma_0^2/n}.$$

□

## 5.7 Approksimativ likelihood teori

Som bemærket i Afsnit 5.5 er det undertiden vanskeligt eller umuligt, at beregne den eksakte værdi af testsandsynligheden  $\varepsilon(\mathbf{x})$  for likelihood ratio testet i (5.7) eller (5.22). I dette afsnit diskuterer vi, hvorledes testsandsynligheden  $\varepsilon(\mathbf{x})$  kan approksimeres. Desuden omtales approksimationer af fordelingen af maksimum likelihood estimatoren  $\hat{\boldsymbol{\omega}} = \hat{\boldsymbol{\omega}}(\mathbf{X})$ . Bemærk, at testsandsynligheden i (5.18) præcis er værdien af fordelingsfunktionen for likelihood ratio testoren  $Q(\mathbf{X})$  beregnet i den observerede værdi  $Q(\mathbf{x})$ , det vil sige

$$\varepsilon(\mathbf{x}) = F_{Q(\mathbf{X})}(Q(\mathbf{x})). \quad (5.27)$$

Spørgsmålet om at approksimere testsandsynligheden i (5.18) eller (5.27) er derfor ækvivalent med at finde approksimationer til fordelingen - under  $H_0$  - af likelihood ratio testoren. Lignende bemærkninger gælder i det tilfælde, hvor  $H_0$  er en sammensat hypotese, det vil sige i det tilfælde, hvor testsandsynligheden beregnes ved hjælp af (5.22).

Vi indskrænker os her til en detaljeret omtale af resultaterne i det tilfælde hvor parameteren er endimensional, det vil sige  $k = 1$ .

Approksimationerne, der omtales i det følgende, er baseret på anden ordens Taylor udviklinger af log likelihood funktionen. Disse er gyldige, idet det er antaget, at parameterrummet  $\Omega$  er et område i  $R$  samt at log likelihood funktionen  $l$  er mindst to gange differentiabel med kontinuerte (partielle) afledede. Mere præcist har vi

$$l(\omega) - l(\hat{\omega}) \doteq \frac{dl}{d\omega}(\hat{\omega})(\omega - \hat{\omega}) + \frac{1}{2} \frac{d^2l}{d\omega^2}(\hat{\omega})(\omega - \hat{\omega})^2, \quad (5.28)$$

hvor  $\doteq$  antyder approksimationen, og hvor udtrykket på højre side er Taylor polynomiet af anden grad for  $l$  omkring maksimum likelihood estimatet  $\hat{\omega}$ . Lad  $j(\omega; \mathbf{x})$  betegne tallet

$$j(\omega; \mathbf{x}) = -\frac{d^2l}{d\omega^2}(\omega; \mathbf{x}). \quad (5.29)$$

Idet  $\hat{\omega}$  er en løsning til likelihoodligningen (5.11), det vil sige  $\frac{dl}{d\omega}(\hat{\omega}) = 0$ , fås af (5.28) at

$$l(\omega) - l(\hat{\omega}) \doteq -\frac{1}{2} j(\hat{\omega}; \mathbf{x})(\hat{\omega} - \omega)^2. \quad (5.30)$$

Funktionen  $\bar{l}(\cdot) = l(\cdot) - l(\hat{\omega})$  kaldes den *normerede log likelihood funktion* og tallet  $j(\omega; \mathbf{x})$  omtales som den *observerede information* svarende til data  $\mathbf{x}$ . Middelværdien af den tilsvarende stokastiske variable  $j(\omega; \mathbf{X})$ , det vil sige

$$i(\omega) = E_{\omega}\{j(\omega; \mathbf{X})\}, \quad (5.31)$$

kaldes den *forventede information* eller *Fishers informationen*.

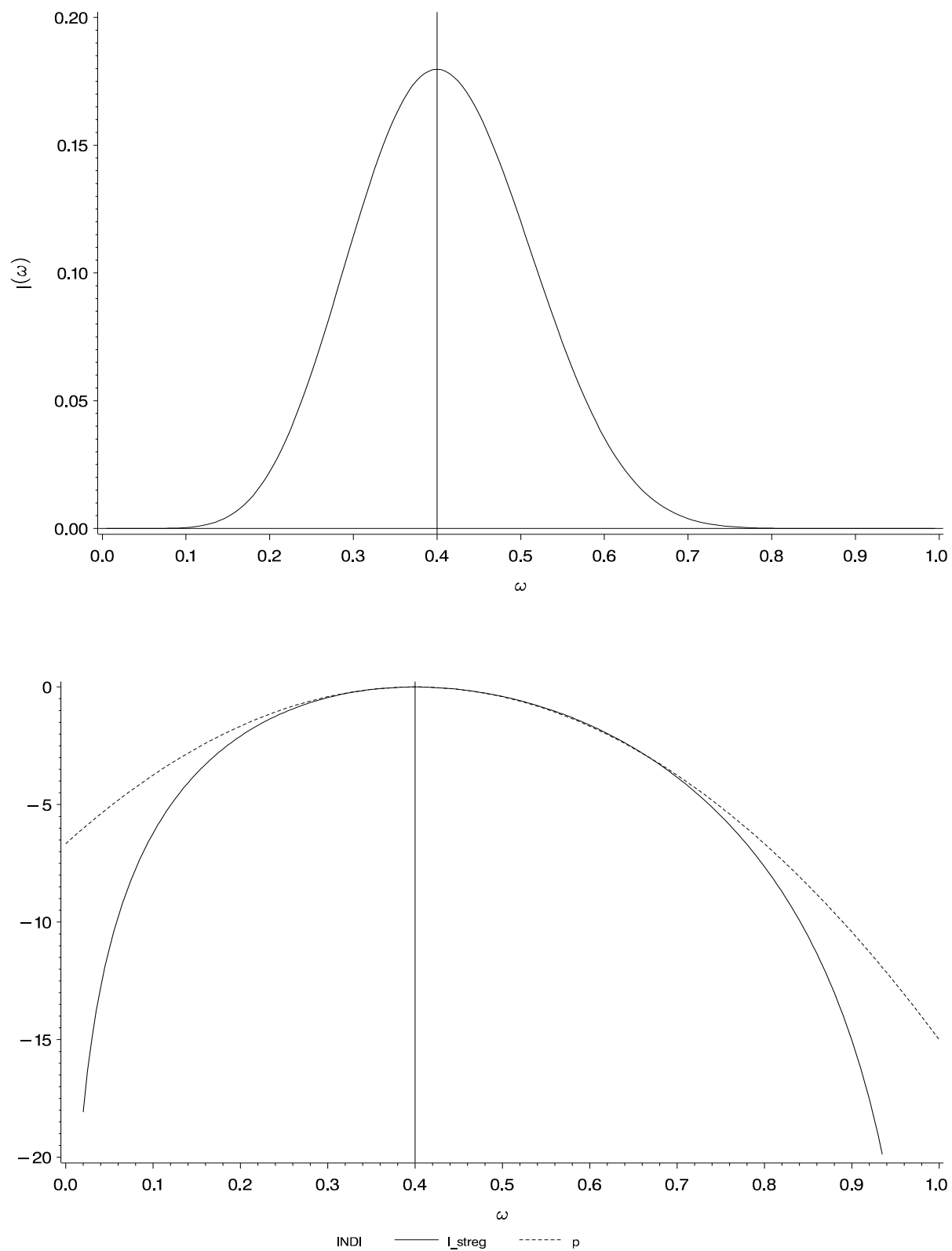
For at forklare, hvorfor ordet 'information' benyttes i denne sammenhæng, bemærker vi, at det fra (5.30) ses, at den normerede log likelihood funktion  $\bar{l}$  i en omegn af  $\hat{\omega}$  kan approksimeres ved parabeln

$$p(\omega) = -\frac{1}{2} j(\hat{\omega}; \mathbf{x})(\hat{\omega} - \omega)^2, \quad (5.32)$$

se Figur 5.5. I (5.32) er  $j(\hat{\omega}; \mathbf{x}) > 0$ , idet  $\hat{\omega}$  er et maksimumspunkt for  $l$ . Jo større  $j(\hat{\omega}; \mathbf{x})$  er, jo mere koncentrerer denne parabel sig om punktet  $\hat{\omega}$ , og kun for værdier af  $\omega$ , der ligger meget tæt på  $\hat{\omega}$ , er  $l(\omega)$  (eller  $L(\omega)$ ) af samme størrelsesorden som  $l(\hat{\omega})$  (eller  $L(\hat{\omega})$ ). Følgelig er  $j(\hat{\omega}; \mathbf{x})$  et mål for den information, som data  $\mathbf{x}$  giver om værdien af den ukendte parameter  $\omega$ .

Vi vender os nu mod en diskussion af, hvorledes fordelingen af henholdsvis maksimum likelihood estimatoren  $\hat{\omega} = \hat{\omega}(\mathbf{X})$  og likelihood ratio testoren  $Q = Q(\mathbf{X})$  kan approksimeres. Det kan vises, at fordelingen af  $\hat{\omega}$  - beregnet under fordelingen svarende til parameteren  $\omega$  - kan approksimeres ved normalfordeling med middelværdi  $\omega$  og varians  $i(\omega)^{-1}$ , som er den inverse til den forventede information  $i(\omega)$ . Dette resultat skrives på følgende måde:

$$\hat{\omega} \approx N(\omega, i(\omega)^{-1}). \quad (5.33)$$



**Figur 5.5** Øverst likelihood funktionen svarende til observationen  $x = 8$  i binomialmodellen med sandsynlighedsparameter  $\omega$  og antalsparameter  $n = 20$ . Nederst den normerede log likelihood funktion  $\bar{l}(\cdot) = l(\cdot) - l(\hat{\omega})$  og den approksimerende parabel  $p(\cdot)$ .

Approksimationen kan vises at være speciel god i det tilfælde, hvor data  $\mathbf{x}$  er én observationsrække  $x_1, \dots, x_n$  fra en fordeling og hvor  $n$  er stor.

Af resultaterne i Afsnit 3.2.1 og (5.33) fås at

$$\frac{\hat{\omega} - \omega}{\sqrt{i(\omega)^{-1}}} \approx N(0, 1)$$

og dermed følgende approksimation:

$$i(\omega)(\hat{\omega} - \omega)^2 \approx \chi^2(1). \quad (5.34)$$

Yderligere kan man undertiden i dette udtryk erstatte den forventede informationsmatriks  $i(\omega)$  med den forventede eller den observerede informationsmatriks beregnet i  $\hat{\omega}$ , det vil sige med  $i(\hat{\omega})$  eller  $j(\hat{\omega}) = j(\hat{\omega}; \mathbf{x})$ . Benyttes den sidstnævnte, opnås approksimationen

$$j(\hat{\omega})(\hat{\omega} - \omega)^2 \approx \chi^2(1). \quad (5.35)$$

Igen er denne approksimation god, hvis  $\mathbf{x}$  er én observationsrække  $x_1, \dots, x_n$  fra en fordeling og  $n$  er stor.

I stedet for at approksimere fordelingen for likelihood ratio testoren  $Q(\mathbf{X})$ , betragter man sædvanligvis approksimationer for fordelingen af størrelsen  $-2 \ln Q(\mathbf{X})$ . Man har følgende approksimative resultat for fordelingen af  $-2 \ln Q(\mathbf{X})$  i det tilfælde, hvor hypotesen  $H_0$  er en *simple* hypotese, der siger, at værdien af parameteren er  $\omega$

$$-2 \ln Q(\mathbf{X}) \approx \chi^2(1). \quad (5.36)$$

Approksimationen er en konsekvens af formlerne (5.30) og (5.35), idet man ved hjælp af disse formler finder, at

$$\begin{aligned} -2 \ln Q(\mathbf{X}) &= -2 \ln \frac{L(\omega)}{L(\hat{\omega})} \\ &= -2(l(\omega) - l(\hat{\omega})) \\ &\doteq j(\hat{\omega})(\hat{\omega} - \omega)^2 \\ &\approx \chi^2(1). \end{aligned} \quad (5.37)$$

Små værdier af likelihood ratio testoren  $Q$  er kritiske for  $H_0$ , hvilket er ækvivalent med at store værdier af  $-2 \ln Q(\mathbf{X})$  er kritiske. Af formel (5.36) får vi derfor følgende vigtige *approksimation for testsandsynligheden for likelihood ratio testet for den simple hypotese  $\omega$*

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2(1)}(-2 \ln Q(\mathbf{x})), \quad (5.38)$$

idet vi ved hjælp af (5.18) finder, at

$$\begin{aligned}
 \varepsilon(\mathbf{x}) &= P_{\omega}(Q(\mathbf{X}) \leq Q(\mathbf{x})) \\
 &= P_{\omega}(-2 \ln Q(\mathbf{X}) \geq -2 \ln Q(\mathbf{x})) \\
 &= 1 - P_{\omega}(-2 \ln Q(\mathbf{X}) < -2 \ln Q(\mathbf{x})) \\
 &\doteq 1 - F_{\chi^2(1)}(-2 \ln Q(\mathbf{x})).
 \end{aligned}$$

Her har vi ved  $\doteq$  brugt formel (5.36) samt den kendsgerning, at fordelingsfunktionen for  $\chi^2(1)$ -fordelingen er kontinuert.

#### Eksempel 4.1 (Fortsat)

Af formel (5.23) ses, at

$$-2 \ln Q(\mathbf{x}) = -2(l(\omega) - l(\bar{x})) = u^2(\mathbf{x}) \sim \chi^2(1),$$

idet  $u(\mathbf{x}) \sim N(0, 1)$ . I dette tilfælde gælder resultatet i (5.37) altså eksakt og ikke blot approksimativt.  $\square$

I det generelle tilfælde, hvor parameteren  $\omega$  er  $k$ -dimensional gælder der for likelihood ratio testoren af en *sammensat hypotese*  $H_0 : \omega \in \Omega_0$ , hvor  $\Omega_0 \subseteq \Omega$ , approksimationer analoge til (5.36) og (5.38). For at formulere disse resultater behøver vi følgende notation. En hypotese  $H_0 : \omega \in \Omega_0$  siges at have  $d$  frie parametre  $\theta_1, \dots, \theta_d$ , hvis der eksisterer et område  $\Theta \subset R^d$  og en en-entydig afbildning af  $\Theta$  på  $\Omega_0$ , det vil sige

$$\begin{aligned}
 \Theta \subseteq R^d &\quad \rightarrow \quad \Omega \subseteq R^k \\
 \boldsymbol{\theta} = (\theta_1, \dots, \theta_d) &\quad \rightarrow \quad \boldsymbol{\omega}(\boldsymbol{\theta}) = (\omega_1(\boldsymbol{\theta}), \dots, \omega_k(\boldsymbol{\theta})).
 \end{aligned} \tag{5.39}$$

Bemærk, at idet vi har antaget, at parameterrummet  $\Omega$  er et område, kan grundmodellen betragtes som en hypotese med de  $k$  frie parametre  $\omega_1, \dots, \omega_k$ . Bemærk endvidere, at for en simpel hypotese er  $d = 0$ .

Under visse regularitetsbetingelser, som stort set altid er opfyldt i praksis, har vi følgende *approksimationer for likelihood ratio testoren af en sammensat hypotese med  $d$  frie parametre*

$$-2 \ln Q(\mathbf{X}) \approx \chi^2(k - d), \tag{5.40}$$

og

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2(k-d)}(-2 \ln Q(\mathbf{x})). \tag{5.41}$$

Det ses af (5.40), at antallet af frihedsgrader i den approksimerende  $\chi^2$ -fordeling er lig med  $k - d$ , hvor  $k$  er antallet af frie parametre i grundmodellen (svarende til  $\Omega$ ) og  $d$  er antallet af frie parametre i hypotesen (svarende til  $\Omega_0$ ).

## 5.8 Afsluttende bemærkninger

Som nævnt i indledningen opfatter vi hovedbestanddelene i en statistisk analyse som

- i) modelopstilling
- ii) modelkontrol
- iii) statistisk inferens.

Som regel gennemløber analysen en eller flere cykliske faser, idet man ved ii) eller iii) opdager utilfredsstillende træk ved modellen og derfor går tilbage til i) for at revidere den.

Som beskrevet i indledningen til kapitlet betragter vi næsten udelukkende statistiske modeller, hvor fordelingerne er en parametriseret familie af fordelinger, og den statistiske inferens er baseret på likelihood funktionen. Vi beskæftiger os med ikke-parametrisk statistik i Kapitel 8. Vi møder ofte den opfattelse, at ikke-parametrisk statistik er fri for forudsætninger, og derfor sikker at bruge. Det er en alvorlig misforståelse. Ofte er et ikke-parametriske test udledt under strenge forudsætninger om uafhængighed, identiske fordelinger, og undertiden endda symmetriske fordelinger. Disse forudsætninger er således fælles for den parametriske statistik, som vi præsenterer her, og den ikke-parametriske statistik, og der er kun et lille skridt til at formulere en parametriske statistisk model. Det ekstra arbejde med at finde en gyldig parametriske statistisk model lønner sig som regel i den sidste ende, idet det giver anledning til formulere mere detaljerede matematiske modeller, hvilket som regel er motivationen bag det meste eksperimentelle arbejde i naturvidenskaben, herunder også i idræt.

Vi har tidligere i dette afsnit bemærket at det ofte er en del af en statistisk analyse at undersøge, om en enklere statistisk model end den, der som udgangspunkt blev opstillet, giver en tilfredsstillende beskrivelse af data, og vi har skitseret hvordan vi bruger statistiske tests til at vurdere det. Det er her meget vigtigt at være opmærksom på, at man aldrig med statistiske tests kan bevise noget. Man kan kun modbevise i den forstand, at man kan overbevise sig om, at data strider mod en simple model. Hvis man har begrænsede data kan man risikere ikke at kunne afvise reduktionen til en simpel model, som måske i virkeligheden er forkert. Det er derfor et moralsk krav helst på forhånd at sikre sig at man har en chance for at opdage at en hypotese er falsk.

Her kommer *forsøgsplanlægning* ind i billedet. Denne disciplin beskæftiger sig med, hvorledes man, under hensyntagen til ressourcer, kan tilrettelægge eksperimenter, herunder indsamling af data, for at opnå mest mulig information om den relevante faglige sammenhæng. På grund af kursets omfang kan vi ikke beskæftige os indgående med dette aspekt af en statistisk analyse.

## Opgaver til Kapitel 5

**Opgave 5.1** Antag at  $X$  er binomialfordelt med antalsparameter  $n$  og sandsynlighedsparameter  $p$ ,  $X \sim b(n, p)$ , det vil sige at

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

samt at  $x$  er en observation af  $X$ .

a) Vis, at log likelihood funktionen for  $p$  er

$$l(p) = \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p).$$

b) Vis, at likelihood ligningen for  $p$  er

$$\frac{x}{p} - \frac{n-x}{1-p} = 0$$

samt at

$$\hat{p} = \hat{p}(x) = \frac{x}{n}.$$

c) Vis ved hjælp af resultaterne i Afsnit 3.2.1, at

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

**Opgave 5.2** Antag, at  $x_1, \dots, x_n$  er en observationsrække fra Poissonfordelingen med parameter  $\lambda$ , som har sandsynlighedsfunktion

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, \dots$$

a) Vis, at log likelihood funktionen for  $\lambda$  er

$$l(\lambda) = -n\lambda + x \cdot \ln \lambda - \sum_{i=1}^n \ln x_i!,$$

hvor  $x = \sum_{i=1}^n x_i$ .

b) Vis, at likelihood ligningen for  $\lambda$  er

$$-n + \frac{x}{\lambda} = 0$$

samt at

$$\hat{\lambda} = \hat{\lambda}(\mathbf{x}) = \bar{x} = \frac{x}{n}.$$

c) Vis ved hjælp af resultaterne i Afsnit 3.2.3 - idet  $x. \sim\sim po(n\lambda)$  - at

$$E\hat{\lambda}(\mathbf{X}) = \lambda \quad \text{og} \quad \text{Var}\hat{\lambda}(\mathbf{X}) = \frac{\lambda}{n}.$$

## 6 Multinomialfordelte data

Multinomialfordelingen kan introduceres på følgende måde. Betragt et eksperiment for hvilket de følgende fire betingelser er opfyldt:

- Eksperimentet består af  $n$  identiske delforsøg.
- Hvert delforsøg kan resultere i præcis én af  $k$  hændelser,  $B_1, \dots, B_j, \dots, B_k$ .
- Sandsynligheden for de  $k$  hændelser er den samme i alle de  $n$  delforsøg,  $P(B_1) = \pi_1, \dots, P(B_j) = \pi_j, \dots, P(B_k) = \pi_k$ .
- Udfaldene af de  $n$  delforsøg er uafhængige.

Hvis  $\mathbf{X}$  betegner den  $k$ -dimensionale diskrete stokastiske vektor  $(X_1, \dots, X_j, \dots, X_k)$ , hvor den  $j$ 'te komponent  $X_j$  angiver, hvor mange gange hændelsen  $B_j$  indtræffer i de  $n$  delforsøg, er  $\mathbf{X}$  multinomialfordelt med antalsparameter  $n$  og sandsynlighedsvektor  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_j, \dots, \pi_k)$ , kort  $\mathbf{X} \sim m(n, \boldsymbol{\pi})$ . Det kan vises, at sandsynlighedsfunktionen for  $\mathbf{X}$  er

$$P(\mathbf{X} = \mathbf{x}) = \frac{n!}{x_1! \cdots x_j! \cdots x_k!} \pi_1^{x_1} \cdots \pi_j^{x_j} \cdots \pi_k^{x_k}. \quad (6.1)$$

Her er  $\mathbf{x} = (x_1, \dots, x_j, \dots, x_k)$  en vektor, således at

$$x_j \in \{0, 1, \dots, n\}, \quad j = 1, \dots, k \quad \text{og} \quad \sum_{j=1}^k x_j = n.$$

### Eksempel 6.1

Følgende tre "eksperimenter" kan beskrives ved hjælp af multinomialfordelingen.

i) Antag, at vi kaster en "ærlig" mønt 100 gange. Lad  $x_1$  og  $x_2$  være antallet af gange mønten viser henholdsvis "plat" og "krone". Betingelserne a) - d) ovenfor kan da antages at være opfyldte, idet eksperimentet består af  $n = 100$  identiske delforsøg, nemlig et kast med mønten. Hvert af de 100 delforsøg har  $k = 2$  udfald, nemlig "plat" eller "krone", og præcis

ét af dem indtræffer og det må antages at mønten har samme sandsynlighed,  $1/2$ , for at vise henholdsvis ”plat” og ”krone” i de 100 delforsøg. Desuden må det kunne antages, at udfaldene i de 100 kast med mønten er uafhængige. Sammenfattende kan  $(x_1, x_2)$  opfattes som udfald af en diskrete stokastisk vektor  $\mathbf{X} = (X_1, X_2)$  som er multinomialfordelt med antalsparameter  $n = 100$  og sandsynlighedsvektor  $\boldsymbol{\pi} = (1/2, 1/2)$ .

ii) Antag, at vi kaster en ”ærlig” terning  $n$  gange. Hvis  $x_i$  betegner antallet af gange terningen viser  $i$  øjne,  $i = 1, \dots, 6$ , kan vektoren  $(x_1, \dots, x_6)$  opfattes som et udfald af den stokastiske vektor  $\mathbf{X} = (X_1, \dots, X_6) \sim m(n, \boldsymbol{\pi})$ , hvor  $\boldsymbol{\pi} = (1/6, \dots, 1/6)$ .

iii) Antag, at vi nummererer de 52 spillekort og at vi  $n$  gange tilfældigt udtrækker ét af de 52 kort, det vil sige trækker et kort og lægger det tilbage inden vi trækker det næste kort. Lad  $x_i$ ,  $i = 1, \dots, 52$ , være antallet af gange kort nummer  $i$  trækkes. Vektoren  $\mathbf{x} = \{x_i\}_{i=1, \dots, 52}$  kan da opfattes som et udfald af vektoren  $\mathbf{X} = \{X_i\}_{i=1, \dots, 52}$ , som er multinomialfordelt med  $k = 52$ , antalsparameter  $n$  og sandsynlighedsvektor  $\boldsymbol{\pi} = \{\pi_i\}_{i=1, \dots, 52}$ , hvor  $\pi_i = 1/52$ .  $\square$

Indenfor langt de fleste fag er der utallige eksempler på data fra eksperimenter, der opfylder de ovenstående fire betingelser, og for hvilke den statistiske analyse derfor - uden videre kontrol - kan foretages ved hjælp af en model baseret på multinomialfordelingen. I Afsnit 6.1 introduceres tre datasæt fra idræt, der benyttes som illustrationer senere i kapitlet. Afsnit 6.2 vedrører statistisk inferens baseret på én multinomialfordeling og illustrerer blandt andet test af simple hypoteser og test af hypotesen om uafhængighed af inddelingskriterier. I Afsnit 6.3 illustreres teorien for en model baseret på flere uafhængige multinomialfordelinger med testet for hypotesen om identitet af sandsynlighedsvektorerne i uafhængige multinomialfordelinger. Alle de nævnte test er baseret på den approksimative likelihood teori, som er omtalt i Afsnit 5.7. Afsnit 6.4 viser et eksempel på anvendelsen af Fishers eksakte test i en situation, hvor forudsætningerne for at bruge den approksimative teori ikke er opfyldt.

I Kapitel 4 så vi på forskellige grafiske metoder til kontrol af fordelingsantagelserne i en statistisk model. Undertiden kan denne kontrol suppleres med numeriske test, der som regel omtales som test for *goodness of fit*. Test af denne type er emnet for Afsnit 6.5.

Alle testene i dette kapitel kan foretages ved hjælp af en programpakke. I et annekst til dette kapitel gives eksempler på beregninger foretaget ved hjælp af *Excel*.

## 6.1 Eksempler

I dette afsnit introduceres tre eksempler, som vil blive brugt til at illustrere statistisk inferens i modeller baseret på multinomialfordelingen.

**Eksempel 6.2**

Af Tabel 1.3 fremgår det, at antallet af sejre, uafgjorte og nederlag i AB's 33 kampe i Faxe Kondi Ligaen 1999-2000 var:

<i>sejr</i>	<i>uafgjort</i>	<i>nederlag</i>	<i>i alt</i>
14	10	9	33

Lad  $\mathbf{x} = (x_1, x_2, x_3)$  betegne disse antal, det vil sige  $\mathbf{x} = (14, 10, 9)$ . Som model for resultaterne i de 33 kampe vil vi antage, at  $\mathbf{x}$  er udfald af en diskret stokastisk vektor  $\mathbf{X}$  som er multinomialfordelt med antalsparameter  $n = 33$  og sandsynlighedsvektor  $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ , hvor for eksempel  $\pi_2$  er sandsynligheden for uafgjort. Vi antager altså implicit a) at sandsynligheden for henholdsvis sejr, uafgjort og nederlag er den samme i alle kampene, det vil sige, at disse sandsynligheder afhænger ikke af modstanderen og heller ikke af om en kamp spilles på hjemmebane eller på udebane, b) resultatet i en kamp influerer ikke på resultaterne i de andre kampe.

I modellen  $\mathbf{X} \sim m(33, \boldsymbol{\pi})$ , svarer hypotesen  $H_0 : \boldsymbol{\pi} = (1/3, 1/3, 1/3)$  til at for AB er sandsynligheden den samme for at vinde, spille uafgjort eller tabe i en tilfældig kamp.  $\square$

**Eksempel 6.3**

(Andersen 1998) I forskningsprojektet *Idræt og Ungdom* er 3869 unge klassificeret efter idrætsaktivitet (timer per uge) og status med hensyn til rygning. Resultatet ses i tabellen nedenfor.

		<i>rygerstatus</i>	
		<i>ryger</i>	<i>ikke-ryger</i>
<i>idrætsaktivitet</i>	<i>timer per uge</i>		
	0.0-0.5	181	603
	0.5-2.0	158	591
	2.0-4.0	162	713
	4.0-7.0	150	697
7.0-	83	531	

Vi kunne naturligvis opskrive de observerede antal som en vektor af længde 10, men det viser sig bekvemt at vælge en notation i overensstemmelse med den måde, observationerne er angivet i den ovenstående tabel. Vi lader derfor  $x_{ij}$  angive antallet af unge i den  $i$ 'te kategori af den variable *idrætsaktivitet* og i den  $j$ 'te kategori af den variable *rygerstatus*. Med denne notation forekommer det da rimeligt at antage, at matricen  $\{x_{ij}\}$  er en realisation af en stokastisk matris  $\{X_{ij}\}$ , som er multinomialfordelt med antalsparameter 3869 og sandsynlighedsmatris  $\{\pi_{ij}\}$ .

Vi vil undersøge spørgsmålet, om de unges rygevaner er uafhængig af idrætsaktiviteten. Lad  $\rho_i$  betegne sandsynligheden for at en ung tilhører den  $i$ 'te idrætsaktivitetskategori og lad tilsva-

rende  $\sigma_j$  betegne sandsynligheden for at en ung tilhører den  $j$ 'te rygerkategori. Idet  $\pi_{ij}$  betegner sandsynligheden for at ung tilhører den  $i$ 'te idrætaktivitetskategori og den  $j$ 'te rygerkategori, kan spørgsmålet formuleres som en hypotese i multinomialmodellen på følgende måde:

$$H_{01} : \pi_{ij} = \rho_i \sigma_j, \quad i = 1, \dots, 5, \quad j = 1, 2. \quad (6.2)$$

□

### Eksempel 6.4

I tabellen nedenfor ses antallet af hjemmesejre, uafgjorte og udesejre i de 198 kampe Faxe Kondi Ligaen 1999-2000 optalt i henholdsvis første, anden og tredje tredjedel af turneringen.

kamp nr.	hjemmesejr	uafgjort	udesejr	$i$ alt
1-66	32	17	17	66
67-132	25	20	21	66
133-198	33	15	18	66

Lad  $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$  betegne de observerede antal af henholdsvis hjemmesejre, uafgjorte og udesejre i den  $i$ 'te tredjedel af turneringen. I modellen

$$M_0 : \mathbf{X}_i \sim m(66, \boldsymbol{\pi}_i), \quad i = 1, 2, 3$$

$\mathbf{X}_1, \mathbf{X}_2$  og  $\mathbf{X}_3$  er stokastisk uafhængige

svarer hypotesen

$$H_{01} : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = \boldsymbol{\pi}_3 (= \boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3))$$

til at sandsynlighederne for henholdsvis hjemmesejr, uafgjort og udesejr er den samme i de tre dele af turneringen. □

## 6.2 Inferens i én multinomialfordeling.

Lad os indledningsvis repetere de egenskaber ved multinomialfordelingen - omtalt i Afsnit 3.2.2 - der benyttes i det følgende. En diskret stokastisk vektor  $\mathbf{X} = (X_1, \dots, X_j, \dots, X_k)$  er multinomialfordelt med antalsparameter  $n$  og sandsynlighedsvektor  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_j, \dots, \pi_k)$ ,  $\mathbf{X} \sim m(n, \boldsymbol{\pi})$ , hvis sandsynlighedsfunktionen for  $\mathbf{X}$  er

$$P(\mathbf{X} = \mathbf{x}) = \frac{n!}{x_1! \cdots x_j! \cdots x_k!} \pi_1^{x_1} \cdots \pi_j^{x_j} \cdots \pi_k^{x_k}, \quad (6.3)$$

hvor  $\mathbf{x} = (x_1, \dots, x_j, \dots, x_k)$  er en vektor, således at

$$x_j \in \{0, 1, \dots, n\}, \quad j = 1, \dots, k \quad \text{og} \quad \sum_{j=1}^k x_j = n.$$

Undertiden vil vi også benytte notationen  $(X_1, \dots, X_j, \dots, X_k) \sim m(n, (\pi_1, \dots, \pi_j, \dots, \pi_k))$  til at angive at  $\mathbf{X} \sim m(n, \boldsymbol{\pi})$ .

Sandsynlighedsvektoren  $\boldsymbol{\pi}$  tilhører mængden

$$\boldsymbol{\Pi} = \{\boldsymbol{\pi} \in R^k : \pi_j > 0, j = 1, \dots, k, \sum_{j=1}^k \pi_j = 1\}.$$

Bemærk, at selvom  $\boldsymbol{\pi}$  er en  $k$ -dimensional vektor, varierer dens komponenter ikke frit. Kender vi for eksempel  $\pi_1, \dots, \pi_{k-1}$  kan  $\pi_k$  beregnes som  $1 - \pi_1 - \dots - \pi_{k-1}$ . Med andre ord - i terminologien fra Afsnit 5.7 - har multinomialmodellen  $k - 1$  frie parametre.

Fra Afsnit 3.2.2 ved vi desuden, at middelværdivektoren for  $\mathbf{X}$  er

$$E\mathbf{X} = n\boldsymbol{\pi} = (n\pi_1, \dots, n\pi_j, \dots, n\pi_k) \quad (6.4)$$

samt at kovariansmatricen for  $\mathbf{X}$  har elementerne

$$\begin{aligned} (\text{Cov } \mathbf{X})_{jj} &= \text{Var}X_j = n\pi_j(1 - \pi_j), \quad j = 1, \dots, k \\ (\text{Cov } \mathbf{X})_{ij} &= \text{Cov}(X_i, X_j) = -n\pi_i\pi_j, \quad i \neq j, \quad i, j = 1, \dots, k. \end{aligned} \quad (6.5)$$

Endelig ved vi, at den marginale fordeling for den  $j$ 'te komponent  $X_j$  af  $\mathbf{X}$  er binomialfordelingen med antalsparameter  $n$  og sandsynlighedsparameter  $\pi_j$ ,  $X_j \sim b(n, \pi_j)$ ,  $j = 1, \dots, k$ , det vil sige

$$P(X_j = x_j) = \binom{n}{x_j} \pi_j^{x_j} (1 - \pi_j)^{n-x_j}, \quad x_j = 0, 1, \dots, n. \quad (6.6)$$

(De mest basale egenskaber ved binomialfordelingen er omtalt i Afsnit 3.2.1.)

I det følgende får vi flere gange brug for et matematisk resultat, der gives i nedenstående sætning, som vi ikke vil bevise.

**Sætning 6.1** Antag, at  $x_j > 0$  for  $j = 1, \dots, k$  samt at  $x_1 + \dots + x_k = n$ . Da antager funktionen

$$\begin{aligned} g : \boldsymbol{\Pi} &\rightarrow R \\ \boldsymbol{\pi} &\rightarrow \pi_1^{x_1} \dots \pi_j^{x_j} \dots \pi_k^{x_k} \end{aligned}$$

sin maksimale værdi i punktet

$$\hat{\boldsymbol{\pi}} = \left( \frac{x_1}{n}, \dots, \frac{x_j}{n}, \dots, \frac{x_k}{n} \right).$$

◆

Efter disse indledende bemærkninger er vi nu klar til at betragte den statistiske inferens i multinomialmodellen

$$M_0 : \mathbf{X} = (X_1, \dots, X_j, \dots, X_k) \sim m(n, \boldsymbol{\pi}).$$

### Estimation

Af formel (6.3) ses, at likelihood funktionen for  $\boldsymbol{\pi}$  er

$$L(\boldsymbol{\pi}) = \frac{n!}{x_1! \cdots x_j! \cdots x_k!} \pi_1^{x_1} \cdots \pi_j^{x_j} \cdots \pi_k^{x_k} \quad (6.7)$$

og det følger af Sætning 6.1, at maksimum likelihood estimatet for  $\boldsymbol{\pi}$  er

$$\boldsymbol{\pi} \leftarrow \hat{\boldsymbol{\pi}}(\mathbf{x}) = \left( \frac{x_1}{n}, \dots, \frac{x_j}{n}, \dots, \frac{x_k}{n} \right); \quad (6.8)$$

med andre ord er maksimum likelihood estimatet  $\hat{\pi}_j$  af  $\pi_j$  den *relative hyppighed*, hvormed hændelsen  $B_j$  indtræffer i de  $n$  delforsøg. Fordelingen af  $\hat{\boldsymbol{\pi}}$  angives som oftest på følgende måde:

$$n\hat{\boldsymbol{\pi}} = \mathbf{X} \sim m(n, \boldsymbol{\pi}).$$

### Hypoteser

Hypoteser i multinomialmodellen  $M_0$  testes ved hjælp af approksimative  $-2 \ln Q$ -test som beskrevet i Afsnit 5.7. Som det fremgår af Afsnit 5.7, bestemmes antallet af frihedsgrader i den  $\chi^2$ -fordeling, der approksimerer  $-2 \ln Q$ -testorens fordeling, som differensen mellem antallet af frie parametre i  $M_0$  og antallet af frie parametre i hypotesen, der testes. Det er derfor vigtigt præcist at kunne angive antallet af frie parameter i en hypotese. For multinomialmodellen gøres dette på følgende måde. Lad  $\boldsymbol{\pi}$  være en-entydig afbildning af et område  $\Theta$  i  $R^d$  på en delmængde  $\Pi_0$  af parametermængden  $\Pi$

$$\begin{aligned} \boldsymbol{\pi} : \Theta \subseteq R^d &\rightarrow \Pi_0 \subseteq \Pi \\ \boldsymbol{\theta} = (\theta_1, \dots, \theta_d) &\rightarrow \boldsymbol{\pi}(\boldsymbol{\theta}) = (\pi_1(\boldsymbol{\theta}), \dots, \pi_j(\boldsymbol{\theta}), \dots, \pi_k(\boldsymbol{\theta})). \end{aligned} \quad (6.9)$$

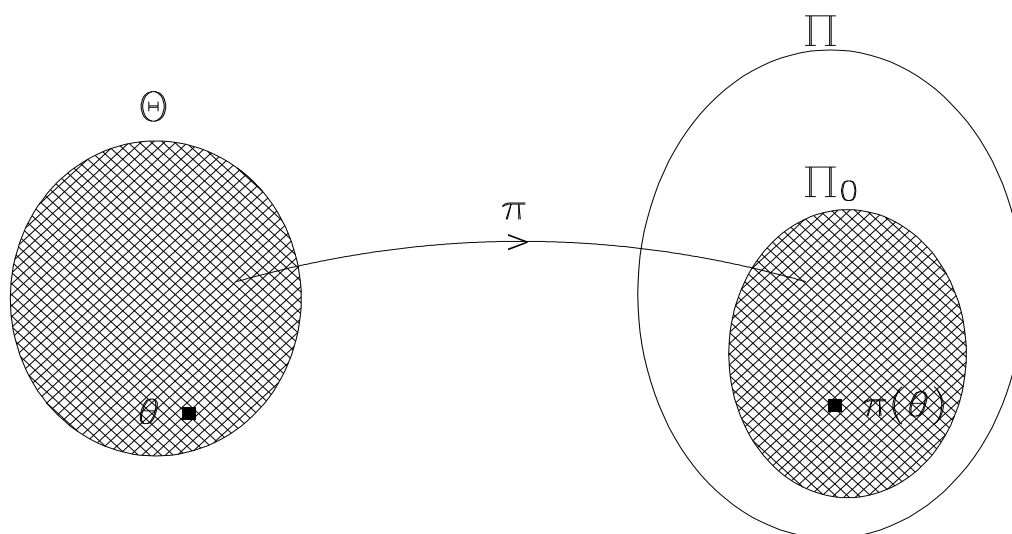
Hypotesen

$$H_{01} : \boldsymbol{\pi} \in \Pi_0 = \boldsymbol{\pi}(\Theta) (\subseteq \Pi) \quad (6.10)$$

siges da at have  $d$  frie parametre. Den generelle definition er illustreret i Figur 6.1. Da afbildningen  $\boldsymbol{\pi}$  er defineret på  $\Theta$ , er en-entydig og har værdimængde  $\Pi_0$ , betyder ovenstående blot, at der til ethvert element  $\boldsymbol{\theta}$  i  $\Theta$  findes et og kun et element  $\boldsymbol{\pi}(\boldsymbol{\theta})$  i  $\Pi_0$  og *vice versa*; med andre ord bruges mængden  $\Theta$  til at navngive elementer i  $\Pi_0$  med.

Hypotesen  $H_{01}$  reducerer modellen  $M_0$  til

$$M_1 : \mathbf{X} = (X_1, \dots, X_j, \dots, X_k) \sim m(n, \boldsymbol{\pi}), \quad \boldsymbol{\pi} \in \Pi_0.$$



**Figur 6.1** Illustration af definitionen af en hypotese med  $d$  frie parametre. Mængden  $\Theta$  antages at være et område i  $R^d$ . Mængden  $\Pi$  symboliserer parametermængden i grundmodellen, mens  $\Pi_0$  symboliserer parametermængden svarende til hypotesen  $H_0$ .

### Estimation under hypotese

Vi betragter nu maksimum likelihood estimation i  $M_1$ . Under  $M_1$  er sandsynlighedsvektoren af formen  $\boldsymbol{\pi}(\boldsymbol{\theta}) = (\pi_1(\boldsymbol{\theta}), \dots, \pi_j(\boldsymbol{\theta}), \dots, \pi_k(\boldsymbol{\theta}))$ , hvor  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_i, \dots, \theta_d)$ . Af (6.7) fås, at likelihood funktionen for  $\boldsymbol{\theta}$  er

$$L(\boldsymbol{\theta}) = \frac{n!}{x_1! \cdots x_j! \cdots x_k!} \pi_1(\boldsymbol{\theta})^{x_1} \cdots \pi_j(\boldsymbol{\theta})^{x_j} \cdots \pi_k(\boldsymbol{\theta})^{x_k}. \quad (6.11)$$

Log likelihood funktionen og likelihood ligningerne bliver derfor henholdsvis

$$l(\boldsymbol{\theta}) = \ln \left( \frac{n!}{x_1! \cdots x_j! \cdots x_k!} \right) + \sum_{j=1}^k x_j \ln(\pi_j(\boldsymbol{\theta}))$$

og

$$\frac{\partial l}{\partial \theta_i}(\boldsymbol{\theta}) = \sum_{j=1}^k x_j \frac{1}{\pi_j(\boldsymbol{\theta})} \frac{\partial \pi_j}{\partial \theta_i}(\boldsymbol{\theta}), \quad i = 1, \dots, d.$$

Hvorledes likelihood ligningerne løses afhænger naturligvis af hypotesen  $H_{01}$  og kan derfor ikke diskuteres generelt. Det er ofte muligt - som illustreret i det følgende - at maksimere likelihood funktion  $L(\boldsymbol{\theta})$  ved hjælp af Sætning 6.1, og i de tilfælde er likelihood ligningerne uden interesse.

Lad  $\hat{\boldsymbol{\theta}}$  betegne maksimum likelihood estimatet for  $\boldsymbol{\theta}$ . Middelværdivektoren for  $\mathbf{X}$  beregnet i fordelingen svarende til sandsynlighedsvektoren  $\boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ , som ifølge (6.4) er

$$\mathbf{e} = (e_1, \dots, e_j, \dots, e_k) = (n\pi_1(\hat{\boldsymbol{\theta}}), \dots, n\pi_j(\hat{\boldsymbol{\theta}}), \dots, n\pi_k(\hat{\boldsymbol{\theta}})), \quad (6.12)$$

omtales som vektoren af *forventede antal under  $H_{01}$* .

### Test af hypotese

Som bekendt er  $\hat{\boldsymbol{\theta}}$  den værdi af parameteren  $\boldsymbol{\theta}$ , som tilordner den største sandsynlighed til observationen  $\mathbf{x}$ , og de forventede antal  $\mathbf{e} = n\boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$  er maksimum likelihood estimatet - under  $H_{01}$  - for middelværdivektoren for  $\mathbf{X}$ . Om hypotesen  $H_{01}$  er sand eller ej, må derfor kunne afgøres ved at undersøge, om vektoren  $\mathbf{e}$  af forventede antal "ligner" observationen  $\mathbf{x}$  eller ej. Tilbage er blot spørgsmålet, om hvorledes sammenligningen af  $\mathbf{e}$  og  $\mathbf{x}$  skal foretages. Lad os se på hvilket svar likelihood metoden giver på dette spørgsmål.

Af (6.7), (6.8) og (6.11) fås, at likelihood ratio testoren for  $H_{01}$  er

$$\begin{aligned} Q(\mathbf{x}) &= \frac{L(\hat{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\pi}})} \\ &= \frac{\pi_1(\hat{\boldsymbol{\theta}})^{x_1} \dots \pi_j(\hat{\boldsymbol{\theta}})^{x_j} \dots \pi_k(\hat{\boldsymbol{\theta}})^{x_k}}{\left(\frac{x_1}{n}\right)^{x_1} \dots \left(\frac{x_j}{n}\right)^{x_j} \dots \left(\frac{x_k}{n}\right)^{x_k}} \\ &= \left(\frac{n\pi_1(\hat{\boldsymbol{\theta}})}{x_1}\right)^{x_1} \dots \left(\frac{n\pi_j(\hat{\boldsymbol{\theta}})}{x_j}\right)^{x_j} \dots \left(\frac{n\pi_k(\hat{\boldsymbol{\theta}})}{x_k}\right)^{x_k} \\ &= \left(\frac{e_1}{x_1}\right)^{x_1} \dots \left(\frac{e_j}{x_j}\right)^{x_j} \dots \left(\frac{e_k}{x_k}\right)^{x_k} \end{aligned}$$

og dermed bliver

$$-2 \ln Q(\mathbf{x}) = 2 \sum_{j=1}^k x_j \ln \left( \frac{x_j}{e_j} \right). \quad (6.13)$$

Hvis de forventede antal alle er større end eller lig med 5 kan approksimationen i (5.41) benyttes, det vil sige, at vi for testsandsynligheden  $\varepsilon(\mathbf{x})$  har følgende approksimation

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2(k-1-d)}(-2 \ln Q(\mathbf{x})), \quad (6.14)$$

idet modellerne  $M_0$  og  $M_1$  har henholdsvis  $k-1$  og  $d$  frie parametre. Bemærk, at hypotesen  $H_{01}$  kun kan testes, hvis  $d < k-1$ , da antallet af frihedsgrader i den approksimerende  $\chi^2$ -fordeling naturligvis skal være positivt.

Et par bemærkninger vedrørende beregning af  $-2 \ln Q$ -testoren ved hjælp af lommeregner. Den hyppigst forekommende fejl er, at **2-tallet** på højresiden i formel (6.13) glemmes. Desuden er det vigtigt at gentage, at  $Q(\mathbf{x})$  er et likelihood ratio test så  $0 < Q \leq 1$ , og derfor er  $-2 \ln Q >$

0. Hvis en beregning resulterer i en negativ værdi af  $-2\ln Q$ , er der altså kun én forklaring: **regnefejl!**

Sammenligningen af de observerede antal  $\mathbf{x}$  og de forventede antal  $\mathbf{e}$  foretages undertiden ved hjælp af  $X^2$ -testoren (læs: chi-i-anden testoren)

$$X^2(\mathbf{x}) = \sum_{j=1}^k \frac{(x_j - e_j)^2}{e_j}. \quad (6.15)$$

Hvis de forventede antal alle er større end eller lig med 5, kan testsandsynligheden  $\varepsilon^*(\mathbf{x})$  for  $X^2$ -testet for en hypotese  $H_{01}$  med  $d$  frie parametre approksimeres på følgende måde

$$\varepsilon^*(\mathbf{x}) \doteq 1 - F_{\chi^2(k-1-d)}(X^2(\mathbf{x})). \quad (6.16)$$

Af de to test for  $H_{01}$  foretrækker vi  $-2\ln Q$ -testet, idet  $X^2$ -testoren blot er en approksimation af  $-2\ln Q$ -testoren. I litteraturen, specielt den ældre, ser man dog ofte  $X^2$ -testoren anvendt, hvilket muligvis blandt andet skyldes, at ln-tasten ikke fandtes på den tids lommeregner og at det derfor var besværligt at beregne  $-2\ln Q$ .

### Konfidensintervaller

Vi vil ikke diskutere konfidensområder for sandsynlighedsvektoren  $\boldsymbol{\pi}$  i modellen  $M_0$ , men nøjes med at angive konfidensintervallet for den  $j$ 'te komponent  $\pi_j$  af  $\boldsymbol{\pi}$ . Konstruktionen af dette tager sit udgangspunkt i formel (6.6), ifølge hvilken den  $j$ 'te komponent  $X_j$  af  $\mathbf{X}$  er binomialfordelt med antalsparameter  $n$  og sandsynlighedsparameter  $\pi_j$ . Problemet er hermed reduceret til at finde konfidensintervallet for sandsynlighedsparameteren  $\pi$  i en binomialmodel

$$M_b : X \sim b(n, \pi).$$

I modellen  $M_b$  kan hypotesen  $H_0 : \pi = \pi_0$ , hvor  $\pi_0$  er kendt, testes ved hjælp af  $u$ -fordelingen. Testet er baseret på, at fordelingen for  $X$  kan approksimeres med en normalfordeling som har samme middelværdi og varians som  $X$ , det vil sige  $X \approx N(n\pi, n\pi(1-\pi))$ , hvilket medfører, at vi har følgende approksimation for fordelingsfunktionen  $F_X$  for  $X$ :

$$F_X(x) = P(X \leq x) \doteq \Phi\left(\frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}\right).$$

Testet af  $H_0 : \pi = \pi_0$  baseret på  $u$ -fordelingen kan vises at være ækvivalent med  $-2\ln Q$ -testet, og det giver anledning til følgende  $(1 - \alpha)$  konfidensinterval for  $\pi$  beregnet ud fra observationen  $x$ :

$$C_{1-\alpha}(x) = \{\pi_0 \mid H_0 : \pi = \pi_0 \text{ accepteres med niveau } \alpha \text{ test}\} = [\pi_-, \pi_+], \quad (6.17)$$

hvor

$$\pi_- = \frac{1}{n + u_{1-\alpha/2}^2} \left[ x + \frac{1}{2}u_{1-\alpha/2}^2 - u_{1-\alpha/2} \sqrt{\frac{x(n-x)}{n} + \frac{1}{4}u_{1-\alpha/2}^2} \right]$$

og

$$\pi_+ = \frac{1}{n + u_{1-\alpha/2}^2} \left[ x + \frac{1}{2}u_{1-\alpha/2}^2 + u_{1-\alpha/2} \sqrt{\frac{x(n-x)}{n} + \frac{1}{4}u_{1-\alpha/2}^2} \right].$$

I disse formler betegner  $u_{1-\alpha/2}$  ( $1 - \alpha/2$ )-fraktilen i  $u$ -fordelingen. Hvis  $\alpha = 0.05$  er fraktilen  $u_{0.975} = 1.960$ .

Der findes mange anvendelser af teorien for én multinomialfordeling som beskrevet i dette afsnit. Vi har her valgt at indskrænke os til at illustrere teorien ved at omtale test af en simpel hypotese og test for uafhængighed af inddelingskriterier. Dette gøres i de følgende to underafsnit til Afsnit 6.2 ved hjælp af Eksempel 6.2 og Eksempel 6.3.

### 6.2.1 Test af simpel hypotese

Vi betragter nu den situation hvor sandsynlighedsvektoren  $\boldsymbol{\pi}$  er fuldstændigt specificeret under hypotesen, det vil sige en såkaldt simpel hypotese.

#### Eksempel 6.2 (Fortsat)

I multinomialmodellen med  $k = 3$

$$M_0 : (X_1, X_2, X_3) \sim m(33, (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3))$$

er maksimum likelihood estimatet - med fire decimalers nøjagtighed - for sandsynlighedsvektoren  $\boldsymbol{\pi}$  givet ved

$$\hat{\boldsymbol{\pi}} = \left( \frac{14}{33}, \frac{10}{33}, \frac{9}{33} \right) = (0.4242, 0.3030, 0.2727).$$

Vi vil undersøge hypotesen om resultaterne sejr, uafgjort og nederlag forekommer lige hyppigt i  $AB$ 's kampe, det vil sige hypotesen  $H_{01} : \boldsymbol{\pi} = (1/3, 1/3, 1/3)$ . Det ses, at  $\hat{\boldsymbol{\pi}}$  ligger tæt på denne værdi. Hypotesen  $H_{01}$  er simpel - den har  $d = 0$  frie parametre - så de forventede antal under  $H_0$  kan beregnes uden videre. Vi finder

	sejr	uafgjort	nederlag	$i$ alt
observeret $\mathbf{x}$	14	10	9	33
forventet $\mathbf{e}$	11	11	11	33

Ved hjælp af formel (6.13) fås

$$-2 \ln Q(\mathbf{x}) = 1.2343,$$

og da de forventede antal alle er større end 5, bliver testsandsynligheden ifølge (6.14)

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(2)}(1.2343) = 0.5395,$$

og hypotesen  $H_{01}$  accepteres. Vi kan altså ikke afvise, at der er samme sandsynlighed for resultaterne sejr, uafgjort og nederlag i  $AB$ 's kampe.  $\square$

## 6.2.2 Uafhængighed af inddelingskriterier

Problemstillingen omtalt i Eksempel 6.3 er et specialtilfælde af følgende generelle situation. Antag, at  $n$  objekter klassificeres efter to inddelingskriterier, hvoraf det første har  $r$  kategorier og det andet  $s$  kategorier. Data  $\mathbf{x}$  kan da opfattes som en  $r \times s$  matrix  $\{x_{ij}\}$ , hvor  $x_{ij}$  betegner antallet af objekter i den  $(i, j)$ 'te klasse svarende til den  $i$ 'te kategori ved det første kriterium og den  $j$ 'te kategori ved det andet kriterium.

	1	...	$j$	...	$s$	$\Sigma$
1	$x_{11}$	...	$x_{1j}$	...	$x_{1s}$	$x_{1\cdot}$
.	.	...	.	...	.	.
.	.	...	.	...	.	.
$i$	$x_{i1}$	...	$x_{ij}$	...	$x_{is}$	$x_{i\cdot}$
.	.	...	.	...	.	.
.	.	...	.	...	.	.
$r$	$x_{r1}$	...	$x_{rj}$	...	$x_{rs}$	$x_{r\cdot}$
$\Sigma$	$x_{\cdot 1}$	...	$x_{\cdot j}$	...	$x_{\cdot s}$	$n$

I tabellen betegner  $x_{i\cdot}$  og  $x_{\cdot j}$  henholdsvis summen af observationerne i den  $i$ 'te række og den  $j$ 'te søjle, altså

$$x_{i\cdot} = \sum_{j=1}^s x_{ij} \quad \text{og} \quad x_{\cdot j} = \sum_{i=1}^r x_{ij}.$$

Multinomialmodellen for den  $rs$ -dimensionale diskrete stokastiske matrix  $\mathbf{X} = \{X_{ij}\}$ ,

$$M_0 : \mathbf{X} = \{X_{ij}\} \sim m(n, \{\pi_{ij}\}), \quad (6.18)$$

har  $rs - 1$  frie parametre.

Lad  $\rho_i$  betegne sandsynligheden for at et objekt tilhører den  $i$ 'te kategori ved det første kriterium,  $i = 1, \dots, r$ , og lad tilsvarende  $\sigma_j$  betegne sandsynligheden for den  $j$ 'te kategori ved det andet kriterium,  $j = 1, \dots, s$ . Hypotesen

$$H_{01} : \pi_{ij} = \rho_i \sigma_j, \quad i = 1, \dots, r, \quad j = 1, \dots, s, \quad (6.19)$$

omtales som hypotesen om *uafhængighed mellem de to inddelingskriterier*, og den har  $d = (r - 1) + (s - 1) = r + s - 2$  frie parametre, idet

$$\sum_{i=1}^r \rho_i = 1 \quad \text{og} \quad \sum_{j=1}^s \sigma_j = 1.$$

Under  $M_0$  er likelihood funktionen for  $\boldsymbol{\pi} = \{\pi_{ij}\}$

$$L(\boldsymbol{\pi}) = \frac{n!}{x_{11}! \cdots x_{rs}!} \prod_{i=1}^r \prod_{j=1}^s \pi_{ij}^{x_{ij}} \quad (6.20)$$

og maksimum likelihood estimatet for  $\boldsymbol{\pi}$  er givet ved

$$\hat{\pi}_{ij} = \frac{x_{ij}}{n}, \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Likelihood funktionen for  $\boldsymbol{\rho} = (\rho_1, \dots, \rho_i, \dots, \rho_r)$  og  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_j, \dots, \sigma_s)$  findes af (6.20) ved at indsætte  $\pi_{ij} = \rho_i \sigma_j$ , og vi får

$$\begin{aligned} L(\boldsymbol{\rho}, \boldsymbol{\sigma}) &= \frac{n!}{x_{11}! \cdots x_{rs}!} \prod_{i=1}^r \prod_{j=1}^s (\rho_i \sigma_j)^{x_{ij}} \\ &= \frac{n!}{x_{11}! \cdots x_{rs}!} \prod_{i=1}^r \rho_i^{x_{i.}} \prod_{j=1}^s \sigma_j^{x_{.j}}. \end{aligned}$$

Det ses, at  $L(\boldsymbol{\rho}, \boldsymbol{\sigma})$  indeholder en faktor, som kun afhænger af  $\boldsymbol{\rho}$ , samt en faktor, som kun afhænger af  $\boldsymbol{\sigma}$ . Da  $\boldsymbol{\rho}$  og  $\boldsymbol{\sigma}$  desuden varierer uafhængigt af hinanden, kan vi anvende Sætning 6.1 på hver af disse faktorer. Vi finder, at maksimum likelihood estimaterne for  $\boldsymbol{\rho}$  og  $\boldsymbol{\sigma}$  er bestemt ved

$$\hat{\rho}_i = \frac{x_{i.}}{n}, \quad i = 1, \dots, r \quad \text{og} \quad \hat{\sigma}_j = \frac{x_{.j}}{n}, \quad j = 1, \dots, s, \quad (6.21)$$

altså som de *relative hyppigheder* for henholdsvis den  $i$ 'te kategori ved det første inddelingskriterium og den  $j$ 'te ved det andet.

Matricen  $\mathbf{e} = \{e_{ij}\}$  af forventede antal under  $H_{01}$  har elementer

$$\begin{aligned} e_{ij} &= n \hat{\rho}_i \hat{\sigma}_j \\ &= \frac{x_{i.} x_{.j}}{n}. \end{aligned} \quad (6.22)$$

Det forventede antal i den  $(i, j)$ 'te klasse findes således som produktet af den  $i$ 'te rækkesum og den  $j$ 'te søjlesum divideret med totalsummen.

Ved hjælp af (6.22) finder vi nu følgende beregningsformel for  $-2 \ln Q$ -testoren for  $H_{01}$  :

$$\begin{aligned} -2 \ln Q(\mathbf{x}) &= 2 \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln \left( \frac{x_{ij}}{e_{ij}} \right) \\ &= 2 \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln \left( \frac{x_{ij}}{x_{i.} x_{.j} / n} \right) \\ &= 2 \sum_{i=1}^r \sum_{j=1}^s x_{ij} [\ln(x_{ij}) - \ln(x_{i.}) - \ln(x_{.j}) + \ln(n)] \\ &= 2 \left[ \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln(x_{ij}) - \sum_{i=1}^r x_{i.} \ln(x_{i.}) - \sum_{j=1}^s x_{.j} \ln(x_{.j}) + n \ln(n) \right]. \end{aligned} \quad (6.23)$$

Ved beregninger af  $-2\ln Q$  i hånden er det nyttigt at have et lille program på lommeregneren, som beregner  $\sum x \ln(x)$ , idet den kantede parentes fremkommer ud fra tabellen over observerede antal som denne størrelse beregnet for indmaden af tabellen minus størrelsen beregnet for rækkesummerne minus størrelsen beregnet for søjlesummerne plus størrelsen beregnet for totalsummen. Igen er det vigtigt at **huske 2-tallet** på højresiden i denne formel.

Som nævnt ovenfor er antallet af frie parametre i grundmodellen  $M_0$  lig med  $rs - 1$ , og da antallet af frie parametre i  $H_{01}$  er  $r + s - 2$ , bliver antallet af frihedsgrader i  $-2\ln Q$ -testet for  $H_{01}$  lig med  $f = (rs - 1) - (r + s - 2) = (r - 1)(s - 1)$ . Hvis de forventede antal under uafhængighedshypotesen alle er større end eller lig med 5, kan testsandsynligheden derfor beregnes som

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2(r-1)(s-1)}(-2\ln Q(\mathbf{x})). \quad (6.24)$$

Accept af  $H_{01}$  reducerer modellen  $M_0$  til modellen

$$M_1 : \mathbf{X} = \{X_{ij}\} \sim m(n, (\{\rho_i \sigma_j\})). \quad (6.25)$$

I  $M_1$  kan det vises, at vektorerne af rækkesummer  $\mathbf{X}_{*} = (X_{1.}, \dots, X_{i.}, \dots, X_{r.})$  og søjlesummer  $\mathbf{X}_{. * } = (X_{.1}, \dots, X_{.j}, \dots, X_{.s})$  er stokastisk uafhængige og multinomialfordelte. Mere præcist har vi

$$\begin{aligned} \mathbf{X}_{*} &= (X_{1.}, \dots, X_{i.}, \dots, X_{r.}) \sim m(n, (\rho_1, \dots, \rho_i, \dots, \rho_r)) \\ \mathbf{X}_{. *} &= (X_{.1}, \dots, X_{.j}, \dots, X_{.s}) \sim m(n, (\sigma_1, \dots, \sigma_j, \dots, \sigma_s)) \\ \mathbf{X}_{*} \text{ og } \mathbf{X}_{. *} &\text{ er stokastisk uafhængige.} \end{aligned} \quad (6.26)$$

Inferens i  $M_1$  vedrørende henholdsvis vektoren  $\boldsymbol{\rho}$  af rækkesandsynligheder og vektoren  $\boldsymbol{\sigma}$  af søjlesandsynligheder kan derfor foretages ved at betragte fordelingen af henholdsvis rækkesummer  $\mathbf{X}_{*}$  og søjlesummer  $\mathbf{X}_{. *}$ .

### Eksempel 6.3 (Fortsat)

Suppleres tabellen over observerede antal med rækkesummer, søjlesummer og totalsum får vi følgende:

	timer per uge	rygerstatus		i alt
		ryger	ikke-ryger	
idrætsaktivitet	0.0-0.5	181	603	784
	0.5-2.0	158	591	749
	2.0-4.0	162	713	875
	4.0-7.0	150	697	847
	7.0-	83	531	614
	<i>i alt</i>	734	3135	3869

Ud fra denne tabel beregnes de forventede antal  $e$  ved hjælp af formel (6.22) og med tre decimalers nøjagtighed finder vi:

	<i>timer per uge</i>	<i>rygerstatus</i>		<i>i alt</i>
		<i>ryger</i>	<i>ikke-ryger</i>	
<i>idrætsaktivitet</i>	0.0-0.5	148.735	635.265	784.000
	0.5-2.0	142.095	606.905	749.000
	2.0-4.0	165.999	709.001	875.000
	4.0-7.0	160.687	686.313	847.000
	7.0-	116.484	497.516	614.000
	<i>i alt</i>	734.000	3135.000	3869

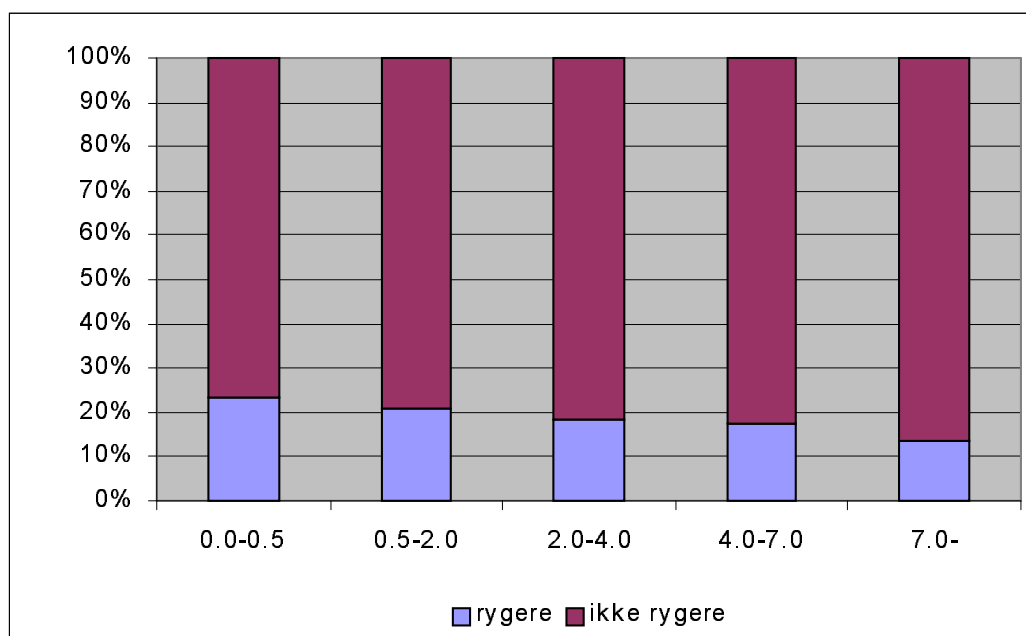
Af formel (6.23) fås, at

$$-2 \ln Q(\mathbf{x}) = 2[23894.4253 - 25761.8754 - 30081.2609 + 31960.8470] = 24.2719,$$

og da de forventede antal alle er større end 5, kan testsandsynligheden for uafhængighedshypotesen ved hjælp af (6.23) beregnes til

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(4)}(24.2719) = 0.00007,$$

og hypotesen  $H_{01}$  forkastes. På grundlag af denne undersøgelse kan vi altså konkludere, at der er en sammenhæng mellem idrætsaktivitet og rygestatus. Af figuren nedenfor ses, at procentdelen af rygere aftager når idrætsaktiviteten vokser.



□

### 6.3 Inferens i flere multinomialfordelinger

Teorien for statistisk inferens i én multinomialfordeling, omtalt i Afsnit 6.2, kan uden videre generaliseres til flere multinomialfordelinger. Vi vil ikke gennemgå denne teori her, men blot illustrere den ved hjælp af et enkelt eksempel.

#### 6.3.1 Homogenitet af flere multinomialfordelinger

Problemstillingen skitseret i Eksempel 6.4 er et specialtilfælde af følgende generelle situation. Antag, at data kan beskrives som udfald af  $r$  uafhængige  $s$ -dimensionale diskrete stokastiske vektorer  $\mathbf{X}_i = (X_{i1}, \dots, X_{ij}, \dots, X_{is})$ , som er multinomialfordelt med antalsparameter  $n_i$  og sandsynlighedsvektor  $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{ij}, \dots, \pi_{is})$ ,  $i = 1, \dots, r$ , samt at vi ønsker at undersøge om sandsynlighedsvektorerne i de  $r$  fordelinger kan antages at være identiske. Observationerne kan da opstilles i et  $r \times s$  skema som nedenfor, hvor

$$x_{.j} = \sum_{i=1}^r x_{ij} \quad \text{og} \quad n_{.} = \sum_{i=1}^r n_i.$$

I modellen

$$M_0: \mathbf{X}_i = (X_{i1}, \dots, X_{ij}, \dots, X_{is}) \sim m(n_i, \boldsymbol{\pi}_i) = m(n_i, (\pi_{i1}, \dots, \pi_{ij}, \dots, \pi_{is})) \quad (6.27)$$

$\mathbf{X}_1, \dots, \mathbf{X}_i, \dots, \mathbf{X}_r$  er stokastisk uafhængige

omtales hypotesen

$$H_{01}: \boldsymbol{\pi}_1 = \dots = \boldsymbol{\pi}_i = \dots = \boldsymbol{\pi}_r = \boldsymbol{\pi} = (\pi_1, \dots, \pi_j, \dots, \pi_s) \quad (6.28)$$

som hypotesen om *homogenitet*.

	1	...	$j$	...	$s$	$\Sigma$
1	$x_{11}$	...	$x_{1j}$	...	$x_{1s}$	$n_1$
.	.	...	.	...	.	.
.	.	...	.	...	.	.
$i$	$x_{i1}$	...	$x_{ij}$	...	$x_{is}$	$n_i$
.	.	...	.	...	.	.
.	.	...	.	...	.	.
$r$	$x_{r1}$	...	$x_{rj}$	...	$x_{rs}$	$n_r$
$\Sigma$	$x_{.1}$	...	$x_{.j}$	...	$x_{.s}$	$n_{.}$

Idet en  $s$ -dimensional multinomialfordeling har  $s - 1$  frie parametre og idet modellen  $M_0$  består af  $r$  uafhængige fordelinger af denne slags, har  $M_0$  i alt  $r(s - 1)$  frie parametre. Likelihood

funktionen under  $M_0$  er

$$L(\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_i, \dots, \boldsymbol{\pi}_r) = \prod_{i=1}^r \frac{n_i!}{x_{i1}! \cdots x_{ij}! \cdots x_{is}!} \pi_{i1}^{x_{i1}} \cdots \pi_{ij}^{x_{ij}} \cdots \pi_{is}^{x_{is}}, \quad (6.29)$$

og maksimum likelihood estimatet under  $M_0$  er givet ved

$$\hat{\pi}_{ij} = \frac{x_{ij}}{n_i}, \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Modellen svarende til homogenitetshypotesen er

$$M_1: \mathbf{X}_i = (X_{i1}, \dots, X_{ij}, \dots, X_{is}) \sim m(n_i, \boldsymbol{\pi}) = m(n_i, (\pi_1, \dots, \pi_j, \dots, \pi_s)) \quad (6.30)$$

$\mathbf{X}_1, \dots, \mathbf{X}_i, \dots, \mathbf{X}_r$  er stokastisk uafhængige

Denne model har  $s - 1$  frie parametre og likelihood funktionen for  $\boldsymbol{\pi}$  fås ved at sætte  $\pi_{ij} = \pi_j$  i (6.29), det vil sige

$$L(\boldsymbol{\pi}) = \left\{ \prod_{i=1}^r \frac{n_i!}{x_{i1}! \cdots x_{ij}! \cdots x_{is}!} \right\} \pi_1^{x_{.1}} \cdots \pi_j^{x_{.j}} \cdots \pi_s^{x_{.s}}$$

Ved hjælp af Sætning 6.1 ses det, at maksimum likelihood estimatet for den fælles sandsynlighedsparameter  $\boldsymbol{\pi}$  er givet ved

$$\hat{\pi}_j = \frac{x_{.j}}{n}, \quad j = 1, \dots, s. \quad (6.31)$$

De forventede antal under  $M_1$  bliver derfor

$$\begin{aligned} e_{ij} &= n_i \hat{\pi}_j \\ &= \frac{n_i x_{.j}}{n}, \quad i = 1, \dots, r, \quad j = 1, \dots, s; \end{aligned} \quad (6.32)$$

altså det forventede antal i den  $j$ 'te kategori i den  $i$ 'te fordeling er produktet af den  $i$ 'te rækkesum og den  $j$ 'te søjlesum divideret med totalsummen.

Beregningsformlen for  $-2 \ln Q$ -testoren er

$$\begin{aligned} -2 \ln Q(\mathbf{x}) &= 2 \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln \left( \frac{x_{ij}}{e_{ij}} \right) \\ &= 2 \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln \left( \frac{x_{ij}}{n_i x_{.j} / n} \right) \\ &= 2 \sum_{i=1}^r \sum_{j=1}^s x_{ij} [\ln(x_{ij}) - \ln(n_i) - \ln(x_{.j}) + \ln(n)] \\ &= 2 \left[ \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln(x_{ij}) - \sum_{i=1}^r n_i \ln(n_i) - \sum_{j=1}^s x_{.j} \ln(x_{.j}) + n \ln(n) \right]. \end{aligned} \quad (6.33)$$

Antallet af frihedsgrader i  $-2 \ln Q$ -testet er  $f = r(s - 1) - (s - 1) = (r - 1)(s - 1)$  så hvis de forventede antal er større end eller lig med 5, kan testsandsynligheden for homogenitetshypotesen beregnes som

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2(r-1)(s-1)}(-2 \ln Q(\mathbf{x})). \quad (6.34)$$

I modellen  $M_1$  kan det vises, at vektorsummen  $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_r$  er multinomialfordelt med antalsparameter  $n$ . og sandsynlighedsvektor  $\boldsymbol{\pi}$ ,

$$\mathbf{X} = (X_{.1}, \dots, X_{.j}, \dots, X_{.s}) \sim m(n, \boldsymbol{\pi}) = m(n, (\pi_1, \dots, \pi_j, \dots, \pi_s)). \quad (6.35)$$

Dette resultat kan benyttes, hvis man ønsker at drage yderligere inferens om den fælles sandsynlighedsvektor  $\boldsymbol{\pi}$ .

### Forskelle og ligheder mellem testene for uafhængighed og homogenitet

Sammenlignes formlerne (6.22)-(6.24) og (6.32)-(6.34) ses, at *beregningerne er identiske* for de to test. Testene vedrører *forskellige hypoteser* og foretages i *forskellige modeller*. Uafhængighedstestet foretages i en model, der kun involverer én multinomialfordeling, kun det totale antal observationer  $n$  er ikke-stokastisk. Homogenitetstestet foretages i en model, der omfatter  $r$  multinomialfordelinger, og i denne model er antallene af observationer  $n_1, \dots, n_i, \dots, n_r$  i de  $r$  fordelinger ikke-stokastiske. Der anvendes med andre ord forskellige strategier ved indsamlingen af data i de to situationer.

#### Eksempel 6.4 (Fortsat)

Spørgsmålet om der er forskel på sandsynlighederne for henholdsvis hjemmesejr, uafgjort og udesejr i de tre dele af Faxe Kondi Ligaen 1999-2000 kan besvares i modellen

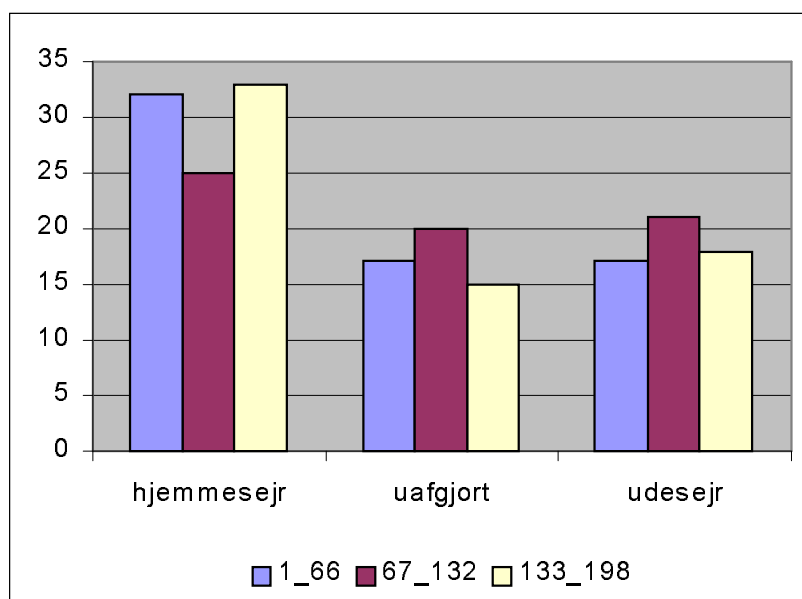
$$M_0 : \mathbf{X}_i \sim m(66, \boldsymbol{\pi}_i), \quad i = 1, 2, 3,$$

$\mathbf{X}_1, \mathbf{X}_2$  og  $\mathbf{X}_3$  er stokastisk uafhængige

ved at teste hypotesen om identitet af sandsynlighedsvektorerne, altså hypotesen

$$H_{01} : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = \boldsymbol{\pi}_3 (= \boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)).$$

Figuren nedenfor antyder, at  $H_{01}$  kan accepteres.



Suppleres tallene side 6.4 med søjlesummerne får vi

<i>kamp nr.</i>	<i>hjemmesejr</i>	<i>uafgjort</i>	<i>udesejr</i>	<i>i alt</i>
1 - 66	32	17	17	66
67 - 132	25	20	21	66
133 - 199	33	15	18	66
<i>i alt</i>	90	52	56	198

og ved hjælp af (6.32) beregnes de forventede antal - med tre decimalers nøjagtighed - til:

<i>kamp nr.</i>	<i>hjemmesejr</i>	<i>uafgjort</i>	<i>udesejr</i>	<i>i alt</i>
1 - 66	30.000	17.333	18.667	66.000
67 - 132	30.000	17.333	18.667	66.000
133 - 199	30.000	17.333	18.667	66.000
<i>i alt</i>	90.000	51.999	56.001	198.000

Da de forventede antal alle er større end 5, finder vi ved hjælp af (6.33) og (6.34), at

$$-2\ln Q(\mathbf{x}) = 2[619.5865 - 829.5516 - 835.8672 + 1047.0769] = 2.4890$$

og at

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(4)}(2.4890) = 0.6466.$$

Hypotesen om homogenitet accepteres, det vil sige, at sandsynlighederne for henholdsvis hjemmesejr, uafgjort og udesejr i de tre dele af turneringen kan antages at være ens.

Den simple hypotese

$$H_{02} : \boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3) = (1/3, 1/3, 1/3),$$

det vil sige hypotesen om at de tre udfald - hjemmesejr, uafgjort, udesejr - af en kamp er lige sandsynlige ser ikke ud til at kunne accepteres ud fra figuren ovenfor. Antallet af hjemmesejre ser ud til at være signifikant større end antallet af uafgjorte og udesejre. Hypotesen kan ifølge bemærkningen efter formel (6.35) testes i modellen

$$M : \mathbf{X} \sim m(198, \boldsymbol{\pi}),$$

hvor  $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3$ . Ved hjælp af tabellen

	<i>hjemmesejr</i>	<i>uafgjort</i>	<i>udesejr</i>	<i>i alt</i>
<i>observeret <math>\mathbf{x}</math></i>	90	52	56	198
<i>forventet <math>\mathbf{e}</math></i>	66	66	66	198

beregnes  $-2 \ln Q$ -teststørrelsen og den tilsvarende testsandsynlighed til

$$-2 \ln Q(\mathbf{x}) = 12.6312$$

og

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(2)}(12.6312) = 0.0018,$$

så  $H_{02}$  forkastes.

95% konfidensintervallet for  $\pi_1$  - sandsynligheden for hjemmesejr - kan beregnes ved hjælp af (6.17). Da  $x = 90$  og  $n = 198$  bliver 95% konfidensintervallet for  $\pi_1$

$$[0.3867, 0.5241].$$

Vi kan således ikke afvise hypotesen om at sandsynligheden  $\pi_1$  for hjemmesejr er  $\frac{1}{2}$ . □

## 6.4 Fishers eksakte test

Alle de test, som vi har benyttet i Afsnit 6.2 og 6.3, har været approksimative test, der er baseret på den approksimative likelihood teori omtalt i Afsnit 5.7. For de betragtede test har vi brugt kriteriet, at de forventede antal  $\mathbf{e}$  alle skulle være større end eller lig 5, for at testet kunne anvendes. Dette kriterium er baseret på numeriske simulationer, og for nogle modeller gælder der, at det kan slækkes, således at den udregnede testsandsynlighed er troværdig, selvom nogle af de

forventede antal er noget mindre end 5. Spørgsmålet om, hvad man skal gøre, hvis de forventede antal er for små, trænger sig dog ofte på i anvendelser af teorien. Vi skal nu omtale Fishers eksakte test, der ofte benyttes i forbindelse med  $r \times s$  tabeller, hvor nogle af de forventede antal er for små. Metoden kan altså anvendes for test af uafhængighed mellem inddelingskriterier og test for homogenitet. Vi giver en detaljeret beskrivelse af metoden for  $2 \times 2$  tabeller, som giver det princip, der benyttes i det generelle tilfælde. Beregningerne i Fishers eksakte test er ofte for omfattende til at kunne gøres manuelt, men nogle statistikpakker (dog ikke *Excel*) er i stand til at udføre testet. Lad os indledningsvis betragte et eksempel, som ikke har noget med idræt at gøre, men som er interessant idet det var genstand for meget stor opmærksomhed i medierne.

### Eksempel 6.5

I en undersøgelse, foretaget af Kræftens Bekæmpelses Cancerregister, beskæftiger man sig med spørgsmålet, om børn med bopæl tæt ved højspændingsanlæg har en forøget risiko for at få kræft. Undersøgelsen er en såkaldt *case-kontrol undersøgelse*. I Cancerregisteret er der i perioden 1968-1986 registreret 1707 tilfælde af sygdommene leukæmi, hjernesvulst eller lymfom blandt børn, der på tidspunktet for diagnosen var under 15 år. Disse børn udgør *casegruppen*. For hvert af børnene i denne gruppe er der tilfældigt udvalgt et antal børn af samme køn og alder. Disse børn udgør *kontrolgruppen*. For samtlige børn har man derefter vurderet, om de har boet så tæt ved højspændingsledninger, at de på årsbasis har været udsat for et gennemsnitligt magnetfelt på  $0.10 \mu T$  (microTesla) eller mere. Vi skal her betragte casegruppen for lymfomer og den tilsvarende kontrolgruppe, som blev valgt fem gange så stor. For denne gruppe er de *observerede antal*:

<i>eksponering</i>	$\geq 0.10 \mu T$	$< 0.10 \mu T$	<i>i alt</i>
<i>case</i>	3	247	250
<i>kontrol</i>	3	1247	1250
<i>i alt</i>	6	1494	1500

Princippet i en case-kontrol undersøgelse er at sammenligne hyppigheder. Hvis for eksempel hyppigheden af tilfælde med eksponering  $\geq 0.10 \mu T$  er signifikant større i casegruppen end i kontrolgruppen, konkluderes det, at eksponering medfører en øget risiko for kræft. Hvis størrelserne af de to grupper anses for faste, er det rimeligt at betragte modellen

$$M_0: X_i \sim b(n_i, p_i) \quad i = C, K, \quad (6.36)$$

$X_C$  og  $X_K$  er stokastisk uafhængige

og i denne teste hypotesen

$$H_0: p_C = p_K. \quad (6.37)$$

Af tabellen over observerede antal ses, at  $\hat{p}_C = 0.0120$  og  $\hat{p}_K = 0.0024$ , dvs. hyppigheden for de eksponerede i casegruppen er fem gange så stor som hyppigheden i kontrolgruppen. Spørgsmålet er nu, om denne forskel er signifikant.

De forventede antal under  $H_0$  findes ved hjælp af (6.32) til

eksponering	$\geq 0.10 \mu T$	$< 0.10 \mu T$	<i>i alt</i>
<i>case</i>	1	249	250
<i>kontrol</i>	5	1245	1250
<i>i alt</i>	6	1494	1500

I (1,1)-cellen er det forventede antal 1, og vi kan derfor ikke bruge det approksimative test.  $\square$

### Fishers eksakte test i $2 \times 2$ tabeller

Af hensyn til den senere omtale af Fishers test i den generelle  $r \times s$  tabel formulerer vi modellen i (6.36) og hypotesen (6.37) ved hjælp af multinomialfordelingen i stedet for binomialfordelingen. Vi betragter altså modellen

$$M_0: \mathbf{X}_i = (X_{i1}, X_{i2}) \sim m(n_i, (\pi_{i1}, \pi_{i2})), \quad i = 1, 2,$$

$\mathbf{X}_1$  og  $\mathbf{X}_2$  er stokastisk uafhængige

og i denne hypotesen om identitet af de to sandsynlighedsvektorer, dvs.

$$H_0: \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = \boldsymbol{\pi} = (\pi_1, \pi_2).$$

Fisher foreslog at teste hypotesen  $H_0$  ved at betragte den betingede fordeling, under  $H_0$ , af  $(\mathbf{X}_1, \mathbf{X}_2)$  givet  $\mathbf{X}_. = \mathbf{X}_1 + \mathbf{X}_2$ . Benyttes (6.35), findes den betingede fordeling af følgende beregninger:

$$P((\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{x}_1, \mathbf{x}_2) | \mathbf{X}_. = \mathbf{x}_.) = \frac{P((\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{x}_1, \mathbf{x}_2))}{P(\mathbf{X}_. = \mathbf{x}_.)}$$

$$= \frac{\frac{n_1!}{x_{11}!(n_1 - x_{11})!} \pi_1^{x_{11}} \pi_2^{n_1 - x_{11}} \frac{n_2!}{x_{21}!(n_2 - x_{21})!} \pi_1^{x_{21}} \pi_2^{n_2 - x_{21}}}{\frac{n!}{x_{.1}!(n - x_{.1})!} \pi_1^{x_{.1}} \pi_2^{n - x_{.1}}},$$

hvilket efter passende forkortelser kan skrives ved hjælp af binomialkoefficienter på følgende måde

$$P((\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{x}_1, \mathbf{x}_2) | \mathbf{X}_. = \mathbf{x}_.) = \frac{\binom{x_{.1}}{x_{11}} \binom{n - x_{.1}}{n_1 - x_{11}}}{\binom{n}{n_1}}.$$

Bemærk, at da vi har betinget med  $\mathbf{X} = \mathbf{x}$ , er størrelserne  $x_{.1}, n_1, n_2$  og derfor også  $n$ , faste i dette udtryk, således at kun  $x_{11}$  varierer.

Den diskrete fordeling med sandsynlighedsfunktionen

$$h(x : M, N, n) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}, \quad x = K_0, \dots, K_1,$$

hvor  $K_0 = \max\{0, n + M - N\}$  og  $K_1 = \min\{M, n\}$ , kaldes den *hypergeometriske fordeling med parametre  $M, N$  og  $n$* . Det ovenstående viser derfor, at den betingede fordeling af  $X_{11}$  givet søjle- og rækkesummer er den hypergeometriske fordeling med parametre  $x_{.1}, n$ . og  $n_1$ .

Testsandsynligheden for Fishers eksakte test er

$$\varepsilon_F(\mathbf{x}) = \sum_y^* h(y; x_{.1}, n, n_1), \quad (6.38)$$

hvor \* over summationstegnet antyder, at summationen skal foretages over alle  $y$  for hvilke  $h(y; x_{.1}, n, n_1) \leq h(x_{11}, x_{.1}, n, n_1)$ . Denne definition af testsandsynligheden kan forklares på følgende måde. Benyttes sandsynlighederne i den betingede fordeling som et mål for, hvor ekstreme de forskellige observationer er, får vi den sædvanlige fortolkning af testsandsynligheden, som sandsynligheden for de udfald  $y$ , der er ligeså ekstreme eller mere ekstreme end det observerede udfald  $x_{11}$ .

Vi har tidligere bemærket, at beregningerne i testet for uafhængighed mellem inddelingskriterier er identiske med beregningerne i testet for homogenitet. Det er derfor ikke overraskende, at de ovenstående beregninger giver samme resultat, hvis man ønsker at teste uafhængighed i en  $2 \times 2$  tabel. Den eneste forskel er, at parametrene i den betingede hypergeometriske fordeling bliver  $x_{.1}, n$  og  $x_1$ . i stedet for  $x_{.1}, n$ . og  $n_1$ .

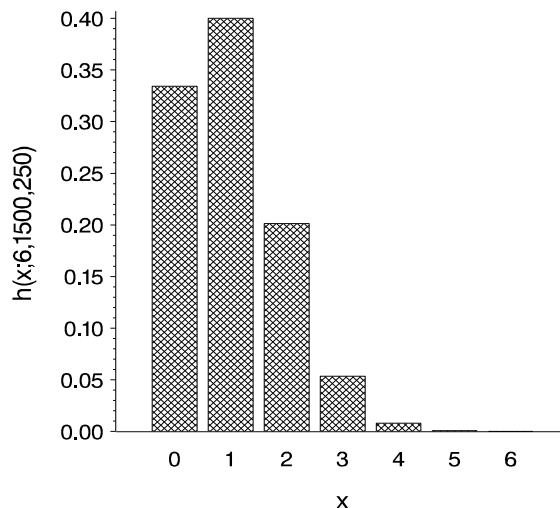
Manuelt kan det være besværligt at beregne testsandsynligheden i (6.38), men som omtalt tidligere har alle de gængse statistikpakker en procedure til at beregne denne.

### Eksempel 6.5 (Fortsat)

Den relevante hypergeometriske fordeling for dette datasæt har parametrene 6, 1500, og 250. Fordelingen er vist i Figur 6.2. I dette tilfælde bliver

$$\varepsilon_F(\mathbf{x}) = 0.062,$$

hvilket ikke giver anledning til at forkaste hypotesen (6.37). Med andre ord, gruppen af lymfomer er ikke signifikant forskellig fra kontrolgruppen med hensyn til eksponering fra magnetfelter. Lad os til sidst i dette eksempel se på, hvilke konklusioner vi havde fået, hvis vi fejlagtigt



**Figur 6.2** Sandsynlighedsfunktionen for den hypergeometriske fordeling med parametre  $(M, N, n) = (6, 1500, 250)$ .

havde brugt  $-2 \ln Q$ -testet eller  $X^2$ -testet. Af størrelserne

$$\begin{aligned} -2 \ln Q(\mathbf{x}) &= 3.546 & \varepsilon(\mathbf{x}) &= 1 - F_{\chi^2(1)}(3.546) = 0.060 \\ X^2(\mathbf{x}) &= 4.819 & \varepsilon^*(\mathbf{x}) &= 1 - F_{\chi^2(1)}(4.819) = 0.028 \end{aligned}$$

ses det, at  $-2 \ln Q$ -testet giver samme konklusion som Fishers eksakte test. Brug af  $X^2$ -testet i denne situation medfører derimod, at man fejlagtigt konstaterer en signifikant forskel på casegruppen og kontrolgruppen. Årsagen til den megen omtale i medierne var, at konklusionen i undersøgelsen var baseret på  $X^2$ -testet.  $\square$

### Fishers eksakte test for $r \times s$ tabeller

Princippet for dette test er det samme som for  $2 \times 2$  tabellen. Testsandsynligheden beregnes i den betingede fordeling af  $r \times s$  tabellen givet række- og søjlesummer ved at summere de betingede sandsynligheder for alle tabeller, der har en mindre betinget sandsynlighed end den observerede tabel. Regnearbejdet her er næsten altid så besværligt, at det er nødvendigt at benytte en statistikpakke for at udføre testet.

## 6.5 Test for goodness of fit

I Afsnit 4.1 så vi, hvorledes man ved hjælp af fraktildiagrammer grafisk er i stand til at undersøge, om data  $\mathbf{x} = (x_1, \dots, x_n)$  kan betragtes som en stikprøve fra en klasse af fordelinger, der er karakteriseret ved en positions- og/eller en skalaparameter. Vurderinger af fraktildiagrammer, og andre grafiske metoder til kontrol af en statistisk model, er naturligvis i et vist omfang subjektive, selvom man ved hjælp af simulationer, som i Appendiks B, kan opnå indsigt i, hvorledes de relevante tegninger skal vurderes. Hvis antallet af observationer  $n$  i stikprøven er tilstrækkelig stort, kan den grafiske kontrol suppleres med et numerisk test, et såkaldt test for *goodness of fit*.

Denne form for modelkontrol er generel, det vil sige, at den også kan anvendes i situationer, hvor den betragtede fordelingsklasse ikke er en positions-skala familie. Mere præcist ønsker vi at undersøge, om data kan opfattes som en stikprøve fra en fordeling  $F_{\theta}$ , der tilhører en fordelingsklasse

$$\mathcal{F} = \{F_{\theta} : \theta \in \Theta\},$$

som er parametriseret ved en  $d$ -dimensional parameter  $\theta$ , altså  $\theta \in \Theta \subseteq R^d$ . Ønsker man eksempelvis at undersøge, om  $\mathbf{x}$  er en stikprøve fra normalfordelingen er  $d = 2$ , idet normalfordelingen parametriseres ved  $(\mu, \sigma^2)$  middelværdi og varians. Hvis fordelingsklassen  $\mathcal{F}$  er mængden af Poisson fordelinger, er  $d = 1$ , da disse fordelinger parametriseres ved middelværdien  $\lambda$ , etc.

Antag, at  $-\infty \leq y_0 < y_1 < \dots < y_j < \dots < y_k \leq \infty$  bestemmer en inddeling af  $R$  i  $k$  intervaller  $I_1, \dots, I_j, \dots, I_k$ , hvor

$$I_j = ]y_{j-1}, y_j], \quad j = 1, \dots, k.$$

Lad  $a_j$  betegne antallet af observationer, som tilhører det  $j$ 'te interval, altså

$$a_j = \#\{i : x_i \in I_j\}, \quad j = 1, \dots, k.$$

Da observationerne  $x_1, \dots, x_n$  antages at være uafhængige og identisk fordelte, vælger vi følgende model for de *observerede antal*  $\mathbf{a} = (a_1, \dots, a_j, \dots, a_k)$ :

$$M_0 : (a_1, \dots, a_j, \dots, a_k) \sim m(n, (\pi_1, \dots, \pi_j, \dots, \pi_k)), \quad (6.39)$$

hvor  $\pi_j$  er sandsynligheden for at observationen tilhører det  $j$ 'te interval  $I_j$ , det vil sige

$$\pi_j = P(X_1 \in I_j), \quad j = 1, \dots, k.$$

Under hypotesen

$$H_0 : X_i \sim F_{\theta}, \quad i = 1, \dots, n,$$

er

$$\pi_j = \pi_j(\boldsymbol{\theta}) = F_{\boldsymbol{\theta}}(y_j) - F_{\boldsymbol{\theta}}(y_{j-1}), \quad j = 1, \dots, k, \quad (6.40)$$

så  $H_0$  kan betragtes som en hypotese i multinomialmodellen  $M_0$ . Da  $\boldsymbol{\theta}$  er  $d$ -dimensional, er antallet af frie parametre i  $H_0$  netop  $d$ .

Likelihood funktionen under  $H_0$  svarende til de observerede antal  $\mathbf{a}$  er

$$L(\boldsymbol{\theta}) = \frac{n!}{a_1! \cdots a_j! \cdots a_k!} \pi_1(\boldsymbol{\theta})^{a_1} \cdots \pi_j(\boldsymbol{\theta})^{a_j} \cdots \pi_k(\boldsymbol{\theta})^{a_k}.$$

Da udtrykkene i (6.40) ofte kan være komplicerede, vælger man at estimere  $\boldsymbol{\theta}$  på grundlag af de oprindelige observationer  $x_1, \dots, x_n$ , altså ved hjælp af likelihood funktionen

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i; \boldsymbol{\theta}), \quad (6.41)$$

hvor  $f(\cdot; \boldsymbol{\theta})$  betegner tæthedsfunktionen svarende til  $F_{\boldsymbol{\theta}}$ . Lad  $\hat{\boldsymbol{\theta}}$  betegne maksimum likelihood estimatet for  $\boldsymbol{\theta}$  beregnet ved hjælp af (6.41). Estimererne for parametrene i multinomialfordelingen bliver da

$$\pi_j(\hat{\boldsymbol{\theta}}) = F_{\hat{\boldsymbol{\theta}}}(y_j) - F_{\hat{\boldsymbol{\theta}}}(y_{j-1}),$$

og de forventede antal  $\mathbf{e}$  under  $H_0$  bliver dermed

$$\begin{aligned} e_j &= n\pi_j(\hat{\boldsymbol{\theta}}) \\ &= n\{F_{\hat{\boldsymbol{\theta}}}(y_j) - F_{\hat{\boldsymbol{\theta}}}(y_{j-1})\}, \quad j = 1, \dots, k. \end{aligned} \quad (6.42)$$

Testsandsynligheden for  $-2 \ln Q$ -testoren for  $H_0$

$$-2 \ln Q(\mathbf{a}) = 2 \sum_{j=1}^k a_j \ln \left( \frac{a_j}{e_j} \right) \quad (6.43)$$

approksimeres ved

$$\varepsilon(\mathbf{a}) \doteq 1 - F_{\chi^2(k-1-d)}(-2 \ln Q(\mathbf{a})), \quad (6.44)$$

forudsat at de forventede antal  $\mathbf{e}$  alle er større end eller lig med 5.

I litteraturen ser man meget ofte  $X^2$ -testet anvendt i forbindelse med test for goodness of fit. Med notationen her er  $X^2$ -testoren og den tilsvarende testsandsynlighed:

$$X^2(\mathbf{a}) = \sum_{j=1}^k \frac{(a_j - e_j)^2}{e_j} \quad (6.45)$$

$$\varepsilon^*(\mathbf{a}) \doteq 1 - F_{\chi^2(k-1-d)}(X^2(\mathbf{a})), \quad (6.46)$$

hvor approksimationen kan anvendes, hvis de forventede antal er større end eller lig med 5.

**Eksempel 6.6**

Hvis vi ved hjælp af et test for goodness of fit ønsker at undersøge, om observationerne  $x_1, \dots, x_n$  kan opfattes som en stikprøve fra normalfordelingen, som parametriseres ved sin middelværdi  $\mu$  og varians  $\sigma^2$ , det vil sige  $d = 2$ , estimerer vi først disse to parametre:

$$\begin{aligned}\mu &\leftarrow \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ \sigma^2 &\leftarrow s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.\end{aligned}$$

Idet fordelingsfunktionen for normalfordelingen med middelværdi  $\bar{x}$  og varians  $s^2$  kan udtrykkes ved hjælp af fordelingsfunktionen for den standardiserede normalfordeling på følgende måde

$$F_{(\bar{x}, s^2)}(y) = \Phi\left(\frac{y - \bar{x}}{s}\right),$$

bliver de forventede antal ifølge (6.42)

$$e_j = n\left\{\Phi\left(\frac{y_j - \bar{x}}{s}\right) - \Phi\left(\frac{y_{j-1} - \bar{x}}{s}\right)\right\}, \quad j = 1, \dots, k. \quad (6.47)$$

Som illustration af test for goodness of fit for normalfordelingen ser vi igen på målingerne i Eksempel 1.1 af højden hos 247 astmaplagede piger i alderen 10-12 år. Vi betragter den grupperede version af disse data, som er givet i Tabel 1.4, idet dog intervallerne  $]112, 116]$  og  $]164, 168]$  erstattes med henholdsvis  $] -\infty, 116]$  og  $]164, \infty[$ . Fra omtalen af Eksempel 1.1 i Afsnit 4.3 ved vi, at

$$\begin{aligned}\mu &\leftarrow \bar{x} = 140.13 \\ \sigma^2 &\leftarrow s^2 = 85.8317.\end{aligned}$$

De observerede og forventede antal, beregnet ved hjælp af (6.47), er angivet i Tabel 6.1.

Indledningsvis betragter vi 14 intervaller, men for at imødekomme kravet, om at de forventede antal skal være større end eller lig med 5, bliver vi nødt til at slå nogle af intervallerne sammen som antydet. Efter dette har vi  $k = 10$  intervaller, og da  $d = 2$  bliver antallet af frihedsgrader i testet for goodness of fit lig med  $k - 1 - d = 10 - 1 - 2 = 7$ .

For  $-2 \ln Q$ -testet får vi fra (6.43) og (6.44), at

$$-2 \ln Q(\mathbf{a}) = 7.2653$$

og

$$\varepsilon(\mathbf{a}) = 1 - F_{\chi^2(7)}(7.2653) = 0.4018.$$

Formlerne (6.45) og (6.46) medfører, at vi for  $X^2$ -testet finder

$$X^2(\mathbf{a}) = 7.5472$$

<i>interval</i>	<b>a</b>	<b>e</b>
$] -\infty, 116]$	1	1.136
$] 116, 120]$	0	2.544
$] 120, 124]$	8	6.407
$] 124, 128]$	20	13.432
$] 128, 132]$	24	23.435
$] 132, 136]$	32	34.032
$] 136, 140]$	49	41.132
$] 140, 144]$	41	41.378
$] 144, 148]$	26	34.646
$] 148, 152]$	21	24.148
$] 152, 156]$	14	14.005
$] 156, 160]$	6	6.761
$] 160, 164]$	4	2.716
$] 164, \infty[$	1	1.233

**Tabel 6.1** Beregning af test for goodness of fit for normalfordelingen for data i Tabel 2.2.

og dermed

$$\varepsilon^*(\mathbf{a}) = 1 - F_{\chi^2(7)}(7.5472) = 0.3742.$$

Intet af testene, af hvilke vi - som tidligere omtalt - foretrækker  $-2\ln Q$ -testet, giver altså anledning til at betvivle, at observationerne kan betragtes som en stikprøve fra normalfordelingen, hvilket er i overensstemmelse med fraktilsammenligningen i Figur 4.3.  $\square$

## Anneks til Kapitel 6

### Beregninger i Excel

I *Excel* er der ikke dialogbokse, der foretager beregningerne i modellerne i dette kapitel. Beregningerne foretages dog forholdsvis let og i disse er funktionen SUMPRODUKT meget nyttig, idet teststørrelsen

$$-2 \ln Q(\mathbf{x}) = 2 \sum_{i=1}^k x_i \ln\left(\frac{x_i}{e_i}\right)$$

bortset fra faktoren 2 (**husk denne**) netop er en sum af produkter mellem de observerede antal  $x_i$  og logaritmen  $\ln\left(\frac{x_i}{e_i}\right)$  til forholdet mellem de observerede antal  $x_i$  og de forventede antal  $e_i$ .

Endvidere er funktionen CHITEST som vist nedenfor ofte nyttig.

#### Eksempel 6.2 (Fortsat)

I regnearket nedenfor indeholder cellerne B4:D5 de observerede og forventede antal.

	A	B	C	D	E
1	Eksempel 6.2				
2					
3		sejr	uafgjort	nederlag	i alt
4	observeret	14	10	9	33
5	forventet	11	11	11	33
6					
7	ln(x/e)	0,241162	-0,09531	-0,200671	
8					
9	-2lnQ:	1,234261			
10	teststørrelse:	0,53949			

Indholdet af cellen B7 beregnes som

$$= \text{LN}(B\$4/B\$5) \quad (= \ln(x_1/e_1))$$

og analoge formler oprettes i cellerne C7 og D7. (De to \$ tegn letter oprettelsen af de analoge formler). Teststørrelsen i B9 beregnes som

$$= 2 * \text{SUMPRODUKT}(B4 : D4; B7 : D7) \quad (= 2 \sum_{i=1}^k x_i \ln\left(\frac{x_i}{e_i}\right))$$

og testsandsynligheden i B10 som

$$= \text{CHIFORDELING}(B9; 2) \quad (= 1 - F_{\chi^2(k-1)}(-2 \ln Q(\mathbf{x}))).$$

□

**Eksempel 6.3 (Fortsat)**

Data, suppleret med række- og søjlesummer samt totalsum, ses i cellerne A4:D10.

	A	B	C	D	E	F	G
1	Eksempel 6.3						
2							
3	observeret					forventet	
4	timer per uge	rygere	ikke rygere	i alt			
5	0.0-0.5	181	603	784		148,735	635,265
6	0.5-2.0	158	591	749		142,095	606,905
7	2.0-4.0	162	713	875		165,999	709,001
8	4.0-7.0	150	697	847		160,687	686,313
9	7.0-	83	531	614		116,484	497,516
10	i alt	734	3135	3869			
11							
12	ln(x/e)						
13		0,196330	-0,052125				
14		0,106098	-0,026556				
15		-0,024385	0,005624				
16		-0,068823	0,015452				
17		-0,338912	0,065134				
18							
19							
20	_2lnQ:	24,27192108					
21	testss:	0,000070					
22							
23	X^2:	23,70707007					
24	testss:	0,000091					

Ud fra disse celler beregnes de forventede antal i cellerne F5:G9. Først beregnes indholdet af F5 som

$$= B5 * D5 / D510 \quad (= x_{1.} x_{.1} / n)$$

og derefter oprettes analoge formel i de øvrige celler (De fire \$ tegn letter oprettelsen af de analoge formler).

Cellerne B13:C17 indeholder størrelserne  $\ln(x_{ij}/e_{ij})$ . Først beregnes indholdet af B13 som

$$= LN(B5/F5) \quad (= \ln(x_{11}/e_{11}))$$

hvorefter analoge formler oprettes i de resterende celler.

Teststørrelsen i B20 beregnes som

$$= 2 * \text{SUMPRODUKT}(B5 : C9; B13 : C17) \quad (= 2 \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln(\frac{x_{ij}}{e_{ij}}))$$

og testsandsynligheden i B21 som

$$= \text{CHIFORDELING}(B20; 4) \quad (= 1 - F_{\chi^2((r-1)(s-1))}(-2 \ln Q(\mathbf{x})).$$

Funktionen CHITEST beregner ud fra de observerede antal  $\{x_{ij}\}$  og de forventede antal  $\{e_{ij}\}$  testsandsynligheden for  $X^2$ -testet, det vil sige

$$\varepsilon^*(\mathbf{x}) = 1 - F_{\chi^2((r-1)(s-1))}(X^2(\mathbf{x})),$$

hvor

$$X^2(\mathbf{x}) = \sum_{i=1}^r \sum_{j=1}^s \frac{(x_{ij} - e_{ij})^2}{e_{ij}}.$$

Funktionen kaldes via ruten Indsæt  $\rightarrow$  Funktion  $\rightarrow$  Statistik  $\rightarrow$  CHITEST som giver en boks, hvor de observerede værdier angives efter Observeret\_værdi og de forventede efter Forventet\_værdi. Testsandsynligheden i B24 er fremkommet således eller ved direkte at indtaste

$$= \text{CHITEST}(B5 : C9; F5 : G9).$$

Værdien af  $X^2$ -teststørrelsen i B23 er derefter beregnet som

$$= \text{CHIINV}(B24; 4).$$

□

#### Eksempel 6.4 (Fortsat)

Da beregningerne for et homogenitetstest er de samme som beregningerne for testet for uafhængighed af inddelingskriterier, kan *Excel* beregningerne i dette eksempel udføres på samme måde som i Eksempel 6.3 ovenfor.

□

## Hovedpunkter til Kapitel 6

### Generelle modeller og hypoteser

#### Model:

Grundmodellen er baseret på den  $k$ -dimensionale multinomialfordeling med antalsparameter  $n$  og sandsynlighedsvektor  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_j, \dots, \pi_k)$

$$M_0 : \mathbf{X} = (X_1, \dots, X_j, \dots, X_k) \sim m(n, \boldsymbol{\pi}), \quad \boldsymbol{\pi} \in \boldsymbol{\Pi}.$$

#### Modelkontrol:

Check, at betingelser a) - d) side 6.1 med rimelighed kan antages at være opfyldt.

#### Estimat:

Sandsynlighedsvektoren  $\boldsymbol{\pi}$  estimeres på grundlag af de observerede antal  $\mathbf{x} = (x_1, \dots, x_j, \dots, x_k)$  som vektoren af relative hyppigheder

$$\boldsymbol{\pi} \leftarrow \hat{\boldsymbol{\pi}}(\mathbf{x}) = \left( \frac{x_1}{n}, \dots, \frac{x_j}{n}, \dots, \frac{x_k}{n} \right).$$

Fordelingen af estimatet angives ved

$$n\hat{\boldsymbol{\pi}} = \mathbf{X} \sim m(n, \boldsymbol{\pi}).$$

#### Konfidensintervaller:

Under  $M_0$  er  $X_j \sim b(n, \pi_j)$ , så konfidensintervallet for  $\pi_j$  kan beregnes på samme måde, som konfidensintervallet for sandsynlighedparameteren  $\pi$  i binomialmodellen  $X \sim b(n, \pi)$  beregnes på grundlag af observationen  $x$ . Dette interval, som er baseret på en approksimation, er

$$C_{1-\alpha}(x) = [\pi_-, \pi_+],$$

hvor

$$\pi_- = \frac{1}{n + u_{1-\alpha/2}^2} \left[ x + \frac{1}{2}u_{1-\alpha/2}^2 - u_{1-\alpha/2} \sqrt{\frac{x(n-x)}{n} + \frac{1}{4}u_{1-\alpha/2}^2} \right]$$

og

$$\pi_+ = \frac{1}{n + u_{1-\alpha/2}^2} \left[ x + \frac{1}{2}u_{1-\alpha/2}^2 + u_{1-\alpha/2} \sqrt{\frac{x(n-x)}{n} + \frac{1}{4}u_{1-\alpha/2}^2} \right].$$

#### Hypoteser:

En hypotese om sandsynlighedsvektoren har  $d$  frie parametre, hvis den er af formen

$$H_0 : \boldsymbol{\pi} \in \boldsymbol{\Pi}_0 = \boldsymbol{\pi}(\boldsymbol{\Theta}) (\subseteq \boldsymbol{\Pi}),$$

hvor  $\Pi_0$  er værdimængden for en en-entydig afbildning  $\pi$  fra en åben delmængde  $\Theta$  af  $R^d$  ind i  $\Pi$ . Mængden omtales som parametermængden for hypotesen  $H_0$

**Test af hypoteser:**

Hvis  $\hat{\theta}$  er maksimum likelihood estimatet for parameteren  $\theta$  under  $H_0$ , er vektoren  $\mathbf{e}$  af forventede antal under  $H_0$  givet ved

$$\mathbf{e} = (e_1, \dots, e_j, \dots, e_k) = (n\pi_1(\hat{\theta}), \dots, n\pi_j(\hat{\theta}), \dots, n\pi_k(\hat{\theta})).$$

Af de to approksimative test,  $-2 \ln Q$ -testet og  $X^2$ -testet, for  $H_0$  foretrækker vi  $-2 \ln Q$ -testet. Begge test er baseret på en sammenligning af de observerede antal  $\mathbf{x}$  og de forventede antal  $\mathbf{e}$ . Hvis **de forventede antal alle er større end eller lig med 5**, kan følgende teststørrelser og de tilsvarende approksimative testsandsynligheder benyttes.

$-2 \ln Q$ -testet:

$$-2 \ln Q(\mathbf{x}) = 2 \sum_{j=1}^k x_j \ln \left( \frac{x_j}{e_j} \right)$$

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2(k-1-d)}(-2 \ln Q(\mathbf{x})),$$

$X^2$ -testet:

$$X^2(\mathbf{x}) = \sum_{j=1}^k \frac{(x_j - e_j)^2}{e_j}$$

$$\varepsilon^*(\mathbf{x}) \doteq 1 - F_{\chi^2(k-1-d)}(X^2(\mathbf{x})).$$

## Specielle modeller og hypoteser

### Uafhængighed af inddelingskriterier:

De observerede antal  $\mathbf{x}$  er en  $r \times s$  tabel  $\{x_{ij}\}$ , som beskrives ved

$$M_0 : \mathbf{X} = \{X_{ij}\} \sim m(n, (\{\pi_{ij}\})).$$

Hypotesen

$$H_0 : \pi_{ij} = \rho_i \sigma_j, \quad i = 1, \dots, r, \quad j = 1, \dots, s,$$

omtales som hypotesen om uafhængighed af inddelingskriterier. Maksimum likelihood estimaterne for vektorene  $\boldsymbol{\rho}$  og  $\boldsymbol{\sigma}$  af række- og søjlesandsynligheder er

$$\hat{\rho}_i = \frac{x_{i.}}{n}, \quad i = 1, \dots, r \quad \text{og} \quad \hat{\sigma}_j = \frac{x_{.j}}{n}, \quad j = 1, \dots, s,$$

og de forventede antal  $\mathbf{e} = \{e_{ij}\}$  beregnes som

$$e_{ij} = \frac{x_{i.} x_{.j}}{n}.$$

Hvis disse **alle er større end eller lig med 5** bliver teststørrelsen (husk 2-tallet) og testsandsynligheden

$$\begin{aligned} -2 \ln Q(\mathbf{x}) &= 2 \left[ \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln(x_{ij}) - \sum_{i=1}^r x_{i.} \ln(x_{i.}) - \sum_{j=1}^s x_{.j} \ln(x_{.j}) + n \ln(n) \right] \\ \varepsilon(\mathbf{x}) &\doteq 1 - F_{\chi^2((r-1)(s-1))}(-2 \ln Q(\mathbf{x})). \end{aligned}$$

Accepteres  $H_0$  er modellen  $M_0$  reduceret til

$$M_1 : \mathbf{X} = \{X_{ij}\} \sim m(n, (\{\rho_i \sigma_j\})),$$

i hvilken der gælder

$$\begin{aligned} \mathbf{X}_{*} &= (X_{1.}, \dots, X_{i.}, \dots, X_{r.}) \sim m(n, (\rho_1, \dots, \rho_i, \dots, \rho_r)) \\ \mathbf{X}_{*} &= (X_{.1}, \dots, X_{.j}, \dots, X_{.s}) \sim m(n, (\sigma_1, \dots, \sigma_j, \dots, \sigma_s)) \\ \mathbf{X}_{*} &\text{ og } \mathbf{X}_{*} \text{ er stokastisk uafhængige.} \end{aligned}$$

### Homogenitet af flere multinomialfordelinger:

I modellen

$$\begin{aligned} M_0 : \mathbf{X}_i &= (X_{i1}, \dots, X_{ij}, \dots, X_{is}) \sim m(n_i, \boldsymbol{\pi}_i) = m(n_i, (\pi_{i1}, \dots, \pi_{ij}, \dots, \pi_{is})) \\ \mathbf{X}_1, \dots, \mathbf{X}_i, \dots, \mathbf{X}_r &\text{ er stokastisk uafhængige} \end{aligned}$$

testes hypotesen om homogenitet, eller identitet, af de  $r$  multinomialfordelinger

$$H_0 : \boldsymbol{\pi}_1 = \dots = \boldsymbol{\pi}_i = \dots = \boldsymbol{\pi}_r = \boldsymbol{\pi} = (\pi_1, \dots, \pi_j, \dots, \pi_s).$$

Maksimum likelihood estimatet for komponenterne i den fælles sandsynlighedsvektor er

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

og de forventede antal beregnes som

$$e_{ij} = \frac{n_i x_{.j}}{n}.$$

$-2 \ln Q$ -teststørrelsen er (husk 2-tallet)

$$-2 \ln Q(\mathbf{x}) = 2 \left[ \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln(x_{ij}) - \sum_{i=1}^r n_i \ln(n_i) - \sum_{j=1}^s x_{.j} \ln(x_{.j}) + n \ln(n) \right]$$

og testsandsynligheden kan beregnes som

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2((r-1)(s-1))}(-2 \ln Q(\mathbf{x})),$$

hvis de forventede antal **alle er større end eller lig med 5**.

Accepteres  $H_0$  er modellen reduceret til

$$M_1 : \mathbf{X}_i = (X_{i1}, \dots, X_{ij}, \dots, X_{is}) \sim m(n_i, \boldsymbol{\pi}) = m(n_i, (\pi_1, \dots, \pi_j, \dots, \pi_s))$$

$\mathbf{X}_1, \dots, \mathbf{X}_i, \dots, \mathbf{X}_r$  er stokastisk uafhængige

og i denne gælder der, at

$$\mathbf{X} = (X_{.1}, \dots, X_{.j}, \dots, X_{.s}) \sim m(n, \boldsymbol{\pi}) = m(n, (\pi_1, \dots, \pi_j, \dots, \pi_s)).$$

## Opgaver til Kapitel 6

**Opgave 6.1** I tabellen nedenfor ses en grupperet version af antallet af mål scoret i Faxe Kondi Ligaen 1999-2000 (Gruppen 0\_1 svarer til kampe hvor der blev scoret 0 eller 1 mål, mens gruppen >5 svarer til kampe hvor der blev scoret mere end 5 mål). Tabellen er desuden opdelt efter den første, anden og tredje tredjedel af turneringen.

	0_1	2	3	4	>5	i alt
kamp 1-66	15	18	13	15	5	66
kamp 67-132	12	22	10	10	12	66
kamp 133-198	10	20	13	9	14	66
i alt	37	60	36	34	31	198

- Vis, at fordelingen af mål er den samme i de tre dele af turneringen.
- Angiv et 95% konfidensinterval for sandsynligheden for at der scores mere end 5 mål i en kamp.

**Opgave 6.2** For de fleste af holdene i Faxe Kondi Ligaen 1999-2000 er det umiddelbart ud fra Tabel 1.3 let at bedømme om der er forskel på holdenes resultater på hjemmebane og på udebane, mens det for andre klubber ikke er oplagt, om der er en forskel. Tabellen nedenfor viser resultaterne for *OB*.

	<i>sejr</i>	<i>uafgjort</i>	<i>nederlag</i>	<i>i alt</i>
<i>hjemme</i>	4	6	6	16
<i>ude</i>	7	4	6	17
<i>i alt</i>	11	10	12	33

- Er det rimeligt, at antage at der ikke er forskel på *OB*'s resultater hjemme og ud?
- Antag, at der ikke er forskel på *OB*'s resultater hjemme og ude. Gør rede for at sejr, uafgjort og nederlag er lige sandsynlige udfald af *OB*'s kampe.

(For *AB*'s vedkommende gik vi i Eksempel 6.2 direkte til spørgsmål b) her, idet det ud fra resultaterne i Tabel 1.3 er oplagt at der ikke er forskel på resultaterne hjemme og ude.)

**Opgave 6.3** På side 6.19 så vi, at fordelingen af hjemmesejre, uafgjorte og udesejre i de 198 kampe i Faxe Kondi Ligaen 1999-2000

<i>hjemmesejr</i>	<i>uafgjort</i>	<i>udesejr</i>	<i>i alt</i>
90	52	56	198

kunne beskrives ved modellen

$$(X_1, X_2, X_3) \sim m(198, (\pi_1, \pi_2, \pi_3)).$$

Af tallene antyder, at hypotesen om at sandsynligheden  $\pi_2$  for en uafgjort er lig med sandsynligheden  $\pi_3$  for udesejr kan accepteres. Opgaven her vedrører test af denne hypotese. Lad  $p$  betegne den fælles værdi af  $\pi_2$  og  $\pi_3$  under hypotesen.

- a) Vis, at hypotesen kan formuleres som følgende hypotese med 1 fri parameter om sandsynlighedsvektoren  $(\pi_1, \pi_2, \pi_3)$ :

$$H_0 : (\pi_1, \pi_2, \pi_3) = (1 - 2p, p, p), \quad p \in ]0, 0.5[.$$

- b) Vis, at likelihood funktionen for  $p$  under  $H_0$  er

$$\begin{aligned} L(p) &= \frac{n!}{x_1!x_2!x_3!} (1 - 2p)^{x_1} p^{x_2+x_3} \\ &= \frac{n!}{x_1!x_2!x_3!} 2^{-(x_2+x_3)} (1 - 2p)^{x_1} (2p)^{x_2+x_3}, \end{aligned}$$

hvor  $n = 198$ ,  $(x_1, x_2, x_3) = (90, 52, 54)$ .

- c) Vis - eventuelt ved hjælp af Sætning 6.1 - at maksimum likelihood estimatet for  $p$  er

$$\hat{p} = \frac{x_2 + x_3}{2n}.$$

- d) Vis, at de forventede antal under  $H_0$  er

<i>hjemmesejr</i>	<i>uafgjort</i>	<i>udesejr</i>	<i>i alt</i>
90	54	54	198

og test hypotesen.

**Opgave 6.4** Ved de olympiske lege i Sydney blev der uddelt 301 guldmedaljer, 299 sølvmedaljer og 328 bronzemedaljer. Nedenfor ses medaljerne fordelt på de seks områder Afrika, Asien (inklusive Rusland og de baltiske lande), Australien (inklusive New Zeeland), Europa, Nordamerika

og Sydamerika (inklusive Mellemamerika).

	guld	sølv	bronze
Afrika	9	11	16
Asien	94	88	115
Australien	17	25	20
Europa	125	119	110
Nordamerika	42	28	41
Sydamerika	14	28	26
i alt	301	299	328

- a) Vis, at det antages, at fordelingen af medaljer på de seks områder er den samme for de tre slags medaljer.
- b) Angiv et 95% konfidensområde for sandsynligheden for at en medalje tilfalder Europa.

**Opgave 6.5** Tallene nedenfor vedrører de 404 trækninger i *Viking Lotto*, der er foretaget indtil den 25. 10. 2000. Ved hver trækning udtrækkes der 6 vindertal blandt tallene fra 1 til 48. I de første 230 trækninger blev der udtrukket 3 tillægstal og der efter 2. I de 404 trækninger er der derfor 2424 vindertal og 1038 tillægstal. Fordelingen af vindertal og tillægstal ses i tabellen nedenfor.

tal	gevinsttal	tillægstal	i alt
1	48	22	70
2	60	21	81
3	51	19	70
4	42	20	62
5	51	19	70
6	45	24	69
7	59	20	79
8	51	19	70
9	52	15	67
10	41	20	61
11	57	25	82
12	44	21	65
13	47	27	74
14	52	19	71
15	35	21	56
16	53	11	64
17	49	22	71
18	52	18	70
19	46	28	74
20	40	18	58
21	65	22	87
22	52	22	74
23	48	16	64
24	60	23	83

tal	gevinsttal	tillægstal	i alt
25	53	22	75
26	49	21	70
27	55	26	81
28	53	24	77
29	47	23	70
30	51	30	81
31	51	30	81
32	58	33	91
33	46	19	65
34	43	24	67
35	49	19	68
36	49	20	69
37	49	18	67
38	51	18	69
39	55	27	82
40	59	26	85
41	70	19	89
42	59	20	79
43	39	29	68
44	60	26	86
45	44	20	64
46	45	19	64
47	49	17	66
48	40	16	56
i alt	2424	1038	3462

Besvar ved hjælp af *Excel* følgende spørgsmål:

- a) Vis, at fordelingen af tallene fra 1 til 48 er den samme for vindertallene og for tillægstallene.

b) Undersøg, om tallen fra 1 til 48 udtrækkes lige hyppigt.

**Opgave 6.6** I *Lotto* er der indtil den 25.10. 2000 foretaget i alt 595 trækninger. I en trækning udtrækkes der 7 vindertal. I de første trækninger blandt tallene fra 1 til 34, senere kom tallet 35 til og endnu senere tallet 36. Antallet af tillægstal har også varieret og i de 595 trækninger er der udtrukket i alt 1534 tillægstal. Fordelingen af vindertal og tillægstal ses i tabellen nedenfor.

tal	vindertal	tillægst	i alt
1	109	44	153
2	120	36	156
3	121	36	157
4	133	39	172
5	114	39	153
6	114	39	153
7	98	50	148
8	124	29	153
9	122	33	155
10	109	38	147
11	112	52	164
12	114	35	149
13	132	38	170
14	116	58	174
15	132	38	170
16	112	48	160
17	137	42	179
18	111	56	167

tal	vindertal	tillægst	i alt
19	127	50	177
20	123	49	172
21	112	39	151
22	126	42	168
23	125	47	172
24	107	44	151
25	127	37	164
26	118	35	153
27	127	49	176
28	121	44	165
29	116	39	155
30	115	52	167
31	111	55	166
32	100	56	156
33	122	49	171
34	120	41	161
35	75	37	112
36	63	19	82
i alt	4165	1534	5699

Besvar ved hjælp af *Excel* følgende spørgsmål:

a) Vis, at det ved test på 5% niveau ikke kan antages, at fordelingen af tallene fra 1 til 36 er den samme for vindertallene og for tillægstallene.

Betragt nu kun tallene fra 1 til 34.

b) Vis at fordelingen af tallene fra 1 til 34 er den samme for vindertallene og for tillægstallene.

c) Undersøg, om tallene fra 1 til 34 udtrækkes lige hyppigt.

## 7 Poissonfordelte data

Én af grundene til at Poissonfordelingen ofte optræder i praksis er Poisson processen, som er en sandsynlighedsteoretisk model, der beskriver hvorledes hændelser indtræffer tilfældigt i for eksempel tid, plan eller rum. Ifølge modellen er antallet af hændelser, der indtræffer i en delmængde af den betragtede mængde, for eksempel i et tidsinterval eller i et område af planen eller rummet, Poissonfordelt.

I Afsnit 7.2 gives en kort beskrivelse af Poisson processen. Desuden omtales nogle få egenskaber ved Poissonfordelingen, som benyttes ved analyse af en statistisk model baseret på denne fordeling. I afsnit 7.1 introduceres de eksempler, der benyttes ved illustrationerne af teorien i dette kapitel. Statistisk analyse af én observationsrække ved hjælp af Poissonfordelingen diskuteres i Afsnit 7.3, mens vi i Afsnit 7.4 som to eksempler på analyse ved hjælp af flere Poissonfordelinger omtaler Poissonmodellen med proportionale parametre og den multiplikative Poissonmodel.

Vi afslutter denne introduktion med en generel bemærkning vedrørende observationsrækker fra *diskrete fordelinger*. Hvis antallet  $n$  i én observationsrække  $x_1, \dots, x_n$  er meget stort, angives observationer gerne - af pladshensyn - på *tabelform*, det vil sige, at man for enhver observeret værdi  $j$  angiver antallet  $a_j$  af  $x$ -er i observationsrækken, der antager værdien  $j$ , altså

$$a_j = \#\{i : x_i = j\}$$

Bemærk, at man ved at angive observationerne på tabelform bevarer information om *hvilke værdier* man har observeret. Derimod kan rækkefølgen af de enkelte observationer  $x_i$ ,  $i = 1, \dots, n$ , naturligvis ikke rekonstrueres ud fra  $a$ -erne; dette er dog uden betydning, idet  $x$ -erne antages at være udfald af uafhængige og identisk fordelte stokastiske variable og nummereringen af de enkelte observationer er derfor uden betydning. Angivelse af diskrete observationer på tabelform kan naturligvis betragtes som en form for gruppering, der i modsætning til den sædvanlige gruppering af kontinuerte data ikke giver anledning til tab af information vedrørende værdierne af de enkelte observationer.

## 7.1 Eksempler

I dette afsnit introduceres de datasæt, der vil blive brugt til at illustrere statistisk analyse i modeller baseret på Poissonfordelingen.

### Eksempel 7.1

Tabellen nedenfor viser - på tabelform - fordelingen af mål i de 198 kampe i Faxe Kondi Ligaen 1999-2000 delt op efter første, anden og tredje tredjedel af turneringen. Fordelingerne er vist i Figur 7.1.

<i>antal mål</i>	<i>kamp 1-66</i>	<i>kamp 67-132</i>	<i>kamp 133-198</i>
<i>j</i>	<i>a<sub>j</sub></i>	<i>a<sub>j</sub></i>	<i>a<sub>j</sub></i>
0	3	4	3
1	12	8	7
2	18	22	20
3	13	10	13
4	15	10	9
5	2	11	6
6	3	0	3
7	0	0	1
8	0	0	2
9	0	1	1
10	0	0	1
<i>i alt</i>	66	66	66

Vi ønsker her dels at beskrive fordelingen af mål i de tre dele af turneringen og dels at undersøge om de tre fordelinger kan antages at være identiske, det vil sige vi ønsker at undersøge om fordelingen af mål kan antages at være den samme i de tre dele af turneringen.

□

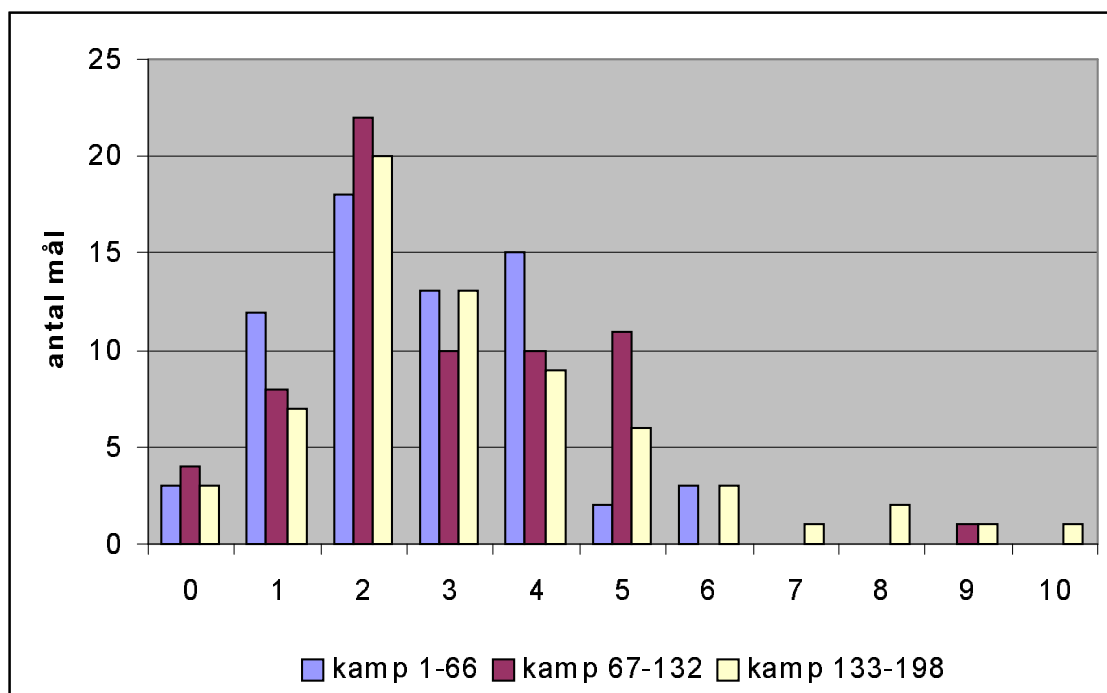
### Eksempel 7.2

Ved de olympiske lege i Sidney 2000 var de nordiske lande medaljehøst: Danmark 6, Finland 4, Norge 10, Sverige 12. Vi ønsker at belyse spørgsmålet om der er forskel på landenes medaljehøst, eventuelt også i lyset af antal indbyggere i landene. Indbyggerantallene er (i millioner): Danmark 5.3, Finland 5.2, Norge 4.5, Sverige 8.9.

□

### Eksempel 7.3

I tabellen nedenfor ses medaljefordelingen for de seks nationer, der fik flest medaljer ved de



**Figur 7.1** Fordelingen af mål i første, anden og tredje tredjedel af Faxe Kondi Ligaen 1999-2000.

olympiske lege i Sydney 2000, se også Figur 7.2.

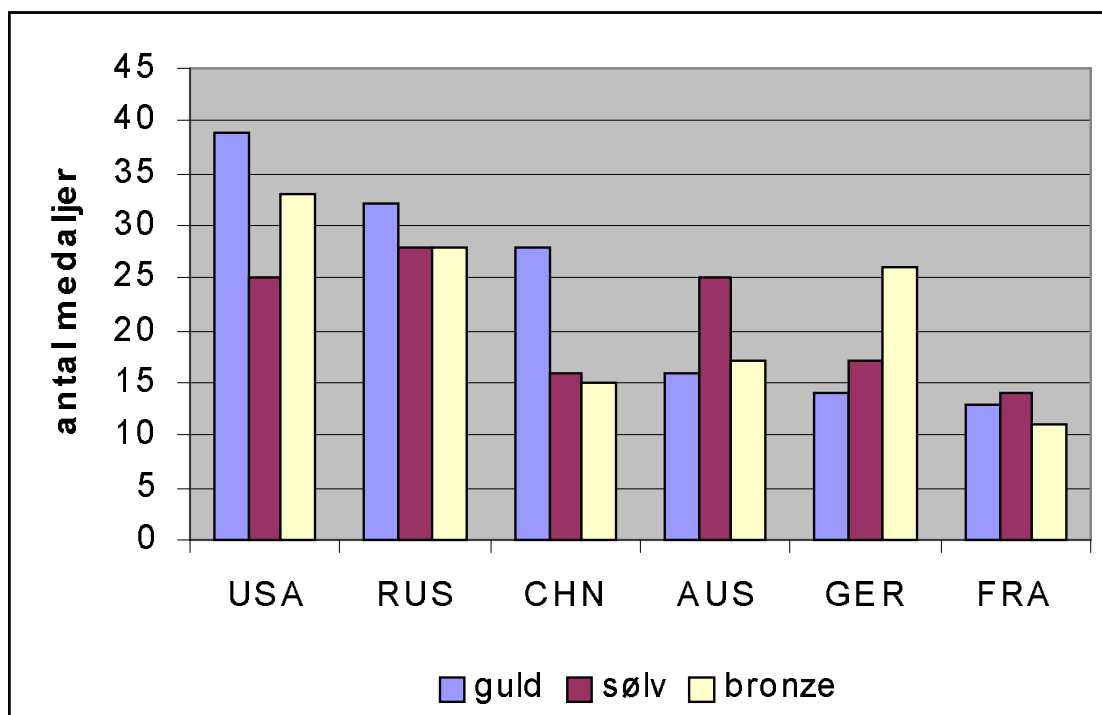
<i>land</i>	<i>guld</i>	<i>sølv</i>	<i>bronze</i>
USA	39	25	33
RUS	32	28	28
CHN	28	16	15
AUS	16	25	17
GER	14	17	26
FRA	13	14	11

Vi ønsker blandt andet at undersøge om medaljernes karat afhænger af de seks nationer.

□

## 7.2 Sandsynlighedsteoretiske resultater vedrørende Poissonfordelingen

Poissonfordelingen er omtalt i Afsnit 3.2.3 og i dette afsnit resumeres de sandsynlighedsteoretiske resultater vedrørende Poissonfordelingen, som benyttes i diskussionen af statistik analyse



**Figur 7.2** Medaljefordelingen for de seks nationer, der fik flest medaljer ved de olympiske lege i Sydney 2000.

i modeller baseret på denne fordeling. Desuden gives en kort introduktion af Poisson processen, der er en matematisk model til beskrivelse af, hvorledes hændelser indtræffer tilfældigt i blandt andet tid, plan og rum. Endelig nævnes i Sætning 7.1 et matematisk resultat, som vil blive benyttet flere gange i dette kapitel.

En diskret stokastisk variabel  $X$  er Poissonfordelt med parameter  $\lambda > 0$ , kort  $X \sim po(\lambda)$ , hvis sandsynlighedsfunktionen (tæthedsfunktionen) for  $X$  er

$$po(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \dots \quad (7.1)$$

Hvis  $X \sim po(\lambda)$  er middelværdien og variansen for  $X$

$$EX = \lambda \quad (7.2)$$

$$Var X = \lambda, \quad (7.3)$$

og dermed er dispersionsindekset (eller dispersionskoefficienten)

$$cd = \frac{Var X}{EX} = 1. \quad (7.4)$$

På dette punkt adskiller Poissonfordelingen sig fra andre diskrete fordelinger. For eksempel fås det af resultaterne i Afsnit 3.2.1 at dispersionsindekset for binomialfordelingen  $b(n, \pi)$  er lig

med  $1 - \pi$  og derfor mindre end 1, mens det for den negative binomialfordeling  $b^-(\kappa, \pi)$  er  $(1 - \pi)^{-1}$  og derfor større end 1, se Afsnit 3.2.5.

Følgende resultat forbinder sandsynlighedsfunktionen for  $b(n, \pi)$  fordelingen med sandsynlighedsfunktionen for Poissonfordelingen:

$$b(x; n, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \rightarrow e^{-\lambda} \frac{\lambda^x}{x!} = po(x; \lambda), \quad \text{for } n \rightarrow \infty \text{ og } \pi \rightarrow 0 \text{ så } n\pi \rightarrow \lambda. \quad (7.5)$$

Resultatet benyttes i modelovervejelser til at skifte fra en model baseret på binomialfordelingen til en model baseret på Poissonfordelingen.

Halesandsynligheder i Poissonfordelingen kan approksimeres med halesandsynligheder i normalfordelingen med samme middelværdi og varians. Hvis  $X \sim po(\lambda)$  gælder

$$\lim_{\lambda \rightarrow \infty} P\left(a \leq \frac{X - \lambda}{\sqrt{\lambda}} \leq b\right) = \Phi(b) - \Phi(a). \quad (7.6)$$

Bemærk, at vi approksimerer sandsynligheder i en diskret fordeling, Poissonfordelingen, med sandsynligheder i en kontinuert fordeling, normalfordelingen. Vi skriver kort

$$X \sim po(\lambda) \quad \text{og } \lambda \text{ stor} \quad \Rightarrow \quad X \approx N(\lambda, \lambda). \quad (7.7)$$

Mange approksimative resultater i dette afsnit kan forstås ved at tænke på at man regner i den approksimerende normalfordeling til Poissonfordelingen. I praksis kan man anvende den approksimerende normalfordeling for  $\lambda > 5$ .

Antag, at  $X_1, \dots, X_i, \dots, X_n$  er uafhængige stokastiske variable, således at  $X_i \sim po(\lambda_i)$ ,  $i = 1, \dots, n$ . Lad  $X$ . betegne summen af de variable, det vil sige  $X = X_1 + \dots + X_i + \dots + X_n$ , og lad tilsvarende  $\lambda$ . betegne summen af parametrene,  $\lambda = \lambda_1 + \dots + \lambda_i + \dots + \lambda_n$ . Da gælder følgende resultater for fordelingen af summen og for den betingede fordeling af de variable givet summen:

$$X. \sim po(\lambda.) \quad (7.8)$$

og

$$(X_1, \dots, X_i, \dots, X_n) \mid X. = x. \sim m(x., \frac{\lambda_1}{\lambda.}, \dots, \frac{\lambda_i}{\lambda.}, \dots, \frac{\lambda_n}{\lambda.}). \quad (7.9)$$

Betingningsresultatet i (7.9) er nøglen til at forstå mange ligheder mellem tests i multinomialfordelingen og i Poissonfordelingen.

Vi giver nu en ganske kort beskrivelse af *Poisson processen*, der er en af grundene til, at Poissonfordelingen ofte optræder i praksis. Poisson processen er en sandsynlighedsteoretisk model for, hvorledes hændelser indtræffer tilfældigt. Antag, at vi betragter hændelser i en delmængde  $S$  af den reelle akse, planen eller rummet, for eksempel tidspunkter for registreringer på en Geiger-tæller, nedslagssteder for meteoritter, positioner af bakteriekolonier på

en agarplade, fangstpositioner for fisk, positioner for indsamlede sten *etc.* Lad  $N(A)$  betegne antallet af hændelser i mængden  $A \subseteq S$ . Antag, at de følgende tre forudsætninger er opfyldt:

- a) Sandsynligheden, for at der indtræffer præcis  $n$  hændelser i  $A$ , afhænger kun af  $|A|$  ( $A$ 's længde, areal eller rumfang), det vil sige, at  $P(N(A) = n)$  afhænger kun af  $|A|$  og  $n$ .
- b) Antallet af hændelser i disjunkte mængder er uafhængige, det vil sige, at  $N(A)$  og  $N(B)$  er uafhængige stokastiske variable, hvis mængderne  $A$  og  $B$  er disjunkte, det vil sige

$$P(N(A) = n, N(B) = m) = P(N(A) = n)P(N(B) = m), \quad \text{hvis } A \cap B = \emptyset.$$

- c) Sandsynligheden, for at der indtræffer mere end én hændelse i  $A$ , er lille, hvis  $|A|$  er lille, eller mere præcist

$$\frac{P(N(A) \geq 2)}{|A|} \rightarrow 0, \quad \text{for } |A| \rightarrow 0.$$

Det kan da vises, at der eksisterer et  $\lambda > 0$ , så det for alle delmængder  $A$  af  $S$  gælder, at antallet af hændelser i  $A$  er Poissonfordelt med parameter  $\lambda |A|$ , altså

$$N(A) \sim po(\lambda |A|). \quad (7.10)$$

Parameteren  $\lambda$  omtales som *intensiteten* af Poisson processen på  $S$ .

Ved hjælp af formlerne (7.8) og (7.9) samt betingelse b) kan det vises, at hvis

$$A = \bigcup_{i=1}^k A_i, \quad \text{hvor } A_i \cap A_j = \emptyset \text{ hvis } i \neq j,$$

da er

$$(N(A_1), \dots, N(A_k)) | N(A) = n \sim m\left(n, \frac{|A_1|}{|A|}, \dots, \frac{|A_k|}{|A|}\right);$$

med andre ord, givet at der indtræffer  $n$  hændelser i  $A$ , er antallene af hændelser i de disjunkte delmængder  $A_1, \dots, A_k$  (som tilsammen udgør  $A$ ) multinomialfordelt med antalsparameter  $n$  og en sandsynlighedsvektor, der angiver, hvor stor en del de enkelte delmængder  $A_1, \dots, A_k$  udgør af  $A$ .

I det følgende skal vi flere gange bruge et matematisk resultat, der er formuleret nedenfor i Sætning 7.1.

**Sætning 7.1** Antag, at  $x > 0$  og  $c > 0$ . Da antager funktionen

$$\begin{aligned} g : ]0, \infty[ &\rightarrow \mathbb{R} \\ \lambda &\rightarrow e^{-c\lambda} \lambda^x \end{aligned}$$

sin maksimale værdi i punktet

$$\hat{\lambda} = \frac{x}{c}.$$



### 7.3 Én observationsrække

I dette afsnit betragter vi én observationsrække fra Poissonfordelingen. Vi antager altså, at observationerne  $x_1, \dots, x_n$  kan betragtes som udfald af uafhængige stokastiske variable  $X_1, \dots, X_n$ , som alle er Poissonfordelte med parameter  $\lambda$ , det vil sige, at vi betragter modellen

$$M_0 : X_i \sim po(\lambda), \quad i = 1, \dots, n. \quad (7.11)$$

#### Estimation

Likelihood funktionen for  $\lambda$  er

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda} \lambda^x \prod_{i=1}^n \frac{1}{x_i!}, \end{aligned}$$

hvor  $x = x_1 + \dots + x_n$ . Heraf finder vi ved hjælp af Sætning 7.1, at maksimum likelihood estimatet  $\hat{\lambda}$  for  $\lambda$  er

$$\hat{\lambda} = \bar{x} = \frac{1}{n}x = \frac{1}{n} \sum_{i=1}^n x_i. \quad (7.12)$$

Parameteren  $\lambda$ , som ifølge (7.2) er middelværdien i  $po(\lambda)$ -fordelingen, estimeres altså ved den empiriske middelværdi. Af (7.8) ses, at der gælder følgende resultat vedrørende fordelingen af maksimum likelihood estimatoren:

$$n\hat{\lambda} = X \sim po(n\lambda). \quad (7.13)$$

Inferens vedrørende værdien af  $\lambda$ , for eksempel test af hypotesen  $\lambda = \lambda_0$ , kan foretages ved at betragte fordelingen af  $X$ .

#### Modelkontrol

Modellen  $M_0$  kan kontrolleres ved et  $\chi^2$ -test for goodness of fit, som beskrevet i Afsnit 6.5, hvis stikprøvestørrelsen  $n$  er tilstrækkelig stor.

En alternativ kontrol af modellen  $M_0$  baserer sig på, at dispersionsindekset for Poissonfordelingen er 1, se formel (7.4). Det må derfor forventes, at forholdet

$$t = \frac{s^2}{\bar{x}} \quad (7.14)$$

mellem den empiriske varians  $s^2$  og den empiriske middelværdi  $\bar{x}$  er tæt på 1. For store værdier af  $\lambda$  eller for store værdier af  $n$  gælder følgende approksimation for fordelingen af den til  $t$  svarende stokastiske variabel

$$t \sim \chi^2(n-1)/(n-1).$$

Dette udsagn læses ” $t$  er en realisation af en stokastisk variabel hvis fordeling kan approksimeres med  $\chi^2(n-1)/(n-1)$ -fordelingen”. Approksimationen kan benyttes, hvis  $n \geq 15$  eller  $\bar{x} \geq 5$ . Resultat benyttes ofte til test af  $M_0$ , idet modellen *accepteres* ved et test på niveau  $\alpha$ , hvis

$$\chi_{\alpha/2}^2(n-1)/(n-1) \leq t \leq \chi_{1-\alpha/2}^2(n-1)/(n-1), \quad (7.15)$$

Testet i (7.15) omtales som *Fishers dispersionsindeks* for Poissonfordelingen.

Hvis Poissonmodellen  $M_0$  forkastes, fordi den observerede værdi af  $t$  er for stor, kan man på grund af bemærkningen efter formel (7.4) forsøge at beskrive observationsrækken ved hjælp af en model baseret på den negative binomialfordeling. Hvis  $M_0$  forkastes på grund af en for lille værdi af  $t$ , peger bemærkningen på en binomialmodel, hvis rimelighed dog som regel checkes ved at undersøge om betingelserne a) - d) side 6.1 er opfyldt for  $k = 2$ .

Beregningen af den empiriske middelværdi og varians, ved hjælp af hvilke dispersionsindekset er defineret, afhænger af, om alle de enkelte observationer er til rådighed eller om observationerne er givet på tabelform. Med indlysende betegnelser har vi

$$S = \sum_{i=1}^n x_i = \sum_j j a_j,$$

$$SK = \sum_{i=1}^n x_i^2 = \sum_j j^2 a_j$$

og

$$\bar{x} = \frac{1}{n}S \quad \text{og} \quad s^2 = \frac{1}{n-1}(SK - \frac{S^2}{n}).$$

I det næste afsnit omtales endnu to approksimative test for modellen  $M_0$ , nemlig  $-2 \ln Q$ -testet i formel (7.38) og det hermed ækvivalente  $X^2$ -test, som vises at være beslægtet med testet baseret på dispersionsindekset, jævnfør formel (7.39) og bemærkningerne derefter.

### Eksempel 7.1 (Fortsat)

For fordelingen af målene i de tre dele af turneringen har vi følgende beregninger, med fire decimalers nøjagtighed:

	$n$	$S$	$SK$	$\bar{x}(\hat{\lambda})$	$s^2$	$t$
<i>kamp 1-66</i>	66	175	599	2.6515	2.0767	0.7832
<i>kamp 67-132</i>	66	186	702	2.8182	2.7357	0.9707
<i>kamp 133-198</i>	66	212	964	3.2121	4.3543	1.3556

I alle tre tilfælde er antallet  $n = 66$  af observationer stort nok til at Fishers dispersionsindeks  $t$  kan benyttes. De observerede værdier af  $t$  skal vurderes i en  $\chi^2(f)/f$ -fordeling med  $f = 65$ . På side 11 i *Statistical Tables* ses, at 2.5% og 97.5% fraktilen i denne fordeling er henholdsvis 0.6862 og 1.3720. Ifølge (7.15) accepteres modellen  $M_0$  derfor ved et test på niveau 5% i alle tre tilfælde.

Antallet af observationer  $n = 66$  er tilpas stort til at vi i denne situation også kan benytte test for goodness of fit til kontrol af  $M_0$ . De forventede antal under  $M_0$ , der beregnes som

$$e_j = ne^{-\hat{\lambda}} \hat{\lambda}^j / j!, \quad j = 0, 1, \dots, 10,$$

er

antal mål $j$	kamp 1-66		kamp 67-132		kamp 133-198	
	$a_j$	$e_j$	$a_j$	$e_j$	$a_j$	$e_j$
0	3	4.6559	4	3.9412	3	2.6579
1	12	12.3452	8	11.1069	7	8.5375
2	18	16.3668	22	15.6506	20	13.7117
3	13	14.4656	10	14.7021	13	14.6812
4	15	9.5889	10	10.3583	9	11.7895
5	2	5.0850	11	5.8383	6	7.5738
6	3	2.2472	0	2.7422	3	4.0547
7	0	0.8512	0	1.1040	1	1.8606
8	0	0.2821	0	0.3889	2	0.7471
9	0	0.0831	1	0.1218	1	0.2666
10	0	0.0220	0	0.0343	1	0.0856
<i>i alt</i>	66	65.9930	66	65.9886	66	65.9662

For at imødekomme kravet om at de forventede antal skal være større end eller lig med 5 er det i alle tre tilfælde nødvendigt at slå grupperne 0 og 1 sammen til en gruppe (0-1) samt at slå grupperne 5, 6, 7, 8, 9, 10 sammen til en gruppe ( $\geq 5$ ). Idet

$$e_{\geq 5} = n(1 - \sum_{j=0}^4 e^{-\hat{\lambda}} \hat{\lambda}^j / j!) = n - \sum_{j=0}^4 e_j,$$

får vi

<i>antal mål</i> <i>j</i>	<i>kamp 1-66</i>		<i>kamp 67-132</i>		<i>kamp 133-198</i>	
	<i>a<sub>j</sub></i>	<i>e<sub>j</sub></i>	<i>a<sub>j</sub></i>	<i>e<sub>j</sub></i>	<i>a<sub>j</sub></i>	<i>e<sub>j</sub></i>
0-1	15	17.0012	12	15.0480	10	11.1954
2	18	16.3668	22	15.6506	20	13.7117
3	13	14.4656	10	14.7021	13	14.6812
4	15	9.5889	10	10.3583	9	11.7895
≥5	5	8.5775	12	10.2410	14	14.6223
<i>i alt</i>	66	66.0000	66	66.0000	66	66.0000

I alle tre tilfælde gælder, at antallet af grupper er  $k = 5$  og modellen  $M_0$  har én fri parameter  $\lambda$ , så antallet af frihedsgrader i testet for goodness of fit bliver  $f = k - 1 - 1 = 3$ . Vi finder, at

$$\begin{aligned} -2 \ln Q_1 &= 4.9160 & \varepsilon_1 &= 1 - F_{\chi^2(3)}(4.9160) = 0.1781 \\ -2 \ln Q_2 &= 4.9434 & \varepsilon_2 &= 1 - F_{\chi^2(3)}(4.9434) = 0.1760 \\ -2 \ln Q_3 &= 3.6016 & \varepsilon_3 &= 1 - F_{\chi^2(3)}(3.6016) = 0.3078. \end{aligned}$$

Testet for goodness of fit giver derfor heller ingen anledning til at betvivle modellen  $M_0$  i nogen af tilfældene.  $\square$

### Konfidensinterval

Vi starter med at give formelen for et approksimativt  $1 - \alpha$  konfidensinterval for middelværdien  $\lambda$  baseret på én observation  $x$  fra en  $po(\lambda)$  fordelt stokastisk variabel  $X$ . Konfidensintervallet er approksimativt, fordi det bygger på den approksimerende  $N(\lambda, \lambda)$  fordeling, jævnfør (7.7). I denne holder uligheden

$$-u_{1-\alpha/2} < \frac{X - \lambda_0}{\sqrt{\lambda_0}} < u_{1-\alpha/2} \quad (7.16)$$

med sandsynlighed  $1 - \alpha$ . Løses uligheden (7.16) med hensyn til  $\lambda_0$  fås den ækvivalente ulighed

$$X + \frac{1}{2}u_{1-\alpha/2}^2 - u_{1-\alpha/2}\sqrt{X + \frac{1}{4}u_{1-\alpha/2}^2} < \lambda_0 < X + \frac{1}{2}u_{1-\alpha/2}^2 + u_{1-\alpha/2}\sqrt{X + \frac{1}{4}u_{1-\alpha/2}^2}, \quad (7.17)$$

som også holder med sandsynlighed  $1 - \alpha$ . Indsættes den aktuelle observation i (7.17) fås  $1 - \alpha$  konfidensintervallet for middelværdien i en Poissonfordeling som

$$C_{1-\alpha}(x) = [\lambda_-, \lambda_+], \quad (7.18)$$

hvor

$$\lambda_- = x + \frac{1}{2}u_{1-\alpha/2}^2 - u_{1-\alpha/2}\sqrt{x + \frac{1}{4}u_{1-\alpha/2}^2}, \quad (7.19)$$

og

$$\lambda_+ = x + \frac{1}{2}u_{1-\alpha/2}^2 + u_{1-\alpha/2} \sqrt{x + \frac{1}{4}u_{1-\alpha/2}^2}. \quad (7.20)$$

Bemærk, at formlen (7.17) understreger, at det er grænserne for konfidensintervallet, der er stokastiske, og at fortolkningen af et konfidensinterval baseret på observationen  $x$  er, at enten er  $\lambda_0$  i konfidensintervallet, eller der er indtruffet en hændelse med sandsynlighed mindre end  $\alpha$ .

Da desuden  $(X - \lambda_0)/\sqrt{\lambda_0}$  er testor for hypotesen  $H_0: \lambda = \lambda_0$  har konfidensintervallet ifølge (7.16) også fortolkningen som de værdier af parameteren, som ikke vil blive forkastet som nulhypotese på grundlag af observationen  $x$ .

Undertiden er man interesseret i at beregne et konfidensinterval for en parameter  $\lambda$ , i situationer hvor den Poissonfordelte stokastiske variabel  $X$  har middelværdi  $c\lambda$ , hvor  $c$  betegner en kendt konstant. I de tilfælde beregnes konfidensintervallet for middelværdien  $c\lambda$  efter formlerne (7.19) og (7.20), og det transformeres til et konfidensinterval for  $\lambda$ .

Det første eksempel på den situation er netop én observationsrække, hvor  $x. \sim \text{po}(n\lambda)$  og (7.19) og (7.20) er grænserne for  $1 - \alpha$  konfidensintervallet for  $n\lambda$ , som transformeres til et konfidensinterval for  $\lambda$  med grænserne

$$\frac{1}{n}(n\lambda)_- = \frac{1}{n} \left[ x. + \frac{1}{2}u_{1-\alpha/2}^2 - u_{1-\alpha/2} \sqrt{x. + \frac{1}{4}u_{1-\alpha/2}^2} \right], \quad (7.21)$$

og

$$\frac{1}{n}(n\lambda)_+ = \frac{1}{n} \left[ x. + \frac{1}{2}u_{1-\alpha/2}^2 + u_{1-\alpha/2} \sqrt{x. + \frac{1}{4}u_{1-\alpha/2}^2} \right]. \quad (7.22)$$

### Eksempel 7.1 (Fortsat)

Ved hjælp af (7.21) og (7.22) beregnes 95% konfidensintervallet for  $\lambda$ , middelværdien af antal scorede mål i én kamp til:

	$n$	$x.(S)$	$\bar{x}(\hat{\lambda})$	$\lambda_-$	$\lambda_+$
<i>kamp 1-66</i>	66	175	2.6515	2.2867	3.0745
<i>kamp 67-132</i>	66	186	2.8182	2.4412	3.2533
<i>kamp 133-198</i>	66	212	3.2121	2.8079	3.6746

□

## 7.4 Inferens i flere fordelinger

I dette afsnit giver vi et par eksempler på statistisk analyse af modeller, der involverer flere Poissonfordelinger. Desuden vises det, at der på grund af resultatet i formel (7.9) er en intim

forbindelse mellem analyse af sådanne modeller og analysen af modeller baseret på multinomialfordelingen.

#### 7.4.1 Poissonmodellen med proportionale parametre

Udgangspunktet for den følgende diskussion er, at datasættet  $\mathbf{x}$  består af observationerne  $x_1, \dots, x_k$ , der kan betragtes som udfald af uafhængige stokastiske variable  $X_1, \dots, X_k$ , som alle er Poissonfordelt men med hver sin parameter, det vil sige, at grundmodellen er

$$M_0 : X_i \sim po(\lambda_i), \quad i = 1, \dots, k. \quad (7.23)$$

Antag, at  $m_1, \dots, m_k$  er kendte tal, og at vi er interesseret i at teste hypotesen, om at parametrene i modellen  $M_0$  er proportionale med  $m_1, \dots, m_k$  som proportionalitetsfaktorer, det vil sige hypotesen

$$H_{01} : \lambda_i = m_i \lambda, \quad i = 1, \dots, k. \quad (7.24)$$

Den tilsvarende model

$$M_1 : X_i \sim po(m_i \lambda), \quad i = 1, \dots, k. \quad (7.25)$$

har én fri parameter  $\lambda$ .

Bemærk, at man i modellen  $M_0$  kan undersøge, om  $x_1, \dots, x_k$  kan betragtes som én observationsrække fra  $po(\lambda)$ -fordelingen, ved at teste hypotesen svarende til at  $m_1 = \dots = m_k = 1$ .

Likelihood funktionen under  $M_0$  er

$$\begin{aligned} L(\lambda_1, \dots, \lambda_k) &= \prod_{i=1}^k e^{-\lambda_i} \frac{\lambda_i^{x_i}}{x_i!} \\ &= e^{-\lambda} \cdot \prod_{i=1}^k \lambda_i^{x_i} \prod_{i=1}^k \frac{1}{x_i!}, \end{aligned}$$

hvor  $\lambda = \lambda_1 + \dots + \lambda_k$ . Log likelihood funktionen under  $M_0$  bliver derfor

$$l(\lambda_1, \dots, \lambda_k) = -\lambda + \sum_{i=1}^k x_i \ln(\lambda_i) - \sum_{i=1}^k \ln(x_i!). \quad (7.26)$$

Da parametrene  $\lambda_1, \dots, \lambda_k$  er variationsuafhængige, ses det af det første udtryk for likelihood funktionen ved hjælp af Sætning 7.1, at maksimum likelihood estimatet for  $\lambda_i$  under  $M_0$  er

$$\hat{\lambda}_i = x_i, \quad i = 1, \dots, k.$$

Log likelihood funktionen for  $\lambda$  under  $H_{01}$  fås ved i (7.26) at erstatte  $\lambda_i$  med  $m_i \lambda$ . Vi finder

$$\begin{aligned} l(\lambda) &= -\sum_{i=1}^k \lambda m_i + \sum_{i=1}^k x_i \ln(m_i \lambda) - \sum_{i=1}^k \ln(x_i!) \\ &= -\lambda m + x \ln(\lambda) + \sum_{i=1}^k x_i \ln(m_i) - \sum_{i=1}^k \ln(x_i!), \end{aligned} \quad (7.27)$$

hvor  $m. = m_1 + \dots + m_k$ . Likelihood ligningen for  $\lambda$  bliver derfor

$$\frac{dl}{d\lambda} = -m. + \frac{x.}{\lambda} = 0,$$

som har løsning  $\lambda = \frac{x.}{m.}$ . Det ses, at hvis  $x. > 0$  er maksimum likelihood estimatet for  $\lambda$  under  $M_1$

$$\hat{\lambda} = \frac{x.}{m.} \quad (7.28)$$

Det forventede antal svarende til observationen  $x_i$  - det vil sige middelværdien af  $X_i$  beregnet under sandsynlighedsmålet svarende til  $\hat{\lambda}$  - er derfor

$$e_i = m_i \hat{\lambda} = x. \frac{m_i}{m.} \quad (7.29)$$

Af formlerne (7.26) - (7.29) ses det, at  $-2 \ln Q$ -teststørrelsen for  $H_{01}$  er

$$\begin{aligned} -2 \ln Q(\mathbf{x}) &= 2[l(\hat{\lambda}_1, \dots, \hat{\lambda}_k) - l(\hat{\lambda})] \\ &= 2\left[\sum_{i=1}^k x_i \ln(x_i) - \sum_{i=1}^k x_i \ln(m_i \hat{\lambda})\right] \\ &= 2 \sum_{i=1}^k x_i \ln\left(\frac{x_i}{e_i}\right). \end{aligned} \quad (7.30)$$

Antallet af frihedsgrader i den  $\chi^2$ -fordeling, der approksimerer fordelingen af  $-2 \ln Q$  under  $H_{01}$ , er  $k - 1$ , idet der er  $k$  frie parametre i  $M_0$  og én fri parameter i  $M_1$ . Hvis *de forventede antal alle er større end eller lig med 5*, har vi følgende approksimation af testsandsynligheden for  $H_{01}$  :

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2(k-1)}(-2 \ln Q(\mathbf{x})). \quad (7.31)$$

Hypotesen  $H_{01}$  kan også testes ved hjælp af  $X^2$ -teststørrelsen, som er

$$X^2(\mathbf{x}) = \sum_{i=1}^k \frac{(x_i - e_i)^2}{e_i}. \quad (7.32)$$

Den tilsvarende approksimation af testsandsynligheden er

$$\varepsilon^*(\mathbf{x}) \doteq 1 - F_{\chi^2(k-1)}(X^2(\mathbf{x})). \quad (7.33)$$

Fordelingen af maksimum likelihood estimatoren for  $\lambda$  under  $H_{01}$  angives som regel på følgende måde:

$$m. \hat{\lambda} = X. \sim po(m. \lambda). \quad (7.34)$$

Erstattes  $x$  med  $x.$  i (7.19) og (7.20) fås grænserne for  $1 - \alpha$  konfidensintervallet for  $m. \lambda$ . Det kan transformeres til et konfidensinterval for  $\lambda$  med grænserne

$$\frac{1}{m.} (m. \lambda)_- = \frac{1}{m.} \left[ x. + \frac{1}{2} u_{1-\alpha/2}^2 - u_{1-\alpha/2} \sqrt{x. + \frac{1}{4} u_{1-\alpha/2}^2} \right], \quad (7.35)$$

og

$$\frac{1}{m_{\cdot}}(m_{\cdot}\lambda)_{+} = \frac{1}{m_{\cdot}} \left[ x_{\cdot} + \frac{1}{2}u_{1-\alpha/2}^2 + u_{1-\alpha/2} \sqrt{x_{\cdot} + \frac{1}{4}u_{1-\alpha/2}^2} \right]. \quad (7.36)$$

Hvis  $m_1 = \dots = m_k = 1$  svarer modellen  $M_1$  til én observationsrække fra Poissonfordelingen. Denne model blev betegnet med  $M_0$  i Afsnit 7.3. I dette tilfælde er

$$\hat{\lambda} = \frac{x_{\cdot}}{k} = \bar{x}_{\cdot},$$

så de forventede antal er ens, idet

$$e_i = \bar{x}_{\cdot}, \quad i = 1, \dots, k, \quad (7.37)$$

og formel (7.30) kan derfor reduceres til

$$-2 \ln Q(\mathbf{x}) = 2 \left[ \sum_{i=1}^k x_i \ln(x_i) - x_{\cdot} \ln(\bar{x}_{\cdot}) \right]. \quad (7.38)$$

Yderligere gælder der i denne situation, at

$$X^2(\mathbf{x}) = \sum_{i=1}^k \frac{(x_i - \bar{x}_{\cdot})^2}{\bar{x}_{\cdot}} = \frac{(k-1)s^2}{\bar{x}_{\cdot}} = (k-1)t, \quad (7.39)$$

hvor  $t$  er Fishers dispersionsindeks, som defineret i formel (7.14). Der er altså en sammenhæng mellem Fishers dispersionsindeks  $t$  og  $X^2(\mathbf{x})$ , men det er bemærkelsesværdigt, at mens  $t$  forkaster for både store og små værdier, så forkaster  $X^2$  kun for store værdier af  $X^2(\mathbf{x})$ . Forklaringen er, at Fishers dispersionsindeks og  $X^2$  er udledt i forskellige modeller og tester forskellige hypoteser. For  $X^2$  er modellen, at observationerne er uafhængige og Poissonfordelt, men ikke nødvendigvis identisk fordelt, og i den model testes netop nulhypotesen, at observationerne er identisk fordelt. Fishers dispersionsindeks udledes derimod i en model, hvor observationerne er uafhængige og identisk fordelt, men iøvrigt har en uspecificeret fordeling. Her betragtes nulhypotesen, at den fælles fordeling er Poissonfordelingen.

En illustration af brugen af  $-2 \ln Q$ -testet til kontrol af modellen, der svarer til én observationsrække fra Poissonfordelingen, bliver givet i fortsættelsen af Eksempel 7.2 nedenfor.

#### Relation til multinomialmodellen

Der er en tæt forbindelse mellem test i Poissonmodellen og test i multinomialmodellen. For de to test i (7.30) og (7.32) kan dette forklares ved hjælp af formel (7.9). Betingelser vi i modellen  $M_0$  med summen af observationerne  $x_{\cdot}$ , får vi ifølge (7.9) den betingede model

$$\tilde{M}_0 : (X_1, \dots, X_k) \mid X_{\cdot} = x_{\cdot} \sim m(x_{\cdot}, (\pi_1, \dots, \pi_k)), \quad (7.40)$$

hvor

$$(\pi_1, \dots, \pi_k) = \left( \frac{\lambda_1}{\lambda_{\cdot}}, \dots, \frac{\lambda_k}{\lambda_{\cdot}} \right).$$

Da  $\lambda$ -erne varierer frit, er der heller ingen bånd på variationen af sandsynlighedsvektoren  $(\pi_1, \dots, \pi_k)$  i (7.40); med andre ord er den betingede model i (7.40) grundmodellen for en multinomialfordeling med  $k$  kategorier og med antalsparameter  $x$ . Hypotesen  $H_{01}$  svarer i denne model til den simple hypotese

$$\tilde{H}_{01} : (\pi_1, \dots, \pi_k) = \left( \frac{m_1}{m}, \dots, \frac{m_k}{m} \right).$$

Af (7.29) ses, at de forventede antal under hypotesen  $H_{01}$  i modellen  $M_0$  er præcis de samme som de forventede antal under hypotesen  $\tilde{H}_{01}$  i modellen  $\tilde{M}_0$ , og dermed er også  $-2 \ln Q$ -testene (eller  $X^2$ -testene) identiske.

Selvom beregningerne i de to modeller er identiske er modellerne forskellige. Forskellen mellem modellerne består i den måde, hvorpå data er indsamlet. I multinomialmodellen har man på forhånd lagt sig fast på at betragte observationer med en given sum, som angives ved antalsparameteren sædvanligvis betegnet med  $n$ ; men i modellen  $\tilde{M}_0$  betegnet med  $x$ . I Poissonmodellen derimod har man ikke på forhånd lagt restriktioner på summen af observationerne.

### Eksempel 7.2 (Fortsat)

For at undersøge om de fire nordiske landes medaljehøst ved de olympiske lege i Sydney 2000 kan antages at være ens, når vi ikke tager hensyn til landenes befolkningstal, betragter vi modellen

$$M_0 : X_i \sim po(\lambda_i), \quad i = 1, 2, 3, 4$$

$X_1, X_2, X_3$  og  $X_4$  er stokastiske uafhængige

og tester i denne hypotesen

$$H_{01} : \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 (= \lambda),$$

som er af formen (7.24) med  $m_1 = m_2 = m_3 = m_4 = 1$ . Accepteres  $H_{01}$  reduceres modellen  $M_0$  til

$$M_1 : X_i \sim po(\lambda), \quad i = 1, 2, 3, 4$$

$X_1, X_2, X_3$  og  $X_4$  er stokastiske uafhængige.

Ved hjælp af formel (7.28) bliver maksimum likelihood estimatet for  $\lambda$  i modellen  $M_1$

$$\hat{\lambda} = \frac{32}{4} = 8.$$

De forventede antal beregnes derefter som angivet i (7.29). Vi finder

<i>land</i>	$m_i$	$x_i$	$e_i$
<i>Danmark</i>	1	6	8
<i>Finland</i>	1	4	8
<i>Norge</i>	1	10	8
<i>Sverige</i>	1	12	8
<i>i alt</i>	4	32	32

og af (7.30) og (7.31) fås

$$-2\ln Q(\mathbf{x}) = 5.1967$$

og

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(3)}(5.1967) = 0.1579,$$

så hypotesen  $H_{01}$  accepteres.

Ved hjælp af (7.35) og (7.36) finder vi, at 95% konfidensintervallet for  $\lambda$  er:

$$[5.6671, 11.2933].$$

Ønsker vi at undersøge om medaljehøsten per indbygger er den samme i de fire nordiske lande tager vi igen udgangspunkt i modellen  $M_0$ , men nu tester vi hypotesen

$$H_{01}^* : \lambda_i = m_i \lambda, \quad i = 1, 2, 3, 4,$$

som er af formen (7.24) hvor  $m$ -erne er befolkningstallene (i millioner) i de fire lande, det vil sige  $m_1 = 5.3$ ,  $m_2 = 5.2$ ,  $m_3 = 4.5$  og  $m_4 = 8.9$ . Accepteres  $H_{01}^*$  reduceres modellen  $M_0$  til

$$M_1^* : X_i \sim po(m_i \lambda), \quad i = 1, 2, 3, 4$$

$X_1, X_2, X_3$  og  $X_4$  er stokastiske uafhængige.

Ved hjælp af formel (7.28) bliver maksimum likelihood estimatet for  $\lambda$  i modellen  $M_1^*$

$$\hat{\lambda} = \frac{32}{23.9} = 1.3389.$$

De forventede antal beregnes derefter som angivet i (7.29). Vi finder - med fire decimalers nøjagtighed -

<i>land</i>	$m_i$	$x_i$	$e_i$
<i>Danmark</i>	5.3	6	7.0962
<i>Finland</i>	5.2	4	6.9623
<i>Norge</i>	4.5	10	6.0251
<i>Sverige</i>	8.9	12	11.9163
<i>i alt</i>	23.9	32	31.9999

Formlerne (7.30) og (7.31) medfører, at

$$-2\ln Q(\mathbf{x}) = 3.8535$$

og

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(3)}(3.8535) = 0.2777,$$

så hypotesen  $H_{01}^*$  accepteres.

95% konfidensintervallet for  $\lambda$ , som i  $M_1^*$  er middelværdien af medaljer per 1 million indbyggere i de fire nordiske lande, er findes ved hjælp af (7.35) og (7.36) til:

$$[0.9485, 1.8901].$$

Eksemplet viser altså, at der ikke er signifikant forskel på de fire nordiske landes medaljehøst ved de olympiske lege i Sydney 2000 hverken absolut eller når befolkningsantallene tages i betragtning.  $\square$

### Eksempel 7.1 (Fortsat)

For at undersøge om parametrene i de tre Poissonfordelinger - én for hver tredjedel af turneringen - er identiske, benytter vi (7.13) og betragter modellen

$$M_0 : X_i \sim po(66\lambda_i), \quad i = 1, 2, 3$$

$X_1, X_2$  og  $X_3$  er stokastiske uafhængige.

Hypotesen

$$H_{01} : \lambda_1 = \lambda_2 = \lambda_3$$

er derfor af formen som i (7.24) med  $m_i = 66$ ,  $i = 1, 2, 3$ . Af beregningerne i skemaet

	$m_i$	$x_{i\cdot}$	$e_i$
<i>kamp 1-66</i>	66	175	191
<i>kamp 67-132</i>	66	186	191
<i>kamp 133-198</i>	66	212	191
<i>i alt</i>	198	573	573

og formlerne (7.28) - (7.31) finder vi, at

$$\hat{\lambda} = \frac{573}{198} = 2.8939,$$

$$-2\ln Q(\mathbf{x}) = 3.7401$$

og

$$\varepsilon = 1 - F_{\chi^2(2)}(3.7401) = 0.1541.$$

Vi accepterer derfor hypotesen, om at parametrene i de tre Poissonfordelinger er identiske, hvilket her betyder, at der er ikke signifikant forskel på fordelingen af mål i kampe i de tre dele af turneringen.

Under  $H_{01}$  er summen af alle observationerne Poissonfordelt, idet

$$X_{.j} \sim po(m, \lambda_j),$$

Da  $x_{.j} = 573$  og  $m = 198$ , fås ved hjælp af formlerne (7.35) og (7.36) at 95% kondidensintervallet for  $\lambda_j$  - middelværdien af antal scorede mål i en tilfældig kamp i Faxe Kondi Ligaen 1999-2000 - er

$$[\lambda_{-j}, \lambda_{+j}] = [2.6665, 3.1408].$$

□

## 7.4.2 Den multiplikative Poissonmodel

Denne model benyttes i situationer, hvor observationerne - som vist nedenfor - kan opskrives i en  $r \times s$  tabel svarende til to inddelingskriterier med henholdsvis  $r$  og  $s$  kategorier. Observationen svarende til den  $i$ 'te kategori ved det første kriterium og den  $j$ 'te kategori ved det andet kriterium betegnes med  $x_{ij}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, s$ . Strukturen af data er altså den samme som ved en tosidet variansanalyse uden gentagelser, og som det fremgår af det følgende, er der visse lighedspunkter mellem denne model og den multiplikative Poissonmodel. Modellerne er dog meget forskellige. Den førstnævnte er en model for kontinuerte variable, hvor man betragter en hypotese om *additiv* struktur af middelværdierne, mens Poissonmodellen er en model for diskrete data, hvor man - som det ses nedenfor - betragter en hypotese om en *multiplikativ* struktur af middelværdierne.

Vi illustrerer teorien ved hjælp af data i Eksempel 7.3, som er angivet i en  $6 \times 3$  tabel.

	1	...	$j$	...	$s$	$\Sigma$
1	$x_{11}$	...	$x_{1j}$	...	$x_{1s}$	$x_{1\cdot}$
.	.	...	.	...	.	.
.	.	...	.	...	.	.
$i$	$x_{i1}$	...	$x_{ij}$	...	$x_{is}$	$x_{i\cdot}$
.	.	...	.	...	.	.
.	.	...	.	...	.	.
$r$	$x_{r1}$	...	$x_{rj}$	...	$x_{rs}$	$x_{r\cdot}$
$\Sigma$	$x_{\cdot 1}$	...	$x_{\cdot j}$	...	$x_{\cdot s}$	$x_{\cdot\cdot}$

I tabellen betegner  $x_i$  og  $x_j$  henholdsvis summen af observationerne i den  $i$ 'te række og den  $j$ 'te søjle, mens  $x_{\cdot\cdot}$  er summen af alle observationerne, det vil sige

$$x_i = \sum_{j=1}^s x_{ij}, \quad x_j = \sum_{i=1}^r x_{ij}, \quad x_{\cdot\cdot} = \sum_{i=1}^r \sum_{j=1}^s x_{ij}.$$

Forudsætter vi, at alle observationer er udfald af uafhængige stokastiske variable, kan de modeller, vi vil betragte, skrives på følgende måde.

Grundmodellen

$$M_0 : x_{ij} \sim\sim po(\lambda_{ij}), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Den *multiplikative* model eller modellen for *ingen vekselvirkning*

$$M_1 : x_{ij} \sim\sim po(\alpha_i \beta_j), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Modellen med *kun rækkevirkning*

$$M_2 : x_{ij} \sim\sim po(\alpha_i \beta), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Modellen med *kun søjlevirkning*

$$M_2^* : x_{ij} \sim\sim po(\alpha \beta_j), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Modellen for *homogenitet*

$$M_3 : x_{ij} \sim\sim po(\alpha \beta), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

De fire sidstnævnte modeller svarer alle til hypoteser, om hvorledes de to inddelingskriterier påvirker fordelingerne i grundmodellen. Modellen  $M_1$  svarer til hypotesen  $H_{01} : \lambda_{ij} = \alpha_i \beta_j$ , ifølge hvilken de to kriterier virker *uafhængigt* af hinanden. Fortolkningen af modellerne  $M_2$ ,  $M_2^*$  og  $M_3$  i relation til de to inddelingskriterier er indlysende.

### Parametrisering af modellerne

Modellen  $M_1$  har  $r + s - 1$  frie parametre, hvilket dog ikke fremgår af opskrivningen af modellen ovenfor. Der findes adskillige måder at parametrisere  $M_1$  på. Den måde, vi har valgt, er bekvem for teoretiske overvejelser men adskiller sig fra den, som programpakker benytter. Lad  $\alpha$ .,  $\beta$  og  $\lambda$ .. betegne summen af henholdsvis  $\alpha$ -erne,  $\beta$ -erne og  $\lambda$ -erne og lad endvidere

$$\rho_i = \frac{\alpha_i}{\alpha.}, \quad i = 1, \dots, r \quad \text{og} \quad \sigma_j = \frac{\beta_j}{\beta.}, \quad j = 1, \dots, s.$$

Idet

$$\lambda_{..} = \sum_{i=1}^r \sum_{j=1}^s \lambda_{ij} = \sum_{i=1}^r \sum_{j=1}^s \alpha_i \beta_j = \sum_{j=1}^s \alpha_i \sum_{i=1}^r \beta_j = \alpha. \beta.,$$

har vi følgende omskrivning af parameteren under  $M_1$  :

$$\alpha_i \beta_j = \alpha. \beta. \frac{\alpha_i}{\alpha.} \frac{\beta_j}{\beta.} = \lambda_{..} \rho_i \sigma_j.$$

Da  $\boldsymbol{\rho} = (\rho_1, \dots, \rho_r)$  og  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_s)$  er henholdsvis en  $r$ -dimensional og en  $s$ -dimensional sandsynlighedsvektor (komponenterne i de to vektorer er positive og summer sammen til 1), ses det, at antallet af frie parametre i  $M_1$  er  $d_1 = 1 + (r - 1) + (s - 1) = r + s - 1$ .

Med denne parametrisering bliver modellerne  $M_1$ ,  $M_2$ ,  $M_2^*$  og  $M_3$  og deres indbyrdes forhold som angivet i nedenstående skema:

$$\begin{array}{ccc} & M_2 : X_{ij} \sim po(\lambda_{..} \rho_i / s) & \\ \nearrow & & \searrow \\ M_1 : X_{ij} \sim po(\lambda_{..} \rho_i \sigma_j) & & M_3 : X_{ij} \sim po(\lambda_{..} / (rs)) \\ \searrow & & \nearrow \\ & M_2^* : X_{ij} \sim po(\lambda_{..} \sigma_j / r) & \end{array}$$

I anvendelser af den multiplikative Poissonmodel er det altid spørgsmålet om eventuel virkning af de to inddelingskriterier der har interesse. Det vil sige hypotesen om *ingen rækkevirking*

$$H_{0R} : \boldsymbol{\rho} = \left( \frac{1}{r}, \dots, \frac{1}{r} \right)$$

og hypotesen om *ingen søjlevirkning*

$$H_{0S} : \boldsymbol{\sigma} = \left( \frac{1}{s}, \dots, \frac{1}{s} \right).$$

Som det fremgår af oversigten over modellerne kan begge hypoteser testes i to modeller. Således svarer både reduktionen  $M_1 \rightarrow M_2$  og reduktionen  $M_2^* \rightarrow M_3$  til hypotesen  $H_{0S}$  om ingen søjlevirkning. Sagt på en anden måde kan  $H_{0S}$  testes både i  $M_1$  og i  $M_2^*$ , og hvis hypotesen ikke

forkastes svarer det til reduktionen til henholdsvis  $M_2$  og  $M_3$ . Vi skal nedenfor se, at uanset om hypotesen  $H_{0S}$  om ingen søjlevirkning testes i  $M_1$  eller i  $M_2^*$ , så er testet det samme.

Tilsvarende bemærkninger kan gøres om hypotesen  $H_{0R}$  om ingen rækkevirkning.

### Estimation

Test for de forskellige modelreduktioner udføres ved hjælp af approksimative  $-2 \ln Q$ -test, som beskrevet i Afsnit 5.7. For at udføre disse test skal vi for hver model kende maksimum likelihood estimatet, værdien  $\hat{l}$  af log likelihood funktionen beregnet i maksimum likelihood estimatet og antallet  $d$  af frie parametre i modellen. Desuden skal vi beregne de forventede antal  $\mathbf{e}$ , for at kunne afgøre om det approksimative test kan benyttes. Disse størrelser beregnes i det følgende for de fem betragtede modeller.

#### $M_0$ :

Likelihood funktionen er

$$\begin{aligned} L(\{\lambda_{ij}\}) &= \prod_{i=1}^r \prod_{j=1}^s e^{-\lambda_{ij}} \lambda_{ij}^{x_{ij}} \frac{1}{x_{ij}!} \\ &= e^{-\lambda_{..}} \prod_{i=1}^r \prod_{j=1}^s \lambda_{ij}^{x_{ij}} \prod_{i=1}^r \prod_{j=1}^s \frac{1}{x_{ij}!}. \end{aligned} \quad (7.41)$$

Da  $\lambda_{ij}$ -erne er variationsuafhængige, får vi af det øverste udtryk ved hjælp af Sætning 7.1, at maksimum likelihood estimatet for  $\lambda_{ij}$  er

$$\hat{\lambda}_{ij} = x_{ij},$$

og af det nederste udtryk ses, at den tilsvarende værdi af log likelihood funktionen er

$$\hat{l}_0 = -x_{..} + \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln(x_{ij}) - \sum_{i=1}^r \sum_{j=1}^s \ln(x_{ij}!). \quad (7.42)$$

Endelig er antallet af frie parametre  $d_0 = rs$ , og de forventede antal  $(\mathbf{e}_0)_{ij} = x_{ij}$ .

#### $M_1$ :

Likelihood funktionen svarende til modellen er

$$\begin{aligned} L(\lambda_{..}, \boldsymbol{\rho}, \boldsymbol{\sigma}) &= \prod_{i=1}^r \prod_{j=1}^s e^{-\lambda_{..} \rho_i \sigma_j} (\lambda_{..} \rho_i \sigma_j)^{x_{ij}} \frac{1}{x_{ij}!} \\ &= e^{-\lambda_{..} x_{..}} \prod_{i=1}^r \rho_i^{x_{i.}} \prod_{j=1}^s \sigma_j^{x_{.j}} \prod_{i=1}^r \prod_{j=1}^s \frac{1}{x_{ij}!}. \end{aligned} \quad (7.43)$$

Da  $\lambda_{..}$ ,  $\boldsymbol{\rho}$  og  $\boldsymbol{\sigma}$  varierer uafhængigt af hinanden, finder vi ved at bruge Sætning 7.1 på den første faktor og Sætning 6.1 på de næste to faktorer, at maksimum likelihood estimatet er givet ved

$$\hat{\lambda}_{..} = x_{..}, \quad \hat{\rho}_i = \frac{x_{i.}}{x_{..}}, \quad \hat{\sigma}_j = \frac{x_{.j}}{x_{..}}.$$

Værdien af log likelihood funktionen i maksimumspunktet er

$$\begin{aligned}\hat{l}_1 &= -x_{..} + x_{..} \ln(x_{..}) + \sum_{i=1}^r x_i \ln\left(\frac{x_i}{x_{..}}\right) + \sum_{j=1}^s x_{.j} \ln\left(\frac{x_{.j}}{x_{..}}\right) - \sum_{i=1}^r \sum_{j=1}^s \ln(x_{ij}!) \\ &= -x_{..} + \sum_{i=1}^r x_i \ln(x_i) + \sum_{j=1}^s x_{.j} \ln(x_{.j}) - x_{..} \ln(x_{..}) - \sum_{i=1}^r \sum_{j=1}^s \ln(x_{ij}!).\end{aligned}\quad (7.44)$$

Antallet af frie parametre er  $d_1 = r + s - 1$ , og de forventede antal er

$$(\mathbf{e}_1)_{ij} = \frac{x_i \cdot x_{.j}}{x_{..}}; \quad (7.45)$$

det vil sige, at det forventede antal i den  $(i, j)$ 'te celle er produktet af den  $i$ 'te rækkesum og den  $j$ 'te søjlesum divideret med totalsummen.

**$M_2$ :**

Anvendes Sætning 7.1 på den første faktor og Sætning 6.1 på den anden faktor i likelihood funktionen

$$\begin{aligned}L(\lambda_{..}, \boldsymbol{\rho}) &= \prod_{i=1}^r \prod_{j=1}^s e^{-\lambda_{..} \rho_i / s} (\lambda_{..} \rho_i / s)^{x_{ij}} \frac{1}{x_{ij}!} \\ &= e^{-\lambda_{..} x_{..}} \lambda_{..}^{x_{..}} \prod_{i=1}^r \rho_i^{x_i} \left(\frac{1}{s}\right)^{x_{..}} \prod_{i=1}^r \prod_{j=1}^s \frac{1}{x_{ij}!},\end{aligned}$$

findes maksimum likelihood estimatet under  $M_2$  til

$$\hat{\lambda}_{..} = x_{..}, \quad \hat{\rho}_i = \frac{x_i}{x_{..}}.$$

Den maksimale værdi af log likelihood funktionen under  $M_2$  er

$$\begin{aligned}\hat{l}_2 &= -x_{..} + x_{..} \ln(x_{..}) + \sum_{i=1}^r x_i \ln\left(\frac{x_i}{x_{..}}\right) + x_{..} \ln\left(\frac{1}{s}\right) - \sum_{i=1}^r \sum_{j=1}^s \ln(x_{ij}!) \\ &= -x_{..} + \sum_{i=1}^r x_i \ln(x_i) - x_{..} \ln(s) - \sum_{i=1}^r \sum_{j=1}^s \ln(x_{ij}!).\end{aligned}\quad (7.46)$$

Antallet af frie parametre er  $d_2 = r$ , og de forventede antal bliver

$$(\mathbf{e}_2)_{ij} = \frac{x_i}{s}; \quad (7.47)$$

de forventede antal i den  $i$ 'te række er altså alle lig med det gennemsnitlige antal observationer i den  $i$ 'te række.

**$M_2^*$ :**

For denne model findes i analogi med  $M_2$ , at

$$\hat{\lambda}_{..} = x_{..}, \quad \hat{\sigma}_j = \frac{x_{.j}}{x_{..}},$$

$$\hat{l}_2^* = -x_{..} + \sum_{j=1}^s x_{.j} \ln(x_{.j}) - x_{..} \ln(r) - \sum_{i=1}^r \sum_{j=1}^s \ln(x_{ij}!), \quad (7.48)$$

$d_2^* = s$  og de forventede antal i den  $j$ 'te søjle er alle lig med gennemsnittet af observationerne i den  $j$ 'te søjle, det vil sige

$$(\mathbf{e}_2^*)_{ij} = \frac{x_{.j}}{r}. \quad (7.49)$$

### M<sub>3</sub>:

Anvendes Sætning 7.1 på den første faktor i likelihood funktionen

$$\begin{aligned} L(\lambda_{..}) &= \prod_{i=1}^r \prod_{j=1}^s e^{-\lambda_{..}/(rs)} (\lambda_{..}/(rs))^{x_{ij}} \frac{1}{x_{ij}!} \\ &= e^{-\lambda_{..}} \lambda_{..}^{x_{..}} \left(\frac{1}{r}\right)^{x_{..}} \left(\frac{1}{s}\right)^{x_{..}} \prod_{i=1}^r \prod_{j=1}^s \frac{1}{x_{ij}!}, \end{aligned}$$

ses, at

$$\hat{\lambda}_{..} = x_{..},$$

og

$$\hat{l}_3 = -x_{..} + x_{..} \ln(x_{..}) - x_{..} \ln(r) - x_{..} \ln(s) - \sum_{i=1}^r \sum_{j=1}^s \ln(x_{ij}!), \quad (7.50)$$

Endvidere er  $d_3 = 1$  og

$$(\mathbf{e}_3)_{ij} = \frac{x_{..}}{rs}; \quad (7.51)$$

med andre ord er det forventede antal i alle celler lig med gennemsnittet af alle observationer.

### Test af hypoteser

Af formlerne (5.40), (7.42) og (7.44) fås, at hypotesen om *multiplikativ* virkning (eller *ingen vekselvirkning*) af de to inddelingskriterier

$$H_{01} : \lambda_{ij} = \lambda_{..} \rho_i \sigma_j, \quad i = 1, \dots, r, \quad j = 1, \dots, s, \quad (7.52)$$

testes ved hjælp af størrelsen

$$\begin{aligned} -2 \ln Q(\mathbf{x}) &= 2[\hat{l}_0 - \hat{l}_1] \\ &= 2\left[\sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln(x_{ij}) - \sum_{i=1}^r x_{i.} \ln(x_{i.}) - \sum_{j=1}^s x_{.j} \ln(x_{.j}) + x_{..} \ln(x_{..})\right], \end{aligned} \quad (7.53)$$

som skal vurderes i en  $\chi^2$ -fordeling med  $d_0 - d_1 = (r-1)(s-1)$  frihedsgrader. Hvis de forventede antal  $\mathbf{e}_1$  i (7.45) er større end eller lig med 5 kan testsandsynligheden beregnes som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2((r-1)(s-1))}(-2 \ln Q(\mathbf{x})). \quad (7.54)$$

Hypotesen om *ingen søjlevirkning* kan specificeres ved

$$H_{0S} : \boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_j, \dots, \sigma_s) = \left(\frac{1}{s}, \dots, \frac{1}{s}, \dots, \frac{1}{s}\right).$$

I modellen  $M_1$  svarer hypotesen  $H_{0S}$  til reduktionen til  $M_2$  og testes ved at betragte størrelsen

$$\begin{aligned} -2\ln Q(\mathbf{x}) &= 2[\hat{l}_1 - \hat{l}_2] \\ &= 2\left[\sum_{j=1}^s x_{.j} \ln(x_{.j}) - x_{..} \ln\left(\frac{x_{..}}{s}\right)\right]. \end{aligned} \quad (7.55)$$

Ved sammenligning med (7.38) ses, at  $-2\ln Q(\mathbf{x})$  er identisk med teststørrelsen for hypotesen om identitet af parametrene for de  $s$  søjlesummer  $X_1, \dots, X_s$ . Testsandsynligheden for hypotesen om ingen søjlevirkning - svarende til reduktionen  $M_1 \rightarrow M_2$  - kan derfor beregnes som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(s-1)}(-2\ln Q(\mathbf{x})), \quad (7.56)$$

hvis det fælles forventede antal for søjlesummerne  $x_{.j}/s$  er større end eller lig med 5.

Hypotesen om *ingen rækkevirkning* er

$$H_{0R} : \boldsymbol{\rho} = (\rho_1, \dots, \rho_i, \dots, \rho_r) = \left(\frac{1}{r}, \dots, \frac{1}{r}, \dots, \frac{1}{r}\right).$$

I modellen  $M_2$  svarer hypotesen til reduktionen til  $M_3$  og testes i denne model ved at betragte

$$\begin{aligned} -2\ln Q(\mathbf{x}) &= 2[\hat{l}_2 - \hat{l}_3] \\ &= 2\left[\sum_{i=1}^r x_i \ln(x_i) - x_{..} \ln\left(\frac{x_{..}}{r}\right)\right]. \end{aligned} \quad (7.57)$$

Af (7.38) ses, at denne teststørrelse er identisk med teststørrelsen for identitet af parametrene for de  $r$  rækkesummer. Hvis det fælles forventede antal for rækkesummerne  $x_{.j}/r$  er større end eller lig med 5, beregnes testsandsynligheden som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(r-1)}(-2\ln Q(\mathbf{x})). \quad (7.58)$$

Vi har nu vist, hvordan man kan foretage reduktioner i modellen  $M_0$  via "ruten"  $M_0 \rightarrow M_1 \rightarrow M_2 \rightarrow M_3$ . Af formlerne (7.42), (7.44), (7.46), (7.48) og (7.50) ovenfor ses, at vi har følgende identiteter:

$$\begin{aligned} 2[\hat{l}_1 - \hat{l}_2] &= 2[\hat{l}_2^* - \hat{l}_3] & d_1 - d_2 &= d_2^* - d_3 \\ &= 2\left[\sum_{j=1}^s x_{.j} \ln(x_{.j}) - x_{..} \ln\left(\frac{x_{..}}{s}\right)\right], & &= s - 1, \\ \\ 2[\hat{l}_1 - \hat{l}_2^*] &= 2[\hat{l}_2 - \hat{l}_3] & d_1 - d_2^* &= d_2 - d_3 \\ &= 2\left[\sum_{i=1}^r x_i \ln(x_i) - x_{..} \ln\left(\frac{x_{..}}{r}\right)\right], & &= r - 1. \end{aligned}$$

Heraf ses, at testet for hypotesen om ingen søjlevirkning er givet ved formlerne (7.55) og (7.56), uanset hvilken af ruterne  $M_0 \rightarrow M_1 \rightarrow M_2 \rightarrow M_3$  eller  $M_0 \rightarrow M_1 \rightarrow M_2^* \rightarrow M_3$ , vi betragter.

Med andre ord; har vi accepteret modellen om multiplikativ virkning, påvirker en eventuel rækkevirkning ikke testet for ingen søjlevirkning - testet er det samme selvom det udføres i de to forskellige modeller  $M_1$  og  $M_2^*$ .

En lignende bemærkning gælder naturligvis for testet af hypotesen om ingen rækkevirkning.

### Fordelingsresultater og relation til multinomialmodellen

Ved hjælp af formel (7.8) kan det vises, at *komponenterne*  $X_i$  i den stokastiske vektor bestående af rækkesummerne  $\mathbf{X}_* = (X_{1.}, \dots, X_{i.}, \dots, X_{r.})$  er stokastisk uafhængige samt at

$$X_i \sim po(\lambda_{.i} \rho_i), \quad i = 1, \dots, r. \quad (7.59)$$

Tilsvarende er *komponenterne*  $X_j$  i vektoren af søjlesummer  $\mathbf{X}_* = (X_{.1}, \dots, X_{.j}, \dots, X_{.s})$  stokastisk uafhængige og

$$X_j \sim po(\lambda_{.j} \sigma_j), \quad j = 1, \dots, s. \quad (7.60)$$

De to vektorer  $\mathbf{X}_*$  og  $\mathbf{X}_*$  er imidlertid **ikke** stokastisk uafhængige, idet summen af komponenterne i begge tilfælde er  $X_{..}$ . Betingelser vi i modellen  $M_0$  med summen af alle observationerne  $x_{..}$  får vi ifølge (7.9) den betingede model

$$\tilde{M}_0 : \{X_{ij}\} \mid X_{..} = x_{..} \sim m(x_{..}, \left\{ \frac{\lambda_{ij}}{\lambda_{.j}} \right\}). \quad (7.61)$$

Da  $\lambda$ -erne varierer frit, er der ingen bånd på sandsynlighedsmatricen i multinomialfordelingen og modellen  $\tilde{M}_0$  svarer til grundmodellen baseret på multinomialfordelingen med antalsparameter  $x_{..}$  for et  $r \times s$  skema. Hypotesen  $H_{01}$  svarer i denne model til hypotesen

$$\tilde{H}_{01} : \frac{\lambda_{ij}}{\lambda_{.j}} = \rho_i \sigma_j, \quad i = 1, \dots, r, \quad j = 1, \dots, s, \quad (7.62)$$

det vil sige til hypotesen om uafhængighed af inddelingskriterier. Det ses af formlerne (6.23), (6.24), (7.53) og (7.54), at testene for  $\tilde{H}_{01}$  og  $H_{01}$  er identiske, selvom det drejer sig om test af forskellige hypoteser i forskellige modeller. Vi har altså hermed set endnu et eksempel på, at man ved at betinge med summen af alle observationer i en model baseret på Poissonfordelingen ”kommer tilbage” til en velkendt multinomialfordelingsmodel.

Når man i en konkret situation skal afgøre, om man skal benytte Poissonmodellen eller multinomialfordelingsmodellen, skal man benytte sig af information om, hvorledes observationerne i  $r \times s$  skemaet er indsamlet. Som tidligere nævnt skal multinomialmodellen benyttes, hvis man på forhånd har lagt sig fast på at betragte, hvorledes et *givet* antal objekter klassificeres efter de to inddelingskriterier; Poissonmodellen benyttes derimod, hvis antallet af objekter, der klassificeres, ikke er kendt på forhånd.

Da analysen af data forløber på samme måde i de to modeller, er det ikke vigtigt at erkende hvilken af de to modeller, man har for så vidt strukturen af data angår. Forskellen mellem de to

modeller ligger kun i, at i Poissonmodellen er der information i det totale antal observationer  $x_{..}$  om intensiteten af det fænomen man observerer, mens det totale antal  $n$  i multinomialmodellen ikke indeholder information.

Af formel (6.26) ses, at vektorerne af rækkesummer  $\mathbf{X}_{*}$  og søjlesummer  $\mathbf{X}_{\cdot}$  er *betinget uafhængige givet* totalsummen  $X_{..}$ , det vil sige, at  $\mathbf{X}_{*}$  og  $\mathbf{X}_{\cdot}$  er uafhængige i den *betingede* fordeling givet  $X_{..} = x_{..}$ .

Som afslutning på omtalen af teorien for den multiplikative Poissonmodel understreger vi, at det af formlerne (7.31), (7.38), (7.57) og (7.58) ses, at testet for ingen rækkevirkning i denne model er ækvivalent med testet for identitet af parametrene i Poissonmodellen for de  $r$  rækkesummer, se formel (7.59).

### Eksempel 7.3 (Fortsat)

Suppleres tabellen side 7.3 med række- og søjlesummer samt totalsum får vi:

<i>land</i>	<i>guld</i>	<i>sølv</i>	<i>bronze</i>	<i>i alt</i>
USA	39	25	33	97
RUS	32	28	28	88
CHN	28	16	15	59
AUS	16	25	17	58
GER	14	17	26	57
FRA	13	14	11	38
<i>i alt</i>	142	125	130	397

I denne situation forekommer det rimeligt at betragte en model for  $6 \times 3$  skemaet baseret på Poissonfordelingen. (At multinomialmodellen ikke forekommer at være korrekt for disse data skyldes, at hverken rækkesummer, søjlesummer eller totalsummen er kendt på forhånd)

Som grundmodel betragter vi altså modellen  $M_0$  med  $r = 6$  og  $s = 3$ . De forventede antal i den multiplikative model  $M_1$  findes ved hjælp af (7.45) og ovenstående skema til:

<i>land</i>	<i>guld</i>	<i>sølv</i>	<i>bronze</i>	<i>i alt</i>
USA	34.695	30.542	31.763	97.000
RUS	31.476	27.708	28.816	88.000
CHN	21.103	18.577	19.320	59.000
AUS	20.746	18.262	18.992	58.000
GER	20.388	17.947	18.665	57.000
FRA	13.592	11.965	12.443	38.000
<i>i alt</i>	142.000	125.000	130.000	397.000

Da de forventede antal alle er større end eller lig med 5 kan reduktionen til den multiplikative

model testes ved hjælp af formlerne (7.53) og (7.54). Vi finder

$$-2 \ln Q(\mathbf{x}) = 14.1514$$

og

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(10)}(14.1514) = 0.1662.$$

Hypotesen  $H_{01}$  accepteres altså og dermed reduktionen til den multiplikative model  $M_1$ .

Som nævnt ovenfor kan hypotesen om ingen søjlevirkning undersøges ved at teste om parametrene for søjlesummerne er identiske. Ingen søjlevirkning betyder i dette tilfælde, at medaljernes karat ingen indflydelse har på antallet af medaljer. Da det forventede antal observationer i den  $j$ 'te søjle under hypotesen om ingen søjlevirkning er  $x_{..}/s$  - her  $x_{..}/3$  - er de observerede og forventede antal følgende:

<i>søjlesummer</i>	<i>guld</i>	<i>sølv</i>	<i>bronze</i>	<i>i alt</i>
<i>observeret</i>	142	125	130	397
<i>forventet</i>	132.333	132.333	132.333	396.999

Af (7.55) og (7.56) fås, at

$$-2 \ln Q(\mathbf{x}) = 1.1450,$$

og

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(2)}(1.1450) = 0.5641,$$

og hypotesen om ingen søjlevirkning accepteres. Fordelingen af medaljer kan altså antages at være uafhængig af medaljernes karat.

For at vurdere om fordelingen af medaljer er den samme for de seks lande betragtes rækkesummerne, og det undersøges, om parametrene i fordelingerne for rækkesummerne er identiske. Vi finder - med tre decimalers nøjagtighed:

<i>rækkesummer</i>	<i>USA</i>	<i>RUS</i>	<i>CHN</i>	<i>AUS</i>	<i>GER</i>	<i>FRA</i>	<i>i alt</i>
<i>observeret</i>	97	88	59	58	57	38	397
<i>forventet</i>	66.167	66.167	66.167	66.167	66.167	66.167	397.002

Ved hjælp af (7.57) og (7.58) finder vi, at

$$-2 \ln Q(\mathbf{x}) = 36.4380,$$

og

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(5)}(36.4380) \approx 0,$$

og hypotesen om ingen rækkevirkning forkastes. Antallet af medaljer afhænger ikke overraskende af landene.

Slutmodellen for disse data er således

$$M_2 : x_{ij} \sim\sim po(\alpha_i \beta) = po(\lambda_{..} \rho_i / s), \quad i = 1, \dots, 6 \quad j = 1, 2, 3.$$

I denne model gælder, at søjlesummerne er uafhængige og

$$x_{i.} \sim\sim po(\lambda_{..} \rho_i), \quad i = 1, \dots, 6.$$

Estimaterne er  $\hat{\lambda}_{..} = x_{..}$  og  $\hat{\rho}_i = x_{i.} / x_{..}$ ,  $i = 1, \dots, 6$ , det vil sige

$$\begin{aligned} \hat{\lambda}_{..} = 397, \quad \hat{\rho}_1 = \frac{97}{397} = 0.244, \quad \hat{\rho}_2 = \frac{88}{397} = 0.222, \quad \hat{\rho}_3 = \frac{59}{397} = 0.149 \\ \hat{\rho}_4 = \frac{58}{397} = 0.146, \quad \hat{\rho}_5 = \frac{57}{397} = 0.144, \quad \hat{\rho}_6 = \frac{38}{397} = 0.096. \end{aligned}$$

□

## Anneks til Kapitel 7

### Beregninger i Excel

Excel har ikke specielle dialogbokse der udfører beregninger i de modeller for Poissonfordelte data, der er omtalt i dette kapitel. Beregningerne udføres dog let som vist nedenfor.

#### Eksempel 7.1 (Fortsat)

Regnearket nedenfor viser beregningen af Fishers dispersion indeks samt testet for goodness of fit til kontrol af modellen  $M_0$  for fordelingen af målene i kampene 1-66.

	A	B	C	D	E	F	G	H
1	antal mål	kamp 1-66		antal mål^2		S:	175	
2	0	3		0		SK:	599	
3	1	12		1		middelværdi:	2,6515	
4	2	18		4		varians:	2,0767	
5	3	13		9		t:	0,7832	
6	4	15		16		grænser:	0,6862	1,3720
7	5	2		25				
8	6	3		36				
9	7	0		49				
10	8	0		64				
11	9	0		81				
12	10	0		100				
13								
14								
15	antal mål	kamp 1-66	forventet		grupperet:	observeret	forventet	ln(x/e)
16	0	3	4,6559		0-1	15	17,0012	-0,125232
17	1	12	12,3452		2	18	16,3668	0,095117
18	2	18	16,3668		3	13	14,4656	-0,106824
19	3	13	14,4656		4	15	9,5889	0,447440
20	4	15	9,5889		>4	5	8,5775	-0,539703
21	5	2	5,0850		i alt	66	66,0000	
22	6	3	2,2472					
23	7	0	0,8512		_2lnQ:	4,9160		
24	8	0	0,2821		testss:	0,1781		
25	9	0	0,0831					
26	10	0	0,0220					
27	i alt	66	65,9932					

Først beregningen af Fishers dispersionsindeks på side 7.8. Data på tabelform er i cellerne A2:B12 og for sådanne data beregnes summen  $S$  og kvadratsummen  $SK$  som

$$S = \sum_j ja_j \quad \text{og} \quad SK = \sum_j j^2 a_j.$$

Værdierne  $j^2$  er beregnet i cellerne D2:D12. Værdien i D2 beregnes som

$$A2 * A2$$

og analoge formler oprettes i cellerne D3:D12. Summen i G1 og kvadratsummen i G2 beregnes ved hjælp af funktionen SUMPRODUKT som

$$= \text{SUMPRODUKT}(A2 : A12; B2 : B12) \quad (= \sum_j j a_j)$$

og

$$= \text{SUMPRODUKT}(B2 : B12; D2 : D12) \quad (= \sum_j j^2 a_j).$$

Da antallet af observationer er  $n = 66$  beregnes empirisk middelværdi  $\bar{x}$  og varians  $s^2$  i G3 og G4 som

$$G1/66 \quad (= S/n)$$

og

$$= (G2 - G1 * G1/66)/65 \quad (= \frac{1}{n-1}(SK - \frac{S^2}{n}))$$

og værdien af Fishers dispersionsindeks  $t$  i G5 som

$$G4/G3 \quad (= \frac{s^2}{\bar{x}}).$$

Grænserne for acceptområdet i G6 og G7 beregnes i en  $\chi^2(65)/65$ -fordeling som henholdsvis

$$= \text{CHIINV}(1 - 0,025; 65)/65 \quad (\chi_{0,025}^2(65)/65)$$

og

$$= \text{CHIINV}(1 - 0,975; 65)/65 \quad (\chi_{0,975}^2(65)/65).$$

Cellerne A15:H27 vedrører testet for goodness of fit, side 7.9. I cellerne A16:B26 ser vi igen data på tabelform, mens cellerne C16:C26 indeholder de forventede antal. Disse er beregnet ved i C16 at beregne

$$= \text{POISSON}(A16; \$G\$3; \text{FALSK}) * 66 \quad (n \cdot po(0; \bar{x}))$$

for derefter at oprette analoge formler i C17:C26.

Cellerne E15:G20 indeholder den grupperede version af data og de forventede værdier, der opfylder at de forventede antal er større end eller lig med 5. Indholdet af F16 og G16 er beregnet som henholdsvis

$$= B16 + B17 \quad (= a_0 + a_1)$$

og

$$= C16 + C17 \quad (= e_0 + e_1).$$

Herefter kopieres indholdet af cellerne B18:C20 til cellerne F17:G19. Endelig beregnes værdien i F20 som

$$= \text{SUM}(B21 : B26) \quad (= \sum_{j=5}^{10} a_j)$$

og værdien i G20 som

$$= 66 - \text{SUM}(G16 : G19) \quad (= n - \sum_{j=0}^4 e_j).$$

Vi mangler nu kun at beregne  $-2 \ln Q$ -teststørrelsen for testet for goodness of fit og den tilsvarende testsandsynlighed. Hertil beregnes værdien i H16 som

$$= \text{LN}(F16/G16) \quad (= \ln(\frac{a_{0-1}}{e_{0-1}}))$$

og analoge formler oprettes i celleren H17:H20. Herefter beregnes værdien i F23 som

$$= 2 * \text{SUMPRODUKT}(F16 : F20; H16 : H20) \quad (= 2[a_{0-1} \ln(\frac{a_{0-1}}{e_{0-1}}) + \sum_{j=2}^4 a_j \ln(\frac{a_j}{e_j}) + a_{\geq 5} \ln(\frac{a_{\geq 5}}{e_{\geq 5}})])$$

og testsandsynligheden i F24 som

$$= \text{CHIFORDELING}(F23; 3) \quad (= 1 - F_{\chi^2(5-1-1)}(-2 \ln Q)).$$

□

### Eksempel 7.2 (Fortsat)

Vi viser her, hvorledes beregningerne i Poissonmodellen med proportionale parametre for data på side 7.16 kan udføres. Resultatet er vist nedenfor.

	A	B	C	D	E	F
1	land	antal medaljer	antal indbyggere	forventet	observeret-forventet	$\ln(x/e)$
2	Danmark	6	5,3	7,0962	-1,0962	-0,16780
3	Finland	4	5,2	6,9623	-2,9623	-0,55422
4	Norge	10	4,5	6,0251	3,9749	0,50665
5	Sverige	12	8,9	11,9163	0,0837	0,00700
6	i alt	32	23,9	32,0000	0,0000	
7						
8	$-2 \ln Q$ :	3,8535				
9	testss:	0,2777				

Cellerne B2:B5 indeholder de observerede antal medaljer,  $x_i$ , og cellerne C2:C5 indbyggertallene,  $m_i$ , i millioner. Først beregnes summerne i B6 og C6. Derefter beregnes de forventede antal,  $e_i$ , i D2:D5. Først beregnes værdien i D2 som

$$= \text{\$B\$6} * \text{\$C2}/\text{\$C\$6} \quad (= x \cdot \frac{m_1}{m}).$$

og analoge formler oprettes i D3:D5. Herefter beregnes værdien i F2 som

$$= \text{LN}(B2/D2) \quad (= \ln(\frac{x_1}{e_1}))$$

og analoge formler oprettes i F3:F5. Endelig beregnes  $-2 \ln Q$ -teststørrelsen i B8 som

$$= 2 * \text{SUMPRODUKT}(B2 : B5; F2 : F5) \quad (= 2 \sum_{i=1}^k x_i \ln(\frac{x_i}{e_i}))$$

og testsandsynligheden i B9 som

$$= \text{CHIFORDELING}(B8; 3) \quad (= 1 - F_{\chi^2(k-1)}(-2 \ln Q)).$$

□

### Eksempel 7.3 (Fortsat)

Som nævnt ovenfor er beregningerne i testet for den multiplikative Poissonmodel identiske med beregningerne i testet for uafhængighed mellem inddelingskriterier i en multinomialmodel. Beregningerne i *Excel* kan derfor udføres som vist i Eksempel 6.3 på side 6.29. □

## Hovedpunkter til Kapitel 7

### Én observationsrække

*Model:*

Observationerne  $x_1, \dots, x_n$  betragtes som udfald af uafhængige stokastiske variable  $X_1, \dots, X_n$ , som alle er Poissonfordelte med parameter  $\lambda$ , det vil sige

$$M_0: X_i \sim po(\lambda), \quad i = 1, \dots, n.$$

*Estimat:*

Maksimum likelihood estimatet  $\hat{\lambda}$  for  $\lambda$  er

$$\hat{\lambda} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Fordelingen af maksimum likelihood estimatoren angives ved

$$n\hat{\lambda} = X \sim po(n\lambda).$$

*Modelkontrol:*

Hvis stikprøvestørrelsen  $n$  er tilstrækkelig stor, kan  $M_0$  kontrolleres ved et  $\chi^2$ -test for goodness of fit, som beskrevet i Afsnit 6.5.

En alternativ kontrol af modellen  $M_0$  baserer sig på *Fishers dispersionsindeks*, som er forholdet

$$t = \frac{s^2}{\bar{x}}$$

mellem den empiriske varians  $s^2$  og den empiriske middelværdi  $\bar{x}$ . Beregningen af den empiriske middelværdi og varians afhænger af, om alle de enkelte observationer er til rådighed eller om observationerne er givet på tabelform. Med indlysende betegnelser har vi

$$S = \sum_{i=1}^n x_i = \sum_j j a_j,$$

$$SK = \sum_{i=1}^n x_i^2 = \sum_j j^2 a_j$$

og

$$\bar{x} = \frac{1}{n} S \quad \text{og} \quad s^2 = \frac{1}{n-1} (SK - \frac{S^2}{n}).$$

Modellen  $M_0$  accepteres ved et test på niveau  $\alpha$ , hvis

$$\chi_{\alpha/2}^2(n-1)/(n-1) \leq t \leq \chi_{1-\alpha/2}^2(n-1)/(n-1),$$

Testet er baseret på en approksimation, som kan anvendes hvis  $n \geq 15$  eller  $\bar{x} \geq 5$ .

*Konfidensintervaller:*

*Middelværdien i en Poissonfordelt stokastisk variabel:*

Et approksimativt  $1 - \alpha$  konfidensinterval for parameteren  $\lambda$  baseret på én observation  $x$  fra  $po(\lambda)$  fordelingen er af formen

$$C_{1-\alpha}(x) = [\lambda_-, \lambda_+],$$

hvor

$$\lambda_- = x + \frac{1}{2}u_{1-\alpha/2}^2 - u_{1-\alpha/2}\sqrt{x + \frac{1}{4}u_{1-\alpha/2}^2}$$

og

$$\lambda_+ = x + \frac{1}{2}u_{1-\alpha/2}^2 + u_{1-\alpha/2}\sqrt{x + \frac{1}{4}u_{1-\alpha/2}^2}.$$

I formlerne betegner  $u_{1-\alpha/2}$  ( $1 - \alpha/2$ )-fraktilen i  $u$ -fordelingen. Hvis  $\alpha = 0.05$  er fraktilen  $u_{0.975} = 1.960$ .

*Middelværdien i én observationsrække fra Poissonfordelingen:*

Her er summen  $x \sim po(n\lambda)$  og  $1 - \alpha$  konfidensintervallet for  $\lambda$  har grænserne

$$\lambda_- = \frac{1}{n} \left[ x + \frac{1}{2}u_{1-\alpha/2}^2 - u_{1-\alpha/2}\sqrt{x + \frac{1}{4}u_{1-\alpha/2}^2} \right]$$

og

$$\lambda_+ = \frac{1}{n} \left[ x + \frac{1}{2}u_{1-\alpha/2}^2 + u_{1-\alpha/2}\sqrt{x + \frac{1}{4}u_{1-\alpha/2}^2} \right].$$

## Flere fordelinger

*Poissonmodellen med proportionale parametre:*

Datasættet  $\mathbf{x}$  består af observationerne  $x_1, \dots, x_k$  der kan betragtes som udfald af uafhængige stokastiske variable  $X_1, \dots, X_k$ , som alle er Poissonfordelt men med hver sin parameter, det vil sige, at grundmodellen er

$$M_0 : X_i \sim po(\lambda_i), \quad i = 1, \dots, k.$$

Vi er interesseret i at teste hypotesen, om at parametrene i modellen  $M_0$  er proportionale med de kendte tal  $m_1, \dots, m_k$  som proportionalitetsfaktorer, det vil sige hypotesen

$$H_{01} : \lambda_i = m_i\lambda, \quad i = 1, \dots, k.$$

Accepteres hypotesen reduceres  $M_0$  til modellen

$$M_1 : X_i \sim po(m_i\lambda), \quad i = 1, \dots, k.$$

Bemærk, at man i modellen  $M_0$  kan undersøge, om  $x_1, \dots, x_k$  kan betragtes som én observationsrække fra  $po(\lambda)$ -fordelingen, ved at teste hypotesen svarende til at  $m_1 = \dots = m_k = 1$ .

Maksimum likelihood estimatet for  $\lambda$  under  $M_1$  er

$$\hat{\lambda} = \frac{x.}{m.},$$

hvor  $x. = x_1 + \dots + x_k$  og  $m. = m_1 + \dots + m_k$ . Det forventede antal i  $M_1$  svarende til observationen  $x_i$  er

$$e_i = m_i \hat{\lambda} = x. \frac{m_i}{m.}.$$

og  $-2 \ln Q$ -teststørrelsen for  $H_{01}$  er

$$-2 \ln Q(\mathbf{x}) = 2 \sum_{i=1}^k x_i \ln \left( \frac{x_i}{e_i} \right).$$

Hvis **de forventede antal alle er større end eller lig med 5**, gælder der følgende approksimation af testsandsynligheden for  $H_{01}$  :

$$\varepsilon(\mathbf{x}) \doteq 1 - F_{\chi^2(k-1)}(-2 \ln Q(\mathbf{x})).$$

Fordelingen af maksimum likelihood estimatoren for  $\lambda$  under  $H_{01}$  angives som regel på følgende måde:

$$m. \hat{\lambda} = X. \sim po(m. \lambda).$$

*Konfidensintervaller for parameteren i Poissonmodellen med proportionale parametre:*

Erstattes  $x$  med  $x.$  i (7.19) og (7.20) fås grænserne for  $1 - \alpha$  konfidensintervallet for  $m. \lambda$ .

Det transformeres til et konfidensinterval for  $\lambda$  med grænserne

$$\lambda_- = \frac{1}{m.} \left[ x. + \frac{1}{2} u_{1-\alpha/2}^2 - u_{1-\alpha/2} \sqrt{x. + \frac{1}{4} u_{1-\alpha/2}^2} \right]$$

og

$$\lambda_+ = \frac{1}{m.} \left[ x. + \frac{1}{2} u_{1-\alpha/2}^2 + u_{1-\alpha/2} \sqrt{x. + \frac{1}{4} u_{1-\alpha/2}^2} \right].$$

Den multiplikative Poissonmodel:

Observationerne kan - som vist side 7.19 - opskrives i en  $r \times s$  tabel svarende til to inddelingskriterier med henholdsvis  $r$  og  $s$  kategorier. Observationen svarende til den  $i$ 'te kategori ved det første kriterium og den  $j$ 'te kategori ved det andet kriterium betegnes med  $x_{ij}$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, s$ . Endvidere betegner  $x_{i.}$  og  $x_{.j}$  henholdsvis summen af observationerne i den  $i$ 'te række og den  $j$ 'te søjle, mens  $x_{..}$  er summen af alle observationerne, det vil sige

$$x_{i.} = \sum_{j=1}^s x_{ij}, \quad x_{.j} = \sum_{i=1}^r x_{ij}, \quad x_{..} = \sum_{i=1}^r \sum_{j=1}^s x_{ij}.$$

Idet observationerne antages at være udfald af uafhængige stokastiske variable, betragtes følgende modeller:

Grundmodellen

$$M_0 : x_{ij} \sim \text{po}(\lambda_{ij}), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Den multiplikative model eller modellen for ingen vekselvirkning

$$M_1 : x_{ij} \sim \text{po}(\alpha_i \beta_j), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Modellen for kun rækkevirkning

$$M_2 : x_{ij} \sim \text{po}(\alpha_i \beta), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Modellen for kun søjlevirkning

$$M_2^* : x_{ij} \sim \text{po}(\alpha \beta_j), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Modellen for homogenitet

$$M_3 : x_{ij} \sim \text{po}(\alpha \beta), \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

Benyttes omskrivning af parameteren under  $M_1$

$$\alpha_i \beta_j = \alpha \beta \cdot \frac{\alpha_i}{\alpha} \frac{\beta_j}{\beta} = \lambda_{..} \rho_i \sigma_j.$$

kan modellerne  $M_1$ ,  $M_2$ ,  $M_2^*$  og  $M_3$  og deres indbyrdes forhold angives på følgende måde :

$$\begin{array}{ccc}
 & M_2 : X_{ij} \sim \text{po}(\lambda_{..} \rho_i / s) & \\
 \nearrow & & \searrow \\
 M_1 : X_{ij} \sim \text{po}(\lambda_{..} \rho_i \sigma_j) & & M_3 : X_{ij} \sim \text{po}(\lambda_{..} / (rs)) \\
 \searrow & & \nearrow \\
 & M_2^* : X_{ij} \sim \text{po}(\lambda_{..} \sigma_j / r) &
 \end{array}$$

Hypotesen om *multiplikativ* virkning (eller ingen vekselvirkning) af de to inddelingskriterier,

$$H_{01} : \lambda_{ij} = \lambda_{.i} \rho_i \sigma_j, \quad i = 1, \dots, r, \quad j = 1, \dots, s,$$

svarer til reduktionen fra  $M_0$  til  $M_1$  og testes ved hjælp af størrelsen

$$-2 \ln Q(\mathbf{x}) = 2 \left[ \sum_{i=1}^r \sum_{j=1}^s x_{ij} \ln(x_{ij}) - \sum_{i=1}^r x_{i.} \ln(x_{i.}) - \sum_{j=1}^s x_{.j} \ln(x_{.j}) + x_{..} \ln(x_{..}) \right],$$

Hvis de forventede antal under  $M_1$

$$(\mathbf{e}_1)_{ij} = \frac{x_{i.} x_{.j}}{x_{..}},$$

**alle er større end eller lig med 5**, kan testsandsynligheden beregnes som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2((r-1)(s-1))}(-2 \ln Q(\mathbf{x})).$$

Hypotesen om *ingen søjlevirkning*,

$$H_{0S} : \boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_j, \dots, \sigma_s) = \left( \frac{1}{s}, \dots, \frac{1}{s}, \dots, \frac{1}{s} \right),$$

svarer i modellen  $M_1$  til reduktionen til  $M_2$  og testes her ved hjælp af størrelsen

$$-2 \ln Q(\mathbf{x}) = 2 \left[ \sum_{j=1}^s x_{.j} \ln(x_{.j}) - x_{..} \ln\left(\frac{x_{..}}{s}\right) \right].$$

Testsandsynligheden beregnes som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(s-1)}(-2 \ln Q(\mathbf{x})),$$

forudsat, at  $x_{..}/s \geq 5$ .

Hypotesen om *ingen rækkevirkning*,

$$H_{0R} : \boldsymbol{\rho} = (\rho_1, \dots, \rho_i, \dots, \rho_r) = \left( \frac{1}{r}, \dots, \frac{1}{r}, \dots, \frac{1}{r} \right),$$

svarer i modellen  $M_1$  til reduktionen til  $M_2^*$  og testes her ved at betragte

$$-2 \ln Q(\mathbf{x}) = 2 \left[ \sum_{i=1}^r x_{i.} \ln(x_{i.}) - x_{..} \ln\left(\frac{x_{..}}{r}\right) \right].$$

Hvis  $x_{..}/r \geq 5$ , beregnes testsandsynligheden som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(r-1)}(-2 \ln Q(\mathbf{x})).$$

I modellen  $M_2$  svarer hypotesen om *ingen rækkevirkning* til reduktionen til  $M_3$  og testes i denne model ved at betragte størrelsen

$$-2 \ln Q(\mathbf{x}) = 2 \left[ \sum_{i=1}^r x_{i.} \ln(x_{i.}) - x_{..} \ln\left(\frac{x_{..}}{r}\right) \right].$$

Hvis  $\mathbf{x}_{..}/\mathbf{r} \geq 5$ , beregnes testsandsynligheden som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(r-1)}(-2 \ln Q(\mathbf{x})).$$

I modellen  $M_2^*$  svarer hypotesen om *ingen søjlevirkning* til reduktionen til  $M_3$ . Hvis  $\mathbf{x}_{..}/\mathbf{s} \geq 5$  benyttes teststørrelsen

$$-2 \ln Q(\mathbf{x}) = 2 \left[ \sum_{j=1}^s x_{.j} \ln(x_{.j}) - x_{..} \ln\left(\frac{x_{..}}{s}\right) \right],$$

og den tilsvarende testsandsynlighed bliver

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(s-1)}(-2 \ln Q(\mathbf{x})).$$

## Opgaver til Kapitel 7

**Opgave 7.1** Antag, at  $X_1, \dots, X_n$  er uafhængige og identisk Poissonfordelte med parameter  $\lambda$ ,  $X_j \sim po(\lambda)$ ,  $j = 1, \dots, n$ . I Opgave 5.2 viste vi, at log likelihood funktionen for  $\lambda$  er

$$l(\lambda) = -n\lambda + x \cdot \ln \lambda - \sum_{i=1}^n \ln x_i!$$

samt at maksimum likelihood estimatet for  $\lambda$  er  $\bar{x}$ .

- a) Vis, at  $-2 \ln Q$ -teststørrelsen for den simple hypotese

$$H_0 : \lambda = \lambda_0,$$

hvor  $\lambda_0$  er en kendt værdi er

$$-2 \ln Q(\mathbf{x}) = 2 \left[ x \cdot \ln \left( \frac{\bar{x}}{\lambda_0} \right) + n\lambda_0 - n\bar{x} \right] \sim \chi^2(1).$$

- b) Vis, at dette også er  $-2 \ln Q$ -teststørrelsen for hypotesen  $H_0 : \lambda = \lambda_0$  i modellen  $X \sim po(n\lambda)$ .
- c) Test hypotesen  $\lambda = 5$  hvis  $x = 7$ ,  $n = 1$  og hvis  $x = 70$ ,  $n = 10$ .

**Opgave 7.2** En cykelrytter har i løbet af sin karriere på 10 år 13 styrt, mens en anden i løbet af en karriere på 5 år er udsat for 11 styrt.

- a) Vis, idet antallet af styrt antages at være Poissonfordelt, at der ikke er signifikant forskel på antallet af styrt per år som de to ryttere har været udsat for.
- b) Angiv et estimat og 95% konfidensintervallet for antallet af styrt per år.

**Opgave 7.3** Faxe Kondi Divisionen 1999-2000 omfattede 16 hold, der mødtes to gange i turneringen, i alt 240 kampe. Nedenfor ses på tabelform fordelingen af mål i de 240 kampe.

Faxe Kondi Divisionen 1999-2000

antal mål	antal kampe
0	11
1	25
2	48
3	53
4	53
5	23
6	15
7	3
8	6
9	3
i alt	<u>240</u>

- a) Vis dels ved hjælp af Fishers dispersionsindeks og dels ved hjælp af et test for goodness of fit, at antallet af mål i de 240 kampe kan betragtes som én Poissonfordelt observationsrække.
- b) Undersøg ved hjælp af tallene her for Faxe Kondi Divisionen og tallene for Faxe Kondi Ligaen på side 7.18 om der er forskel på antal scorede mål per kamp i de to rækker.

**Opgave 7.4** I et bachelorprojekt fra Institut for Idræt, Københavns Universitet, med titlen *Fysiske krav i elitefodbold for ungdomsspillere* undersøger Berg og Blæsild (2000) blandt andet løbemønstre hos spillerne. Man skelner mellem *lavintensiv aktivitet*, som omfatter *stå stille*, *gå*, *jog*, *let løb* og *baglæns løb*, og *højintensiv aktivitet*, som omfatter *halvhurtigt løb*, *hurtigt løb* og *sprint*. En bestemt spiller videofilmes i en hel kamp og det optælles hvor mange gange i løbet af kampen en spiller har været i hver af de otte kategorier, nævnt ovenfor.

De første tal nedenfor vedrører en sammenligning af en spiller fra Faxe Kondi Ligaen (*senior*) og en spiller fra Ynglinge Ligaen (*yngling*).

niveau	lavintensitet					højintensitet		
	stå	gå	jog	let	baglæns	halvhurtigt	hurtigt	sprint
senior	143	339	302	250	35	140	66	23
yngling	145	370	342	242	39	105	34	7

Antag, at de observerede tal er Poissonfordelte.

- a) Illuster de observerede antal for såvel lavintensitet som højintensitet aktiviteter ved hjælp af figurer lavet i *Excel*.
- b) Vis for såvel lavintensitet som højintensitet aktiviteter at data kan beskrives ved en multiplikativ Poissonmodel.

- c) Undersøg for såvel lavintensitet som højintensitet aktiviteter om der er forskel på senioren og ynglingen.

De nedenstående tal vedrører en sammenligning af løbemønstre for ynglinge spillere med forskellige positioner på banen.

placering	lavintensitet					højintensitet		
	stå	gå	jog	let	baglæns	halvhurtigt	hurtigt	sprint
forsvar	182	413	330	216	77	90	24	5
midtbane	124	373	372	287	70	121	38	8
angreb	128	323	314	222	32	104	41	9

Antag igen, at de observerede tal er Poissonfordelte.

- d) Illustrer de observerede antal for såvel lavintensitet som højintensitet aktiviteter ved hjælp af figurer lavet i *Excel*.
- e) Vis, at højintensitet aktiviteterne kan beskrives ved en multiplikativ Poissonmodel mens dette ikke er tilfældet for lavintensitet aktiviteterne.
- f) Undersøg, om spillerenes højintensitet aktiviteter afhænger af positionen på banen.

De følgende tre opgaver har ikke noget med idræt at gøre men vedrører andre interessante anvendelser af Poissonfordelingen.

**Opgave 7.5** Data i denne opgave vedrører bombing af den sydlige del af London under Anden Verdenskrig. Området er opdelt i 576 delområder hvert på  $1/4 \text{ km}^2$ , og for hvert delområde er det registreret, hvor mange bomber der faldt i det pågældende område. Registreringerne er gengivet i Tabel 7.1. nedenfor.

- a) Vis ved at betragte Fishers dispersionsindeks, at det kan antages, at de 576 observationer kan betragtes som en Poissonfordelt observationsrække.
- b) Angiv et estimat for antallet af bomber, der faldt i et delområde, samt et 95% konfidensinterval for dette antal.
- c) Angiv et estimat for sandsynligheden for, at der ingen bombe faldt i et delområde, samt et 95% konfidensinterval for denne sandsynlighed.

**Opgave 7.6** Data i denne opgave består af registreringer af ”store” jordskælv i en periode på 75 år fra 1903 til og med 1977. Et jordskælv betegnes som ”stort”, hvis dets størrelse på Richter skalaen er mindst 7.5 eller hvis mere end 1000 mennesker er omkommet ved jordskælvet. Tabel 7.2 nedenfor viser på tabelform det årlige antal jordskælv for de 75 år.

$j$	$a_j$
0	229
1	211
2	93
3	35
4	7
5	0
6	0
7	1
$n$	576
$S$	537
$SK$	1059

**Tabel 7.1** På tabelform er angivet antallet af bomber, der faldt i de 576 delområder på hvert  $1/4$  km<sup>2</sup> i det sydlige London under Anden Verdenskrig. Endvidere er antal observationer  $n$ , sum  $S$  og kvadratsum  $USS$  angivet.

$j$	$a_j$
0	31
1	28
2	14
3	1
4	1
$n$	75
$S$	63
$SK$	109

**Tabel 7.2** På tabelform er angivet det årlige antal “store“ jordskælv for de 75 år fra 1903 til og med 1977. Endvidere er antal observationer  $n$ , sum  $S$  og kvadratsum  $USS$  angivet.

- a) Vis ved at betragte Fishers dispersionsindeks, at det kan antages, at de 75 observationer kan betragtes som en Poissonfordelt observationsrække.
- b) Angiv et estimat for det årlige antal "store" jordskælv i den betragtede periode samt et 95% konfidensinterval for dette antal.

Antag, at der i de kommende 25 år vil indtræffe 23 "store" jordskælv.

- c) Undersøg om det kan antages, at det årlige antal "store" jordskælv er det samme for de næste 25 år som for perioden fra 1903 til og med 1977.

**Opgave 7.7** På en større arbejdsplads har man over en periode på 5 uger registreret antallet af tilskadekomster opdelt efter faggruppe og tid på dagen.

- a) Undersøg, hvorledes ulykkesantallet afhænger af faggruppe og tid på dagen.
- b) På virksomheden var der i den pågældende periode ansat 2413 faglærte, 988 ufaglærte og 539 lærlinge. Er der samme ulykkeshyppighed i de tre faggrupper?

<i>Tid</i>	<i>Faggruppe</i>		
	<i>Faglærte</i>	<i>Ufaglærte</i>	<i>Lærlinge</i>
<i>Før frokost</i>	203	90	90
<i>Efter frokost</i>	250	98	93

**Tabel 7.3** Arbejdsulykker inddelt efter faggruppe og tid på dagen.



## 8 Ikke-parametriske test

Som det fremgår af Kapitel 4 - Kapitel 7 er den statistiske inferens i denne bog parametrisk, idet den er baseret på parametriserede klasser af fordelinger. Udgangspunktet er at data  $\mathbf{x}$  opfattes som udfald af en stokastisk vektor  $\mathbf{X}$ , hvis fordelingsfunktion antages at tilhøre en parametriseret klasse af fordelingsfunktioner  $\mathcal{F} = \{F_{\boldsymbol{\omega}} : \boldsymbol{\omega} \in \Omega\}$ . Her er parameteren  $\boldsymbol{\omega}$  valgt, således at den er relevant for den saglige sammenhæng, der ligger til grund for det eksperiment, hvis resultat var data  $\mathbf{x}$ .

Undertiden kritiseres parametrisk inferens for at være for følsom overfor afvigelser for den valgte fordelingsklasse, eksempelvis hævdes det nu og da, at test udledt i en statistisk model baseret på normalfordelingen er for følsomme over for afvigelser fra antagelsen om normalitet. Argumentet er ofte at den empiriske varians  $s^2$  påvirkes meget af ekstreme værdier. Hvis for eksempel nogle få af observationerne har meget ekstreme værdier vokser den empiriske varians  $s^2$ , og da  $s$  eller  $s^2$  optræder i nævneren i henholdsvis  $t$ - og  $F$ -test bliver disse teststørrelser tilsvarende små, hvilket igen betyder, at signifikante afvigelser fra de betragtede hypoteser ikke afsløres.

I modsætning til en del andre bøger i elementær statistik har vi i denne bog beskæftiget os en del med det punkt i en statistisk analyse der hedder modelkontrol og som netop vedrører spørgsmålet om data kan beskrives ved hjælp af modellens parametriserede fordelingsklasse. Rimeligheden af fordelingsklassen er i alle eksempler blevet vurderet ved hjælp af grafiske eller numeriske test baseret på de oprindelige observationer eller på residualer i modellen. Hvis denne kontrol af modellen falder negativt ud skal man naturligvis ikke drage inferens i den betragtede model, da konklusioner baseret på en forkert model sjældent er rigtige. Hvis man ved modelkontrollen får det indtryk, at en eller flere observationer er ekstreme i forhold til de øvrige er det naturligt at kontakte personen, der har udført eksperimentet, for at få en forklaring. Hvis det viser sig, at de ekstreme observationer skyldes ændrede forsøgsbetingelser kan man udelade disse observationer fra beregningerne. Hvis fagmanden derimod bekræfter gyldigheden af de ekstreme observationer må modellen forkastes og en ny opstilles.

Hvis det viser sig helt umuligt af finde en parametriseret fordelingsklasse, der giver en rimelig beskrivelse af data, kan man ty til ikke-parametrisk statistik, som er baseret på min-

dre specifikke antagelser vedrørende observationernes fordeling. Det hævdes undertiden, at ikke-parametrisk statistik er fri for forudsætninger, hvilket ikke er korrekt. De fleste ikke-parametriske test er udledt under forudsætninger såsom uafhængighed, identiske fordelinger og undertiden også symmetriske fordelinger af observationerne.

Formålet med kapitlet her er at give et indtryk af tankegangen i ikke-parametrisk statistik. I Afsnit 8.1 omtales *fortegnstestet*, som nok er det simpleste af de ikke-parametriske test. I Afsnit 8.2, der er baseret på Lehmann (1975), omtales de simpleste eksempler på *rangtest*, det vil sige test baseret på observationernes rang. I afsnittene 8.2.1 - Afsnit 8.2.3 betragtes rangtest for henholdsvis én, to og flere observationsrækker. Endelig vises det i et annekst til dette kapitel hvorledes nogle af beregningerne kan foretages ved hjælp af *Excel*.

## 8.1 Fortegnstestet

Gennemgangen af fortegnstestet er baseret på Eksempel 8.1.

### Eksempel 8.1

Andersen(1998) Konditallet før og efter et intensivt træningsprogram for 15 idrætsudøvere.

<i>idrætsudøver nr.</i>	<i>før</i>	<i>efter</i>	<i>differens</i>	<i> differens </i>	<i>rang</i>
1	71.52	72.12	+0.60	0.60	6
2	69.33	72.09	+2.76	2.76	10
3	75.27	74.98	-0.29	0.29	2
4	66.78	73.75	+6.67	6.67	13
5	71.30	71.32	+0.02	0.02	1
6	72.96	72.54	-0.42	0.42	4
7	75.13	75.72	+0.59	0.59	5
8	69.09	76.81	+7.72	7.72	15
9	71.82	73.05	+1.23	1.23	8
10	73.88	74.20	+0.32	0.32	3
11	75.11	73.45	-1.66	1.66	9
12	75.01	78.45	+3.44	3.44	11
13	67.66	74.81	+7.15	7.15	14
14	73.69	72.74	-0.95	0.95	7
15	74.34	78.03	+3.69	3.69	12

Det er her af interesse at afgøre om træningsprogrammet har haft en virkning på konditallet.

Problemstillingen i Eksempel 8.1 kender vi fra det parrede  $t$ -test i Afsnit 4.4. Hvis  $d_i$  er differensen mellem konditallet efter og før træningen for den  $i$ 'te person, undersøgte vi virkningen af træningen ved i modellen

$$M_0 : D_i \sim N(\delta, \sigma_D^2), \quad i = 1, \dots, n, \quad (8.1)$$

at teste hypotesen  $\delta = 0$  ved hjælp af

$$t(\mathbf{d}) = \frac{\bar{d}\sqrt{n}}{\sqrt{s_d^2}} \sim t(n-1).$$

I fortegnstestet droppes antagelsen om normalitet af  $D$ -erne og vi betragter hypotesen

$$H_0 : D_i \text{ har en kontinuert fordeling som er symmetrisk om } 0, \quad i = 1, \dots, n,$$

hvor som i (8.1) implicit antager, at  $D$ -erne er uafhængige og identisk fordelte. Under hypotesen  $H_0$  er

$$P(D_i > 0) = P(D_i < 0) = 1/2,$$

så hvis  $S_+$  betegner antallet af differenser med positivt fortegn har vi

$$S_+ \sim b(n, 1/2).$$

Lad  $s_+$  betegne det observerede antal af positive differenser. Da binomialfordelingen  $b(n, 1/2)$  er symmetrisk med middelværdi  $n/2$  gælder der, at hvis  $s_+ \leq n/2$  så er værdien  $n - s_+$  lige så kritisk for  $H_0$  som  $s_+$  mens værdierne  $0, 1, \dots, s_+ - 1$  og  $n - s_+ + 1, \dots, n$  er mere kritiske for  $H_0$  end  $s_+$ . Tilsvarende, hvis  $s_+ \geq n/2$  er værdien  $n - s_+$  lige så kritisk som  $s_+$  mens værdierne  $0, 1, \dots, n - s_+ - 1$  og  $s_+ + 1, \dots, n$  er mere kritiske for  $H_0$  end  $s_+$ . Idet der tages hensyn til at  $b(n, 1/2)$ -fordelingen er diskret beregnes testsandsynligheden i binomialfordelingen som

$$\varepsilon_F(\mathbf{d}) = b(s_+; n, 1/2) + 2 \sum_{i=0}^{\min(s_+, n-s_+)-1} b(i; n, 1/2), \quad (8.2)$$

det vil sige som sandsynligheden for det observerede udfald  $s_+$  plus to gange sandsynligheden for udfald der er mere kritiske end  $s_+$ .

**Bemærkning 8.1** I fortegnstestet betragtes kun observationer hvis differens er forskellig fra 0, det vil sige at  $n = \#\{i : d_i \neq 0\}$ . ▼

### Eksempel 8.1 (Fortsat)

Af tabellen side 8.1 ses, at det observerede antal positive differenser er  $s_+ = 11$ . Da alle differenser er forskellige fra 0 er  $n = 15$  og af (8.2) fås, at testsandsynligheden for  $H_0$  er

$$\varepsilon_F(\mathbf{d}) = b(11; 15, 1/2) + 2 \sum_{i=0}^3 b(i; 15, 1/2) = 0.0768.$$

Lad os tilsammenligning udføre beregningerne for det parrede  $t$ -test. Efter grafisk kontrol af at forudsætningerne for anvendelsen af testet er opfyldt, beregnes teststørrelsen til

$$t(\mathbf{d}) = 2.6058 \sim\sim t(14),$$

og den tilsvarende testsandsynlighed for hypotesen  $\delta = 0$  bliver

$$\varepsilon_t(\mathbf{d}) = 0.0207.$$

Konklusionen vedrørende hypotesen om at differenserne har en fordeling, der er symmetrisk omkring 0, er altså forskellig ved de to test. Fortegnstestet accepterer hypotesen mens det parrede  $t$ -test forkaster hypotesen. Denne forskel kommenteres i en forsættelse af eksemplet nedenfor.  $\square$

## 8.2 Rangtest

Testene i dette afsnit er alle baseret på rangen af observationerne i en observationsrække som defineret i Definition 1.1. Vi minder om, at hvis

$$x_{(1)} \leq \dots \leq x_{(i)} \leq \dots \leq x_{(n)}$$

betegner den ordnede stikprøve for en observationsrække  $x_1, \dots, x_n$  så defineres rangen af observationerne således:

$$\begin{aligned} \text{rang}(x_{(i)}) &= i, & \text{hvis } x_{(i-1)} < x_{(i)} < x_{(i+1)} \\ \text{rang}(x_{(i)}) &= \dots = \text{rang}(x_{(i+k-1)}) = i + (k-1)/2, & \text{hvis } x_{(i)} = \dots = x_{(i+k-1)} \end{aligned} \quad (8.3)$$

Rangen af observationen  $x_{(i)}$  er altså  $i$ , hvis  $x_{(i)}$  er den eneste observation med denne værdi, det vil sige hvis  $x_{(i-1)} < x_{(i)} < x_{(i+1)}$ . Hvis  $k$  observationer  $x_{(i)}, x_{(i+1)}, \dots, x_{(i+k-1)}$  er lige store, det vil sige hvis  $x_{(i)} = x_{(i+1)} = \dots = x_{(i+k-1)}$ , tildeles de alle rangen  $i + (k-1)/2$ , som er gennemsnittet af de  $k$  tal  $i, i+1, \dots, i+k-1$ . I engelsk sproget litteratur betegnes de her betragtede range som "midranks".

De *ordnede værdier* i stikprøven er de forskellige værdier  $y_1, y_2, \dots, y_m$ , som observationerne i stikprøven antager, ordnet efter størrelse, det vil sige

$$y_1 < y_2 < \dots < y_m. \quad (8.4)$$

For  $j = 1, \dots, m$  betegnes *antallet* af observationer med værdien  $y_j$  med  $a_j$ . Der er altså  $m$  forskellige værdier i stikprøven. Hvis  $m = n$  er alle observationerne forskellige og alle  $a$ -erne

har værdien 1. Hvis  $a_j = k > 1$  forekommer værdien  $y_j$  altså  $k$  gange blandt  $x$ -erne. Hvis  $x_{(i)} = x_{(i+1)} = \dots = x_{(i+k-1)} = y_j$  siges observationerne  $x_{(i)}, x_{(i+1)}, \dots, x_{(i+k-1)}$  at være *sammenfaldende* og de tilordnes ifølge ovenstående alle rangen  $i + (k - 1)/2$ . I engelsk sproget litteratur omtales sammenfaldende observationer som "ties".

I Afsnit 8.2.1 - Afsnit 8.2.3 betragtes rangtest for henholdsvis én, to og flere observationsrækker. I gennemgangen af rangtestene vil vi indledningsvis antage at alle observationerne er forskellige for senere i bemærkninger at angive modifikationer af testene i tilfælde af sammenfaldende observationer.

### 8.2.1 Wilcoxon's test for én observationsrække

For observationsrækken  $x_1, \dots, x_n$  betragter vi hypotesen

$$H_0 : X_i \text{ har en kontinuert fordeling symmetrisk om } 0, \quad i = 1, \dots, n,$$

hvor vi også antager at  $X$ -erne er uafhængige og identisk fordelte.

Et ikke-parametrisk test for  $H_0$  er *Wilcoxon's test for én observationsrække*, som undertiden omtales som *Mann-Whitney testet*. Testet er baseret på rangene af de numeriske værdier  $|x_1|, \dots, |x_n|$  af observationerne. Testet involverer kun de observationer  $x$  som er forskellige fra 0. Lad  $N$  betegne dette antal, det vil sige  $N = \#\{i : x_i \neq 0\}$  og lad  $r_i^+$  betegne rangen af  $|x_i|$ . Teststørrelsen, der betragtes er

$$W = \sum_{\{i: x_i > 0\}} r_i^+, \quad (8.5)$$

det vil sige summen af rangene for de positive observationer. Hvis hypotesen  $H_0$  er korrekt skal de positive og negative observationer falde tilfældigt mellem hinanden og have nogenlunde de samme numeriske værdier. Summen af rangene for de positive observationer skal derfor stort set være lig med summen af rangene for de negative observationer. Hvis summen af rangene af de positive observationer er meget større end summen af rangene for de negative observationer tyder dette på at der signifikant flere positive observationer end negative eller de positive observationer er signifikant større end de negative. Store værdier af  $W$  er derfor kritiske for  $H_0$ . Et symmetriargument viser tilsvarende at også små værdier er kritiske for  $H_0$ .

Hvis alle observationer, der er forskellige fra 0, er negative er  $W = 0$ , og hvis alle observationer, der er forskellige fra 0, er positive er  $W = N(N + 1)/2$ , som er summen af tallene  $1, 2, \dots, N$ . Det kan vises, at fordelingen af  $W$  ikke afhænger af  $X$ -ernes fælles fordeling, samt at

$$EW = \frac{N(N + 1)}{4}$$

og

$$Var W = \frac{N(N + 1)(2N + 1)}{24}.$$

For små værdier af  $N$  kan testsandsynligheden findes ved hjælp af tabeller, mens man for store værdier af  $N$  benytter at  $W$  er normalfordelt. Standardiseres fordelingen af  $W$  i sådanne tilfælde er

$$U_1(\mathbf{X}) = \frac{W - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} \approx N(0,1) \quad (8.6)$$

og testsandsynligheden for Wilcocons test kan beregnes som

$$\varepsilon_W(\mathbf{x}) = 2(1 - \Phi(|u_1(\mathbf{x})|)), \quad (8.7)$$

hvor  $u_1(\mathbf{x})$  er den observerede værdi af  $U_1(\mathbf{X})$  og hvor  $\Phi$  er fordelingsfunktionen for  $N(0,1)$ -fordelingen.

**Bemærkning 8.2** Hvis der er sammenfaldende observationer modificeres teststørrelsen  $U_1(\mathbf{X})$  til

$$U_1^*(\mathbf{X}) = \frac{W - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24} - \frac{1}{48} \sum_j (a_j^3 - a_j)}} \approx N(0,1) \quad (8.8)$$

og testsandsynligheden beregnes som

$$\varepsilon_W(\mathbf{x}) = 2(1 - \Phi(|u_1^*(\mathbf{x})|)). \quad (8.9)$$

Bemærk, at observerede værdier  $y_j$  for hvilke der er ikke er sammenfaldende observationer, det vil sige for hvilke  $a_j = 1$ , bidrager ikke til summen  $\sum_j (a_j^3 - a_j)$ , idet vi for sådanne observationer har  $a_j^3 - a_j = 1^3 - 1 = 0$ . ▼

### Eksempel 8.1 (Fortsat)

Vi udfører nu Wilcocons test på observationsrækken af differenser i dette eksempel. Af tabellen side 8.2 ses at der ikke er sammenfaldende observationer blandt de numeriske differenser som desuden også alle er forskellige fra 0, det vil sige  $N = 15$ . I tabellens sjette søjle ses rangene for de numeriske værdier af differenserne i femte søjle. Ved hjælp af fjerde og sjette søjle ses det, at summen af rangene for de positive differenser er  $w = 98$ . Formlerne (8.6) og (8.7) medfører, at

$$u_1(\mathbf{x}) = \frac{98 - \frac{15 \cdot 16}{4}}{\sqrt{\frac{15 \cdot 16 \cdot 31}{24}}} = 2.1582$$

og

$$\varepsilon_W(\mathbf{x}) = 0.0309,$$

så hypotesen  $H_0$  om at differenserne har en fordeling der er symmetrisk om 0 forkastes.

Vi har i dette eksempel fundet tre forskellige testsandsynligheder for hypotesen  $H_0$ , nemlig  $\varepsilon_F(\mathbf{d}) = 0.0768$ ,  $\varepsilon_W(\mathbf{x}) = 0.03090$  og  $\varepsilon_t(\mathbf{d}) = 0.0207$ . Det bør ikke undre. Fortegnstestet, som accepterer  $H_0$ , er baseret udelukkende på differensernes fortegn. I Wilcoxons test, der forkaster  $H_0$ , benyttes differensernes fortegn samt størrelsesforholdet af deres numeriske værdier, mens det parrede  $t$ -test, der forkaster  $H_0$ , beregnes ved hjælp af den empiriske middelværdi og varians for differenserne. Testene udnytter altså forskellige aspekter ved differenserne og desuden udnyttes den information, som differenserne indeholder, i forskellig grad.

I eksemplet her er det ikke urimeligt at lave ensidede test for  $H_0$ , idet man nok mest er interesseret i om træningen har en positiv indflydelse på konditallene, det vil sige om tallene efter træningen er signifikant større end tallene før træningen. Testes  $H_0$  ved ensidede test, forkastes  $H_0$  ved alle tre test, idet testsandsynligheden er det halve af testsandsynligheden ved testene ovenfor, da størrelserne  $s_+ = 11 (> 15/2)$ ,  $u_1 = 2.1582 (> 0)$  og  $t = 2.6058 (> 0)$  alle indikerer afvigelse i retning af at konditallene forøges ved træningen.  $\square$

### 8.2.2 Wilcoxons test for to observationsrækker

Gennemgangen er baseret på data i Eksempel 8.2.

#### Eksempel 8.2

Vi betragter igen data i Eksempel 4.2, som består af konditallene for 20 aktive og 17 ikke-aktive idrætsudøvere. Konditallene er gengivet i tabellen nedenfor hvor også observationernes range i den samlede stikprøve er vist.

<i>kondital</i>		<i>rang</i>	
<i>aktive</i>	<i>ikke-aktive</i>	<i>aktive</i>	<i>ikke-aktive</i>
68.9	56.0	13.5	1
75.2	61.8	29	2
74.3	64.1	28	4.5
72.9	64.9	21	6.5
72.0	65.2	18	8
63.9	66.3	3	10
76.3	66.9	33.5	11
76.3	68.9	33.5	13.5
75.4	70.6	31	16
66.0	70.8	9	17
68.4	72.4	12	20
64.1	73.1	4.5	22.5
73.1	73.9	22.5	25
64.9	74.1	6.5	26.5
73.4	74.1	24	26.5
76.2	75.3	32	30
79.4	78.7	36	35
69.4		15	
79.8		37	
72.1		19	
		428	275

Som i Eksempel 4.2 er vi interesserede i at undersøge om der er forskel på konditallene for de ikke-aktive og de aktive idrætsudøvere. □

Som tidligere lader vi  $x_{ij}$  betegne den  $j$ 'te observation i den  $i$ 'te observationsrække,  $j = 1, \dots, n_i$ ,  $i = 1, 2$ . Her betragter vi en ikke-parametrisk hypotese for de to observationsrækker, nemlig

$$H_0 : X_{ij} \text{ har en kontinuert fordeling } F, j = 1, \dots, n_i, i = 1, 2,$$

hvor vi også antager at  $X$ -erne er uafhængige. Med  $H_0$  ønsker vi at undersøge om samtlige  $n. = n_1 + n_2$  observationer kan betragtes som én observationsrække med den fælles fordeling  $F$ .

Et ikke-parametrisk test for denne hypotese er *Wilcoxon's test for to observationsrækker*, som tager udgangspunkt i rangene af observationerne i den samlede stikprøve. Lad  $R_{ij}$  betegne

rangen af  $X_{ij}$  i den samlede stikprøve og lad

$$R_{1.} = \sum_{j=1}^{n_1} R_{1j}$$

være summen af rangene i den første observationsrække. Den mindste værdi af  $R_{1.}$  fremkommer hvis de  $n_1$  observationer i den første observationsrække alle er mindre end observationerne i den anden række og i så tilfælde er  $R_{1.} = n_1(n_1 + 1)/2$  som er summen af tallene  $1, 2, \dots, n_1 - 1, n_1$ . Omvendt fremkommer den største værdi af  $R_{1.}$  ved at alle observationer i den første række er større end observationerne i den anden række og i så tilfælde er  $R_{1.} = (2n. - n_1 + 1)n_1/2$ , som er summen af tallene  $n_2 + 1, n_2 + 2, \dots, n_2 + n_1 - 1, n_2 + n_1$ . Det kan vises, at under  $H_0$  er middelværdien af  $R_{1.}$  gennemsnittet at den mindste og største værdi denne variabel kan antage, det vil sige

$$ER_{1.} = \frac{n_1(n. + 1)}{2}$$

samt at

$$Var R_{1.} = \frac{n_1 n_2 (n. + 1)}{12}.$$

Hvis  $R_{1.}$  er lille, er observationerne i den første række stort set alle mindre end observationerne i den anden række, hvilket harmonerer dårligt med at observationerne i de to rækker under  $H_0$  har samme fordeling. Tilsvarende antyder en stor værdi af  $R_{1.}$  en afvigelse fra  $H_0$ , da det svarer til at alle observationer i den første række stort set er større end observationerne i den anden række. Sammenfattende forkastes  $H_0$  for små og store værdier af  $R_{1.}$ .

For små værdier af  $n_1$  og  $n_2$  kan testsandsynligheden ved Wilcoxons test for to observationsrækker findes i tabeller. For store værdier af  $n_1$  og  $n_2$  kan det vises at  $R_{1.}$ 's fordeling kan approksimeres ved en normalfordeling. Standardiseres denne normalfordeling får vi at

$$U_2(\mathbf{X}) = \frac{R_{1.} - \frac{n_1(n. + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n. + 1)}{12}}} \approx N(0, 1). \quad (8.10)$$

Approksimationen kan vises at være tilfredsstillende hvis blot  $n_1 \geq 10$  og  $n_2 \geq 10$  og i sådanne tilfælde kan testsandsynligheden for Wilcoxons test for to observationsrækker beregnes som

$$\varepsilon(\mathbf{x}) = 2(1 - \Phi(|u_2(\mathbf{x})|)), \quad (8.11)$$

hvor  $u_2(\mathbf{x})$  er den observerede værdi af  $U_2(\mathbf{X})$  og  $\Phi$  er fordelingsfunktionen for  $N(0, 1)$ -fordelingen.

**Bemærkning 8.3** I tilfælde af sammenfaldende observationer betragtes følgende modifikation af  $U_2(\mathbf{X})$ :

$$U_2^*(\mathbf{X}) = \frac{R_{1.} - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12} \left[ 1 - \frac{\sum_j (a_j^3 - a_j)}{(n+1)n(n-1)} \right]}} \approx N(0, 1). \quad (8.12)$$

Den tilsvarende testsandsynlighed beregnes som

$$\varepsilon(\mathbf{x}) = 2(1 - \Phi(|u_2^*(\mathbf{x})|)). \quad (8.13)$$

Bemærk, at observerede værdier  $y_j$  for hvilke der er ikke er sammenfaldende observationer, det vil sige for hvilke  $a_j = 1$ , bidrager ikke til summen  $\sum_j (a_j^3 - a_j)$ , idet vi for sådanne observationer har  $a_j^3 - a_j = 1^3 - 1 = 0$ . ▼

### Eksempel 8.2 (Fortsat)

Af tabellen side 8.8 ses vi for disse data har sammenfaldende observationer, idet de seks værdier 64.1, 64.9, 68.9, 73.1, 74.1 og 76.3 alle er observeret to gange. Vi har derfor, at

$$\sum_j (a_j^3 - a_j) = 6(2^3 - 2) = 36.$$

Af tredje søjle i tabellen ses at  $r_{1.} = 428$ . Da  $n_1 = 20$  og  $n_2 = 17$  er  $n = 37$  og ved hjælp af (8.12) fås, at

$$u_2^*(\mathbf{x}) = \frac{428 - \frac{20 \cdot 38}{2}}{\sqrt{\frac{20 \cdot 17 \cdot 38}{12} \left[ 1 - \frac{36}{38 \cdot 37 \cdot 36} \right]}} = 1.4634$$

og af (8.13) fås testsandsynligheden

$$\varepsilon(\mathbf{x}) = 2(1 - \Phi(1.4634)) = 0.143.$$

Der er altså - overraskende nok - ikke forskel på konditallene for de ikke-aktive og de aktive. Samme konklusion nåede vi frem til i Afsnit 4.4, hvor vi analyserede data ved hjælp af modellen for to normalfordelte observationsrækker. Her blev hypotesen om ens varianser accepteret ved et  $F$ -test med en testsandsynlighed på 0.464 mens hypotesen om ens middelværdier blev accepteret ved et  $t$ -test med en testsandsynlighed på 0.110. □

### 8.2.3 Kruskal-Wallis test

Gennemgangen af Kruskal-Wallis test, som er den ikke-parametriske analog til ensidet variansanalyse, er baseret på Eksempel 8.3.

#### Eksempel 8.3

Vi betragter igen data i Eksempel 4.5 vedrørende resultaterne af pigernes længdespring ved atletikstævnet for 1. års studerende ved Institut for Idræt, Københavns Universitet i årene 1998 - 2000.

1998		1999		2000	
<i>længde</i>	<i>rang</i>	<i>længde</i>	<i>rang</i>	<i>længde</i>	<i>rang</i>
3.72	19	4.32	35	3.96	28
3.65	13	3.79	23	3.43	4
3.90	25	3.53	6	4.30	34
3.74	21	3.54	7	4.22	31.5
3.32	3	4.27	33	3.56	8.5
4.22	31.5	3.75	22	3.70	17.5
3.58	10	4.21	30	3.70	17.5
4.56	36	3.66	16	3.56	8.5
3.65	13	4.58	37		
2.99	1	3.73	20		
3.91	26.5	5.18	38		
3.65	13	3.00	2		
3.65	13	3.91	26.5		
3.88	24	3.52	5		
3.65	13				
4.20	29				
<i>sum</i>	291	<i>sum</i>	300.5	<i>sum</i>	149.5

Som tidligere er vi interesseret i at afgøre om længden af springene er uafhængig af årene.  $\square$

Vi lader  $x_{ij}$  betegne den  $j$ 'te observation i den  $i$ 'te observationsrække,  $j = 1, \dots, n_i$ ,  $i = 1, \dots, k$ . Her betragter vi en ikke-parametrisk hypotese for  $k$  observationsrækker, nemlig

$$H_0 : X_{ij} \text{ har en kontinuert fordeling } F, j = 1, \dots, n_i, i = 1 \dots k,$$

hvor vi også antager at  $X$ -erne er uafhængige. Vi ønsker altså at undersøge om samtlige  $n. = n_1 + \dots + n_k$  observationer kan betragtes som én observationsrække.

Et ikke-parametrisk test for denne hypotese er *Kruskal-Wallis test for  $k$  observationsrækker*, som tager udgangspunkt i rangene af observationerne i den samlede stikprøve. Lad  $R_{ij}$  betegne rangen af  $X_{ij}$  i den samlede stikprøve og lad

$$R_i = \sum_{j=1}^{n_i} R_{ij}$$

være summen af rangene i den  $i$ 'te observationsrække. Den mindste værdi af  $R_i$  fremkommer hvis de  $n_i$  observationer i den  $i$ 'te observationsrække alle er mindre end observationerne i de øvrige rækker og i så tilfælde er  $R_i = n_i(n_i + 1)/2$  som er summen af tallene  $1, 2, \dots, n_i - 1, n_i$ . Omvendt fremkommer den største værdi af  $R_i$  ved at alle observationer i den  $i$ 'te række er større end observationerne i de øvrige rækker og i så tilfælde er  $R_i = (2n - n_i + 1)n_i/2$ , som er summen af de  $n_i$  tal  $n - n_i + 1, n - n_i + 2, \dots, n - 1, n$ . Det kan vises, at under  $H_0$  er middelværdien af  $R_i$  gennemsnittet af den mindste og største værdi denne variabel kan antage, det vil sige

$$ER_i = \frac{n_i(n + 1)}{2}$$

og dermed at gennemsnittet  $\bar{R}_i = R_i/n_i$  har middelværdi

$$E\bar{R}_i = \frac{n + 1}{2} = \bar{R}..,$$

som er gennemsnittet af de  $n$  tal  $1, 2, \dots, n - 1, n$ . Under  $H_0$  må det forventes at de  $k$  ranggennemsnit  $\bar{R}_i$  varierer tilfældigt omkring  $\bar{R}..$  og Kruskal-Wallis introducerede derfor følgende teststørrelse for  $H_0$ :

$$KW(\mathbf{X}) = \frac{12}{n(n + 1)} \sum_{i=1}^k n_i (\bar{R}_i - \bar{R}..)^2 \quad (8.14)$$

og viste at denne teststørrelse approksimativt er  $\chi^2(k - 1)$ -fordelt for moderate værdier af alle stikprøvestørrelserne  $n_i$ . Da store værdier af  $KW(\mathbf{X})$  er kritiske for  $H_0$  beregnes testsandsynligheden som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(k-1)}(KW(\mathbf{x})), \quad (8.15)$$

hvor  $KW(\mathbf{x})$  er den observerede værdi af  $KW(\mathbf{X})$ .

Hvis  $k = 2$  kan det vises, at Kruskal-Wallis testet er ækvivalent med Wilcoxon's test for to observationsrækker, idet der da gælder, at

$$KW(\mathbf{X}) = U_2(\mathbf{X})^2 \quad (8.16)$$

samt af kvadratet af en  $N(0, 1)$ -fordelt stokastisk variabel er  $\chi^2(1)$ -fordelt.

Teststørrelsen i (8.14) beregnes let i hånden ud fra rangsummerne i de  $k$  observationsrækker idet

$$\sum_{i=1}^k n_i (\bar{R}_i - \bar{R}..)^2 = \sum_{i=1}^k \frac{R_i^2}{n_i} - \frac{R_{..}^2}{n}. \quad (8.17)$$

**Bemærkning 8.4** I tilfælde af sammenfaldende observationer erstattes teststørrelsen  $KW(\mathbf{X})$  med

$$KW^*(\mathbf{X}) = \frac{KW(\mathbf{X})}{1 - \frac{\sum_j (a_j^3 - a_j)}{(n+1)n(n-1)}} \quad (8.18)$$

og testsandsynligheden beregnes som

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(k-1)}(KW^*(\mathbf{x})). \quad (8.19)$$

Bemærk igen, at observerede værdier  $y_j$  for hvilke der er ikke er sammenfaldende observationer, det vil sige for hvilke  $a_j = 1$ , bidrager ikke til summen  $\sum_j (a_j^3 - a_j)$ , idet vi for sådanne observationer har  $a_j^3 - a_j = 1^3 - 1 = 0$ . ▼

### Eksempel 8.3 (Fortsat)

Af tabellen side 8.11 ses at der er sammenfaldende observationer, idet de fire værdier 3.56, 3.70, 3.91 og 4.22 alle er observeret to gange mens værdien 3.65 er observeret fem gange. Vi har derfor at

$$\sum_j (a_j^3 - a_j) = 4(2^3 - 2) + (5^3 - 5) = 24 + 120 = 144.$$

Da observationsantallene og rangsummerne - ifølge tabellen - er

$i$	$n_i$	$r_i$
1	16	291.0
2	14	300.5
3	8	149.5
<i>sum</i>	38	741.0

finder vi ved hjælp af (8.14) og (8.17), at

$$KW(\mathbf{x}) = \frac{12}{38 \cdot 39} 86.8616 = 0.7033$$

og dermed af (8.18) at

$$KW^*(\mathbf{x}) = \frac{0.7033}{1 - \frac{144}{39 \cdot 38 \cdot 37}} = 0.7052.$$

Af (8.19) fås, at

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(2)}(0.7052) = 0.7029.$$

Hypotesen  $H_0$  accepteres, så som i Afsnit 4.5 finder vi, at fordelingen af længderne i de tre år kan antages at være identiske. Der blev tallene analyseret som tre normalfordelte observationsrækker og hypotesen om ens varianser blev accepteret ved et Bartlett test med testsandsynlighed

0.1941 mens hypotesen om ens middelværdier blev accepteret ved et  $F$ -test med testsandsynlighed 0.5865.  $\square$

### Eksempel 8.2 (Fortsat)

Af tabellen side 8.8 ses, at for disse data er observationsantal og rangsummer:

$i$	$n_i$	$r_i$
1	20	428
2	17	275
<i>sum</i>	37	703

Fra tidligere ved vi at  $\sum_j (a_j^3 - a_j) = 36$ . Formlerne (8.14), (8.17) og (8.18) medfører, at Kruskal-Wallis teststørrelsen er

$$KW^*(\mathbf{x}) = 2.1415.$$

Af (8.19) fås, at testsandsynligheden for  $H_0$  er

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(1)}(2.1415) = 0.143,$$

altså samme testsandsynlighed som ved Wilcoxon's test for to observationsrækker. Dette skyldes at  $\sqrt{2.1415} = 1.4634$  samt bemærkningen efter formel (8.16).  $\square$

## Anneks til Kapitel 8

### Beregninger i Excel

Excel har ikke specielle dialogbokse til beregning af ikke-parametriske test. Hvis der ikke er sammenfaldende observationer beregnes de dog let ved hjælp af funktionen PLADS. Med notationen på side 8.4 er definitionen af denne funktion

$$\text{PLADS}(x_{(i)}) = i, \quad \text{hvis } x_{(i-1)} < x_{(i)} < x_{(i+1)}$$

$$\text{PLADS}(x_{(i)}) = \dots = \text{PLADS}(x_{(i+k-1)}) = i, \quad \text{hvis } x_{(i)} = \dots = x_{(i+k-1)}.$$

Hvis der ikke er sammenfaldende observationer, ses det at  $\text{rang}(x_{(i)}) = \text{PLADS}(x_{(i)})$  hvis  $x_{(i-1)} < x_{(i)} < x_{(i+1)}$ , mens  $\text{rang}(x_{(i)}) = \text{PLADS}(x_{(i)}) + (k-1)/2$ , hvis  $x_{(i)} = \dots = x_{(i+k-1)}$  i tilfælde af sammenfaldende observationer.

Excel har en dialogboks Rang og fraktil, der beregner fraktiler som funktionen PLADS, idet dog observationerne ordnes i aftagende rækkefølge, så denne dialogboks er ikke til megen hjælp her.

Vi indskrænker os her til at vise beregningerne i et eksempel hvor der ikke er sammenfaldende observationer.

#### Eksempel 8.1 (Fortsat)

Beregningerne af fortegnstestet ses nedenfor:

	A	B	C	D	E	F	G	H
1	Eksempel 8.1							
2								
3	idrætsudøver nr.	før	efter	differens	fortegn			
4	1	71,52	72,12	0,60	1		S+	11
5	2	69,33	72,09	2,76	1		epsilon	0,076813
6	3	75,27	74,98	-0,29	0			
7	4	66,78	73,45	6,67	1			
8	5	71,3	71,32	0,02	1			
9	6	72,96	72,54	-0,42	0			
10	7	75,13	75,72	0,59	1			
11	8	69,09	76,81	7,72	1			
12	9	71,82	73,05	1,23	1			
13	10	73,88	74,2	0,32	1			
14	11	75,11	73,45	-1,66	0			
15	12	75,01	78,45	3,44	1			
16	13	67,66	74,81	7,15	1			
17	14	73,69	72,74	-0,95	0			
18	15	74,34	78,03	3,69	1			

Værdierne for konditallene før og efter træningen findes i cellerne B4:C18, mens differenserne for tallene efter og før er beregnet i D4:D18 ved i D4 at beregne

$$C4 - B4$$

og oprette analoge formler i cellerne D5:D18. Herefter beregnes i E4:E18 en variabel, som er 1, hvis fortegnet af differensen er positivt, og 0, hvis fortegnet er negativt. Indholdet af E4 beregnes som

$$= \text{HVIS}(D4 > 0; 1; 0)$$

og analoge formler oprettes i E5:E18. Herefter kan antallet af positive differenser,  $s_+$ , i H4 beregnes som

$$= \text{SUM}(E4 : E18)$$

og testsandsynligheden  $\varepsilon_F(\mathbf{x})$  i H5 som

$$= \text{BINOMIALFORDELING}(11; 15; 0,5; \text{FALSK}) + 2 * \text{BINOMIALFORDELING}(3; 15; 0,5; \text{SAND}),$$

det vil sige  $\varepsilon_F(\mathbf{x}) = b(s_+; n, 1/2) + 2 \sum_{i=0}^{\min(s_+, n-s_+)} b(i; n, 1/2)$ .

Wilcoxon's test for én observationsrække beregnes på differenserne som vist nedenfor:

	A	B	C	D
1				
2	differenser	abs differenser	rang	rang af positive
3	0,60	0,6	6	6
4	2,76	2,76	10	10
5	-0,29	0,29	2	0
6	6,67	6,67	13	13
7	0,02	0,02	1	1
8	-0,42	0,42	4	0
9	0,59	0,59	5	5
10	7,72	7,72	15	15
11	1,23	1,23	8	8
12	0,32	0,32	3	3
13	-1,66	1,66	9	0
14	3,44	3,44	11	11
15	7,15	7,15	14	14
16	-0,95	0,95	7	0
17	3,69	3,69	12	12
18			W:	98
19				
20				
21	u:	2,15825497		
22	tosidet:	0,030907903		
23	ensidet:	0,015453951		

De numeriske eller absolutte værdier i B3:B17 af differenserne i A3:A17 beregnes ved i B3 at indsætte formlen

$$= \text{ABS}(A3) \quad (= |d_1|)$$

og oprette analoge formler i B4:B17. Rangen af de numeriske værdier af differenserne,  $r_i^+$ , er beregnet i C3:C17 ved hjælp af funktionen PLADS ved i C3 at beregne

$$= \text{PLADS}(B3; \$B\$3 : \$B\$17; 1) \quad (= r_1^+)$$

og dernæst oprette analoge formler i C4:C17. Herefter beregnes i D3:D17 en variabel hvis værdi er rangen af den numeriske værdi, hvis differensen er positiv, og 0, hvis differensen er negativ. I D3 indtastes formlen

$$= \text{HVIS}(A3 > 0; C3; 0)$$

og analoge formler oprettes i D4:D17. Herefter findes værdien af  $W$  i D18 som

$$= \text{SUM}(D3 : D17) \quad (= \sum_i r_i^+).$$

Teststørrelsen  $u_1(\mathbf{x})$  i (8.6) beregnes i B21 som

$$= (D18 - 15 * 16 / 4) / \text{KVRØD}(15 * 16 * 31 / 24)$$

og den tilsvarende testsandsynlighed  $\varepsilon_W(\mathbf{x})$  i (8.7) beregnes i B22 som

$$= 2 * (1 - \text{NORMFORDELING}(B21; 0; 1; \text{SAND})).$$

Testsandsynligheden for det ensidede test i B23 er blot halvdelen af  $\varepsilon_W(\mathbf{x})$ . □

## Hovedpunkter til Kapitel 8

### Én observationsrække

For observationsrækken  $x_1, \dots, x_n$  betragter vi hypotesen

$$H_0 : X_i \text{ har en kontinuert fordeling symmetrisk om } 0, \quad i = 1, \dots, n,$$

hvor vi også antager at  $X$ -erne er uafhængige og identisk fordelte.

### Fortegnstestet

Teststørrelse:  $S_+$ , antallet af positive observationer.

Testsandsynlighed:

$$\varepsilon_F(\mathbf{x}) = b(s_+; n, 1/2) + 2 \sum_{i=0}^{\min(s_+, n-s_+)-1} b(i; n, 1/2).$$

### Wilcoxon's test

Testet involverer kun de observationer  $x$  som er forskellige fra 0. Lad  $N = \#\{i : x_i \neq 0\}$ .

Teststørrelse: For store værdier af  $N$  betragtes

$$U_1(\mathbf{X}) = \frac{W - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}.$$

Her er

$$W = \sum_{\{i: x_i > 0\}} r_i^+,$$

hvor  $r_i^+$  er rangen af  $|x_i|$  i observationsrækken af de numeriske værdier  $|x_1|, \dots, |x_n|$ .

Testsandsynlighed:

$$\varepsilon_W(\mathbf{x}) = 2(1 - \Phi(|u_1(\mathbf{x})|)),$$

Hvis der er sammenfaldende observationer betragtes følgende modifikation:

Teststørrelse:

$$U_1^*(\mathbf{X}) = \frac{W - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24} - \frac{1}{48} \sum_j (a_j^3 - a_j)}}.$$

Testsandsynlighed:

$$\varepsilon_W(\mathbf{x}) = 2(1 - \Phi(|u_1^*(\mathbf{x})|)).$$

### To observationsrækker

Lad  $x_{ij}$  betegne den  $j$ 'te observation i den  $i$ 'te observationsrække,  $j = 1, \dots, n_i, i = 1, 2$ .

Hypotese:

$$H_0 : X_{ij} \text{ har en kontinuert fordeling } F, j = 1, \dots, n_i, i = 1, 2,$$

hvor  $X$ -erne er uafhængige.

### Wilcoxon's test

Teststørrelse: Hvis  $n_1 \geq 10$  og  $n_2 \geq 10$  betragtes

$$U_2(\mathbf{X}) = \frac{R_{1\cdot} - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}},$$

hvor  $n = n_1 + n_2$  og hvor

$$R_{1\cdot} = \sum_{j=1}^{n_1} R_{ij}$$

er summen af rangene i den første observationsrække, idet  $R_{ij}$  betegner rangen af  $X_{ij}$  i den samlede stikprøve.

Testsandsynlighed:

$$\varepsilon(\mathbf{x}) = 2(1 - \Phi(|u_2(\mathbf{x})|)),$$

hvor  $u_2(\mathbf{x})$  er den observerede værdi af  $U_2(\mathbf{X})$  og  $\Phi$  er fordelingsfunktionen for  $N(0, 1)$ -fordelingen.

I tilfælde af sammenfaldende observationer betragtes følgende modifikationer:

Teststørrelse:

$$U_2^*(\mathbf{X}) = \frac{R_{1\cdot} - \frac{n_1(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12} \left[ 1 - \frac{\sum_j (a_j^3 - a_j)}{(n+1)n(n-1)} \right]}}$$

Testsandsynlighed:

$$\varepsilon(\mathbf{x}) = 2(1 - \Phi(|u_2^*(\mathbf{x})|)).$$

### Flere observationsrækker

Lad  $x_{ij}$  betegne den  $j$ 'te observation i den  $i$ 'te observationsrække,  $j = 1, \dots, n_i, i = 1, \dots, k$ .

Hypotese:

$$H_0 : X_{ij} \text{ har en kontinuert fordeling } F, j = 1, \dots, n_i, i = 1 \dots k,$$

hvor  $X$ -erne er uafhængige.

**Kruskal-Wallis test**

Teststørrelse:

$$KW(\mathbf{X}) = \frac{12}{n.(n. + 1)} \sum_{i=1}^k n_i (\bar{R}_i. - \bar{R}..)^2,$$

hvor  $n. = n_1 + \dots + n_k$  og hvor

$$\bar{R}_i. = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}$$

og

$$\bar{R}.. = \frac{1}{n.} \sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij}$$

betegner henholdsvis gennemsnittet af rangene i det  $i$ 'te række og det totale gennemsnit af rangene  $R_{ij}$  af  $X_{ij}$  i den samlede stikprøve.

Testsandsynlighed:

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(k-1)}(KW(\mathbf{x})),$$

Beregningsformel:

$$\sum_{i=1}^k n_i (\bar{R}_i. - \bar{R}..)^2 = \sum_{i=1}^k \frac{R_i.^2}{n_i} - \frac{R..^2}{n.},$$

hvor  $R_i.$  og  $R..$  er henholdsvis summen af rangene i den  $i$ 'te række og totalsummen.

I tilfælde af sammenfaldende observationer betragtes følgende modifikationer:

Teststørrelse:

$$KW^*(\mathbf{X}) = \frac{KW(\mathbf{X})}{1 - \frac{\sum_j (a_j^3 - a_j)}{(n. + 1)n.(n. - 1)}}$$

Testsandsynlighed:

$$\varepsilon(\mathbf{x}) = 1 - F_{\chi^2(k-1)}(KW^*(\mathbf{x})).$$

## Opgaver til Kapitel 8

**Opgave 8.1** Beregn testsandsynligheden for  $-2 \ln Q$ -testet for hypotesen  $\boldsymbol{\pi} = (1/2, 1/2)$  i modellen

$$(X_1, X_2) \sim m(15, \boldsymbol{\pi})$$

på grundlag af observationen  $(x_1, x_2) = (11, 4)$  og sammenlign denne med  $\varepsilon_F$  på side 8.3.

**Opgave 8.2** Betragt tallene i Opgave 4.11. Undersøg ved hjælp af fortegnstestet og Wilcoxon test for én observationsrække om vægttabet kan antages at være 6.5 kg ved at betragte  $x_1, \dots, x_{12}$ , hvor  $x_i = d_i - 6.5$ ,  $i = 1, \dots, 12$ .

**Opgave 8.3** Betragt data i Opgave 4.15 og undersøg ved hjælp af Wilcoxon test for to observationsrækker om observationerne i grupperne 2 og 3 kan antages at have samme fordeling.

**Opgave 8.4** Undersøg for såvel piger som for drenge ved hjælp af Kruskal-Wallis's test om resultaterne i kuglestød i Opgave 4.12 kan antages at have en fordeling, der er uafhængig af årene.

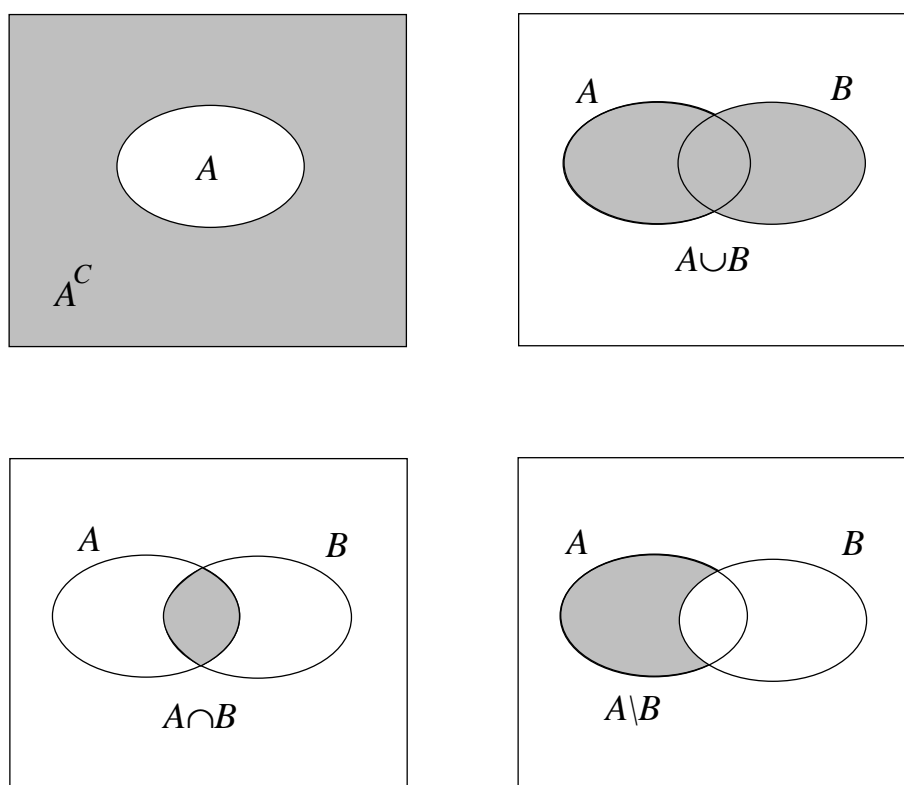


# A Forskellige matematiske begreber

## A.1 Notation fra mængdelæren

Hvis  $A$  og  $E$  er to mængder, er  $A$  en *delmængde* af  $E$ , kort  $A \subseteq E$ , hvis alle elementer i  $A$  også er elementer i  $E$ , det vil sige

$$e \in A \Rightarrow e \in E.$$



**Figur A.1** Illustration af mængderne  $A^C$ ,  $A \cup B$ ,  $A \cap B$  og  $A \setminus B$ .

Hvis  $A \subseteq E$ , er *komplementærmængden* til  $A$  (inden for  $E$ ) mængden

$$A^C = \{e \in E : e \notin A\}$$

Hvis  $A$  og  $B$  er delmængder af  $E$ , er *foreningsmængden* af  $A$  og  $B$  mængden

$$A \cup B = \{e \in E : e \in A \text{ og/eller } e \in B\},$$

*fællesmængden* af  $A$  og  $B$  er mængden

$$A \cap B = \{e \in E : e \in A \text{ og } e \in B\}$$

og *mængdedifferensen* mellem  $A$  og  $B$  er

$$A \setminus B = \{e \in A : e \notin B\} = (A \cap B^C).$$

Hvis  $A_1, A_2, \dots, A_n, \dots$  er en følge af delmængder af  $E$ , omtales mængden

$$\bigcup_{i=1}^n A_i = A_1 \cup \dots \cup A_n = \{e \in E : e \in A_i \text{ for mindst et } i = 1, \dots, n\}$$

som en *endelig foreningsmængde* og mængden

$$\bigcap_{i=1}^n A_i = A_1 \cap \dots \cap A_n = \{e \in E : e \in A_i \text{ for alle } i = 1, \dots, n\}$$

som en *endelig fællesmængde*, mens mængderne

$$\bigcup_{i=1}^{\infty} A_i = \{e \in E : e \in A_i \text{ for mindst et } i = 1, 2, \dots\}$$

og

$$\bigcap_{i=1}^{\infty} A_i = \{e \in E : e \in A_i \text{ for alle } i = 1, 2, \dots\}$$

kaldes henholdsvis en *tællelig foreningsmængde* og en *tællelig fællesmængde*.

Den *tomme mængde*  $\emptyset$  er mængden uden elementer. Den opfattes som en delmængde af enhver anden mængde.

To delmængder  $A$  og  $B$  af  $E$  siges at være *disjunkte*, hvis

$$A \cap B = \emptyset,$$

og elementerne i en følge af delmængder,  $A_1, A_2, \dots$ , siges at være *parvis disjunkte*, hvis

$$A_i \cap A_j = \emptyset, \quad \text{hvis } i \neq j, \quad i, j = 1, 2, \dots$$

## A.2 Rækker

Hvis  $a_1, a_2, \dots, a_n, \dots$  er en uendelig følge af reelle tal kaldes

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + \dots + a_n + \dots$$

en *uendelig række*.

Rækkens  $n$ 'te led er  $a_n$  og rækkens  $n$ 'te afsnitssum er

$$s_n = a_1 + a_2 + \dots + a_n.$$

Hvis  $s_n \rightarrow s$ , når  $n \rightarrow \infty$ , er rækken *konvergent med sum  $s$* , hvilket vi kort skriver

$$s = \sum_{n=1}^{\infty} a_n,$$

ellers kaldes rækken *divergent*.

Hvis  $a_n = 0$  for  $n > i$  kaldes rækken en *endelig række* med  $i$  led. Endelige rækker er konvergente da  $s_n = s_i$  for  $n \geq i$ .

(Undertiden har man - som i to af eksemplerne nedenfor - en følge startende i 0,  $a_0, a_1, a_2, \dots, a_n, \dots$ . Rækken  $\sum_{n=0}^{\infty} a_n$  er da konvergent med sum  $s$ , hvis  $s_{n+1} \rightarrow s$ , når  $n \rightarrow \infty$ , hvor  $s_{n+1} = a_0 + a_1 + a_2 + \dots + a_n$ .)

Rækken  $\sum_{n=1}^{\infty} a_n$  siges at være *absolut konvergent*, hvis rækken af absolutte (numeriske) værdier  $\sum_{n=1}^{\infty} |a_n|$  er konvergent.

Der gælder, at

$$\sum_{n=1}^{\infty} |a_n| \text{ konvergent} \quad \Rightarrow \quad \sum_{n=1}^{\infty} a_n \text{ konvergent},$$

det vil sige, at absolut konvergens medfører konvergens.

### Eksempler

*Endelige rækker:*

Hvis  $a$  og  $b$  er reelle tal og  $i$  et helt positivt tal er

$$(a+b)^i = \sum_{n=0}^i \binom{i}{n} a^n b^{i-n}, \quad (\text{binomialrækken})$$

hvor

$$\binom{i}{n} = \frac{i!}{n!(i-n)!} = \frac{i \cdot (i-1) \cdot \dots \cdot 2 \cdot 1}{(n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1)((i-n) \cdot (i-n-1) \cdot \dots \cdot 2 \cdot 1)}. \quad (\text{A.1})$$

Hvis  $q \neq 1$  er et reelt tal og  $i$  et helt positivt tal er

$$\sum_{n=0}^i q^n = 1 + q + \cdots + q^i = \frac{1 - q^{i+1}}{1 - q} \quad (\text{endelig kvotientrække}) \quad (\text{A.2})$$

Uendelige rækker:

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1 - q}, \quad \text{hvis } |q| < 1 \quad (\text{uendelig kvotientrække}) \quad (\text{A.3})$$

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x, \quad \text{for } x \in \mathbb{R}, \quad (\text{eksponentialrækken}) \quad (\text{A.4})$$

$$\sum_{n=1}^{\infty} \frac{x^n}{n} = -\ln(1 - x), \quad \text{for } |x| < 1, \quad (\text{logaritmisk række}) \quad (\text{A.5})$$

### Regnerækker for uendelige rækker

Hvis

$$A = \sum_{n=1}^{\infty} a_n \quad \text{og} \quad B = \sum_{n=1}^{\infty} b_n$$

er

$$A + B = \sum_{n=1}^{\infty} (a_n + b_n) \quad (\text{A.6})$$

og hvis  $k$  er en konstant er

$$kA = \sum_{n=1}^{\infty} ka_n. \quad (\text{A.7})$$

Hvis rækkerne  $A = \sum_{n=1}^{\infty} a_n$  og  $B = \sum_{n=1}^{\infty} b_n$  er absolut konvergente og

$$c_n = a_1 b_n + a_2 b_{n-1} + \cdots + a_n b_1$$

er  $\sum_{n=1}^{\infty} c_n$  absolut konvergent og

$$AB = \sum_{n=1}^{\infty} c_n. \quad (\text{A.8})$$

## A.3 Dobbeltintegraler og partiel differentiation

I forbindelse med beregninger relateret til kontinuerte to-dimensionale stokastiske vektorer er de matematiske begreber *dobbeltintegraler* og *partiel differentiation* vigtige. Begreberne omtales henholdsvis i Afsnit A.3.1 og A.3.2 nedenfor.

### A.3.1 Dobbeltintegraler

Lad  $f$  være en funktion af to variable og lad  $A = ]a, b[ \times ]c, d[$  være en delmængde af  $\mathbf{R}^2$ , hvor  $-\infty \leq a < b \leq \infty$  og  $-\infty \leq c < d \leq \infty$ . Værdien af *dobbeltintegralet*

$$I = \iint_A f(x_1, x_2) dx_2 dx_1 = \int_a^b \int_c^d f(x_1, x_2) dx_2 dx_1$$

beregnes da på følgende måde: Hvis  $g(x_1)$  betegner værdien af det inderste integral, det vil sige

$$g(x_1) = \int_c^d f(x_1, x_2) dx_2,$$

er

$$I = \int_a^b \int_c^d f(x_1, x_2) dx_2 dx_1 = \int_a^b g(x_1) dx_1.$$

Værdien af et dobbeltintegral bestemmes altså ved at integrere to gange. Først integreres funktionen  $f(x_1, x_2)$  med  $x_1$  fastholdt med hensyn til  $x_2$ , hvorefter resultatet af denne integration  $g(x_1)$  integreres med hensyn til  $x_1$ .

For alle de funktioner  $f$ , som vi skal integrere, kan dobbeltintegralet også beregnes ved at *ombytte integrationsordenen*, det vil sige som

$$I = \int_c^d \int_a^b f(x_1, x_2) dx_1 dx_2 = \int_c^d h(x_2) dx_2,$$

hvor

$$h(x_2) = \int_a^b f(x_1, x_2) dx_1.$$

### A.3.2 Partiel differentiation

Lad  $F(x_1, x_2)$  være en funktion af to variable. Lad  $x_2$  være fast og antag, at funktionen  $G_{x_2}$  af den variable  $x_1$  givet ved

$$G_{x_2}(x_1) = F(x_1, x_2)$$

er differentiabel. Den *partielt afledede* af  $F$  med hensyn til  $x_1$  defineres da som den afledede af  $G_{x_2}$  med hensyn til  $x_1$ , hvilket skrives således

$$\frac{\partial}{\partial x_1} F(x_1, x_2) = \frac{d}{dx_1} G_{x_2}(x_1).$$

Partiel differentiation angives altså ved hjælp af symbolet  $\partial$ .

Tilsvarende defineres den partielt afledede af  $F$  med hensyn til  $x_2$  som

$$\frac{\partial}{\partial x_2} F(x_1, x_2) = \frac{d}{dx_2} H_{x_1}(x_2),$$

hvis funktionen

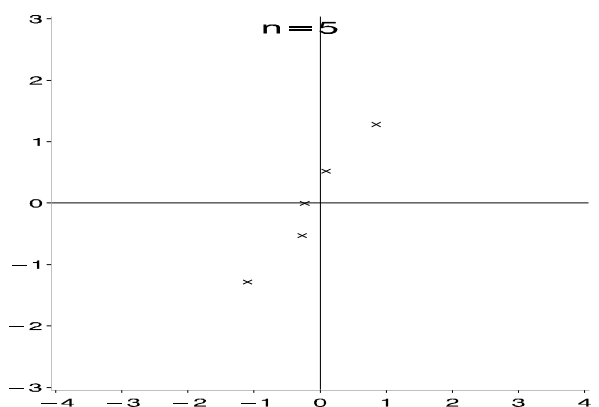
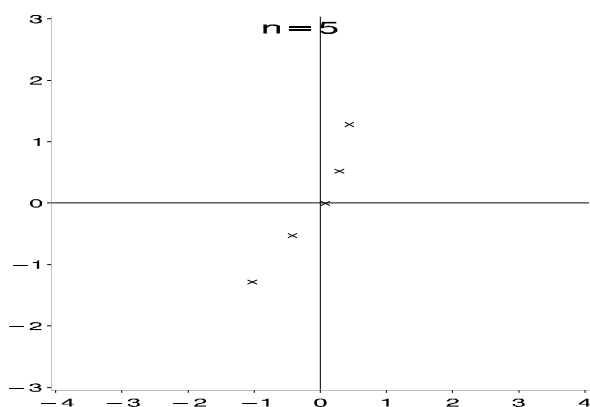
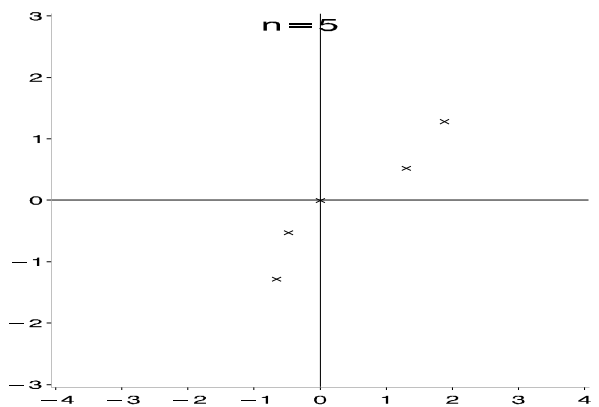
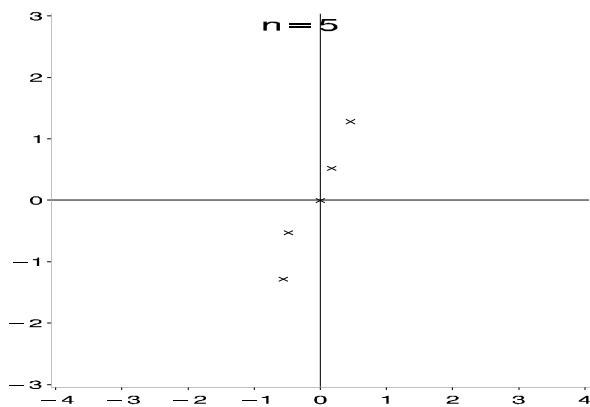
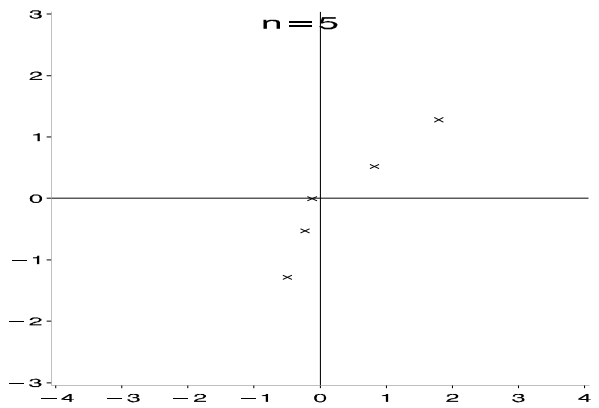
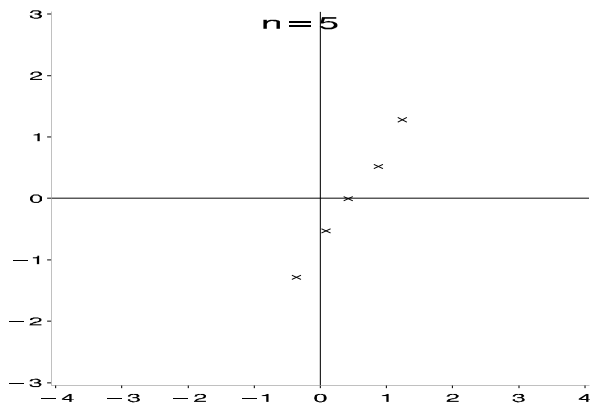
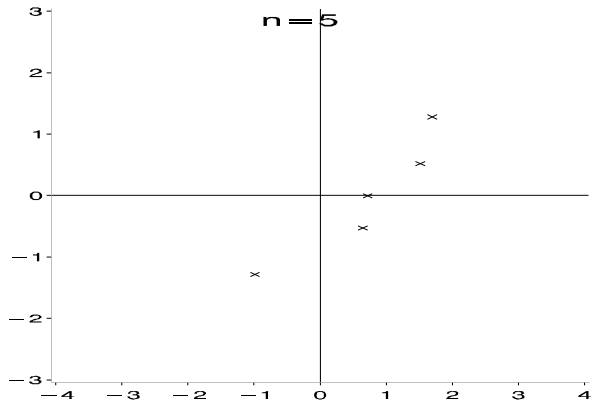
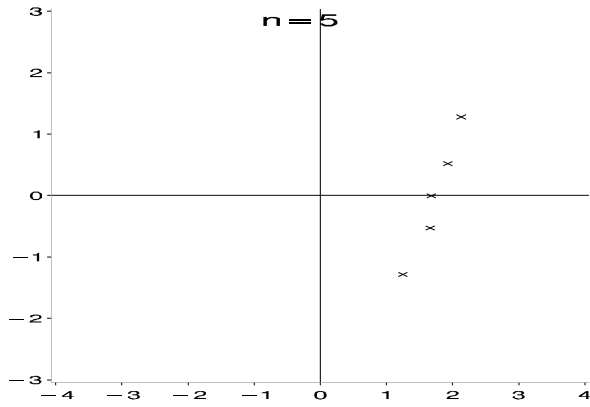
$$H_{x_1}(x_2) = F(x_1, x_2)$$

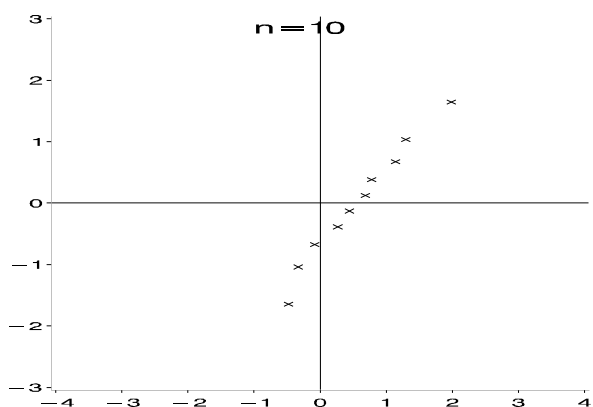
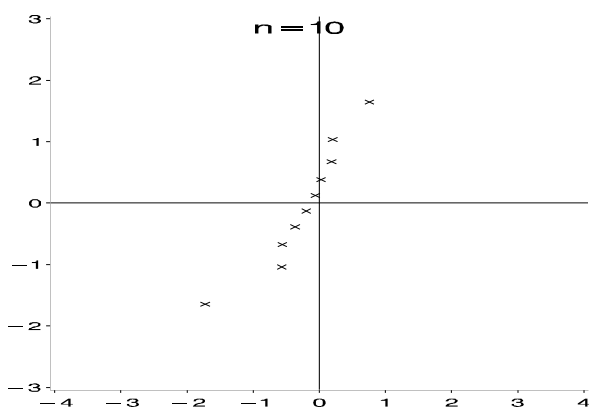
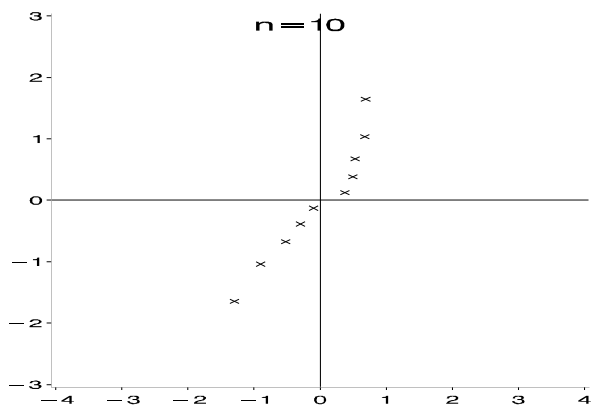
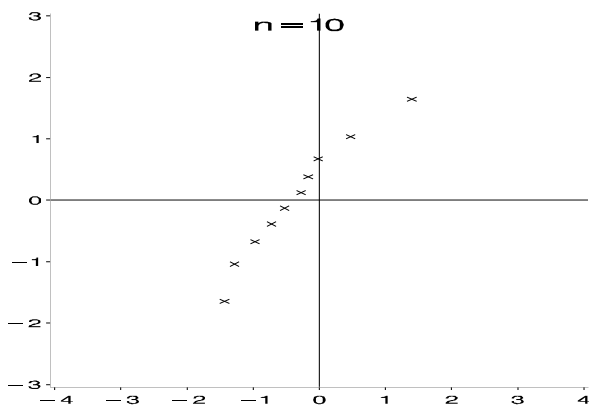
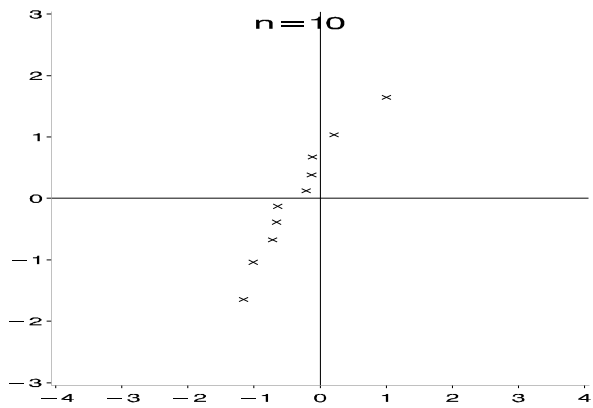
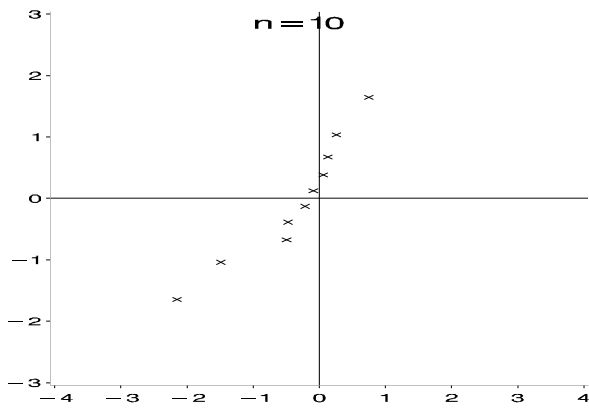
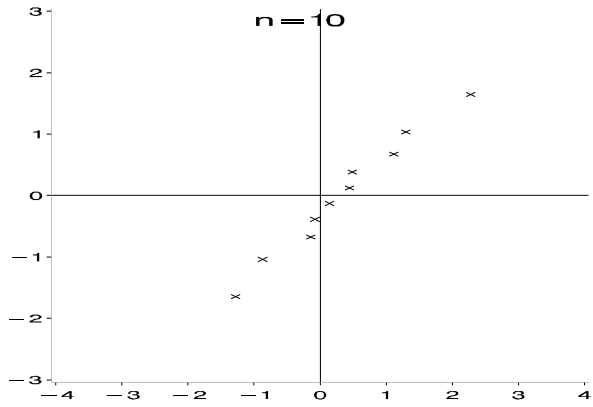
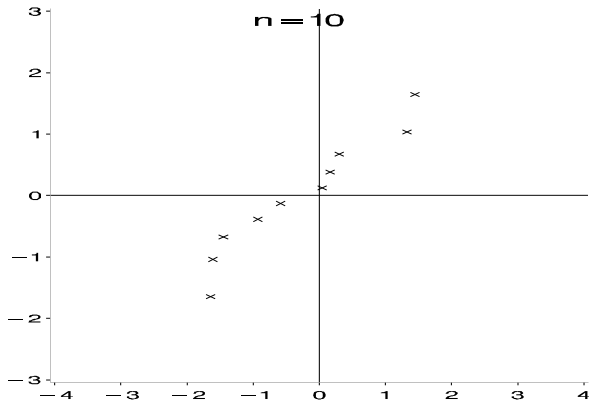
er differentiabel med hensyn til  $x_2$ .

## B Simulerede fraktildiagrammer

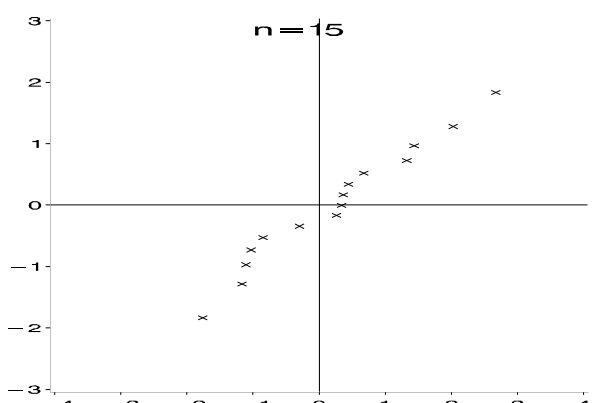
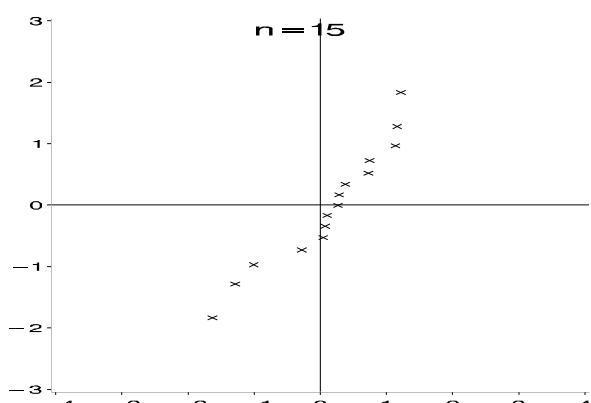
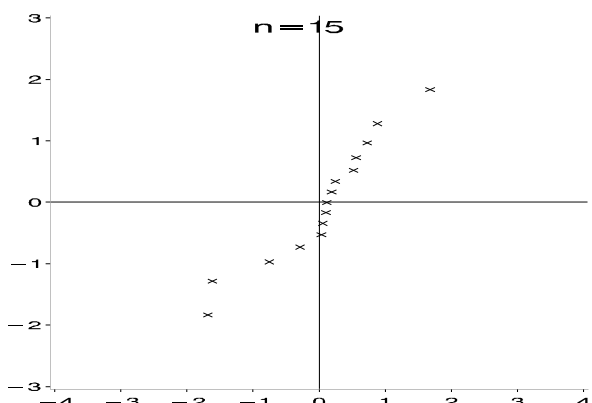
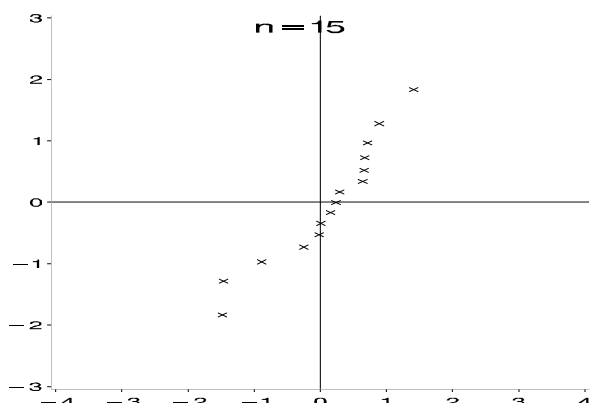
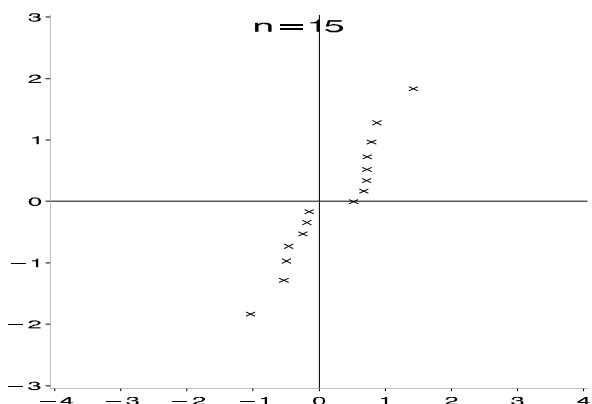
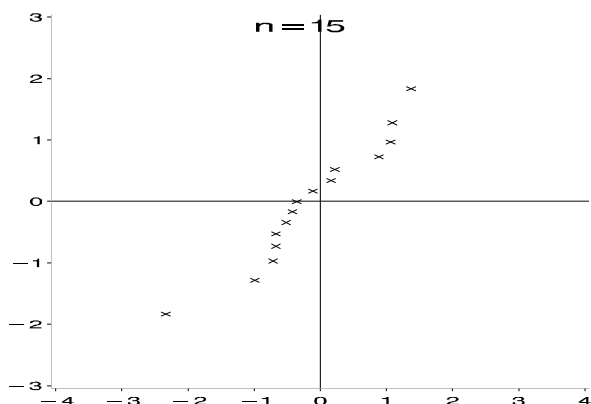
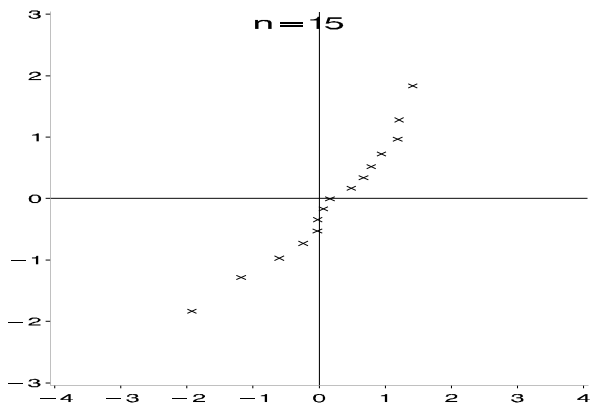
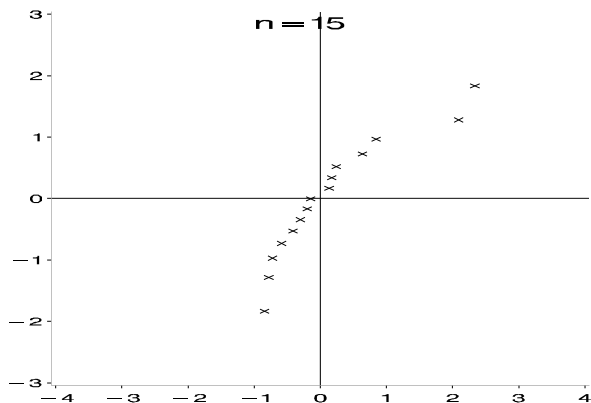
For at give læseren nogen erfaring i at vurdere fraktildiagrammer viser vi i dette appendiks fraktildiagrammer for forskellige stikprøver  $u_1, \dots, u_n$  fra standard normalfordelingen  $N(0, 1)$ . Stikprøverne er frembragt ved numerisk simulation ved hjælp af funktionen NORMAL i den statistiske programpakke SAS. For hver af stikprøvestørrelserne  $n = 5, 10, 15, 25, 50, 100, 250$  er der simuleret otte stikprøver. De tilsvarende fraktildiagrammer er vist på de følgende sider. Størrelsen af stikprøverne fremgår af de enkelte diagrammers overskrift.

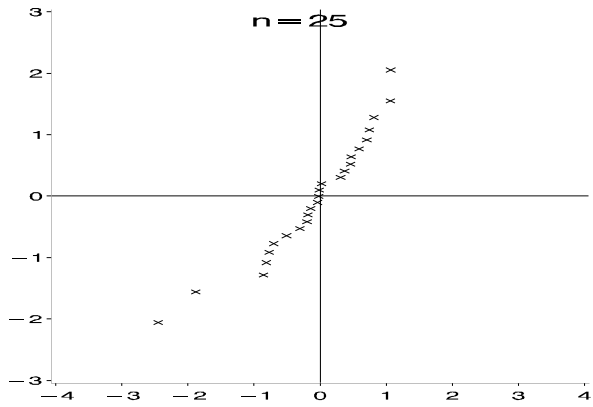
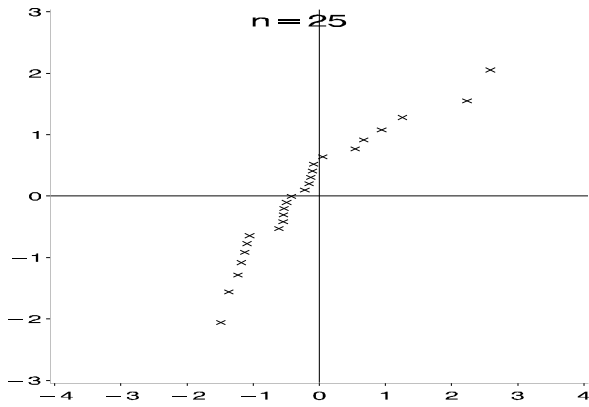
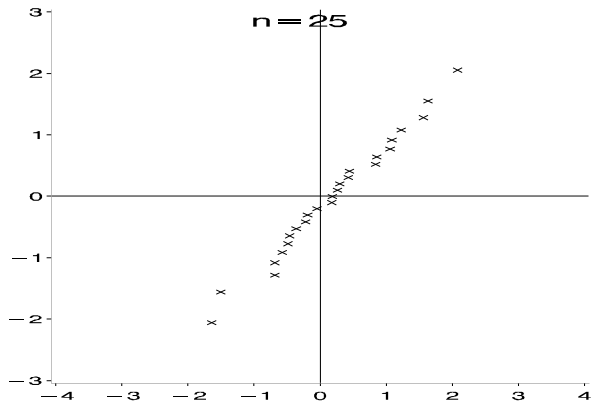
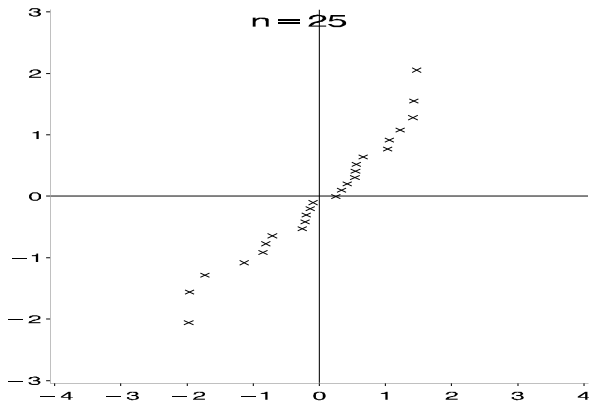
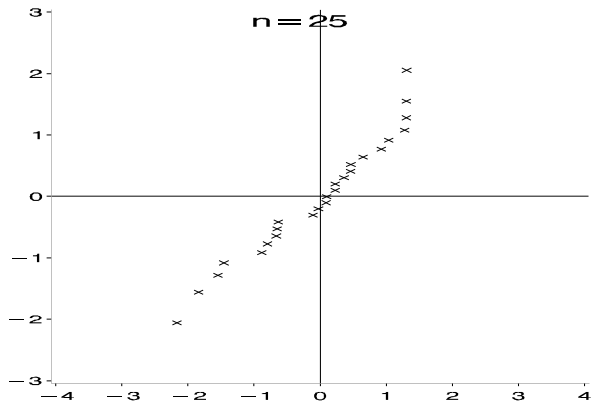
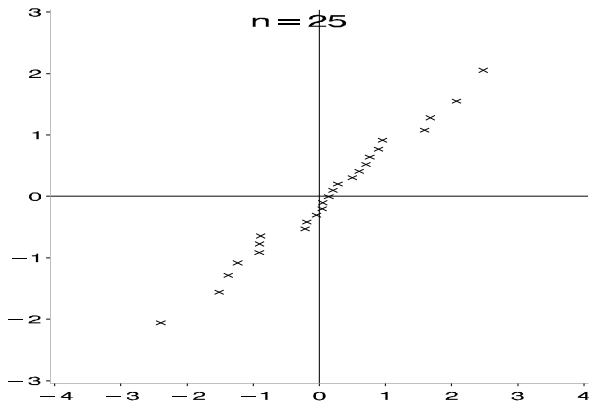
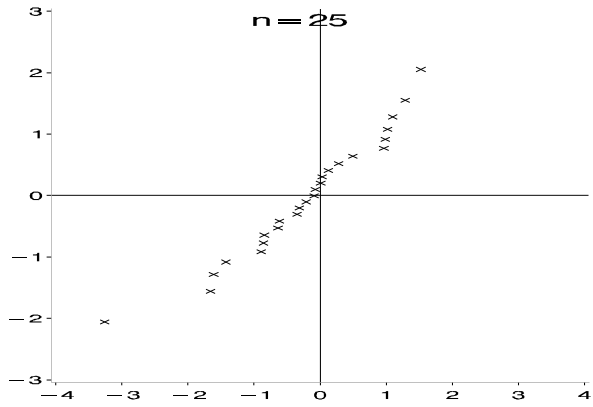
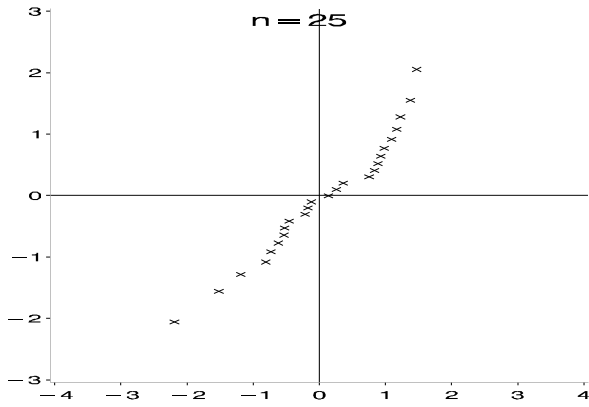
# B.2



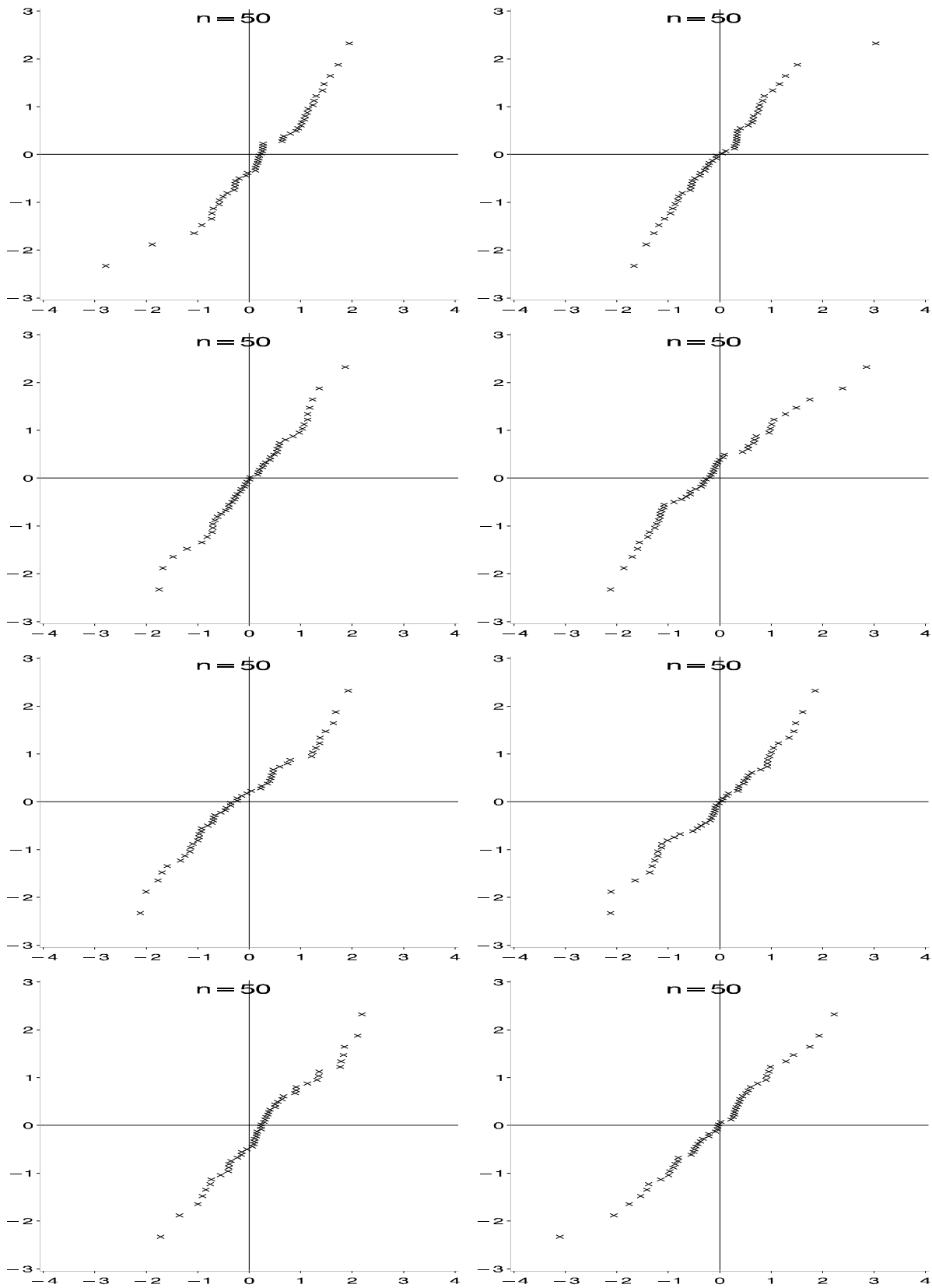


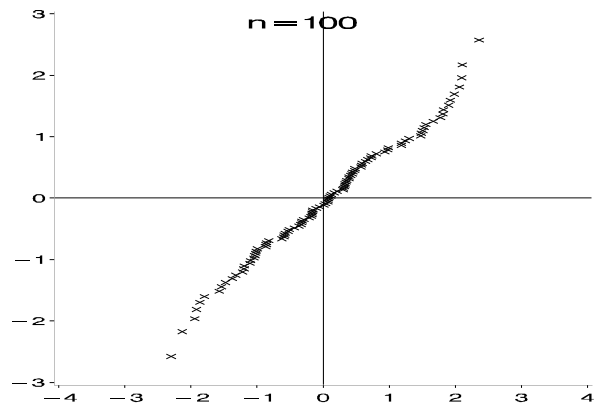
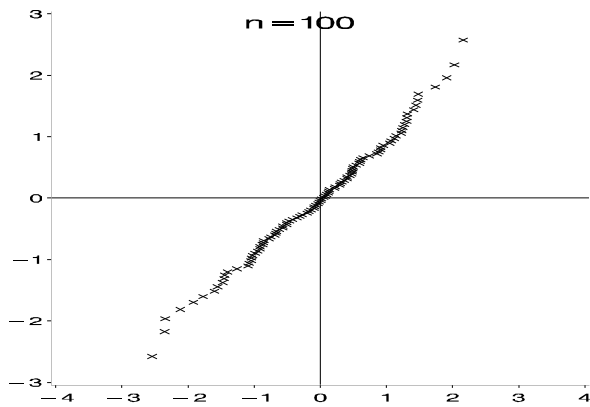
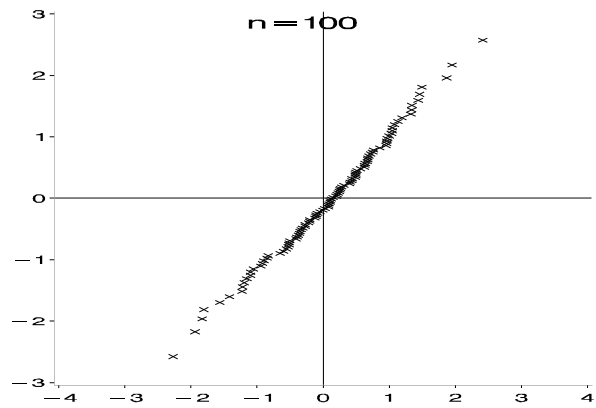
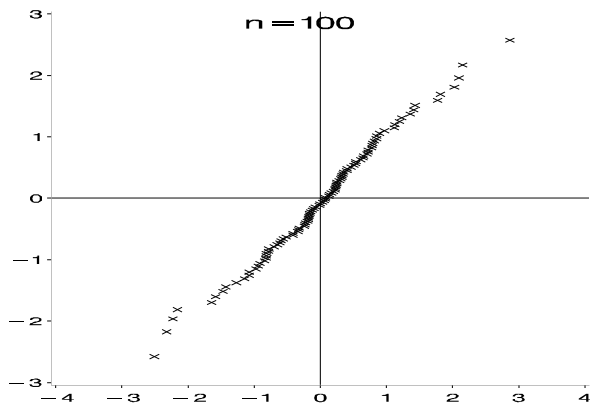
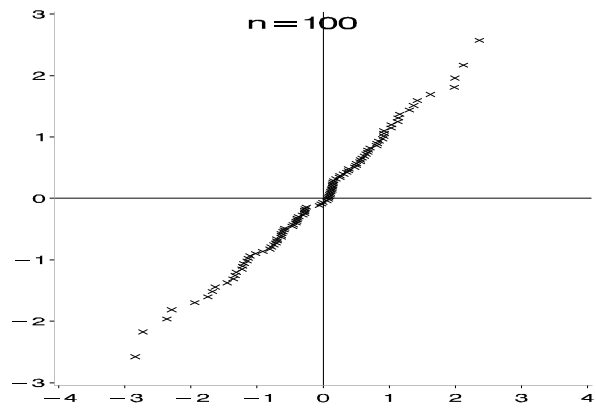
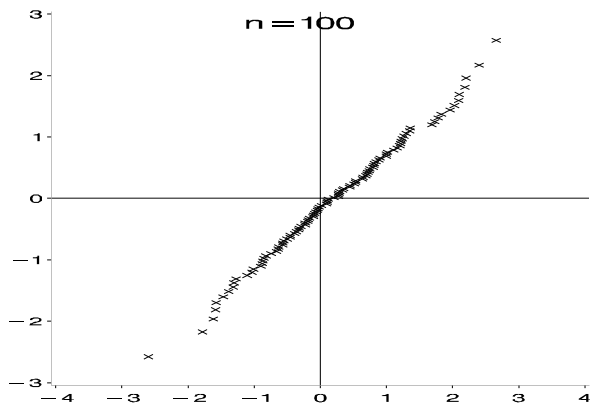
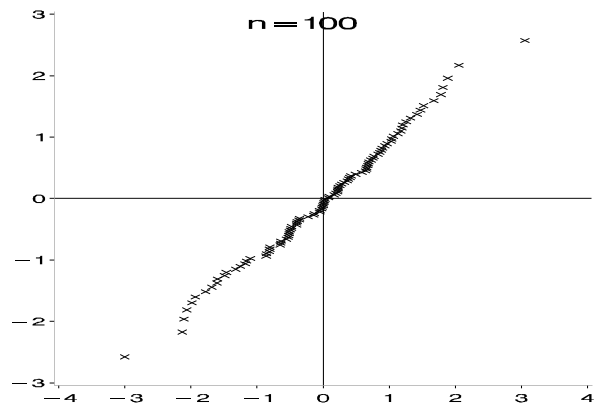
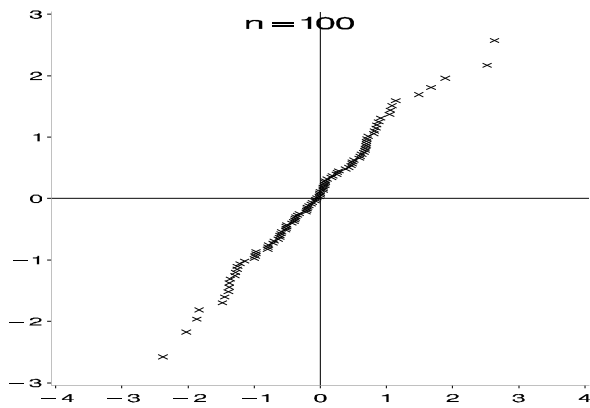
# B.4



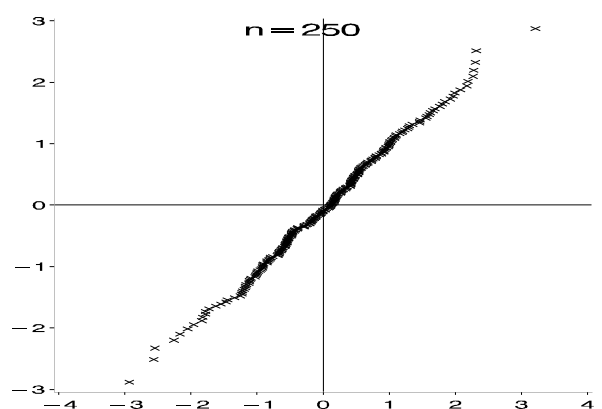
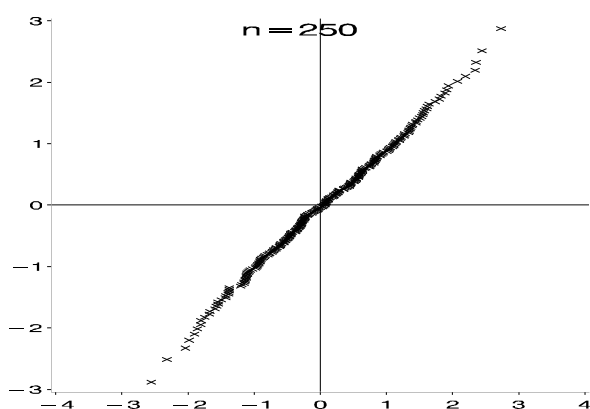
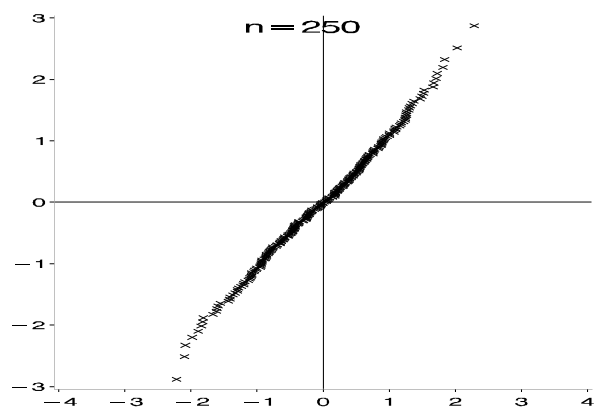
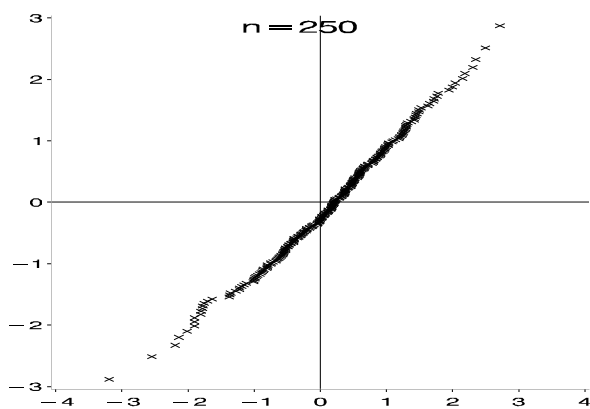
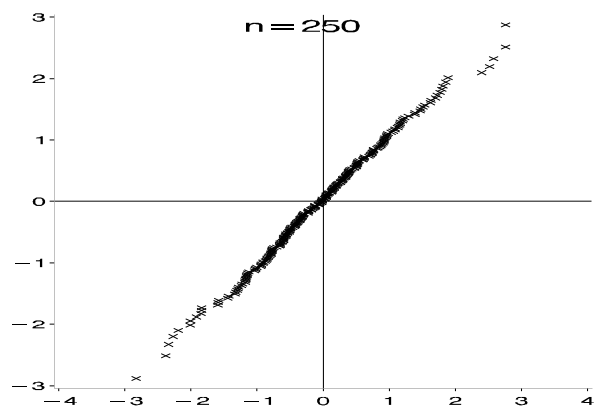
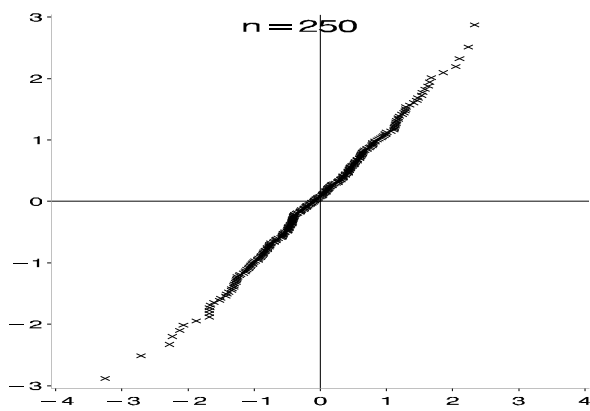
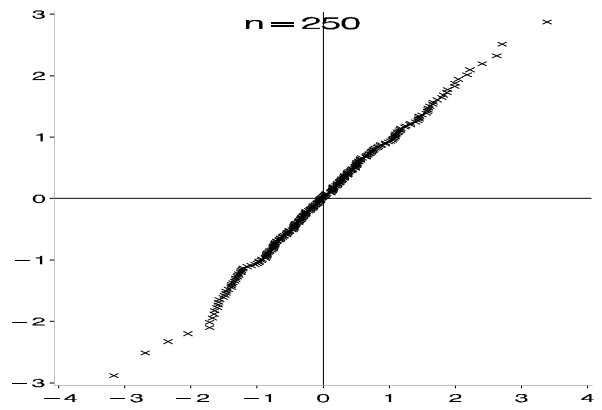
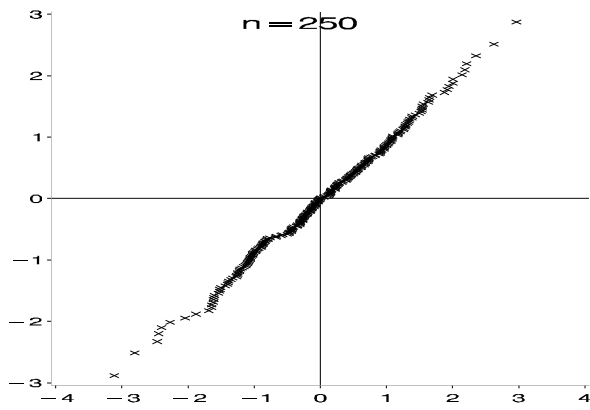


# B.6





B.8



## C Matematiske symboler

$\in$  tilhører, som i  $e \in E$  -  $e$  tilhører  $E$

$\forall$  for alle, som i  $\forall e \in E$  - for alle  $e$  i  $E$

$\exists$  eksisterer, som i  $\exists e \in E$  - der eksisterer  $e$  i  $E$

$\{ \}$  mængde, som i  $\{e \in E : e = 2\}$  - mængden af  $e$  i  $E$  således at  $e = 2$

$\#$  antal elementer i en mængde, som i  $\#E = 7$  - antallet af elementer i  $E$  er 7

$\subseteq$  delmængde af, som i  $A \subseteq B$  - alle elementer i  $A$  er elementer i  $B$

$\supseteq$  indeholder, som i  $A \supseteq B$  - alle elementer i  $B$  er elementer i  $A$

$\cap$  fællesmængde, som i  $A \cap B$  - alle elementer i  $A$  som også er elementer i  $B$

$\cup$  foreningsmængde, som i  $A \cup B$  - elementer som tilhører enten  $A$  eller  $B$

$\emptyset$  den tomme mængde

$^C$  komplementærmængde, som i  $A^C$  - alle elementer som ikke tilhører  $A$

$\setminus$  mængdedifferens, som i  $A \setminus B$  - alle elementer i  $A$  som ikke er elementer i  $B$

$[, ]$  lukket interval, som i  $[a, b]$  - alle elementer  $e$  hvorom det gælder  $a \leq e \leq b$

$], [$  åbent interval, som i  $]a, b[$  - alle elementer  $e$  hvorom det gælder  $a < e < b$

$\times$  produkt af mængder, som i  $A \times B$  - alle par af elementer  $(a, b)$  hvor  $a$  tilhører  $A$  og  $b$  tilhører  $B$

$\rightarrow$  konvergens, som i  $s_n \rightarrow s$  - følgen med elementer  $s_n$  konvergerer mod  $s$

fra til, som i  $f : R \rightarrow [0, 1]$  -  $f$  er en funktion defineret på  $R$  med værdier i  $[0, 1]$

$||$  numerisk (absolut) værdi, som i  $|-7| = 7$  - den numeriske værdi af  $-7$  er 7

## C.2

indhold af mængde, som i  $|A|$  - længden (eller arealet eller rumfanget) af mængden  $A$

$\vee$  logisk eller, som i  $A \vee B$  - enten er udsagnet  $A$  eller udsagnet  $B$  (eller begge) sandt

maksimum, som i  $8 \vee 4 = 8$  - maksimum af 8 og 4 er 8

$\wedge$  logisk og, som i  $A \wedge B$  - begge udsagn  $A$  og  $B$  er sande

minimum, som i  $8 \wedge 4 = 4$  - minimum af 8 og 4 er 4

$\Sigma$  sum, som i  $\sum_{i=1}^n x_i$  - summen  $x_1 + x_2 \cdots + x_n$

$\Pi$  produkt, som i  $\prod_{i=1}^n x_i$  - produktet  $x_1 x_2 \cdots x_n$

$\infty$  uendelig

lim grænseværdi, som i  $\lim_{n \rightarrow \infty} s_n = s$  - grænseværdien for følgen  $s_n$  når  $n \rightarrow \infty$  er  $s$

$\partial$  partielt afledet, som i  $\partial f(x, y) / \partial x$  - funktionen  $f$  af  $x$  og  $y$  differentieret med hensyn til  $x$  for fastholdt  $y$

$\sim$  fordelt som, som i  $X \sim N(0, 1)$  -  $X$  er normalfordelt med middelværdi 0 og varians 1

$\approx$  approksimativt fordelt som, som i  $X \approx N(0, 1)$  - fordelingen for  $X$  kan approksimeres ved en normalfordeling med middelværdi 0 og varians 1

$\doteq$  approksimativt lig med, som  $f(x) \doteq a$  - værdien af funktionen  $f$  beregnet i  $x$  kan approksimeres ved  $a$

## D Det græske alfabet

Da vi i teksten ofte bruger græske bogstaver bringes her en oversigt over bogstaverne i det græske alfabet.

<i>navn</i>	<i>lille</i>	<i>stort</i>	<i>navn</i>	<i>lille</i>	<i>stort</i>
alfa	$\alpha$	<i>A</i>	ny	$\nu$	<i>N</i>
beta	$\beta$	<i>B</i>	xi	$\xi$	$\Xi$
gamma	$\gamma$	$\Gamma$	omicron	$o$	<i>O</i>
delta	$\delta$	$\Delta$	pi	$\pi$	$\Pi$
epsilon	$\varepsilon$	<i>E</i>	rho	$\rho$	<i>P</i>
zeta	$\zeta$	<i>Z</i>	sigma	$\sigma$	$\Sigma$
eta	$\eta$	<i>H</i>	tau	$\tau$	<i>T</i>
theta	$\theta$	$\Theta$	upsilon	$\upsilon$	<i>Y</i>
iota	$\iota$	<i>I</i>	phi	$\varphi$	$\Phi$
kappa	$\kappa$	<i>K</i>	chi	$\chi$	<i>X</i>
lambda	$\lambda$	$\Lambda$	psi	$\psi$	$\Psi$
my	$\mu$	<i>M</i>	omega	$\omega$	$\Omega$



## Referencer

Andersen, E. B. (1998): *Statistik for idrætsstuderende*. Noter fra kursus afholdt ved Institut for Idræt, Københavns Universitet.

Berg, F. og Blæsild, K. (2000): *Fysiske krav i elitefodbold for ungdomsspillere*. Bachelorprojekt, Institut for Idræt, Københavns Universitet.

Blæsild, P. og Granfeldt, J. (2000): *Statistik for biologer og geologer*. Institut for Matematiske Fag, Aarhus Universitet.

Lehmann, E. L. (1975) . Nonparametrics: *Statistical Methods Based on Ranks*. Holden-Day, San Francisco.



# Indeks

<b>A</b>	
acceptområde	5.6
additivetsmodellen	
tosidet variansanalyse	4.113
afhængig variabel	
lineær regression	4.78
afskæring	
lineær regression	4.78
<b>B</b>	
Bartlett test	
$-2 \ln Q$	4.61
for identitet af $k > 2$ varianser	4.61
hovedpunkter	4.73
konstanten $C$	4.61
testsandsynlighed	4.62
Bayes formel	2.7
beregninger	
$s^2$	7.8
$\bar{x}$	7.8
beregninger i <i>Excel</i>	1.31, 4.10, 4.19, 4.28, 4.50, 4.70, 4.135, 6.28, 7.29, 8.15
beta funktion	3.8
betinget fordeling	2.25
betinget sandsynlighed	2.6
binomialfordeling	
beregning af punktsandsynligheder	3.13
definition	3.12
<i>Excel</i>	3.13
middelværdi og varians	3.13
binomialrækken	A.3
<b>C</b>	
$\chi^2$ -fordeling	5.21
definition	3.5
<i>Excel</i>	
fordelingsresultater	3.6
middelværdi og varians	3.6
tabel	3.7
<b>D</b>	
data	
flerdimensionale	1.27
grafisk repræsentation	1.2
grupperede	1.2
grupperede, ugrupperet version	1.19
gruppering	1.5
idræt	1.1, 5.2
kvalitative	1.2
kvantitative	1.2
sæt	1.1, 5.2
tabelform	7.1
tabellering	1.2
todimensionale	1.27
ugrupperede	1.2
delmængde	A.1
område	5.3
sammenhængende	5.3
åben	5.3

diagram	Eksempel 2.5	2.9
blok	superligaholds hjemmekampe	2.9
fraktil	Eksempel 2.6	2.12
kasse	uniform fordeling	2.12
lagkage	Eksempel 2.7	2.14, 2.29
pinde	to kampe på tipskuponen, point fordeling	2.14
prik	Eksempel 2.8	2.15
probit	antal hjemmekampe inden første sejr	2.15
profil	Eksempel 2.9	2.18, 2.30
søjle	tæthedsfunktion for uniform fordeling	2.18
disjunkte mængder	Eksempel 2.10	2.19, 2.22, 2.24, 2.29
parvis	to kampe på tipskuponen	2.19
diskret stokastisk variabel	Eksempel 2.11	2.21, 2.23, 2.24
diskret stokastisk vektor	uniform fordeling på delmængde af $R^2$	2.21
dobbeltintegral	Eksempel 2.12	2.23, 2.25, 2.31
	uniform fordeling på trekant	2.23
	Eksempel 3.1	3.13
	sandsynlighedsfunktion for binomialforde-	
	ling	3.13
	Eksempel 3.2	3.16
	sandsynlighedsfunktion for poissonfordeling	
		3.16
	Eksempel 3.3	3.18
	sandsynlighedsfunktion for hypergeometrisk	
	fordeling	3.18
	Eksempel 3.4	3.20
	sandsynlighedsfunktion for negativ binomial-	
	fordeling	3.20
	Eksempel 4.1	
		4.13, 4.19, 4.21, 5.4, 5.10, 5.13, 5.16, 5.21
	bestemmelse af laktatkonzentration	4.13
	Eksempel 4.2	4.34, 4.35, 4.38
	<i>Excel</i>	4.50
	kondital for ikke-aktive og aktive	4.34
	Eksempel 4.3	4.42
	<i>Excel</i>	4.52
	tider i semifinalerne i kvindernes 100 m løb	4.42
	Eksempel 4.4	4.46
<b>E</b>		
Eksempel 1.1	1.3, 1.5, 4.7, 4.26, 6.26	
højde af piger	1.3	
Eksempel 1.2	1.3, 4.3, 4.5, 4.10, 4.27	
kondital for eliteidrætsudøvere	1.3	
Eksempel 1.3	1.4, 1.24, 1.26	
<i>Excel</i>	1.37	
resultatet af Faxe Kondi Ligaen	1.4	
Eksempel 1.4	1.11, 1.16, 1.31, 1.35	
hypotetiske kondital	1.11	
Eksempel 1.5	1.28	
glycogen indhold i muskler	1.28	
Eksempel 2.1	2.4	
uniforme sandsynlighedsmål på endelig		
mængde	2.4	
Eksempel 2.2	2.4, 2.19	
to kampe på tipskuponen	2.4	
Eksempel 2.3	2.5	
uniforme sandsynlighedsmål på interval	2.5	
Eksempel 2.4	2.8	
superligaholds chancer på hjemme- og ude-		
bane	2.8	

<i>Excel</i> . . . . .	4.53	Eksempel 7.1 . . . . .	7.2, 7.8, 7.11, 7.17
muskelglucogen før og efter træning . . .	4.46	antal mål i Faxe Kondi Ligaen . . . . .	7.2
Eksempel 4.5 . . . . .	4.59, 4.62, 4.67	<i>Excel</i> . . . . .	7.29
<i>Excel</i> . . . . .	4.70	Eksempel 7.2 . . . . .	7.2, 7.15
længdespring . . . . .	4.59	de nordiske landes medaljehøst ved OL i Syd-	
Eksempel 4.6 . . . . .	4.76, 4.82, 4.93	ney . . . . .	7.2
<i>Excel</i> . . . . .	4.101	<i>Excel</i> . . . . .	7.31
lineær regression af puls på tid . . . . .	4.76	Eksempel 7.3 . . . . .	7.2, 7.26
Eksempel 4.7 . . . . .	4.83, 4.93	<i>Excel</i> . . . . .	7.32
<i>Excel</i> . . . . .	4.103	medaljefordeling ved OL i Sydney . . . . .	7.2
finaletider i kvindernes 200 m, 400 m og		Eksempel 8.1 . . . . .	8.2, 8.3, 8.6
800 m løb . . . . .	4.83	<i>Excel</i> . . . . .	8.15
Eksempel 4.8 . . . . .	4.94	kondital før og efter træning . . . . .	8.2
puls og iltoptagelse . . . . .	4.94	Eksempel 8.2 . . . . .	8.7, 8.10, 8.14
Eksempel 4.9 . . . . .	4.111, 4.117, 4.127	kondital for ikke-aktive og aktive . . . . .	8.7
<i>Excel</i> . . . . .	4.135	Eksempel 8.3 . . . . .	8.11, 8.13
tosidet variansanalyse uden gentagelser	4.111	længdespring . . . . .	8.11
Eksempel 4.10 . . . . .	4.111, 4.117, 4.131	eksperiment	
<i>Excel</i> . . . . .	4.136	datasæt . . . . .	1.1, 5.2
tosidet variansanalyse med gentagelser .	4.111	eksponentialfordeling	
Eksempel 4.11 . . . . .	4.134	definition . . . . .	3.22
tosidet variansanalyse og det parrede <i>t</i> -test	4.134	<i>Excel</i> . . . . .	3.23
Eksempel 6.1 . . . . .	6.1	middelværdi og varians . . . . .	3.22
multinomialfordelte data . . . . .	6.1	eksponentialrækken . . . . .	A.4
Eksempel 6.2 . . . . .	6.3, 6.10	empirisk fordelingsfunktion . . . . .	1.12
AB's kampe på hjemme- og udebane . . . .	6.3	empirisk korrelationskoefficient . . . . .	1.28
<i>Excel</i> . . . . .	6.28	empirisk middelværdi . . . . .	1.16
Eksempel 6.3 . . . . .	6.3, 6.13	empirisk spredning . . . . .	1.16
<i>Excel</i> . . . . .	6.29	empirisk varians . . . . .	1.16
idrættaktivitet og rygning . . . . .	6.3	én observationsrække	
Eksempel 6.4 . . . . .	6.17	Poissonfordelingen . . . . .	7.7
<i>Excel</i> . . . . .	6.30	tabelform . . . . .	7.1
opdeling af resultater i Faxe Kondi Ligaen	6.4	endelig række . . . . .	A.3
Eksempel 6.5 . . . . .	6.20, 6.22	eksempler . . . . .	A.3
undersøgelse af sammenhæng mellem kræft		estimat . . . . .	5.5
og magnetfelter . . . . .	6.20	interval . . . . .	5.16
Eksempel 6.6 . . . . .	6.26	maksimum likelihood . . . . .	5.9
test for goodness of fit . . . . .	6.26	notation . . . . .	5.5

estimation		<i>F</i> -fordeling	3.12
én observationsrække, Poissonfordelingen	7.7	fraktildiagram	4.10
lineær regression	4.78	funktionen BINOMIALFORDELING	3.13
maksimum likelihood	5.9	funktionen CHIFORDELING	3.7
middelværdien i én observationsrække	4.13	funktionen CHIINV	3.7, 6.30
multiplikativ Poissonmodel	7.21	funktionen CHITEST	6.28, 6.30
proportionale parametre i Poissonmodel	7.13	funktionen EKSPFORDELING	3.23
teori	5.5	funktionen FAST	1.40
todimensional normalfordeling	4.96	funktionen FFORDELING	3.12
tosidet variansanalyse	4.114	funktionen FINV	3.12
variansen i én observationsrække	4.21	funktionen FRAKTIL	1.34
estimator	5.5	funktionen HYPGEOFORDELING	3.19
maksimum likelihood	5.6, 5.9	funktionen KOMBIN	2.34
<i>Excel</i>		funktionen NEGBINOMFORDELING	3.20
analoge formler	1.35	funktionen NORMFORDELING	3.4
Beskrivende statistik	1.32	funktionen NORMINV	3.4
binomialfordeling	3.13	funktionen PLADS	8.15
$\chi^2$ -fordeling	3.7	funktionen POISSON	3.16
Diagram	1.35	funktionen POTENS	2.33
100 (procent) stablet søjlediagram		funktionen SLUMP	2.35
.....	1.39	funktionen SUMPRODUKT	6.28
Grupperet søjle	1.38, 1.39	funktionen TFORDELING	3.10
Punktdiagram	1.36	funktionen TINV	3.10
dialogboksen Anava:		Histogram	1.33
Enkelt faktor	4.70	hypergeometrisk fordeling	3.19
To-faktor med gentagelse	4.135	negativ binomialfordeling	3.20
To-faktor uden gentagelse	4.135	normalfordeling	3.4
dialogboksen F-test:		numeriske variable	1.39
Dobbelt stikprøve for ens varians		Poissonfordeling	3.16
.....	4.50	<i>t</i> -fordeling	3.10
dialogboksen Regression	4.101	<i>t</i> -test for kendt middelværdi	4.28
dialogboksen <i>t</i> -test:		tekst variable	1.39
Parvis dobbelt stikprøve for		test for kendt varians	4.28
middelværdi	4.50	<i>u</i> -test	4.19
To stikprøver med ens varians	4.50		
To stikprøver med forskellig		<b>F</b>	
varians	4.50	<i>F</i> -fordeling	
eksponentialfordelingen	3.23	definition	3.10



græsk alfabet . . . . .	D.1	hændelser . . . . .	2.2
<b>H</b>		hændelsessystem . . . . .	1.1, 5.2
histogram . . . . .	1.5	højreskæv . . . . .	1.16
homogenitet		<b>I</b>	
tosidet variansanalyse . . . . .	4.113	inferens	
homogenitet af flere multinomialfordelinger	6.15	likelihood . . . . .	5.7
$-2 \ln Q$ -testet, testsandsynlighed . . . . .	6.16	statistisk . . . . .	5.5
beregningsformel for $-2 \ln Q$ -testet . . . . .	6.16	information	
eksempel . . . . .	6.20	Fisher . . . . .	5.18
fordelingsresultat . . . . .	6.17	forventet . . . . .	5.18
maksimum likelihood estimat . . . . .	6.16	observeret . . . . .	5.18
hovedpunkter		intervalestimat . . . . .	5.16
én observationsrække med kendt varians . . . . .	4.20	<b>K</b>	
én observationsrække med ukendt varians . . . . .	4.30	$k$ observationsrækker . . . . .	4.59
ikke-parametriske test . . . . .	8.18	estimation . . . . .	4.63
$k$ observationsrækker . . . . .	4.73	notation . . . . .	4.32
lineær regression . . . . .	4.106	statistisk model . . . . .	4.32
multinomialmodel . . . . .	6.31	kassediagram . . . . .	1.16
Poissonfordelte data . . . . .	7.33	kategori . . . . .	1.23
to observationsrækker . . . . .	4.55	numerisk . . . . .	1.23
tosidet variansanalyse . . . . .	4.137	komplementærmængde . . . . .	A.1
hypergeometrisk fordeling . . . . .	6.22	konfidens	
beregning af punktsandsynligheder . . . . .	3.18	interval . . . . .	5.16
definition . . . . .	3.17	område . . . . .	5.16
<i>Excel</i> . . . . .	3.19	konfidensinterval	
middelværdi og varians . . . . .	3.18	binomialmodel . . . . .	6.9
hypotese		for afskæringen i lineær regression . . . . .	4.107
kritisk observation . . . . .	5.11	for forskel mellem to middelværdier; ens va-	
multinomialmodel . . . . .	6.6	rianser . . . . .	4.56
punkt . . . . .	5.6	for forskel mellem to middelværdier; forskel-	
sammensat . . . . .	5.6	lige varianser . . . . .	4.57
simpel . . . . .	5.6	for hældningen i lineær regression . . . . .	4.107
test af . . . . .	5.6	for middelværdien; kendt varians . . . . .	4.16
hyppighed		for middelværdien; ukendt varians . . . . .	4.23
relativ . . . . .	1.7	for regressionslinjen . . . . .	4.107
hældning		for spredningen i en normalfordeling . . . . .	4.26
lineær regression . . . . .	4.78	for variansen i en normalfordeling . . . . .	4.26

for variansen i lineær regression . . . . .	4.107	ratio test, approksimativ testsandsynlighed	5.20
multinomialmodel . . . . .	6.9	ratio testor, approksimativ fordeling . . . . .	5.21
Poissonmodel . . . . .	7.10	lineær regression	
middelværdien $\lambda$ baseret på én Poissonfordelt variabel . . . . .	7.11	estimaternes fordeling . . . . .	4.106
middelværdien $\lambda$ i én Poissonfordelt observationsrække . . . . .	7.11	hypoteser om regressionsparametrene . . . . .	4.90, 4.107
parameteren i modellen med proportionale parametre . . . . .	7.13	konfidensintervaller for parametrene . . . . .	4.106
kontinuert stokastisk variabel . . . . .	2.16	med gentagelser . . . . .	4.83
kontinuert stokastisk vektor . . . . .	2.20, 5.2	modelkontrol . . . . .	4.106
kontrast . . . . .	4.126	test af hypotesen om lineær regression . . . . .	4.86, 4.106
korrelation . . . . .	2.28	uden gentagelser . . . . .	4.78
empirisk . . . . .	1.28	log likelihood	
kovarians . . . . .	2.28	funktion . . . . .	5.9
regneregler . . . . .	2.28	funktion, normeret . . . . .	5.18
kumulerede antal . . . . .	1.18	logaritmisk række . . . . .	A.4
kvartil		loven om total sandsynlighed . . . . .	2.7
afstand, empirisk . . . . .	1.15		
nedre . . . . .	2.11	<b>M</b>	
nedre, empirisk . . . . .	1.12	maksimum likelihood	
øvre, empirisk . . . . .	1.12	estimat . . . . .	5.9
øvre . . . . .	2.11	estimation . . . . .	5.9
kvotientrække		estimator . . . . .	5.6, 5.9
endelig . . . . .	A.4	marginal fordeling . . . . .	2.22
uendelig . . . . .	A.4	median . . . . .	2.11
<b>L</b>		empirisk . . . . .	1.12
likelihood . . . . .	5.1	middelværdi	
approksimativ teori . . . . .	5.17	af funktion af diskret stokastisk vektor . . . . .	2.26
estimat, maksimum . . . . .	5.9	af funktion af kontinuert stokastisk vektor . . . . .	2.27
estimation, maksimum . . . . .	5.9	af gennemsnit . . . . .	2.29
estimator, maksimum . . . . .	5.9	diskret stokastisk variabel . . . . .	2.26
funktion . . . . .	5.6	empirisk . . . . .	1.16
inferens . . . . .	5.7	kontinuert stokastisk variabel . . . . .	2.26
ligninger . . . . .	5.9	regneregler . . . . .	2.27
maksimum, estimator . . . . .	5.6	middelværdivektor . . . . .	2.26
ratio test . . . . .	5.11	mindste kvadraters metode . . . . .	4.78
		model	
		funktion . . . . .	5.3, 5.7

inferens . . . . .	5.1	<b>N</b>	
kontrol . . . . .	5.1, 5.4	negativ binomialfordeling	
opstilling . . . . .	1.1, 5.1, 5.2	beregning af punktsandsynligheder . . . . .	3.19
parametrisk . . . . .	5.3	definition . . . . .	3.19
sandsynlighedsteoretisk . . . . .	1.1, 5.2	<i>Excel</i> . . . . .	3.20
statistisk . . . . .	5.2	middelværdi og varians . . . . .	3.20
multinomialfordeling		normalfordeling . . . . .	1.7, 4.1
betingelser for . . . . .	6.1	definition . . . . .	3.1
definition . . . . .	3.15	<i>Excel</i> . . . . .	3.4
egenskaber ved . . . . .	6.4	fordelingsresultater . . . . .	3.3
marginale fordelinger . . . . .	3.15	middelværdi og varians . . . . .	3.2
middelværdivektor og kovariansmatriks . . . . .	3.15	standard . . . . .	3.1
multinomialmodel		tabeller . . . . .	3.3
$-2 \ln Q$ -testor . . . . .	6.8	todimensional . . . . .	3.4
$-2 \ln Q$ -testor, testsandsynlighed . . . . .	6.8	normalfordelte data . . . . .	4.1
$X^2$ -testoren . . . . .	6.9	notation	
$X^2$ -testoren, testsandsynlighed . . . . .	6.9	følge af modeller . . . . .	4.69
estimation . . . . .	6.6	numerisk	
estimation under hypotese . . . . .	6.7	undersøgelse . . . . .	5.4
flere multinomialfordelinger . . . . .	6.15	<b>O</b>	
forventede antal under hypotese . . . . .	6.8	observation	
frie parametre . . . . .	6.6	kritisk . . . . .	5.11
homogenitet af flere multinomialfordelinger		observationer	
. . . . .	6.15	sammenfaldende . . . . .	8.5
hovedpunkter . . . . .	6.31	observationsrække . . . . .	1.2, 5.9
hypotese . . . . .	6.6	én; normalfordeling . . . . .	4.1
hypotese, frie parametre . . . . .	6.6	én; Poissonfordelingen . . . . .	7.7
konfidensinterval . . . . .	6.9	én; todimensional normalfordeling . . . . .	4.94
statistisk inferens . . . . .	6.5	område . . . . .	5.3
test af simpel hypotese, eksempel . . . . .	6.10	accept . . . . .	5.6
uafhængighed af inddelingskriterier . . . . .	6.11	kritisk . . . . .	5.6
mængdedifferens . . . . .	A.2	omvendt betinget sandsynlighed . . . . .	2.7
mængdelære . . . . .	2.1, A.1	opgaver . . . . .	
måle . . . . .	1.2	. . . . .	1.41, 2.32, 3.22, 4.142, 5.23, 6.35, 7.39, 8.21
målelig mængde . . . . .	2.2	ordnede stikprøve . . . . .	1.11
		ordnede værdier . . . . .	1.11, 8.4

<b>P</b>	
<i>p</i> -fraktil	2.10
empirisk fordeling	1.12
<i>p</i> -værdi	5.12
parameter	5.3
fri	5.21
mængde	5.3
rum	5.3
parrede <i>t</i> -test	4.46
partiell differentiation	A.5
pindediagram	1.5
Poisson processen	7.1, 7.5
intensitet	7.6
Poissonfordeling	
approksimeret med normalfordeling	7.5
beregning af punktsandsynligheder	3.16
definition	3.16
egenskaber ved	7.3
<i>Excel</i>	3.16
grænsefordeling for binomialfordeling	7.5
middelværdi og varians	3.16
relation til multinomialfordeling ved betingning	7.5
Poissonmodel	
én observationsrække, $-2 \ln Q$ -test	7.14
én observationsrække, estimation	7.7
én observationsrække, Fishers dispersionsindeks	7.7
én observationsrække, modelkontrol	7.7
én observationsrække, test for goodness of fit	7.7
én observationsrække, $X^2$ -test	7.14
konfidensinterval	7.10
konfidensinterval for middelværdien $\lambda$ i Poissonfordeling	7.11
konfidensinterval for middelværdien $\lambda$ i Poissonfordelt observationsrække	7.11
konfidensinterval for parameteren i Poissonmodellen med proportionale parametre	7.13
multiplikativ	7.18
multiplikativ, estimation	7.21
multiplikativ, homogenitet	7.19
multiplikativ, ingen vekselvirkning	7.19
multiplikativ, kun rækkevirkning	7.19
multiplikativ, kun søjlevirkning	7.19
multiplikativ, parametrisering	7.20
multiplikativ, relation til multinomialmodel	7.25
multiplikativ, test af hypoteser	7.23
proportionale parametre	7.12
proportionale parametre, $-2 \ln Q$ -test	7.13
proportionale parametre, $-2 \ln Q$ -test, testsandsynlighed	7.13
proportionale parametre, estimation	7.13
proportionale parametre, fordeling af estimator	7.13
proportionale parametre, relation til multinomial model	7.15
proportionale parametre, $X^2$ -test	7.13
proportionale parametre, $X^2$ -test, testsandsynlighed	7.13
position	
lineær regression	4.78
prikdiagram	1.4
probit	4.5
profildiagram	4.117
programpakker	1.7
<b>R</b>	
rang	
af observationer	1.11, 8.4
rangtest	8.4
reduktion	
statistisk model	5.6
regneregler	
betinget sandsynlighed	2.7
middelværdi	2.27
sandsynligheder	2.3

- uendelige rækker . . . . . A.4
- varians og kovarians . . . . . 2.28
- regressionskoefficient
  - lineær regression . . . . . 4.78
- regressionslinje
  - lineær regression . . . . . 4.78
- relativ hyppighed . . . . . 1.7, 6.6, 6.12
- residual
  - kvadratsum . . . . . 4.79
  - tosidet variansanalyse . . . . . 4.117
- respons
  - lineær regression . . . . . 4.78
- rækkevirkning . . . . . 4.113
- S**
- S*
  - sum af observationer . . . . . 1.17, 7.8
- $s^2$ 
  - beregningsformel . . . . . 7.8
- $s^2_{(i)}$ 
  - $k$  observationsrækker . . . . . 4.32
- SAK*
  - sum af afvigelses kvadrater . . . . . 1.17
- $SAK_{(i)}$ 
  - $k$  observationsrækker . . . . . 4.32
- sammenfaldende observationer . . . . . 8.5
- sandsynlighed
  - test . . . . . 5.12
- sandsynlighedsfunktion
  - diskret stokastisk variabel . . . . . 2.12
  - diskret stokastisk variabel, egenskaber ved . . . . . 2.14
  - diskret stokastisk vektor . . . . . 2.19
  - marginal fordeling . . . . . 2.22
- sandsynlighedsmål . . . . . 1.1, 5.2
- sandsynlighedsmål . . . . . 2.1
- sandsynlighedspapir . . . . . 4.5
- sandsynlighedsrum . . . . . 2.2
- sandsynlighedsteori . . . . . 1.1, 5.2
- SAP*
  - sum af afvigelses produkter . . . . . 1.28
- signifikansniveau . . . . . 5.12, 5.14
  - observeret . . . . . 5.12
- simultan fordeling . . . . . 2.22
- SK*
  - sum af kvadrater . . . . . 1.17, 7.8
- SP*
  - sum af produkter . . . . . 1.28
- spredning . . . . . 2.27
  - empirisk . . . . . 1.16
- standardafvigelse . . . . . 2.27
- statistik
  - beskrivende . . . . . 1.2
  - deskriptiv . . . . . 1.2
  - ikke-parametrisk . . . . . 5.22
  - inferens . . . . . 5.5
  - parametrisk model . . . . . 5.3
- statistikens slutningsregel . . . . . 4.17
- statistisk
  - analyse . . . . . 1.1, 5.1, 5.2
  - approksimativ metode . . . . . 5.1
  - metode . . . . . 5.1
- stikprøve . . . . . 1.2
  - ordnet . . . . . 1.11
  - størrelse . . . . . 1.2
- stokastisk
  - variation . . . . . 1.1, 5.2
  - vektor, multinomialfordelt . . . . . 6.1
- stokastisk variabel . . . . . 2.10
  - diskret . . . . . 2.12
  - kontinuert . . . . . 2.16
- stokastisk vektor . . . . . 2.19
  - diskret . . . . . 2.19, 5.2
  - kontinuert . . . . . 2.20, 5.2
- styrkefunktion . . . . . 5.14
- støtte
  - diskret stokastisk variabel . . . . . 2.14





undersøgelse	
grafisk . . . . .	5.4
numerisk . . . . .	5.4
uniform fordeling . . . . .	2.12
tæthedsfunktion . . . . .	2.18
uniforme sandsynlighedsmål	
på endelig mængde . . . . .	2.4
på interval . . . . .	2.5

**V**

varians . . . . .	2.27
af gennemsnit . . . . .	2.29
empirisk . . . . .	1.16
regneregler . . . . .	2.28
variansanalyse . . . . .	4.67
variansanalysetabel . . . . .	4.67
ensidet variansanalyse . . . . .	4.67
tosidet variansanalyse med gentagelser .	4.124
tosidet variansanalyse uden gentagelser	4.125
venstreskæv . . . . .	1.16