En Introduktion til Statistik

Bind 3C

Hierarkiske modeller

Poul Thyregod

LYNGBY 1998

IMM
Trykt af IMM - DTU
Bogbinder Hans Meyer
## Indhold

1 Stikprøver fra endelige populationer, Repræsentative undersøgelser 1

1.1 Grundlæggende begreber .................................................. 1

1.1.0 Indledning ................................................................. 1

1.1.1 Oversigt ................................................................. 4

1.2 Endelige populationer og tilfældige stikprøver ...................... 5

1.2.1 Populationsparametre .................................................. 5

1.3 Stikprøver fra endelige populationer ................................ 17

1.3.1 Målgruppe, stikprøveramme, stikprøve og tilfældig stikprøve ................................................................. 17

1.3.2 Stikprøveudtagning ved simpel tilfældig udvælgelse .......... 18

1.4 Estimation af populationstotalen eller populationsgennemsnit 22

1.5 Estimation af populationsvarians ...................................... 25

1.5.1 Momenter for stikprøvevariansen .................................. 25

1.5.2 Konfidensgrænser: .................................................... 27

1.6 Stikprøver fra populationer med flere værdier pr analyseenhed 33

1.6.1 Stikprøvekovarians ..................................................... 33

1.6.2 Relativ værdi pr analyseenhed .................................... 36

1.7 Kvantentralskøn ............................................................. 38
1.7.1 Det simple kvotientskøn ............................................. 39
1.7.2 Korrigerede kvotientskøn ............................................. 43
1.7.3 Kvotientskøn for populationsgennemsnittet .......... 45
1.7.4 Regressionsskøn for populationsgennemsnittet .... 51
1.7.5 Sammenligning mellem regressionsskøn, kvotientskøn og direkte estimation ved stikprøvegennemsnittet. 59

1.8 Udvælgelse med varierende sandsynligheder ............. 62
1.8.1 Indledning ................................................................. 62
1.8.2 Fordelingsforhold ved udvælgelse med varierende sandsynligheder ..................................................... 62
1.8.3 Udvælgelse proportional med størrelse (PPS)-sampling 68

1.9 Udnyttelse af populationens struktur, stratifikation ...... 73
1.9.1 Vilkårlig allokering ..................................................... 74
1.9.2 Proportional fordeling af stikprøven på strata .......... 75
1.9.3 Optimal fordeling på strata ........................................... 77
1.9.4 Sammenligning mellem simpel tilfældig og stratifice-
ret udvælgelse ................................................................. 82

1.10 Udnyttelse af populationens struktur,
Klyngeudvælgelse ............................................................ 85
1.10.1 Udvælgelse af klynger med varierende sands. ....... 96

1.11 Totrinsudvælgelse .......................................................... 101
1.12 Referencer ................................................................. 102

2 Likelihoodfunktion, generaliserede lineære modeller for end- 

dimensionale eksponentielle dispersionsparameterfamilier 103

2.0 Indledning ................................................................. 103
2.1 Likelihoodfunktionen ..................................................... 105
2.1.1 Sufficiens ................................................................. 109
2.1.2 Scorefunktionen og Informationsmatricen .......... 112
Indhold

2.1.3 Maksimum likelihood estimat .................................. 117
2.2 Eksponentielle familier og dispersionsmodeller ............. 121
  2.2.1 Naturlige eksponentielle familier af fordelinger .......... 121
  2.2.2 Eksponentielle dispersionsmodeller ....................... 131
  2.2.3 Oversigt over enhedsvariansfunktioner, dispersions- 
        parametre og enhedsdevianer for sædvanlige ekspo-
        nentielle dispersionsmodeller ............................ 141
  2.2.4 Lidt om likelihoodfunktionen svarende til observatio-
        ner fra eksponentielle dispersionsmodeller ............ 143
2.3 Linkfunktioner .................................................. 150
  2.3.1 Sædvanlige linkfunktioner .................................. 152
  2.3.2 Illustration af afbildningerne ved forskellige linkfunk-
        tioner .................................................. 154
  2.3.3 Hyperbelfunktioner ......................................... 156
  2.3.4 Logaritmfunktioner ......................................... 157
  2.3.5 Eksponentialfunktioner .................................... 159
  2.3.6 Potensfunktioner ........................................... 162
2.4 Generaliserede lineære modeller .............................. 164
  2.4.0 Indledning .................................................. 164
  2.4.1 Definition af en generaliseret lineær model ............. 164
  2.4.2 Eksempel på generaliserede lineære modeller .......... 170
2.5 Estimation i generaliseret lineær model, fordeling af estimater 177
  2.5.1 Maksimum likelihood estimat, observeret og forven-
        tet information ........................................ 177
  2.5.2 Fittede værdier ........................................... 181
  2.5.3 Asymptotisk fordeling af maksimum likelihood esti-
        matet .................................................. 182
  2.5.4 Iterative metoder til estimation i generaliserede li-
        neære modeller ....................................... 189
2.5.5 Eksempler på estimation i generaliserede lineære modeller ..... 193
2.5.6 Residualer ..... 200
2.5.7 Fordeling af fittede værdier og residualer ..... 204
2.5.8 Residualer, standardisering og studentisering ..... 211
2.5.9 Forudsigelse, prædiktion ..... 214

2.6 Test for modeltilpasning i generaliseret lineær model ..... 216
2.6.1 Residualdevians svarende til generaliseret lineær model ..... 217
2.6.2 Estimation af dispersionsparameteren $\sigma^2$ ..... 222

2.7 Eksempler på regressions- og homogenitetsmodeller ..... 223
2.7.1 Regressionsmodeller ..... 223
2.7.2 Homogenitetshypotesen, den minimale model ..... 230

2.8 Parametrisk repræsentation af modeller ..... 243
2.8.1 Introduktion ..... 243
2.8.2 Kontinuerte kovariable ..... 246
2.8.3 Intercept led ..... 247
2.8.4 Kvalitative kovariable, faktorvariable ..... 248
2.8.5 Parametrisk repræsentation af blandede led ..... 254

2.9 Modellmatrix, kontraster ..... 254
2.9.1 Modellmatrix for kontinuerte kovariable ..... 255
2.9.2 Incidensmatrix for faktorvariabel ..... 255
2.9.3 Parametrisering af faktormodell ved kontraster ..... 258
2.9.4 Modellmatrix svarende til blandede led ..... 260
2.9.5 Incidensmatrix svarende til to klassifikationskriterier ..... 261
2.9.6 Klassifikationer med hierarkisk ordnet indeksmængde ..... 265
2.9.7 Partiel ordning af klassifikationer ..... 265
2.9.8 Aliasrelationer mellem parametre, marginalitet ..... 268
2.10 Modelformler .................................................. 275
  2.10.1 Hierarkisk organiseret indeksmængde, underordnede faktorer .................................. 278
2.11 Test for modelreduktion ....................................... 280
  2.11.1 Indledning, strategier for modeltilpasning .................. 280
  2.11.2 Test af enkelte parametre ................................. 282
  2.11.3 Test af delhypotese ...................................... 286
  2.11.4 Modelreduktion ved successiv testning i hierarkiske hypoteser ................................ 299
  2.11.5 Modelreduktion ved partielle tests ........................ 302
  2.11.6 Total deviansopspaltning svarende til successiv tilføjelse eller fjernelse af led ............ 307
  2.11.7 Successiv testning ved estimation af dispersionsparameter ...................................... 311
2.12 Vekselvirkning .................................................. 312
2.13 Tosidig inddeling ............................................... 318
2.14 Forklaringsgrad $R^2$ .......................................... 330
  2.14.1 Korrigeret forklaringsgrad $R'^2$ ......................... 330
  2.14.2 Akaike’s informationskriterium $A_H$ .................... 331
2.15 Valg af model og modelkontrol ............................... 332
  2.15.1 Generelt om modelvalg og kontrol ....................... 332
  2.15.2 Brug af residualer til kontrol af systematiske afvigelser fra modellen ..................... 335
  2.15.3 Kontrol af enkeltobservationer, leverage ................ 338
  2.15.4 Kontrol af enkeltobservationers overensstemmelse, residual ..................................... 340
  2.15.5 Kontrol af enkeltobservationers indflydelse (influens) .............................................. 342
  2.15.6 Vurdering af enkeltobservationer, sammenfatning .. 346
2.16 Referencer: ..................................................... 348
3 Modeller for binære responsvariable

3.1 Binomialfordelingen som eksponentiel dispersionsparameterfamilie, kanonisk link ........................................ 351
3.1.1 Odds, logit .......................................................... 351
3.1.2 Sammenligning af hændelser .................................. 353
3.1.3 Generaliserede lineære modeller for binomialt fordelt variable .................................................. 354

3.2 Regressionsmodeller .................................................. 357
3.2.1 Logistisk regression ............................................. 358
3.2.2 Regression ved andre link-funktioner ...................... 365
3.2.3 Regressionsmodeller med flere forklarende variable .. 371

3.3 Faktorielle opstillinger med binært respons .................. 375
3.3.1 Opstillinger med to faktorer .................................. 375
3.3.2 Vekselvirkning og valg af linkfunktion ...................... 380
3.3.3 Yule’s krydsprodukt ratio og betingede odds .......... 387
3.3.4 Rasch model for itemanalyse, latente parametre ....... 388

3.4 Tovejs antalstabeller svarende til binært respons .......... 391
3.4.1 Indledning .......................................................... 391
3.4.2 Konfidentsintervaller ved sammenligning af to hyppigheder .................................................. 392
3.4.3 Prospektive og retrospektive undersøgelser ............. 399
3.4.4 Modeller for prospektive studier ......................... 401
3.4.5 Retrospektive studier .......................................... 403
3.4.6 Modeller for gentagne målinger ......................... 410

3.5 Modeller for parvise sammenligninger ....................... 418
3.5.1 Bradley-Terry modellen .................................... 419

3.6 Referencer ......................................................... 422
<table>
<thead>
<tr>
<th>4 Modeller for flerdimensionale antalstabeller</th>
<th>427</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduktion til modeller med kategorisk respons</td>
<td>427</td>
</tr>
<tr>
<td>4.1.1 Uafhængige Poisson-fordelte observationer</td>
<td>429</td>
</tr>
<tr>
<td>4.1.2 Modeller for Multinomial stikprøveudvælgelse</td>
<td>431</td>
</tr>
<tr>
<td>4.1.3 Produkt-multinomial stikprøveudvælgelse</td>
<td>432</td>
</tr>
<tr>
<td>4.2 Modeller med endimensionalt respons, Multinomialfordelingen</td>
<td>433</td>
</tr>
<tr>
<td>4.2.1 Indledning</td>
<td>433</td>
</tr>
<tr>
<td>4.2.2 Odds- og oddsratioer, ét klassifikationskriterium</td>
<td>437</td>
</tr>
<tr>
<td>4.2.3 Baseline odds</td>
<td>439</td>
</tr>
<tr>
<td>4.2.4 Nabokategori odds</td>
<td>443</td>
</tr>
<tr>
<td>4.2.5 Fortsættelses-odds</td>
<td>445</td>
</tr>
<tr>
<td>4.2.6 Kumulative logit'ær</td>
<td>447</td>
</tr>
<tr>
<td>4.2.7 Andre linkfunktioner</td>
<td>449</td>
</tr>
<tr>
<td>4.2.8 Regressionsmodeller</td>
<td>450</td>
</tr>
<tr>
<td>4.3 Modeller med flere klassifikationskriterier</td>
<td>461</td>
</tr>
<tr>
<td>4.3.1 Flere klassifikationskriterier, Yule's krydsprodukt-ratio</td>
<td>461</td>
</tr>
<tr>
<td>4.3.2 Tovejs antalstabeller, multinomial stikprøveudvælgelse</td>
<td>463</td>
</tr>
<tr>
<td>4.4 Log-lineære modeller</td>
<td>466</td>
</tr>
<tr>
<td>4.5 Betinget uafhængighed</td>
<td>467</td>
</tr>
<tr>
<td>4.5.1 Uafhængighedsgrafer</td>
<td>468</td>
</tr>
<tr>
<td>4.6 Tovejs antalstabeller</td>
<td>469</td>
</tr>
<tr>
<td>4.6.1 Multinomial stikprøveudvælgelse</td>
<td>470</td>
</tr>
<tr>
<td>4.7 Grafiske modeller</td>
<td>476</td>
</tr>
<tr>
<td>4.7.1 Faktorisering, Reducible komponenter</td>
<td>477</td>
</tr>
<tr>
<td>4.7.2 Dekomposable modeller</td>
<td>477</td>
</tr>
<tr>
<td>4.7.3 Strategier for modelvalg</td>
<td>478</td>
</tr>
</tbody>
</table>
4.8 Generel formulering af modeller for flerdimensionalt respons 479
  4.8.1 Relation til teorien for Markovfelter ............... 479
  4.8.2 Grafiske modeller og Gibbs tilstande ............... 482
  4.9 Referencer ........................................ 483

5 Hierarkiske modeller for endimensionale normalfordelinger 485
  5.1 Indledning og notation ................................ 485
  5.2 Ensidet variansanalyse i den systematiske model ...... 488
  5.3 Ensidet variansanalyse i den tilfældige model .......... 496
    5.3.1 Estimation af parametre i den tilfældige model ... 502
    5.3.2 Test af homogenitetshypotese i den tilfældige model 505
  5.4 Likelihoodbaseret estimation i den tilfældige model .... 509
  5.5 SAS® procedure til analyse af den tilfældige model ...... 517
    5.5.1 GLM ........................................... 517
    5.5.2 Mixed ........................................ 521
    5.5.3 Varcomp ....................................... 526
  5.6 Eksempler på den tilfældige model ..................... 526
  5.7 Normalfordelingsmodeller med tilfældigt varierende varians. 532
  5.8 Referencer: ........................................ 545

6 Hierarkiske modeller for eksponentielle dispersionsmodeller 547
  6.1 Indledning ......................................... 547
    6.1.1 Den systematiske model .......................... 548
    6.1.2 Den tilfældige model ........................... 549
  6.2 Bernoullifordelingen ................................ 556
  6.3 Den geometriske fordeling ............................ 571
  6.4 Poissonfordelingen .................................. 579
6.5 Eksponentialfordelingen .................................. 589
6.6 Fordeling af empiriske varianser for normalfordelte variable 599
  6.6.1 Den systematiske model .............................. 599
  6.6.2 Den tilfældige model ................................ 600
  6.6.3 Fortolkning af parametre i strukturfordelingen af $\sigma^2$ 601
  6.6.4 Marginal fordeling af stikprøvevariansen ............... 602
  6.6.5 Estimation af parametre i strukturfordeling ........... 607
6.7 Den flerdimensionale normalfordeling ................... 609
  6.7.1 Den systematiske model ............................. 609
  6.7.2 Den tilfældige model ................................ 611
6.8 Oversigtstabeller ....................................... 626
6.9 Referencer ............................................. 631

7 Lineære normalfordelingsmodeller ......................... 633
  7.1 Balancerede regressionsmodeller med varierende koefficienter 633
    7.1.1 Indledning ......................................... 633
    7.1.2 Den systematiske model ............................ 634
    7.1.3 Den tilfældige model ................................ 646
  7.2 Ubalancerede regressionsmodeller med varierende koefficienter661
    7.2.1 Den systematiske model ............................ 661
    7.2.2 Den tilfældige model ................................ 665
  7.3 Tidsrække modeller .................................... 670
    7.3.1 Den endimensionale autoregressive proces af første
          orden. .............................................. 670
    7.3.2 Flerdimensionale tidsrække modeller ................ 672
  7.4 Blandede modeller ..................................... 675
  7.5 Referencer ............................................. 676
# Indhold

8 Aposteriorifordelinger 677

8.1 Betingede fordelinger, Bayes’ sætning 678
   8.1.1 Bayes’ sætning 678

8.2 Apriori- og aposteriorifordelinger 679

8.3 Aposteriorifordelinger for eksponentielle dispersionsmodeller 683
   8.3.1 Resume af afsnit 6 683
   8.3.2 Generelle resultater vedrørende aposteriorifordelinger 684
   8.3.3 Binomial-beta sampling 691
   8.3.4 Negativ binomial-beta sampling 699
   8.3.5 Poisson-Gamma sampling 700
   8.3.6 Exponential reciprok gamma sampling 703
   8.3.7 Normalfordeling med samme varians 705
   8.3.8 Empiriske varianser fra normalfordelte observationer 707
   8.3.9 Normalfordelingsmodeller med tilfældigt varierende varians: 711

8.4 Filtrering af en tidsrække 713

8.5 Den flerdimensionale normalfordeling 715

8.6 Regressionsmodeller 728

8.7 Tidsrækkeomodeller 735
Afsnit 4

Modeller for flerdimensionale antalstabeller

4.1 Introduktion til modeller med kategorisk respons

I dette afsnit vil vi betragte modeller for flerdimensionale antalstabeller, dvs. flerdimensionale tabeller, hvor elementerne i de enkelte celler er et antal. Hver af tabellens dimensioner tages som udtryk for en klassifikation med hensyn til et klassifikationsskriterium \((A, B, \ldots)\). Antallet af klassifikationskriterier kaldes tabellens dimension.

Som vi så i det foregående afsnit, kan en sådan antalstabel fremkomme på flere måder, fx som resultat af

- en prospektiv undersøgelse
- eller en retrospektiv undersøgelse, fx en tværsnitsundersøgelse
Ved modelleringen er det ydermere af betydning at sondre mellem variable (klassifikationer), der har karakter af forklarende variable og variable, der kan opfattes som responsvariable. I en prospektiv undersøgelse er det sædvanligvis relativt enkelt at foretage denne sondring, hvorimod det kan være mere vanskeligt i tværsnitsundersøgelser. Selv om likelihoodfunktionen sædvanligvis vil være den samme, hvadenten en variabel optræder i en model som en responsvariable, eller som en forklarende variabel, vil det naturligvis være af betydning for fortolkningen af en model, om en given klassifikation optræder som en forklarende variabel eller som en responsvariable.

Nedenstående skema (modifikation af Bhapker og Koch (1968)) angiver forskellige muligheder for fortolkning af en flerdimensional antalstable:

<table>
<thead>
<tr>
<th>Forklarende variable</th>
<th>Responsvariable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antal faktorer</td>
<td>Antal klassifikationskriterier</td>
</tr>
<tr>
<td>Ingen</td>
<td>Flere</td>
</tr>
<tr>
<td>Én eller flere</td>
<td>Én eller flere</td>
</tr>
<tr>
<td>Flere</td>
<td>Ingen</td>
</tr>
</tbody>
</table>

Endelig kan der som forklarende variable også indgå kontinuerne kovariable.

I afsnit 4.2 vil vi diskutere multinomialfordelingen og en række forskellige parametriseringer af multinomialfordelingen, når den optræder som fordeling for et enddimensionalt respons under en række forskellige kombinationer af forklarende variable.

I afsnittene 4.3 til 4.8 diskuteres modellering af et flerdimensional respons. Specielt introducerer vi i afsnit 4.4 de såkaldte log-lineære modeller.

Som et eksempel på problemstillingerne vil vi betragte en todimensional antalstable som illustreret i nedenstående generelle opstilling:

\[
\text{OBSERVEREDE ANTAL, } x_{ij}
\]
4.1 Modeller med kategorisk respons

$$\begin{array}{|c|cccc|c|}
\hline
\text{Klassifikation A} & \text{Klassifikation B} & & & & \\
\hline
\text{1} & x_{1,1} & x_{1,2} & \ldots & x_{1,J} & x_{1,+} \\
\text{2} & x_{2,1} & x_{2,2} & \ldots & x_{2,J} & x_{2,+} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\text{I} & x_{I,1} & x_{I,1} & \ldots & x_{I,J} & x_{I,+} \\
\text{Ialt} & x_{+,1} & x_{+,2} & \ldots & x_{+,J} & x_{+,+} \\
\hline
\end{array}$$

Sædvanligvis vil det underliggende design svare til en af følgende tre modeller:

- Observation af $I \times J$ uafhængige $P(\lambda_{i,j})$-fordelte variable (To faktorer, ingen klassifikation af respons, kun optælling).
- Observation af én multinomialt fordelt størrelse (af dimensionen $I \times J$) (Ingen faktorer, todimensionalt respons).
- Observationer af $I$ uafhængige multinomialt fordelt størrelse (hver af dimensionen $J$) (Én faktor, endimensionalt respons)

4.1.1 Uafhængige Poisson-fordelede observationer:

Under denne model antages ethvert element $x_{i,j}$ i tabellen at være fremkommet som udfald af en Poisson fordelt variabel $X_{i,j} \in P(\lambda_{i,j})$, hvor de enkelte udvalg er uafhængige.

Modellen svarer til eksempel 2.13.2

Frekvensfunktionen svarende til denne model er

$$f(x) = \prod_{i,j} \frac{\lambda_{i,j}^{x_{i,j}} \exp(-\lambda_{i,j})}{x_{i,j}!} \quad (4.1.1)$$

De forventede værdier af celleværdierne er $E[X_{i,j}] = \lambda_{i,j}$. Vi bemærker at marginalsommerne ligeledes er Poisson-fordelede. Eksempelvis har vi for $X_{i,+} = \sum_{j=1}^{J} x_{i,j}$, at $X_{i,+} \in P(\lambda_{i,+})$ med $\lambda_{i,+} = \sum_{j=1}^{J} \lambda_{i,j}$.

Log-likelihoodfunktionen svarende til modellen (4.1.1) er

$$l(\lambda; x) = \sum_{i,j} x_{i,j} \ln(\lambda_{i,j}) - \sum_{i,j} \lambda_{i,j} \quad (4.1.2)$$
Afsnit 4. Modeller for flerdimensionale antalstabeller

Modellen udgør en eksponentiel familie med den kanoniske parameter \( \theta_{i,j} = \ln(\lambda_{i,j}) \). Den fulde model tillader \( \lambda_{i,j} \) at variere frit, dvs \( \theta \in \mathbb{R}^{I \times J} \).

Naturlige reduktioner af denne model er reduktioner svarende til to-faktor modeller, som betragtet tidligere, dvs. først en vurdering af "vekselvirkningen", nemlig et forsøg på tilpasning af modellen

\[
\theta_{i,j} = \alpha_i + \beta_j ,
\]

Modellen kan formelt formuleres ved kontraster mellem \( \theta_{i,j} \) ved

\[
\Delta_{i_1,j_1,i_2,j_2}^{A,B} = (\theta_{i_1,j_1} - \theta_{i_1,j_2}) - (\theta_{i_2,j_1} - \theta_{i_2,j_2}) = 0
\]

Eksempel 4.1.1 Trafikuheld, klassificeret efter kvartal og uheldskategori

Tabel 4.1 viser politiets registreringer af motorkøretøjsuheld med personskader for dagtimerne i 1990. De registrerede uheld er klassificeret efter uheldskategori og kvartal.

Tabel 4.1. Uheld med personskade i 1990, klassificeret efter kvartal og uheldskategori

<table>
<thead>
<tr>
<th>Kvartal</th>
<th>Ene - uheld</th>
<th>Indhentnings - uheld</th>
<th>Møde - uheld</th>
<th>Sving og kryds</th>
<th>Andre</th>
<th>Ialt</th>
</tr>
</thead>
<tbody>
<tr>
<td>januar</td>
<td>105</td>
<td>67</td>
<td>54</td>
<td>310</td>
<td>29</td>
<td>565</td>
</tr>
<tr>
<td>april</td>
<td>146</td>
<td>76</td>
<td>85</td>
<td>485</td>
<td>37</td>
<td>829</td>
</tr>
<tr>
<td>juli</td>
<td>114</td>
<td>94</td>
<td>100</td>
<td>461</td>
<td>32</td>
<td>801</td>
</tr>
<tr>
<td>oktober</td>
<td>105</td>
<td>62</td>
<td>74</td>
<td>314</td>
<td>24</td>
<td>579</td>
</tr>
<tr>
<td>Ialt</td>
<td>470</td>
<td>299</td>
<td>313</td>
<td>1570</td>
<td>122</td>
<td>2774</td>
</tr>
</tbody>
</table>

Det vil naturligt at opfatte de 4 \( \times \) 5 antal i tabellen som realisationer af 4 \( \times \) 5 uafhængige \( P(\lambda_{i,j}) \)-fordelte variable, da ingen af antallene kunne være fastlagt på forhånd.
4.1.2 Modeller for Multinomial stikprøveudvælgelse:

Som et eksempel på modeller med flerdimensionalt respons vil vi betragte nedenstående tabel:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kvind</td>
<td>234</td>
<td>559</td>
<td>1 157</td>
<td>5 826</td>
<td>2 553</td>
<td>10 329</td>
</tr>
<tr>
<td></td>
<td>2.3 %</td>
<td>5.4 %</td>
<td>11.2 %</td>
<td>56.4 %</td>
<td>24.7 %</td>
<td></td>
</tr>
<tr>
<td>Mænd</td>
<td>183</td>
<td>446</td>
<td>1 005</td>
<td>3 841</td>
<td>1 428</td>
<td>6 903</td>
</tr>
<tr>
<td></td>
<td>2.7 %</td>
<td>6.5 %</td>
<td>14.6 %</td>
<td>55.6 %</td>
<td>20.7 %</td>
<td></td>
</tr>
<tr>
<td>Ialt</td>
<td>417</td>
<td>1 005</td>
<td>2 162</td>
<td>9 667</td>
<td>3 981</td>
<td>17 232</td>
</tr>
<tr>
<td></td>
<td>2.4 %</td>
<td>5.8 %</td>
<td>12.5 %</td>
<td>56.1 %</td>
<td>23.1 %</td>
<td></td>
</tr>
</tbody>
</table>

Tabellen repræsenterer en række buspassagers svar på et spørgsmål vedrørende deres tilfredshed med overholdelsen af køreplanen. Svarene er indsamlet blandt passagerer i busser, der var rettidige.

Svarene er klassificeret såvel efter køn som efter tilfredshed. Mens klassifikationen efter køn blot er kategorisk, er klassifikationen efter tilfredshed en ordnet klassifikation.

Svarene kan tænkes at være fremkommet ved en enkelt stikprøveudtagning, hvor totalsummen $N = \sum_{i,j} x_{i,j}$ af observationerne er vær fastlagt af stikprøveplanen, og de $N$ observationer er derefter krydsklassificeret i henhold til faktorerne beskrevet ved $I$ og $J$.

Frekvensfunktionen svarende til denne model er multinomialfordelingen,

$$f(x) = \frac{N!}{\prod_{i,j} x_{i,j} !} \prod_{i,j} p_{i,j}^{x_{i,j}}$$ (4.1.3)

hvor $\sum_{i,j} p_{i,j} = 1$, og log-likelihoodfunktionen er

$$l(p; x) = \frac{N!}{\prod_{i,j} x_{i,j} !} \sum_{i,j} x_{i,j} \ln(p_{i,j})$$ (4.1.4)

hvor $\sum_{i,j} p_{i,j} = 1$.

De forventede værdier af celleværdierne er $E[X_{i,j}] = Np_{i,j}$.
Modellen (4.1.3) er en eksponentiel familie.

De marginale tabeller er de to endimensionale tabeller, der fremkommer ved at klassificere alene med hensyn til køn (uden hensyntagen til tilfredshed) og ved at klassificere alene med hensyn til tilfredshed (uden hensyntagen til køn).

Den multinomiale stikprøvemodel (4.1.3) kaldes i en række sammenhænge en model for en tværsnitsundersøgelse, (afsnit 3.4.5), svarende til at man på et givet tidspunkt undersøger et udsnit af en population og klassificerer denne efter en række karakteristika.

I andre sammenhænge, hvor man fokuserer på faktor-respons sammenhænge, ser man undertiden modellen betegnet som en model uden faktorer (blot den undersøgte population) og et todimensionalt respons, svarende til den todimensionale klassifikation.

### 4.1.3 Produkt-multinomial stikprøveudvælgelse:

Under denne stikprøvemodel tænkes stikprøveudtagningen at være stratificeret således, at én af de variable optræder som design-variable (eller forklarende variable). Antallet af observationer fra hvert stratum er fastlagt på forhånd gennem stikprøveplanen, og observationerne er derefter klassificeret efter værdierne af den anden variable. I eksemplet med buspassagerne, kunne man eventuelt have fastsat antallet af mandlige resp. kvindelige respondenter på forhånd.

Såfremt eksempelvis de \( I \) rækker i tabellen repræsenterer \( I \) værdier af en stratifikationsvariabel med fastlagte værdier \( x_{i,+} \), finder vi frekvensfunktionen svarende til den \( i \)-te række som

\[
f(x_i) = \frac{x_{i,+}!}{\prod_{j=1}^{I} x_{i,j}!} \prod_{j=1}^{J} p_{i,j}^{x_{i,j}}
\]

(4.1.5)

hvor \( \sum_{j=1}^{J} p_{i,j} = 1 \) for \( i = 1, 2, \ldots, I \)

Idet stikprøverne fra de \( I \) rækker er indbyrdes uafhængige, finder man frekvensfunktionen for hele observationssættet som

\[
f(x) = \prod_{i=1}^{I} \frac{x_{i,+}!}{\prod_{j=1}^{I} x_{i,j}!} \prod_{j=1}^{J} p_{i,j}^{x_{i,j}}
\]

(4.1.6)
4.1 Multinomialfordelingen

De forventede værdier af celleværdierne er $E[X_{i,j}] = n_{i,+}p_{i,j}$.

Log-likelihoodfunktionen svarende til produkt-multinomialfordelingsmodellen er

$$l(p) = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j} \ln(p_{i,j}) \quad (4.1.7)$$

hvor $\sum_{j} p_{i,j} = 1$

**Bemærkning 1** *Produkt-multinomial modellen svarer til den betingede fordeling af celleværdier i den rene multinomial model*

Vi bemærker, at frekvensfunktionen svarende til produkt-multinomial modellen angiver den betingede fordeling af observationerne under multinomialmodellen (4.1.3) for givne søjlesummer $X_{i,+} = n_{i,+}$.

Uafhængige Binomialt-fordelte observationer:

Et specialtilfælde af produktmodellen fås for kontrollerede forsøg med en binær responsvariabel som betragtet i afsnit 3.3.

4.2 Modeller med endimensionalt respons, Multinomialfordelingen

4.2.1 Indledning

Vi vil først betragte modeller, hvor responset er resultatet af en klassifikation efter et enkelt kriterium.

Vi vil her under ét betragte de forskellige muligheder for kombinationer af forklarende variable, dvs fra en situation med en enkelt multinomial stikprøve uden forklarende variable til en situation med flere stikprøver, svarende til forskellige værdier af kontinuerte kovariable og/eller faktorvariable.
Afsnit 4. Modeller for flerdimensionale antalstabeller

Vi vil karakterisere responsfordelingen ved \( r - 1 \) prædiktorer, \( \vartheta_1, \ldots, \vartheta_{r-1} \), hvor den \( j \)te prædiktor er en funktion af sædet, \( (p_1, \ldots, p_r) \) af responssandsynligheder. I lighed med formuleringen af de generaliserede lineære modeller vil vi formulere en model for prædiktoren \( \vartheta_j \), som er lineær i de forklarende variable. I afsnittene 4.2.3 til 4.2.7 vil vi diskutere forskellige valg af prædiktorer, \( \vartheta_j \).

I overensstemmelse med betragtningerne i afsnit 2.8 vil vi forestille os afhængigheden af de forklarende variable beskrevet ved en modelmatrix, \( X \), hvis \( i \)te række \( x_i^T \) beskriver de forklarende variable svarende til den \( i \)te multinomialfordeling.

Principielt kan man vælge et sæt parametre, \( \beta_j \) til hver af de lineære prædiktorer, \( \vartheta_j \), (og eventuelt også et sæt forklarende variable, \( x_{ij}^* \) til hver af disse). Man vil da opstille modeller af formen

\[
\vartheta_{ij} = x_{ij}^T \beta_j \tag{4.2.1}
\]

for hver af de \( r - 1 \) prædiktorer, \( \vartheta_j \).

Sædvanligvis vil man dog tilstræbe at benytte ét sæt af forklarende variable til beskrivelse af hele parametersættet; \( (p_{i1}, \ldots, p_{ir}) \), for den \( i \)te multinomialfordeling, dvs udtryk af formen

\[
\vartheta_{ij} = x_i^T \beta_j \tag{4.2.2}
\]

For en responsvariabel, som er en ren tællevariabel, der tæller antallet af gange, en given hændelse indtræffer, vil responsfordelingen ofte kunne beskrives ved en Poissonfordeling, og man vil bruge en generaliseret lineær model til modellering af tabellen, som fx beskrevet i eksempel 2.13.2.

Hvis den variable kun har to kategorier, \( A \) og \( A^c \) og tabellen indeholder antallene af responser i hver af de to kategorier, har man en situation med binært respons, som beskrevet i afsnit 3.

For en responsvariabel med flere end to kategorier, vil den naturlige fordeling af responset ofte være multinomialfordelingen.

Multinomialfordelingen er et eksempel på en flerdimensional eksponentiel familie, der i en vis forstand kan opfattes som en generalisering af binomialfordelingen. Imidlertid giver multinomialfordelingen mulighed for væsentligt mere struktur, hvilket øger kompleksiteten af modellerne, også selv om vi holder os til affine hypoteser for de kanoniske parametre.
4.2 Multinomialfordelingen

Vi minder om, at familien af multinomialfordelinger, Mult\( (n, p_1, p_2, \ldots, p_r) \) er en fuld ekspontiel familie med kanonisk stikprøvefunktion \( t(\cdot) \) givet ved:

\[
t(x) = \begin{pmatrix} n_1(x) \\ \vdots \\ n_r(x) \end{pmatrix} \in \mathbb{R}^r,
\]

hvor \( n_j(x) \) angiver antallet af udfald af kategori \( j \), og med kanonisk parameter

\[
\vartheta = \begin{pmatrix} \ln(p_1) \\ \vdots \\ \ln(p_r) \end{pmatrix} \in \mathbb{R}^r
\]

Relationen \( \sum_j p_j = 1 \) indebærer, at det kanoniske parameterrum kun har dimensionen \( m - 1 \).

Enhedsmiddelværdiafbildningen er

\[
\tau_j(\vartheta) = \frac{\exp(\vartheta_j)}{\sum_{\nu=1}^{r} \exp(\vartheta_\nu)} = p_j,
\]

og elementerne i enhedsdispersionsmatricen er

\[
V_{i,j}(p) = \begin{cases} 
p_i(1-p_i) & \text{for } i = j \\
-p_i^2 p_j & \text{for } i \neq j
\end{cases}
\]

Dispersionsmatricen \( V(p) \) er singulær, da der gælder \( \sum_j p_j = 1 \).

Middelværdi og dispersion svarende til antalsparameteren \( n \) fås ved at multiplikere enhedsmiddelværdien og enhedsdispersionsmatricen med antalsparameteren \( n \) i analogi med de endimensionale ekspontielle dispersionsmodeller.

Parametreriseringen ved \( \vartheta \) er en overparametrerisering. Således vil sættet \( \vartheta_0 \), og sættet

\[
\vartheta_1 = \vartheta_0 + c1,
\]

hvor \( c \) er et vilkårligt reelt tal, give anledning til samme multinomialfordeling.

I almindelighed fortæller den enkelte komponent af \( \vartheta \) således ikke umiddelbart noget om den absolutte værdi af sandsynlighederne \( p_j \).

Oftest vil man imidlertid også være mere interesseret i relationer mellem sandsynlighederne, end i deres absolutte værdier.
Eksempel 4.2.1 Tilfredshedsangivelser fra buspassagerer

Som led i de løbende undersøgelser af kundetilfredsheden foretager et bus-
selskab stikprøveundersøgelser blandt passagerer, hvor passagererne bli-
ver bedt om at svare på en række spørgsmål vedrørende servicekvaliteten.
Samtidig med registreringen af kundernes tilfredshed registrerer man også
forskellige objektive størrelser, herunder bussens eventuelle forsinkelse.

Et af spørgsmålene vedrører kundens tilfredshed med overholdelsen af køre-
planen. Spørgsmålet har svarkategorierne "Meget utilfreds", "Utilfreds", "Bå-
de og", "Tilfreds", "Meget tilfreds".

Nedenstående tabel viser fordelingen af svarene på dette spørgsmål for
10 329 kvindelige passagerer i rettidige busser.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Antal</td>
<td>234</td>
<td>559</td>
<td>1 157</td>
<td>5 826</td>
<td>2 553</td>
<td>10 329</td>
</tr>
<tr>
<td>Andel (pct)</td>
<td>2.27</td>
<td>5.41</td>
<td>11.20</td>
<td>56.40</td>
<td>24.72</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Idet vi antager, at passagerne er udvalgt uafhængigt af hinanden, kan man
opfatte fordelingen af disse 10 329 svar som udfaldet af en
Mult(10 329, p₁, p₂, . . ., p₅)-fordelt variabel, hvor (p₁, p₂, . . ., p₅) kan for-
tolkes som andelen af kvindelige passagerer i rettidige busser, som vil svare
hhv "Meget utilfreds", "Utilfreds", etc. på dette spørgsmål.

Fordelingen er illustreret grafisk i nedenstående figur.
Svarkategoriene er ordnede, gående fra "Meget utilfreds" til "Meget tilfreds". Ordningen er afspejlet i figuren ved at andelen af besvarelser af kategorien "Meget utilfreds" er markeret umiddelbart til venstre for klokken 12, derefter fortsættes mod uret indtil kategorien "Meget tilfreds", der er afbildet umiddelbart til højre for klokken 12.

Bemærkning 1 *Likert skalaen*


4.2.2 Odds- og oddsratioer, ét klassifikationskriterium

Betragt et endeligt udfaldsrum $\Omega$ for et forsøg.

Vi betragter en klassifikation $\Omega = \bigcup_{j=1}^{n} A_j$ i disjunkte kategorier med de
tilknyttede sandsynligheder

\[ p_j = P [A_j], \quad j = 1, \ldots, r. \]

I tilfældet, hvor der kun er to mulige kategorier, \( r = 2 \), og der udføres \( n \) uafhængige eksperimenter, kan man som nævnt beskrive fordelingen af antallet af indtrufne hændelser i kategorien \( A_1 \) ved en binomialfordeling. I afsnit 3.1.1 indførte vi begrebet "odds" til beskrivelse af sandsynligheden \( p_1 = P [A_1] \) som

\[ \theta = \frac{p_1}{1 - p_1} \quad (4.2.5) \]

Vi så, at logaritmen til odds, logit'en \( \vartheta = \ln(\theta) \) netop var den kanoniske parameter i familien af Bernoulli-fordelinger, (eller familien af binomialfordelinger).

I tilfældet, hvor der er flere end to kategorier, \( 2 < r \), vil den tilsvarende fordeling af antallet af hændelser i kategoriene \( A_1, A_2, \ldots, A_r \) være en multinomialfordeling med den \( r \)-dimensionale kanoniske parameter \( \vartheta \) bestemt ved \( \vartheta_j = \ln(p_j) \). Modellen har imidlertid kun dimensionen \( r - 1 \), idet der jo gælder, at \( \sum p_j = 1 \). Modellen er således defineret af \( r - 1 \) kontraster mellem disse kanoniske parametre.

Ovenstående odds-begreb, (4.2.5), der blev indført i afsnit 3.1.1, var dækkende for situationer, hvor man blot betragtede en enkelt hændelse, \( A \) og dens komplement \( A^c \). Når der er flere disjunkte hændelser, der sammenlignes, er det ofte af interesse at betragte forhold mellem sandsynlighederne for forskellige hændelser, og man har derfor udvidet odds-begrebet fra afsnit 3.1.1 til at dække situationer med flere disjunkte hændelser. I det følgende vil vi betragte forskellige parametriseringer af multinomialfordelingen baseret på sådanne udvidelser af odds-begrebet. Disse generaliserede odds vil ligeledes blive betegnet odds, og de tilsvarende logaritmerede værdier vil blive betegnet logit'\( \text{er}.\)

Grunden til at vi anfører disse forskellige parametriseringer af multinomialfordelingen er, at i nogle anvendelser er det muligt at etablere lineære modeller svarende til én parametrering, i andre anvendelser er det en anden parametrising, der muliggør en model, der er lineær i de forklarende variable. I praksis er det derfor fordelagtigt at have flere forskellige parametrisinger til rådighed.
Agresti (1990) beskriver en lang række modeller til analyse af flerdimensio-
nalt kategorisk respons. McCullagh (1980) har specielt diskuteret forskellige 
regressionsmodeller.

4.2.3 Baseline odds

Definition 4.2.1 Baseline odds, baseline logit

Lad $A^0$ angive en vilkårlig referencekategori. Vi definerer da odds for hændelserne $A_j$ ved

$$
\theta_j = \frac{P [A_j]}{P [A^0]} = \frac{p_j}{p^0}
$$

(4.2.6)

Den valgte referencekategori, $A^0$, kaldes undertiden baseline kategorien og de tilsvarende odds kaldes for baseline odds.

Ofte vælger man at organisere kategorierne sådan at referencekategori'nen netop er den sidste kategori, $A_r$, dvs sådan at

$$
\theta_j = \frac{P [A_j]}{P [A_r]} = \frac{p_j}{p_r}
$$

(4.2.7)

Det tilsvarende logaritmiske mål:

$$
\vartheta_j = \ln(\theta_j), \quad j = 1, 2, \ldots, r - 1
$$

(4.2.8)

kalder baseline logit'eren. Forskellen mellem logit'erne, $\vartheta_j$ og $\vartheta_{j'}$ for to hændelser, afhænger ikke af baselinehændelsen $A^0$.

Sammenligning mellem sandsynligheder svarende til de forskellige kategorier kan da udføres ved sammenligning mellem de tilsvarende odds. Vi indfører odds-ratioen for $A_i$ mod $A_j$, $i \neq j$ som

$$
\omega_{i,j} = \frac{\theta_i}{\theta_j} = \frac{p_i/p^0}{p_j/p^0} = \frac{P [A_i]}{P [A_j]},
$$

(4.2.9)
Asfnt 4. Modeller for flerdimensionale antalstabeller

Det tilsvarende logaritminiske mål, differensen mellem baseline logit’erne,

\[ \Delta_{i,j} = \ln(\omega_{i,j}) = \ln(p_i) - \ln(p_j) = \vartheta_i - \vartheta_j \]  \hspace{1cm} (4.2.10)

bliver da netop en kontrast mellem logaritmen til odds for de to hændelser. Kontrasterne \( \Delta_{i,j} \) afhænger ikke af hvilken hændelse \( A^0 \), man har valgt som referencehændelse (baseline).

Man kan således reparametrisere en multinomialfordeling ved at betrægte odds-ratioerne, eller logaritmen til odds-ratioerne.

Hvis der er i alt \( r \) forskellige hændelser, er der i alt \( \binom{r}{2} \) forskellige par af hændelser, for hvilke man kan konstruere odds-ratioer. Man kan vise, at såfremt man har valgt \( r - 1 \) af disse par, er de resterende kontraster bestemt.

Eksempel 4.2.2 Tilfredshedsangivelser fra buspassagerer, baseline odds

Vi betrægter atter situationen fra eksempel 4.2.1.

En naturlig referencekategori (baseline) kunne her være svarkategorien "Både og". Nedenstående tabel viser de beregnede (observerede) odds og logit’er for de forskellige svarkategorier bestemt med kategorien "Både og" som referencekategori.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Antal</td>
<td>234</td>
<td>559</td>
<td>1 157</td>
<td>5 826</td>
<td>2 553</td>
<td>10 329</td>
</tr>
<tr>
<td>Odds</td>
<td>0.2022</td>
<td>0.4831</td>
<td>1.0000</td>
<td>5.035</td>
<td>2.2066</td>
<td></td>
</tr>
<tr>
<td>Logit</td>
<td>-1.5983</td>
<td>-0.7274</td>
<td>0.0000</td>
<td>1.6165</td>
<td>0.7914</td>
<td></td>
</tr>
</tbody>
</table>

Havde man i stedet valgt den sidste kategori, "Meget tilfreds", som referencekategori (baseline), havde man fået følgende tabel

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Antal</td>
<td>234</td>
<td>559</td>
<td>1 157</td>
<td>5 826</td>
<td>2 553</td>
<td>10 329</td>
</tr>
<tr>
<td>Odds</td>
<td>0.0917</td>
<td>0.2190</td>
<td>0.4532</td>
<td>2.2821</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Logit</td>
<td>-2.3897</td>
<td>-1.5189</td>
<td>0.7914</td>
<td>0.8251</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Vi bemærker, at differensen mellem logit for to kategorier er den samme ved de to forskellige valg af referencekategori.
Multinomiale logit modeller

Antag, at den i'te observerede fordeling har tilknyttet værdierne $x_i^*$ af de forklarende variable, og at man ønsker at beskrive sættet, $(p_{i1}, \ldots, p_{ir})$, af responsandsynligheder som funktion af disse forklarende variable.

Lineære modeller for baseline logit'erne kaldes multinomiale logitmodeller. Modellerne er af formen

$$p_j(x_i^*) = \frac{\exp(x_i^* T \beta_j)}{\sum_{\nu=1}^{r} \exp(x_i^* T \beta_{\nu})}, \quad (4.2.11)$$

hvor parametervektoren $\beta_j$ svarer til baseline kategorien sættes til 0.

Såfremt man har valgt den sidste kategori, $A_r$, som referencekategori, dvs $\beta_r = 0$, bliver modellen jvf. (4.2.8)

$$\theta_j(x_i^*) = \ln \left( \frac{p_j(x_i^*)}{p_r(x_i^*)} \right) = x_i^* T \beta_j \quad (4.2.12)$$

Man modellerer således med en separat parametervektor, $\beta_j$, for hver af kategoriene.

Eksempel 4.2.3 Brug af multinomiale logit modeller til beskrivelse af forbrugeres valg af forbrugsvarer.

De multinomiale logit modeller bruges ofte i forbindelse med modeller til beskrivelse af forbrugeres valg af forbrugsvarer.


McFadden (1982) beskriver en multinomial logit model, hvor de $r$ kategorier angiver $r$ forskellige (disjunkte) valgmuligheder for forbrugeren. Modellen svarer til (4.2.11), idet det dog antages, at parametervektoren $\beta$ er den samme for alle responskategorier. Modellen er på formen:

$$p_j(x_i^*) = \frac{\exp(x_i^* T \beta_j)}{\sum_{\nu=1}^{r} \exp(x_i^* T \beta_{\nu})}, \quad (4.2.13)$$

Idet $R_1, \ldots, R_r$ angiver $r$ uafhængige stokastiske variable, der hver for sig følger en $\text{Max}_1(x_i^* T \beta_j, 1)$-fordeling. Forbrugeren vælger nu kategorien
j, hvis \( R_j \) er større end de øvrige variable \( R_\nu, \nu \neq j \). Sandsynligheden for denne hændelse er netop \( p_j \). De variable \( R_j \) kaldes undertiden for latente variable.

Størrelsen \( v_{ij} = x_i^T \beta_j \) fortolkes som utiliteten eller nytten af valget \( j \) for den \( i \) te person. Modellen anvendes også inden for psykometrien, hvor den betegnes Luce’s strict utility model, se Luce (1959, 1977). Modellen er en udvidelse af Bradley-Terry modellen, som vi betragtede i afsnit 3.5.1.

En variant af ovenstående model fås ved at tillade at de forskellige individer kun har muligheder for at vælge blandt en begrænset mængde af alternativerne \( \{A_1, \ldots, A_r\} \).

Lad \( C_i \) angive mængden af tilgængelige valgmuligheder for det \( i \) te individ.

I analogi med (4.2.13) modellerer man da sandsynligheden for at den \( i \) te person vælger \( A_j \in C_i \) ved

\[
p_j(x_{ij}^*) = \frac{\exp(x_{ij}^* T \beta)}{\sum_{\nu \in C_i} \exp(x_{i\nu}^* T \beta)} ,
\]

(4.2.14)

For ethvert par af muligheder \( A_j \) og \( A_\nu \) finder man

\[
\ln \left( \frac{p_j(x_{ij}^*)}{p_\nu(x_{i\nu}^*)} \right) = (x_{ij}^* - x_{i\nu}^*)^T \beta
\]

(4.2.15)

Betydningen af de enkelte forklarende variable for valget mellem \( A_j \) og \( A_\nu \) afhænger af afstanden mellem personens værdi af denne variabel for disse to alternativer. Specielt ser man, at odds for at vælge \( A_j \) frem for \( A_\nu \) ikke afhænger af de øvrige valgmuligheder i mængden af alternativer, \( C_i \). Denne egenskab kaldes undertiden for uafhængighed af irrelevante alternativer.


Vi afbryder til slut en anden modelklasse, der undertiden benyttes til beskrivelse af diskrete valg, nemlig den såkaldte conjoint analysis (se Green og Srinivasan (1990)). Conjoint analyse kan opfattes som en flerdimensionalskaleringsteknik, eng. Multidimensional scaling, der har til formål at beskrive individets præferencer for de enkelte valgmuligheder, \( \{A_1, \ldots, A_r\} \) som en linearkombination af utiliteterne for de egenskaber, som er tilknyttet de enkelte valgmuligheder. Ideen er her, at den \( j \) te valgmulighed har
4.2 Multinomialfordelingen

tilknyttet en række (sædvanligvis binære) egenskaber, beskrevet ved vektoren \( x_j^* \). Den samlede ytte, som individet tillægger den \( j \)-te valgmulighed tænkes fremkommet som en linearkombination \( x_j^* \beta \) af utiliteterne (part-worth nytten) \( \beta_p \) for hver af egenskaberne \( p \). Individet vælger den mulighed, der giver den største samlede ytte.

Conjoint analyse bruges ofte i forbindelse med forsøgsplanlægning til at bestemme den kombination af egenskaber, der giver den største ytte. Ofte bruges resultatet af en conjoint analyse (utiliteterne) som input til en simuleringssmodel for at analysere effekten af at indføre nye produkter med andre kombinationer af de betragtede egenskaber.

4.2.4 Nabokategori odds

I stedet for at parametrisere ved baseline odds kan man parametrisere ved nabokategori odds

Forholdet,

\[
\theta_j^N = \frac{p_j}{p_{j+1}} = \frac{p_j}{p_{j+1}}, \quad j = 1, \ldots, r - 1
\]  

(4.2.16)

mellem sandsynlighederne for at den \( j \)-te kategori indtræffer, og sandsynligheden for at dens nabokategori, \( A_{j+1} \) indtræffer, kaldes nabokategori odds.

Det tilsvarende logaritmiske mål:

\[
\varphi_j^N = \ln(\theta_j^N), \quad j = 1, 2, \ldots, r - 1
\]  

(4.2.17)

kalder nabologit’en.

Relationen

\[
\theta_j = \frac{p_j}{p_r} = \frac{p_j}{p_{j+1}} \times \frac{p_{j+1}}{p_{j+2}} \cdots \frac{p_{r-1}}{p_r} = \theta_j^N \theta_{j+1}^N \cdots \theta_r^N
\]

viser, at baseline odds (4.2.7) kan udtrykkes ved nabokategori-odds, og tilsvarende kan baseline logit’en (4.2.8) udtrykkes ved nabologit’erne som
\[ \vartheta_j = \vartheta_j^N + \vartheta_{j+1}^N + \cdots + \vartheta_{r-1}^N \] (4.2.18)

Eksempel 4.2.4 Tilfredshedsangivelser fra buspassagerer, nabokategori odds

Vi betragter atter situationen fra eksempel 4.2.1.

Nedenstående tabel viser de beregnede (observerede) nabokategori odds og logit’er for de forskellige svarkategorier.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Antal</td>
<td>234</td>
<td>559</td>
<td>1 157</td>
<td>5 826</td>
<td>2 553</td>
<td>10 329</td>
</tr>
<tr>
<td>Odds</td>
<td>0.4186</td>
<td>0.4831</td>
<td>0.1986</td>
<td>2.2820</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Logit</td>
<td>-0.8708</td>
<td>-0.7274</td>
<td>-1.6165</td>
<td>0.8251</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Der er ikke nogen nabokategori odds (eller logit) svarende til den sidste kategori, Meget tilfreds, men det betyder ikke noget, da der jo kun skal fire odds til at bestemme en multinomialfordeling med fem kategorier. □

Bemærkning 1 En lineær model for nabologit’erne er ækvivalent med en lineær model for baseline-logit’erne

Antag nemlig, at der gælder den lineære (affine) model

\[ \vartheta_j^N = \alpha_j + x^{*T} \beta \quad j = 1, \ldots, r - 1, \]

hvor \( x^* \) angiver en (sojle)-vektor af kendte koefficienter, fælles for samtlige kategorier. Det følger da af (4.2.18), at

\[ \vartheta_j = \sum_{\nu=j}^{r-1} \vartheta_{\nu}^N = \sum_{\nu=j}^{r-1} \alpha_{\nu} + (r - j)x^{*T} \beta \]

der er på formen

\[ \vartheta_j = \alpha_j^* + u_j^* x^* \beta \]

med \( u_j^* = (r - j)x^* \). □
4.2.5 Fortsættelses-odds

En anden parametrizering fås ved at indføre de såkaldte fortsættelses-odds

\[ \theta_j^F = \frac{\Pr[A_j]}{\Pr[A_{j+1}] + \cdots + \Pr[A_r]} = \frac{p_j}{p_{j+1} + \cdots + p_r}, \quad j = 1, \ldots, r - 1 \]  

(4.2.19)

eller, svarende til den omvendte orientering,

\[ \theta_j^{*F} = \frac{\Pr[A_{j+1}]}{\Pr[A_1] + \cdots + \Pr[A_j]} = \frac{p_{j+1}}{p_1 + \cdots + p_j}, \quad j = 1, \ldots, r - 1 \]  

(4.2.20)

Fortsættelses-odds kan fortolkes som odds svarende til betingede sandsynligheder:

Lad nemlig \( \pi_j \) angive den betingede sandsynlighed for at responset er \( A_j \), givet responset er \( A_j \) eller højere (dvs. \( A_j, \ldots, A_r \)). Da bestemmes \( \pi_j \) ved

\[ \pi_j = \frac{p_j}{p_j + \cdots + p_r}, \quad j = 1, \ldots, r - 1, \]  

(4.2.21)

og der gælder

\[ 1 - \pi_j = \frac{p_{j+1} + \cdots + p_r}{p_j + \cdots + p_r} \]

hvorfor vi har

\[ \theta_j^F = \frac{\pi_j}{1 - \pi_j}, \]  

(4.2.22)

alså netop de sædvanlige odds (4.2.5) svarende til disse betingede sandsynligheder.

Tilsvarende kan \( \theta_j^{*F} \) fortolkes som odds svarende til de betingede sandsynligheder, \( \pi_j^* \), for at responset er \( A_{j+1} \) givet at responset er \( A_{j+1} \) eller lavere.
De logaritmenerede værdier af fortsættelses odds, 
\[ \psi_j^F = \ln(\theta_j^F) = \ln\left(\frac{\pi_j}{1 - \pi_j}\right), \quad j = 1, \ldots, r - 1, \]  
(4.2.23)
kaldes fortsættelses logit'er.

Eksempel 4.2.5 Tilfredshedsangivelser fra buspassagerer, fortsættelses odds

Vi betragter atter situationen fra eksempel 4.2.1.
Nedenstående tabel viser de beregnede (observerede) fortsættelses odds og logit'er for de forskellige svarkategorier.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Antal</td>
<td>234</td>
<td>559</td>
<td>1 157</td>
<td>5 826</td>
<td>2 553</td>
<td>10 329</td>
</tr>
<tr>
<td>Antal forts.</td>
<td>10 095</td>
<td>9 536</td>
<td>8 379</td>
<td>2 553</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odds</td>
<td>0.0232</td>
<td>0.0586</td>
<td>0.1381</td>
<td>2.2820</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logit</td>
<td>-3.7645</td>
<td>-2.8367</td>
<td>-1.9799</td>
<td>0.8251</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bemærkning 1 Likelihoodfunktionen svarende til en parametrisering ved fortsættelses odds faktoriserer i et produkt af binomialfordelings likelihoodfunktioner

Lad nemlig \( n_1, n_2, \ldots, n_r \) angive de observerede antal i de \( r \) kategorier med \( n = \sum n_j \). Man kan da udtrykke multinomialfordelingssandsynligheden for dette observationssæt som 
\[ b(n, n_1, \pi_1) \times b(n - n_1, n_2, \pi_2) \times \cdots \times b(n - n_1 - \cdots - n_{r-2}, n_{r-1}, \pi_{r-1}) \]  
(4.2.24)

Såfremt man har en model, hvor parametrene i modelspecificationen for den \( i \)’te fortsættelseslogit er forskellige fra parametrene i modelspecificationen for den \( j \)’te fortsættelseslogit \( (i \neq j) \), kan man altså bestemme maksimum-likelihood estimatorne ved at maksimere likelihoodfunktionen svarende til
4.2 Multinomialfordelingen

hver fortsættelseslogit for sig. Ydermere fås kvotientteststørrelsen for modeltilpasning for hele modellen ved at summere kvotientteststørrelsen for modeltilpasning for de enkelte fortsættelseslogit’er.

I eksempel 4.2.7 vil vi nærmere illustrere dette forhold. □

4.2.6 Kumulative logit’er

Vi betragter endelig parametrizeringen, der fremkommer ved at aggregere kategorierne. Vi indfører de kumulerede sandsynligheder, \( \Pi_j \) ved

\[
\Pi_j = p_1 + p_2 + \cdots + p_j, \quad j = 1, 2, \ldots, r - 1
\]  

(4.2.25)

De kumulative odds indføres da som

\[
\theta^K_j = \frac{\Pi_j}{1 - \Pi_j}, \quad j = 1, 2, \ldots, r - 1
\]  

(4.2.26)

og de tilsvarende kumulative logit’er er bestemt ved

\[
\psi^K_j = \ln(\theta^K_j) = \ln \left( \frac{\Pi_j}{1 - \Pi_j} \right), \quad j = 1, 2, \ldots, r - 1
\]  

(4.2.27)

Vi bemærker, at alle \( r \) kategorier indgår i definitionen af de enkelte kumulative logit’er.

Eksempel 4.2.6 Tilfredshedsangivelser fra buspassagerer, kumulative odds

Vi betragter atter situationen fra eksempel 4.2.1.

Nedenstående tabel viser de beregnede (observerede) kumulative odds og logit’er for de forskellige svarkategorier.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Antal</td>
<td>234</td>
<td>559</td>
<td>1 157</td>
<td>5 826</td>
<td>2 553</td>
<td>10 329</td>
</tr>
<tr>
<td>Kumu leren antal</td>
<td>234</td>
<td>793</td>
<td>1 950</td>
<td>7 776</td>
<td>10 329</td>
<td></td>
</tr>
<tr>
<td>Odds</td>
<td>0.0232</td>
<td>0.0832</td>
<td>0.2327</td>
<td>3.0458</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Logit</td>
<td>-3.7645</td>
<td>-2.4870</td>
<td>-1.4579</td>
<td>1.1138</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Så fremt der optræder forklarende variable, \( x^* \), kan man betræge en lineær (affin) model for de kumulative logit'er,

\[
v_j^K(x) = \alpha_j + x^*T\beta,
\]

hvor \( x^* \) angiver en (søjle)-vektor af kendte koefficienter, fælles for samtlige kategorier.

Sættet af parametre, \( (\alpha_1, \ldots, \alpha_{r-1}) \) kaldes afskæringspunkter (eng. cut-points). Det gælder, at \( \alpha_i \) er ikke-aftagende som funktion af \( i \).

For modellen bestemt ved (4.2.28) gælder, at odds-ratioen svarende til to forskellige værdier, \( x_1^* \) og \( x_2^* \) af de forklarende variable er den samme for alle kategorier, \( j \). Der gælder nemlig

\[
\vartheta_j^K(x_1^*) - \vartheta_j^K(x_1^*) = \ln \left( \frac{\Pi_j(x_1^*)/(1 - \Pi_j(x_1^*))}{\Pi_j(x_2^*)/(1 - \Pi_j(x_2^*))} \right) = (x_1^* - x_2^*)T\beta
\]

Odds ratioen i udtrykket (4.2.29) kaldes en kumulativ odds ratio.

Logaritmen til den kumulative odds ratio er proportional med forskellen mellem værdierne af de forklarende variable. Modellen (4.2.28) kaldes derfor en proportional odds model.

**Bemærkning 1 Fortolkning af modeller for kumulative logit'er**

Betragt den simple model for de kumulative logit'er

\[
v_j^K = \alpha_j , j = 1, \ldots, r - 1
\]

Modellen svarer til, at de kumulative sandsynligheder, \( \Pi_j \) er parametriseret som

\[
\Pi_j = \frac{\exp(\alpha_j)}{1 + \exp(\alpha_j)} , j = 1, \ldots, r - 1
\]

Lad \( R \in L(0, 1) \) være en logistisk fordelt stokastisk variabel. Der gælder da

Definer nu den variable \( J \) ved
\[ J = j \quad \text{hvis} \quad \alpha_{j-1} < R \leq \alpha_j, \quad (4.2.32) \]

hvor \(-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_r = \infty\). Da gælder netop, at \(P[R \leq \alpha_j] = \Pi_j, j = 1, \ldots, r - 1\), hvorfor vi har, at \(P[J = j] = P[A_j] = p_j\) for \(j = 1, \ldots, r - 1\).

Den variable \(J\) angiver således indeks for den hændelse, der indtræffer i et enkelt forsøg, og man kan opfatte \(R\) som en finere angivelse af responsset. \(R\) betegnes undertiden den latente variable.

Hvis responsfordelingen i stedet havde været givet ved (4.2.28) ville man kunne anstille analoge betragtninger for \(R \in L (-x^T \beta, 1)\), dvs en positionsforskydning \(x^T \beta\) enheder.

Specielt, hvis der kun er en enkelt, kontinuer, forklarende variabel, \(x\), kan vi betrakte responskurven svarende til det \(j\)te kumulative respons,

\[ F_j(x) \overset{\text{DEF}}{=} P[A_1 \cup A_2 \cdots \cup A_j] = P[R \leq \alpha_j] \quad (4.2.33) \]

Responskurven for de kumulative responser svarer til de logistiske regressionsmodeller, der blev diskuteret i afsnit 3.2. For modellen (4.2.28) med proportionale odds med en enkelt, kontinuer, forklarende variabel gælder altså

\[ F_\nu(x) = F_j(x + (\alpha_\nu - \alpha_j)/\beta) \quad (4.2.34) \]

De enkelte responskurver fremkommer altså af hinanden ved translation \((\alpha_\nu - \alpha_j)/\beta\) enheder i \(x\)-aksens retning.

I eksempel 4.2.8 på side 460 vil vi give et eksempel på en sådan situation med proportionale odds. \(\Box\)

### 4.2.7 Andre linkfunktioner

Valget af logitfunktionen som linkfunktion vil sædvanligvis indebære at modeltilpasningen er relativt simpel - i det mindste i forhold til andre, mere komplicerede linkfunktioner.

I forbindelse med kontinuerede forklarende variable kan den aktuelle situation dog undertiden begrunde, at man vil overveje andre linkfunktioner. For
de parametriseringer, hvor odds er bestemt som forholdet mellem sandsynligheden for en hændelse og sandsynligheden for den komplementære hændelse, kan man eksempelvis overveje at benytte nogle af de linkfunktioner, der blev betragtet i binomialfordelingssituationen, afsnit 3.2.2.

Nabologit'er, fortsættelseslogit'er og kumulative logit'er er af speciel interesse i forbindelse med ordnede responskategorier, hvor naboegenskaber kan tillægges en fortolkning.

I sådanne situationer med ordnede responskategorier og kontinuerede forklarende variable kan man vælge at modellere de betingede sandsynligheder, \( \pi_j \) (4.2.21), eller de kumulerede sandsynligheder \( \Pi_j \) (4.2.25) ved en log-log, eller en komplementær log-log link, eller eventuelt ved en probit-link.

### 4.2.8 Regressionsmodeller

**Eksempel 4.2.7 Tilfredshedsangivelser fra buspassagerer i afhængighed af bussens forsinkelser**

Vi betragter atter situationen fra eksempel 4.2.1. Som nævnt i eksemplet registrerede man såvel kundernes tilfredshed med overholdelsen af køreplanen som bussens eventuelle forsinkelser.

Nedenstående tabel viser fordelingen af svarene på 12 233 spørgeskemaer indsamlet blandt kvindelige passagerer. Svarene er opdelt efter bussens forsinkelser i minutter.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>234</td>
<td>559</td>
<td>1 157</td>
<td>5 826</td>
<td>2 553</td>
<td>10 329</td>
</tr>
<tr>
<td></td>
<td>2.3 %</td>
<td>5.4 %</td>
<td>11.2 %</td>
<td>56.4 %</td>
<td>24.7 %</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>41</td>
<td>100</td>
<td>145</td>
<td>602</td>
<td>237</td>
<td>1 125</td>
</tr>
<tr>
<td></td>
<td>3.6 %</td>
<td>8.9 %</td>
<td>12.9 %</td>
<td>53.5 %</td>
<td>21.1 %</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>42</td>
<td>76</td>
<td>89</td>
<td>254</td>
<td>72</td>
<td>533</td>
</tr>
<tr>
<td></td>
<td>7.9 %</td>
<td>14.3 %</td>
<td>16.7 %</td>
<td>47.7 %</td>
<td>13.5 %</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>48</td>
<td>39</td>
<td>95</td>
<td>27</td>
<td>244</td>
</tr>
<tr>
<td></td>
<td>14.3 %</td>
<td>19.7 %</td>
<td>16.0 %</td>
<td>38.9 %</td>
<td>11.1 %</td>
<td></td>
</tr>
</tbody>
</table>

Nedenstående figur illustrerer disse fordelinger grafisk.
Svarkategoriene er ordnede, gælende fra "Meget utilfreds" til "Meget tilfreds". Ordningen er afspjølet i figuren på samme måde som i eksempel 4.2.1 ved at andelen af besvarelser af kategorien "Meget utilfreds" er markeret umiddelbart til venstre for klokken 12, derefter fortsættes mod uret indtil kategorien "Meget tilfreds", der er afbildet umiddelbart til højre for klokken 12. Det ses, at andelen af utilfredse kunder stiger, jo større forsinkelsen er, og tilsvarende falder andelen af kunder, der erklærer sig tilfredse eller meget tilfredse.

I dette eksempel vil vi illustrere modelleringen af disse besvarelser hhv. ved brug af fortsættelseslogit' er og ved kumulerede logit' er.

Modellering ved fortsættelseslogit' er

De beregnede fortsættelsesandele er angivet i nedenstående tabel:
Modellen er

\[ \ln \left( \frac{\pi_j(x_i)}{1 - \pi_j(x_i)} \right) = \alpha_j + \beta_j x_i, \quad (4.2.35) \]

hvor \( x_i \) angiver forsinkelsen i minutter.

Estimationen kan udføres for hver svarkategori for sig, fx. ved SAS proceduren GENMOD.

Nedenstående tabel angiver for hver svarkategori estimaterne \( \alpha_j \) og \( \beta_j \) samt de tilsvarende residualdevianser, \( D(y, \hat{\pi}) \). Deviansbidragene er bestemt som deviansbidrag svarende til binomialfordelingen jvf udtrykket (4.2.24). Residualdeviansen, \( D(y, \hat{\pi}) \) for hver af kategorierne skal sammenlignes med fraktilerne i en \( \chi^2(2) \)-fordeling. Endvidere er residualdevianserne svarende til de forskellige kategorier uafhængige af hinanden, hvorfor vi kan tillade os at lægge dem sammen til en total med 10 frihedsgrader.

<table>
<thead>
<tr>
<th>Svar-kategori</th>
<th>( \alpha_j )</th>
<th>( \beta_j )</th>
<th>( D(y, \hat{\pi}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Meget Utilfreds&quot;</td>
<td>-3.7733</td>
<td>0.2736</td>
<td>0.3744</td>
</tr>
<tr>
<td>&quot;Utilfreds&quot;</td>
<td>-2.8287</td>
<td>0.2328</td>
<td>0.3152</td>
</tr>
<tr>
<td>&quot;Både og&quot;</td>
<td>-1.9814</td>
<td>0.1269</td>
<td>0.3297</td>
</tr>
<tr>
<td>&quot;Tilfreds&quot;</td>
<td>0.8234</td>
<td>0.0721</td>
<td>0.3271</td>
</tr>
<tr>
<td>&quot;Meget Tilfreds&quot;</td>
<td>-1.1101</td>
<td>-0.1367</td>
<td>1.0090</td>
</tr>
<tr>
<td>Ialt</td>
<td></td>
<td></td>
<td>2.3554</td>
</tr>
</tbody>
</table>

Residualdevianserne giver ikke anledning til afvisning af modellen.

De estimerede relationer mellem fortsættelseslogit’erne og forsinkelsen er illustreret i nedenstående figur.
På figuren er endvidere indtegnet de observerede logit’er og et 95 % konfidensinterval for de observerede logit’er bestemt ved at udføre logit-transformationen på de sædvanlige konfidensintervaller for $\pi_j$. Disse konfidensintervaller kan jvf bemærkning 1 på side 446 fortolkes i relation til den betingede sandsynlighed for fortsættelse.

Til illustration af tilpasningen viser nedenstående figur tilpasningen af den lineære logitmodel for kategorien "Utilfreds". Der ses at være en god tilpasning til den lineære model.
Betrægter man deviansresidualerne

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1335</td>
<td>-0.1829</td>
<td>0.0477</td>
<td>0.0712</td>
</tr>
<tr>
<td>2</td>
<td>-0.3088</td>
<td>0.7232</td>
<td>-0.3103</td>
<td>-0.4598</td>
</tr>
<tr>
<td>5</td>
<td>-0.3355</td>
<td>-0.2636</td>
<td>0.4061</td>
<td>0.5798</td>
</tr>
<tr>
<td>7</td>
<td>0.3856</td>
<td>-0.0669</td>
<td>-0.2573</td>
<td>-0.3186</td>
</tr>
</tbody>
</table>

Ser man, at ingen af residualerne er numerisk større end 1, hvilket støtter indtrykket af den gode tilpasning.

Forløbet af de estimerede sandsynligheder for svar i de respektive kategorier er illustreret i nedenstående figur
4.2 Multinomialfordelingen

Modellering ved kumulative logit’er

Nedenstående tabel viser for hver svarkategori estimererne $\alpha_j$ og $\beta_j$ samt de tilsvarende residualdevianser, $D(y, \hat{\Pi})$ beregnet svarende til et binomialt respons for hver kategori for sig. Vi bemærker, at residualdevianserne svarende til de forskellige kategorier ikke er uafhængige. Vi tillader os alligevel som en grov modelkontrol at beregne en total ved addition.

<table>
<thead>
<tr>
<th>Svarkategori</th>
<th>$\alpha_j$</th>
<th>$\beta_j$</th>
<th>$D(y, \hat{\Pi})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meget Utilfreds</td>
<td>-3.7733</td>
<td>0.2736</td>
<td>0.3744</td>
</tr>
<tr>
<td>Utilfreds</td>
<td>-2.4849</td>
<td>0.2550</td>
<td>0.4157</td>
</tr>
<tr>
<td>Både og</td>
<td>-1.4603</td>
<td>0.2033</td>
<td>0.2157</td>
</tr>
<tr>
<td>Tilfreds</td>
<td>1.1101</td>
<td>0.1367</td>
<td>1.0090</td>
</tr>
<tr>
<td>Ialt</td>
<td></td>
<td></td>
<td>1.7991</td>
</tr>
</tbody>
</table>

Residualdevianserne giver ikke anledning til afvisning af modellen.
De estimerede relationer mellem de kumulative logit’er og forsinkelsen er illustreret i nedenstående figur.

De tilsvarende deviansresidualer er angivet i nedenstående tabel:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1335</td>
<td>-0.0565</td>
<td>0.0956</td>
<td>0.1610</td>
</tr>
<tr>
<td>2</td>
<td>-0.3088</td>
<td>0.3551</td>
<td>-0.3280</td>
<td>-0.8505</td>
</tr>
<tr>
<td>5</td>
<td>-0.3355</td>
<td>-0.4577</td>
<td>-0.1140</td>
<td>0.5020</td>
</tr>
<tr>
<td>7</td>
<td>0.3856</td>
<td>-0.2773</td>
<td>0.2933</td>
<td>-0.0837</td>
</tr>
</tbody>
</table>

Også her ses at være en god tilpasning.

Forløbet af de estimerede sandsynligheder for svar i de respektive kategorier er illustreret i nedenstående figur.
Forløbet adskiller sig ikke synligt fra forløbet af de estimerede sandsynligheder under den lineære model for fortsættelseslogit'erne. For praktiske formål er der således ingen forskel på de to modeller i det betragtede variationsområde for forsinkelserne.

For at illustrere den kvalitative forskel på de to modeller viser nedenstående figurer det extrapolerede forløb af sandsynlighederne for de forskellige svarkategorier ved forsinkelser op til 20 minutter.
Modellering ved model med proportionale odds

Vi vil her betragte modellen (4.2.28) svarende til at effekten af forsinkelsen på den kumulative logit er den samme for alle responskategorier, dvs modellen med proportionale odds.
Nedenstående tabel viser for hver svarkategori estimatorne, $\hat{\alpha}_j$ for cut-points, estimatet, $\hat{\beta}$ for den fælles hældning, samt residualdeviansen, $D(y, \hat{\Pi})$.

<table>
<thead>
<tr>
<th>Svar-kategori</th>
<th>$\alpha_j$</th>
<th>$\beta$</th>
<th>$D(y, \hat{\Pi})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meget Utilfreds</td>
<td>-3.6859</td>
<td>0.2129</td>
<td>6.9736</td>
</tr>
<tr>
<td>Utilfreds</td>
<td>-2.4390</td>
<td>0.2129</td>
<td>7.8757</td>
</tr>
<tr>
<td>Både og</td>
<td>-1.4675</td>
<td>0.2129</td>
<td>0.7657</td>
</tr>
<tr>
<td>Tilfreds</td>
<td>1.0804</td>
<td>0.2129</td>
<td>18.1811</td>
</tr>
<tr>
<td>Ialt</td>
<td></td>
<td></td>
<td>33.7961</td>
</tr>
</tbody>
</table>

Tilpasningen er ikke så god som for modellen med individuelle hældninger (naturligvis). Selv om der er flere frihedsgrader for hver af residualdevianserne, (totalen har 18 frihedsgrader) er tilpasningen næppe tilfredsstillende, modellens gode egenskaber til trods.

De estimerede relationer mellem de kumulative logit'er og forsinkelsen er illustreeret i nedenstående figur.

De tilsvarende deviansresidualer er angivet i nedenstående tabel
Endelig viser vi i nedenstående figur det extrapolerede forløb af sandsynlighederne for de forskellige svarkategorier ved forsinkelser op til $20$ minutter under denne model med proportionale odds.

![Diagram](image)

**Eksempel 4.2.8 Fortolkning af modellen med proportionale odds ved latente variable**

Vi illustrerer endelig fortolkningen af modellen med proportionale odds ved såkaldte latente variable.

Nedenstående figur viser tæthedsfunktionen for fordelingen af den latente variable (lodret akse) svarende til forskellige forsinkelser (forsinkelserne angivet ud af den vandrette akse). De vandrette linier på figuren angiver cutpoints svarende til de forskellige tilfredshedsgrader.
4.3 Modeller med flere klassifikationskriterier

4.3.1 Flere klassifikationskriterier, Yule’s krydsprodukt-ratio

Ved en flerdimensional klassifikation har man mulighed for en række forskellige parametrimeringer. Det afgørende i parametrimeringen er de definerende kontraster.

Nedenfor skal vi - lidt kompakt - gøre rede for forskellige sæt af definerende kontraster.

For en klassifikation \( \Omega = \bigcup_{i=1}^{r} A_i \) i disjunkte hændelser \( A_i \) og en anden klassifikation \( \Omega = \bigcup_{j=1}^{s} B_j \) i disjunkte hændelser med de tilknyttede sandsynligheder

\[
p_{i,j} = P \left[ A_i \cap B_j \right], \quad i = 1, \ldots, r, \quad j = 1, \ldots, s
\]

kan vi definere odds-ratioen svarende til cellerne \( i_1, i_2 \) og \( j_1, j_2 \) som forholdet mellem odds-ratioen for odds for \( B_{j_1} \) mod \( B_{j_2} \) svarende til \( A_{i_1} \) og
odds-ratioen for odds for $B_{j_1}$ mod $B_{j_2}$ svarende til $A_{i_2}$, dvs

$$\omega_{i_1,i_2;j_1,j_2}^{A:B} = \frac{\omega_{i_1,j_1}^{A:B}}{\omega_{i_2,j_1}^{A:B}} = \frac{p_{i_1,j_1}}{p_{i_2,j_1}} \frac{p_{i_2,j_2}}{p_{i_1,j_2}} = \frac{P[A_{i_1,j_1}]}{P[A_{i_1,j_2}]} / \frac{P[A_{i_2,j_1}]}{P[A_{i_2,j_2}]}$$

(4.3.1)

hvor vi har indført symbolet

$$\omega_{i_1,i_2;j_1,j_2}^{A:B} = \frac{p_{i_1,j_1}}{p_{i_1,j_2}}$$

til at angive oddsratioen mellem cellerne $(A, B) = (i_1, j_1)$ og $(A, B) = (i_1, j_2)$

Forholdet (4.3.1) er en generalisering af definitionen på Yule’s krydsprodukt ratio i afsnit 3.3.3. Forholdet (4.3.1) kaldes derfor også Yules-krydsprodukt ratio.

Den logaritmerede form af Yules krydsprodukt-ratio er

$$\Delta_{i_1,i_2;j_1,j_2}^{A:B} = \ln(\omega_{i_1,i_2;j_1,j_2}^{A:B}) = \frac{p_{i_1,j_1}}{p_{i_1,j_2}} / \frac{p_{i_2,j_1}}{p_{i_2,j_2}}$$

$$= \frac{P[A_{i_1,j_1}]}{P[A_{i_1,j_2}]} / \frac{P[A_{i_2,j_1}]}{P[A_{i_2,j_2}]}$$

(4.3.2)

Vi har

$$\Delta_{i_1,i_2;j_1,j_2}^{A:B} = \Delta_{i_2,j_1,j_2}^{B} - \Delta_{i_2,j_1,j_2}^{A} = \Delta_{i_1,i_2;j_1,j_2}^{A} - \Delta_{i_1,i_2;j_1,j_2}^{B}$$

som umiddelbart kan generaliseres til flere klassifikationer.

For et tosidet skema $\{A_1, \ldots, A_r\} \times \{B_1, \ldots, B_s\}$ med sættet af sandsynligheder $\{p_{i,j}\}$ kan der dannes $\binom{r}{2} \binom{r}{2}$ kontraster af denne type.

Denne mængde af odds-ratioer indeholder imidlertid megen redundant information.

Sættet $\{p_{i,j}\}$ er bestemt, såfremt der er givet en delmængde af $(r-1)(s-1)$ lokale kontraster.

Eksempelvis er sættet $\{p_{i,j}\}$ bestemt af den minimale mængde

$$\omega_{i,j}^* = \omega_{i,i+1;j,j+1} = \frac{p_{i,j}}{p_{i,j+1}} / \frac{p_{i+1,j}}{p_{i+1,j+1}}, \ i = 1, \ldots, r - 1, \ j = 1, \ldots, s - 1$$

(4.3.3)
4.3 Modeller med flere klassifikationskriterier

En anden minimal mængde fås som

\[ \omega_{i:j} = \omega_{i:r:j:s} = \frac{p_{i:j}}{p_{i:s}} \cdot \frac{p_{r:j}}{p_{r:s}} \quad i = 1, \ldots, r-1, \quad j = 1, \ldots, s-1 \]  

(4.3.4)

Såfremt alle elementerne \( \{p_{i:j}\} \) er strengt positive, er der enentydig sammenhæng mellem kontrasterne (4.3.3) eller (4.3.4).

Såfremt sættene af marginale sandsynligheder

\[ P[A_i] = \sum_{j=1}^{s} p_{i:j}, \quad i = 1, \ldots, r \]

og

\[ P[B_j] = \sum_{i=1}^{r} p_{i:j}, \quad j = 1, \ldots, s \]

er givet, da vil sættet \( \{p_{i:j}\} \) være fastlagt ved et sæt af definerende kontraster (4.3.3) eller (4.3.4).

4.3.2 Tovejs antalstabeller, multinomial stikprøveudvælgelse

Vi vil indledningsvis betragte tovejs-antalstabeller, dvs inddelinger af et antal individer efter to kriterier.

Omdend den anvendte stikprøvemodel ikke altid fremgår af en given antalstabel \( \{x_{i:j}\} \), vil det være relevant at vurdere, hvorledes stikprøven er fremkommet ved formulering af den parametriske model for analysen.

I det følgende vil vi kort omtale de sædvanlige modeller svarende til todimensionale tabeller.

Under multinomial stikprøveudvælgelse tænkes totalsummen \( N = \sum_{i,j} x_{i:j} \) af observationerne at være fastlagt af stikprøveplanen, og de \( N \) observationer er derefter krydsklassificeret i henhold til faktorerne beskrevet ved \( I \) og \( J \).

Frekvensfunktionen svarende til denne model er multinomialfordelingen, \( \text{Mult}(N, p_{1,1}, \ldots, p_{1,J}, \ldots, p_{I,1}, \ldots, p_{I,J}) \)

\[ f(x) = \frac{N!}{\prod_{i,j} x_{i:j}!} \prod_{i,j} p_{i,j}^{x_{i:j}} \]  

(4.3.5)
hvor $\sum_{i,j} p_{i,j} = 1$, og log-likelihoodfunktionen er
\[ l(p; x) = \frac{N!}{\Pi_{i,j} x_{i,j}!} \sum_{i,j} x_{i,j} \ln(p_{i,j}) \] (4.3.6)

hvor $\sum_{i,j} p_{i,j} = 1$.

De forventede værdier af celleværdierne er $E [X_{i,j}] = N p_{i,j}$.

En naturlig reduktion af modellen er, at antage, at der er uafhængighed mellem de to klassifikationer, dvs. at
\[ H_0 : p_{i,j} = p_{i,+} p_{+,j} \quad \text{for alle} \quad (i, j) \] (4.3.7)

Der gælder

**Sætning 4.3.1 Uafhængighedshypoteze for todimensional tabel**

Under den multinomiale stikprøvemodel gælder, at $0 < p_{i,j} = p_{i,+} p_{+,j}$ for alle $(i, j)$ hvis og kun hvis
\[ \frac{p_{i_1 j_1} p_{i_2 j_2}}{p_{i_1 j_2} p_{i_2 j_1}} = \omega_{i_1, i_2; j_1, j_2}^A = 1 \] (4.3.8)

for alle $i_1, i_2 \in (1, 2, \ldots, J)^2$ og alle $j_1, j_2 \in (1, 2, \ldots, J)^2$

**Bevis:**
Det følger umiddelbart, at uafhængighed indebærer at forholdet mellem odds-ratioerne er 1.

Den modsatte implikation bevises ved udnyttelse af relationen $\sum_{i,j} p_{i,j} = 1$.

**Bemærkning 1 Det er nok at kræve, at forholdet mellem odds-ratioer for de tilsvarende definerende kontraster er 1**

Det er ikke nødvendigt at forlange, at alle forhold mellem odds-ratioer er 1. For eksempel gælder, at såfremt
\[ \omega_{1,2; 2,3}^A = \frac{p_{1,2} p_{2,3}}{p_{1,3} p_{2,2}} = 1 \]
og
\[ \omega_{1,2; 2,4}^A = \frac{p_{1,2} p_{2,4}}{p_{1,4} p_{2,2}} = 1, \]
da vil også
\[
\omega_{1,2;3,4}^A = \frac{P_{1,4}P_{2,3}}{P_{1,3}P_{2,4}} = \frac{P_{1,2}P_{2,3}}{P_{1,3}P_{2,2}} \left/ \frac{P_{1,2}P_{2,4}}{P_{1,4}P_{2,2}} \right. = 1
\]
Generelt gælder, at
\[
\omega_{i_1,i_2; j_1,j_2}^{A,B} = 1
\]
for alle \(i_1, i_2\) og \(j_1, j_2\), hvis og kun hvis
\[
\omega_{1,i;1,j}^{A,B} = \frac{P_{11}P_{i,j}}{P_{1j}P_{i1}} = 1
\]
for alle \(i\) og \(j\).

Dette resultat hænger sammen med fremstillingen ved definerende kontraster i afsnit 4.3.1.

Udtrykt ved de kanoniske parametre for multinomialfordelingen har vi,

Sætning 4.3.2 Uafhængshedshypotese for todimensional tabel ved kanoniske parametre

Under den multinomiale stikprøvemodel gælder, at \(0 < p_{i,j} = p_{i}.p_{.,j}\) for alle \((i, j)\) hvis og kun hvis log odds kan udtrykkes på formen
\[
\vartheta_{i,j} = \alpha_i + \beta_j
\]

Bevis:
Følger umiddelbart

Bemærkning 1 Multinomialmodellen svarer til den betingede fordeling af celleværdier i Poissonmodellen

Vi bemærker, at såfremt \(\{x_{i,j}\}\) er fremkommet under en Poisson-model (4.1.1), da vil den betingede fordeling af celleværdierne \(\{X_{i,j}\}\) givet totalværdien \(X_{+,+} = N\) være en multinomial-fordeling med
\[
p_{i,j} = \frac{\lambda_{i,j}}{\lambda_{+,+}}
\]
svarende til
\[
\vartheta_{i,j} = \ln(\lambda_{i,j}) - \ln(\lambda_{+,+})
\]
4.4 Log-lineære modeller

Ganske som vi kunne forestille os flere forskellige stikprøvemodeller for en todimensional antalstabel, kan man for en flerdimensional antalstabel forestille sig en række forskellige stikprøvemodeller.

Uden hensyn til stikprøvemodellen kan de dog alle bringes på en såkaldt log-lineær form.

**Definition 4.4.1 Log-lineær model for en antalstabel**

Betragt en antalstabel svarende til de r klassifikationer \( A, B, \ldots, H \) og lad det forventede antal i den \( (i, j, k, \ldots) \)'te celle være \( \lambda_{i,j,k,\ldots} \).

Ved en log-lineær model for antalstabellen vil vi forstå en model af formen

\[
\ln(\lambda_{i,j,k,\ldots}) = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + \ldots + (\alpha \beta)_i^{AB} + \ldots + (\alpha \beta \gamma)_i^{ABC} + \ldots
\]

(4.4.1)

Modelstrukturen svarer til de sædvanlige modeller for faktorforsøg.

For at undgå overparametrizering kan man eksempelvis vælge at sætte \( \alpha_i^A = \beta_j^B = \ldots = 0 \) og \( (\alpha \beta)_i^{AB} = (\alpha \beta)_i^{AB} = \ldots = 0 \) etc. i overensstemmelse med betragtningerne i afsnit 2.9.3
4.5 Betinget uafhængighed

Når man betragter log-lineære modeller, vil man sædvanligvis indskrænke sig at undersøge såkaldte hierarkiske modeller.

**Definition 4.4.2 Hierarkisk log-lineær model**

En log-lineær model kaldes hierarkisk, hvis modellen er sådan, at såfremt et af leddene i udtrykket (4.4.1) er nul, da vil også alle højere ordens led, der indeholder de samme faktorkombinationer, være nul.

Eller, sagt på en anden måde: Hvis modellen eksempelvis indeholder led af formen \((\alpha \beta)_{i,j}^{AB}\), da indeholder den også led, der er marginale i forhold til dette led, dvs. leddene \(\alpha_i^A\) og \(\beta_j^B\).

**Definition 4.4.3 Frembringere for log-lineær model**

Samlingen af led i en hierarkisk model, som ikke er marginale i forhold til nogle af de øvrige led i modellen kaldes modellens frembringere, eller den frembringende klasse for modellen.

**Bemærkning 1 Frembringerne for en log-lineær model svarer til de sufficiente marginaler**

Vi bemærker, at de led, der frembringer en log-lineær model, netop svarer til de sufficiente marginaler, dvs de række-, søjle-, eller lagsummer, der er sufficiente for estimation af parametrene i modellen.

4.5 Betinget uafhængighed

Inden vi går over til at betragte flervejsantalstabeller vil vi indføre begrebet betinget uafhængighed, der skal bruges til at beskrive statistiske sammenhænge i sådanne, lidt mere komplicerede situationer.

**Definition 4.5.1 Betinget uafhængige stokastiske variable**

Lad \(X, Y\) og \(Z\) være stokastiske variable med simultan tæthed \(f_{X,Y,Z}(\cdot,\cdot,\cdot)\) og tilsvarende marginale og betingede tætheder

\(X\) og \(Z\) siges at være betinget uafhængige givet \(Y\), såfremt der gælder

\(f_{X,Z|Y}(x,z;y) = f_X|Y(x;y)f_Z|Y(z;y)\) for alle \(y\) som opfylder \(f_Y(y) > 0\)

Hvis \(X\) og \(Z\) er betinget uafhængige for givet \(Y\) skriver vi \(X \perp Z \mid Y\).
Der gælder

**Sætning 4.5.1 Faktoriseringssætning for uafhængige variable**

De stokastiske variable $X$ og $Y$ er uafhængige hvis og kun hvis der findes to funktioner $g(\cdot)$ og $h(\cdot)$ sådan at

$$f_{X,Y}(x,y) = g(x)h(y) \quad \text{for alle } x \text{ og } y$$

**Bevis:**
Bevises direkte

**Sætning 4.5.2 Simultan uafhængighed medfører marginal uafhængighed**

Såfremt der for en opdelt stokastisk variabel, $(X,Y,Z)$, gælder at $X \perp (Y,Z)$, da vil også $X \perp Y$.

**Bevis:**
Bevises direkte

**Sætning 4.5.3 Faktoriseringssætning for betinget uafhængige variable**

De stokastiske variable $Y$ og $Z$ er betinget uafhængige givet $X$, hvis og kun hvis der findes to funktioner $g(\cdot, \cdot)$ og $h(\cdot, \cdot)$ sådan at

$$f_{X,Y,Z}(x,y,z) = g(x,y)h(x,y) \quad \text{for alle } y \text{ og } z \text{ og alle } x \text{ for hvilke } f_X(x) > 0$$

**Bevis:**
Bevises direkte

### 4.5.1 Uafhængighedsgrafer

Uafhængighedsgrafer er et nyttigt værktøj fra grafteorien, der kan bruges til at illustrere de statistiske sammenhænge i en antalstabel.

En graf er et matematisk objekt, bestående af to mængder, en mængde af knuder, og en mængde af kanter, hvor kanterne er par af knuder. Hvis parrene opfattes som ordnede par, siger man at man har en orienteret graf. Vi vil her betragte ikke-orienterede grafer.

Graferne repræsenteres ofte ved et diagram, hvor en knude repræsenteres ved et punkt, og en kant repræsenteres ved en linie, der forbinder de to punkter, som tilhører kanten.
4.6 Trevejs antalstabeller

Definition 4.5.2 Uafhængighedsgraf for stokastiske variable

Lad $X_1, X_2, \ldots, X_r$ angive $r$ stokastiske variable, og lad $R = \{1, 2, \ldots, r\}$ angive den tilsvarende mængde af knuder.

En graf siges at være en uafhængighedsgraf for $X_1, X_2, \ldots, X_r$, hvis der ikke er en kant mellem to variable, hvis disse to variable er stokastisk uafhængige givet resten af de variable.

Formelt skriver man $(i, j)$ er ikke en kant, hvis og kun hvis

$$X_i \perp X_j \mid X_{R \setminus \{i, j\}}$$

Vi bemærker, at definitionen vedrører betinget uafhængighed. Der er ingen velegnet teori, der bygger på marginal uafhængighed.

På tilsvarende måde defineres en uafhængighedsgraf svarende til en klassifikation efter de $r$ kriterier, $A, B, \ldots, H$.

Definition 4.5.3 Uafhængighedsgraf for antalstabell

Betragt en antalstabell svarende til klassifikationerne $A, B, \ldots, H$. Ved uafhængighedsgrafen svarende til denne antalstabell forstås en graf med knuderne $1, 2, \ldots, r$ repræsenterede de $r$ klassifikationer, og sådan, at hvis to klassifikationer er stokastisk uafhængige givet resten af tabellen, da er det ikke nogen kant imellem knuderne svarende til disse to klassifikationer.

Bemærkning 1 Uafhængighedsgrafen kaldes også en vekselvirkningsgraf

4.6 Trevejs antalstabeller

For tovejs-klassificerede data havde vi - ud over den endimensionale Poissonmodel - essentielt to mulige stikprøvemodeller, nemlig enten en model svarende til en enkelt multinomialfordeling, eller en model svarende til sammenligning af et antal ($I$) uafhængige multinomialfordelinger.
Ved modellen svarende til en enkelt \((I \times J\)-dimensional) multinomialfordeling var der kun én relevant modelreduktion, nemlig til en model, hvor der var uafhængighed mellem rækker og søjler.

Ved tredimensionale tabeller er der væsentligt flere relevante modelreduktioner. Vi vil her kun betragte situationen svarende til en enkelt stikprøve, klassificeret efter tre inddelingskriterier.

Vi vil lade de tre klassifikationer betegne med \(A\), \(B\) og \(C\), med de respektive klasseantal \(I\), \(J\) og \(K\), og med kasserne indiceret ved \(i\), \(j\), og \(k\) henholdsvis.

Vi vil opfatte elementerne i den tredimensionale tabel som elementer i en terning med rækker \((A)\), søjler \((B)\), og lag \((C)\).

### 4.6.1 Multinomial stikprøveudvælgelse:

Under denne stikprøvemodel tænkes totalsummen \(N = \sum_{i,j,k} x_{i,j,k}\) af observationerne at være fastlagt af stikprøveplanen, og de \(N\) observationer er derefter krydsklassificeret i henhold til faktorerne beskrevet ved \(I, J, K\).

Frekvensfunktionen svarende til denne model er multinomialfordelingen (Polynomialfordelingen)

\[
f(x) = \frac{N!}{\prod_{i,j,k} x_{i,j,k}!} \prod_{i,j,k} p_{i,j,k}^{x_{i,j,k}} \tag{4.6.1}
\]

hvor \(\sum_{i,j,k} p_{i,j,k} = 1\)

De forventede værdier af celleværdierne er

\[
E[X_{i,j,k}] = Np_{i,j,k}.
\]

Vi bemærker at såfremt \(\{x_{i,j,k}\}\) er fremkommet under en Poisson-model, da vil den betingede fordeling af celleværdierne \(\{X_{i,j,k}\}\) givet totalværdien \(X_{i,j,k} = N\) være en multinomial-fordeling med \(p_{i,j,k} = \lambda_{i,j,k}/\lambda_{i,j}\).

Log-likelihoodfunktionen svarende til multinomialfordelingsmodellen er

\[
l(p) = \frac{N!}{\prod_{i,j,k} x_{i,j,k}!} \sum_{i,j,k} x_{i,j,k} \ln(p_{i,j,k}) \tag{4.6.2}
\]

hvor \(\sum_{i,j,k} p_{i,j,k} = 1\)
Vi har følgende principielt forskellige hypoteser:

\[ H_0 : p_{ijk} = p_i^A p_j^B p_k^C \]  \hspace{1cm} (4.6.3)

\[ H_{1a} : p_{ijk} = p_i^A p_{jk}^B p_k^C \]  \hspace{1cm} (4.6.4)

\[ H_{2a} : p_{ijk} = \frac{p_{ij}^A p_{ik}^A}{p_i^A} \]  \hspace{1cm} (4.6.5)

Hypotesen \( H_0 \) udtrykker total uafhængighed mellem de tre inddelingskriterier. Hypotesen udtrykkes symbolsk som

\[ H_0 : A \perp B \perp C \]

Uafhængighedsgrafen er:

```
  Ĉ
 /     \
 /       \
A  B̂
```

Hypotesen \( H_{1a} \) udtrykker, at rækkerne er uafhængige af søjler og lag. Hypotesen udtrykkes symbolsk som

\[ H_{1a} : A \perp B, C \]

Uafhængighedsgrafen er:

```
  Ĉ
 /    \ 
 /      \
A  B̂
```

Ved permutation af indices får man de tilsvarende hypoteser:

\[ H_{1b} : p_{ijk} = p_j^B p_{ik}^A C \]

\[ H_{1c} : p_{ijk} = p_k^C p_{ij}^A B \]

der udtrykker at søjlerne er uafhængige af rækker og lag \( H_{1b} : B \perp A, C \), og at lagene er uafhængige af rækker og søjler \( H_{1c} : C \perp A, B \)

Hypotesen \( H_{2a} \) er en hypoteze om betinget uafhængighed. Hypotesen udtrykker, at søjler og lag er uafhængige, givet rækkepositionen. Hypotesen udtrykkes symbolsk som

\[ H_{2a} : B \perp C|A \]

Uafhængighedsgrafen er:
De analoge hypoteser er

\[ H_{2b} : p_{ijk} = \frac{p_{ij}^B p_{jk}^C}{p_j^B} \]
\[ H_{2c} : p_{ijk} = \frac{p_{ik}^A p_{jk}^C}{p_k^C} \]

svarende til rækker og lag uafhængige for givet søjle \((H_{2b} : A \perp C|B)\) og at rækker og søjl er uafhængige for givet lag \((H_{2c} : A \perp B|C)\).

Denne hypotese er den væsentligste hypotese i en trevejsklassifikation, fordi den udtrykker, at sammenhængen mellem to af klassifikationerne kan forklares ved sammenhængen med den tredje klassifikation.

Det skal bemærkes, at hypoteser af formen \(H_{2a}\) ikke indebærer, at \(B\) og \(C\) er uafhængige. Selv om hypoteseen \(H_{2a}\) er accepteret, indebærer dette ikke, at man vil finde uafhængighed mellem \(B\) og \(C\) i den marginale tovejsklassifikation efter \(B\) og \(C\).

Der er en sidste hypoteseform, \(H_3\), som vi skal betragte i forbindelse med denne multinomialfordelingsmodel. Denne sidste hypotese kan imidlertid ikke udtrykkes som en uafhængighedshypotese.

Hypotesen \(H_3\) udtrykker, at Yules krydsproduktratio (4.3.1)

\[ \omega_{i_1,i_2;j_1,j_2}^{A:B} = \frac{p_{i_1,j_1,k}}{p_{i_1,j_2,k}} \frac{p_{i_2,j_1,k}}{p_{i_2,j_2,k}} \]

ikke afhænger af laget \(k\).

Ved at ombytte lag med søjl ser man, at hypotesen tilsvarende udtrykker, at

\[ \omega_{i_1,i_2;j;k_1,k_2}^{A:C} = \frac{p_{i_1,j,k_1}}{p_{i_1,k_2}} \frac{p_{i_2,j,k_1}}{p_{i_2,j,k_2}} \]

ikke afhænger af søjlen \(j\).
Ved i stedet at ombytte lag med rækker kan hypotesen endelig udtrykkes ved, at

\[ \omega_{i;j_1,j_2;k_1,k_2} = \frac{p_{i,j_1,k_1}}{p_{i,j_2,k_2}} \frac{p_{i,j_1,k_2}}{p_{i,j_2,k_2}} \]

ikke afhænger af rækken \( i \).

Dvs. andenordens differenser \( \Delta_{i_1,i_2}^A \Delta_{j_1,j_2}^B \) er konstante, svarende til, at tredieordensdifferenser er nul.

Uafhængighedsgrafen svarende til denne hypotese er:

```
    C
   /\  \
  /   \ /
 /     \|
A-------B
```

Modellerne kan udtrykkes som såkaldte log-lineære modeller, ved at udtrykke logaritmen til de forventede værdier af celleværdierne

\[ \eta_{i,j,k} = \ln (E [X_{i,j,k}]) = Np_{i,j,k} \]

som et additivt udtryk.

Vi skriver dem først på den mættede (saturated) form

\[
H_0 : \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C \\
H_{1a} : \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + (\beta \gamma)_j^{BC} \\
H_{1b} : \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + (\alpha \gamma)_i^{AC} \\
H_{1c} : \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + (\alpha \beta)_i^{AB} \\
H_{2a} : \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + (\alpha \beta)_{i,j}^{AB} + (\alpha \gamma)_i^{AC} \\
H_{2b} : \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + (\alpha \beta)_{i,j}^{AB} + (\beta \gamma)_j^{BC} \\
H_{2c} : \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + (\alpha \gamma)_{i,k}^{AC} + (\beta \gamma)_j^{BC} \\
H_3 : \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + (\alpha \beta)_{i,j}^{AB} + (\beta \gamma)_{j,k}^{BC} + (\alpha \gamma)_{i,k}^{AC} 
\]

Vi kan fjerne overparametriseringen og opskrive modellerne ved frembringerne, angivet til højre i nedenstående oversigt.
Hypoteserne kan organiseres hierarkisk, som illustreret i nedenstående inklusionsdiagram, hvor symbolerne $A.B$ angiver at leddet $(\alpha \beta)_{i,j}^{AB}$ indgår i modellen, og hvor den øverste modell, $A.B.C$ symboliserer den fulde model, hvor hver celle har sin egen middelværdi:

\[
\begin{align*}
H_0 & \quad \eta_{i,j,k} = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C & [A][B][C] \\
H_{1a} & \quad \eta_{i,j,k} = \alpha_i^A + \beta_j^B + \gamma_k^C + (\beta \gamma)_{j,k}^{BC} & [A][BC] \\
H_{1b} & \quad \eta_{i,j,k} = \beta_j^B + (\alpha \gamma)_{i,k}^{AC} & [B][AC] \\
H_{1c} & \quad \eta_{i,j,k} = \gamma_k^C + (\alpha \beta)_{j,k}^{AB} & [C][AB] \\
H_{2a} & \quad \eta_{i,j,k} = (\alpha \beta)_{i,j}^{AB} + (\alpha \gamma)_{i,k}^{AC} & [AB][AC] \\
H_{2b} & \quad \eta_{i,j,k} = (\alpha \beta)_{i,j}^{AB} + (\beta \gamma)_{j,k}^{BC} & [AB][BC] \\
H_{2c} & \quad \eta_{i,j,k} = (\alpha \gamma)_{i,k}^{AC} + (\beta \gamma)_{j,k}^{BC} & [AC][BC] \\
H_3 & \quad \eta_{i,j,k} = (\alpha \beta)_{i,j}^{AB} + (\beta \gamma)_{j,k}^{BC} + (\alpha \gamma)_{i,k}^{AC} & [AB][AC][BC]
\end{align*}
\]

Eksempel 4.6.1 *Simpsons paradox, the National Halothane study*
Andel overlevende ved to behandlinger

<table>
<thead>
<tr>
<th></th>
<th>Behandling</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>Ialt</td>
</tr>
<tr>
<td>Køn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mænd</td>
<td>60/80</td>
<td>100/150</td>
<td>160/230</td>
</tr>
<tr>
<td>Kvinder</td>
<td>40/120</td>
<td>10/40</td>
<td>50/160</td>
</tr>
<tr>
<td></td>
<td>100/200</td>
<td>110/190</td>
<td>210/390</td>
</tr>
</tbody>
</table>

Den marginale tabel antyder, at behandling II er bedre end behandling I, selv om behandling I er bedre end behandling II, såvel for mænd som for kvinder. Denne tilsyneladende modstrid betegnes Simpson’s paradoks (Simpson (1951), allerede påpeget af Yule (1903)).

Simpson’s paradoks kan optræde fordi aggregering kan føre til uhensigtsmæssig vægtning af de forskellige populationer.

Behandling I blev givet til 80 mænd og 120 kvinder. Den aggregerede marginal er et vejet gennemsnit af succesraten for mænd og succesraten for kvinder med en svag overvægt på succesraten for kvinder.

Behandling II blev givet til 150 mænd og kun til 40 kvinder, hvorfor den aggregerede succesrate er et vejet gennemsnit af succesraten for mænd og kvinder med den største vægt givet til mændenes succesrate. Groft taget kan man sige, at den marginale sammenligning er en sammenligning mellem succesraten for behandling I, som er gennemsnittet af mænds og kvinders succesrate, med succesraten for behandling II, som essentielt er mændenes succesrate. Da succesraten for mænd er meget større, end for kvinder, giver den marginale tabel den illusion at behandling II er den bedste.

Moralen af dette eksempel er, at man ikke nødvendigvis kan stole på konklusioner, der er truffet på baggrund af marginale tabeller. Sådanne aggregerede sammenligninger kan være forplumrede af forskellige sammenvejninger af de indgående delpopulationer. I almindelighed er det derfor nødvendigt, at betragte samtlige dimensioner i en tabel, dvs samtlige delpopulationer.

Dette resultat adskiller sig fra resultaterne fra fordelinger for kontinuerte variable.
Således har man, hvis man betræger den partielle korrelationskoefficient mellem to (blant tre) normalfordelte variable:

\[ \rho_{12|3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1 - \rho_{13}^2)(1 - \rho_{23}^2)}}. \]

Man har derfor, at hvis én af de marginale korrelationskoefficienter, \( \rho_{13} = 0 \) eller \( \rho_{23} = 0 \), da er \( \rho_{12|3} \) et multiplum af \( \rho_{12} \), og vi kan teste hypotesen \( \rho_{12|3} = 0 \) ved at teste \( \rho_{12} = 0 \).

I en tredimensionale tabell er man kun bruge de marginale summer opnået ved at marginalisere (collapse) over en tredie variabel til at måle vekselvirkningen mellem de to variable, hvis den tredie variabel er uafhængig af mindst én af de to variable, der indgår i denne vekselvirkning. Specielt kan en tilsyneladende afhængighed mellem to variable i marginale tabeller skyldes indflydelsen af en tredje variabel.

### 4.7 Grafske modeller

**Definition 4.7.1 Grafsk model**

En model for en tabell siges at være en grafsk model, såfremt den er sådan, at hvis den indeholder alle tofaktorvekselevirkningerne svarende til en højere ordens vekselvirkning, da indeholder den også denne højere ordens vekselvirkning.

Enhver grafsk model er også en hierarkisk model, mens det omvendte ikke altid er tilfældet.

Ved analyse af tredimensionale tabeller vil det ofte være ønskeligt at indskrænke sig til grafske modeller, da sådanne modeller kan fortolkes ved relationer om betinget uafhængighed af visse variable, givet resten af de variable.

**Eksempel 4.7.1 Vekselvirkningsgraf for trefaktormodel**
Betrægt trefaktormodellen givet ved de sufficiente marginaler \([AB][BC][AC]\) svarende til den loglineære repræsentation

\[ \vartheta_{ijk} = (\alpha \beta)_{ij}^{AB} + (\beta \gamma)_{jk}^{BC} + (\alpha \gamma)_{jk}^{AC} \]

Modellens vekselvirkningsgraf er

![Diagram](image)

Modellen er en hierarkisk model.

Modellen er ikke grafisk, da den ikke indeholder det højere ordens vekselvirkningsled \((\alpha \beta \gamma)_{ijk}^{ABC}\)

Den grafiske model udspændt af \([AB][BC][AC]\) er i dette tilfælde den fulde model \([ABC]\) svarende til \((\alpha \beta \gamma)_{ijk}^{ABC}\).

\[\square\]

### 4.7.1 Faktorisering, Reducible komponenter

Reducibilitet er en egenskab ved en fordeling. Egenskaben er imidlertid isomorf med en kendt grafteoretisk egenskab: En graf, \(G\), er reducibel, hvis og kun hvis der findes en opdeling, \(G = a \cup b \cup c\), af graffen, hvor hverken \(b\) eller \(c\) er tomme, og hvor \(a\) separerer \(b\) og \(c\) og delgraferne på \(a\) er fuldstændig. Komponenterne af \(G\) er delgraferne på \(a \cup b\) og \(a \cup c\).

### 4.7.2 Dekomposable modeller

Introduceret af Goodman og Haberman. Dekomposable modeller kan karakteriseres på flere forskellige måder:

- Multiplikative modeller, hvor den simultane tæthed faktoriserer i et produkt af marginale tætheder. En sådan faktorisering er entydig og bestemmer alle modellens egenskaber.

- Den simultane tæthed faktoriserer i et produkt af marginale tæthedsfunktioner på klikker. De irreducible komponenter er fuldstændige.
Afsnit 4. Modeller for flerdimensionale antalstabeller

- Modellerne er rekursive modeller, dvs at deres kanter kan ordnes sådan at den rekursive faktorisering af den simultane tæthed forenkles. (Markov kæde egenskab)

- Modellerne har triangulerede uafhængighedsgrafer

- Maksimum likelihood estimerne for parametrene kan udtrykkes på lukket form.

Definition 4.7.2 Dekomposabel model

En fordeling siges at være dekomposabel, hvis og kun hvis den kan reduceres til fuldstændige irreducible komponenter.

Det gælder, at enhver dekomposabel model er en grafisk model, mens det omvendte ikke altid er tilfældet.

4.7.3 Strategier for modelvalg

I det foregående har vi indført forskellige egenskaber for log-lineære modeller.

Hierarkiske modeller, som tillader successiv testning af hypoteser og fjernelse af led.
Imidlertid er ikke alle hierarkiske modeller lige lette at fortolke, eksempelvis er modellen \([AB],[BC],[AC]\) drilagtig.

Grafiske modeller, som tillader fortolkning ved betinget uafhængighed, der undertiden ligefrem kan tillægges kausal betydning

Dekomposable modeller, som muliggør eksplicitte estimatorer for samtlige parametre, og som endvidere tillader en repræsentation som en rekursiv model, dvs. som en Markov kæde.

Whittaker (1990) giver en grundig gennemgang af grafiske modeller, der også omfatter grafiske modeller svarende til kontinuerte variable.

4.7 Generel formulering

Vi anfører endelig en helt anden angrebsvinkel til modellering af antalstabeller med flerdimensionalt respons, nemlig den kaldte korrespondanceanalyse se fx Greenacre (1984). Korrespondanceanalyseen sigter mod at reducere antallet af dimensioner i en antalstabel, i lighed med principal komponent analyse, eller faktoranalyse for flerdimensionale normalfordelinger.

4.8 Generel formulering af modeller for flerdimensionalt respons

Vi slutter dette afsnit med en (ganske abstrakt) formulering af modeller for flerdimensionale antalstabeller.

Vi betragter en mængde $F$ af klassifikationskriterier eller faktorer. For enhver faktor $\phi$ i $F$ lader vi $I_\phi$ angive mængden af niveauer for faktoren $\phi$.

Mængden af celler i antalstabellen er mængden

$$I = \prod_{\phi \in F} I_\phi$$

En bestemt celle vil blive benævnt $i = (i_\phi, \phi \in F)$.

En antalstabel er en klassifikation af $n$ objekter i oversættelse med kriterierne.

4.8.1 Relation til teorien for Markovfelter

Teorien for endelige Markovfelter opererer med en endelig mængde $\Gamma$ af sites. Mængden af sites svarer her til mængden af faktorer.

For enhver site, $\gamma \in \Gamma$, er der en endelig mængde, $I_{\gamma}$ af elementære tilstande. Mængden

$$I = \prod_{\gamma \in \Gamma} I_{\gamma}$$

kaldes mængden af konfigurationer. En given konfiguration betegnes med $i = (i_\gamma, \gamma \in \Gamma)$. 
Afsnit 4. Modeller for flerdimensionale antalstabeller

Endelig har man en ikke-orienteret graf, Π på Γ, dvs et par Π = (V(Π), E(Π)) bestående af en mængde, V(Π) af knuder (vertices), og en mængde E(Π) af kanter (edges), hvor E(Π) er en mængde af ikke-ordnede par af distinkte elementer i Γ.

To knuder α og β siges at være nabøer (adjacent), hvis og kun hvis \{α, β\} ∈ E(Π). I dette tilfælde skriver vi α ∼ β.

Lad α ⊆ Γ. Randen af α, δα er mængden af kanter i Γ\α, som er nabøer til en knude i α. Afslutningen af α er mængden α = α ∪ δα.

En fuldstændig delmængde er en delmængde α ⊆ Γ hvor alle elementer er indbyrdes nabøer. En klike er en maksimal fuldstændig delmængde.

Betragt nu en sandsynlighed P på I sådan at P(i) > 0 foralde i ∈ I, og betragt de stokastiske variable defineret ved koordinatprojektionerne:

\[
\begin{align*}
X_γ(i) &= i_γ, \quad \text{for } γ ∈ Γ \\
X_α(i) &= i_α, \quad \text{for } α ⊆ Γ, α ≠ Ø
\end{align*}
\]

**Definition 4.8.1 Markov felt**

Det stokastiske felt (Xγ, γ ∈ Γ) siges at være et Markov felt med hensyn til P og Γ (eller P er Markov med hensyn til Γ) hvis en af nedensstående ækvivalente betingelser er opfyldt:

a) For ethvert γ ∈ Γ er Xγ og XΓ\γ betinget uafhængige givet Xδγ

b) For ethvert par α, β ∈ Γ med οnsim β, er Xα og Xβ betinget uafhængige givet XΓ\{α, β\}

c) For ethvert α ⊆ Γ er Xα og XΓ\α betinget uafhængige givet Xδα

d) Hvis to disjunkte delmængder α ⊆ Γ og β ⊆ Γ er separeret af en delmængde δ ⊆ Γ, sådan at enhver sti fra α til β i Γ går gennem δ, da er Xα og Xβ betinget uafhængige givet Xδ.

Et potential er en reel funktion, $\Phi$ på $I$ af formen

$$\Phi(i) = \sum_{a \in \Gamma} \xi_a(i_a)$$

hvor funktionerne $\xi_a$ alene afhænger af $i$ gennem $i_a$. Funktionerne $\xi_a$ kaldes vekselvirkningspotentialerne.

Et sandsynlighedsmålt $P$ på $I$ kaldes en Gibbs tilstand med potential $\Phi$, såfremt der gælder

$$P(i) = \exp(\Phi(i))$$

Tilsvarende kaldes et sandsynlighedsmålt $P$ på $I$ for en Gibbs tilstand (med potential $\Phi(i) = \ln(P(i))$).

$\Phi$ kaldes et nærmeste-nabo potential, hvis det er opbygget af vekselvirkninger mellem indbyrdes naboer, dvs hvis $\xi_a = 0$, når ikke alle knuder i $a$ er indbyrdes naboer, dvs hvis $a$ ikke er en fuldstændig delmængde af $\Gamma$.

Et sandsynlighedsmålt $P$ kaldes en nærmeste-nabo Gibbs tilstand hvis og kun hvis $P$ er en Gibbs tilstand med potential $\Phi$, hvor $\Phi$ er et nærmeste nabo potential.

Et fundamentalt resultat i teorien for Markovfelter og nærmeste nabo Gibbs tilstande udsiger, at disse to begreber er identiske. Der gælder:

**Sætning 4.8.1** Aeqvivalens mellem nærmeste nabo Gibbs-tilstand og Markovfelter

Et sandsynlighedsmålt $P$ på $I$ er en nærmeste nabo Gibbs-tilstand hvis og kun hvis det tilsvarende stokastiske felt er et Markovfelt.

**Bevis:**
Se f.eks. Speed (1978)

4.8.2 Grafiske modeller og Gibbs tilstande

Betragt en antalstabel med en mængde $F$ af faktorer og antag, at vi har givet en graf $F$ på faktorerne, specificeret ved mængden $V(F)$ af knuder, og mængden $E(F)$ af kanter.

Lad $\mathcal{F}$ være mængden af kliker i $F$, dvs de maksimale fuldstændige delmængder.

Den grafiske model frembragt af $F$ er da den hierarkiske model med frembringende klasse $\mathcal{F}$.

Den frembringende klasse $\mathcal{F}$ definerer entydigt grafen $F$ ved relationen

$$\alpha \sim \beta \iff \exists f \in \mathcal{F} \quad \text{sådan at} \quad \{\alpha, \beta\} \subseteq c$$

Grafen $F$ er således blot en anden repræsentation af den frembringende klasse $\mathcal{F}$.

Den frembringende klasse $\mathcal{F}$ modsvarer nogle bånd på vekselvirkningerne mellem faktorerne i $F$.

Det følger således af definitionen på en hierarkisk model, at $\xi_a \equiv 0$ med mindre $a$ er indeholdt i en maksimal fuldstændig delmængde, dvs. med mindre $a$ er en fuldstændig delmængde.

Mængden af sandsynligheder $P$ i modellen er altså netop mængden af nærmeste-nabo Gibbs tilstande svarende til $F$.

Det følger således af sætning 4.8.1, at sandsynlighederne $P$, der er indeholdt i den grafiske model, netop er de sandsynligheder, der gør $(X_{\gamma}, \gamma \in F)$ til et markovfelt.

Den grafiske model kan således beskrives ved de betingelser vedrørende betinget uafhængighed, der er udtrykt i de fire ækvivalente formuleringer af Markov egenskaben i definition 4.8.1.

Det fremgår således specielt, at hvis to grupper af faktorer er i forskellige - hver for sig sammenhængende - dele af grafen, da er de uafhængige.

Hvis to faktorer ikke er naboer, da er de betinget uafhængige givet de andre faktorer. Hvis to grupper af faktorer $a$ og $b$ er separeret af en gruppe $d$, da er de betinget uafhængige, givet faktorerne i $d$. 
4.9 Referencer

 Ikke alle hierarkiske modeller er grafiske. Til en vilkårlig frembringende klasse kan man dog altid knytte en graf. En sådan graf vil delvist fastlægge veckelvirkningsstrukturen.

 Lad $\mathcal{F}$ være en frembringende klasse og sæt $F = \bigcup_{f \in \mathcal{X}}$. Vi kan da definere en graf $G = (V(\mathcal{F}), E(\mathcal{F}))$ ved at sætte $V(\mathcal{F}) = F$ og definere en kant ved at der er førsteordensveckelvirkning mellem de indgående faktorer, dvs. $\{f_1, f_2\} \in E(\mathcal{F})$ hvis og kun hvis der findes et $f \in \mathcal{F}$ sådan at $\{f_1, f_2\} \subseteq c$. En sådan graf svarer netop til de grafiske repræsentationer, vi tidligere har betragtet, med hovedeffekterne som knuder, og første-ordens veckelvirkninger som kanter.

 $\mathcal{F}$ svarer til en grafisk model, netop hvis $\mathcal{F}$ består af alle klikerne i denne graf. Hvis dette er tilfældet, kalder vi $\mathcal{F}$ for en grafisk frembringende klasse.

 Hvis der er kliker i grafen, som ikke er i $\mathcal{F}$, da er $\mathcal{F}$ ikke grafisk, og veckelvirkningsstrukturen i modellen kan ikke beskrives fuldstændigt alene ved grafen.

 Dette indebærer at veckelvirkningsstrukturen i en grafisk model er fastlagt gennem første-ordens veckelvirkningerne, da disse veckelvirkninger definerer grafen, som så bestemmer klikerne - og dermed veckelvirkningerne af højere orden.

4.9 Referencer


Afsnit 5

Hierarkiske modeller for endimensionale normalfordelinger

5.1 Indledning og notation

Vi skal i dette og de følgende afsnit betragte modeller for data, der kan opfattes som $k$ grupper af observationer, hvor forsøgsomstændighederne, eller -objekterne er tilstræbt rimeligt ensartede inden for hver af grupperne, men hvor omstændighederne antages at variere fra gruppe til gruppe.

Vi repræsenterer observationerne ved skemaet

<table>
<thead>
<tr>
<th>Gruppe</th>
<th>Observationer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$</td>
</tr>
<tr>
<td>2</td>
<td>$X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$k$</td>
<td>$X_{k,1}, X_{k,2}, \ldots, X_{k,n_k}$</td>
</tr>
</tbody>
</table>
svarende til en klassifikation i \( k \) grupper med \( n_i, (i = 1, 2, \ldots, k) \) gentagne observationer i hver af de \( k \) grupper. (Klassifikationen svarende til gentagelserne er således underordnet gruppeklassifikationen).

Et udtryk for niveauet af observationerne svarende til den \( i \)’te gruppe er gennemsnittet af disse observationer,

\[
\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}
\]  

(5.1.1)

og variabiliteten af disse observationer kan udtrykkes ved kvadratafvigelsessummen for den \( i \)’te række

\[
\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,
\]

normeret på passende måde.

Et fælles udtryk for variabiliteten af observationerne ved gentagelse under ensartede omstændigheder (indenfor en gruppe) fås ved at addere disse kvadratafvigelsessummer. Vi benytter betegnelsen

\[
SAK_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2
\]  

(5.1.2)

som udtryk for variationen indenfor grupper.

Et udtryk for det fælles niveau, der er karakteristisk for hele samlingen af observationer, er det fælles gennemsnit,

\[
\bar{X}_* = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}}{N} = \frac{\sum_{i=1}^{k} n_i \bar{X}_i}{\sum_{i=1}^{k} n_i}
\]  

(5.1.3)

hvor \( N \) angiver det totale antal observationer,

\[
N = \sum_{i=1}^{k} n_i
\]  

(5.1.4)

Vi bemærker, at det fælles gennemsnit kan beregnes som det vejede gennemsnit af gruppegennemsnitlerne med de respektive observationsantal som vægte.
5.1 Indledning og notation

Gruppegennemsnittenes variabilitet udtrykkes naturligt ved kvadratafvigelsessummen

\[ SAK_2 = \sum_{i=1}^{k} n_i (X_{i.} - \bar{X}.)^2 \quad (5.1.5) \]

der måler gruppegennemsnittenes afvigelse fra totalgennemsnittet, vægtet med antallet af observationer, der indgår i det pågældende gruppegennemsnit, idet dette antal jo er et udtryk for det pågældende gennemsnits præcision.

Vi erindrer om den pythagoræiske relation

\[ \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}.)^2 = SAK_1 + SAK_2 \quad (5.1.6) \]

der viser hvorledes observationernes variation omkring det fælles gennemsnit kan spaltes i et udtryk for den interne variation i grupperne, \( SAK_1 \), og et udtryk for gruppegennemsnittenes variation, \( SAK_2 \).

Til støtte for beregningerne kan benyttes

\[
SAK_1 = \sum_{i=1}^{k} SK_i - \sum_{i=1}^{k} \frac{S_i^2}{n_i}
\]

\[
SAK_2 = \sum_{i=1}^{k} \frac{S_i^2}{n_i} - \left( \sum_{i=1}^{k} S_i \right)^2 / N
\]

med

\[ S_i = \sum_{j=1}^{n_i} X_{ij} \quad \text{og} \quad SK_i = \sum_{j=1}^{n_i} X_{ij}^2 \]

Sætning 5.1.1 Forventningsværdi af variationen mellem gruppegennemsnittene

Såfremt gruppegennemsnittene \( \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k \) er indbyrdes uafhængige med samme forventningsværdi \( E[\bar{X}_i] = \mu \), gælder
\[ E[S_{AK2}] = \sum_{i=1}^{k} n_i(1 - w_i) \sqrt{\frac{X_i - \bar{X}}{N}} \] (5.1.7)

med stikprøveandelen \( w_i \) givet ved

\[ w_i = n_i / N \] (5.1.8)

hvor \( N = \sum n_i \)

**Bevis:**
Følger af resultater i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998 afsnit 0.2.1.

I en række tilfælde vil vi endvidere få brug for den "vægtede gennemsnitlige gruppestørrelse", \( n_0 \)

\[ n_0 = \frac{\sum_{i=1}^{k} n_i - (\sum_{i=1}^{k} n_i^2 / \sum_{i=1}^{k} n_i)}{k - 1} = \left( N - \frac{\sum_{i=1}^{k} n_i^2}{N} \right) / (k - 1) \] (5.1.9)

Størrelsen \( n_0 \) er beskrevet i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998, lemma 0.2.1, formel (0.2.5).

I det balancede tilfælde, hvor der er lige mange observationer i hver gruppe, d.v.s. \( n_1 = n_2 = \cdots = n_k = n \) fås netop

\[ n_0 = n, \]

den fælles gruppestørrelse.

### 5.2 Ensidede variansanalyse i den systematisk-ske model

Vi indleder med at resumere den ensidede variansanalysemodel for normalfordelte data i tilfældet med en såkaldt systematisk klassifikation, dvs. en
klassifikation, hvor man opfatter de \( k \) grupper som faste. Den \( i \)’te gruppe er udvalgt netop med henblik på at repræsentere omstændigheder, der er væsensforskellige fra den \( j \)’te gruppe.

Modellen bygger på antagelsen

\[
X_{ij} \in N(\mu_i, \sigma^2) \quad (5.2.1)
\]

hvor \( X_{ij} \) er indbyrdes uafhængige.

Da den kanoniske link for normalfordelingen er identiteten, og da normalfordelingen udfylder hele den reelle akse ser man ofte modellen formuleret som en additiv model for observationerne,

\[
X_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (5.2.2)
\]

hvor

\[
\epsilon_{ij} \in N(0, \sigma^2) \quad (5.2.3)
\]

er indbyrdes uafhængige.

For at sikre entydigheden af parametriseringen svarende til (5.2.2) og (5.2.3) fastlægges båndet

\[
\sum_{i=1}^{k} n_i \alpha_i = 0 \quad (5.2.4)
\]

Ovenstående model svarer til at den tilfældige variation, der modelleres, alene er gentagelsesvariationen indenfor grupper. Modellen kaldes underti- den Model I.

Sætning 5.2.1 Den systematiske model for ensidet variansanalyse

For modellen givet ved (5.2.1) gælder:

\[
E[X_{ij}] = \mu_i \quad (5.2.5)
\]

\[
\text{COV}[X_{ij}, X_{hl}] = \begin{cases} 
\sigma^2 & \text{for } (i, j) = (h, l) \\
0 & \text{ellers}
\end{cases}
\]
samt

\[ E[\bar{X}_i] = \mu_i \quad \text{(5.2.6)} \]

\[
\text{COV}[\bar{X}_i, \bar{X}_h] = \begin{cases} 
\sigma^2/n_i & \text{for } i = h \\
0 & \text{ellers}
\end{cases}
\]

Endelig har man

\[ SAK_1 \in \sigma^2 \chi^2(N - k) \quad \text{(5.2.7)} \]

og

\[ SAK_2 \in \sigma^2 \chi^2(k - 1, \lambda) \quad \text{(5.2.8)} \]

hvor \( SAK_1 \) og \( SAK_2 \) er stokastisk uafhængige, og hvor ikke-centralitetsparametrene \( \lambda \) er givet ved

\[
\lambda = \frac{\sum_i n_i(\mu_i - \mu_0)^2}{(k - 1)\sigma^2} \quad \text{(5.2.9)}
\]

med

\[
\mu_0 = \frac{\sum_{i=1}^k n_i \mu_i}{\sum_{i=1}^k n_i} \quad \text{(5.2.10)}
\]

**Bevis:**
Sætningen bevises direkte

\[ \square \]

Antagelserne kan undersøges grafisk, f.eks. ved indtegning af gruppevise fraktildiagrammer. Antagelsen om varianshomogenitet kan undersøges grafisk ved vurdering af parallelliteten af de gruppevise diagrammer, antagelsen kan testes, f.eks. ved Bartlett’s test (eksempel 2.7.10, side 239).

Ofte er man interesseret i at undersøge kontraster mellem to grupper. Ni-veauforskellen mellem grupperne \( i \) og \( h \) er \( \mu_i - \mu_h \), som estimeres ved \( \bar{X}_i - \bar{X}_h \).

Der gælder:
Sætning 5.2.2 Sammenligning af grupper i den systematiske model for ensidet variansanalyse

For modellen givet ved (5.2.1) gælder:

\[ E \left[ X_i - \bar{X}_h \right] = \mu_i - \mu_h \]  
(5.2.11)

\[ \text{V} \left[ X_i - \bar{X}_h \right] = \sigma^2 \left( \frac{1}{n_i} + \frac{1}{n_h} \right) \text{ for } i \neq h \]  
(5.2.12)

\[ \text{V} \left[ X_i - \bar{X}_h \right] = \sigma^2 \left( \frac{1}{n_i} + \frac{1}{n_h} \right) \text{ for } i \neq h \]  
(5.2.13)

Bevis:
Sætningen bevises direkte.

Ofte ønsker man at foretage en sammenligning af alle grupper under eet. Den hypotese, at de \( k \) forskellige forsøgsomstændigheder ikke giver anledning til påviselige forskelle i observationerne formuleres som

\[ H_I : \mu_1 = \cdots = \mu_k = 0 \]  
(5.2.14)

imod alternativet at \( \mu_i \neq \mu_h \) for mindst ét sæt indices \((i, h)\).

Hypotesen er en homogenitetshypotese i en generaliseret linæær model (se nedenstående bemærkning 1). Det er velkendt, blandt andet fra Introduktion til Statistik, Bind 1, at denne hypotese kan testes ved at splæte den totale variation i en del, der alene kan tilskrives gentagelserne, og en del, der beskriver variationen mellem forsøgsomstændigheder.

Sætning 5.2.3 Test for fuldstændig homogenitet i den systematiske model for ensidet variansanalyse

lad

\[ Z = \frac{SAK_2/(k-1)}{SAK_1/(N-k)} \]  
(5.2.15)

Under modellen givet ved (5.2.1) og (5.2.3) gælder:

\[ Z \in F(k-1, N-k, \lambda), \]  
(5.2.16)

hvor ikke-centralitetsparameteren \( \lambda \) er bestemt ved (5.2.9)
Kvotientteststørrelsen for hypotesen (5.2.14) har det kritiske område:

\[ C = \{ z | z > F(k - 1, N - k)_{1-\alpha} \} \]  

(5.2.17)

**Bevis:**

I praksis udføres testet ved at betragte variansanalyse-skemaet svarende til opspaltningen (5.1.6)

<table>
<thead>
<tr>
<th>Variation</th>
<th>SAK</th>
<th>f</th>
<th>( \mathbb{E}[SAK/f] )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mellem grupper</strong></td>
<td>( \sum_i n_i (\overline{X_i} - \overline{X}.)^2 )</td>
<td>( k - 1 )</td>
<td>( \sigma^2 + \frac{1}{(k - 1)} \sum_i n_i (\mu_i - \mu_0)^2 )</td>
</tr>
<tr>
<td><strong>Indenfor grupper</strong></td>
<td>( \sum_i \sum_j (X_{ij} - \overline{X_i})^2 )</td>
<td>( N - k )</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>( \sum_i \sum_j (X_{ij} - \overline{X}.)^2 )</td>
<td>( N - 1 )</td>
<td>[]</td>
</tr>
</tbody>
</table>

hvor \( \mu_0 \) er givet ved (5.2.10).

**Bemærkning 1** Formulering af modellen som generaliseret lineær model

Modellen i sætning 5.2.1 (den systematiske model) kan formuleres som en generaliseret lineær model ved som vanligt at opstille samtlige observationer i én søjle, organiseret efter grupper.

Vi vil i det følgende betegne observationerne med \( Y_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, k, \) for at kunne reserveve symbolet \( \mathbf{X} \) til modellmatricen.

Opfatter vi observationerne \( \mathbf{Y} \) som en søjlevektor har vi, at den systematiske model svarer til en generaliseret lineær model for \( \mathbf{Y} \), hvor \( Y_{ij} \in N(\mu_i, \sigma^2) \), dvs. at linkfunktionen er den identiske afbildning (den kanoniske link for normalfordelingen) og dispersionsparameteren er \( \sigma^2 \).
I analogi med formuleringen som en generaliseret lineær model vælger vi at benytte symbolet $\beta$ for parameteren. Da linkfunktionen er identiteten har vi altså $\eta \equiv \mu$, dvs. $\beta_i \equiv \mu_i$, dvs
\[ H_0 : E[Y_{ij}] = \beta_i \quad (5.2.18) \]
uden bånd på $\beta$-værdierne.

Dispersionsmatricen for søjlevektoren $Y$ er
\[ D[Y] = \sigma^2 I, \quad (5.2.19) \]
hvor $I$ angiver den $N \times N$ dimensionale enhedsmatrix.

Modelmatricen svarende til denne parametrizering er blot incidensmatricen $U$ (afsnit 2.9.2), hvor den $i$'te søjle i $U$ har ettaller på pladserne svarende til den $i$'te gruppe, og nuller ellers.

Middelværdiligningen (2.5.6) på side 178 bliver derfor
\[ U^T \mu(\beta) = U^T y \]
der ved indsættelse af (5.2.18)
\[ \mu(\beta) = U\beta \]
bliver
\[ U^T U\beta = U^T y. \quad (5.2.20) \]
Idet
\[ U^T U = \text{diag}\{n_i\} \]
har man
\[ [U^T U]^{-1} = \text{diag}\{n_i^{-1}\} \]
hvorfor
\[ \hat{\beta} = [U^T U]^{-1} U^T y = \text{diag}\{n_i^{-1}\} U^T y \]
Man har derfor parameterestimaterne
\[ \hat{\beta}_i = \bar{y}_{i+} \]
Man kan vise, at hat-matricen $H = U[U^T U]^{-1} U^T$ kan udtrykkes som

$$H = U[U^T U]^{-1} U^T = \text{Blok diag}\{n_i^{-1} J_{n_i}\}, \quad (5.2.21)$$

hvor Blok diag betyder en blokdiagonal matrix, og hvor $J_n$ angiver en $n \times n$ matrix med ludder i etaller.

De fittede værdier er således

$$\hat{\mu} = Hy$$

Hypotesen (5.2.14) formuleres her som

$$H_M : \mu_1 = \mu_2 = \cdots = \mu_k \ (= \alpha) \quad (5.2.22)$$

dvs med modelmatricen svarende til den konstante faktor,

$$U_M = 1_N$$

hvorfor middelværdiligningen bliver

$$1_N^T \mu(\alpha) = 1_N^T y$$

dvs

$$1_N^T 1_N \alpha = 1_N^T y$$

eller

$$N \alpha = \sum_i \sum_j y_{ij}$$

med løsningen

$$\hat{\alpha} = [1_N 1_N]^{-1} 1_N^T y = \bar{y}_{++} \quad (5.2.23)$$

De fittede værdier bliver

$$\hat{\mu} = H_0 y \quad (5.2.24)$$

med hat-matricen

$$H_0 = \frac{1}{N} J_N$$
5.2  Ensidet variansanalyse i systematisk model

Deviansbidragene er netop de kvadratiske afvigelser

\[ d(y; \hat{\mu}) = (y - \hat{\mu})^2 \]

Man får derfor

\[ G^2(H_0) = y^T(I_N - H)y = \sum_i \sum_j (y_{ij} - \bar{y}_{i+})^2 \]

og

\[
G^2(H_M | H_0) = y^T(I_N - H)y - y^T(I_N - H_0)y \\
= y^T(H - H_0)y \\
= \sum_{i=1}^{k} n_i(\bar{y}_{i+} - \bar{y}_{++})^2
\]

alså netop kvadratafviselsessummen mellem grupper.

Udtrykt ved disse matricer bliver variansanalysekemaet

<table>
<thead>
<tr>
<th>Variation</th>
<th>( f )</th>
<th>SAK matrixform</th>
<th>SAK analytisk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mellem grupper</td>
<td>( k - 1 )</td>
<td>( y^T(H - H_0)y )</td>
<td>( \sum_i n_i(\bar{y}<em>{i+} - \bar{y}</em>{++})^2 )</td>
</tr>
<tr>
<td>Indenfor grupper</td>
<td>( N - k )</td>
<td>( y^T(I_N - H)y )</td>
<td>( \sum_i \sum_j (y_{ij} - \bar{y}_{i+})^2 )</td>
</tr>
<tr>
<td>Total</td>
<td>( N - 1 )</td>
<td>( y^T(I_N - H_0)y )</td>
<td>( \sum_i \sum_j (y_{ij} - \bar{y}_{++})^2 )</td>
</tr>
</tbody>
</table>

Maksimum-likelihood estimatoren for \( \sigma^2 \) under \( H_0 \) bliver

\[
(\sigma^*)^2 = \frac{1}{N} y^T(I_N - H)y \\
= \frac{1}{N} \sum_i \sum_j (y_{ij} - \bar{y}_{i+})^2 \quad (5.2.25)
\]

Det gælder imidlertid, at størrelsen \( y^T(I_N - H)y \) er likelihood-sufficient (def. 2.1.5). for \( \sigma^2 \).
Man vælger derfor at estimere $\sigma^2$ ud fra fordelingen af den likelihoodsuffici- ciente størrelse.

Idet

$$y^T (I_N - H)y \in \sigma^2 \chi^2(N - k)$$

finder man ved betragtning af likelihood'en svarende til denne fordeling, at maksimum-likelihood estimatet for $\sigma^2$ netop bliver det centrale estimat,

$$\hat{\sigma}^2 = \frac{1}{N - k} y^T (I_N - H)y = \frac{SAK_1}{N - k}$$

$\square$

5.3 Ensidet variansanalyse i den tilfældige model

I modsætning til den systematiske model, hvor effekten af den grupperende faktor blev modelleret som en systematisk (deterministisk) effekt, og kun gentagelseseffekten fortolkes som tilfældig, modellerer man i den tilfældige model også effekten af den grupperende faktor som tilfældig.

Den tilfældige model kaldes undertiden Model II.

Vi antager i lighed med det foregående afsnit, at

$$Y_{ij} | \mu_i \in N(\mu_i, \sigma^2) , \quad (5.3.1)$$

men i modsætning til den systematiske model, beskriver vi gruppeniveauet $\mu_i$ som en realisation af en stokastisk variabel.

Vi antager nemlig, at

$$\mu_i \in N(\mu_0, \sigma_0^2) , \quad (5.3.2)$$

hvor $\mu_i$'erne antages indbyrdes uafhængige, og $Y_{ij}$ antages at være indbyr- des uafhængige i den betingede fordeling af $Y_{ij}$ for givet $\mu_i$. 
I lighed med den systematiske model kan man også formulere denne model som en additiv model for observationerne $Y_{ij}$ ved
\[
Y_{ij} = \mu_0 + \alpha_i + \epsilon_{ij},
\]
(5.3.3)
med $\epsilon_{ij} \in N(0, \sigma^2)$ og $\alpha_i \in N(0, \sigma^2_0)$, hvor $\epsilon_{ij}$ er indbyrdes uafhængige, og $\alpha_i$ ligeledes er indbyrdes uafhængige og endelig er $\alpha_i$'erne uafhængige af $\epsilon_{ij}$.

Da det ofte vil være af interesse at betragte forholdet mellem de to varianser, indfører vi symbolet $\gamma$ for dette forhold,
\[
\gamma = \frac{\sigma^2_0}{\sigma^2}
\]
(5.3.4)

Parameteren $\gamma$ udtrykker således inhomogeniteten mellem grupper i forhold til gruppernes interne variation.

Vi vil ofte bruge betegnelsen **signal/støj forholdet** for parameteren $\gamma$.

Den tilfældige model vil være rimelig i situationer, hvor interessen ikke er indskrænket til alene de betragtede forsøgsomstændigheder (grupper), men hvor disse omstændigheder snarere opfattes som repræsentative for en større samling (population) af varierende forsøgsomstændigheder, principielt udtaget tilfældigt fra denne population.

Vi siger ofte, at vi har **hierarkisk variation**, eller at vi har en hierarkisk model, svarende til at vi har en tilfældig variation mellem grupper (underordnet den konstante faktor), og gentagelsen inden for grupper er underordnet inddelingen i grupper.

Til illustration af forskellen mellem den tilfældige og den systematiske model bemærker vi yderligere, at analysen af den systematiske model lægger vægt på vurderingen af resultaterne i de enkelte grupper, $\mu_i$, og eventuelle forskelle, $\mu_i - \mu_h$, på resultaterne i specifikke grupper, mens analysen af den tilfældige model i første række sigter mod at beskrive variationen mellem grupperne, $V[\mu_i] = \sigma^2_0$. Den tilfældige model er et specialtilfælde af den mere generelle **varianskomponentmodel**, og $\sigma^2_0$, variansen for den tilfældigt modellerede effekt, kaldes en **varianskomponent**.
Undertiden kan det have interesse nøjere at betragte resultaterne under netop de forsøgsomstændigheder, der tilfældigvis indgår i stikprøven. Det vil da være naturligt at betragte de betingede fordelinger af en given række i skemaet for en fastholdt værdi af μ\(_i\) (dvs svarende til gentagelser inden for denne (tilfældigt valgte) gruppe). Der gælder

**Sætning 5.3.1** Betingede fordelinger i den tilfældige model for ensidet variansanalyse

Under modellen givet ved (5.3.1) og (5.3.2) vil den betingede fordeling af \(Y_{ij}\) givet \(μ_i\) være en normalfordeling med

\[
E [Y_{ij}|μ_i] = μ_i
\]

\[
COV[Y_{ij}, Y_{il}|μ_i] = \begin{cases} 
0 & \text{for } j \neq l \\
σ^2 & \text{for } j = l
\end{cases}
\]

og endvidere er

\[
E [\bar{Y}_i|μ_i] = μ_i, \quad \text{og} \quad V [\bar{Y}_i|μ_i] = σ^2/n_i
\]

**Bevis:**

Beviset følger direkte af antagelserne

Betrænger vi derimod de marginale fordelinger, d.v.s. fordelingerne svarende til gentagelser af hele forsøget, finder vi

**Sætning 5.3.2** Marginale fordelinger i den tilfældige model for ensidet variansanalyse

Den marginale fordeling af \(Y_{ij}\) er en normal fordeling med

\[
E [Y_{ij}] = μ_0
\]

\[
COV[Y_{ij}, Y_{hl}] = \begin{cases} 
σ_0^2 + σ^2 & \text{for } (i, j) = (h, l) \\
σ_0^2 & \text{for } i = h, \ j \neq l \\
0 & \text{for } i \neq h
\end{cases}
\]
5.3 Ensidet variansanalyse i tilfældig model

Bevis:
Beviset følger direkte af antagelserne

Bemærkning 1 Observationer fra samme gruppe er korrelerede
Vi bemærker, at der en positiv kovarians mellem observationer fra samme gruppe. Denne positive kovarians udtrykker netop, at observationer inden for en gruppe vil afvige i samme retning fra den marginale middelværdi \( \mu_0 \), nemlig i retning mod den pågældende gruppemiddelværdi.

Korrelationskoefficienten

\[
\rho = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} = \frac{\gamma}{1 + \gamma}
\] (5.3.8)

der beskriver korrelationen indenfor gruppe, benævnes ofte intraklasseskorrelationen.

Bemærkning 2 Korrelations- og dispersionsmatrix for observationer fra samme gruppe
Betragter vi observationssættet svarende til den \( i \)’te gruppe som en \( n_i \)-dimensional søjlevektor,

\[
Y_i = \begin{pmatrix}
Y_{i1} \\
Y_{i2} \\
\vdots \\
Y_{in_i}
\end{pmatrix}
\]

har vi, at korrelationsmatricen i den marginale fordeling af \( Y_i \) er en equikorrelationsmatrix (se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998, afsnit 0.2.1) af formen

\[
E_{n_i} = (1 - \rho)I_{n_i} + \rho J_{n_i} = \begin{pmatrix}
1 & \rho & \cdots & \rho \\
\rho & 1 & \cdots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \cdots & 1
\end{pmatrix}
\] (5.3.9)
hvor $J_n$ er en $n \times n$-dimensional matrix bestående af lutter ettaller

Observationssættene $Y_i$, $i = 1, 2, \ldots, k$ kan således beskrives som $k$ uafhængige observationer af en $n_i$ dimensional variabel $Y_i \in N_{n_i}(\mu, \sigma^2 I_{n_i} + \sigma_n^2 J_{n_i})$, dvs at dispersionsmatricen for $Y_i$ er

$$V_i = D[Y_i] = E[(Y_i - \mu)(Y_i - \mu)^T]$$

$$= \begin{pmatrix}
\sigma^2 + \sigma^2 & \sigma^2 & \cdots & \sigma^2 \\
\sigma^2 & \sigma^2 + \sigma^2 & \cdots & \sigma^2 \\
\vdots & \vdots & \ddots & \vdots \\
\sigma^2 & \sigma^2 & \cdots & \sigma^2 + \sigma^2
\end{pmatrix}$$

(5.3.10)

En sådan matrix betegnes undertiden en sammensat symmetrisk matrix (eng. compound symmetric).

**Bemærkning 3** **Kovariansstruktur for hele observationssættet**

Opstiller vi samtlige observationer i én søje, organiseret efter grupper, ser vi, at den $N \times N$-dimensionale dispersionsmatrix $D[Y]$ er

$$V = D[Y] = \text{Blok diag}\{V_i\}$$

(5.3.11)

hvor $V_i$ er givet ved (5.3.10)

Tilsvarende finder man, at korrelationsmatricen for hele observationssættet er en $N \times N$-dimensional blokmatrix med matricerne $E_{n_i}$ i diagonalen, og nuller udenfor, hvilket illustrerer, at observationer fra forskellige grupper er uafhængige, mens observationer fra samme gruppe er korrelerede.

**Bemærkning 4** **Simultanfordeling af gruppegennemsnittene**

Vi bemærker endelig, at den simultane fordeling af gruppegennemsnittene er karakteriseret ved

$$\text{COV}[\bar{Y}_i, \bar{Y}_h] = \begin{cases}
\sigma^2 + \sigma^2/n_i & \text{for } i = h \\
0 & \text{ellers}
\end{cases}$$

(5.3.12)
5.3  Ensidet variansanalyse i tilfældig model

Det vil sige, at de $k$ gruppegennemsnit $\bar{Y}_i$, $i = 1, 2, \ldots, k$ er indbyrdes uafhængige, og variansen på gruppegennemsnittet,

$$\text{Var} [\bar{Y}_i] = \sigma^2_0 + \frac{\sigma^2}{n_i} = \sigma^2 (\gamma + 1/n_i)$$

omfatter både variansen på den tilfældige komponent, $\gamma_i$, og residualvariansen på gennemsnittet.

En forøgelse af stikprøvestørrelsen i de enkelte grupper vil således forøge præcisionen ved bestemmelse af gruppeforventningsværdien $\alpha_i$, men variationen mellem de enkelte gruppeforventningsværdier formindskes naturligvis ikke ved denne gennemsnitsdannelse.

Bemærkning 5  Indlejring af modellen i model for kvadratafvigelsessummer

I nogle fremstillinger indlejres modellen givet ved (5.3.1) og (5.3.2) i en mere generel model, der tager sit udgangspunkt i den ortogonale opspaltung af kvadratafvigelsessummen.

Under modellen (5.3.1) og (5.3.2) har vi

$$E [SAK_1/(N-k)] = \sigma^2 \text{ og } E [SAK_2/(k-1)] = \sigma^2_0 + n_0 \sigma^2$$

hvor den vægtede gennemsnitlige gruppestørrelse $n_0$ er givet ved (5.1.9).

Introducerer vi parameteren $\tau^2 > 0$ ved

$$E [SAK_2/(k-1)] = \tau^2 ,$$

kan $\tau^2$ og $\sigma^2$ estimeres uafhængigt af hinanden ved de respektive kvadratafvigelsessummer.

Vi betragter nu den mere generelle normalfordelingsmodel givet ved

$$\text{COV}[Y_{ij}, Y_{hl}] = \begin{cases} 
\nu^2 & \text{for } (i,j) = (h,l) \\
\lambda & \text{for } i = h, j \neq l \\
0 & \text{for } i \neq h
\end{cases} \quad (5.3.13)$$

med $\nu^2 > 0$ og $\lambda$ vilkårlig.
Hierark. mod. for endim. normalford.

Dispersionsmatricen svarende til den $n_i$-dimensionale observationsvektor $Y_i$ fra den $i$-te gruppe med middelværdivektoren $\mu$ (en vektor bestående af $n_i$ $\mu'$er) bliver således

$$D[Y_i] = E[(Y_i - \mu)(Y_i - \mu)^T] = \begin{pmatrix} \nu^2 & \lambda & \cdots & \lambda \\ \lambda & \nu^2 & \cdots & \lambda \\ \vdots & \vdots & \ddots & \vdots \\ \lambda & \lambda & \cdots & \nu^2 \end{pmatrix}$$

altså en equikorrelationsmatrix af formen

$$D_i = (\nu^2 - \lambda)I_{n_i} + \lambda J_{n_i}$$

Idet

$$\nu^2 = V[Y_{ij}] = \sigma^2 + (\tau^2 - \sigma^2)/n_0$$
$$\lambda = COV[Y_{ij}, Y_{i,i}] = (\tau^2 - \sigma^2)/n_0,$$

kan parametrene $\sigma^2$ og $\tau^2$ udtrykkes ved $\nu^2$ og $\lambda$ som

$$\sigma^2 = \nu^2 - \lambda, \quad \text{og} \quad \tau^2 = \nu^2 + (n_0 - 1)\lambda$$

Betingelsen $\sigma^2 > 0$ og $\tau^2 > 0$ fører da til begrænsningen på kovariansparameteren $\lambda$

$$-\frac{\nu^2}{n_0 - 1} \leq \lambda \leq \nu^2$$

Idet modellen givet ved (5.3.1) og (5.3.2) indebærer $0 \leq \lambda \leq \nu^2$ ser vi, at modellen (5.3.13) er en reel udvidelse. Udvidelsen består i, at der tillades negative kovarianser indenfor grupper. (Sådanne negative kovarianser er vanskelige at tolke, og vi vil derfor betragte den oprindelige model (5.3.1) og (5.3.2)).

Antagelserne kan undersøges grafisk, f.eks. ved indtegning af gruppevise fraktildiagrammer. Antagelsen om varianshomogenitet kan undersøges grafisk ved vurdering af paralleliteten af de gruppevise diagrammer. Antagelsen kan testes, f.eks. ved Bartlett’s test, eksempel 2.7.10, side 239.

5.3.1 Estimation af parametre i den tilfældige model

Såfremt man ikke har edb-programmer til rådighed benyttes ofte de såkaldte variansanalyseestimater, bestemt ved hjælp af momentmetoden. Der gælder
Sætning 5.3.3 Momentestimator (variansanalyseestimator) i den tilfældige model

Under modellen givet ved (5.3.1) og (5.3.2) finder man momentestimaterne for parametrene $\mu_0, \sigma^2$ og $\sigma_0^2$ ved

\[
\begin{align*}
\tilde{\mu}_0 &= \bar{Y}.
\tilde{\sigma}^2 &= \frac{SAK_1}{(N-k)} \\
\tilde{\sigma}_0^2 &= \frac{SAK_2/(k-1) - SAK_1/(N-k)}{n_0} = \frac{SAK_2/(k-1) - \tilde{\sigma}^2}{n_0}
\end{align*}
\]

hvor den vægtede gennemsnitlige gruppestørrelse $n_0$ er givet ved (5.1.9)

Bevis:

\[
\begin{align*}
\text{E} [SAK_1/(N-k)] &= \sigma^2 \\
\text{og} \\
\text{E} [SAK_2/(k-1)] &= \sigma^2 + n_0\sigma_0^2
\end{align*}
\]

Bemærkning 1 Trunkering af variansestimat til ikke-negative værdier

I stedet for den centrale estimator (5.3.14) benyttes ofte

\[
\tilde{\sigma}_0^2 = \max\{\tilde{\sigma}_0^2, 0\}
\]

da vi jo har, at $\sigma_0^2 > 0$.

Sætning 5.3.4 Varianser for momentestimatorer i den tilfældige model
Estimatorerne $\bar{\mu}, \bar{\sigma}^2$ og $\bar{\sigma}_0^2$ givet ved (5.3.14) er centrale, og varianserne for estimatorerne er givet ved

\[ V[\bar{\sigma}^2] = \frac{2\sigma^4}{N - k} \]

\[ V[\bar{\sigma}_0^2] = \frac{2\sigma^4}{n_0^2} A \]

samt

\[ \text{COV}[\bar{\sigma}^2, \bar{\sigma}_0^2] = -\frac{V[\bar{\sigma}^2]}{n_0} \]  \hspace{1cm} (5.3.16)

med

\[ A = \frac{1}{(k - 1)^2} \left\{ \left[ \sum_i \left\{ \frac{n_i}{w_i(\gamma)} \right\}^2 \right]^2 + \sum_i \left\{ \frac{n_i}{w_i(\gamma)} \right\}^2 \right\} + \frac{1}{N - k}, \]

hvor

\[ w_i(\gamma) = \frac{1}{1 + n_i \gamma} \]  \hspace{1cm} (5.3.18)

I det balancerede tilfælde, $n_1 = n_2 = \cdots = n_k = n$, reduceres udtrykkene til

\[ V[\bar{\sigma}^2] = \frac{2\sigma^4}{k(n - 1)} \]

\[ V[\bar{\sigma}_0^2] = \frac{2}{n^2} \left[ \frac{\sigma^4}{k(n - 1)} + \frac{(\sigma^2 + n\sigma_0^2)^2}{k - 1} \right] \]

Bevis:
Sætningen bevises direkte ud fra udtrykkene (5.3.14) \hfill \square

Bemærkning 1 Centrale estimatorer for varianskvotient i balanceret tilfælde

I det balancerede tilfælde, $n_1 = n_2 = \cdots = n_k = n$, kan vi angive eksplisitte centrale estimatorer for $\gamma$ og $w(\gamma) = 1/(1 + n\gamma)$. Der gælder
\[ \tilde{w} = \frac{SAK_1}{k(n-1)} \frac{SAK_2}{k-3} \]

og

\[ \tilde{\gamma} = \frac{1}{n} \left\{ \frac{SAK_2}{k-1} \frac{SAK_1}{k(n-1)-2} - 1 \right\} \]

er centrale estimatorer for henholdsvis \( w(\gamma) = 1/(1+n\gamma) \) og for \( \gamma = \sigma_0^2/\sigma^2 \)

\[ \square \]

### 5.3.2 Test af homogenitetshypotese i den tilfældige model

Den hypotese, at de varierende forsøgsomstændigheder er uden påviselig indflydelse på observationerne, formuleres under den tilfældige model som

\[ H_{II} : \sigma_0^2 = 0. \]

Hypotesen testes ved at sammenligne varianskvotienten

\[ Z = \frac{SAK_2/(k-1)}{SAK_1/(N-k)} \]

med fraktilerne i en F\((k-1, N-k)\)-fordeling. Der gælder

**Sætning 5.3.5 Test af homogenitetshypotese i den tilfældige model**
Hierark. mod. for endim. normalford.

Under modellen givet ved (5.3.1) og (5.3.2) har kvotienttestet for hypotesen (5.3.21) det kritiske område

\[ C = \{ z | z > F(k - 1, N - k)_{1 - \alpha} \} \]

hvor \( z \) er givet ved (5.3.22)

I det balancerede tilfælde, \( n_1 = n_2 = \ldots = n_k = n \), gælder

\[ Z \in (1 + n \gamma)F(k - 1, N - k) \]  \hspace{1cm} (5.3.23)

**Bevis:**
Se f.eks. Scheffé (1959)

Testet er således det samme som i den systematiske model

**Bemærkning 1 Konfidentsinterval for varianskvotienten**

I det balancerede tilfælde, \( n_1 = n_2 = \ldots = n_k = n \), kan man benytte (5.3.23) til at konstruere et konfidentsinterval for varianskvotienten \( \gamma \). Ved benyttelse af (5.3.23) finder man, at et \( 1 - \alpha \) konfidentsinterval for \( \gamma \), dvs. et interval \( (\gamma_L, \gamma_U) \), der tilfredsstiller

\[ P [\gamma_L < \gamma < \gamma_U] \]

fås ved at benytte

\[ \gamma_L = \frac{1}{n} \left( \frac{Z}{F(k - 1, N - k)_{1 - \alpha/2}} - 1 \right) \]

og

\[ \gamma_U = \frac{1}{n} \left( \frac{Z}{F(k - 1, N - k)_{\alpha/2}} - 1 \right) \]  \hspace{1cm} (5.3.24)

hvor \( Z \) er givet ved (5.3.22).
Eksempel 5.3.1 *Balancede data, variasランスeskema*


Med henblik på at vurdere forskellige stikprøveplaner til estimation af renheden af et parti bestående af adskillige baller uld har U.S. Customs Laboratory, Boston, blandt andet udtaget 4 prøver tilfældigt fra hver af 7 baller uruguyansk uld.

Prøveresultaterne er angivet i tabel 5.1.

| Prøve | Balle nr. | | | | | |
| --- | --- | --- | --- | --- | --- | --- | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 52.33 | 56.99 | 54.64 | 54.90 | 59.89 | 57.76 | 60.27 |
| 2 | 56.26 | 58.69 | 57.48 | 60.08 | 57.76 | 59.68 | 60.30 |
| 3 | 62.86 | 58.20 | 59.29 | 58.72 | 60.26 | 59.58 | 61.09 |
| 4 | 50.46 | 57.35 | 57.51 | 55.61 | 57.53 | 58.08 | 61.45 |

| Balle gennemsnit | 55.48 | 57.81 | 57.23 | 57.33 | 58.86 | 58.78 | 60.78 |

Vi udfører nu beregninger til en ensidet variasランスanalyse, og får variasランスeskemaet

<table>
<thead>
<tr>
<th>Variation</th>
<th>SAK</th>
<th>f</th>
<th>$s^2 = SAK/f$</th>
<th>$E[S^2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mellem baller</td>
<td>65.9628</td>
<td>6</td>
<td>10.9938</td>
<td>$\sigma^2 + 4\sigma_0^2$</td>
</tr>
<tr>
<td>Indenfor baller</td>
<td>131.4726</td>
<td>21</td>
<td>6.2606</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>
Udfører vi et test for hypotesen $H_0 : \sigma^2 = 0$, finder vi teststørrelsen

$$z = \frac{10.9938}{6.2606} = 1.76 < F_{0.95}(6, 21) = 2.57$$

Det kan således ikke afvises - ved test på et 5 % niveau - at variationen mellem renheden af ballerne ikke overstiger den interne variation i ballernes renhed.

Vort formål var imidlertid at beskrive variationerne i renheden af en sending, og vi vælger derfor den fulde model. Ved brug af momentestimaterne (5.3.14) finder vi estimatet for variationen indenfor en balle $\tilde{\sigma}^2 = 6.261$ og variationen mellem baller estimeres ved

$$\tilde{\sigma}^2 = \frac{10.9938 - 6.2606}{4} = 1.183$$

Partiets middelrenhed estimeres ved $\tilde{\mu}_0 = \tilde{\bar{y}}_\cdot = 58.04$. Usikkerheden på skønnene kan f. eks. estimeres ved indsættelse af de fundne værdier i udtrykkene (5.3.16) for varianserne på estimaterne,

$$\tilde{V}[\tilde{\sigma}^2] = \frac{2 \cdot (6.261)^2}{21} = (1.932)^2$$

og

$$\tilde{V}[\tilde{\sigma}_0^2] = \frac{2}{16} \left[ \frac{(6.261)^2}{21} + \frac{(10.994)^2}{6} \right] = (1.659)^2$$

der indikerer, hvorfor vi ikke kunne påvisse $\sigma_0^2 > 0$.

Ønsker vi et 95 % konfidensinterval for varianskvotienten $\gamma = \sigma_0^2/\sigma^2$ finder vi ved benyttelse af (5.3.24), at intervallet er bestemt som

$$\gamma_L = \frac{1}{4} \left( \frac{1.76}{F(6, 21)_{0.975}} - 1 \right) = 0.25 \times \left( \frac{1.76}{3.09} - 1 \right) = -0.11$$

$$\gamma_U = \frac{1}{4} \left( \frac{1.76}{F(6, 21)_{0.025}} - 1 \right) 0.25 \times (1.76 \times 6.31 - 1) = 2.53$$

idet $F(6, 21)_{0.025} = 1/F(21, 6)_{0.975}$
5.4 Likelihoodbaseret estimation i den tilfældige model

Sætning 5.4.1 Maksimum-likelihood estimator for parametrene under den tilfældige model

Under modellen givet ved (5.3.1) og (5.3.2) er maximum-likelihood estima-terne for $\mu$, $\sigma^2$ og $\sigma_0^2 = \sigma^2 \gamma$ bestemt ved

a) For $\sum_i n_i^2 (\bar{y}_i - \bar{y}_.)^2 < sak_1 + sak_2$ fås

$$\hat{\mu} = \bar{y}_. = \sum_i n_i \bar{y}_i / N$$

$$\hat{\sigma}^2 = (sak_1 + sak_2) / N$$

og

$$\hat{\gamma} = 0$$

(5.4.1)

b) For $\sum_i n_i^2 (\bar{y}_i - \bar{y}_.)^2 > sak_1 + sak_2$ bestemmes estima-terne som løsning til

$$\hat{\mu} = \sum_{i=1}^{k} n_i w_i(\hat{\gamma}) \bar{y}_i / W(\hat{\gamma})$$

(5.4.3)

$$\hat{\sigma}^2 = \frac{1}{N} \{ sak_1 + \sum_{i=1}^{k} n_i w_i(\hat{\gamma})(\bar{y}_i - \mu^*)^2 \}$$

(5.4.4)

$$\sum_{i=1}^{k} n_i^2 w_i(\hat{\gamma})^2 (\bar{y}_i - \hat{\mu})^2 / W(\hat{\gamma})$$

$$= \{ sak_1 + \sum_{i=1}^{k} n_i w_i(\hat{\gamma})(\bar{y}_i - \hat{\mu})^2 \} / N$$

(5.4.5)

hvor $w_i(\gamma)$ er givet ved (5.3.18) og

$$W(\gamma) = \sum_{i=1}^{k} n_i w_i(\gamma)$$

(5.4.6)
Løsningen til (5.4.3) til (5.4.5) bestemmes ved iteration.

**B**evis:

Logaritmen til likelihoodfunktionen bliver - på nær en additiv konstant -

\[
l(\mu, \sigma^2, \gamma) = -\frac{sak_1}{2\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{k} n_i w_i(\gamma)(\bar{y}_i - \mu)^2
\]

\[= -\frac{N}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{k} \log(w_i(\gamma)) \quad (5.4.7)
\]

Vi ønsker at finde de værdier, \(\hat{\mu}, \hat{\sigma}^2, \hat{\gamma}\), der maximerer \(l\).

Såfremt maksimum findes i et indre punkt, findes disse værdier ved at betragte

\[
\frac{\partial l}{\partial \sigma^2} = \frac{sak_1}{2\sigma^4} + \frac{1}{2\sigma^4} \sum_{i=1}^{k} n_i w_i(\gamma)(\bar{y}_i - \mu)^2 - \frac{N}{2\sigma^2}
\]

\[
\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i w_i(\gamma)(\bar{y}_i - \mu)
\]

\[
\frac{\partial l}{\partial \gamma} = \frac{1}{2\sigma^2} \sum_{i=1}^{k} \left( n_i w_i(\gamma)(\bar{y}_i - \mu) \right)^2 - \frac{1}{2} \sum_{i=1}^{k} n_i w_i(\gamma)
\]

Sættes \(\frac{\partial l}{\partial \mu} = 0\) fås (5.4.3).

Af \(\frac{\partial l}{\partial \sigma^2} = 0\) fås

\[
\sigma^2 = \left[ sak_1 + \sum_{i} n_i w_i(\gamma)(\bar{y}_i - \mu)^2 \right] / N
\]

der netop er (5.4.4)

For at bestemme variansforholdet \(\gamma\) betragter vi

\[
\frac{\partial l}{\partial \gamma} = \frac{1}{2\sigma^2} \sum_{i=1}^{k} \left( n_i w_i(\gamma)(\bar{y}_i - \mu) \right)^2 - \frac{1}{2} \sum_{i=1}^{k} n_i w_i(\gamma)
\]
5.4  Estimation i tilfældig model

Vi undersøger først om maximumspunktet er et indre punkt. Dette undersøges ved at betragte \( \frac{\partial}{\partial \gamma} \) for \( \gamma = 0 \). Hvis \( \left[ \frac{\partial l}{\partial \gamma} \right]_{\gamma=0} < 0 \) findes maximum på randen, d.v.s. for \( \gamma = 0 \). Hvis \( \left[ \frac{\partial l}{\partial \gamma} \right]_{\gamma=0} > 0 \) findes maximum i et indre punkt.

For \( \gamma = 0 \) bliver \( w_i(\gamma) = 0 \), hvorfor \( \frac{\partial l}{\partial \mu} = 0 \) fører til

\[
\hat{\mu} = \bar{y}_. = \frac{\sum_{i=1}^{k} n_i \bar{y}_i}{\sum_{i=1}^{k} n_i}
\]

og \( \frac{\partial l}{\partial \sigma^2} = 0 \) fører til

\[
\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}_.)^2
\]

\[
= \left[ s ak_1 + \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y}_.)^2 \right] / N = (s ak_1 + s ak_2) / N
\]

Man får derfor, at løsningen svarer til \( \frac{\partial l}{\partial \gamma} < 0 \) i løsningspunktet, hvis

\[
\sum_{i=1}^{k} n_i^2 (\bar{y}_i - \bar{y}_.)^2 < s ak_1 + s ak_2
\]

Såfremt dette er opfyldt fås \( \hat{\gamma} = 0 \) og \( \hat{\mu} \) og \( \hat{\sigma}^2 \) bestemmes ved (5.4.3) og (5.4.4).

For \( \sum_i n_i^2 (\bar{y}_i - \bar{y}_.)^2 > s ak_1 + s ak_2 \), bestemmes ML-estimatet \( \hat{\gamma} \) som løsning til \( \frac{\partial l}{\partial \gamma} = 0 \), hvilket netop fører til (5.4.5). \( \square \)

Bemærkning 1  Maksimum-likelihood estimatet \( \hat{\mu} \) er et vejlet gennemsnit af gruppegennemsnittene

Vi ser af (5.4.3), at \( \hat{\mu} \) er et vejlet gennemsnit af gruppegennemsnittene, \( \bar{y}_i \), med de marginale præcissioner

\[
\sigma^2 n_i w_i(\gamma) = \sigma^2 / V [\bar{Y}_i] \]
som vægte. Der gælder nemlig

$$V[\bar{Y}_i] = \sigma_0^2 + \sigma^2/n_i = \frac{\sigma_0^2}{n_i}(1 + n_i\gamma) = \sigma^2/\{n_iw_i(\gamma)\}$$

Hvis eksperimentet er balanceret, d.v.s. $n_1 = n_2 = \cdots = n_k$, bliver alle vægtene ens, og man får det simple resultat, at $\hat{\mu}$ blot er det simple gennemsnit af gruppegennemsnittene.

**Bemærkning 2** Estimatet for $\sigma^2$ udnytter også variationen mellem grupper

Vi bemærker, at estimatet for $\sigma^2$ ikke kun er baseret på variationen indenfor grupper, sak_1, men estimatet udnytter desuden kendskabet til variationen imellem grupper, idet nemlig

$$E[(\bar{Y}_i - \mu)^2] = V[\bar{Y}_i] = \sigma^2/\{n_iw_i(\gamma)\}$$

hvorfor leddene $(\bar{Y}_i - \mu)^2$ indeholder information, såvel om $\sigma^2$, som information om $\gamma$.

**Bemærkning 3** Estimatet for $\sigma_0^2$ er ikke nødvendigvis centralt

Vi bemærker endvidere, at - som vanligt ved ML-estimatet- er estimatet for $\sigma_0^2$ ikke nødvendigvis centralt.

I stedet for maksimum-likelihood estimatet benytter man derfor undtagelserne et estimat baseret på likelihoodfunktionen svarende til fordelingen af residualerne, det såkaldte REML-estimat. Lemma 5.4.2 angiver profillikelihooden for denne likelihoodfunktion.

**Bemærkning 4** I det balancerede tilfælde er momentestimatorne for $\mu$ og $\sigma^2$ de samme som maksimum-likelihoodestimatorne.

I det balancerede tilfælde, $n_1 = n_2 = \cdots = n_k$ afhænger vægtene

$$w_i(\gamma) = \frac{1}{1 + n\gamma}$$

ikke af $i$, og (5.4.3) bliver

$$\hat{\mu} = \sum_{i=1}^{k} \bar{Y}_{i+}/k = \bar{Y}_{++},$$
alså netop momentestimatet.

Hvis \((n - 1)sak_2 > sak_1\) svarer maksimum-likelihood estimatoren til et indre punkt i parameterrummet.

Ligningerne (5.4.4) og (5.4.5) til bestemmelse af \(\sigma^2\) og \(\gamma = \sigma_0^2 / \sigma^2\) bliver

\[
N \sigma^2 = sak_1 + \frac{1}{1 + n\gamma} sak_2 \\
N \frac{n}{1 + n\gamma} \frac{sak_2}{k} = sak_1 + \frac{1}{1 + n\gamma} sak_2
\]

med løsningen

\[
\hat{\sigma}^2 = \frac{sak_1}{N - k} \\
\hat{\gamma} = \frac{1}{n} \left( \frac{sak_2}{k\hat{\sigma}^2} - 1 \right)
\]

dvs.

\[
\sigma_0^2 = \frac{sak_2/k - \hat{\sigma}^2}{n}
\]

Ved sammenligning med (5.3.14) ser man at maksimum-likelihood estimatet for \(\sigma^2\) er den samme som momentestimatet, men maksimum-likelihood estimatet for \(\sigma_0^2\) er systematisk mindre end det centrale momentestimat. Maksimum-likelihood estimatet tilgodeser ikke at der kun er \(k - 1\) frihedsgrader for \(sak_2\).

Hvis \((n - 1)sak_2 < sak_1\), bliver også maksimum-likelihood estimatet for \(\sigma^2\) forskelligt fra momentestimatet.

\[\square\]

**Bemærkning 5 Startværdi for iterationen**

En god startværdi for iterationen fås ved at tage udgangspunkt i den sædvanlige \(F\)-teststørrelse for homogenitet imellem grupper (5.2.15)

\[
z = \frac{sak_2/(k - 1)}{sak_1/(N - k)} \approx 1 + n\gamma,
\]

d.v.s. at et hurtigt og nemt bud på \(\mu\) fås ved at benytte (5.4.3) med

\[
\gamma = \frac{1}{n} \left\{ \frac{sak_2/(k - 1)}{sak_1/(N - k)} - 1 \right\}.
\]

\[\square\]
Lemma 5.4.1 Profilloglikelihood’en med hensyn til middelværdien μ

Under modellen givet ved (5.3.1) og (5.3.2) er profil-loglikelihoodfunktionen med hensyn til μ bestemt ved

$$\tilde{l}_y(\sigma^2, \gamma) = -\frac{1}{2} \left[ N \ln(\sigma^2) + \sum_{i=1}^{k} \ln(1 + n_i \gamma) + \frac{\mathbf{r}^T(\mathbf{V}^*)(\gamma)^{-1}\mathbf{r}}{\sigma^2} + N \ln(2\pi) \right] ,$$

hvor

$$\mathbf{V}^* = \text{Blok diag}\{\mathbf{I}_{n_i} + \gamma \mathbf{J}_{n_i}\} \quad (5.4.8)$$

$$\mathbf{r} = \mathbf{y} - \hat{\mu}(\gamma) \mathbf{1} \quad (5.4.9)$$

$$\hat{\mu}(\gamma) = \sum_{i=1}^{k} n_i w_i(\gamma) \bar{y}_i / \sum_{i=1}^{k} n_i w_i(\gamma) \quad (5.4.10)$$

med

$$w_i(\gamma) = \frac{1}{1 + n_i \gamma}$$

Bevis:

Følger ved at bemærke, at modellen for den N-dimensionale observationsvektor y er $y \in \mathbb{N}(\mu \mathbf{1}, \sigma^2 \mathbf{V}^*(\gamma))$, hvor

$$\mathbf{V}^*(\gamma) = \text{Blok diag}\{\mathbf{I}_{n_i} + \gamma \mathbf{J}_{n_i}\} , \quad (5.4.11)$$

jvf (5.3.11).

For fastholdt γ minimeres kvadratafvigelsessummen

$$(\mathbf{y} - \mu \mathbf{1})^T[\mathbf{V}^*(\gamma)]^{-1}(\mathbf{y} - \mu \mathbf{1})$$

jvf Sætning 0.5.3 i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998 for

$$\hat{\mu}(\gamma) = \left[\mathbf{1}^T(\mathbf{V}^*(\gamma))^{-1}\mathbf{1}\right]^{-1} \mathbf{1}^T(\mathbf{V}^*(\gamma))^{-1}\mathbf{y} ,$$

der netop er (5.4.10).

Idet matricen

$$\mathbf{V}_i^* = \text{Blok diag}\{\mathbf{I}_{n_i}\}$$
har egenværdien 1 med multipliciteten $n_i - 1$ og egenværdien $1 + n_i \gamma$ med multipliciteten 1, har man

$$\det(V^*) = \prod_{i=1}^{k} \det(V_i^*) = \prod_{i=1}^{k} (1 + n_i \gamma),$$

hvoraf resultatet følger.

**Lemma 5.4.2** Profilloglikelihood'en svarende til fordelingen af residualerne

Under modellen givet ved (5.3.1) og (5.3.2) er profil-loglikelihoodfunktionen med hensyn til $\mu$, svarende til fordelingen af residualerne bestemt ved

$$\tilde{l}_r(\sigma^2, \gamma) = -\frac{1}{2} \left[ (N - 1) \ln(\sigma^2) + \sum_{i=1}^{k} \ln(1 + n_i \gamma) + \ln \left( \sum_{i=1}^{k} n_i w_i(\gamma) \right) \right]
+ \frac{r^T (V^*)^{-1} r}{\sigma^2} + N \ln(2\pi),$$

(5.4.12)

hvor $V^*$ og $r$ er givet ved (5.4.8) og (5.4.9).

Estimatet $\tilde{\mu}, \tilde{\sigma}^2$ og $\tilde{\gamma}$, der giver anledning til maksimum af (5.4.12) kaldes REML-estimatet (Residual-maksimum likelihood).

Undertiden ses også betegnelsen "Restricted maksimum likelihood" benyttet.

Vi bemærker, at fordelingen af residualerne er en $N - 1$-dimensional fordeling.

**Bevis:**

**Sætning 5.4.2** Approximativ varians for maksimum-likelihood estimatorerne under den tilfældige model
Den asymptotiske varsians for maximaliseringsestimatorerne er

\[ \text{V}[\sigma^2] = 2\sigma^4 \sum_{i=1}^{k} (n_i w_i(\gamma))^2 / D \]

\[ \text{V}[\sigma_0^2] = 2\sigma^4 (N - k + \sum_{i=1}^{k} n_i w_i(\gamma)^2) / D \]

og

\[ \text{COV}[\sigma^2, \sigma_0^2] = -2\sigma^2 (\sum_{i=1}^{k} n_i w_i(\gamma))^2 / D \]

hvor

\[ D = N \sum_{i=1}^{k} (n_i w_i(\gamma))^2 - (\sum_{i=1}^{k} n_i w_i(\gamma))^2 \]

og hvor \( w_i(\gamma) \) er givet ved (5.3.18).

**Bevis:**
Resultatet følger ved at betragte informationsmatricen (den forventede krumning af likelihoodfunktionen i maximumumspunktet) \( \square \)

**Eksempel 5.4.1 Varianskvotienten \( \gamma \)'s betydning for vægtningen af gruppegennemsnittene**

For at belyse effekten af varianskvotienten \( \gamma \) ved vægtningen af observationerne i den ubalancerede situation betragter vi først grænsetilfældene \( \gamma = 0 \) og \( \gamma = \infty \)

Såfremt \( \gamma = 0 \) finder man \( w_i(\gamma) = 1 \), hvorfør man får det sædvanlige estimat:

\[ \mu_{\gamma=0}^* = \sum_{i=1}^{k} n_i \bar{y}_i / \sum_{i=1}^{k} n_i \]

d.v.s. alle \( N \) enkeltobservationer vægtes ens.

Jo større værdier af \( \gamma \) (d.v.s. jo større variation der er mellem gruppemiddelværdierne i forhold til \( \sigma^2 \)), desto mere nærmer \( \gamma \) sig til \( \infty \) og estimatet nærmer sig til

\[ \mu_{\gamma=\infty}^* = \sum_{i=1}^{k} \bar{y}_i / k \]
d.v.s. det simple gennemsnit af gruppegennemsnittene uden hensyn til de enkelte gruppestørrelser.

5.5 **SAS® procedurer til analyse af den tilfældige model**

Programsystemet SAS® indeholder forskellige procedurer, der blandt andet kan benyttes til analyse af den tilfældige model. I det følgende skal vi kort illustrere brugen af disse procedurer.

Vi vil betragte såvel situationer med balancede data (lige mange observationer i alle grupper), som situationer med ubalancede data.

Vi vil benytte data fra eksempel 5.3.1 til at illustrere brugen af de forskellige procedurer. Vi vil antage, at data er indlæst i de variable `renh`, `balle` og `prove`.

For at illustrere beregningerne i tilfældet med ubalancerede data, benytter vi de samme data, blot antager vi, at de to sidste prøver fra hver af de tre første baller mangler, dvs. de observationer, der er markeret med • i tabel 5.1.

5.5.1 **GLM**

Proceduren er i det væsentlige rettet mod analyse af såkaldte "generelle lineære modeller" for normalfordelte data, dvs. normalfordelingsmodeller, hvor middelværdien kan beskrives som en lineær funktion af de forklarende variable.

Proceduren kan derfor specielt benyttes til at udføre en ensidet variansanalyse i den systematiske model.

Da momentestimaterne svarende til den tilfældige model kan bestemmes ud fra variansanalyseskemaet, kan proceduren derfor også benyttes til momentestimation i den tilfældige model. Da endvidere testet for betydning af den pågældende variable er det samme i den tilfældige model, som i den systematiske model, kan proceduren derfor også benyttes til bestemmelse af dette test.
Eksempel 5.5.1 Balancedede data, beregning ved SAS®-proceduren GLM

Betragt situationen i eksempel 5.3.1 og antag at data er indlæst i de variable renh, balle og prove, i alt 28 observationer. SAS® programmet

PROC GLM;
CLASS balle prove;
MODEL renh = balle;
RANDOM balle;
RUN;

definerer de variable balle og prove som klassifikationsvariable, og angiver modelformlen svarende til en balleeffekt:

MODEL renh = balle

Endelig i sætningen

RANDOM balle;

erklæres balleeffekten som tilfældig.

Programmet resulterer i følgende udskrift:

General Linear Models Procedure
Class Level Information

Class Levels Values
BALLE 7 1 2 3 4 5 6 7
PROVE 4 1 2 3 4

Number of observations in data set = 28
### 5.5 SAS® procedurer til analyse af den tilfældige model

**General Linear Models Procedure**

Dependent Variable: RENH

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>65.96264286</td>
<td>1.76</td>
<td>0.1573</td>
</tr>
<tr>
<td>Error</td>
<td>21</td>
<td>131.47220000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>27</td>
<td>197.43484286</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square: 0.334098  C.V.: 4.311284  RENH Mean: 58.0364286

**General Linear Models Procedure**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALLE</td>
<td>6</td>
<td>65.96264286</td>
<td>1.76</td>
<td>0.1573</td>
</tr>
</tbody>
</table>

Man genfindes det sædvanlige variantsanalyseskema i udskriftens p.2. (Da der kun optræder én forklarende variabel, er der ingen forskel på type I og type III kvadratafvigelser).

Rubrikkens Model refererer til kvadratafvigelsessummen $sak_2$ svarende til den betragtede model (som her kun indeholder faktoren Balle). Rubrikkens Error refererer til variationen indenfor baller ($sak_1$).

Under overskriften RENH Mean er angivet gennemsnittet af renhedsmålingerne.

Testet for hypotesen $\sigma^2_0 = 0$ er det samme som testet for balleeffekt i den systematiske model. Denne teststørrelse er anført såvel i variantsanalyseskemaet, som på en separat linie længere nede i udskriften.

Alle disse størrelser ville også blive udskrevet, selv om man ikke havde angivet specifikationen RANDOM balle.

Udskriftenes side 3 er et resultat af denne specifikation. Udskriften viser, at den gennemsnitlige variation hidrørende fra baller ($sak_2/(k - 1)$) har
forventningsværdien

\[ E \left[ \frac{S \bar{A} \bar{K}_2}{(k - 1)} \right] = \sigma^2 + 4\sigma_0^2 \]

(jvf beviset for sætning 5.3.3).

Herved kan \( \sigma_0^2 \) estimeres som i eksempel 5.3.1.

**Eksempel 5.5.2 Ubalancerede data, beregning ved SAS-proceduren GLM**

For sættet af ubalancerede data vil programmet

```
PROC GLM;
CLASS balle prove;
MODEL renh = balle;
RANDOM balle;
RUN;
```

give udskriften:

```
General Linear Models Procedure

Dependent Variable: RENH

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>72.55028636</td>
<td>4.36</td>
<td>0.0096</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>41.57370000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>21</td>
<td>114.12396366</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square    C.V.    RENH Mean
0.635715    2.862840 58.1522727

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALLE</td>
<td>6</td>
<td>72.55028636</td>
<td>4.36</td>
<td>0.0096</td>
</tr>
</tbody>
</table>

General Linear Models Procedure
```

Udskriftenes struktur adskiller sig ikke fra udskriften fra de balancerede data.
5.5 SAS® procedurer til analyse af den tilfældige model

Vi bemærker specielt, at udskriftens side 3 viser, at den anførte gennemsnitlige variation hidrørende fra baller \((sak_2/(k-1))\) har forventningsværdien

\[
E[SAX/(k-1)] = \sigma^2 + 3.0903\sigma_0^2
\]

(jvf beviset for sætning 5.3.3), dvs \(n_0 = 3.0903\).

Idet \(\hat{\sigma}^2 = 41.5737/15 = 2.77158\) får man derfor momentestimatet

\[
\hat{\sigma}_0^2 = (12.091714 - 2.771580)/3.0909 = 3.0153
\]

\[\Box\]

### 5.5.2 Mixed

Proceduren er direkte rettet mod analyse af "blandede normalfordelingsmodeller", dvs lineære modeller, der både indeholder systematiske og tilfældige komponenter.

Proceduren kan derfor specielt benyttes til at udføre en ensidet variantanalyse i den tilfældige model.

Proceduren giver mulighed for at vælge mellem maksimum-likelihood estimatorne (Sætning 5.4.1) og estimatorer bestemt ved maksimering af likehoodfunktionen svarende til residualerne, (REML)-estimaterne, lemma 5.4.2

**Eksempel 5.5.3** Balanceerede data, maksimum-likelihood estimation ved SAS®-proceduren MIXED

Programmet:

```sas
PROC MIXED METHOD=ML ASYCOV ;
CLASS balle prove;
MODEL renh = ;
RANDOM balle ;
RUN;
```

kalder procedure MIXED. Nøgleordet METHOD =ML angiver, at man ønsker maksimum-likelihood estimaterne, og ordet ASYCOV angiver, at man ønsker den asymptotiske varians-kovariansmatrix for estimaterne.

I procedure MIXED specificeres kun de systematiske (såkaldte "fixed") effekter i modelformlen. Det eneste systematiske led i den tilfældige model for
den ensidede variansanalyse er interceptleddet, som altid er underforstået i modelspecifikationen. Modelformlen bliver derfor blot MODEL renh=, der angiver, at den afhængige variable er renh.

Sætningen RANDOM balle angiver de tilfældige komponenter i modellen. Her altså blot den variable balle.

Proceduren giver anledning til følgende udskrift:

The MIXED Procedure

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALLE</td>
<td>7</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>PROVE</td>
<td>4</td>
<td>1 2 3 4</td>
</tr>
</tbody>
</table>

til kontrol af de indlæste specifikationer.

Endvidere udkrives iterationsforløbet:

The MIXED Procedure

ML Estimation Iteration History

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>Objective</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>82.68971508</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>82.22198159</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Convergence criteria met.

Da data er balancerede, behøves kun én iteration.

Derefter udkrives estimatorne for $\sigma^2$ og $\sigma_0^2$:

Covariance Parameter Estimates (MLE)

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
</table>
De to varianser $\sigma^2$ og $\sigma_0^2$ kaldes kovariansparametre, da modellen jo er karakteriseret ved dispersionsmatricen, $V$, (5.3.11)

Udskriftens første søjle Ratio angiver det estimerede forhold $\hat{\gamma}$ mellem variansen svarende til den tilfældige effekt og residualvariansen.

I søjlen Estimate angives estimererne $\hat{\sigma}_0^2$ (svarende til BALLE) og $\hat{\sigma}^2$ (svarende til residualvariansen). Søjlen Std Error angiver den estimerede spredning, hhv $\sqrt{V[\hat{\sigma}_0^2]}$ og $\sqrt{V[\hat{\sigma}^2]}$.

Søjlen Z angiver

$$\frac{\hat{\sigma}_0^2}{\sqrt{V[\hat{\sigma}_0^2]}}$$

god den tilsvarende størrelse for residualvariansen.

Såfremt den sande varians er nul, vil $Z$ approximativt følge en standardiseret normalfordeling, og man kan derfor bruge $Z$ til et approximativt test for en hypotese $\sigma_0^2 = 0$, og evt. også en hypotese $\sigma^2 = 0$.

Størrelsen $Pr > |Z|$ angiver netop testsandsynligheden svarende til dette test. Hypotesen forkastes for små værdier af $Pr > |Z|$.

Som en konsekvens af ordren ASYCOV udskrives:

**Asymptotic Covariance Matrix of Estimates**

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Row</th>
<th>COVP1</th>
<th>COVP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALLE</td>
<td>1</td>
<td>1.81896982</td>
<td>-0.93321128</td>
</tr>
<tr>
<td>Residual</td>
<td>2</td>
<td>-0.93321128</td>
<td>3.73284513</td>
</tr>
</tbody>
</table>

Estimatet er den inverse observerede informationsmatrix. Den inverse observerede informationsmatrix beregnes som 2 gange den inverse Hessian-matrix i maksimumspunktet.

Endelig udskrives et resume af tilpasningen,

The MIXED Procedure
Model Fitting Information for RENH

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observations</td>
<td>28.0000</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-66.8413</td>
</tr>
<tr>
<td>Akaike’s Information Criterion</td>
<td>-68.8413</td>
</tr>
<tr>
<td>Schwarz’s Bayesian Criterion</td>
<td>-70.1735</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>133.6825</td>
</tr>
</tbody>
</table>

Vi noterer således, at værdien af log likelihood, og tilsvarende af -2 log likelihood udskrives. Disse værdier kan eventuelt bruges i mere komplicerede situationer, hvor man har en række hierarkisk organiserede hypoteser.

**Eksempel 5.5.4 Balancede data, REML estimation ved SAS-proceduren MIXED**

Såfremt vi i stedet havde benyttet residual-likelihood funktionen til estimationen ved procedurekaldet

```
PROC MIXED METHOD=REML ASYCOV ;
CLASS balle prove;
MODEL renh = ;
RANDOM balle ;
RUN;
```

havde vi fået estimaterne

**Covariance Parameter Estimates (REML)**

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALLE</td>
<td>1.18329821</td>
</tr>
<tr>
<td>Residual</td>
<td>6.26058095</td>
</tr>
</tbody>
</table>

med de tilsvarende varians-kovarianseestimater:

**Asymptotic Covariance Matrix of Estimates**

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Row</th>
<th>COVP1</th>
<th>COVP2</th>
</tr>
</thead>
</table>
5.5 SAS® procedurer til analyse af den tilfældige model

\[
\begin{array}{ccc}
\text{BALLE} & 1 & 2.75128329 & -0.93321128 \\
\text{Residual} & 2 & -0.93321128 & 3.73284513
\end{array}
\]

Vi ser, at REML-estimatet for \( \sigma^2 \) er det samme, som ML-estimatet, men REML-estimatet for \( \sigma_0^2 \) er større end ML-estimatet, svarende til at ML-estimatet ikke korrigerer for frihedsgraderne i estimationen af \( \sigma_0^2 \).

Eksempel 5.5.5 Ubalancerede data, maksimum-likelihood estimation ved SAS-proceduren MIXED

For sættet af ubalancerede data vil programmet

```
PROC MIXED METHOD=ML ASYCOV;
CLASS balle prove;
MODEL renh = ;
RANDOM balle;
RUN;
```

give estimaterne

\[
\begin{array}{cc}
\text{Covariance Parameter Estimates (MLE)} \\
\text{Cov Parm} & \text{Estimate} \\
\text{BALLE} & 2.62615356 \\
\text{Residual} & 2.80080383
\end{array}
\]

Sammenigner vi med momentestimatet i eksempel 5.5.2 ser vi, at såvel estimatet for \( \sigma_0^2 \) som for \( \sigma^2 \) adskiller sig fra momentestimatet.

Eksempel 5.5.6 Ubalancerede data, REML-estimation ved SAS-proceduren MIXED

Vi anfører endelig REML-estimaterne svarende til den ubalancerede situation.

\[
\begin{array}{cc}
\text{Covariance Parameter Estimates (REML)} \\
\text{Cov Parm} & \text{Estimate} \\
\text{BALLE} & \text{(REML estimate)} \\
\text{Residual} & \text{(REML estimate)}
\end{array}
\]
<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>BALLE</td>
<td>3.26121525</td>
</tr>
<tr>
<td>Residual</td>
<td>2.79158363</td>
</tr>
</tbody>
</table>

Vi ser, at estimatet for $\sigma^2_0$ som ventet er større end det tilsvarende maksimum-likelihood estimat, og desuden bemærker vi, at der er en lille forskel på estimatorne for $\sigma^2$.

5.5.3 Varcomp

Vi anfører endelig, at SAS-systemet desuden indeholder en procedure, VARCOMP, der er rettet mod analyse af modeller for varianskomponenter i normalfordelingsammenhænge.

Da proceduren MIXED i det store og hele er mere generel end VARCOMP, vil vi ikke her gå nærmere ind på proceduren VARCOMP.

Vi skal blot nævne, at ved brug af VARCOMP skal alle effekter, såvel tilfældige, som systematiske, anføres i modelformlen.

Et program svarende til eksempel 5.5.3 men med brug af proceduren VARCOMP ville derfor have formen:

```plaintext
PROC VARCOMP METHOD=ML;
CLASS balle prove;
MODEL renh = balle;
RANDOM balle;
```

5.6 Eksempler på den tilfældige model

Eksempel 5.6.1 Repeterbarhed og reproducerbarhed for en prøvningsmetode

De usikkerhedsmål, man oftest benytter til beskrivelse af prøvningsmetoders nøjagtighed, knytter sig til de omstændigheder, hvorunder prøvningen tænkes gentaget. Begreberne er blandt andet beskrevet i ASTM E 177 Use of
5.6 Eksempler på den tilfældige model

the terms precision and bias in ASTM test methods, i ASTM E 456 Terminology for statistical methods, I ISO 3534-1 Terminology og i ISO 5725-1 Determination of repeatability and reproduceability.

Usikkerheds målne deles sædvanligvis op i to hovedkategorier, nemlig usikkerhed knyttet til repeterbarhedsbetingelser, og usikkerhed knyttet til reproducerbarhedsbetingelser. De to hovedkategorier afgrænses som anført nedenfor.

Repeterbarhedsbetingelser: Betingelser, hvorunder indbyrdes uafhængige prøvningsresultater er opnået med den samme metode

- i) på identisk prøvemateriale
- ii) i det samme laboratorium
- iii) af den samme operatør
- iv) under benyttelse af samme udstyr
- v) indenfor et kort tidsinterval

Reproducerbarhedsbetingelser: Betingelser, hvorunder prøvningsresultater er opnået med den samme metode

- i) på identisk prøvemateriale
- ii) på forskellige laboratorier
- iii) med forskellige operatører
- iv) under benyttelse af forskelligt udstyr

De kvantitative størrelser, der benyttes til beskrivelse af nøjagtigheden er sædvanligvis repeterbarhedsvarians og reproducerbarhedsvarians. Disse størrelser kan bestemmes ved udsendelse af ens prøver til en række laboratorier (n prøver til det i'te laboratorium). Prøvningsresultaterne $Y_{ij}$ modelleres da ved (5.3.1), hvor $\mu$ angiver den vedtagne sande værdi plus en eventuel bias; $\sigma^2$ angiver repeterbarhedsvariansen, og $\sigma^2 + \sigma_0^2$ angiver reproducerbarhedsvariansen.

ISO 5725-3 beskriver en række mellemliggende mål for repeterbarheden.
Nedenstående tabel viser data fra en ringprøvning, hvor prøver fra et homogent materiale (papir fra en bestemt produktion) blev udsendt til hvert af 9 laboratorier med henblik på bestemmelse af papirets reflektans. Alle laboratorier benyttede samme prøvningsteknik. Hvert laboratorium udførte 4 gentagne prøvninger på prøvningsmaterialet.

Tabellen angiver resultaterne af reflektansmålinger af et homogent prøvningsmateriale udsendt til 9 laboratorier. Resultaterne er angivet som % reflektans ved 440 [nm].
(Data fra ASTM-E 691)

<table>
<thead>
<tr>
<th>i</th>
<th>Laboratory 1</th>
<th>Laboratory 2</th>
<th>Laboratory 3</th>
<th>Laboratory 4</th>
<th>Laboratory 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>97.4</td>
<td>92.6</td>
<td>96.2</td>
<td>95.2</td>
<td>93.0</td>
</tr>
<tr>
<td></td>
<td>97.8</td>
<td>93.0</td>
<td>96.8</td>
<td>95.5</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>98.8</td>
<td>92.5</td>
<td>96.8</td>
<td>95.7</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>98.0</td>
<td>92.7</td>
<td>96.2</td>
<td>95.7</td>
<td>95.1</td>
</tr>
<tr>
<td>$\bar{y}_i$</td>
<td>98.000</td>
<td>92.700</td>
<td>96.500</td>
<td>95.525</td>
<td>94.425</td>
</tr>
<tr>
<td>$s_i$</td>
<td>0.589</td>
<td>0.216</td>
<td>0.346</td>
<td>0.236</td>
<td>0.974</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>i</th>
<th>Laboratory 6</th>
<th>Laboratory 7</th>
<th>Laboratory 8</th>
<th>Laboratory 9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>94.8</td>
<td>96.3</td>
<td>99.8</td>
<td>99.5</td>
</tr>
<tr>
<td></td>
<td>95.2</td>
<td>96.3</td>
<td>100.0</td>
<td>99.3</td>
</tr>
<tr>
<td></td>
<td>95.0</td>
<td>96.1</td>
<td>99.9</td>
<td>99.8</td>
</tr>
<tr>
<td></td>
<td>94.9</td>
<td>96.3</td>
<td>99.7</td>
<td>99.7</td>
</tr>
<tr>
<td>$\bar{y}_i$</td>
<td>94.975</td>
<td>96.250</td>
<td>99.850</td>
<td>99.575</td>
</tr>
<tr>
<td>$s_i$</td>
<td>0.171</td>
<td>0.100</td>
<td>0.129</td>
<td>0.222</td>
</tr>
</tbody>
</table>

Rækken $\bar{y}_i$, angiver gennemsnittet af de fire prøvningsresultater på det $i$'te laboratorium og $s_i$ angiver tilsvarende den empiriske spredning på dette laboratoriums resultater, $s_i^2 = \sum_j (y_{ij} - \bar{y}_i)^2 / 3$.

Idet vi benytter modellen (5.3.1) og (5.3.2), hvor $\mu$ angiver den ved prøvningsmetoden bestemte værdi af prøvningsmaterialets reflektans, $\sigma^2$ angiver metodens repeterbarhedsvarians, og $\sigma^0_2$ angiver variansen mellem laboratorier, finder vi variansanalyseskemaet:

<table>
<thead>
<tr>
<th>Variation</th>
<th>SAK</th>
<th>$f$</th>
<th>$s^2 = SAK/f$</th>
<th>$E[S^2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mellem laboratorier</td>
<td>179.8323</td>
<td>8</td>
<td>22.4790</td>
<td>$\sigma^2 + 4\sigma^0_2$</td>
</tr>
<tr>
<td>Indenfor laboratorier</td>
<td>4.8700</td>
<td>27</td>
<td>0.1804</td>
<td>$\sigma^2$</td>
</tr>
</tbody>
</table>
5.6 Eksempler på den tilfældige model

Man finder således estimatet $\tilde{\sigma}^2 = 0.1804 = (0.42 [%\text{reflektans}])^2$ for re- 

peterbarhedsvariansen, og $\tilde{\sigma}_0^2 = 5.5747 = (2.36 [%\text{reflektans}])^2$ for vari- 

ansen mellem laboratorier. Reproducerbarhedsvariansen, dvs. usikkerhe- 

den $\sigma_0^2 + \sigma^2$ på en reflektansbestemmelse estimeres da som $\tilde{\sigma}_0^2 + \tilde{\sigma}^2 = 

5.5747 + 0.1804 = 5.7551 = (2.39 [%\text{reflektans}])^2$.

Usikkerheden på gennemsnittet af $m$ gentagne prøvninger samme labora- 

torium kan således udtrykkes ved variantestimatet $V[\overline{Y}] \approx \tilde{\sigma}_0^2 + \tilde{\sigma}^2/m$.

Selv ved vilkårligt mange gentagne prøvninger på samme laboratorium vil 

usikkerheden på gennemsnittet af gentagne prøvninger ikke kunne blive 

mindre end spredningen mellem laboratorier, $\sigma \approx 2.36 [%\text{reflektans}]$.

Eksempel 5.6.2 Optimal allokering af stikprøveindsats

En produktion af bulkvarer leveres i partier bestående af et stort antal 

sække.

Modtageren af et parti ønsker ved en stikprøvekontrol at vurdere, hvor-

vidt kvaliteten er tilfredsstillende. Der udtales derfor tilfældigt $m$ sække, 

hvorefter der udtales $n$ prøver tilfældigt fra indholdet af hver sæk og kva-

litetsegenskaben for den pågældende prøve bestemmes ved en laboratorie-

undersøgelse.

Lad $\overline{Y}_i$, betegne det gennemsnitlige indhold af de $n$ prøver fra den $i$'te sæk, 

og lad tilsvarende $\overline{Y}_j$ betegne det totale gennemsnit af de ialt $m$ $n$ prøver.

Antag, at variansen på prøvningsresultatet ved gentagne prøver fra samme 

sæk er $\sigma^2$, og antag endvidere, at sækkens middelkvalitet varierer fra sæk 

til sæk med variansen $\gamma \sigma^2$.

Man finder da variansen på gennemsnittet fra den $i$'te sæk,

$$V[\overline{Y}_i.] = \sigma^2(\gamma + 1/n)$$

og da prøverne fra de $m$ sække er uafhængige, finder man variansen på det 

totale gennemsnit

$$V[\overline{Y}_i.] = \frac{\sigma^2}{m}(\gamma + 1/n)$$

Såfremt omkostningerne ved udtagning og laboratorieundersøgelse af en 

prøve er $k$ omkostningsenheder, mens omkostningerne ved udvælgelse af
en sæk og forberedelse af sækken til prøvning er \( \kappa k \), finder man, at de samlede omkostninger ved udvælgelse og analyse er

\[
K(m, n) = m \times \kappa \times k + m \times n \times k = (\kappa + n) \times m \times k
\]

For et givet budget, \( K(m, n) = C \times k \), finder man da, at den optimale allokering af stikprøveressuourcerne, dvs. den allokering, der giver den mindste varians, bestemmes ved

\[
n \simeq \sqrt{\kappa / \gamma} \quad m = \frac{C}{\kappa + n}
\]

Antallet af laboratorieprøver pr sæk bestemmes således af forholdet \( \kappa \) mellem omkostningen ved udvælgelse af en sæk og omkostningen ved undersøgelse af en prøve fra sækken samt af forholdet \( \gamma \) mellem variansen imellem sækkene og variansen indenfor en sæk.

Jo større værdi af \( \kappa \), desto flere prøver vil man udtae pr. sæk; jo større værdi af \( \gamma \), desto færre prøver vil man udtae pr. sæk.

\[\square\]

**Eksempel 5.6.3 Stratifikation**

Ved stikprøveundersøgelser af partier, der er opdelt i naturlige enheder benyttes ofte følgende fremgangsmåde: Der udtages tilfældigt \( k \) enheder fra partiet. Hver enhed udsettes for \( n \) uafhængige prøvninger (oftest er \( n = 1 \)), og det totale gennemsnit \( \overline{Y}. = \sum_i \sum_j Y_{ij} / (nk) \) udregnes. Ved sammenligning af dette fundne gennemsnit med en specifiseret værdi \( \mu_0 \) for middelkvaliteten benyttes variansen

\[
V[\overline{Y}.] = \sigma_2^2 / k + \sigma_3^2 / (nk),
\]

hvor \( \sigma_2^2 \) angiver variansen mellem kvalitetsmålet for de enkelte enheder og \( \sigma_3^2 \) angiver variansen for gentagne prøvninger på samme enhed.

Såfremt prøvningerne er udført under reproducerbarhedsbetingelser benyttes \( \sigma_3^2 = \sigma_0^2 + \sigma^2 \), i modsat fald vil det være rigtigst at benytte \( \sigma_3^2 = \sigma^2 \) (reproducerbarhedsvariansen), og at addere reproducerbarhedsbidraget \( \sigma_0^2 \) til udtrykket for \( V[\overline{Y}.] \).

\[\square\]

**Eksempel 5.6.4 Bulk sampling**
5.6 Eksempler på den tilfældige model

Ved stikprøveundersøgelse af bulkprodukter (f. eks. jernindhold i malm, fugtighedsprocent af korn etc.) benyttes ofte en variant af følgende fremgangsmåde:

Der udtages tilfældigt \( k \) prøver (inkremente) af samme størrelse fra partiet. Der foretages en fysisk gennemsnitsdannelse af disse inkremente ved at prøverne blandes til en basisbunke (gross sample), hvorefter der udtages en laboratorieprøve fra denne bunke (oftest ved neddeling eller ved benyttelse af et specielt instrument). Fra laboratorieprøven udtages nu \( m \) analyseprøver af en størrelse og konsistens, der tillader analyse. Som skøn over partiets middelindhold benyttes da gennemsnittet \( \bar{X} = \frac{\sum_{\nu} X_{\nu}}{m} \) af de \( m \) analyseresultater.

Til beskrivelse af usikkerheden på denne størrelse benyttes følgende model:

Middelindholdet \( Y_i \) i et tilfældigt udtaget inkrement antages at være

\[
Y_i = \mu + \alpha_i \quad i = 1, 2, \ldots, k
\]

, hvor \( \mu \) angiver partiets middelindhold og \( \alpha_i \) antages at kunne beskrives som uafhængige \( N(0, \sigma^2_{\alpha}) \)-fordelte variable. (Den tilfældige udvælgelse af inkremente tjener til at sikre, at eventuelle systematiske variationer af \( \alpha_i \) elimineres.)

Middelindholdet \( \bar{Y} \) i basisbunken er da

\[
\bar{Y} = \mu + \bar{\alpha},
\]

hvor \( \bar{\alpha} \in N(0, \sigma^2_{\alpha}/k) \), og middelindholdet \( Z \) i laboratorieprøven bliver

\[
Z = \bar{Y} + B,
\]

hvor \( B \) antages at være uafhængig af \( \alpha_i \) og \( B \in N(0, \sigma^2_B) \). Variansbidraget \( \sigma^2_B \) beskriver variationen hidrørende fra forberedelsen af laboratorieprøven.

Endelig antages analyseresultatet \( X_{\nu} \) at kunne beskrives som

\[
X_{\nu} = Z + \epsilon_{\nu}, \quad \nu = 1, 2, \ldots, m
\]

hvor \( \epsilon_{\nu} \) antages indbyrdes uafhængige og uafhængige af \( \alpha_i \) og \( B \), og hvor \( \epsilon_{\nu} \in N(0, \sigma^2_3) \). Variansbidraget \( \sigma^2_3 \) beskriver variationen hidrørende fra forberedelse og prøvning af analyseprøven.
Under disse antagelser finder man endelig usikkerhedsvariansen på estima-
tet $\bar{X}$. (variansen svarende til gentagelser af hele proceduren):

$$\text{V}[\bar{X}] = \sigma^2_0/k + \sigma^2_B + \sigma^2_3/m.$$ 

Bestemmelsen af de indgående varianser må foregå ved planlagte forsøg, 
der inddrager flere basisbunker etc., og deres størrelse må løbende estime-
res med henblik på en overvågning af usikkerheden. Gy (1992) har formu-
leret en sammenhængende teori for bulksampling. Dele af teorien bygger 
imidlertid på nogle antagelser om partikelformen, og teorien anses derfor 
of nogle brugere for at være kontroversiel.

5.7 **Normalfordelingsmodeller med tilfældigt varierende varians.**

I nogle situationer er modellen med tilfældigt varierende middelværdier 
(varianskomponentmodellen) ikke tilstrækkelig til at beskrive variationen i 
data. Vi skal derfor her angive en udvidelse af varianskomponentmodellen, 
der muliggør beskrivelse af observationer, hvor variationen inden for den 
enkelt gruppe varierer tilfældigt fra gruppe til gruppe.

Modellen bygger på antagelsen

$$Y_{ij}|(\mu_i, \sigma^2_i) \in \text{N}(\mu_i, \sigma^2_i) \quad (5.7.1)$$

hvor

$$\mu_i|\sigma^2_i \in \text{N}(\mu_0, \sigma^2_i/m), \quad \text{og} \quad (5.7.2)$$

$$\sigma^2_i \in \text{RGam}(\alpha, \beta)$$

og hvor $Y_{ij}$ er betinget uafhængige (se definition 4.5.1 side 467) for givet 
$(\mu_i, \sigma^2_i)$, og endvidere er $\sigma^2_i$ indbyrdes uafhængige.

Antagelsen indebærer en udvidelse i forhold til den tilfældige model, der 
blev beskrevet i afsnitene 5.3 til 5.6.
5.7 Normalfordelingsmodeller med tilfældig varians

Den væsentligste udvidelse består i antagelsen om at gruppevarianser, $\sigma_i^2$ varierer tilfældigt fra gruppe til gruppe. Vi har valgt at antage, at variansen følger en såkaldt reciprok gammafordeling, RGam-fordelingen. Fordelingen er beskrevet i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

Den marginale fordeling af empiriske varianser under denne antagelse er beskrevet i afsnit 6.6 på side 599.

Den marginale fordeling af de empiriske varianser er en såkaldt reciprok betafordeling.

Fordelingsforholdene under denne model er anført i

Sætning 5.7.1 Betingede og marginale fordelinger i modellen med varierende varians

Under modellen givet ved (5.7.1) og (5.7.2) vil den betingede fordeling af $Y_{ij}$ givet $\mu_i$ og $\sigma_i^2$ være en normalfordeling med

$$E[Y_{ij}|\mu_i, \sigma_i^2] = \mu_i$$

$$\text{COV}[Y_{ij}, Y_{il}|\mu_i, \sigma_i^2] = \begin{cases} 0 & \text{for } j \neq l \\ \sigma_i^2 & \text{for } j = l \end{cases}$$

Endvidere er de betingede fordelinger af $\overline{Y}_{i+}$ normale med

$$E[\overline{Y}_{i+}|\mu_i, \sigma_i^2] = \mu_i$$

$$\text{VAR}[\overline{Y}_{i+}|\mu_i, \sigma_i^2] = \sigma_i^2/n_i$$

$$E[\overline{Y}_{i+}|\sigma_i^2] = \mu_0$$

$$\text{VAR}[\overline{Y}_{i+}|\sigma_i^2] = \sigma_i^2\left(\frac{1}{n_i} + \frac{1}{m}\right)$$

Den marginale fordeling af gruppeniveauet, $\mu_i$ er givet ved

$$\mu_i \in T(2\alpha, \mu_0, \sqrt{\beta/(m\alpha)})$$

hvor $T(\nu, \mu_0, \beta_1)$ angiver $t$-fordelingen med $\nu$ frihedsgrader og positionsparameter $\mu_0$ og skalaparameter $\beta_1$ (se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998).

Den marginale fordeling af $\overline{Y}_{i+}$ er givet ved

$$T\left(2\alpha, \mu_0, \sqrt{\frac{\beta}{\alpha} \left(\frac{1}{n_i} + \frac{1}{m}\right)}\right)$$
For variansskønnet, $S_i^2$ for den $i$'te gruppe

$$S_i^2 = \sum_j (Y_{ij} - \bar{Y}_{i+})^2 / (n_i - 1), \quad (5.7.5)$$

gælder

$$(n_i - 1)S_i^2 | \sigma^2_i \in \chi^2(n_i - 1),$$

og den marginale fordeling er givet ved

$$S_i^2 \in \text{RBet} \left( \alpha, \frac{1}{2}(n_i - 1), \frac{2\beta}{n_i - 1} \right), \quad (5.7.6)$$

hvor RBet-fordelingen er den reciproke betafordeling.

**Bevis:**

Overspringes.

\[ \square \]

**Bemærkning 1** *Fordelingen af gruppegenennemsnittene har tykkere halser end normalfordelingen*

Vi bemærker, at den marginale fordeling af gruppengennemsnittet, $\mu_i$ og af gruppegenennemsnittet $\bar{Y}_{i+}$ er $t$-fordelinger, der har tykkere halser end normalfordelingen.

Dette skyldes, at det ikke kun er middelværdierne, $\mu_i$, der varierer, men også de tilsvarende varianser.

\[ \square \]

**Bemærkning 2** *Fortolkning af parametrene $\alpha$ og $\beta$ i fordelingen af varianserne.*

Vi bemærker, at fordelingen af $\sigma^2$ kan udtrykkes ved at

$$\frac{1}{\sigma^2} \in \frac{1}{\sigma^2_0} (\nu - 2) \chi^2(\nu), \quad (5.7.7)$$

med $\sigma^2_0 = \text{E} [\sigma^2] = \beta / (\alpha - 1)$ og $\nu = 2\alpha$. se sætning 6.6.1 på side 601

\[ \square \]

Såfremt fraktildiagrammerne for de enkelte grupper i en variansanalyse har forskellige hældninger, såfremt Bartlett's test for varianshomogenitet giver anledning til mistanke om at varianserne er forskellige, eller såfremt fordelingen af gruppegenennemsnittet har tykkere halser end normalfordelingens, samtidig med at fordelingerne inden for grupper kan beskrives ved normale fordelinger, kan det være rimeligt at forsøge at beskrive data ved denne model.
5.7 Normalfordelingsmodeller med tilfældig varians

Bemærkning 3 Variansanalyseeskema ved tilfældigt varierende varians
Lader vi $S^2$ betegne det sædvanlige vejede gennemsnit af de gruppevise variansskøn,

$$S^2 = SAK_1 / (N - k)$$  (5.7.8)

har vi

$$E[S^2] = \beta / (\alpha - 1) = \sigma_0^2$$

således at variansanalyseeskemaet bliver

<table>
<thead>
<tr>
<th>Variation</th>
<th>SAK</th>
<th>f</th>
<th>$E[SAK/f]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mellem grupper</td>
<td>$\sum_i n_i (\bar{Y}<em>{i+} - \bar{Y}</em>{++})^2$</td>
<td>$k - 1$</td>
<td>$\sigma_0^2 (1 + \frac{n_0}{m})$</td>
</tr>
<tr>
<td>Indenfor grupper</td>
<td>$\sum_i \sum_{ij} (Y_{ij} - \bar{Y}_{i+})^2$</td>
<td>$N - k$</td>
<td>$\sigma_0^2$</td>
</tr>
</tbody>
</table>

Det ses, at det sædvanlige variansanalyseeskema kun giver mulighed for estimation af middelvariansen $E[\sigma^2] = \sigma_0^2$. For at bestemme parameteren $\nu$ er det nødvendigt yderligere at betragte variationen mellem de gruppevise varianser.

Sætning 5.7.2 Forventningsværdi af kvadratafvigelsessummen mellem empiriske varianser
Lad

$$SAK_s = \sum_i (n_i - 1)(S_i^2 - S^2)^2$$  (5.7.9)

hvor $S_i^2$ er givet ved (5.7.5) og $S^2$ ved (5.7.8), angive variationen mellem de gruppevise empiriske varianser, da gælder under modellen givet ved (5.7.1) og (5.7.2)

$$E[SAK_s/(k - 1)] = (\sigma_0^2)^2 \{2 \frac{\nu/2 - 1}{\nu/2 - 2} + \frac{1}{(\nu/2 - 2)(k - 1)} \left[ N - k - \sum_i \frac{(n_i - 1)^2}{N - k} \right] \}$$  (5.7.10)

Bevis:
Overspringes.  □
Sætning 5.7.3 Momentestimation i modellen med tilfældigt varierende varians

Momentestimaterne for parametrene $\mu_0, \sigma^2_0, \nu$ og $m$ i modellen (5.7.1) og (5.7.2) er givet ved

\[
\tilde{\mu}_0 = \bar{Y}_{++} \\
\tilde{\nu} = 2 \left[ 1 + \left\{ Q_1 + \frac{1}{k-1} \left[ N - k - \frac{\sum_i (n_i - 1)^2}{N - k} \right] \right\} / (Q_1 - 2) \right] \\
\tilde{\sigma}^2_0 = \frac{SAK_1}{N - k} \quad \text{og} \quad \tilde{m} = \frac{1}{(k - 1)(Q_2 - 1)} \left( N - \frac{\sum_i n_i^2}{N} \right) = \frac{n_0}{Q_2 - 1}
\]

hvor

\[
Q_1 = \frac{SAK_s / (k - 1)}{[SAK_1 / (N - k)]^2} \quad \text{(5.7.11)} \\
\text{og} \quad Q_2 = \frac{SAK_2 / (k - 1)}{SAK_1 / (N - k)} \quad \text{(5.7.12)}
\]

og $n_0$ er givet ved (5.1.9).

**Bevis:**
Overspringes

\[\square\]

**Bemærkning 1 Momentestimation i det balancede tilfælde**

I det balancede tilfælde, $n_1 = n_2 = \cdots = n_k = n$, får vi

\[
\tilde{\nu} = 4 + 2(n + 1) / (Q_1 - 2) \\
\tilde{\sigma}^2_0 = SAK_1 / [k(n - 1)] \\
\text{og} \quad \tilde{m} = n / (Q_2 - 1)
\]
iden der i dette tilfældes gælder

\[ E \left[ \frac{SAK_2}{(k - 1)} \right] = \sigma_0^2 \left( 1 + \frac{n}{m} \right) \]

\[ E \left[ \frac{SAK_s}{(k - 1)} \right] = \frac{(\sigma_0^2)^2}{(\nu/2 - 2)} \left[ 2(\nu/2 - 1) + n - 1 \right] \]

\[ \square \]

**Eksempel 5.7.1 Variation af træffepunkt for 35 skytter**

Nedenstående figur viser den observerede fordeling af det gennemsnitlige træfpunkt for 20 skud for hver af 35 skytter. Den indtegnet kurve viser normalfordelingen med samme middelværdi og samme varians som de 35 resultater.

Det anes, at den observerede fordeling har noget tykkere halder, end normalfordelingen.

Den observerede fordeling af det gennemsnitlige træfpunkt for hver af 35 skytter

![Diagram](image_url)

Fordelingen er sammenlignet med en normalfordeling med samme middelværdi og varians.
For nærmere at undersøge dette, viser nedenstående figur det tilsvarende fraktildiagram (Q-Q-plot).

Man ser også her en antydning af, at fordelingen har tykkere haler, end normalfordelingen. For at belyse fænomenet yderligere, har man i den følgende figur tegnet et histogram over de empiriske varianser for hver skytte. I figuren er indtegnet tætheden for den tilsvarende $\sigma_0^2\chi^2(19)/19$-fordeling. Det ses, at den observerede fordeling har noget tykkere haler, end $\chi^2$-fordelingen.
Det tilsvarende Q-Q plot giver også en antydning af de tykkere haler og den manglende masse i den centrale del:

Og nedenstående P-P plot (kumulerede hyppigheder for de empiriske varianser mod tilsvarende kumulerede sandsynligheder i \( \chi^2 \)-fordelingen, se Shapiro og Wilk (1965) og Wilk og Gnanadesikan (1968)) viser en tydelig afvigelse fra \( \chi^2 \)-fordelingen.
Man har derfor valgt at modellere resultaterne ved en normalfordelingsmodel med en tilfældig middelværdi, $\mu_i$, og en tilfældig varians, $\sigma^2$, for hver skytte. I figuren er med en fuldt optrukket kurve indtegnet tætheden i den marginale fordeling (5.7.4) svarende til denne model.

5.7 Normalfordelingsmodeller med tilfældig varians

Det tilsvarende Q-Q plot viser ligeledes en tilfredsstillende tilpasning
Tabel 5.2. Oversigt over marginale fordelinger i endimensionale varianskomponentfordelinger

Normalfordelte observationer, samme varians i alle grupper.

$k$ grupper og $n_i$ gentagelser i hver gruppe.

Observationer: $Y_{i1}, Y_{i2}, \ldots, Y_{in_i};$

\[
\bar{Y}_{i+} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} ; \quad \bar{Y}_{++} = \frac{\sum_{i=1}^{k} n_i \bar{Y}_{i+}}{N} ; \quad S_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i+})^2 / (n_i - 1)
\]

\[
SAK_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i+})^2 ; \quad SAK_2 = \sum_{i=1}^{k} n_i (\bar{Y}_{i+} - \bar{Y}_{++})^2 ; \quad s_2^2 = SAK_2 / (k - 1)
\]

Model: $Y_{ij} | \mu_i \in N(\mu_i, \sigma^2);$

Parameteren $\mu_i$ er $N(\mu_0, \sigma_0^2)$ fordelt, $(\sigma_0^2 = \gamma \sigma^2).$

| Stikprøveford. af $\bar{Y}_{i+} | \mu_i$ | Fordeling af $\mu_i$ | Marginalford. af $\bar{Y}_{i+}$ | $\mathbb{V} [\bar{Y}_{i+}]$ | Fordeling af $S_i^2$ | $\mathbb{E} [S_i^2]$ | $\mathbb{V} [S_i^2]$ |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| $N(\mu_i, \sigma_i^2/n_i)$ | $N(\mu_0, \gamma \sigma_i^2)$ | $N(\mu_0, \sigma_i^2 (\gamma + 1/n_i))$ | $\sigma_i^2 (\gamma + 1/n_i)$ | $\sigma_i^2 \chi_i^2 (n_i - 1) / (n_i - 1)$ | $\sigma_i^2$ | $2\sigma_i^4 / (n_i - 1)$ |

Momentestimation af parametre i endimensionale varianskomponentmodeller:

\[
\tilde{\mu}_0 = \bar{Y}_{++} ; \quad \tilde{\sigma}^2 = SAK_1 / (N - k) ; \quad \tilde{\sigma}_0^2 = \left\{ SAK_2 - \frac{k - 1}{N - k} SAK_1 \right\} / \{(k - 1) n_0 \}
\]

Den vægtede gennemsnitlige gruppestikprøvestørrelse, $n_0$ er bestemt ved (5.1.9).
Tabel 5.3. Oversigt over marginale fordelinger i endimensionale varianskomponentfordelinger

Normalfordelte observationer, varierende varians i grupper.

\( k \) grupper og \( n_i \) gentagelser i hver gruppe.

Observationer: \( Y_{i1}, Y_{i2}, \ldots, Y_{in_i} \);

\[
\bar{Y}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} ; \quad \bar{Y}_{++} = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{Y}_{i+} ; \quad S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i+})^2 /(n_i - 1)
\]

\[
SAK_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i+})^2 ; \quad SAK_2 = \sum_{i=1}^{k} n_i (\bar{Y}_{i+} - \bar{Y}_{++})^2 ; \quad s^2 = SAK_2/(k - 1) ; \quad S^2_+ = SAK_1/(N - k)
\]

\[
SAK_s = \sum_{i=1}^{k} (n_i - 1)(S_i^2 - S_+^2)^2 ; \quad N = \sum_{i=1}^{k} n_i
\]

Model: \( Y_{ij}|(\mu_i, \sigma_i^2) \in N(\mu_i, \sigma_i^2) \);

For givet \( \sigma_i^2 \) er \( \mu_i | \sigma_i^2 \in N(\mu_0, \sigma_i^2/m) \).

\( \sigma_i^2 \) er \( RGamma(\alpha, \beta) \)-fordelt

| Marginal fordel. af \( \bar{Y}_{i+} | \sigma_i^2 \) | Marginal fordeling af \( \bar{Y}_{i+} \) | Marginal fordeling af \( S_i^2 \) | \( E \left[ \frac{SAK_1}{N-k} \right] \) | \( E \left[ SAK_2/(k - 1) \right] \) |
|-------------------------------------------------|---------------------------------|-----------------|-------------------|-------------------|
| \( N\left(\mu, \sigma_i^2 \left( \frac{1}{n_i} + \frac{1}{m} \right) \right) \) | \( T \left( 2\alpha, \mu_0, \sqrt{\frac{\beta}{\alpha} \left( \frac{1}{n_i} + \frac{1}{m} \right)} \right) \) | \( RBet \left( \alpha, \frac{n_i - 1}{2}, \frac{2\beta}{n_i - 1} \right) \) | \( \frac{\beta}{\alpha - 1} \) | \( \frac{\beta}{\alpha - 1} \left( 1 + \frac{n_0}{m} \right) \) |

\( \sigma_0^2 = \beta/(\alpha - 1) \), \( \nu = 2\alpha \)
Momentestimation af parametre i endimensionale varianskomponentmodeller:
Normaltfordelte observationer, varierende varians i grupper.

\[
\tilde{\mu} = \overline{Y}_{++} \\
\tilde{\alpha} = 1 + \left\{ Q_1 + \frac{1}{k-1} \left[ N - k - \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1)^2 \right] \right\} / (Q_1 - 2) \\
\tilde{\beta} = (\tilde{\alpha} - 1) SAK_1 / (N - k) \\
\tilde{m} = n_0 / (Q_2 - 1)
\]

med

\[
Q_1 = \{ SAK_s / (k - 1) \} / \{ SAK_1 / (N - k) \}^2 \\
Q_2 = \{ SAK_2 / (k - 1) \} / \{ SAK_1 / (N - k) \}
\]

Den vægtede gennemsnitlige gruppestikprøvestørrelse, \( n_0 \), er bestemt ved (5.1.9).
5.8 Reference:


Hierarkiske modeller for eksp. disp. modeller
Afsnit 6

Hierarkiske modeller for eksponentielle dispersionsmodeller

6.1 Indledning

Den problemstilling, vi har behandlet for normalfordelte observationer ved den ensidede variansmodel med tilfældig variation, nemlig en modellering af eventuelle forskelle mellem grupper ved en tilfældig variation, lader sig forholdsvis simpelt generalisere til modeller, hvor den naturlige variation indenfor grupper beskrives ved eksponentielle dispersionsmodeller.

Vi vil i dette afsnit diskutere sådanne modeller, hvor variationen inden for grupper kan beskrives ved en eksponentiel dispersionsmodel, og hvor variationen mellem grupper kan beskrives ved en tilfældig model. Vi vil dog indskrænke os til et enkelt niveau af denne tilfældige variation mellem grupper, ligesom vi vil antage en vis form for "homogenitet" også i variationen mellem grupper, nemlig af fordelingen af variationen mellem grupper har en unimodal tæthed.
Vi vil indlede med at betragte

### 6.1.1 Den systematiske model

Vi betragter følgende model:

For fastholdt \( i \) er de variable \( X_{i1}, X_{i2}, \ldots, X_{in_i} \) indbyrdes uafhængige og identisk fordelt med tæthed af formen

\[
g(x; \vartheta_i) = d(x) \exp\{[\vartheta_i x - \kappa(\vartheta_i)]/\sigma^2\},
\]

(6.1.1)

der er fuldstændigt specificeret på nær den ukendte parameter \( \vartheta_i \) som karakteriserer den \( i \)te gruppe.

Modellen specificerer således, at for fastholdt \( i \) vil fordelingen af \( X_{i1}, X_{i2}, \ldots, X_{in_i} \) tilhøre en eksponentiel dispersionsmodel med middelværdiparameter \( \mu_i = \tau(\vartheta_i) \), og med varians \( \nu_i = \sigma^2 V(\mu) \), hvor middelværdiafbildningen \( \tau(\cdot) \) og enhedsvariansfunktionen \( V(\cdot) \) som vanligt bestemmes ved

\[
\begin{align*}
\mu &= \tau(\vartheta) = \kappa'(\vartheta) \\
V(\mu) &= \kappa''(\tau^{-1}(\mu))
\end{align*}
\]

(6.1.2)

svarende til (2.2.27) og (2.2.28).

Modellen (6.1.1) kan i en række tilfælde verificeres, f.eks. ved betragtning af de \( k \) fraktildiagrammer svarende til observationssættene \( X_{i1}, X_{i2}, \ldots, X_{in_i} \) for \( i = 1, 2, \ldots, k \).

Under modellen (6.1.1) er

\[
Z_i = \sum_{j=1}^{n_i} X_{ij} = n_i \bar{X}_i
\]

sufficient for parameteren \( \vartheta_i \), og tætheden for \( Z_i \) er på formen

\[
g_z(z; \vartheta_i) = h_z(z, n_i) \exp\{[\vartheta_i z - n_i \kappa(\vartheta_i)]/\sigma^2\},
\]

(6.1.3)

altså en eksponentiel dispersionsmodel med middelværdi \( \mu_i = E[Z_i] = n_i \tau(\vartheta_i) \) og variansfunktion \( V_Z(\mu) = n_i V(\mu/n_i) \).
I stedet for at betragte summerne $Z_i$, vil man ofte betragte gennemsnittene

$$\overline{X}_{i+} = Z_i/n_i = \sum_{j=1}^{n_i} X_{ij}/n_i$$

Det følger af sætning 2.2.5, at fordelingen af $Y_i = \overline{X}_{i+}$ tilhører den samme eksponentielle dispersionsmodel som $X_{ij}$, (dvs med samme dispersionsparameter $\sigma^2$ og variansfunktion $V(\mu)$) men med vægten $w_i = n_i$.

Tætheden for $\overline{X}_{i+}$ er således

$$g_{\varphi}(\overline{x}; \varphi_i) = h_{\varphi}(\overline{x}, n_i) \exp\{n_i[\varphi_i(\overline{x} - \kappa(\varphi_i)]/\sigma^2\}$$

og der gælder

$$\mu_i = E[\overline{X}_{i+}] = \tau(\varphi_i)$$

$$V[\overline{X}_{i+}] = \sigma^2 V(\mu_i)/n_i,$$

hvor middelværdiafbildningen $\tau(\cdot)$ og variansfunktionen $V(\cdot)$ er bestemt ved (6.1.2).

I lighed med situationen i afsnit 2 benytter vi hovedsageligt den kanoniske form og de kanoniske parametre til at beskrive den basale struktur. Ved analysen af data vil man ofte benytte en middelværdiparametrerisering. Specielt vil man således formulere homogenitetshypotesen ved middelværdiparametrene $\mu$.

I afsnit 2.7.1 har vi i sætning 2.7.1 diskuteret test af homogenitetshypotesen

$$H_I: \mu_1 = \mu_2 = \cdots = \mu_k$$

for de sædvanlige fordelinger under en systematisk model.

### 6.1.2 Den tilfældige model

Såfremt hypotesen $H_I$ må afvises, kan det i en række situationer være naturligt at modellere data ved en tilfældig model, d.v.s. at antage at $\varphi_i$, $i = 1, \ldots, k$ er realiserede værdier fra en fordeling af $\varphi$'er. Fordelingen af $\varphi$ kaldes strukturfordelningen eller apriorifordelningen af $\varphi$.

Under den tilfældige model vil vi som nævnt opfatte parameteren $\varphi_i$, der karakteriserer den betingede fordeling af observationerne $X_{i1}, X_{i2}, \ldots, X_{in}$, i den $i$’te gruppe, som en stokastisk variabel.
Vi vil yderligere antage, at $X_{i1}, X_{i2}, \ldots, X_{in_i}$ er betinget uafhængige (jvf definition 4.5.1 side 467) af $\vartheta_i$ i den simultane fordeling af $X_{i1}, X_{i2}, \ldots, X_{in_i}$ og $\vartheta_i$.

For en udvalgt gruppe, $i$, er den betingede fordeling af $X_{i1}, X_{i2}, \ldots, X_{in_i}$, givet gruppeparameteren $\vartheta_i$, altså sådan at $X_{i1}, X_{i2}, \ldots, X_{in_i}$ er indbyrdes uafhængige med tæthed (6.1.1).

Den simultane fordeling af $X_{i1}, X_{i2}, \ldots, X_{in_i}$, for givet $\vartheta_i$ har da tætheden

$$f(x_{i1}, x_{i2}, \ldots, x_{in_i} | \vartheta_i) = \prod_{j=1}^{n_i} g(x_{ij}; \vartheta_i)$$

**Definition 6.1.1 Konjugeret klasse af fordelinger**

Betragt en eksponentiel dispersionsmodel med tætheder af formen (6.1.1), dvs

$$g(x; \vartheta) = d(x) \exp\{[\vartheta x - \kappa(\vartheta)]/\sigma^2\} \quad (6.1.6)$$

hvor parameteren $\vartheta \in D$

Lad $\mathcal{M} = \tau(D)$ angive middelværdirummet svarende til denne fordeling.

For $m \in \mathcal{M}$ og $\gamma \in \Delta \subseteq \mathbb{R}_+$ vil

$$g(\vartheta; m, \gamma) = \frac{1}{C(m, \gamma)} \exp\{[\vartheta m - \kappa(\vartheta)]/\gamma\} \quad (6.1.7)$$

med

$$C(m, \gamma) = \int_D \exp\{[\vartheta m - \kappa(\vartheta)]/\gamma\} \, d\vartheta$$

være en tæthed for fordelingen af $\vartheta$.

Familien (6.1.7) for $(m, \gamma) \in \mathcal{M} \times \Delta$ kaldes den konjugerede klasse til (6.1.6).

**Bemærkning 1 Den konjugerede klasse er fastlagt af variansfunktionen**

Vi bemærker, at den konjugerede klasse alene afhænger af den kumulantfrembringerende funktion $\kappa(\cdot)$. Den konjugerede klasse er således fastlagt, blot $\kappa(\cdot)$ er fastlagt. Imidlertid gælder det (jvf sætning 1.2.4 i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998, at $\kappa(\cdot)$ er fastlagt af variansfunktionen $V(\cdot)$.

□
Bemærkning 2 Den konjugerede klasse udgør en en eksponentiel familie

Det fremgår af formen (6.1.7) på tætheden for fordelingen af \( \theta \), at for fastholdt \( \gamma \) vil familien af fordelinger af \( \theta \) udgøre en eksponentiel familie med den kanoniske parameter \( m \in \mathcal{M} \).

Vi vil betragte strukturfordelinger (eller apriorifordelinger), der svarer til, at tætheden for den kanoniske parameter \( \theta \) tilhører den konjugerede klasse. Sådanne fordelinger kaldes konjugerede apriorifordelinger eller konjugerede strukturfordelinger.

De konjugerede strukturfordelinger og de tilsvarende marginale fordelinger for de almindelige stikprøvefordelinger er angivet i tabel 6.1.

I lighed med situationen i afsnit 2 benytter vi hovedsageligt den kanoniske form og de kanoniske parametre til at beskrive den basale struktur, - og her altså yderligere til at bestemme den konjugerede strukturfordeling. Ved analysen vil vi benytte den sædvanlige middelværdiparametrisering \( \mu = \tau(\theta) \) svarende til den eksponentielle dispersionsmodel.

Det er klart, at når strukturfordelingen for \( \theta \) er fastlagt, kan man blot indføre transformationen \( \mu = \tau(\theta) \), hvorved strukturfordelingen for \( \mu \) er bestemt.

Sætning 6.1.1 Fortolkning af parametrene i den konjugerede fordeling

Lad fordelingen af \( X \) tilhøre den eksponentielle dispersionsmodel (6.1.6) og lad strukturfordelingen af \( \theta \) være givet ved (6.1.7). Da gælder under visse regularitetsbetingelser

\[
m = \mathbb{E}[\mu] \tag{6.1.8}
\]

\[
\gamma = \frac{\mathbb{V}[\mu]}{\mathbb{E}[\mathbb{V}(\mu)]}, \tag{6.1.9}
\]

hvor \( \mu = \mathbb{E}[X|\theta] \), og hvor \( \mathbb{V}(\mu) \) angiver variansfunktionen (6.1.2) i den betingede fordeling af observationerne.
Hierarkiske modeller for eksp. disp. modeller

BEMÆRK forskellen mellem symbolet \( V[\mu] \), der angiver variansen i apriorifordelingen af observationernes betingede middelværdier \( \mu \), og funktionen \( V(\mu) \), der angiver variansfunktionen (6.1.2) i den betingede fordeling af observationerne.

**Bevis:**
Fordelingen af en enkelt observation \( X \) har den betingede tæthed

\[
    f(x|\vartheta) = d(x) \exp\{[(x - \kappa(\vartheta))/\sigma^2]\}
\]

og tætheden for \( \vartheta \) er

\[
    g(\vartheta; m, \gamma) = \frac{1}{C(m, \gamma)} \exp\{[(\vartheta - \kappa(\vartheta))/\gamma]\}
\]

Vi antager at parameterområdet \( D \) udgør hele den reelle akse, og at

\( g(\vartheta; m, \gamma) \to 0 \) for \( \vartheta \to \pm \infty \), og ligeledes at \( g(\vartheta; m, \gamma) \to 0 \) for \( \vartheta \to \pm \infty \).

Der gælder da

\[
    \int_{-\infty}^{\infty} g'(\vartheta; m, \gamma) d\vartheta = g(\vartheta; m, \gamma) \bigg|_{\vartheta=-\infty}^{\infty} = 0,
\]

men

\[
    g'(\vartheta; m, \gamma) = \frac{m - \tau(\vartheta)}{\gamma} g(\vartheta; m, \gamma),
\]

hvorfor

\[
    \int_{-\infty}^{\infty} g'(\vartheta; m, \gamma) = \int_{-\infty}^{\infty} \frac{m - \tau(\vartheta)}{\gamma} g(\vartheta; m, \gamma) = \frac{m - \mathbb{E}[\tau(\vartheta)]}{\gamma};
\]

således at vi har

\[
    m = \mathbb{E}[\tau(\vartheta)].
\]

Tilsvarende får man

\[
    g''_1(\vartheta; m, \gamma) = \left[\frac{[m - \tau(\vartheta)]^2}{\gamma^2} - \frac{\kappa''(\vartheta)}{\gamma}\right] g(\vartheta; m, \gamma),
\]

og idet

\[
    \int_{-\infty}^{\infty} g''_1(\vartheta; m, \gamma) d\vartheta = g'(\vartheta; m, \gamma) \bigg|_{\vartheta=-\infty}^{\infty} = 0,
\]

får man da

\[
    V[\tau(\vartheta)] = \gamma \mathbb{E}[\kappa''(\vartheta)].
\]
6.1 Indledning

Men da $\kappa''(\vartheta)$ netop er variansfunktionen, har man (6.1.9)

For en diskussion af betingelserne for sætningen se f.eks. U. Müller-Funk

**Bemærkning 1 Fortolkning ved varians indenfor grupper og varians mellem grupper**

Idet fordelingen af $X$ antages at have dispersionsparameteren $\sigma^2$ (og vægten 1) har vi

$$
E [X|\vartheta] = \mu = \tau(\vartheta)
$$
$$
V [X|\vartheta] = \sigma^2 V(\mu) = \sigma^2 \kappa''(\tau^{-1}(\mu))
$$

og

$$
m = E [\tau(\vartheta)] = E [E [X|\vartheta]] = E [\mu]
$$
$$
\gamma = \frac{V [\tau(\vartheta)]}{E [\kappa''(\vartheta)]} = \frac{V [E [X|\vartheta]]}{E [V [X|\vartheta]]/\sigma^2}
$$

**Sætning 6.1.2 Momenter i marginal fordeling af enkeltobservationer og af gruppegennemsnit**

Såfremt den betingede fordeling af $X_{ij}|\mu$ har tæthed af formen (6.1.1), og såfremt fordelingen af $\mu$ er sådan, at tætheden af $\vartheta = \tau^{-1}(\mu)$ er af formen (6.1.7), da gælder for den marginale fordeling af $X_{ij}$ (jvf. sætning 0.1.1 i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998)

$$
E [X_{ij}] = m
$$

(6.1.10)

$$
COV[X_{ij}, X_{hl}] = \begin{cases} 
E [V(\mu)](\sigma^2 + \gamma) & \text{for } (i, j) = (h, l) \\
V [\mu] = \gamma E [V(\mu)] & \text{for } i = h, j \neq l \\
0 & \text{for } i \neq h
\end{cases}
$$
Observationer $X_{ij}$ og $X_{il}$ i samme gruppe er således indbyrdes korrelerede med intraklassekorrelationen

\[
\rho = \frac{V[\mu]}{\sigma^2 E[V(\mu)] + V[\mu]} = \frac{\gamma E[V(\mu)]}{E[V(\mu)](\sigma^2 + \gamma)} = \frac{\gamma}{\sigma^2 + \gamma} \tag{6.1.11}
\]

Omvendt kan signal/støj forholdet, $\gamma$, udtrykkes ved intragruppekorrelationen $\rho$ som

\[
\gamma = \frac{\rho}{1 - \rho} \sigma^2 \tag{6.1.12}
\]

Momenterne i fordelingen af gruppegennemsnittene, $\overline{X}_{i+}$, er

\[
E[\overline{X}_{i+}] = m \tag{6.1.13}
\]
\[
V[\overline{X}_{i+}] = E[V(\mu)](\gamma + \frac{\sigma^2}{n_i}) \tag{6.1.14}
\]

Endvidere er $\overline{X}_{i+}$ og $\overline{X}_{h+}$ uafhængige for $i \neq h$.

**Bevis:**
Resultatet følger af sætning 0.1.1 i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

Det gælder således, at

\[
E[\overline{X}_{i+}] = E[E[\overline{X}_{i+}|\mu]] = E[\mu] = m
\]

Endvidere har man

\[
V[\overline{X}_{i+}] = E[V[\overline{X}_{i+}|\mu]] + V[E[\overline{X}_{i+}|\mu]] = \frac{\sigma^2}{n_i} E[V(\mu)] + V[\mu] \tag{6.1.15}
\]
\[
= E[V(\mu)](\gamma + \frac{\sigma^2}{n_i})
\]
6.1 Indledning

Til brug for en simpel parameterestimation anfører vi

**Sætning 6.1.3 Forventningsværdi af kvadratafvigelsessummen svarende til variationen mellem grupper under den tilfældige model**

Lad $\overline{X}_{i+}$ være som i ovenstående sætning, og sæt som vanligt

$$SAX_2 = \sum_{i=1}^{k} n_i (\overline{X}_{i+} - \overline{X}_{++})^2. \quad (6.1.16)$$

Da gælder under den tilfældige model

$$E[SAX_2] = (k - 1)E[V(\mu)](\sigma^2 + n_0 \gamma), \quad (6.1.17)$$

hvor $n_0$ er givet ved (5.1.9) og hvor $\gamma$ er givet ved (6.1.9)

**Bevis:**
Resultatet følger af lemma 0.2.1 i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

Udtrykket ses også direkte ved at bemærke, at

$$E[SAX_2] = \sum_i n_i (1 - \frac{n_i}{N}) V[\overline{X}_{i+}] = (k - 1)\{\sigma^2 E[V(\mu)] + n_0 V[\mu]\}$$

Relationen (6.1.17) sammen med

$$E[\overline{X}_{++}] = E[\mu] = m \quad (6.1.18)$$

kan benyttes til at beregne et simpelt estimat for parametrene i apriorifordelingen.

Vi bemærker, at momentmetoden anvendt på $SAX_2$ ikke nødvendigvis sikrer, at observationerne vægtes med deres præcisioner. I en række tilfælde kunne det være naturligere at betragte

$$\sum_i (\overline{X}_{i+} - \overline{X}_{++})^2/(k - 1)$$
som et udtryk for den gennemsnitlige varians $V[\bar{X}_{i+}]$. På grund af den simple opspaltung (5.1.6) vil vi imidlertid hovedsageligt betragte estimation baseret på $SAK_2$.

En mere præcis estimationsmetode er maksimum-likelihood metoden. Maksimum-likelihood estimatet må sædvanligvis yybestemmes ved iteration.

Notationen og begrebsapparatet, der benyttes til beskrivelse af forholdene omkring de sædvanlige diskrete fordelinger, binomialfordelingen, Poissonfordeling m.v. kompliceres af at beskrivelsen af fordelingsforholdene er simpst for de heltallige gruppetotaler, $Z_i$. Den praktiske interesse knytter sig imidlertid sædvanligvis til middelværdiparameteren $\mu$ svarende til en dispersionsmodel for $\bar{X}_{i+} = Z_i/n_i$.

Vi har derfor fundet det nødvendigt at operere med to parallele parametriseringer af apriorifordelingen.

Den ene parametrisering, (sædvanligvis betegnet med symbolerne $\alpha$ og $\beta$) benyttes til beskrivelse af de eksakte fordelingsforhold for de betragtede diskrete fordelinger af $Z_i$ og de tilsvarende sammensatte (compound) fordelinger, da disse fordelinger oftest (blandt andet i Statistik I) er udtrykt ved denne parametrisering.

Den anden parametrisering, der knytter sig til middelværdiparametriseringen (ved $\mu$ og $V(\mu)$) af den betingede fordeling af observationerne, benytter en middelværdiparametrisering for fordelingen af $\mu$ (ved parametrene $m$ og $\gamma$).

### 6.2 Bernoullifordelingen

Såfremt $X_{ij}|\mu_i \in B(1, \mu_i)$, finder vi $Z_i|\mu_i \in B(n_i, \mu_i)$

**Sætning 6.2.1 Test for homogenitet**

Kvotientteststørrelsen for hypotesen (6.1.5) er

$$G^2(H_1) = \sum_i n_i d(h_i; \bar{h}_+)$$

(6.2.1)

hvor

$$h_i = \frac{z_i}{n_i} = \bar{x}_{i+} \quad \text{og} \quad \bar{h}_+ = \frac{\sum z_i}{\sum n_i} = \bar{x}_{++}$$

(6.2.2)
angiver de observerede relative hyppigheder i henholdsvis den $i^{te}$ gruppe og i totalmaterialet og $d(h_i; \bar{h}_+)$ angiver enhedsdeviansen for binomialfordelingen,

$$d(h_i; \bar{h}_+) = 2 \left[ h_i \ln(h_i/\bar{h}_+) + (1 - h_i) \ln((1 - h_i)/(1 - \bar{h}_+)) \right]$$

Under hypotesen (6.1.5) vil $G^2(H_I)$ asymptotisk være fordelt som $\chi^2(k - 1)$. Hypotesen forkastes for store værdier af $G^2(H_I)$.

**Bevis:**
Den anførte teststørrelse er deviansteststørrelsen (Tabel 2.3) svarende til hypotesen $H_I$ imod den fulde model $X_i+ \in B(n_i, \mu_i)/n_i$. □

**Bemærkning 1 Pearson teststørrelsen**

Pearson-teststørrelsen for hypotesen (6.1.5) er

$$X^2 = \sum_{i=1}^{k} n_i \frac{(h_i - \bar{h}_+)^2}{\bar{h}_+(1 - \bar{h}_+)} = \sum_{i=1}^{k} \frac{(Z_i - n_i\bar{h}_+)^2}{n_i\bar{h}_+(1 - \bar{h}_+)} , \quad f = k - 1$$

(6.2.3)

er asymptotisk ækvivalent med kvotienttestet. Dette test måler direkte variationen imellem grupper, $SAK_2$, i forhold til den estimerede binomialvarians, $\bar{h}_+(1 - \bar{h}_+)$. □

Så fremt det findes begrundet at afvise hypotesen (6.1.5), kan man f.eks. vælge at modellere fordelingen af $\mu$ ved en Be($\alpha, \beta$)-fordeling.

**Sætning 6.2.2 Den marginale fordeling af gruppetotal ved betabinomial sampling**

Så fremt $Z|\mu \in B(n, \mu)$ og $\mu \in \text{Be}(\alpha, \beta)$, da er den marginale fordeling af $Z$ en $\text{Pl}(n, \alpha, \alpha + \beta)$-fordeling.

Der gælder
Hierarkiske modeller for eksp. disp. modeller

\[
E[Z] = n\pi \\
V[Z] = n \frac{\pi(1-\pi)}{1+\gamma} [1+n\gamma]
\]

med

\[\pi = \alpha/(\alpha + \beta), \quad \gamma = \frac{1}{\alpha + \beta}\]

(6.2.5)

For den relative hyppighed \(Y = Z/n\) gælder:

\[
E[Y] = \pi \\
V[Y] = \frac{\pi(1-\pi)}{1+\gamma} \left(\gamma + \frac{1}{n}\right)
\]

(6.2.6)

Bevis:
For \(\mu \in \text{Be}(\alpha, \beta)\) har vi

\[
E[\mu] = \pi = \frac{\alpha}{\alpha + \beta} \\
V[\mu] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{\pi(1-\pi)}{\alpha + \beta + 1}
\]

(6.2.7)

(6.2.8)

(Familien af betaafdelinger kan opfattes som en eksponentiel dispersionsmodel med middelværdiparameter \(\pi = \alpha/(\alpha + \beta)\), variansfunktion \(V_{\text{Beta}}(\pi) = \pi(1-\pi)\) og dispersionsparameter \(\sigma^2 = 1/(\alpha + \beta + 1)\)).
6.2 Bernoullifordelingen

Vi betragter nu enhedsvariansfunktionen, \( V_{\text{Bin}}(\mu) \), for binomialfordelingen, bestemt ved \( V_{\text{Bin}}(\mu) = \mu(1 - \mu) \). Der gælder

\[
\gamma = \frac{\mathbb{V} [\mu]}{\mathbb{E} [V_{\text{Bin}}(\mu)]} = \frac{\mathbb{V} [\mu]}{\mathbb{E} [\mu(1 - \mu)]} = \frac{1}{\alpha + \beta}, \tag{6.2.9}
\]

idet

\[
\mathbb{E} [V_{\text{Bin}}(\mu)] = \mathbb{E} [\mu(1 - \mu)] = \pi(1 - \pi) \frac{\alpha + \beta}{\alpha + \beta + 1}
\]

Vi har således

\[
\mathbb{E} [V_{\text{Bin}}(\mu)] = \frac{\pi(1 - \pi)}{1 + \gamma} = \frac{V_{\text{Bin}}(\pi)}{1 + \gamma}.
\]

Sætningen følger nu ved indsættelse i (6.1.14).

\[\square\]

**Bemærkning 1 Overdispersion i forhold til binomialfordelingen**

Vi bemærker indledningsvis, at parameteren \( \pi \) netop angiver \( \mathbb{E} [\mu] \).

Opspaltningen svarende til variationen indenfor grupper og variationen mellem grupper giver størrelserne

\[
\mathbb{E} [V[Z | \mu]] = n\mathbb{E} [V_{\text{Bin}}(\mu)] = nV_{\text{Bin}}(\pi) \frac{1}{\gamma + 1}
\]

\[
\mathbb{V} [\mathbb{E} [Z | \mu]] = n^2\mathbb{V} [\mu] = n^2\pi(1 - \pi) \gamma \frac{\gamma}{\gamma + 1} = n^2V_{\text{Bin}}(\pi) \frac{\gamma}{\gamma + 1}
\]

Udtrykt ved intraklassekorrelationen

\[
\rho = \frac{\gamma}{\gamma + 1} = \frac{\alpha + \beta}{\alpha + \beta + 1}
\]

Finder vi

\[
\mathbb{E} [V[Z | \mu]] = n\mathbb{E} [V_{\text{Bin}}(\mu)] = nV_{\text{Bin}}(\pi)(1 - \rho)
\]

\[
\mathbb{V} [\mathbb{E} [Z | \mu]] = n^2\mathbb{V} [\mu] = n^2V_{\text{Bin}}(\pi)\rho
\]

og
\[ V[Z] = nV_{Bin}(\pi)\{1 + (n - 1)\rho\} \]

Lader vi \( Y = Z/n \) angive den observerede relative hyppighed finder vi tilsvarende

\[
V[Y] = \frac{V_{Bin}(\pi)}{\gamma + 1} \left( \gamma + \frac{1}{n} \right) \\
E[V[Y|\mu]] = \frac{V_{Bin}(\pi)}{n} \left( \frac{1}{\gamma + 1} \right) \\
V[E[Y|\mu]] = V[\mu] = V_{Bin}(\pi) \frac{\gamma}{\gamma + 1}
\]

dvs

\[ V[Y] = V_{Bin}(\pi)\{\rho + \frac{1 - \rho}{n}\} \]

Vi bemærker, at for \( n \rightarrow \infty \) vil \( V[Y] \) nærme sig variansen \( V[\mu] = V_\beta(\pi) \) i fordelingen af \( \mu \), hvilket er i overensstemmelse med at fordelingen af \( Y \) vil nærme sig fordelingen af \( \mu \).

For begrænsede værdier af \( n \) vil fordelingen af \( Z \) have en overdispersion i forhold til binomialfordelingen, \( B(n, \pi) \), med samme antalsparameter og med sandsynlighedsparameter \( \pi \) svarende til \( E[\mu] \). Overdispersionen er

\[ \sigma^2 = \frac{V[Z]}{nV_{Bin}(\pi)} = \frac{1 + n\gamma}{1 + \gamma} \]  

(6.2.11)

Tilsvarende vil fordelingen af \( Y = Z/n \) have overdispersionen \( \sigma^2 = (1 + n\gamma)/(1 + \gamma) \) i forhold til \( B(n, \pi)/n \)-fordelingen.

\[ \square \]

**Bemærkning 2** Overdispersionen udtrykt ved "den effektive stikprøvestørrelse"
Det følger af ovenstående bemærkning, at den marginale varians af den relative hyppighed, $Y$, kan udtrykkes som

$$V[Y] = V_{Bin}(\pi)/n_{eff},$$

hvor den effektive stikprøvestørrelse, $n_{eff}$ er bestemt ved

$$n_{eff} = \frac{n}{\sigma^2} = n \frac{\gamma + 1}{n\gamma + 1} \quad (6.2.12)$$

eller

$$\frac{1}{n_{eff}} = \frac{1}{n(\gamma + 1)} + \frac{\gamma}{\gamma + 1} = \frac{1}{n} (1 - \rho) + \rho$$

Variansen på $Y$ er således den samme, som variansen i en $B(n_{eff}, \pi)/n_{eff}$ fordelt variabel.

Jo større værdi af intraklassekorrelationen $\rho$, desto mindre bliver den effektive stikprøvestørrelse, $n_{eff}$.

Ved analyse af strukturer i middelværdiparameteren $\pi$ kan man altså med rimelig tilnærmede benytte generaliserede lineære modeller svarende til binomialfordelingen, idet man blot erstatte den aktuelle stikprøvestørrelse, $n$ med den mindre, effektive stikprøvestørrelse, $n_{eff}$ bestemt ved (6.2.12) for at tilgodese overdispersionen.

Specielt kan man eksempelvis bestemme approximative konfidensintervaller for $\pi$ ved bestemmelse af konfidensintervaller for en tilsvarende binomialfordelt størrelse med stikprøvestørrelsen $n_{eff}$.

Bemærkning 3 **Fortolkning af parameteren $\gamma$**

Parameteren $\gamma = \sqrt{\mu}/E[V_{Bin}(\mu)]$ er sammen med stikprøvestørrelsen bestemmende for overdispersionen.

Vi har

$$\gamma^{-1} = \left[ \frac{V_{Bin}(E[\mu])}{V[\mu]} - 1 \right]$$
Parameteren $\gamma$ måler således afvigelser fra den rene binomialfordeling af $Z$. For $\pi \in [0, 1]$, vil $\mathbb{V} [\mu] \rightarrow 0$ være enbetydende med $\gamma \rightarrow 0$, og fordelingen af $Z$ vil nærme sig en $B(n, \pi)$-fordeling.

For $\gamma > \pi$ vil fordelingen af $\mu$ være J-formet med modus for $\mu = 0$. For $\gamma > 1 - \pi$ vil fordelingen af $\mu$ være J-formet med modus for $\mu = 1$. For $\gamma < \min\{\pi, 1 - \pi\}$ er fordelingen af $\mu$ unimodal med modus i $\mu = (\pi - \gamma)/(1 - 2\gamma)$. □

**Sætning 6.2.3 Momentestimation i Polyafordelingen**

Lad $Z_1, Z_2, \ldots, Z_k$ være uafhængige variable, hvor $Z_i \in \text{Pl}(n_i, \alpha, \alpha + \beta)$

Momentestimaterne for $\pi = \frac{\alpha}{\alpha + \beta}$ og $\gamma = \frac{1}{\alpha + \beta}$ er:

\[
\tilde{\pi} = \bar{h}_+ \tag{6.2.13}
\]

\[
\tilde{\gamma} = \frac{s_2^2 - \bar{h}_+(1 - \bar{h}_+)}{n_0 \bar{h}_+(1 - \bar{h}_+) - s_2^2}
\]

hvor

\[
s_2^2 = \frac{\text{sk}^2}{(k - 1)} = \sum_{i=1}^{k} n_i (h_i - \bar{h}_+)^2 / (k - 1) \tag{6.2.14}
\]

\[
h_i = z_i / n_i; \quad \bar{h}_+ = \frac{\sum_{i=1}^{k} z_i}{\sum_{i=1}^{k} n_i} \tag{6.2.15}
\]

og hvor den vægtede gennemsnitlige stikprøvestørrelse, $n_0$, er givet ved (5.1.9).

Momentestimaterne for $\alpha$ og $\beta$ bliver derfor

\[
\tilde{\alpha} = \bar{h}_+ \frac{n_0 \bar{h}_+(1 - \bar{h}_+) - s_2^2}{s_2^2 - \bar{h}_+(1 - \bar{h}_+)} \tag{6.2.16}
\]

\[
\tilde{\beta} = (1 - \bar{h}_+) \frac{n_0 \bar{h}_+(1 - \bar{h}_+) - s_2^2}{s_2^2 - \bar{h}_+(1 - \bar{h}_+)}
\]
6.2 Bernoullifordelingen

Bevis:
Overspringes

Bemærkning 1 Singular løsning

Hvis $s^2_2 < \bar{k}_+(1 - \bar{k}_+)$, bliver $\tilde{\gamma}$, $\tilde{\alpha}$ og $\tilde{\beta}$ negative. I dette tilfælde vil det derfor være naturligt at sætte disse størrelser til nul, d.v.s fordelingen af $\mu$ estimeres til at være en étpunktfordeling, og fordelingen af $Z_i$ bliver da en $B(n, \pi)$-fordeling. Parametreringsen ved $(\pi, \gamma)$ tillader netop estimation af $\pi$, selv i tilfældet $\gamma = 0$.

Sætning 6.2.4 Maksimum-likelihood estimation i Polyafordelingen

Lad $Z_1, Z_2, \ldots, Z_k$ være uafhængige variable, hvor $Z_i \in P(n_i, \alpha, \alpha + \beta)$

Maksimum-likelihood estimaterne $(\hat{\pi}, \hat{\gamma})$ for $\pi = \frac{\alpha}{\alpha + \beta}$ og $\gamma = \frac{1}{\alpha + \beta}$ findes da ved at maksimere

$$l(\pi; \gamma; z_1, z_2, \ldots, z_k) = \sum_{i=1}^{k} \left( \sum_{\nu=0}^{z_i-1} \ln(\pi + \nu \gamma) + \sum_{\nu=0}^{n_i-z_i-1} \ln(1 - \pi + \nu \gamma) - \sum_{\nu=0}^{n_i-1} \ln(1 + \nu \gamma) \right)$$

(6.2.17)

med hensyn til $\pi$ og $\gamma$.

Bevis:
Sætningen bevises ved at notere, at - på nær en konstant - er logaritmen til likelihoodfunktionen givet ved (6.2.17)

Bemærkning 1 Bestemmelse af maksimum-likelihood estimaterne

Komponenterne af scorefunktionen er

$$\frac{\partial l}{\partial \pi} = \sum_{i=1}^{k} \left( \sum_{\nu=0}^{z_i-1} \frac{1}{\pi + \nu \gamma} - \sum_{\nu=0}^{n_i-z_i-1} \frac{1}{1 - \pi + \nu \gamma} \right)$$

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{k} \left( \sum_{\nu=0}^{z_i-1} \frac{\nu}{\pi + \nu \gamma} + \sum_{\nu=0}^{n_i-z_i-1} \frac{\nu}{1 - \pi + \nu \gamma} - \sum_{\nu=0}^{n_i-1} \frac{\nu}{1 + \nu \gamma} \right)$$
Såfremt maksimum findes i et indre punkt, fås maksimum-likelihood estimerne ved at sætte scorefunktionen lig nul og løse ligningerne med hensyn til π og γ. Ligningerne må løses iterativt. Som startværdier for iterationen kan benyttes momentestimatorerne \( \bar{\pi} \) og \( \bar{\gamma} \) bestemt ved (6.2.13).

Såfremt \( s_2^2 < \bar{h}_+ (1 - \bar{h}_+) \), må man formode, at maksimum ligger på randen af området svarende til \( \bar{\gamma} = 0 \) (omend dette ikke er alment bevidst). For \( \gamma = 0 \) er maksimum-likelihood estimatorens for π bestemt ved \( \bar{\pi} = \bar{h}_+ \).

**Bemærkning 2 Kvotienttest for homogenitet**

Kvotientteststørrelsen for hypotesen

\[
H_{II} : \gamma = 0 \quad \text{med alternativet} \quad \bar{H}_{II} : \gamma > 0
\]

er

\[
Z = 2 \sum_{i=1}^{k} \left[ \sum_{\nu=0}^{z_i - 1} \ln \left( \frac{\hat{\pi} + \nu \hat{\gamma}}{\bar{h}_+} \right) + \sum_{\nu=0}^{n_i - z_i - 1} \ln \left( \frac{1 - \hat{\pi} + \nu \hat{\gamma}}{1 - \bar{h}_+} \right) - \sum_{\nu=0}^{n_i - 1} \ln(1 + \nu \hat{\gamma}) \right]
\]

(6.2.18)

Under \( H_{II} \) vil \( Z \) approximativt følge en \( \chi^2(1) \)-fordeling.

Teststørrelsen bestemmes ved at bemærke, at maksimum af likelihoodfunktionen (6.2.17) under \( H_{II} \) fås for \( \bar{\pi} = \bar{h}_+ \). Ved indsatse av \( \gamma = 0 \) og \( \bar{\pi} = \bar{h}_+ \) i (6.2.17) fås

\[
l(\bar{\pi}, 0; z_1, z_2, \ldots, z_k) = \sum_{i=1}^{k} \left[ z_i \ln(\bar{h}_+) + (n_i - z_i) \ln(1 - \bar{h}_+) \right]
\]

Subtraheres denne værdi fra den fundne maksimumværdi af (6.2.17) og multipliceres med 2 fås (6.2.18).

Vi bemærker, at under den tilfældige model fører kvotienttestet for homogenitet til en anden teststørrelse, end under den systematiske model.

**Eksempel 6.2.1 Variation mellem andelen af dimensionsafvigende låg i stikprøver fra 229 partier**

Nedenstående tabel viser fordelingen af antallet af dimensionsafvigende låg i stikprøver å 770 låg fra hvert af 229 produktionspartier (Kilde: J.H Ford (1951))
### 6.2 Bernoullifordelingen

<table>
<thead>
<tr>
<th>$z_i$</th>
<th>Antal stikprøver</th>
<th>B(770, 0.0014)</th>
<th>Pl(770, 0.460, 319.1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Forv.</td>
<td>$(\text{obs} - \text{forv})^2$</td>
<td>Forv.</td>
</tr>
<tr>
<td>0</td>
<td>131</td>
<td>75.47</td>
<td>130.64</td>
</tr>
<tr>
<td>1</td>
<td>38</td>
<td>83.83</td>
<td>42.52</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>46.50</td>
<td>21.96</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>17.17</td>
<td>12.74</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4.75</td>
<td>7.79</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1.05</td>
<td>4.91</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0.19</td>
<td>3.16</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.03</td>
<td>2.06</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>-</td>
<td>1.36</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>-</td>
<td>0.90</td>
</tr>
<tr>
<td>Ialt</td>
<td>229</td>
<td>112.68</td>
<td>6.14</td>
</tr>
</tbody>
</table>

Man finder i alt $z_+ = \sum_{i=1}^{229} z_i = 254$ afvигende låg i de 229 prøver, dvs $\bar{z}_+ = 254/229 = 1.11$ afvигende/pr prøve, hvorfor man har $\bar{h}_+ = 1.11/770 = 0.0014$ afvигende/pr låg. Endvidere finder man

$$s_+^2 = \sum_{i=1}^{229} 770(h_i - \bar{h}_+)^2/228 = 0.00449.$$  

Under den systematiske model finder man Pearson-teststørrelsen for samme andel afvигende i de 229 partier $X^2 = 711.71$ med $f = 228$. Idet $E[\chi^2(f)] = f$, og $V[\chi^2(f)] = 2f$, finder man, at den observerede værdi af $X^2$ er væsentligt større, end

$$\chi^2(228)_{0.95} \approx E[\chi^2(228)] + 1.64\sqrt{V[\chi^2(228)]} = 228 + 35 = 263.$$  

Der synes således ikke at være belæg for at modellere observationerne som realisationer af samme binomialfordeling.

Vi bemærker dog, at approximationen ved $\chi^2$-fordelingen er ganske ringe. $X^2$-teststørrelsen består af 229 led af formen $(z_i - 1.11)^2/1.11$ og $\chi^2$-approximationen fremkommer essentielt ved at approximere hvert led med en $N(0, 1)^2$-størrelse. Men de mulige værdier af $z_i$ er 0, 1, ..., 770 med hovedvægten på værdier i nærheden af 1.11, dvs. $z_i = 0$ og $z_i = 1$, hvorfor fordelingen af $X^2$ vil have størstedelen af massen placeret i punkterne...
Hierarkiske modeller for eksp. disp. modeller

\[(0 - 1.11)^2 / 1.11, (1 - 1.11)^2 / 1.11 \text{ og } (2 - 1.11)^2 / 1.11. \] Det er således en grov tilnærmelse at approximere denne fordeling med den kontinuerle \( \chi^2 \)-fordeling.

Havde vi i stedet bestemt kvotientteststørrelsen (6.2.1), havde vi fundet størrelsen \( G^2(H_f) = 273.4 \). Også denne størrelse indikerer en signifikant afvigelse, omend ikke så markant.

Figur 6.1 viser den observerede fordeling af \( Z \) sammenlignet med frekvensfunktionen for en \( B(770, 0.0014) \)-fordeling. Det fremgår klart af figuren, at den observerede fordeling har tykkere haler, end binomialfordelingen.
Figur 6.1. Den observerede fordeling af antallet af afvigende låg i 229 stikprøver af 770 låg sammenlignet med en B(770, 0.0014) fordeling.
I tabellen over observationerne er endvidere angivet de forventede antal stikprøver med henholdsvis 0, 1, 2, ... afvigende låg, beregnet under binomialfordelingsforudsætningen. Det ses, at det sædvanlige $\chi^2$-test for fordelingstype (Statistik 1, afsnit 4.2.2) fører til $\chi^2 = 112.68$ med $f = 3$. Det må således klart afvises, at de 229 observationer kan beskrives som ensfordelte binomialfordelte størrelser.

Det er derfor naturligt at benytte en model med en andel afvigende enheder, der varierer tilfældigt mellem partier. Vælger man at beskrive variationen ved en $\text{Be}(\alpha, \beta)$-fordeling, vil den marginale fordeling af $Z$ (jvf. sætn. 6.2.2) kunne beskrives ved en $\text{Pl}(770, \alpha, \alpha + \beta)$ fordeling.

Momentestimaterne for $\alpha$ og $\beta$ findes af sætning 6.2.3 til $\tilde{\alpha} = 0.5095$ og $\tilde{\beta} = 363.42$, svarende til $\tilde{\pi} = 0.0014$ og $\tilde{\gamma} = 0.00275$.

Ved benyttelse af en numerisk maksimeringsroutine finder man maksimum likelihood estimaterne $\hat{\alpha} = 0.459$ og $\hat{\beta} = 318.64$ svarende til $\hat{\pi} = 0.0014$ og $\hat{\gamma} = 0.00313$. 
Figur 6.2. Den observerede fordeling af antallet af afvigende låg i 229 stikprøver á 770 låg sammenlignet med en Pl(770,0.459,319.1) fordeling.
Figur 6.2 viser den observerede fordeling af $Z$ sammenlignet med frekvens-
funktionen for en $\text{P}(770,0.459,319.1)$-fordeling. Der ses at være en væsent-
lig bedre tilpasning ved Polyafordelingen, end ved binomialfordelingen.

I tabellen over observationerne er endvidere angivet det forventede antal
stikprøver med henholdsvis $0,1,2,...$ dimensionsafvigende, beregnet under
Polyafordelingsforudsætningen. Det ses, at det sædvanlige $\chi^2$- test for for-
delingstype fører til $\chi^2 = 6.14$ med $f = 4$. Testet giver således ingen grund
til at afvise, at de 229 observationer kan beskrives som ensfordelte Polyafor-
delte størrelser, og modellen med den tilfældige variation imellem partiernes
defektandele opretholdes derfor.
6.3 Den geometriske fordeling

fil: hiergen2.tex 1997-04-20

Såfremt $X_{ij} | p_i \in \text{Geo}^*(p_i)$, finder vi for $Z_i = X_{i1} + \cdots + X_{in_i}$, at $Z_i | p_i \in \text{NB}^*(n_i, p_i)$.

Familien af negative binomialfordelinger er en additiv eksponentiel dispersionsmodel med enhedsmiddelværdi

$$\mu = \mathbb{E}[X_{ij} | p] = \frac{p}{1 - p}, \quad (6.3.1)$$

dvs

$$p = \frac{\mu}{1 + \mu}. \quad (6.3.2)$$

Familien har enhedsvariansfunktionen

$$V_{NB}(\mu) = \mu(1 + \mu) = \frac{p}{(1 - p)^2} \quad (6.3.3)$$

Den kanoniske linkfunktion er

$$\eta(\mu) = \ln \left( \frac{\mu}{1 + \mu} \right) \quad (6.3.4)$$

For $Z \in \text{NB}^*(n, p)$ gælder

$$\mathbb{E}[Z] = n\mu$$

$$\mathbb{V}[Z] = nV_{NB}(\mu) = n\mu(1 + \mu)$$

For $Y = Z/n$ har vi derfor:

$$\mathbb{E}[Y] = \mu = \frac{p}{1 - p}$$

$$\mathbb{V}[Y] = \frac{V_{NB}(\mu)}{n} = \frac{\mu(1 + \mu)}{n}$$

med $\mu$ og $V_{NB}(\mu)$ givet ved (6.3.1) og (6.3.3)
Sætning 6.3.1 Test for homogenitet

Kvotientteststørrelsen for hypotesen (6.1.5) er

\[ G^2(H_I) = \sum_{i=1}^{k} n_i d(y_i; \bar{y}_+) \]  

(6.3.5)

med \( y_i = z_i / n_i (= x_{i+}) \) og med enhedsdeviansen \( d(y_i; \bar{y}_+) \) givet ved

\[ d(y_i; \bar{y}_+) = 2[y_i \ln(y_i / \bar{y}_+) - (y_i + 1) \ln[(1 + y_i) / (1 + \bar{y}_+)] \]

Under hypotesen (6.1.5) vil \( G^2(H_I) \) asymptotisk være fordelt som \( \chi^2(k-1) \). Hypotesen forkastes for store værdier af \( G^2(H_I) \).

Bevis:
Testet er det sædvanlige homogenitetstest i en generaliseret lineær model.

\[ \Box \]

**Remærkning 1 Pearson-teststørrelsen**

Pearson-teststørrelsen for hypotesen (6.1.5) er

\[ X^2 = \sum_{i=1}^{k} n_i \frac{(y_i - \bar{y}_+)^2}{\bar{y}_+(1 + \bar{y}_+)} \]  

(6.3.6)

Testet er asymptotisk ækvivalent med kvotienttestet. Testet måler variationen imellem gruppegennemsnittene, \( SAK_2 \), i forhold til den estimerede varians, \( V_{NB}(\bar{y}_+) \) i den negative binomialfordeling.

\[ \Box \]

Såfremt man ønsker at modellere eventuelle forskelle mellem grupperne ved en tilfældig model, kan man vælge at modellere fordelingen af \( p \) ved en \( Be(\alpha, \beta) \)-fordeling. Den marginale fordeling af \( Z_i \) bliver da en \( NPl(n_i, \beta, \alpha+\beta) \)-fordeling.

**Lemma 6.3.1 Momenter i fordelingen af \( \mu = p/(1 - p) \) ved betanegativ binomial sampling**

Såfremt \( Z|p \in NB^*(n, p) \) og \( p \in Be(\alpha, \beta) \), da gælder for fordelingen af \( \mu = p/(1 - p) \):
6.3 Den geometriske fordeling

For $\beta \leq 1$ har fordelingen ikke nogen middelværdi. Såfremt $1 < \beta$, har fordelingen af $\mu$ middelværdien

$$\psi = E [\mu] = E \left[ \frac{p}{1-p} \right] = \frac{\alpha}{\beta - 1} \quad (6.3.7)$$

For $\beta \leq 2$ har fordelingen ingen varians. Såfremt $2 < \beta$, har fordelingen af $\mu$ variansen

$$V [\mu] = \frac{V_{NB}(\psi)}{\beta - 2} = \frac{\gamma}{1 - \gamma} V_{NB}(\psi) \quad (6.3.8)$$

Den marginale middelværdi af variansfunktionen $V_{NB}(\mu)$ er

$$E [V_{NB}(\mu)] = \frac{\beta - 1}{\beta - 2} V_{NB}(\psi) = \frac{1}{1 - \gamma} V_{NB}(\psi) \quad , \quad (6.3.9)$$

hvor "signal/støj forholdet," $\gamma$, eksisterer såfremt $2 < \beta$. Signal/støjforholdet $\gamma$ er givet ved

$$\gamma = \frac{V [\mu]}{E [V_{NB}(\mu)]} = \frac{1}{\beta - 1} \quad (6.3.10)$$

med $0 < \gamma < 1$.

Endelig er intraklassekorrelationen, $\rho$, givet ved

$$\rho = \frac{V [\mu]}{E [V_{NB}(\mu)] + E [V [\mu]]} = \frac{1}{\beta} \quad (6.3.11)$$
for $2 < \beta$.

**Bevis:**
Ved integration i betafordelingen finder man

$$V[\mu] = V\left[ \frac{p}{1-p} \right] = \frac{\alpha}{\beta-1} \left( 1 + \frac{\alpha}{\beta-1} \right) \frac{1}{\beta-2} = \frac{\psi(1+\psi)}{\beta-2} ,$$

der ved indsættelse af $\psi(1+\psi) = V_{NB}(\psi)$ (jvf(6.3.3)), netop er (6.3.8).

Tilsvarende finder man

$$E[V_{NB}(\mu)] = E[\mu(1+\mu)] = \frac{\alpha}{\beta-2} \left( 1 + \frac{\alpha}{\beta-1} \right) = \frac{\beta-1}{\beta-2} \psi(1+\psi) .$$

Idet vi atter udnytter relationen $\psi(1+\psi) = V_{NB}(\psi)$ (jvf (6.3.3)), fås (6.3.9).

Endelig fås udtrykket for intraklassekorrelationen $\rho$ ved indsættelse af udtrykket for $\gamma$ i (6.1.11).

\[\square\]

**Sætning 6.3.2** **Den marginale fordeling af gruppetotal og gruppegennemsnit ved negativ binomial-beta sampling**

Såfremt $Z|p \in NB^*(n,p)$, og $p \in Be(\alpha, \beta)$, da er den marginale fordeling af $Z$ en NPI$^*(n,\alpha,\alpha+\beta)$-fordeling.

Den negative polyafordeling, NPI$(n,\beta,\alpha+\beta)$-fordelingen, er beskrevet i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

Såfremt $\beta > 2$ gælder:

\[
E[Z] = n\psi \quad (6.3.12)
\]
\[
V[Z] = n \frac{\psi(1+\psi)}{1-\gamma}[1+n\gamma] \quad (6.3.14)
\]

med $\psi$ og $\gamma$ givet ved (6.3.7) og (6.3.10).
6.3 Den geometriske fordeling

Middelværdien \( \mathbb{E} [Z] \) eksisterer dog blot \( \beta > 1 \).

For den relative hyppighed \( Y = Z/n \) gælder tilsvarende for \( \beta > 2 \):

\[
\begin{align*}
\mathbb{E} [Y] & = \psi \\
\mathbb{V} [Y] & = \frac{\psi (1 + \psi)}{1 - \gamma} \left[ \gamma + \frac{1}{n} \right]
\end{align*}
\]

(6.3.15) (6.3.16)

Bevis:
Udtrykket for momenterne fås ved indsættelse af resultaterne fra lemma 6.3.1 i det generelle udtryk (6.1.14) i sætning 6.1.2.

\[ \square \]

Bemærkning 1 Oppsættning af den totale varians

Oppsætningen af den totale varians (6.3.14)

\[
\mathbb{V} [Z] = n \frac{\mathbb{V}_{NB} (\psi)}{1 - \gamma} [1 + n\gamma]
\]

i komponenter svarende til den gennemsnitlige varians indenfor grupper og variansen mellem gruppemiddelværdier, giver størrelserne

\[
\begin{align*}
\mathbb{E} [\mathbb{V} [Z|\mu]] & = n \mathbb{E} [\mathbb{V}_{NB} (\mu)] = n \frac{\mathbb{V}_{NB} (\psi)}{1 - \gamma} \\
\mathbb{V} [\mathbb{E} [Z|\mu]] & = n^2 \mathbb{V} [\mu] = n^2 \gamma \frac{\mathbb{V}_{NB} (\psi)}{1 - \gamma}
\end{align*}
\]

\[ \square \]

Bemærkning 2 Udvidelse af parameterområdet

Betafordelingen er defineret for \( 0 < \alpha \) og \( 0 < \beta \). For \( \alpha < 1 \) vil tætheden i fordelingen af \( p \) være J-formet med modus for \( p = 0 \). For \( \beta < 1 \) vil tætheden være J-formet med modus for \( p = 1 \). For \( 1 < \alpha \) og \( 1 < \beta \) vil \( p \)
tætheden være unimodal med modus i \( p = (\alpha - 1)/(\alpha + \beta - 2) \) (svarende til \( p = 1/(1 + \psi - \gamma) \)).

Parametrisingen ved \( \psi \) og \( \gamma \) svarer til området \( \{2 < \alpha\} \times \{2 < \beta\} \). Parametrisingen kan formelt udvides til også at omfatte intervallerne \( \beta \in ]0, 1[ \cup ]1, 2[ \), med billedmængden \( \gamma \in ]-\infty, -1[ \cup ]1, \infty[ \), men parametrisingen er ikke differentiabel i hele parameterrummet \( \alpha > 0, \beta > 0 \), og den vil derfor give anledning til problemer ved estimation i nærheden af polerne \( \beta = 1 \) og \( \beta = 2 \). Det er endvidere klart, at ved denne udvidelse mister man fortolkningen af \( \gamma \) som en kvotient mellem varianser.

\[ \square \]

**Bemærkning 3**  *Parametrising ved momenter i betafordelingen*

I lighed med det parametrisingen ved beta-binomial fordelingen kunne man betragte parametrisingen

\[ \pi = \alpha / (\alpha + \beta), \quad \gamma^* = 1 / (\alpha + \beta) \]  \[ (6.3.17) \]

hvori parameteren \( \pi \) angiver \( E[p] \), og \( \gamma^* \) er bestemt ved

\[ \gamma^* = \frac{V[p]}{E[p(1-p)]} \]

Der gælder

\[ \psi = \frac{1 - \pi}{1 - \pi - \gamma^*} \quad \gamma = \frac{\gamma^*}{1 - \pi - \gamma^*} \]

\[ \pi = \frac{1 + \psi}{1 + \psi + \gamma} \quad \gamma^* = \frac{1}{1 + \psi + \gamma} \]

\[ \square \]

**Bemærkning 4**  *Overdispersion i forhold til den negative binomialfordeling*

For begrænsede værdier af \( n \) vil fordelingen af \( Z \) have en overdispersion i forhold til den negative binomialfordeling med samme antalsparameter og med samme middelværdi, dvs med sandsynlighedsparameter \( p = \psi / (1 + \psi) \).
Overdispersionen er

\[ \sigma^2 = \frac{\text{V}[Z]}{nV_{NB}(\psi)} = \frac{1 + n\gamma}{1 - \gamma} \]  \hspace{1cm} (6.3.18)

Tilsvarende vil fordelinger af \( Y = Z/n \) have overdispersionen \( \sigma^2 = (1 + n\gamma)/(1 - \gamma) \) i forhold til \( \text{NB}^*(n, \psi/(1 + \psi))/n \)-fordelingen.

Den marginale varians af den gennemsnitlige rate \( Y \) kan altså udtrykkes som

\[ \text{V}[Y] = V_{NB}(\psi)/n_{eff} \]

hvori den effektive stikprøvestørrelse, \( n_{eff} \) er bestemt ved

\[ n_{eff} = \frac{n}{\sigma^2} = n \frac{1 - \gamma}{1 + n\gamma} \]  \hspace{1cm} (6.3.19)

For \( \psi \) fastholdt, \( 0 < \psi < \infty \) og \( \gamma \rightarrow 0 \) vil \( \text{V}[Z] \) nærme sig \( n\psi(1 + \psi) \), der netop er variansen i en \( \text{NB}^*(n, \psi/(1 + \psi)) \)-fordeling, svarende til at fordelingen af \( Z \) nærmer sig en \( \text{NB}^*(n, \psi/(1 + \psi)) \)-fordeling.

\[ \square \]

Sætning 6.3.3 Maksimum-likelihood estimation i den negative Polyafordeling

\[ Er \ endnu \ ikke \ udarbejdet \]

\[ \square \]

Bemærkning 1 Bestemmelse af maksimum-likelihood estimatorne

\[ Er \ endnu \ ikke \ udarbejdet \]

\[ \square \]

Bemærkning 2 Khotientteststørrelsen for homogenitetstest

\[ Er \ endnu \ ikke \ udarbejdet \]

\[ \square \]
Sætning 6.3.4 Momentestimation i den negative Polyafordeling

Lad \( Z_1, Z_2, \ldots, Z_k \) være uafhængige variable, hvor \( Z_i \in \mathcal{NP}^+(n_i, \alpha, \alpha + \beta) \). Momentestimatene for \( \psi = \alpha / (\beta - 1) \) og \( \gamma = 1 / (\beta - 1) \) er da

\[
\begin{align*}
\tilde{\psi} &= \bar{y}_+ \\
\tilde{\gamma} &= \frac{s_2^2 - \bar{y}_+(1 + \bar{y}_+)}{s_2^2 + n_0 \bar{y}_+(1 + \bar{y}_+)}
\end{align*}
\]

(6.3.20) (6.3.21)

med

\[
s_2^2 = \frac{1}{k} \sum_{i=1}^{k} n_i (y_i - \bar{y}_+)^2 / (k - 1),
\]

hvor \( y_i = z_i / n_i \), og hvor den vægtede gennemsnitlige stikprøvestørrelse, \( n_0 \), er bestemt ved (5.1.9).

De tilsvarende estimatorer for \( \alpha \) og \( \beta \) er

\[
\begin{align*}
\tilde{\alpha} &= \bar{y}_+ \left( 1 + \frac{n_0 + 1}{s_2^2 - \bar{y}_+(1 + \bar{y}_+)} \right) \\
\tilde{\beta} &= 2 + \frac{n_0 + 1}{s_2^2 - \bar{y}_+(1 + \bar{y}_+)}
\end{align*}
\]

(6.3.22) (6.3.23)

Bevis:
Følger af sætning 6.1.3.

Bemærkning 1 Singularitet ved momentestimation

Hvis \( s_2^2 < \bar{y}_+(1 + \bar{y}_+) \) bliver \( \tilde{\gamma} < 0 \). Dette kan tages som udtryk for enten, at \( \gamma = 0 \) svarende til en etpunktsfordeling af \( \mu \) (og \( p \)), eller at \( \tilde{\alpha} < 2 \) svarende til at fordelingen af \( \mu \) ikke har nogen varians. Det er nødvendigt at foretage en nøjere analyse af data, evt. en maksimum-likelihood estimation, for at skelne mellem disse to situationer.
6.4 Poissonfordelingen

Såfremt $X_{ij} | \mu_i \in P(\mu_i)$ finder vi for $Z_i = X_{i1} + \cdots + X_{in_i}$, at $Z_i | \mu_i \in P(n_i \mu_i)$.

Familien af Poissonfordelinger er en additit eksponentiel dispersionsmodel med middelværdi

$$\mu = \mathbb{E}[X_{ij} | \mu]$$ (6.4.1)

Familien har enhedsvariansfunktionen

$$V_P(\mu) = \mu$$ (6.4.2)

Den kanoniske linkfunktion er

$$\eta(\mu) = \ln(\mu)$$ (6.4.3)

For $Z \in P(n\mu)$ gælder

$$\mathbb{E}[Z] = n\mu$$
$$V[Z] = nV_P(\mu) = n\mu$$

For $Y = Z/n$ har vi derfor:

$$\mathbb{E}[Y] = \mu$$
$$V[Y] = \frac{V_P(\mu)}{n} = \frac{\mu}{n}$$

Sætning 6.4.1 Test for homogenitet

Kvotientteststørrelsen for hypotesen (6.1.5) er

$$G^2(H_I) = 2 \sum_{i=1}^{k} n_i \left[ y_i \ln\left(\frac{y_i}{\bar{y}_+}\right) - (y_i - \bar{y}_+) \right]$$ (6.4.4)

med $y_i = z_i/n_i \ (= \bar{x}_{i+})$. 
Kvotientteststørrelsen svarer til de vægtede enhedsdevianser (med vægtene $n_i$)
$$d(y_i, \bar{y}_+) = 2n_i \left[ y_i \ln(y_i/\bar{y}_+) - (y_i - \bar{y}_+) \right]$$

Under hypotesen (6.1.5) vil $G^2(H_I)$ asympotisk være fordelt som $\chi^2(k-1)$. Hypotesen forkastes for store værdier af $G^2(H_I)$.

**Bevis:**
Testet er det sædvanlige homogenitetstest i en generaliseret lineær model. □

**Bemærkning 1 Pearson-teststørrelsen**

Pearson-teststørrelsen for hypotesen (6.1.5) er

$$X^2 = \sum_{i=1}^{k} n_i \frac{(y_i - \bar{y}_+)^2}{\bar{y}_+} \quad f = k - 1 \quad (6.4.5)$$

Testet er asympotisk ækvivalent med kvotienttestet. Dette test måler direkte variationen imellem grupper, $SAK_2$, i forhold til den estimerede varians, $V_P(\bar{y}_+) = \bar{y}_+$. (Ref.: P.V. Sukhatme (1938)).

Testet fås direkte af kvotienttestet ved at benytte, at $\ln x \leq x - 1$ for $x > 0$, hvor lighedstegnet kun gælder for $x = 1$. Man har derfor

$$\frac{a - b}{a} \leq \ln\left(\frac{a}{b}\right) \leq \frac{a - b}{b}$$

med lighedstegn gyldigt for $a = b$. Approximerer man $\ln(a/b)$ med gennemsnittet af øvre og nedre grænse finder man

$$\ln\left(\frac{a}{b}\right) \approx \frac{a^2 - b^2}{2ab} \quad (6.4.6)$$

Benyttes approximationen (6.4.6) i udtrykket (6.3.6) for $Z$, finder vi Pearson-teststørrelsen (6.4.5). □

Såfremt man ønsker at modellere eventuelle forskelle mellem grupperne ved en tilfældig model, kan man vælge at beskrive fordelingen af $\mu$ ved en $G(\alpha, 1/\beta)$-fordeling.
Lemma 6.4.1 Momenter i fordelingen af μ ved gamma-Poisson sampling

Så fremt $Z|μ \in P(nμ)$ og $μ \in G(α, 1/β)$, da gælder for fordelingen af $μ$:

\begin{align*}
m &= E[μ] = \frac{α}{β} \quad \text{(6.4.7)} \\
V[μ] &= \frac{V_P(m)}{β} = γ V_P(m) \quad \text{(6.4.8)}
\end{align*}

Den marginale middelværdi af variansfunktionen $V_P(μ)$ er

\begin{align*}
E[V_P(μ)] &= V_P(m) = m \quad \text{(6.4.9)}
\end{align*}

hvor "signal/støj forholdet," $γ$, er givet ved

\begin{align*}
γ &= \frac{V[μ]}{E[V_P(μ)]} = \frac{1}{β} \quad \text{(6.4.10)}
\end{align*}

med $0 < γ < ∞$.

Endelig er intraklassekorrelationen, $ρ$, givet ved

\begin{align*}
ρ &= \frac{V[μ]}{E[V[μ]]} = \frac{1}{1 + β} \quad \text{(6.4.11)}
\end{align*}

Bevis:
Beviset følger ved at bemærke, at der for $\mu \in G(\alpha, 1/\beta)$ gælder

$$E[\mu] = \frac{\alpha}{\beta}$$

$$V[\mu] = \frac{\alpha}{\beta^2}$$

\[\square\]

Sætning 6.4.2 Den marginale fordeling af gruppetotal og gruppegennemsnit ved gamma- Poisson sampling

Såfremt $Z|\mu \in P(n\mu)$ og $\mu \in G(\alpha, 1/\beta)$, da er den marginale fordeling af $Z$ en $NB(\alpha, \beta/(\beta + n))$-fordeling.

Der gælder

$$E[Z] = nm$$  \hspace{1cm} (6.4.12)

$$V[Z] = nm(1 + n\gamma)$$  \hspace{1cm} (6.4.13)

med $m$ og $\gamma$ givet ved (6.4.7) og (6.4.10).

For den relative hyppighed $Y = Z/n$ gælder tilsvarende

$$E[Y] = m$$  \hspace{1cm} (6.4.14)

$$V[Y] = m \left[ \gamma + \frac{1}{n} \right]$$  \hspace{1cm} (6.4.15)

Bevis:
Udtrykket for momenterne fås ved indsættelse af resultaterne fra lemma
6.4 Poissonfordelingsmodeller

6.4.1 i det generelle udtryk (6.1.14) i sætning 6.1.2.

Bemærkning 1 Opspaltning af den totale varians

Opspaltningen af den totale varians (6.4.13)

\[ V[Z] = n \cdot V_P(m) \cdot [1 + n\gamma] \]

i komponenter svarende til den gennemsnitlige varians indenfor grupper og variansen mellem gruppegennemsnitværdier, giver størrelserne

\[ E[V[Z|\mu]] = n \cdot E[V_P(\mu)] = n \cdot m \]
\[ V[E[Z|\mu]] = n^2 \cdot V[\mu] = n^2 \cdot \gamma \cdot V_P(m) \]

Bemærkning 2 Overdispersion i forhold til Poissonfordelingen

For begrænsede værdier af \( n \) vil fordelingen af \( Z \) have en overdispersion i forhold til Poissonfordelingen \( P(n \cdot m) \) med samme antalsparameter \( n \) og med middelværdiparameteren \( m \).

Overdispersionen er

\[ \delta = \frac{V[Z]}{n \cdot V_P(m)} = 1 + n\gamma \]  \hspace{1cm} (6.4.16)

Tilsvarende vil fordelingen af \( Y = Z/n \) have overdispersionen \( \delta = (1 + n\gamma) \) i forhold til \( P(n \cdot m)/n \)-fordelingen.

Den marginale varians af den gennemsnitlige rate \( Y \) kan altså udtrykkes som

\[ V[Y] = V_P(m)/n_{eff} \]

hvor den effektive stikprøvestørrelse, \( n_{eff} \) er bestemt ved

\[ n_{eff} = \frac{n}{\delta} = \frac{n}{1 + n\gamma} \]

svarende til
\[
\frac{1}{n_{\text{eff}}} = \frac{1}{n} + \gamma
\]  
(6.4.17)

For \( Y \in \text{NB}(\alpha, \beta/(\beta + n))/n \) gælder altsaa

\[
\mathbb{V}[Y] = \frac{m}{n_{\text{eff}}}
\]

med \( m \) og \( \gamma \) bestemt ved (6.4.7) og (6.4.10).

Variansen for \( Y \) er således den samme som variansen for gennemsnittet af \( n_{\text{eff}} \) uafhængige målinger af en \( P(m) \) fordelt variabel. Såfremt man kender parameteren \( \gamma \), kan eksempelvis approximative konfidensintervaller for \( m \) altså bestemmes som konfidensintervaller for en tilsvarende Poissonfordelt størrelse med stikprøvestørrelsen \( n_{\text{eff}} \).

Parameteren \( \gamma \) udtrykker afvigelsen fra den rene Poissonfordeling af \( Z \). For \( m > 0 \) og \( \gamma \to 0 \) vil \( \mathbb{V}[\mu] \to 0 \), og fordelingen af \( Z \) vil nærme sig en \( P(nm) \)-fordeling.

\[\Box\]

Sætning 6.4.3 Momentestimation i den negative binomialfordeling

Lad \( Z_1, Z_2, \ldots, Z_k \) være uafhængige variable, hvor \( Z_i \in \text{NB}(\alpha, \beta/(\beta + n_i)) \)

Momentestimaterne for \( m = \alpha/\beta \) og \( \gamma = 1/\beta \) er da

\[
\bar{m} = \bar{y}_+ \quad \text{ (6.4.18)}
\]

\[
\tilde{\gamma} = \frac{s_2^2 / \bar{y}_+ - 1}{n_0} \quad \text{ (6.4.19)}
\]

med

\[
s_2^2 = \sum_{i=1}^{k} n_i(y_i - \bar{y}_+)^2 / (k - 1) ,
\]
hvor \( y_i = z_i / n_i \), og hvor den vægtede gennemsnitlige stikprøvestørrelse, \( n_0 \), er bestemt ved (5.1.9).

Momentestimatorerne for \( \alpha \) og \( \beta \) bliver derfor

\[
\hat{\alpha} = n_0 \frac{\bar{y}_+^2}{s_2^2 - \bar{y}_+} \quad (6.4.20)
\]

\[
\hat{\beta} = n_0 \frac{\bar{y}_+}{s_2^2 - \bar{y}_+} \quad (6.4.21)
\]

**Bevis:**
Resultatet følger umiddelbart ved identifikation af momenterne.

\[\square\]

**Bemærkning 1 Singulariteter ved momentestimationen**

Hvis \( s_2^2 < \bar{y}_+ \) bliver \( \bar{y}, \hat{\alpha} \) og \( \hat{\beta} \) negative. I dette tilfælde vil det være naturligt at sætte disse størrelser til nul, d.v.s. at fordelingen af \( \mu \) estimeres til at være en etpunktfordeling, og fordelingen af \( Z_i \) bliver da en \( P(n_i m) \)-fordeling. Parametreringen ved \( (m, \gamma) \) tillader netop estimation af \( m \), selv i tilfældet \( \gamma = 0 \).

\[\square\]

**Sætning 6.4.4 Maksimum-likelihood estimation i den negative binomialfordeling**

Lad \( Z_1, Z_2, \ldots, Z_k \) være uafhængige variable, hvor \( Z_i \in NB(\alpha, \beta/(\beta + n_i)) \).

Maksimum-likelihood estimaterne \((\hat{m}, \hat{\gamma})\) for \( m = \alpha / \beta \) og \( \gamma = 1 / \beta \) findes da ved at maksimere

\[
l(m, \gamma; z_1, z_2, \ldots, z_k) = \sum_{i=1}^{k} \sum_{\nu=0}^{z_i-1} \ln(m + \nu \gamma) - \sum_{i=1}^{k} \left( \frac{m}{\gamma} + z_i \right) \ln(1 + n_i \gamma) \quad (6.4.22)
\]

med hensyn til \( m \) og \( \gamma \).

For \( \gamma = 0 \) fortolkes udtrykket for \( l \) som

\[
l(m, 0; z_1, z_2, \ldots, z_k) = \ln(m) \sum_{i=1}^{k} z_i - m \sum_{i=1}^{k} n_i
\]
Bevis:
Sætningen vises ved at bemærke, at - på nær en konstant - er logaritmen til likelihoodfunktionen givet ved (6.4.22). Udtrykket for $\gamma = 0$ fås ved grænseovergang i (6.4.22).

\[ \frac{\partial l}{\partial m} = \sum_{i=1}^{k} \sum_{\nu=0}^{z_i-1} \frac{1}{m + \nu \gamma} - \sum_{i=1}^{k} \frac{1}{\gamma} \ln(1 + n_i \gamma) \]

\[ \frac{\partial l}{\partial \gamma} = \sum_{i=1}^{k} \left[ \frac{m}{\gamma^2} \ln(1 + n_i \gamma) - \frac{m / \gamma + z_i}{1 + n_i \gamma} + \sum_{\nu=0}^{z_i-1} \frac{\nu}{m + \nu \gamma} \right] \]

Såfremt maksimum findes i et indre punkt, fås maksimum-likelihood estimatorne ved at sætte scorefunktionen lig nul og løse ligningerne med hensyn til parametrene $m$ og $\gamma$.

Ligningerne må løses iterativt. Som udgangsværdier for iterationen kan benyttes momentestimatorne $\hat{m}$ og $\hat{\gamma}$ bestemt ved (6.4.18) og (6.4.19).

Såfremt $s_2^2 < \bar{y}_+$, må man formode, at maksimum ligger på randen af området, svarende til $\gamma = 0$. I det balancerede tilfælde kan det vises, at $s_2^2 < \bar{y}_+ \Rightarrow \hat{\gamma} = 0$. (Se W. Simonsen (1976, 1980) og J. Bonitzer (1978)).

For $\gamma = 0$ er maksimum-likelihood estimatet for $m$ bestemt ved $\hat{m} = \bar{y}_+$.

\[ \frac{\partial l}{\partial \gamma} = \sum_{i=1}^{k} \left[ \frac{m}{\gamma^2} \ln(1 + n_i \gamma) - \frac{m / \gamma + z_i}{1 + n_i \gamma} + \sum_{\nu=0}^{z_i-1} \frac{\nu}{m + \nu \gamma} \right] \]

Bemærkning 1 Bestemmelse af maksimum-likelihood estimatorne

Komponenterne af scorefunktionen er

\[ \frac{\partial l}{\partial m} = \sum_{i=1}^{k} \sum_{\nu=0}^{z_i-1} \frac{1}{m + \nu \gamma} - \sum_{i=1}^{k} \frac{1}{\gamma} \ln(1 + n_i \gamma) \]

\[ \frac{\partial l}{\partial \gamma} = \sum_{i=1}^{k} \left[ \frac{m}{\gamma^2} \ln(1 + n_i \gamma) - \frac{m / \gamma + z_i}{1 + n_i \gamma} + \sum_{\nu=0}^{z_i-1} \frac{\nu}{m + \nu \gamma} \right] \]

Såfremt maksimum findes i et indre punkt, fås maksimum-likelihood estimatorne ved at sætte scorefunktionen lig nul og løse ligningerne med hensyn til parametrene $m$ og $\gamma$.

Ligningerne må løses iterativt. Som udgangsværdier for iterationen kan benyttes momentestimatorne $\hat{m}$ og $\hat{\gamma}$ bestemt ved (6.4.18) og (6.4.19).

Såfremt $s_2^2 < \bar{y}_+$, må man formode, at maksimum ligger på randen af området, svarende til $\gamma = 0$. I det balancerede tilfælde kan det vises, at $s_2^2 < \bar{y}_+ \Rightarrow \hat{\gamma} = 0$. (Se W. Simonsen (1976, 1980) og J. Bonitzer (1978)).

For $\gamma = 0$ er maksimum-likelihood estimatet for $m$ bestemt ved $\hat{m} = \bar{y}_+$.

Bemærkning 2 Kvantienttest af hypotesen $\gamma = 0$

Kvotienttestet for hypotesen

\[ H_{II} : \gamma = 0 \quad \text{med alternativet} \quad \bar{H}_{II} : \gamma > 0 \]

under den tilfældige model, har teststørrelsen

\[ Z = 2 \left[ \sum_{i=1}^{k} \sum_{\nu=0}^{z_i-1} \ln \left( \frac{\hat{m} + \nu \hat{\gamma}}{\bar{y}_+} \right) - \sum_{i=1}^{k} \left( \frac{\hat{m}}{\hat{\gamma}} + t_i \right) \ln(1 + n_i \hat{\gamma}) \right] \]

(6.4.23)
hvor \( \hat{m} \) og \( \hat{\gamma} \) angiver maksimum-likelihood estimerne fundet i sætning 6.4.4

Under \( H_{II} \) vil \( Z \) approximativt følge en \( \chi^2(1) \)-fordeling.

**Eksempel 6.4.1 Variation mellem episoder af tordenvejr ved Cape Kennedy**

Data fra Williford et al. (1974).


**Fordelingen af antal dage med 0,1,2 eller flere episoder af tordenvejr ved Cape Kennedy**

<table>
<thead>
<tr>
<th>Antal episoder ( z_i )</th>
<th>Antal dage ( # i )</th>
<th>Poisson forventet</th>
<th>Negativ binomial forventet</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>803</td>
<td>791.85</td>
<td>802.92</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>118.78</td>
<td>100.15</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>8.91</td>
<td>14.38</td>
</tr>
<tr>
<td>3+</td>
<td>3</td>
<td>0.46</td>
<td>2.55</td>
</tr>
</tbody>
</table>

Alle observationsperioder er på \( n_i = 1 \) dag.

Man finder gennemsnittet, \( \bar{y}_+ = 0.15 \) tordenvejr/dag. Det fremgår klart af tabellen, at den observerede fordeling har en større spredning (tykkere haller), end den tilsvarende Poissonfordeling. Det er derfor naturligt at antage, at antallet af tordenvejr på en given dag varierer som en Poisson fordelt variabel, men at Poisson-intensiteten varierer mellem dagene.

Da \( \bar{y}_+ = 0.15 \) og \( s^2_2 = 0.18 \), finder man momentestimaterne \( \hat{m} = 0.15 \) [tordenvejr/dag] og \( \hat{\gamma} = 0.2 \). Disse estimerer afviger ikke meget fra maksimum-likelihood estimerne, der bestemmes ved en numerisk maksimeringsroutine til at være \( \hat{m} = 0.1489 \) [tordenvejr/dag] og \( \hat{\gamma} = 0.1939 \). Det ses af tabellen, at den negative binomialfordeling med disse parametre giver en meget fin tilpasning til data. Man finder den sædvanlige \( \chi^2 \) -test størrelse for fordelingstype (statistik 1, afsnit 4.2.2) til \( \chi^2 = 0.09 \) med 1 frihedsgrad.
Resultatet er ikke overraskende, da det er velkendt, at meteorologiske fænomener ikke er spredt jævnud over tidsaksen, men at der er en vis træghed i disse fænomener, der bevirker, at tordenvejr ofte kommer i "stimer". Vi bemærker, at modelleringen ved den negative binomialfordeling ikke tilgodeser denne autokorrelation i tordenforekomsten. Den negative binomialfordeling giver en rimelig beskrivelse af hyppigheden af tordenvejr på tilfældigt udtrukne dage, men ønsker man en beskrivelse af hyppigheden af tordenvejr på succesive dage, må den eventuelle autokorrelation inddrages i analysen.

Eksempel 6.4.2 Variation mellem antal fejl ved airconditioneringsanlæg


<table>
<thead>
<tr>
<th>Fly nr</th>
<th>7908</th>
<th>7909</th>
<th>7910</th>
<th>7911</th>
<th>7912</th>
<th>7913</th>
<th>7914</th>
<th>7915</th>
<th>8044</th>
<th>8045</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_i$</td>
<td>8</td>
<td>16</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>4</td>
<td>9</td>
<td>12</td>
</tr>
</tbody>
</table>

Man finder $\bar{y}_+ = 10.30$ [fejl/1000 timer] og $s_2^2 = s^2 = 15.79$. Benyttes Pearson-størrelsen til sammenligning af de 10 observerede fejlintensiteter under den systematiske model finder man

$$(n - 1)s^2 / \bar{y}_+ = 13.80 > \chi^2_{0.95}(9),$$

og man vælger derfor at modellere variationen med en tilfældig model for intensiteten.

Idet det antages, at $\mu \in G(m/\gamma, \gamma)$ med $m = \alpha/\beta$ og $\gamma = 1/\beta$ finder man estimatorne

$\hat{m} = 10.30$ [fejl/1000 timer]
og

\[ \tilde{\gamma} = \frac{15.79}{10.30} - 1 = 0.533 \text{ [1000 timer]} \]

svarende til \( \tilde{\alpha} = 19.32, \tilde{\beta} = 1.88 \)

Måler vi i stedet i enheden [timer] finder vi

\[ \tilde{m}^* = 0.0103 \text{ [fejl/time]} \]

og

\[ \tilde{\gamma}^* = 0.000533 \text{ [timer]} \]

svarende til \( \tilde{\alpha}^* = 19.32, \tilde{\beta}^* = 1876. \)

6.5 Eksponentialfordelingen

Såfremt \( X_{ij} | \mu_i \in \text{Ex}(\mu_i) \), finder vi \( Z_i = X_{i1} + \cdots + X_{in_i} \), at \( Z_i | \mu_i \in \text{G}(n_i, \mu_i) \).

Vi har tidligere set, at familien af gammafordelinger er en eksponentiel dispersionsmodel. Familien, \( \text{G}(n_i, \mu_i) \), af fordelinger af \( Z_i \) er en additiv eksponentiel dispersionsmodel med \( n_i \) som indeksparameter, mens familien, \( \text{G}(n_i, \mu_i/n_i) \), af fordelinger af \( Y_i = Z_i/n_i \) er en reproduktiv eksponentiel dispersionsmodel med \( 1/n_i \) som dispersionsparameter (eller dispersionsparameter \( 1 \) og vægten \( n_i \)).

Vi betragter således en eksponentiel dispersionsmodel med middelværdi

\[ E[X_{ij} | \mu] = \mu, \quad (6.5.1) \]

og variansfunktionen

\[ V_G(\mu) = \mu^2 \quad (6.5.2) \]

Den kanoniske linkfunktion er

\[ \eta(\mu) = \frac{1}{\mu} \quad (6.5.3) \]

For \( Z \in \text{G}(n, \mu) \) gælder

\[ E[Z] = n\mu \]

\[ V[Z] = nV_G(\mu) = n\mu^2 \]
For $Y = Z/n$ har vi derfor:

$$E[Y] = \mu$$

$$V[Y] = \frac{V_G(\mu)}{n} = \frac{\mu^2}{n}$$

med $\mu$ og $V_G(\mu)$ givet ved (6.5.1) og (6.5.2)

**Sætning 6.5.1 Test for homogenitet**

Kvotientteststørrelsen for hypotesen (6.1.5) er

$$G^2(H_I) = 2 \sum_{i=1}^{k} n_i \left[ \ln(\bar{y}_+/y_i) + \frac{y_i - \bar{y}_+}{\bar{y}_+} \right]$$ \hspace{1cm} (6.5.4)

med $y_i = z_i/n_i \ (= \bar{x}_{i+})$ og $\bar{y}_+ = \sum z_i / \sum n_i \ (= \bar{x}_{++})$.

Kvotientteststørrelsen svarer til de vægtede enhedsdevianser (med vægtene $n_i$)

$$d(y_i, \bar{y}_+) = 2n_i \left[ \ln(\bar{y}_+/y_i) + \frac{y_i - \bar{y}_+}{\bar{y}_+} \right]$$

Under hypotesen (6.1.5) vil $G^2(H_I)$ asymptotisk være fordelt som $\chi^2(k-1)$. Hypotesen forkastes for store værdier af $G^2(H_I)$.

**Bevis:**

Testet er det sædvanlige homogenitetstest i en generaliseret lineær model.

**Bemærkning 1 Pearson-teststørrelsen**

Pearson-teststørrelsen for hypotesen (6.1.5) er

$$X^2 = \sum_{i=1}^{k} n_i \frac{(y_i - \bar{y}_+)^2}{\bar{y}_+^2}$$ \hspace{1cm} (6.5.5)

Testet er asymptotisk ækvivalent med kvotienttestet. Testet måler variansen imellem gruppegennemsnittene, $SAK_k$, i forhold til den estimerede varians, $V_G(\bar{y}_+) = \bar{y}_+^2$ i eksponentialfordelingen.
Såfremt man ønsker at modellere eventuelle forskelle imellem grupper ved en tilfældig model, kan man vælge at beskrive fordelingen af $\mu$ ved en $\text{RGam}(\alpha, \beta)$-fordeling svarende til at fordelingen af $1/\mu$ modelleres ved en $\text{G}(\alpha, 1/\beta)$-fordeling.

Den marginale fordeling af $Z_i$ bliver da en $\text{RBet}(\alpha, n_i, \beta)$-fordeling.

**Lemma 6.5.1** Momenter i fordelingen af $\mu$ ved reciprok gamma-gamma sampling

Såfremt $Z|\mu \in \text{G}(n, \mu)$ og $\mu \in \text{RGam}(\alpha, \beta)$, da gælder for fordelingen af $\mu$:

For $\alpha \leq 1$ har fordelingen ikke nogen middelværdi. Såfremt $1 < \alpha$, har fordelingen af $\mu$ middelværdien

\[
m = E[\mu] = \frac{\beta}{\alpha - 1} \quad (6.5.6)
\]

For $\alpha \leq 2$ har fordelingen ingen varians. Såfremt $2 < \alpha$, har fordelingen af $\mu$ variansen

\[
V[\mu] = \frac{V_G(m)}{\alpha - 2} = \frac{\gamma}{1 - \gamma} V_G(m) , \quad (6.5.7)
\]

hvor "signal/støj forholdet," $\gamma$, eksisterer såfremt $2 < \alpha$. Signal/støjforholdet $\gamma$ er givet ved

\[
\gamma = \frac{V[\mu]}{E[V_G(\mu)]} = \frac{1}{\alpha - 1} \quad (6.5.8)
\]

med $0 < \gamma < 1$.

Den marginale middelværdi af variansfunktionen $V_G(\mu)$ er
\[ E[V_G(\mu)] = \frac{\alpha - 1}{\alpha - 2} V_G(m) = \frac{1}{1 - \gamma} V_G(m) \]  

(6.5.9)

Endelig er intraklassekorrelationen, \( \rho \), givet ved

\[ \rho = \frac{V[\mu]}{E[V[\mu]]} = \frac{1}{\alpha} \]  

(6.5.10)

for \( 2 < \alpha \).

\textbf{Bevis:}

Momenterne findes direkte ud fra momenterne i den reciproke gammafordeling.

\[ \square \]

Sætning 6.5.2 Den marginale fordeling af gruppetotal ved Reciprok gamma-eksponentielt sampling

Såfremt \( Z|\mu \in G(n, \mu) \) og \( \mu \in \text{RGam}(\alpha, \beta) \), da er den marginale fordeling af \( Z \) en \( \text{RBet}(\alpha, n, \beta) \)-fordeling.

For \( Z \in \text{RBet}(\alpha, n, \beta) \) vil \( T = \beta/(\beta + Z) \) følge en \( \text{Be}(\alpha, n) \)-fordeling.

Såfremt \( \alpha > 2 \) gælder:

\textbf{For} \( \alpha > 1 \) har fordelingen en forventningsværdi

\[ E[Z] = nm \]  

(6.5.11)

\[ V[Z] = n \frac{m^2}{1 - \gamma}[1 + n\gamma] \]  

(6.5.13)
med $m$ og $\gamma$ givet ved (6.5.6) og (6.5.8).
Middelværdien $E[Z]$ eksisterer dog blot $\alpha > 1$.
For den gennemsnitlige værdi $Y = Z/n$ gælder tilsvarende for $\alpha > 2$:

\[
E[Y] = m \quad (6.5.14)
\]
\[
V[Y] = \frac{m^2}{1-\gamma} \left[ \gamma + \frac{1}{n} \right] \quad (6.5.15)
\]

**Bevis:**
Udtrykket for momenterne fås ved indsættelse af resultaterne fra lemma 6.5.1 i det generelle udtryk (6.1.14) i sætning 6.1.2.

**Bemærkning 1 Opspaltning af den totale varians**

Opspaltningen af den totale varians (6.5.13)

\[
V[Z] = n \frac{V_G(m)}{1-\gamma} [1 + n\gamma]
\]
i komponenter svarende til den gennemsnitlige varians indenfor grupper og variansen mellem gruppemiddelværdier, giver størrelserne

\[
E[V[Z|\mu]] = n \, E[V_G(\mu)] = n \, \frac{V_G(m)}{1-\gamma}
\]
\[
V[E[Z|\mu]] = n^2 \, V[\mu] = n^2 \gamma \, \frac{V_G(m)}{1-\gamma}
\]

**Bemærkning 2 Overdispersion i forhold til gammafordelingen**

For begrænsede værdier af $n$ vil fordelingen af $Z$ have en overdispersion i forhold til Gammafordelingen med samme antalsparameter og med "middelværdiparameter" $m$. 
Overdispersionen er
\[ \delta = \frac{V[Z]}{nV_G(m)} = \frac{1 + n\gamma}{1 - \gamma} \quad (6.5.16) \]

Tilsvarende vil fordelingen af \( Y = Z/n \) have overdispersionen \( \delta = (1 + n\gamma)/(1 - \gamma) \) i forhold til \( G(n, m)/n \)-fordelingen.

Den marginale varians af gennemsnittet \( Y \) kan altså udtrykkes som
\[ V[Y] = V_G(m)/n_{eff}, \]

hvor den effektive stikprøvestørrelse, \( n_{eff} \) er bestemt ved
\[ n_{eff} = \frac{n}{\delta} = n \frac{1 - \gamma}{1 + n\gamma} \quad (6.5.17) \]

For \( 0 < m < \infty \) vil \( \gamma \to 0 \) være ensbetydende med \( V[1/\theta] \to 0 \), og fordelingen af \( Z \) vil nærme sig en \( G(n, m) \)-fordeling.

\[ \square \]

**Sætning 6.5.3 Maksimum-likelihood estimation i den reciproke Betafordeling**

Lad \( T_1, T_2, \ldots, T_k \) være uafhængige variable, hvor \( T_i \in RBet(\alpha, n_i, \beta) \)

Maksimum-likelihood estimaterne \( \hat{m} \) og \( \hat{\gamma} \) for \( m = \beta/(\alpha - 1) \) og \( \gamma = 1/(\alpha - 1) \) findes da ved at maksimere
\[
l(m, \gamma; t_1, t_2, \ldots, t_k) = \sum_{i=1}^{k} \left[ -n_i \ln\left(\frac{m}{1 + \gamma}\right) - \left(n_i + \frac{1 + \gamma}{\gamma}\right) \ln(1 + t_i \gamma/m) \right. \\
\left. + \sum_{\nu=1}^{n_i - 1} \ln\left(1 + \nu \gamma/(1 + \gamma)\right) \right]
\]

med hensyn til \( m \) og \( \gamma \).

**Bevis:**
Sætningen bevises ved at notere, at - på nær en konstant - er logaritmen til likelihoodfunktionen givet ved (6.5.18).

\[ \square \]
Bemærkning 1 Bestemmelse af maksimum-likelihood estimaterne

Er endnu ikke udarbejdet

Sætning 6.5.4 Kvantitativ teststørrelse for homogenitetstest

Er endnu ikke udarbejdet

Sætning 6.5.5 Momentestimation i den reciproke Betafordeling

Lad $Z_1, Z_2, \ldots, Z_k$ være uafhængige variable, hvor $Z_i \in \text{R} \text{Bet}(\alpha, n_i, \beta)$

Momentestimatorne for $m = \beta/(\alpha - 1)$ og $\gamma = 1/(\alpha - 1)$ er da

\[
\tilde{m} = \bar{y}_+ \\
\tilde{\gamma} = \frac{s_2^2 - \bar{y}_+^2}{s_2^2 + n_0 \bar{y}_+^2} \tag{6.5.18}
\]

med

\[
s_2^2 = \sum_{i=1}^{k} n_i (y_i - \bar{y}_+)^2 / (k - 1),
\]

hvor $y_i = z_i / n_i$, og hvor den vægtede gennemsnitlige stikprøvestørrelse, $n_0$, er bestemt ved (5.1.9).

De tilsvarende estimator for $\alpha$ og $\beta$ er

\[
\tilde{\alpha} = 1 + \frac{s_2^2 + n_0 \bar{y}_+^2}{s_2^2 - \bar{y}_+^2} \\
\tilde{\beta} = \frac{s_2^2 + n_0 \bar{y}_+^2}{s_2^2 - \bar{y}_+^2}
\]

Bevis:

Følger af sætning 6.1.3.
Bemærkning 1 *Singularitet ved momentestimation*

Såfremt $s_2^2 < \tilde{y}_+$ bliver $\tilde{\gamma}$ negativ. I dette tilfælde vil det være naturligt, at sætte $\gamma$ til nul, d.v.s. fordelingen af $\mu$ estimeres til at være en etpunktforde-
ling, og fordelingen af $Z$ bliver da en $G(n, m)$-fordeling, hvor $m$ estimeres som $\tilde{m} = \tilde{y}_+$.

\[\square\]

Eksempel 6.5.1 Hændelsesrate

Ved modellering af levetider betragter man ofte den såkaldte hændelsesrate (engelsk: hazard rate). Såfremt levetiden $X$ har fordelingsfunktionen $F(\cdot)$, hvor $F(x) = P[X \leq x]$ med tætheden $f(x) = F'(x)$, da er hændelsesraten $\lambda(\cdot)$ defineret som

$$\lambda(x) = \frac{f(x)}{1 - F(x)}$$

Hændelsesraten angiver således den infinitesimal dødssandsynlighed til ti-
den $x$, givet komponenten er i live til tiden $x$. Forløbet af hændelsesraten beskriver ældningsforholdet for populationen.

Levetiden for elektroniske komponenter modelleres ofte ved eksponential-
fordelingen, i det mindste inden for den operationelle horisont. Hændelses-
raten for eksponentialfordelingen er konstant, svarende til at der ikke finder
nogen væsentlig ældning sted for sådanne komponenter.

Såfremt komponenter placeres i forskellige omgivelser, udsættes de for for-
skellige stresspåvirkninger. Dette kan modelleres ved at lade middellevet-
tiden, $\theta$, afhænge af stressfaktorerne for de specifikke omgivelser. For en
population af komponenter i almindelig brug vil det derfor være naturligt
at modellere den marginale fordeling af levetiden som en mikstur af de
betingede levetidsfordelinger.

Sætter vi $n = 1$ i sætning 6.5.2 får vi specielt, at såfremt $X|\theta \in \text{Ex}(\theta)$, og $\theta \in \text{RGam}(\alpha, \beta)$, da vil den marginale fordeling af $X$ være en $\text{RBet}(\alpha, 1, \beta)$-
fordeling. Det kan vises, at denne fordeling har en aftagende hændelsesrate
(se f.eks. Barlow og Proschan, der viser, at en mikstur af fordelinger med
aftagende hændelsesrater igen har en aftagende hændelsesrate).

Resultatet, der umiddelbart kan synes paradoksalt, kan fortolkes på følgende
måde: for voksende værdier af tiden $x$, vil den information, at komponen-
ten har overlevet til tiden $x$, give en stærkere og stærkere indikation af at
komponenten hidrører fra den del af populationen, der har store værdier af
middellevetiden $\theta$, og man finder da, at jo længere komponenten lever, des mindre bliver den umiddelbare dødsintensitet.

\[
\square
\]

**Eksempel 6.5.2 Ventetid til udskiftning**

Betrægter vi en bestemt komponentposition i et større apparat, f.eks. en air-conditioner i en flyvemaskine, og antages det at komponenten udskiftes med en ny (eller repareres, så den er så god som ny) vil der for et bestemt apparat gælde, at levetiderne $X_1, X_2, \ldots, X_n$ indtil den $n$'te udskiftning kan betragtes som uafhængige ensfordelte variable.

Antag nu, at levetiderne $X_i$ for den betragtede komponent i et givet apparat (fly) varierer i overensstemmelse med en $\text{Ex}(\theta)$ fordeling, og at middeltiden mellem fejl $\mathbb{E}[X|\theta] = \theta$ varierer imellem apparaterne i overensstemmelse med en RGam($\alpha, \beta$)-fordeling. Lad $T_n$ angive ventetiden til den $n$'te udskiftning. Der gælder da

\[
P[T_n \leq t|\theta] = P[G(n, \theta) \leq t] \quad (6.5.20)
\]

og

\[
P[T_n \leq t|\theta] = P[R\text{Bet}(\alpha, n, \beta) \leq t] = P\left[\text{Be}(n, \alpha) \leq \frac{t}{t + \beta}\right] \quad (6.5.21)
\]

Betrægter vi i stedet antallet $N_t$ af udskiftninger i en periode af længden $t$, finder vi

\[
P[N_t \leq x|\theta] = P[P(t/\theta) \leq x] \quad (6.5.22)
\]

og ifølge Sætning 6.4.2 har vi da, idet $1/\theta \in G(\alpha, 1/\beta)$, at

\[
P[N_t \leq x] = P[N\text{B}(\alpha, \beta/(\beta + t)) \leq x] \quad (6.5.23)
\]

Ved benyttelse af den sædvanlige ventetidsrelation

\[
\{T_n \leq t\} = \{N_t \geq n\},
\]

finder vi nu af (6.5.21) og (6.5.23)

\[
P[R\text{Bet}(\alpha, n, \beta) \leq t] = P\left[\text{Be}(n, \alpha) \leq \frac{t}{t + \beta}\right] = 1 - P[N\text{B}(\alpha, \beta/(\beta + t)) \leq n - 1] \quad (6.5.24)
\]
svarende til den velkendte relation

\[ P [G(n, \theta) \leq t] = 1 - P [P(t/\theta) \leq n - 1] \]

I eksempel 6.4.2 betragtede vi antallet af fejl, ved airconditioneringsanlægget i løbet af 1000 flyvetimer for hver af 10 fly. (Data fra F. Proschan (1963). Se også L. J. Bain og F. T. Wright (1982)).

Nedenstående tabel angiver de observerede antal fejl, \( n_i \) i løbet af de betragtede 1000 flyvetimer tillige med tidspunktet \( t_i \) for den senest observerede fejl.

<table>
<thead>
<tr>
<th>Fly nr</th>
<th>7908</th>
<th>7909</th>
<th>7910</th>
<th>7911</th>
<th>7912</th>
<th>7913</th>
<th>7914</th>
<th>7915</th>
<th>8044</th>
<th>8045</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_i )</td>
<td>8</td>
<td>16</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>4</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>( t_i )</td>
<td>865</td>
<td>983</td>
<td>842</td>
<td>944</td>
<td>917</td>
<td>812</td>
<td>991</td>
<td>650</td>
<td>934</td>
<td>921</td>
</tr>
</tbody>
</table>

Såfremt vi ser bort fra den information, der ligger i restflyvetiden \( 100 - t_i \), og antager at observationstiden er ophørt ved den \( n_i \)’te udskiftning, kan vi benytte ventetidsmodellen, der svarer til Poisson-Gamma modellen i afsnit 6.4. Vi vil således antage, at ventetiden \( X_{ij} \) fra den \( (j - 1) \)’te udskiftning til den \( j \)’te udskiftning af anlægget i det \( i \)’te fly kan beskrives ved \( \text{Ex}(\mu) \)-fordelte variable, hvor middeltiden mellem fejl, \( \mu \), varierer mellem fly i overensstemmelse med en \( \text{RGam}(\alpha, \beta) \)-fordeling, svarende til at fejlintensiteten \( \lambda = 1/\mu \) varierer i overensstemmelse med en \( G(\alpha, \beta) \)-fordeling.

Idet \( \bar{y}_+ = 86.01 \) [timer], \( s_2^2 = 9853 \) og \( n_0 = 10.146 \), finder man momentestimaterne \( \bar{m} = 86.01 \) [timer] og \( \bar{y} = 0.0289 \) svarende til \( \bar{\alpha} = 35.58 \) og \( \bar{\beta} = 2974 \) [timer]. Sammenlignes med estimaterne \( \alpha^* = 19.32 \) og \( \beta^* = 1876 \), der blev fundet fra de samme data i eksempel 6.4.2 synes forskellen umiddelbart ganske stor, men betragter man f.eks. skønnet over \( E[\mu] = \alpha/\beta \) og \( V[\mu] = \alpha/\beta^2 \) finder man under ventetidsmodellen \( \bar{\alpha}/\bar{\beta} = 0.012 \) og \( \bar{\alpha}/(\bar{\beta})^2 = (0.002)^2 \), mens estimation under Poisson-Gamma modellen fører til \( \bar{\alpha}^*/\bar{\beta}^* = 0.010 \) og \( \bar{\alpha}^*/(\bar{\beta}^*)^2 = (0.002)^2 \), hvilket viser at den reelle forskel ikke er så stor.
6.6 Fordeling af empiriske varianser for normalfordelte variable

6.6.1 Den systematiske model

Betragt et tosidet skema af observationer:

\[ X_{ij}, j = 1, 2, \ldots, n_i, i = 1, \ldots, k, \] og sæt som vanligt

\[ SAK_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i+})^2 \]

med

\[ \bar{x}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \]

Antag, at \( X_{ij} | (\mu_i, \sigma_i^2) \in N(\mu_i, \sigma_i^2) \), og \( X_{ij}, j = 1, \ldots, n_i \) er betinget uafhængige givet sættet af \( (\mu_i, \sigma_i^2), i = 1, \ldots, k \).

Vi har da, at fordelingen af \( SAK_i \) er

\[ SAK_i | \sigma_i^2 \in \sigma_i^2 \chi^2(f_i) \]

med \( f_i = n_i - 1 \). \( SAK_1, \ldots, SAK_k \) er indbyrdes uafhængige, og fordelingen af \( SAK_i \) afhænger ikke af \( \mu_i \).

Betræger vi specielt de empiriske varianser

\[ S_i^2 = \frac{SAK_i}{f_i} \] (6.6.1)

har vi jvf eksempel 2.2.7, at

\[ S_i^2 | \sigma_i^2 \in \sigma_i^2 \chi^2(f_i)/f_i , \]

eller, udtrykt ved gammafordelingen:

\[ S_i^2 | \sigma_i^2 \in G(f_i/2, \sigma_i^2/(f_i/2)) \]

Behandlingen af fordelingsforholdene for \( S_i^2 \) er således blot et specialtilfælde af betragtningerne i afsnit 6.5. På grund af den særlige interesse,
der knytter sig til de empiriske varianser fra normalfordelte observatio-
ner, vil vi alligevel her give en beskrivelse af fordelingsforholdene i dette
specialtilfælde.

Homogenitetshypotesen

\[ H_I : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \]

dværende til den systematiske model (Bartletts test) blev behandlet i ek-
sempel 2.7.10.

### 6.6.2 Den tilfældige model

Under en tilfældig model vil det i lighed med afsnit 6.5 være naturligt at
betragte

\[ \sigma_i^2 \in \text{RGam}(\alpha, \beta) , \]

hvilket er det samme som

\[ \frac{1}{\sigma_i^2} \in \text{G}(\alpha, 1/\beta) , \]

Det følger af betragtningerne i Oversigt over fordelinger med anvendelser i
Statistik, IMM 1998, at

\[
\begin{align*}
\mathbb{E}[\sigma^2] & = \frac{\beta}{\alpha - 1} \\
\mathbb{V}[\sigma^2] & = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)} = \frac{\mathbb{E}[\sigma^2]^2}{\alpha - 2}
\end{align*}
\]

Dvs den relative spredning i fordelingen af \( \sigma^2 \) er

\[ \sqrt{\mathbb{V}[\sigma^2]/\mathbb{E}[\sigma^2]} = 1/\sqrt{\alpha - 2} \]
6.6.3 Fortolkning af parametre i strukturfordelingen af $\sigma^2$

I stedet for parametreringen af fordelingen af $\sigma^2$ ved $\alpha$ og $\beta$ vil vi indføre en parametreringsfunktion, der er relateret til $\chi^2$-fordelingen.

Der gælder

Sætning 6.6.1 *Fortolkning af strukturfordelingen af $1/\sigma^2$ som en $\chi^2$-fordeling*

Antag, at strukturfordelingen af $\sigma^2$ er som i (6.6.3), dvs $\sigma^2 \in \text{RGam}(\alpha, \beta)$. Da gælder

\[
\frac{1}{\sigma^2} \in \frac{1}{\sigma_0^2} (\nu - 2) \chi^2(\nu), \quad (6.6.5)
\]

med $\sigma_0^2 = \text{E}[\sigma^2]$ og $\nu = 2\alpha$.

Bevis:

Indfører vi $\nu = 2\alpha$ i (6.6.2), har vi

\[
\frac{1}{\sigma^2} \in \text{G}(\nu/2, 1/\beta), \quad (6.6.6)
\]

dvs formparameteren er $\nu/2$ og skalaparameteren er $1/\beta$. Der gælder

\[
\text{E}[\sigma^2] = \frac{\beta}{\nu/2 - 1}, \quad (6.6.7)
\]

hvorför vi kan udtrykke parameteren $\beta$ som

\[
\beta = (\nu - 2) \frac{\sigma_0^2}{2}, \quad (6.6.8)
\]

hvorför vi har sat

\[
\sigma_0^2 = \text{E}[\sigma_i^2] \quad (6.6.9)
\]

Vi kan udtrykke gammafordelingen (6.6.2) som en $\chi^2$-fordeling ved $\chi^2(f) \equiv \text{G}(f/2, 2)$, dvs

\[
\frac{1}{\sigma^2} \in \frac{1}{2\beta} \chi^2(\nu), \quad (6.6.10)
\]
eller, idet vi udtrykker skalaparameteren $\beta$ ved forventningsværdien, $\sigma_0^2$ af $\sigma^2$ (jvf (6.6.8)), har vi at

$$\frac{1}{\sigma^2} \in G(\nu/2, 2/[(\nu - 2)\sigma_0^2]) ,$$

(6.6.11)

hvilket er det samme som (6.6.5).

\[\square\]

**Bemærkning 1** Parameteren $\nu$ betegnes undertiden “frihedsgraderne” i fordelingen af $1/\sigma^2$

På grund af relationen (6.6.5) betegner man undertiden parameteren $\nu$ som “frihedsgraderne i fordelingen af $1/\sigma^2$”. Vi bemærker dog, at parameteren $\nu$ ikke behøver være heltallig.

Det følger af (6.6.4), og af relationen $\nu = 2\alpha$, at parameteren $\nu$ er bestemt ved den relative spredning i fordelingen af $\sigma^2$.

\[\square\]

Figur 6.3 viser et eksempel på fordelingen af $\sigma^2$.

### 6.6.4 Marginal fordeling af stikprøvevariansen

**Sætning 6.6.2** Den marginale fordeling af $S^2$ ved reciprok gamma strukturfordeling

Såfremt $S^2|\sigma^2 \in \sigma^2 \chi^2(f)/f$ og $1/\sigma^2 \in \sigma_0^2(\nu-2) \chi^2(\nu)$, da er den marginale fordeling af $S^2$ givet ved

$$S^2 \in \text{RBet}\left(\nu/2, f/2, \frac{\nu - 2}{f} \sigma_0^2\right)$$

(6.6.12)

Såfremt $\nu \leq 2$ har fordelingen af $S^2$ ingen middelværdi. For $2 < \nu$ har fordelingen af $S^2$ middelværdien

$$E[S^2] = E[\sigma^2] = \sigma_0^2$$

(6.6.13)
Figur 6.3. Strukturfordeling af sand varians, $\sigma^2$ for $\nu = 4$, $\sigma_0^2 = 1$

For $\nu \leq 4$ har fordelingen ingen varians. Såfremt $4 < \nu$, har fordelingen af $S^2$ variansen

$$V[S^2] = \frac{(\sigma_0^2)^2}{\nu/2 - 2} \left[ 1 + \frac{2(\nu/2 - 1)}{f} \right]$$  \hspace{1cm} (6.6.14)

**Bevis:**

Fås af (6.5.6) og (6.5.7)

Den simultane fordeling af variansen, $\sigma^2$, og af stikprøvevariansen (den empiriske varians), $S^2$, er illustreret i figur 6.4
Figur 6.4. Simultan fordeling af empirisk varians, $S^2$, bestemt i stikprøve på $n = 5$ og sand varians $\sigma^2$
(Strukturforselning af $\sigma^2$ som i figur 6.3.)

Figur 6.5 viser den betingede fordeling af den empiriske varians, $S^2$, svarende til en givet værdi af $\sigma^2$, ($\sigma^2 = 1$), og figur 6.6 viser den marginale fordeling af den empiriske varians, $s^2$, svarende til fordelingen i figur 6.4. Det ses, at den marginale fordeling af $S^2$ har tykkere haler, end den betingede fordeling i figur 6.5.

Bemærkning 1 *Den marginale fordeling af $S^2$ udtrykt ved F-fordelingen*

Ved at udnytte relationen mellem RBet-fordelingen og F-fordelingen (se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998) finder
Figur 6.5. Betinget fordeling af empirisk varians, $S^2$, bestemt i stikprøve på $n = 5$ for en sand varians, $\sigma^2 = 1$.

$\frac{1}{S^2} \in \frac{\nu}{(\nu - 2)\sigma_0^2} \text{F}(\nu, f)$, \hspace{1cm} (6.6.15)

hvor $\text{F}(\nu, f)$ angiver en stokastisk variabel, der følger en F-fordeling med frihedsgraderne $(\nu, f)$.

Eksempel 6.6.1 Bestemmelse af sandsynligheder i den marginale fordeling af $S^2$
Figur 6.6. Marginal fordeling af empirisk varians, $S^2$, bestemt i stikprøve på $n = 5$
(Strukturfordeling af $\sigma^2$ som i figur 6.3.)

Antag, at situationen er som beskrevet i sætning 6.6.2, hvor strukturfordelingen af $\sigma^2$ er givet ved (6.6.5) med parametrene $\nu = 4.5$ og $\sigma_0^2 = 9.0$, og at der udtages stikprøver af størrelsen $n = 8$.

Man ønsker nu at bestemme sandsynligheden for at få en værdi af den empiriske varians, $S^2$, der er større end 16.67.

Idet stikprøvestørrelsen er $n = 8$, har man, at frihedsgraderne for $S^2$ er $f = n - 1 = 7$. 
6.6 Empiriske varianser

Vi ønsker således at bestemme sandsynligheden

\[ P \left[ S^2 > 16.67 \right] = P \left[ \frac{1}{S^2} < \frac{1}{16.67} \right] = P \left[ \frac{\nu}{(\nu - 2)\sigma_0^2} F(\nu, f) \leq \frac{1}{16.67} \right], \]

dvs

\[ P \left[ F(\nu, f) \leq \frac{(\nu - 2)\sigma_0^2}{16.67\nu} \right] = P \left[ F(4.5, 7) \leq \frac{2.5 \times 9}{4.5 \times 16.67} \right], \]

eller

\[ P \left[ F(4.5, 7) \leq 0.30 \right], \]

Da tabellen over fraktiler i F-fordelingen kun angiver fraktiler svarende til sandsynligheder, der er større end 50 %, benytter vi relationen

\[ P \left[ F(f_1, f_2) \leq x \right] = 1 - P \left[ F(f_2, f_1) \leq 1/x \right] \]

dvs den søgte sandsynlighed fås som

\[ P \left[ F(4.5, 7) \leq 0.30 \right] = 1 - P \left[ F(7, 4.5) \leq 3.33 \right] \]

Ved opslag i tabellen over fraktiler i F-fordelingen finder vi, at \( P \left[ F(7, 4) \leq 3.98 \right] = 0.90 \) og \( P \left[ F(7, 5) \leq 3.37 \right] = 0.90 \). Den søgte sandsynlighed, \( P \left[ F(4.5, 7) \leq 0.30 \right] \) er altså lidt større end 0.10.

6.6.5 Estimation af parametre i strukturfordeling

Estimationen foregår principielt som beskrevet i afsnit 6.5 og i sætning 5.7.3 på side 536.

Maksimum likelihood estimatet kan bestemmes ved benyttelse af sætning 6.5.3. Maksimum-likelihood estimatorne må bestemmes ved iteration.

Momentestimatorne kan bestemmes direkte ved nedenstående

Sætning 6.6.3 Momentestimation af parametre i strukturfordeling for empiriske varianser

Lad \( S_1^2, \ldots, S_k^2 \) være uafhængige variable, hvor

\[ S_i^2 \in \text{RBet} \left( \nu/2, f_i/2, \frac{\nu - 2}{f_i} \sigma_0^2 \right) \]
svarende til situationen i sætning 6.6.2, hvor \( S_i^2 | \sigma_i^2 \in \sigma_i^2 \chi^2(f_i)/f_i \) og strukturfordelingen af \( \sigma_i^2 \) er bestemt ved

\[
1/\sigma_i^2 \in \frac{1}{\sigma_0^2(\nu - 2)} \chi^2(\nu)
\]

Da er momentestimaterne for parametrene \( \sigma_0^2 \) og \( \nu \) bestemt ved

\[
\tilde{\sigma}_0^2 = \frac{s^2}{\hat{\nu}}
\]

(6.6.16)

\[
\hat{\nu} = 2 \left[ 1 + \left\{ Q_1 + \frac{1}{k-1} \left[ \sum_i f_i - \frac{\sum_i f_i^2}{\sum_i f_i} \right] \right\} / (Q_1 - 2) \right],
\]

hvor

\[
Q_1 = \frac{SAK_s/(k - 1)}{(s^2)^2},
\]

med

\[
SAK_s = \sum_i (n_i - 1)(S_i^2 - \overline{S^2})^2
\]

(6.6.17)

og

\[
\overline{s^2} = \left( \sum_{i=1}^k f_i s_i^2 \right) / \left( \sum_{i=1}^k f_i \right)
\]

(6.6.18)

**Bevis:**

Beviset følger af sætning 6.5.5.

**Bemærkning 1 Momentestimaterne i det balancerede tilfælde**

I det balancerede tilfælde, \( f_1 = f_2 = \cdots = f_k = f \), får vi

\[
\hat{\nu} = 2(2 + (f + 2)/(Q_1 - 2))
\]
6.7 Den flerdimensionale normalfordeling

6.7.1 Den systematiske model

Vi vil betrægde den sædvanlige flerdimensionale variansanalysemodel. Lad observationerne $X_{ij}$ være $p$-dimensionale vektorer, hvor

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \; i = 1, 2, \ldots, k; \; j = 1, 2, \ldots, n_i$$

(6.7.1)

med

$$\sum_{i=1}^{k} n_i \alpha_i = 0$$

hvor $\mu$, $\alpha_i$ og $\epsilon_{ij}$ angiver $p$-dimensionale vektorer med $\epsilon_{ij}$ indbyrdes uafhængige, $\epsilon_{ij} \in N_p(0, \Sigma)$, og hvor $\Sigma$ angiver den fælles $p \times p$-dimensionale kovariansmatrix. For simpelheds skyld antager vi at $\Sigma$ har fuld rang.

Under disse antagelser finder vi at $Z_i \sum_j X_{ij} \in N_p(n_i(\mu + \alpha_i), n_i \Sigma)$.

I dette tilfælde beskrives variationen ved $p \times p$-dimensionale SAK-matricer. Vi vil derfor i lighed med tidligere indførte betegnelserne

$$\overline{X}_{i+} = \sum_{j=1}^{n_i} X_{ij} / n_i$$

(6.7.2)

$$\overline{X}_{i+} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} / N = \sum_{i=1}^{k} n_i \overline{X}_{i+} / \sum_{i=1}^{k} n_i$$

(6.7.3)

til beskrivelse af gruppegennemsnittene og det fælles gennemsnit, ligesom vi indfører

$$SAK_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i+})(X_{ij} - \overline{X}_{i+})^T$$

(6.7.4)

$$SAK_2 = \sum_{i=1}^{k} n_i (\overline{X}_{i+} - \overline{X}_{++})(\overline{X}_{i+} - \overline{X}_{++})^T$$

(6.7.5)

$$SAK_0 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{++})(X_{ij} - \overline{X}_{++})^T$$

(6.7.6)
til beskrivelse af henholdsvis variationen indenfor grupper \((S\overline{A}K_1)\), imellem grupper \((S\overline{A}K_2)\) og den totale variation \((S\overline{A}K_0)\).

Vi bemærker, at også her gælder den sædvanlige pythagorasiske relation

\[
S\overline{A}K_0 = S\overline{A}K_1 + S\overline{A}K_2
\]  

(6.7.7)

**Sætning 6.7.1 Test for homogenitet**

Kvotientteststørrelsen for hypotesen

\[
H_I: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0
\]  

(6.7.8)

mod alternativet

\[
\overline{H}_I: \alpha_i \neq 0 \text{ for mindst ét } i
\]

har teststørrelsen

\[
Z = \frac{\text{det}(S\overline{A}K_1)}{\text{det}(S\overline{A}K_0)}
\]

hvor \(\text{det}(A)\) angiver determinanten af matricen \(A\).

Under hypotesen \(H_I\) er \(Z\) fordelt som Wilk's \(\Lambda\) med parametrene \((p, k - 1, N - k)\)

**Bevis:**
Se f.eks. Rao p. 556.

\[\square\]

**Bemærkning 1 Approximativ fordeling af teststørrelsen**

Wilk's \(\Lambda\) teststørrelse benævnes også Anderson's \(U\)-teststørrelse.

Rao (1971) anfører p.556 at fordelingen af

\[
Z^* = \frac{1 - Z^{1/s}}{Z^{1/s}}
\]

med

\[
m = N - 1 - \frac{p + k}{2}, \quad s = \sqrt{\frac{p^2(k - 1)^2 - 4}{p^2 + (k - 1)^2 - 5}} \quad \text{og} \quad \lambda = \frac{p(k - 1) - 2}{4}
\]

approximativt følger en \(F(p(k - 1), ms - 2\lambda)\)-fordeling.

\[\square\]
Bemærkning 2 Test ved brug af den generaliserede Mahalanobis afstand

Et approximativt test for hypotesen $H_I$ fås ved at betragte den generaliserede Mahalanobis afstand

$$
\chi'^2 = \sum_{i=1}^{k} n_i (\overline{X}_{i+} - \overline{X}_{++})^T S^{-1}_i (\overline{X}_{i+} - \overline{X}_{++})
$$

(6.7.9)

hvor

$$
S_i = S_A K_i / (N - k)
$$

angiver det sædvanlige skøn over kovariansmatricen $\Sigma$.

Den generaliserede Mahalanobis afstand $\chi'^2$ er en skalar. Under hypotesen $H_I$ er $\chi'^2$ approximativt $\chi^2(p(k - 1))$-fordelt. Testet forkaster for store værdier af $\chi'^2$.

6.7.2 Den tilfældige model

Såfremt man ønsker at modellere eventuelle forskelle imellem grupperne ved en tilfældig model kan man vælge at modellere fordelingen af $\alpha_i$ med en $N_p(0, \Sigma_0)$-fordeling.

Sætning 6.7.2 Den marginale fordeling af gruppetotal og gruppegennemsnit ved normal-normal sampling

Såfremt $X_{ij} = \mu + \alpha_i + \epsilon_{ij}, i = 1, 2, \ldots, k; j = 1, 2, \ldots, n_i$ hvor $\alpha_i$ er uafhængige, $\alpha_i \in N_p(0, \Sigma_0), i = 1, 2, \ldots, k$ og hvor $\epsilon_{ij}$ er indbyrdes uafhængige, $\epsilon_{ij} \in N_p(0, \Sigma)$,
da er den marginale fordeling af $Z_i = \sum_j X_{ij}$ en

$$N_p(n_i \mu, n_i \Sigma + n_i^2 \Sigma_0)$$-fordeling

og den marginale fordeling af $\overline{X}_{i+}$ er en

$$N_p(\mu, \frac{1}{n_i} \Sigma + \Sigma_0)$$-fordeling

Endvidere gælder, at

$$SAK_1 \in \text{Wis}_p(N - k, \Sigma)$$

og $SAK_1$ er uafhængig af $\overline{X}_{i+}$, $i = 1, 2, \ldots, k$.

**Bevis:**
Sætningen bevises ved at bemærke, at den marginale fordeling igen er en normal fordeling og ved at benytte relationen (0.1.2) i Sætning 0.1.1 i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

$$D[Z] = D[E[Z|\mu]] + E[D[Z|\mu]] \quad (6.7.10)$$

□

**Bemærkning 1 Det generaliserede signal/støj-forhold**

Indfører vi den $p \times p$-dimensionale matrix $\Gamma$ for forholdet mellem variationen mellem grupper og variansen inden for grupper ("signal/støj-forholdet"), $\Gamma = \Sigma_0 \Sigma^{-1}$, finder vi :

$$D[\overline{X}_{i+}] = \left(\frac{1}{n_i} I + \Gamma\right) \Sigma$$

□

**Sætning 6.7.3 Momentestimation i den flerdimensionale normalfordelingsmodel**
6.7 Flerdimensional normalfordeling

Under antagelserne fra sætning 6.7.2 findes momentestimatorerne for $\mu$, $\Sigma$ og $\Sigma_0$ som

$$
\tilde{\mu} = x_{++} \\
\tilde{\Sigma} = \frac{1}{N-k} \text{sak}_1 \\
\tilde{\Sigma}_0 = \frac{1}{n_0} \left( \frac{\text{sak}_2}{k-1} - \tilde{\Sigma} \right)
$$

(6.7.11)

Bevis:
Sætningen bevises ved at bemærke, at

$$
E [X_{++}] = \mu, \quad E [SAK_1] = (N-k)\Sigma, \quad \text{og} \quad E [SAK_2] = (k-1)(\Sigma+n_0\Sigma_0)
$$

hvor $n_0$ er givet ved (5.1.9).

Bemærkning 1 Momentestimatet er ikke nødvendigvis ikke-negativ definit

Skønnet $\tilde{\Sigma}_0$ er centralt for $\Sigma_0$, men skønnet er ikke nødvendigvis en ikke-negativ definit matrix. Såfremt $\tilde{\Sigma}_0$ ikke har fuld rang, indikerer det, at fordelingen af $\alpha$ kan være udartet.

Sætning 6.7.4 Maksimum-likelihood estimation under den flerdimensionale normalfordelingsmodel

Under antagelserne fra sætning 6.7.2 findes maksimum-likelihood estima
terne for $\mu$, $\Sigma$ og $\Sigma_0$ ved at maksimere

$$
l(\mu, \Sigma, \Sigma_0; x_1+, \ldots, x_k+) = -\frac{N-k}{2} \ln(\det(\Sigma)) - \frac{1}{2} \text{tr}(\text{sak}_1 \Sigma^{-1})
- \sum_{i=1}^{k} \left[ \ln \left( \frac{\Sigma}{n_i} + \Sigma_0 \right) + \frac{1}{2} (\tilde{x}_{i+} - \mu)^T \left( \frac{\Sigma}{n_i} + \Sigma_0 \right)^{-1} (\tilde{x}_{i+} - \mu) \right]
$$

(6.7.12)

med hensyn til $\mu \in \mathbb{R}^p$ og $\Sigma$ og $\Sigma_0$ i mængden af ikke-negativ definite symmetriske $p \times p$-matricer.

Bevis:
Beviset følger ved at bemærke, at $SAK_1$ følger en $\text{Wis}_p(N-k, \Sigma)$-fordeling
og at \( \text{SAK}_1 \) er uafhængig af \( \overline{X}_{i+}, \ i = 1, 2, \ldots, k \), samt at \( \overline{X}_{i+} \in N_p(\mu, \Sigma/n_i + \Sigma_0) \) er indbyrdes uafhængige, \( i = 1, 2, \ldots, k \).

\[
\text{Bemærkning 1 Numerisk bestemmelse af ML-estimatorerne}
\]

Optimeringsproblemet har ikke nogen eksplicit løsning, hvorfor estimatoren må bestemmes ved en iterativ søgeprocedure. Ved en automatiseret procedure vil man ofte parametrinisere varians-kovariansmatricerne ved de enkelte elementer. Afgrænsningen af søgeområdet foretages da ved at teste på de resulterende matricer, og indlægge en passende straff funktion, såfremt determinanten er negativ.

\[
\text{Eksempel 6.7.1 Variation mellem målefejl for flowmåler}
\]

Nedenstående tabel viser resultaterne af 3 gentagne kalibreringer af 6 flowmålere, der er udtaget af en større målerpopulation. De 6 målere blev hver kalibreret ved de samme to flow, henholdsvis 0.1 [m³ /h] og 0.5 [m³ /h].

<table>
<thead>
<tr>
<th>Maaler</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>flow</td>
<td>flow</td>
<td>flow</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>41</td>
<td>-2.0</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>42</td>
<td>5.0</td>
<td>3.0</td>
<td>1.0</td>
</tr>
<tr>
<td>43</td>
<td>2.0</td>
<td>1.0</td>
<td>-3.0</td>
</tr>
<tr>
<td>44</td>
<td>4.0</td>
<td>4.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>45</td>
<td>4.0</td>
<td>2.0</td>
<td>0.0</td>
</tr>
<tr>
<td>46</td>
<td>5.0</td>
<td>9.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Med henblik på beskrivelse af variationen af fejlvisningen i målerpopula
tionen opstiller man følgende model:

\[
X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \ i = 41, 42, \ldots, 46; \ j = 1, 2, 3.
\]

hvor \( \alpha_i \) er uafhængige, \( \alpha_i \in N_2(0, \Sigma_0) \), og hvor \( \epsilon_{ij} \) er indbyrdes uafhængige, \( \epsilon \in N_2(0, \Sigma) \).
Man finder
\[
\bar{x}_{41+} = \begin{pmatrix} -1.00 \\ 1.67 \end{pmatrix}; \quad \bar{x}_{42+} = \begin{pmatrix} 2.67 \\ 2.00 \end{pmatrix}; \quad \bar{x}_{43+} = \begin{pmatrix} 0.00 \\ 0.00 \end{pmatrix};
\]
\[
\bar{x}_{44+} = \begin{pmatrix} 2.00 \\ 3.67 \end{pmatrix}; \quad \bar{x}_{45+} = \begin{pmatrix} 1.00 \\ 1.00 \end{pmatrix}; \quad \bar{x}_{46+} = \begin{pmatrix} 5.00 \\ 9.00 \end{pmatrix};
\]
\[
sak_1 = \begin{pmatrix} 66.67 & 29.00 \\ 29.00 & 15.33 \end{pmatrix}; \quad sak_0 = \begin{pmatrix} 134.28 & 116.22 \\ 116.22 & 171.78 \end{pmatrix}
\]
og
\[
sak_2 = \begin{pmatrix} 67.61 & 87.22 \\ 87.22 & 156.44 \end{pmatrix}
\]

Under den systematiske model finder man kvotientteststørrelsen, Wilks \( \Lambda \) for hypotesen \( \alpha_{41} = \cdots = \alpha_{46} = 0 \) imod alternativet, at mindst to \( \alpha \)-værdier er forskellige er \( \Lambda = 0.018 \).

Gentagelsesvariationen udspænder således kun omkring 2 % af variationen i hele materialet.

Ved indsættelse i (6.7.11) finder man momentestimaterne
\[
\tilde{\mu} = \begin{pmatrix} 1.61 \\ 2.89 \end{pmatrix}
\]
kalibreringsusikkerheden
\[
\tilde{\Sigma} = \frac{sak_1}{12} = \begin{pmatrix} 5.56 & 2.42 \\ 2.42 & 1.28 \end{pmatrix}
\]
og dispersionsmatricen for målerpopulationen
\[
\tilde{\Sigma}_0 = \frac{1}{3}[sak_2/5 - sak_1/12] = \begin{pmatrix} 2.65 & 5.01 \\ 5.01 & 10.00 \end{pmatrix}
\]

Hosstående figurer illustrerer dekomponeringen af variationen.

Vi bemærker af figur 6.8 at målerens "fejlniveau", \( \alpha_{i1} + \alpha_{i2} \) varierer i populationen stort set uafhængigt af målerens "fejldifferens", \( \alpha_{i1} - \alpha_{i2} \).
Man kan forestille sig, at man for hver måler bestemmer målerens kalibreringskurve, nemlig visningsfejlen som funktion af flow'et. I området mellem de to betragtede flow vil kalibreringskurven erfaringsmæssigt kunne tilnærmes med en ret linie. Man har da, at "fejlniveauet" vil være udtryk for kalibreringskurvens afskæring, dvs. målerens justering, mens "fejldifferensen" udtrykker kalibreringskurvens hældning, der afhænger af målerens konstruktion. Det er således ikke overraskende, at disse to størrelser varierer uafhængigt af hinanden.
Samhørende værdier af registreret fejl ved to flow for 3 gentagne prøver på hver af 6 flowmålere. Gentagelser på samme måler er markeret med samme symbol.

Figur 6.7. Illustration af den totale variation af fejlvisningen ved to flow for en række målere.
Samhørende værdier af estimeret målerfejl ved to flow for stikprøve bestående af 6 flowmålere

Figur 6.8. Illustration af variationen mellem grupper (målerfejlen)
Samhørende værdier af kalibreringsfejlen ved to flow
for 3 gentagne prøver på hver af 6 flowmåleres
målerens egenfejl er elimineret

Figur 6.9. Illustration af variationen indenfor grupper (gentagelsesvariationen)
En væsentlig del af beregningerne i eksemplet kunne udføres i SAS®-programsystemet vedproceduren GLM under benyttelse af ordren MANOVA, der bevirket udskrivning af sak-matricer.

Antag, at data fra eksemplet er indlæst i de variable lbnr, fejl1 og fejl2. SAS®-programmet

PROC GLM;
CLASS lbnr;
MODEL fejl1 fejl2 = lbnr /E1;
MANOVA H=lbnr/PRINTH PRINTTE HTYPE=1 ETYPE=1;
RANDOM lbnr;
RUN;

definerer i ordren CLASS lbnr ;, at den variable lbnr som klassifikations- variabel. Modelformlen MODEL fejl1 fejl2 = lbnr / E1 ; angiver, at vi betragter en model svarende til en løbenummereffekt (samt et intercept) for de to variable fejl1 og fejl2. Valget E1 i modelformlen angiver, at vi ønsker en såkaldt type-I kvadratafvigelsessum.

I sætningen

MANOVA H=lbnr/PRINTH PRINTTE;

angiver nogleordet MANOVA, at man ønsker at opfatte de to variable fejl1 og fejl2 på venstre side i modelformlen som en todimensional observation.


Endelig angiver ordren RANDOM lbnr; , at effekten fra lbnr skal opfattes som tilfældig.

Proceduren udfører først de endimensionale analyser på de variable fejl1 og fejl2 hver for sig. Denne del af udskriften er analog til udskriften fra den endimensionale analyse, som blev betragtet i eksempel 5.5.1.
General Linear Models Procedure

Dependent Variable: PELLL2

General Linear Models Procedure

Dependent Variable: PELLL1

General Linear Models Procedure

Number of observations in data set = 18

Class Levels Values

Class Level Information

General Linear Models Procedure
De endimensionale analyser viser, at der er signifikant forskel på fejlvisningen fej12 (ved flow 0.5 [m³/h]), \( F(5,12) = 24.49 \), mens man ikke kan påvise forskel på fejlvisningen fej11 ved flow 0.1 [m³/h] \( F(5,12) = 2.43 \).

sak\textsubscript{1} matricen udskrides under overskriften \( E = \text{Error SS\&CP Matrix} \)

\[
\begin{array}{ccc}
  & \text{FEJL1} & \text{FEJL2} \\
\text{FEJL1} & 66.6667 & 29 \\
\text{FEJL2} & 29 & 15.3333
\end{array}
\]

og endvidere udskrides korrelationsmatricen svarende til sak\textsubscript{1}

General Linear Models Procedure
Multivariate Analysis of Variance

Partial Correlation Coefficients from the Error SS\&CP Matrix/Prob > |r|

\[
\begin{array}{ccc}
  & \text{FEJL1} & \text{FEJL2} \\
\text{FEJL1} & 1.000000 & 0.907038 \\
 & 0.0001 & 0.0001 \\
\text{FEJL2} & 0.907038 & 1.000000 \\
 & 0.0001 & 0.0001
\end{array}
\]

Det ses, at der er en stærk positiv korrelation \( \hat{\rho} = 0.91 \) mellem kalibreringsfejlen ved de to flow.

sak\textsubscript{2}-matricen udskrides under overskriften \( H = \text{Type I SS\&CP Matrix for LBNR} \) som:

General Linear Models Procedure
Multivariate Analysis of Variance

\[
H = \text{Type I SS\&CP Matrix for LBNR}
\]

\[
\begin{array}{ccc}
  & \text{FEJL1} & \text{FEJL2} \\
\end{array}
\]
og endelig udskrives en række teststørrelser for test af hypotesen om forsvindende effekt af lbnr i den systematiske model. Vi bemærker, at kvotientteststørrelsen $Z$ jvf. sætning 6.7.1 udskrives under betegnelsen Wilks’ Lambda.

**Manova Test Criteria and F Approximations for the Hypothesis of no Overall LBNR Effect**

$H = \text{Type I SS&CP Matrix for LBNR} \quad E = \text{Error SS&CP Matrix}$

$$S=2 \quad M=1 \quad N=4.5$$

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks’ Lambda</td>
<td>0.01895960</td>
<td>13.7775</td>
<td>10</td>
<td>22</td>
<td>0.00001</td>
</tr>
<tr>
<td>Pillai’s Trace</td>
<td>1.29172915</td>
<td>4.3771</td>
<td>10</td>
<td>24</td>
<td>0.0015</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>35.35683630</td>
<td>35.3568</td>
<td>10</td>
<td>20</td>
<td>0.0001</td>
</tr>
<tr>
<td>Roy’s Greatest Root</td>
<td>34.88712493</td>
<td>83.7291</td>
<td>5</td>
<td>12</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

**NOTE:** F Statistic for Roy’s Greatest Root is an upper bound.

**NOTE:** F Statistic for Wilks’ Lambda is exact.

Vi ser, at hypotesen klart må forkastes.

Endelig, som et resultat af ordren RANDOM lbnr udskrives udtrykket for

$$E \left[ \frac{s\sigma^2}{(k - 1)} \right] = \Sigma + n_0 \Sigma_0$$

jvf. sætning 6.7.3:

**Source**

<table>
<thead>
<tr>
<th>Type I Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBNR</td>
</tr>
</tbody>
</table>

Var(Errror) + 3 Var(LBNR)

Den vægtede gennemsnitlige stikprøvestørrelse er her $n_0 = 3$, nemlig den fælles stikprøvestørrelse.
Eksempel 6.7.2 Variation mellem næsehøjder hos 12 kaster i Uttar Pradesh

Nedenstående tabel viser de registrerede gennemsnit af fire antropologiske størrelser samt den empiriske dispersionsmatrix indenfor grupper \( S_1 = \frac{sak_1}{(N-k)} \), målt på stikprøver fra 12 kaster og stammer i Uttar Pradesh. Kilde: C.R. Rao (1948).

De fire registrerede størrelser er hovedlængde \( x_1 \), hovedbredde \( x_2 \), bizygomatisk bredde \( x_3 \), og næsehøjde \( x_4 \).

<table>
<thead>
<tr>
<th>Gruppe</th>
<th>( i )</th>
<th>( n_i )</th>
<th>( x_{1i} )</th>
<th>( x_{2i} )</th>
<th>( x_{3i} )</th>
<th>( x_{4i} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basti</td>
<td>1</td>
<td>86</td>
<td>191.92</td>
<td>139.88</td>
<td>133.36</td>
<td>51.24</td>
</tr>
<tr>
<td>Brahmin</td>
<td>2</td>
<td>92</td>
<td>191.35</td>
<td>139.50</td>
<td>132.68</td>
<td>50.40</td>
</tr>
<tr>
<td>Chattri</td>
<td>3</td>
<td>139</td>
<td>191.92</td>
<td>139.88</td>
<td>133.36</td>
<td>51.24</td>
</tr>
<tr>
<td>Muslim</td>
<td>4</td>
<td>167</td>
<td>190.78</td>
<td>137.40</td>
<td>131.52</td>
<td>51.38</td>
</tr>
<tr>
<td>Bhatu</td>
<td>5</td>
<td>148</td>
<td>186.10</td>
<td>138.58</td>
<td>133.55</td>
<td>52.06</td>
</tr>
<tr>
<td>Habru</td>
<td>6</td>
<td>124</td>
<td>186.94</td>
<td>137.40</td>
<td>131.16</td>
<td>50.30</td>
</tr>
<tr>
<td>Bhil</td>
<td>7</td>
<td>187</td>
<td>181.87</td>
<td>137.62</td>
<td>131.18</td>
<td>48.60</td>
</tr>
<tr>
<td>Dom</td>
<td>8</td>
<td>113</td>
<td>186.40</td>
<td>137.52</td>
<td>132.64</td>
<td>50.34</td>
</tr>
<tr>
<td>Ahir</td>
<td>9</td>
<td>68</td>
<td>187.45</td>
<td>138.12</td>
<td>131.70</td>
<td>48.98</td>
</tr>
<tr>
<td>Kurmi</td>
<td>10</td>
<td>94</td>
<td>188.86</td>
<td>137.86</td>
<td>131.82</td>
<td>49.22</td>
</tr>
<tr>
<td>Artisan</td>
<td>11</td>
<td>173</td>
<td>187.69</td>
<td>136.84</td>
<td>131.30</td>
<td>48.72</td>
</tr>
<tr>
<td>Khar</td>
<td>12</td>
<td>57</td>
<td>188.83</td>
<td>136.28</td>
<td>130.70</td>
<td>48.62</td>
</tr>
<tr>
<td>Ialt</td>
<td>( N = 1448 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Vejet gennemsnit, } \bar{x}_{++} = 188.03 \quad 137.25 \quad 131.91 \quad 50.30
\]

\[
S_1 = \begin{pmatrix}
43.6590 & 5.8865 & 8.4396 & 4.0610 \\
20.2500 & 11.1438 & 2.7326 \\
20.9764 & 2.9688 & 12.2500
\end{pmatrix}
\]

Af tabellen over gruppegennemsnit finder man da

\[
sak_2 = \begin{pmatrix}
14695 & -2744 & 730 & 3804 \\
5783 & 1012 & -1305 \\
1046 & 956 & 2906
\end{pmatrix}
\]
og endvidere finder man ved benyttelse af (5.1.9)

\[ n_0 = 119.41. \]

Den generaliserede Mahalanobis afstand (6.7.9), der sammenligner variationen imellem grupper med variationen inden for grupper, udregnes til \( \chi^2 = 991.41 \), der ses at være langt større end \( \chi^2(44,0.999) = 79.1 \).

Til bestemmelse af sak_1 finder man sak_1 = \((N - k)S_1\), således at

\[
sak_1 = \begin{pmatrix} 62681 & 8453 & 12119 & 5832 \\ 29079 & 16002 & 3924 & \\ 30122 & 4263 & & \\ 17591 & & & \end{pmatrix}
\]

hvorved sak_0 kan findes som sak_0 = sak_1 + sak_2.

Da det(sak_1) = \(5.9212 \times 10^{17}\), og det(sak_2) = \(10.82 \times 10^{17}\) finder man kvotientteststørrelsen for identitet af de 12 forventningsværdier til \(z = 0.55\). Variationen indenfor grupper udgør således lidt over halvdelen af det volumen i det 4-dimensionale rum, der udspændes af data.

Vi finder \(z^* = 21.29\), der sammenlignes med en F(44,5484)-fordeling (se bemærkning 1 på side 610). Da \(F(44,5484)_{0.99} = 1.58\) viser også dette test en klar forskel imellem de undersøgte grupper.

Da vi ikke har særlig interesse i netop de undersøgte grupper, men snarere ønsker at belyse variationen imellem grupper, vælger vi at modellere forskellen mellem gruppemiddelværdierne ved en tilfældig model.

Ved indsættelse i (6.7.11) finder vi da momentestimatene

\[
\hat{\mu} = \begin{pmatrix} 188.03 \\ 137.25 \\ 131.91 \\ 50.30 \end{pmatrix}
\]

\[
\hat{\Sigma} = S_1 = \begin{pmatrix} 43.6500 & 5.8865 & 8.4396 & 4.0610 \\ 20.2500 & 11.1438 & 2.7326 \\ 20.9764 & 2.9688 & \\ 12.2500 & & & \end{pmatrix}
\]
og

\[
\mathbf{\tilde{\Sigma}_0} = \begin{pmatrix}
10.8246 & -2.1383 & 0.4854 & 2.8622 \\
4.2341 & 0.6780 & -1.0161 \\
0.6221 & 0.7035 & 2.1103
\end{pmatrix}
\]

For at vurdere, hvorvidt variationen i middelværdier udfylder hele det 4-dimensionale rum, har vi beregnet determinanten \( \det(\mathbf{\tilde{\Sigma}_0}) = 5.44 \) og endvidere har vi bestemt egenværdierne for \( \mathbf{\tilde{\Sigma}_0} \). Vi finder at den mindste egenværdi er 0.0749, hvilket indikerer, at den væsentligste variation er begrænset til en 3-dimensional hyperplan. Vi skal dog ikke her gå nærmere ind på dette forhold.

\[ \square \]

6.8 Oversigtstabeller
| Stikprøveforskelng af \(X_i|\theta\) | \(\mu = E[X|\theta]\) | \(V(\mu)\) | Strukturforskelng \(w(\cdot)\) | \(m = E[\mu]\) | \(E[V(\mu)]\) | \(\gamma\) | \(\frac{V[\mu]}{E[V(\mu)]}\) | Reference |
|-------------------------------|----------------|----------|----------------|----------------|----------------|---|----------------|----------------|
| B(1, \(p\))                  | \(p\)          | \(\mu(1-\mu)\) | \(p \in Be(\alpha,\beta)\) | \(\pi = \frac{\alpha}{\alpha + \beta}\) | \(\frac{\pi(1-\pi)}{1+\gamma}\) | \(\frac{1}{\alpha + \beta}\) | | Afsn. 6.2 |
| Geo(1, \(p\))                | \(\frac{1-p}{p}\) | \(\mu(1+\mu)\) | \(p \in Be(\alpha,\beta)\) | \(\psi = \frac{\beta}{\alpha - 1}\) | \(\frac{\psi(1+\psi)}{1-\gamma}\) | \(\frac{1}{\alpha - 1}\) | | Afsn. 6.3 |
| P(\(\mu\))                   | \(\mu\)         | \(\mu\)     | \(\mu \in G(\alpha,1/\beta)\) | \(m = \frac{\alpha}{\beta}\) | \(m\) | \(\frac{1}{\beta}\) | | Afsn. 6.4 |
| Ex(\(\mu\))                  | \(\mu\)         | \(\mu^2\)   | \(\mu \in RGam(\alpha,1/\beta)\) | \(m = \frac{\beta}{\alpha - 1}\) | \(\frac{m^2}{1-\gamma}\) | \(\frac{1}{\alpha - 1}\) | | Afsn. 6.5 |
| N(\(\mu,\sigma^2\))          | \(\mu\)         | \(\sigma^2\) | \(N(m,\sigma^2)\) | \(m\) | \(\sigma^2\) | \(\frac{\sigma^2}{\sigma^2}\) | | Afsn. 5.3 |

\[E[\overline{X}_+] = m; \quad V[\overline{X}_+] = E[V(\mu)] \left(\gamma + \frac{1}{n}\right)\]

Tabel 6.1. Hierarkiske modeller for endimensionale eksponentielle familier med naturlige konjugerede aprioriforskelinger
| Stikprøvefordeling af $Z$ | Strukturfordeling $w(\cdot)$ | Marginalfordeling af $Z$ | $\mu = \mathbb{E}[X|\theta]$ | $\nu = \mathbb{V}[\mu]$ | $m = \mathbb{E}[\mu]$ | $\gamma = \frac{\mathbb{V}[\mu]}{\mathbb{E}[\mathbb{V}(\mu)]}$ | $\mathbb{E}[\mathbb{V}(\mu)]$ |
|------------------------|--------------------------|-----------------------|------------------|------------------|------------------|--------------------------|------------------|
| $B(n,p)$               | $p \in \text{Be}(\alpha, \beta)$ | $\text{Pl}(n,\alpha,\alpha + \beta)$ | $p$               | $\mu(1-\mu)$    | $\pi = \frac{\alpha}{\alpha + \beta}$ | $\frac{1}{\alpha + \beta}$ | $\frac{V(\pi)}{1 + \gamma}$ |
| $\text{NB}^*(n,p)$     | $p \in \text{Be}(\alpha, \beta)$ | $\text{NPI}^*(n,\beta,\alpha + \beta)$ | $\frac{p}{1-p}$ | $\mu(1+\mu)$    | $\psi = \frac{\alpha}{\beta - 1}$ | $\frac{1}{\beta - 1}$ | $\frac{V(\psi)}{1 - \gamma}$ |
| $P(n\mu)$              | $\mu \in \text{G}(\alpha, 1/\beta)$ | $\text{NB}(\alpha,\beta/(\beta + n))$ | $\mu$             | $\mu$            | $m = \frac{\alpha}{\beta}$ | $\frac{1}{\beta}$ | $V(m)$ |
| $G(n,\mu)$             | $\mu \in \text{RGam}(\alpha, 1/\beta)$ | $\text{RBet}(\alpha,n,\beta)$ | $\mu$             | $\mu^2$          | $m = \frac{\beta}{\alpha - 1}$ | $\frac{1}{\alpha - 1}$ | $\frac{V(m)}{1 - \gamma}$ |

$Z = X_1 + X_2 + \cdots + X_n; \quad Y = Z/n$

$\mathbb{E}[Y] = m \quad \mathbb{V}[Y] = \mathbb{E}[\mathbb{V}(\mu)] \left( \gamma + \frac{1}{n} \right)$

Tabel 6.2. Momenter i de marginale fordelinger ved hierarkisk variation
<table>
<thead>
<tr>
<th>Marginal fordeling af $Z$</th>
<th>$m = E[Y_i]$</th>
<th>$\gamma = \frac{V[\mu]}{E[V(\mu)]}$</th>
<th>$\tilde{m}$</th>
<th>$\tilde{\gamma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL($n, \alpha, \alpha + \beta$)</td>
<td>$\pi = \frac{\alpha}{\alpha + \beta}$</td>
<td>$\frac{1}{\alpha + \beta}$</td>
<td>$\tilde{\pi} = \bar{y}_+$</td>
<td>$\frac{s^2 - \bar{y}<em>+(1 - \bar{y}</em>+)}{n_0 \bar{y}<em>+(1 - \bar{y}</em>+) - s^2}$</td>
</tr>
<tr>
<td>NPI*($n, \beta, \alpha + \beta$)</td>
<td>$\psi = \frac{\alpha}{\beta - 1}$</td>
<td>$\frac{1}{\beta - 1}$</td>
<td>$\tilde{\psi} = \bar{y}_+$</td>
<td>$\frac{s^2 - \bar{y}<em>+(1 + \bar{y}</em>+)}{s^2 + n_0 \bar{y}<em>+(1 + \bar{y}</em>+)}$</td>
</tr>
<tr>
<td>NB($\alpha, \beta/(\beta + n)$)</td>
<td>$m = \frac{\alpha}{\beta}$</td>
<td>$\frac{1}{\beta}$</td>
<td>$\tilde{m} = \bar{y}_+$</td>
<td>$\frac{s^2/\bar{y}_+ - 1}{n_0}$</td>
</tr>
<tr>
<td>RBet($\alpha, n, \beta$)</td>
<td>$m = \frac{\beta}{\alpha - 1}$</td>
<td>$\frac{1}{\alpha - 1}$</td>
<td>$\tilde{m} = \bar{y}_+$</td>
<td>$\frac{s^2 - \bar{y}<em>+^2}{s^2 + n_0 \bar{y}</em>+^2}$</td>
</tr>
</tbody>
</table>

$Z_i = X_{i1} + X_{i2} + \cdots + X_{in_i}; \quad Y_i = Z_i/n_i$

$SAK_2 = \sum_{i=1}^{k} n_i(y_i - \bar{y}_+)^2; \quad s^2 = SAK_2/(k - 1)$

$\bar{y}_+ = \frac{\sum_{i=1}^{k} n_i y_i}{\sum_{i=1}^{k} n_i}$

$n_0$ bestemt ved (5.1.9)

Tabel 6.3. Momentestimation af $m$ og $\gamma$ ved hierarkisk variation
<table>
<thead>
<tr>
<th>Marginal fordeling af Z</th>
<th>$\tilde{\alpha}$</th>
<th>$\tilde{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Pl}(n, \alpha, \alpha + \beta)$</td>
<td>$\bar{y}<em>+ \frac{n_0 \bar{y}</em>+(1 - \bar{y}<em>+)}{s_2^2 - \bar{y}</em>+(1 - \bar{y}_+)}$</td>
<td>$(1 - \bar{y}<em>+) \frac{n_0 \bar{y}</em>+(1 - \bar{y}<em>+)}{s_2^2 - \bar{y}</em>+(1 - \bar{y}_+)}$</td>
</tr>
<tr>
<td>$\text{NPI}^*(n, \beta, \alpha + \beta)$</td>
<td>$\bar{y}<em>+ \left(1 + \frac{n_0 + 1}{s_2^2 - \bar{y}</em>+(1 + \bar{y}_+)}\right)$</td>
<td>$2 + \frac{n_0 + 1}{s_2^2 - \bar{y}<em>+(1 + \bar{y}</em>+)}$</td>
</tr>
<tr>
<td>$\text{NB}(\alpha, \beta/\beta + n)$</td>
<td>$n_0 \frac{\bar{y}<em>+^2}{s_2^2 - \bar{y}</em>+}$</td>
<td>$n_0 \frac{\bar{y}<em>+}{s_2^2 - \bar{y}</em>+}$</td>
</tr>
<tr>
<td>$\text{RBet}(\alpha, n, \beta)$</td>
<td>$1 + \frac{s_2^2 + n_0 \bar{y}<em>+^2}{s_2^2 - \bar{y}</em>+^2}$</td>
<td>$\bar{y}<em>+ \frac{s_2^2 + n_0 \bar{y}</em>+^2}{s_2^2 - \bar{y}_+^2}$</td>
</tr>
</tbody>
</table>

$Z_i = X_{i1} + X_{i2} + \cdots + X_{in_i}$; \hspace{1cm} $Y_i = Z_i/n_i$

$SAK_2 = \sum_{i=1}^{k} n_i (y_i - \bar{y}_+)^2$; \hspace{1cm} $s_2^2 = SAK_2/(k - 1)$

$\bar{y}_+ = \frac{\sum_{i=1}^{k} n_i y_i}{\sum_{i=1}^{k} n_i}$

$n_0$ bestemt ved (5.1.9)

Tabel 6.4. Momentestimation af $\alpha$ og $\beta$ ved hierarkisk variation
6.9 Referencer


Hierarkiske modeller for eksp. disp. modeller
Afsnit 7

Lineære normalfordelingsmodeller

7.1 Balancede regressionsmodeller med varierende koefficienter

7.1.1 Indledning

I eksempel 2.7.1 betragtede vi en regressionsmodel for et sæt normalt fordelt observationer med en forklarende variabel. Fremstillingen kan umiddelbart udvides til en model med $p$ kontinuerte kovariable (forklarende variable).

I dette afsnit vil vi - i lighed med betragtningerne i afsnit 5 og 6 - udvide denne model til at omfatte $k$ grupper af observationer. For hver observation foreligger der udover observationen $y$ en værdi af hver af de $p$ kovariable.

I analogi med de foregående afsnit vil vi dels betragte en systematisk model, dvs en model, hvor vi modellerer de $k$ observationssæt ved en regressionsmodel, hvor regressionskoefficienterne er karakteristiske for den pågældende
gruppe, og dels en tilfældig model, dvs en model, hvor regressionskoeficient-
ernes variation modelleres ved en fordeling, dvs hvor de udvalgte grupper
blot betragtes som en stikprøve fra en fordeling, og hvor interessen samler
sig om beskrivelse af fordelingen af regressionskoeficienter.

Vi vil indledningsvist forudsætte, at der er lige mange observationer, \( n \), i
hver af de \( k \) grupper, og vi vil yderligere antage, at værdierne af de \( p \) for-
klarende variable er de samme i alle \( k \) grupper, dvs. at de \( k \) observationssæt
har tilknyttet samme, kendte \((n \times p)\)-matrix

\[
X = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
\]  
(7.1.1)

af værdier \( x_{rj} \) af de \( p \) forklarende variable, hvor \( x_{rj}, r = 1, 2, \ldots, n, j = 1, 2, \ldots, p \) angiver den værdi af den \( j^{te} \) forklarende variabel, der svarer
til den \( r^{te} \) underobservation. Den \( j^{te} \) søjle i \( X \)-matricen repræsenterer
således værdierne af den \( j^{te} \) kovariabel jvf. afsnit 2.8.2.

Såfremt disse forudsætninger er opfyldt, siger vi, at modellen er balanceret.

For simpelheds skyld vil vi yderligere antage, at matricen \( X \) har fuld rang,
dvs at \((X^TX)^{-1}\) eksisterer.

Disse begrænsende forudsætninger indebærer, at notationen og estimatio-
nen forenkles. I praksis vil det være muligt at sikre disse forudsætninger
opfyldt i kontrollerede forsøg, hvor man kan styre (eller vælge) værdierne
af de forklarende variable. I situationer, hvor dette ikke kan lade sig gøre,
bliver modellen ubalanceret. I afsnit 7.2 vil vi behandle analysen af sådanne
ubalancerede modeller.

### 7.1.2 Den systematiske model

Vi betragter modellen

\[
H_0 : \quad Y_i = X_i \beta_i + \epsilon_i, \quad i = 1, 2, \ldots, k; \quad \epsilon_i \in N_n(0, \sigma^2 I_n),
\]  
(7.1.2)

hvor \( Y_i \) angiver de \( n \) observationer fra den \( i^{te} \) gruppe skrevet som en
\( n \)-dimensional søjlevektor; den fælles modellmatrix \( X \) er en kendt \( n \times p \) ma-
trix, og hvor \( \beta_i, i = 1, \ldots, k \) angiver en \( p \)-dimensional vektor af ukendte
regressionskoefficienter, og hvor de $n$-dimensionale vektorer $\epsilon_i$ af "observationsejler" er indbyrdes uafhængige, $i = 1, \ldots, k$.

Antagelsen $\epsilon_i \in \mathbb{N}_n(0, \sigma^2 I_n)$ indebærer, at observationsfejlene ikke kun er uafhængige imellem grupper, men at også de enkelte observationsfejl $\epsilon_{ir}$, $r = 1, \ldots, n$ indenfor samme gruppe er indbyrdes uafhængige.

Skrevet fuldt ud er modellen svarende til den $i$’te gruppe:

$$
\begin{pmatrix}
Y_{i1} \\
Y_{i2} \\
\vdots \\
Y_{in}
\end{pmatrix} =
\begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix}
\begin{pmatrix}
\beta_{i1} \\
\beta_{i2} \\
\vdots \\
\beta_{ip}
\end{pmatrix}
+ 
\begin{pmatrix}
\epsilon_{i1} \\
\epsilon_{i2} \\
\vdots \\
\epsilon_{in}
\end{pmatrix}
$$

Modellen er en generaliseret lineær model for normalfordelte observationer.

Modellen svarer til variansfunktionen $V(\mu) = 1$; dispersionsparameteren er $\sigma^2$, og linkfunktionen $\eta(\mu) = \mu$ er identiteten, dvs netop den kanoniske linkfunktion.

Modellen (7.1.2) kan udtrykkes ved modelformlen

$$Y = A.X_1 + A.X_2 + \cdots + A.X_p , \quad (7.1.3)$$

hvor $A$ symboliserer den variable (faktorvariabel), der angiver klassifikationen i de $k$ grupper, og $X_1, \ldots, X_p$ symboliserer de $p$ forklarende variable, og hvor prikoperatoren (afsnit 2.10) indikerer, at for hver af de forklarende variable $X_i$ er der en regressionskoefficient for hver af de $k$ grupper.

I modelformlen har vi ikke eksplcit tilgodeset en intercept-parameter. Med mindre man eksplicit specificerer, at man ikke ønsker nogen intercept, vil de fleste programsystemer indføre en fælles interceptparameter ved at tilføje en søjle bestående af lutter etaller til nedenstående modelmatrix (7.1.5) (for modellen for samtlige observationer).

Såfremt man ønsker en model med en individuel interceptparameter for hver af de $k$ grupper, skal man tilføje et led "+$A$" i modelformlen (7.1.3). En sådan individuel interceptparameter kan tilgodeses ved at lade første søjle i $X$-matricen bestå af etaller.

Modellen (7.1.2) har den parametriske fremstilling

$$\mu_{ir} = \beta_{i1} x_{r1} + \beta_{i2} x_{r2} + \cdots + \beta_{ip} x_{rp} , \quad (7.1.4)$$
for $i = 1, \ldots, k$, $r = 1, \ldots, n$.

Modellen udtrykker netop, at regressionskoefficienterne $\beta_i$ for de $k$ grupper tillades at være forskellige.

Modelmatricen for de $kn$ observationer er den $kn \times kp$-dimensionale matrix $X_0$ givet ved

$$X_0 = \begin{pmatrix} X & 0 & \cdots & 0 \\ 0 & X & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X \end{pmatrix}.$$  \hfill (7.1.5)

Modelmatricen (7.1.5) kan udtrykkes mere kompakt som

$$X_0 = X \otimes I_k$$  \hfill (7.1.6)

hvor $X \otimes I_k$ angiver tensorproduktet mellem den $n \times p$-dimensionale matrix $X$ og den $k \times k$-dimensionale enhedsmatrix $I_k$.

**Sætning 7.1.1 Maksimum likelihood estimation i den balancerede regressionsmodel**

Under modellen (7.1.2) er maksimum likelihood estimatoren for $\beta_i$ givet ved

$$\hat{\beta}_i = Py_i$$  \hfill (7.1.7)

med

$$P = (X^TX)^{-1}X^T$$  \hfill (7.1.8)

og den marginale maksimum likelihood estimator for $\sigma^2$ er
7.1 Balancerede regressionsmodeller

\[ \hat{\sigma}^2 = \frac{sak_1.}{k(n-p)} \]  

(7.1.9)

hvor

\[ sak_1. = \sum_{i=1}^{k} sak_{1,i} \]  

(7.1.10)

med

\[ sak_{1,i} = (y_i - X\hat{\beta}_i)^T(y_i - X\hat{\beta}_i) \]

Der gælder

\[ \hat{\beta}_i \in N_p(\beta_i, \sigma^2(X^TX)^{-1}) ; \quad SAK_{1.} \in \sigma^2 \chi^2(k(n-p)) \]  

(7.1.11)

og \( \hat{\beta}_i \) og \( SAK_{1.} \) er indbyrdes uafhængige.

**Bevis:**

Loglikelihoodfunktionen er

\[ l(\beta; y_1, \ldots, y_k) = -\frac{1}{2\sigma^2} \sum_{i=1}^{k} (y_i - X\beta_i)^T(y_i - X\beta_i) - (k/2) \ln[n\sigma^2] \]

Modellen (7.1.2) specifikserer relationen

\[ \mu = X_0\beta, \]  

(7.1.12)

mellem middelværdierne, hvor modelmatricen \( X_0 \) er givet ved (7.1.5).

Middelværdiligningen (2.7.1) svarende til modellen (7.1.12) er da

\[ X_0^T y = X_0^T X_0\beta. \]
Ligningen har løsningen
\[ \hat{\beta} = (X_0^T X_0)^{-1} X_0^T y. \] (7.1.13)

Indsætter vi nu \( X_0 = X \otimes I_k \) fra (7.1.6), får vi
\[ (X_0^T X_0)^{-1} = (X^T X)^{-1} \otimes I_k, \]

således at parameterprojektionsmatricen \( P_0 \), der projicerer den \( kn \)-dimensionale observatioensvektor \( Y \) ned på det \( kp \)-dimensionale parameterrum, kan udtrykkes som
\[
P_0 = [(X^T X)^{-1} \otimes I_k] [I_k \otimes X^T] \\
= [(X^T X)^{-1} X^T] \otimes I_k = P \otimes I_k,
\]

hvor vi har indført de individuelle parameterprojektionsmatricer \( P \) givet ved (7.1.8).

Udtrykket (7.1.13) for estimatet \( \hat{\beta} \) bliver da
\[
\hat{\beta} = P_0 Y = [P \otimes I_k] y,
\]
der netop spalter op i de \( k \) individuelle estimatorer (7.1.7), svarende til enkeltvis estimation af hver af de \( k \) regressionsplaner uden hensyntagen til observationer fra de øvrige grupper. (Dette er ikke overraskende, da modellen netop udsiger, at de \( k \) grupper har hver sin regressionsplan, og at observationerne er uafhængige af hinanden, svarende til at log-likelihoodfunktionen er en sum af \( k \) individuelle bidrag \(- (y_i - X_i \beta_i)^T (y_i - X_i \beta_i)\), der kan maksimeres enkeltvis med hensyn til \( \beta_i \).)

Tilsvarende finder man, at hat-matricen, der fører observatioensvektoren \( y \) over i de fittede værdier, \( \hat{\mu} \), bliver
\[
H_0 = X_0 P_0 = X_0 (X_0^T X_0)^{-1} X_0^T \\
= H \otimes I_k,
\]

hvor vi har indført hat-matricen \( H \) svarende til de individuelle regressionsplaner:
\[
H = X P = X (X^T X)^{-1} X^T. \] (7.1.14)

Vektoren \( R_0 \) af residualer bliver
\[
R_0 = y - X_0 \hat{\beta} = (I_{nk} - H_0) y = (I_n \otimes I_k - H \otimes I_k) y \\
= [(I_n - H) \otimes I_k] y.
\]
Den kvadratiske form \((I_{nk} - H_0)^T(I_{nk} - H_0)\) svarende til residualkvadratsummen (deviansen),

\[
s_{ak1} = D(y; \mu(\hat{\beta}) = (y - X_0\hat{\beta})^T(y - X_0\hat{\beta}) = R_0^T R_0
\]

spalter i en sum af ens former

\[
(I_n - H)^T(I_n - H),
\]

svarende til residualkvadratsummerne

\[
s_{ak1,i} = y_i^T(I_n - H)^T(I_n - H)y_i = (y_i - X_i\hat{\beta}_i)^T(y_i - X_i\hat{\beta}_i).
\]

Profilloglikelihood’en for \(\sigma^2\) (med hensyn til \(\beta\)) er

\[
\tilde{l}(\sigma^2; y) = -\frac{1}{\sigma^2} \sum_{i=1}^{k} s_{ak1,i} - (k/2) \ln(n\sigma^2),
\]

hvorfor \(s_{ak1} = \sum s_{ak1,i}\) er likelihood-sufficient for \(\sigma^2\). Man vil derfor benytte den marginale likelihood svarende til \(s_{ak1}\) og bestemme den marginale likelihood estimator for \(\sigma^2\).

For at betemme fordelingen af \(s_{ak1}\) bemærker vi, at da \(H\) er en projekionsmatrix, er den idempotent. Endvidere er \(H\) symmetrisk, og der gælder derfor

\[
(I_n - H)^T(I_n - H) = I_n - H^T - H + HH^T = I_n - H.
\]

Da \(X^TX\) er antaget at have fuld rang \(p\), har den kvadratiske form bestemt ved \((I_n - H)^T(I_n - H)\) rangen \(n - p\), og residualkvadratsummen \(s_{ak1,i}\) følger derfor en \(\sigma^2\chi^2(n - p)\)-fordeling, og fordelingen af \(s_{ak1,i}\) er uafhængig af \(\hat{\beta}_i = PY_i\), svarende til at residualerne \((I_n - H)y_i = y_i - X_i\hat{\beta}_i\) er ortogonale på underrummet udspændt af søjlerne i \(X\).

Da observationssættet \(y_i\) fra én gruppe er uafhængigt af observationssættet \(y_j\) fra en anden gruppe, er også residualkvadratsummerne \(s_{ak1,i}\) indbyrdes uafhængige. Det følger da, at fordelingen af summen \(s_{ak1}\) er en \(\sigma^2\chi^2(k(n - p))\)-fordeling. Den marginale maksimum likelihood for \(\sigma^2\) bliver derfor (7.1.9).

Da \(\hat{\beta}_i = Py_i\) er en lineær transformation af den \(N_n(X\beta_i, \sigma^2I_n)\)-fordelte størrelse \(y_i\), fås fordelingen af \(\hat{\beta}_i\) umiddelbart ved brug af transformationssætningen for normalfordelingen.
**Bemærkning 1 Estimation ved observationer med vilkårlig, kendt dispersionsmatrix**

I formuleringen af ovenstående sætning antog vi, at observationsfejlene $\xi_i$ havde samme varians $\sigma^2$ for alle værdier af de forklarende variable, og desuden, at observationsfejlene indenfor samme gruppe var indbyrdes uafhængige, nemlig at pånær dispersionsparameteren $\sigma^2$ kunne dispersionsmatricen for vektoren $\xi$ af observationsfejl beskrives ved en enhedsmatrix.

Såfremt der er seriel korrelation mellem observationerne i den enkelte gruppe, eller hvis den interne varians for forsøgsresultaterne afhænger af værdierne af de uafhængige variable, vil en sådan model med homogene observationsfejl imidlertid give en lovlig grov beskrivelse af de forhold, man ønsker at modellere.

For at tilgodese sådanne situationer, kan man udvide modellen (7.1.2) ved at tillade en mere generel varians-kovariansstruktur på bekostning af et lidt mere kompliceret formelapparat.

Den udvidede model fremkømmer ved at man i (7.1.2) erstatter antagelsen $\xi_i \in N_n(0, \sigma^2 I_n)$ med

$$H_0 : \quad Y_i = X\beta_i + \xi_i, \quad i = 1, 2, \ldots, k; \quad \xi_i \in N_n(0, \sigma^2 V),$$

(7.1.15)

hvor $V$ er en kendt symmetrisk positiv definit $n \times n$ matrix.

Modellen (7.1.15) svarer til en vægtet model. Hvis $V$ er en diagonalmatrix, har observation $Y_{ir}$ tilknyttet vægten $w_r = V_{rr}^{-1}$. I det generelle tilfælde er vægtnormen $W = V^{-1}$.

Under modellen (7.1.15) får man loglikelihoodfunktionen

$$l(\beta; y_1, \ldots, y_k) = -\frac{1}{2\sigma^2} \sum_{i=1}^{k} (y_i - X\beta_i)^T V^{-1} (y_i - X\beta_i) - (k/2) \ln[\sigma^2 \det(V)]$$

Maksimaliseringestimatoren er
\[ \hat{\beta}_i = P_w y_i \]  
(7.1.16)

med

\[ P_w = (X^T V^{-1}X)^{-1} X^T V^{-1} \]  
(7.1.17)

og den marginale maksimaliseringsestimator for \( \sigma^2 \) er givet ved (7.1.9), hvor

\[ s a k_{1,i} = (y_i - X\hat{\beta}_i)^T V^{-1} (y_i - X\hat{\beta}_i). \]  
(7.1.18)

Der gælder

\[ \hat{\beta}_i \in N_p (\beta_i, \sigma^2 (X^T V^{-1}X)^{-1}); \quad S A K_{1,i} \in \sigma^2 \chi^2 (k(n-p)) \]

\[ \Box \]

En nærliggende reduktion af modellen (7.1.2) er en hypotese om at regressionskoeficienterne er de samme i alle \( k \) grupper, dvs.

\[ H_1 : \beta_1 = \beta_2 = \cdots = \beta_k (= \beta_0), \]

imod alternativet

\[ H_1^c : \beta_i \neq \beta_j \text{ for mindst et sæt } (i,j) \text{ med } i \neq j \]

Hypotesen svarer til modelformlen

\[ Y = X_1 + X_2 + \cdots + X_p \]

med den parametriske fremstilling

\[ \mu_{ir} = \beta_1 x_{r1} + \beta_2 x_{r2} + \cdots + \beta_p x_{rp} \]

Hypotesen er en delmodel af (7.1.2).

Der gælder:
Sætning 7.1.2 Test for fælles regressionskoefficienter

Under modellen (7.1.2) har kvotienttestet for hypotesen

\[ H_1 : \beta_1 = \beta_2 = \cdots = \beta_k (= \beta_0), \]  
(7.1.19)

teststørrelsen

\[ Z = \frac{SAK_2 / [p(k-1)]}{SAK_1 / (k(n-p))} \]  
(7.1.20)

hvor

\[ SAK_2 = \sum_{i=1}^{k} (\hat{\beta}_i - \hat{\beta}_0)^T X^T X (\hat{\beta}_i - \hat{\beta}_0) \]  
(7.1.21)

med

\[ \hat{\beta}_0 = P \left( \sum_{i=1}^{k} y_i / k \right), \]  
(7.1.22)

hvor parameterprojektionsmatricen \( P \) er givet ved (7.1.8).

Under \( H_1 \) følger \( Z \) en \( F(p(k-1), k(n-p)) \)-fordeling. Testet forkaster for store værdier af \( z \).

Bevis:
Hypotesen \( H_1 \) svarer til at observationerne kan beskrives som \( k \) identisk fordelte gentagelser af \( Y_i \in N_n(X\beta_0, \sigma^2 I_n) \).

Under \( H_1 \) er modelmatricen den \((kn \times p)\)-dimensionale matrix

\[ M_1 = \begin{pmatrix} X \\ X \\ \vdots \\ X \end{pmatrix} = X \otimes 1_k, \]
hvor $1_k$ angiver en $k$-dimensional søjlevektor med etaller.

Der gælder

$$M_1^T M_1 = (X^T \otimes 1_k^T) (X \otimes 1_k) = (X^T X) \otimes (1_k^T 1_k) = k(X^T X),$$

hvorfor man finder parameterprojektionsmatricen

$$P_1 = (M_1^T M_1)^{-1} M_1^T = \frac{1}{k} (X^T X)^{-1} X^T \otimes 1_k = \frac{1}{k} P \otimes 1_k,$$

der netop fører til at estimatet $\hat{\beta}_0$ bestemmes som regressionen på de observerede gennemsnit

$$\bar{y}_r = \sum_{i=1}^k y_{ir}/k$$

svarende til estimatet (7.1.22)

Deviansopspaltningen svarende til den hierarkiske modelreduktion er vist i tabel 7.1 (jvf tabel 2.10)

Deviansopspaltningen svarer til spaltningen

$$\sum_{i=1}^k (y_i - X\hat{\beta}_0)^T (y_i - X\hat{\beta}_0) = sak_1 + sak_2$$

af residualkvadratsummen $G^2(H_1)$

$$G^2(H_1) = SAK_0 - (y - M_1 \hat{\beta}_0)^T (y - M_1 \hat{\beta}_0)$$

svarende til $H_1$.

Det følger af spaltningssætning, at under $H_1$ vil $SAK_2 \in \sigma^2 \chi^2(p(k - 1))$ og at $SAK_2$ er uafhængig af $SAK_1$.

□

**Bemærkning 1 Variansanalyseskema**

Analysen opskrives ofte i et variansanalyseskema svarende til opspaltningen (7.1.23) af kvadratsummen af residualerne omkring den fælles regressionsplan.
<table>
<thead>
<tr>
<th>Variationskilde</th>
<th>$f$</th>
<th>Devians</th>
<th>middeldevians</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mellem $H_1$ og $H_0$</td>
<td>$p(k-1)$</td>
<td>$D(\mu(\widehat{\beta});\mu(\widehat{\beta}_0))$</td>
<td>$\frac{D(\mu(\widehat{\beta});\mu(\widehat{\beta}_0))}{p(k-1)}$</td>
<td>$\frac{D(\mu(\widehat{\beta});\mu(\widehat{\beta}_0))/[p(k-1)]}{D(y;\mu(\widehat{\beta}_0))/[k(n-p)]}$</td>
</tr>
<tr>
<td>Afvigelse fra $H_0$</td>
<td>$k(n-p)$</td>
<td>$D(y;\mu(\widehat{\beta}))$</td>
<td>$\hat{\sigma}^2 = \frac{D(y;\mu(\widehat{\beta}))}{k(n-p)}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$kn-p$</td>
<td>$D(y;\mu(\widehat{\beta}_0))$</td>
<td>$\frac{D(y;\mu(\widehat{\beta}_0))}{kn-p}$</td>
<td></td>
</tr>
</tbody>
</table>

Tabel 7.1. Deviansopspaltning svarende til hierarkisk modelreduktion af regressionsmodellen i sætning 7.1.2
7.1 Balancerede regressionsmodeller

<table>
<thead>
<tr>
<th>Variation</th>
<th>SAK</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mellem regresionsplaner</td>
<td>$sak_2 = \sum_{i=1}^{k} (\hat{\beta}_i - \hat{\beta}_0)^T X^T X (\hat{\beta}_i - \hat{\beta}_0)$</td>
<td>$p(k-1)$</td>
</tr>
<tr>
<td>Omkring individuelle planer</td>
<td>$sak_1 = \sum_{i=1}^{k} (y_i - X\hat{\beta}_i)^T (y_i - X\hat{\beta}_i)$</td>
<td>$k(n-p)$</td>
</tr>
<tr>
<td>Omkring fælles plan</td>
<td>$\sum_{i=1}^{k} (y_i - X\hat{\beta}_0)^T (y_i - X\hat{\beta}_0)$</td>
<td>$kn-p$</td>
</tr>
</tbody>
</table>

Det fremgår ved sammenligning af udtrykket (7.1.21) for $SAK_2$ med udtrykket (7.1.11) for fordelingen af $\hat{\beta}_i$, at afvigelserne $(\hat{\beta}_i - \hat{\beta}_0)$ vægtes med deres præcisioner $X^T X$ (den inverse dispersionsmatrix).

**Bemærkning 2 Test for samme indflydelse af enkelte af de forklarende variable**

Ovenstående test sammenligner hele parametervektoren $\beta$ for de $k$ grupper. Testet forkaster såfremt blot een af komponenterne er forskellig fra gruppe til gruppe. Ved succesiv testning kan man undersøge forskelle mellem grupper for de enkelte komponenter af parametervektoren. Vi skal dog ikke komme nærmere ind herpå i denne fremstilling.

**Bemærkning 3 Test ved observationer med vilkårlig, kendt dispersionsmatrix**

Under den generelle model (7.1.15) finder man estimatet for de fælles regressionskoefficienter $\beta_0$ under hyphotesen (7.1.19)

$$\hat{\beta}_0 = P_w[y_i/k],$$  \hspace{1cm} (7.1.24)

hvor parameterprojektionsmatricen, $P_w$, er givet ved (7.1.17), og teststørrelsen for hyphotesen (7.1.19)

$$Z = \frac{SAK_2/[p(k-1)]}{SAK_1./(N-pk)}$$
hvor
\[ SAK_2 = \sum_{i=1}^{k}(\hat{\beta}_i - \hat{\beta}_0)^T X^T V^{-1} X (\hat{\beta}_i - \hat{\beta}_0) \] (7.1.25)
med \( sak_1 \) bestemt ved brug af (7.1.18).

Under \( H_1 \) følger \( Z \) en \( F(p(k-1), k(n-p)) \)-fordeling. Testet forkaster for store værdier af \( z \).

7.1.3 Den tilfældige model

Såfremt man ønsker at modellere eventuelle forskelle mellem grupper ved en tilfældig model, kan man vælge at udbygge ovenstående model med antagelsen \( \beta_i \in N_p(\beta_0, \sigma^2 \Gamma) \), hvor \( \beta_0 \) angiver den ukendte middel-parametervektor, og den \( p \times p \)-dimensionale symmetriske, positiv definite matrix \( \sigma^2 \Gamma \) angiver dispersionsmatricen for fordelingen af \( \beta_i \) omkring \( \beta_0 \). For en ordens skyld gør vi opmærksom på, at det ikke er nogen indskrænkning i modellen, at vi har valgt at skalere dispersionsmatricen med faktoren \( \sigma^2 \).

Fordelingsforholdene under denne model fremgår af

Sætning 7.1.3 Marginal fordeling af observationer og estimerer under regressionsmodel med tilfældige koefficienter

Lad
\[ Y_i = X \beta_i + \epsilon_i, \ i = 1, 2, \ldots, k \]

hvor
\[ \beta_i \in N_p(\beta_0, \sigma^2 \Gamma), \ \epsilon_i \in N_n(0, \sigma^2 I_n), \] (7.1.26)

og hvor \( \beta_i, \beta_j \) er indbyrdes uafhængige for \( i \neq j \), og \( \epsilon_i \) og \( \epsilon_j \) er indbyrdes uafhængige for \( i \neq j \), og endvidere \( \beta_i \) og \( \epsilon_j \) er uafhængige.

Da er den marginale fordeling af \( Y_i \) givet ved
\[ Y_i \in N_n(X \beta_0, \sigma^2 \{I_n + X \Gamma X^T\}) , \]
og den marginale fordeling af estimaterne, \( \hat{\beta}_i \) (7.1.7), for regressionskoefficienterne er
\[ \hat{\beta}_i \in \mathcal{N}_p(\beta_0, \sigma^2[(X^T X)^{-1} + \Gamma]) \]  

(7.1.27)

**Bevis:**
Udtrykket for den marginale dispersionsmatrix fås ved at bemærke, at

\[ D [Y_i] = E [D [Y_i|\beta]] + D [E [Y_i|\beta]] = \sigma^2 I_n + D [X\beta] \]

jvf Sætning 0.1.1, formel (0.1.2) i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

\[ \square \]

**Bemærkning 1** *Fordelingen af estimaterne for regressionskoefficienterne omfatter både estimationsusikkerheden og variansen mellem grupper*

Fordelingen af estimaterne \( \hat{\beta}_i \) er analog til fordelingen af gruppegennemsnit i de modeller, vi betragtede i afsnit 5 og 6. Fordelingen indeholder et bidrag, \( \sigma^2 (X^T X)^{-1} \) svarende til estimationsusikkerheden, og et bidrag, \( \sigma^2 \Gamma \) svarende til dispersionen mellem gruppernes regressionskoefficienter. Da estimatet \( \hat{\beta}_i \), der beskriver den \( i'te \) gruppe, i dette tilfælde er en \( p \)-dimensional størrelse, beskrives usikkerheden og dispersionen ved \( p \)-dimensionale dispersionsmatricer i analogi med modellen for den flerdimensionale normalfordeling, der blev betragtet i afsnit 6.7.2.

\[ \square \]

**Bemærkning 2** *Fordeling af estimator ved observationer med villkårlig, kendt dispersionsmatrix*

Såfremt man i (7.1.26) antager den mere generelle fordeling af observationsfejlene,

\[ \epsilon_i \in \mathcal{N}_n(0, \sigma^2 V), \]  

(7.1.28)

hvor \( V \) er en kendt symmetrisk positiv definit \( n \times n \) matrix, da gælder at den marginale fordeling af \( Y_i \) er

\[ Y_i \in \mathcal{N}_r(X\beta_0, \sigma^2 \{V + \Sigma_{\Gamma}X^T\}), \]
og den marginale fordeling af estimaterne, \( \hat{\beta}_i \) (7.1.16), for regressionskoefficienterne er

\[
\hat{\beta}_i \in N_p(\beta_0, \sigma^2 \{(X^T V^{-1} X)^{-1} + \Gamma\})
\]  

(7.1.29)

Sætning 7.1.4 Maksimaliseringsestimation for regressionsmodel med tilfældige koefficienter

Under antagelserne fra Sætning 7.1.3 fås maksimaliseringsestimatorerne for parametrene \( \beta_0, \sigma^2 \) og \( \Gamma \) ved at maksimere

\[
l(\beta_0, \sigma^2, \Gamma; y_1, \ldots, y_k) = -\frac{k}{2} \left\{ n \ln(\sigma^2) + \ln(\det(D(\Gamma))) \right\}
\]

\[+ \frac{1}{\sigma^2} \sum_{i=1}^{k} (y_i - X\beta_0)^T (D(\Gamma))^{-1} (y_i - X\beta_0) \]

hvor

\[
D(\Gamma) = X\Gamma X^T + I_n
\]

(7.1.31)

med hensyn til \( \sigma^2 > 0, \beta_0 \in \mathbb{R}^p \) og \( \Gamma \) positiv semidefinit.

Bevis:

Sætningen vises ved at bemærke, at (7.1.30) netop er logaritmen til likelihoodfunktionen.

Bemærkning 1 Likelihoodligningerne ved maksimum i et inter punkt

Såfremt maksimumværdien findes i et indre punkt, tilfredsstiller estima-terne udtrykkene

\[
\hat{\beta}_0 = \frac{1}{k} \left[ X^T (D(\Gamma))^{-1} X \right]^{-1} X^T (D(\Gamma))^{-1} \left( \sum_{i=1}^{k} y_i \right) \]

(7.1.32)

\[
\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{k} (y_i - X\hat{\beta}_0)^T (D(\hat{\Gamma}))^{-1} (y_i - X\hat{\beta}_0) \]

(7.1.33)
hvor \( \hat{\Gamma} \) er bestemt som den positive semidefinitte \( p \times p \)-matrix, der maksimører udtrykket

\[
l(\Gamma) = -N \times \ln(sak^*(\Gamma)) - k \times \ln[\det(D(\Gamma))] \tag{7.1.34}
\]

hvor

\[
sak^*(\Gamma) = \sum_{i=1}^{k} y_i^T (D(\Gamma))^{-1} y_i
\]

\[
\frac{1}{k} \left( \sum_{i=1}^{k} y_i^T \right) (D(\Gamma))^{-1} X [X^T (D(\Gamma))^{-1} X]^{-1} X^T (D(\Gamma))^{-1} \left( \sum_{i=1}^{k} y_i \right)
\]

Maksimaliseringsestimatet for \( \Gamma \) kan da bestemmes ved en numerisk bestemmelse af maksimum for (7.1.34) over rummet af symmetriske positive semidefinite \( p \times p \)-matrizer \( \Gamma \). Maksimaliseringsestimaterne for de øvrige parametre kan heretter bestemmes ved at indsætte den fundne værdi i (7.1.32) og (7.1.33). (Se f.eks. Thyregod (1983))

\[\square\]

**Bemærkning 2 Central estimator, REML-estimation** Det er velkendt, at maksimaliseringsestimatoren ikke nødvendigvis er central. Såfremt \( \Gamma \) var kendt, ville estimatoren

\[
\hat{\sigma}^2 = \frac{1}{N-p} \sum_{i=1}^{k} (y_i - X\hat{\beta}_0)^T (D(\hat{\Gamma}))^{-1} (y_i - X\hat{\beta}_0)
\]

være central for \( \sigma^2 \). Man benytter derfor ofte estimatet baseret på maksimering af likelihoodfunktionen svarende til fordelingen af residualerne (svarende til \( X\hat{\beta} \)), det såkaldte REML-estimat (se bemærkning 3 til sætning 5.4.1)

\[\square\]

**Bemærkning 3 Estimation ved observationer med vilkårlig, kendt dispersionsmatrix**

Sætningen og bemærkningerne gælder også i den generelle situation, hvor dispersionsmatricen for observationsfejlene \( \epsilon_i \) er \( \sigma^2 V \). Man skal blot erstatte udtrykket (7.1.31) for \( D(\Gamma) \) med

\[
D(\Gamma) = X\Gamma X^T + V \tag{7.1.35}
\]
Sætning 7.1.5 *Momentestimation for regressionsmodel med tilfældige koefficienter*

Under antagelserne fra sætning 7.1.3 bestemmes momentestimatorerne for parametrene $\beta_0$, $\sigma^2$ og $\Gamma$ ved

$$\tilde{\beta}_0 = \frac{1}{k} \sum_{i=1}^{k} \beta_i = \mathbf{P} \left( \sum_{i=1}^{k} y_i/k \right) \quad (7.1.36)$$

$$\tilde{\sigma}^2 = \frac{1}{k(n-p)} \text{sak}_1. \quad (7.1.37)$$

$$\tilde{\Gamma} = \frac{\text{sak}_\beta}{(k-1)\tilde{\sigma}^2} - (\mathbf{X}^T \mathbf{X})^{-1}, \quad (7.1.38)$$

hvor $\beta_i$, $\mathbf{P}$ og $\text{sak}_1$ er givet ved (7.1.8), (7.1.7) og (7.1.10), og hvor

$$\text{sak}_\beta = \sum_{i=1}^{k} (\beta_i - \tilde{\beta}_0)(\beta_i - \tilde{\beta}_0)^T$$

Der gælder, at

$$\text{sak}_\beta \in \text{Wis}_p(k-1, \sigma^2[(\mathbf{X}^T \mathbf{X})^{-1} + \Gamma])$$

og at

$$\tilde{\beta}_0 \in \text{N}_p(\beta_0, (\sigma^2/k)[(\mathbf{X}^T \mathbf{X})^{-1} + \Gamma])$$

er central og variansminimal.

**Bevis:**

Sætningen vises ved at bemærke, at $\tilde{\beta}_0$ er gennemsnittet, og $\text{sak}_\beta/(k-1)$ er den empiriske dispersionsmatrix for $\beta_i$.

Idet $\beta_i$ er indbyrdes uafhængige med

$$\mathbb{E} [\hat{\beta}] = \beta_0$$

og

$$\mathbb{D} [\hat{\beta}_i] = \sigma^2[(\mathbf{X}^T \mathbf{X})^{-1} + \Gamma])$$
ifølge (7.1.27), har vi

\[
\begin{align*}
E[\hat{\beta}_0] &= \beta_0, \quad \text{og} \\
E[\text{sak}_\beta] &= \left(1 - \frac{1}{k}\right) \sum_{i=1}^{k} \mathbf{D}[\hat{\beta}_i] = (k - 1)\sigma^2[(X^T X)^{-1} + \Gamma]
\end{align*}
\]

\[\square\]

Bemærkning 1 Estimaterne for \(\sigma^2\) og \(\beta\) udnytter også variationen mellem grupper

Sammenligner vi momentestimaterne (7.1.36) og (7.1.37) for \(\beta_0\) og \(\sigma^2\) med maksimaliseringsestimaterne (7.1.32) og (7.1.33) ser vi, at i modsætning til momentestimaterne tilgodeser maksimaliseringsestimaterne for \(\sigma^2\) og \(\Gamma\) kendskabet til dispersionsmatricen \(\beta\) ved den indbyrdes vægtning af de enkelte komponenter af \(\beta_i\).

Vi vil her ikke komme nærmere ind på test af hypotesser vedrørende \(\Gamma\). I Bondeson (1989) er en række forskellige tests diskuteret.

\[\square\]

Bemærkning 2 Estimation ved observationer med vilkårlig, kendt dispersionsmatrix

Sætningen udvides let til at dække også i den generelle situation, hvor dispersionsmatricen for observationsfejlene \(\epsilon_i\) er \(\sigma^2 \mathbf{V}\).

Lad modellen være som i sætning 7.1.3, men lad fordelingen af \(\epsilon_i\) være givet ved (7.1.28).

Da fås momentestimaterne for \(\beta_0\), \(\sigma^2\) og \(\Gamma\) ved
\[
\tilde{\beta}_0 = \frac{1}{k} \sum_{i=1}^{k} \tilde{\beta}_i = P_w \left( \sum_{i=1}^{k} y_i / k \right) \quad (7.1.39)
\]

\[
\tilde{\sigma}^2 = \frac{1}{k(n - p)} \text{sak}_1. \quad (7.1.40)
\]

\[
\tilde{\Gamma} = \frac{\text{sak}_\beta}{(k - 1)\tilde{\sigma}^2} - (X^T V^{-1} X)^{-1}, \quad (7.1.41)
\]

hvor \( \tilde{\beta}_i \), \( P_w \) og \( \text{sak}_1 \). er givet ved (7.1.17), (7.1.16) og (7.1.18), og hvor

\[
\text{sak}_\beta = \sum_{i=1}^{k} (\tilde{\beta}_i - \tilde{\beta}_0)(\tilde{\beta}_i - \tilde{\beta}_0)^T
\]

Der gælder, at

\[
\text{sak}_\beta \in \text{Wis}_p(k - 1, \sigma^2[(X^T V^{-1} X)^{-1} + \Gamma])
\]

g og at

\[
\tilde{\beta}_0 \in \text{N}_p(\beta_0, (\sigma^2/k)[(X^T V^{-1} X)^{-1} + \Gamma])
\]
er central og variansminimal.

\[\square\]

**Eksempel 7.1.1 Vækst af Ramus-knoglen for 5 drenge, momentestimation**


Tabel 7.2 viser samhørende værdier af højden af kæbebenet (ramus) i [mm] og alderen i [år] hos 5 drenge i aldersgruppen 8-10 år. For hver dreng er ramushøjden registreret fire gange, nemlig ved alder 8, 8 1/2, 9 og 9 1/2 år.

I eksempel 2.7.3 betragtede vi observationssættet svarende til dreng B.

I eksempel 2.7.3 valgte vi at modellere ramushøjden ved den reducerede alder \( x_i \) ved en normalfordelt størrelse med en middelværdi, der var en lineær funktion af den reducerede alder, \( x_i \).
Tabel 7.2. Tabel over samhørende værdier af ramus højde (i mm.) og alder for 5 drenge.

<table>
<thead>
<tr>
<th>Dreng</th>
<th>Alder i år</th>
<th>( x_i = \text{alder} - 8.75 )</th>
<th>8</th>
<th>8 1/2</th>
<th>9</th>
<th>9 1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>52.5</td>
<td>53.2</td>
<td>53.3</td>
<td>53.7</td>
<td>53.175</td>
<td>0.74</td>
</tr>
<tr>
<td>B</td>
<td>51.2</td>
<td>53.0</td>
<td>54.3</td>
<td>54.5</td>
<td>53.250</td>
<td>2.24</td>
</tr>
<tr>
<td>C</td>
<td>51.2</td>
<td>51.4</td>
<td>51.6</td>
<td>51.9</td>
<td>51.525</td>
<td>0.46</td>
</tr>
<tr>
<td>D</td>
<td>52.1</td>
<td>52.8</td>
<td>53.7</td>
<td>55.0</td>
<td>53.400</td>
<td>1.92</td>
</tr>
<tr>
<td>E</td>
<td>50.7</td>
<td>51.7</td>
<td>52.7</td>
<td>53.3</td>
<td>52.100</td>
<td>1.76</td>
</tr>
<tr>
<td>snit</td>
<td>51.54</td>
<td>52.42</td>
<td>53.12</td>
<td>53.68</td>
<td>52.690</td>
<td>1.424</td>
</tr>
</tbody>
</table>

Tilsvarende vil vi her modellere relationen mellem ramushøjde og alder ved en lineær regressionsmodel, der tillader en individuel vækstrelation for hver drong. Viformulerer modellen

\[ Y_{ij} = \beta_{i1} + x_{ij}\beta_{i2} + \epsilon_{ij}, \ i = 1, 2, \ldots, 5; \ j = 1, 2, 3, 4, \]

hvor \( Y_{ij} \) angiver den registrerede ramushøjde i [mm] hos den \( i \)’te dreng ved den reducerede alder \( x_{ij} \), og hvor \( \epsilon_{ij} \) antages uafhængige identisk normalfordelte \( N(0, \sigma^2) \).

Vi har altså modelmatricen

\[ X = \begin{pmatrix} 1.0 & -0.75 \\ 1.0 & -0.25 \\ 1.0 & 0.25 \\ 1.0 & 0.75 \end{pmatrix} \]

med \( X^T X = \begin{pmatrix} 4.00 & 0.00 \\ 0.00 & 1.25 \end{pmatrix} \)

hvorfor

\[ (X^T X)^{-1} = \begin{pmatrix} 0.25 & 0.00 \\ 0.00 & 0.80 \end{pmatrix} \]

Skønnene over de individuelle regressionkoefficienter \( \hat{\beta}_i \) beregnet ved \( \hat{\beta}_i = (X^T X)^{-1} X^T y_i \) er anført i tabel 7.2.

De observerede værdier er afbildet i figur 7.1. I figuren er også de estimerede regressionslinier indtegnet.
Figur 7.1. Samhørende værdier af alder og ramushøjde for 5 drenge. (Data fra tabel

Ved benyttelse af (7.1.10) finder man residualkvadratsummen

\[ sak_1 = 0.8640 \text{ [mm]}^2 \]

hvorfor man får skønnet over variansen omkring de individuelle linier

\[ \hat{\sigma}^2 = sak_1 / 10 = 0.0864 = (0.2939 \text{ [mm]})^2 \]

Den totale variation, \( sak_0 \) kan findes som residualkvadratsummen svarende til den fælles regression. Man finder

\[ sak_0 = 14.9244 \text{ [mm]}^2 \]
hvorfor man får variationen mellem drengenes regressionslinier

\[ sak_2 = sak_0 - sak_1 = 14.0604 \text{ [mm]}^2 \]

Resultatet kan samles i variansanalyseskemaet

<table>
<thead>
<tr>
<th>Variation</th>
<th>( sak )</th>
<th>( f )</th>
<th>( sak/f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mellem individuelle</td>
<td>14.0604</td>
<td>8</td>
<td>1.7575</td>
</tr>
<tr>
<td>regressionslinier</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omkring individuelle</td>
<td>0.8640</td>
<td>10</td>
<td>0.0864</td>
</tr>
<tr>
<td>regressionslinier</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omkring fælles linie</td>
<td>14.9244</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>

Under den systematiske model bliver teststørrelsen for fælles regressionslinie \( z = 20.34 \), der sammenlignes med en \( F(8, 10) \)-fordeling. Den fundne værdi overstiger langt \( F_{0.999}(8, 10) \)-fraktilen, og der er derfor ingen grund til at opretholde en hypotese om en fælles vækstkurve.

Den fælles regressionslinie udtrykker imidlertid en sammenhæng af generel interesse, og vi vælger derfor at benytte en model med tilfældig variation mellem koefficienterne, d.v.s. vi vil antage at \( \beta_i \in \mathcal{N}(\beta_0, \sigma^2 \Gamma) \)

Figur 7.2 viser de samhørende værdier af de estimerede regressionsparametre.

Til estimation af parametrene i denne model beregner vi først momentskønnet over dispersionsmatricen for \( \beta_i \). Man får

\[ sak_\beta = \begin{pmatrix} 2.4219 & 1.4022 \\ 1.4022 & 2.7583 \end{pmatrix} \]

Niveaukurverne svarende til denne fordeling er indtegnet i figur 7.2.

Vi finder nu momentskønnet over \( \Gamma \)

\[ \Gamma = \frac{sak_\beta}{4 \times 0.0864} - (X^TX)^{-1} = \begin{pmatrix} 7.9812 & 4.0573 \\ 4.0573 & 7.0072 \end{pmatrix} - \begin{pmatrix} 0.25 & 0.00 \\ 0.00 & 0.80 \end{pmatrix} = \begin{pmatrix} 7.7312 & 4.0573 \\ 4.0573 & 6.2072 \end{pmatrix} \]

Der er således en positiv samvariation mellem ramushøjden ved 8.75 år, og vækstraten.
Figur 7.2. Samhørende værdier af vækstrate og ramushøjde ved 8.75 år for 5 drenge.

Til vurdering af samvariationens størrelse, kan man bestemme korrelationskoefficienten

$$\hat{\rho} = \frac{4.0573}{\sqrt{7.7312 \times 6.2072}} = 0.54$$

Vi kan altså beskrive ramushøjderne for 8-10 årig drenge ved en lineær funktion af alderen. Den gennemsnitlige ramushøjde ved alderen 8.75 år er estimeret til 52.69 [mm], og den gennemsnitlige vækstrate er estimeret til 1.42 [mm/år].

Drengene har forskellige ramushøjder ved alderen 8.75 år, og de har også hver sin vækstrate. Residualvariancen omkring de individuelle regressionslinier er estimeret til $$\hat{\sigma}^2 = (0.2939 \text{ [mm]})^2$$. 
Der er en positiv samvariation mellem ramushøjde ved alder 8.75 [år] og vækstrate. Den estimerede varians-kovariansmatrix for samvariationen mellem ramushøjde ved alder 8.75 [år] og vækstrate er

\[
\hat{\sigma}^2 \tilde{\Gamma} = 0.2939^2 \begin{pmatrix} 7.7312 & 4.0573 \\ 4.0573 & 6.2072 \end{pmatrix} = \begin{pmatrix} 0.8173^2 & 0.3506 \\ 0.3506 & 0.7323^2 \end{pmatrix}
\]


Eksempel 7.1.2 Vækst af Ramus-knoglen for 5 drenge, estimation ved SAS® proceduren MIXED

Analysen i det foregående eksempel kunne også udføres i programsystemet SAS® ved brug af proceduren MIXED, som vi betragtede i afsnit 5.5.2.

Antag at data fra eksemplet er indlæst i de variable ramus, dreng og ald.

Programmet:

```latex
PROC MIXED METHOD=ML ASYCOV;
CLASS dreng;
MODEL ramus= ald/ S ;
RANDOM INTERCEPT ald /TYPE=UN SUB=dreng ;
RUN;
```

kalder procedure MIXED. Nøgleordet METHOD =ML angiver, at man ønsker maksimaliseringsestimaterne, og ordet ASYCOV angiver, at man ønsker den asymptotiske varians-kovariansmatrix for estimaterne.

I modelformlen MODEL ramus= ald/ S ; specificeres de systematiske effekter i modellen. Modellen indeholder de to systematiske parametre, \( \beta_0 \) (7.1.26), nemlig middelværdien af INTERCEPT og af koefficienten til ald. Optionen S efter skråstregen angiver, at vi ønsker en løsning (solution), dvs. estimatoren for de systematiske effekter udskrevet.

Sætningen RANDOM INTERCEPT ald /TYPE=UN SUB=dreng ; angiver, at vi vil modellere såvel interceptparametrene som koefficienten til alder som tilfældige effekter. Optionen TYPE=UN specificerer at dispersionsmatricen svarende til hele observationssættet

\[
D [Y] = \sigma^2 [I_n + \Gamma\Gamma^T] \otimes I_k
\]
ikke er en diagonalmatrix, og endelig angiver optionen \texttt{SUB=dreng}, at observationer svarende til forskellige værdier af \texttt{dreng} er uafhængige, dvs. at
\[
D \{Y\} \text{ er en blokdiagonalmatrix.}
\]

Proceduren giver anledning til følgende udskrift

\begin{quote}
\textbf{The MIXED Procedure}
\end{quote}

\begin{quote}
\textbf{Class Level Information}
\end{quote}

\begin{quote}
\textbf{Class} \hspace{1cm} \textbf{Levels} \hspace{1cm} \textbf{Values}
\end{quote}

\begin{quote}
\texttt{DRENG} \hspace{1cm} 5 \hspace{0.5cm} A \hspace{0.5cm} B \hspace{0.5cm} C \hspace{0.5cm} D \hspace{0.5cm} E
\end{quote}

\begin{quote}
til kontrol af de indlæste specifikationer
\end{quote}

\begin{quote}
Endvidere udskrives iterationsforløbet
\end{quote}

\begin{quote}
\textbf{The MIXED Procedure}
\end{quote}

\begin{quote}
\textbf{ML Estimation Iteration History}
\end{quote}

\begin{quote}
\begin{tabular}{|c|c|c|c|c|}
\hline
\textbf{Iteration} & \textbf{Evaluations} & \textbf{Objective} & \textbf{Criterion} \\
\hline
0 & 1 & 14.14530368 & \\
1 & 1 & -4.78207975 & 0.00000000 \\
\hline
\end{tabular}
\end{quote}

\begin{quote}
Convergence criteria met.
\end{quote}

Da data er balancerede behøves kun én iteration.

Derefter udskrives estimatorne for \(\sigma^2\) og \(\Gamma\)

\begin{quote}
\textbf{Covariance Parameter Estimates (MLE)}
\end{quote}

\begin{quote}
\begin{tabular}{|c|c|c|}
\hline
\textbf{Cov Parm} & \textbf{Subject} & \textbf{Estimate} \\
\hline
UN(1,1) & DRENG & 0.53005000 \\
UN(2,1) & DRENG & 0.28044000 \\
UN(2,2) & DRENG & 0.41526400 \\
Residual & & 0.08640000 \\
\hline
\end{tabular}
\end{quote}
7.1 Balancerede regressionsmodeller

Estimaterne står i søjlen betegnet \textit{estimate}.

Estimatet \( \hat{\sigma}^2 \) står i linien \textit{Residual}, og elementerne i \( \hat{\sigma}^2 \Gamma \) står i rubrikkerne svarende til hhv. \textit{UN(1,1)}, \textit{UN(2,1)} og \textit{UN(2,2)}.

Da estimaterne er maksimaliseringsestimater, afviger de en smule fra (de centrale) momentestimater, som vi fandt i det foregående eksempel.

Den asymptotiske varians-kovariansmatrix for estimaterne udskrives som:

\[
\begin{array}{cccc}
\text{Cov Parm} & \text{Row} & \text{COVP1} & \text{COVP2} & \text{COVP3} & \text{COVP4} \\
\hline
\text{UN(1,1)} & 1 & 0.12182040 & 0.06188189 & 0.03175724 & -0.00037325 \\
\text{UN(2,1)} & 2 & 0.06188189 & 0.06917141 & 0.05433626 & 0.00000000 \\
\text{UN(2,2)} & 3 & 0.03175724 & 0.05433626 & 0.09480666 & -0.00119439 \\
\text{Residual} & 4 & -0.00037325 & 0.00000000 & -0.00119439 & 0.00149299 \\
\end{array}
\]

Elementerne i den udskrevne matrix identificeres som elementerne i matri
cen

\[
D = \begin{bmatrix}
\hat{\sigma}^2 \\ 
\hat{\gamma}_{11} \\ 
\hat{\sigma}^2 \\ 
\hat{\gamma}_{12} \\ 
\hat{\sigma}^2 \\ 
\hat{\gamma}_{22} \\
\end{bmatrix}
\]

Vi ser, at ingen af komponenterne i \( \hat{\sigma}^2 \Gamma \) synes at afvige signifikant fra nul (ingen af dem ligger mindre end 2 spredninger væk fra 0; således er estimatet for \( \hat{\sigma}^2 \gamma_{1,1} = 0.53005 \) med den estimerede spredning \( \sqrt{0.12182040} = 0.349 \)).

Endelig udskrives et resume af tilpasningen

\[
\begin{array}{ll}
\text{Model Fitting Information for RAMUS} \\
\hline
\text{Description} & \text{Value} \\
\hline
\text{Observations} & 20.0000 \\
\text{Log Likelihood} & -15.9877 \\
\text{Akaike's Information Criterion} & -19.9877 \\
\text{Schwarz's Bayesian Criterion} & -21.9792 \\
-2 \text{Log Likelihood} & 31.9755 \\
\end{array}
\]
Null Model LRT Chi-Square       18.9274
Null Model LRT DF              3.0000
Null Model LRT P-Value         0.0003

samt de ønskede estimator for de systematiske effekter

MIXED ML

Solution for Fixed Effects

| Effect     | Estimate   | Std Error   | DF   | t       | Pr > |t|   |
|------------|------------|-------------|------|---------|------|-----|
| INTERCEPT  | 52.69000000| 0.33215960  | 4    | 158.63  | 0.0001|
| ALD        | 1.42400000 | 0.31125038  | 4    | 4.58    | 0.0102|

Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>NDF</th>
<th>DDF</th>
<th>Type III F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALD</td>
<td>1</td>
<td>4</td>
<td>20.93</td>
<td>0.0102</td>
</tr>
</tbody>
</table>

hvor Std Error angiver den estimerede spredning for den pågældende koeficient, og DF angiver de tilsvarende frihedsgrader (antallet af observationer N minus rangen af produktet af modelmatricen svarende til de systematiske effekter og modelmatricen svarende til de tilfældige effekter). I udskriften af testet for de systematiske effekter angiver NDF og DDF henholdsvis frihedsgraderne for tælleren (eng. numerator) og nævneren (eng. denominator).

Havde vi i stedet benyttet REML-estimationen ved optionen

PROC MIXED METHOD=REML ASYCOV ;

dervedi vi få estimatorne

Covariance Parameter Estimates (REML)

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>DRENG</td>
<td>0.66796250</td>
</tr>
</tbody>
</table>
7.1 Ubalancerede regressionsmodeller

\[ \begin{array}{lcc}
\text{UN(2,1)} & \text{DRENG} & 0.35055000 \\
\text{UN(2,2)} & \text{DRENG} & 0.53636000 \\
\text{Residual} & & 0.08640000 \\
\end{array} \]

der i det balancerede tilfælde netop er de samme som momentestimaterne. Estimaterne for de systematiske effekter er de samme som ved maksimaliseringsmetoden, men skønnet over usikkerheden er ændret en smule:

\textit{MIXED REML}

\textbf{Solution for Fixed Effects}

| Effect   | Estimate  | Std Error | DF | t     | Pr > |t| |
|----------|-----------|-----------|----|-------|-------|---|
| INTERCEPT| 52.69000000 | 0.37136572 | 4  | 141.88 | 0.0001 |
| ALD      | 1.42400000 | 0.34798851 | 4  | 4.09   | 0.0149 |

\textbf{Tests of Fixed Effects}

<table>
<thead>
<tr>
<th>Source</th>
<th>NDF</th>
<th>DDF</th>
<th>Type III F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
</table>
| ALD    | 4   | 16.75 | 0.0149

7.2 Ubalancerede regressionsmodeller med varierende koefficienter

7.2.1 Den systematiske model

Såfremt eksperimentet ikke er balanceret, d.v.s såfremt der er benyttet forskellige modelmatricer i de forskellige grupper, forskelligt antal observationer etc. bliver løsningen i det væsentlige analog med det foregående afsnit, men formulapparatet bliver noget tungere, da regressionerne i de enkelte grupper nu kræver separate vægtninger.
Sætning 7.2.1 Maksimaliseringsestimater for den ubalancerede regressionsmodel

Lad $Y_i$ angive en $n_i$-dimensional vektor af observationer, og $X_i$ en kendt $n_i \times p$ dimensional matrix således at $(X_i^T X_i)^{-1}$ eksisterer. Antag, at

$$Y_i = X_i \beta_i + \epsilon_i, \ i = 1, 2, \ldots, k$$

hvor

$$\epsilon_i \in N_{n_i}(0, \sigma^2 I_{n_i})$$

og $\epsilon_i$ og $\epsilon_j$ er indbyrdes uafhængige for $i \neq j$.

Da er maksimaliseringsestimatoren for $\beta_i$ givet ved

$$\hat{\beta}_i = P_i y_i, \quad (7.2.1)$$

hvor parameterprojektionsmatricen for den $i$'te gruppe er er givet ved

$$P_i = (X_i^T X_i)^{-1} X_i^T \quad (7.2.2)$$

og den centrale estimator for $\sigma^2$ er

$$\hat{\sigma}^2 = \frac{sak_{1.}}{N - kp} \quad (7.2.3)$$

hvor

$$sak_{1.} = \sum_{i=1}^{k} sak_{1,i} \quad (7.2.4)$$

med

$$sak_{1,i} = (y_i - X_i \hat{\beta}_i)^T (y_i - X_i \hat{\beta}_i)$$

og $N = \sum_i n_i$

Der gælder

$$\hat{\beta}_i \in N_p(0, \sigma^2 (X_i^T X_i)^{-1}); \ SAK_{1.} \in \sigma^2 \chi^2(N - kp)$$

og $\hat{\beta}_i$ og $SAK_{1.}$ er indbyrdes uafhængige.

Bevis:
7.2 Ubalancerede regressionsmodeller

Beviset følger i analogi med beviset for sætning 7.1.1 ved at bemærke, at maksimaliseringsestimatoren fås ved at minimerere kvadratafvigelsessummen

\[ l(\beta; y_1, \ldots, y_k) = -\sum_{i=1}^{k}(y_i - X_i\beta_i)^T(y_i - X_i\beta_i) \]

Uafhængigheden følger af spaltningssætningen.

\[ \square \]

Sætning 7.2.2 Test for fælles regressionskoefficienter

Kvotienttestet for hypotesen

\[ H_1: \quad \beta_1 = \beta_2 = \cdots = \beta_k (= \beta_0), \quad (7.2.5) \]

imod alternativet

\[ H_1^c: \beta_i \neq \beta_j \text{ for mindst ett sæt } (i, j) \text{ med } i \neq j \]

har teststørrelsen

\[ Z = \frac{SAK_2/[p(k - 1)]}{SAK_1/(N - k)} \quad (7.2.6) \]

hvor

\[ SAK_2 = \sum_{i=1}^{k}(\hat{\beta}_i - \hat{\beta}_0)^TX_i^TX_i(\hat{\beta}_i - \hat{\beta}_0) \quad (7.2.7) \]

med

\[ \hat{\beta}_0 = \frac{1}{k} \sum_{i=1}^{k}P_iy_i \quad (7.2.8) \]

Under \( H_1 \) følger \( Z \) en \( F(p(k - 1), N - kp) \)-fordeling. Testet forkaster for store værdier af \( z \).

Bevis:
Sætningen vises analogt med sætning 7.1.2 ved at bemærke, at der gælder opspaltningen

\[ \sum_{i=1}^{k} (y_i - X_i\hat{\beta}_0)^T(y_i - X_i\hat{\beta}_0) = sak_1 + sak_2 \]

og benytte spaltningssætningen.

\[ \Box \]

Bemærkning 1 *Estimation og test ved observationer med vilkårlig, kendt dispersionsmatrix*

Såfremt

\[ \epsilon_i \in N_{n_i}(0, \sigma^2 V_i) \]

fås maksimaliseringsestimatoren ved at minimere kvadratafvigelsessummen

\[ l(\beta; y_1, \ldots, y_k) = -\sum_{i=1}^{k} (y_i - X_i\beta_i)^T V_i^{-1} (y_i - X_i\beta_i) \]

Man finder estimatet

\[ \hat{\beta}_i = P_{wi} y_i \] (7.2.9)

med parameterprojektionsmatricen

\[ P_{wi} = (X_i^T V_i^{-1} X_i)^{-1} X_i^T V_i^{-1} . \] (7.2.10)

Den centrale estimator for \( \sigma^2 \) er givet ved (7.2.3) med

\[ sak_{1,i} = (y_i - X_i\hat{\beta}_i)^T V_i^{-1} (y_i - X_i\hat{\beta}_i) \]

Der gælder

\[ \hat{\beta}_i \in N_p(0, \sigma^2 (X_i^T V_i^{-1} X_i)^{-1}); \ SAK_{1,i} \in \sigma^2 \chi^2(N - kp) \]

Kvotienttestet for hypotesen (7.2.5) har teststørrelsen (7.2.6), hvor

\[ SAK_2 = \sum_{i=1}^{k} (\hat{\beta}_i - \hat{\beta}_0)^T X_i^T V_i^{-1} X_i(\hat{\beta}_i - \hat{\beta}_0) \] (7.2.11)
med

\[ \hat{\beta}_0 = \frac{1}{k} (X_i^T V_i^{-1} X_i)^{-1} X_i^T \sum_{i=1}^{k} y_i \]  

(7.2.12)

svarende til opspaltningen

\[ \sum_{i=1}^{k} (y_i - X_i \hat{\beta}_0)^T V_i^{-1} (y_i - X_i \hat{\beta}_0) = s_{ak_1} + s_{ak_2} \]

■

7.2.2 Den tilfældige model

I lighed med det foregående afsnit kan man vælge at udbygge modellen med en tilfældig komponent ved antagelsen \( \beta_i \in N_p(\beta_0, \sigma^2 \Gamma) \). For denne model gælder

Sætning 7.2.3 Marginal fordeling af observationer og estimater under regressionsmodel med tilfældige koefficienter

Lad

\[ Y_i = X_i \beta_i + \epsilon_i, \quad i = 1, 2, \ldots, k \]

hvor

\[ \beta_i \in N_p(\beta_0, \sigma^2 \Gamma) \]

og hvor \( \beta_i, \beta_j \) er indbyrdes uafhængige for \( i \neq j \), og \( \epsilon_i \) og \( \epsilon_j \) er indbyrdes uafhængige for \( i \neq j \) og endvidere \( \beta_i \) og \( \epsilon_j \) er uafhængige.

Da er den marginale fordeling af \( Y_i \) en

\[ N_{n_i}(X_i \beta_0, \sigma^2, [I_{n_i} + X_i \Gamma X_i^T]) \] i-fordeling,

og den marginale fordeling af estimatet \( \hat{\beta}_i \) (7.2.1) for regressionskoefficienterne svarende til den \( i \)te gruppe er
\[ \tilde{\beta}_i \in N_p(\beta_0, \sigma^2[(X_i^T X_i)^{-1} + \Lambda]) \] (7.2.13)

**Bevis:**
Beviset forløber i analogi med beviset for sætning 7.1.3.

\[ \square \]

**Sætning 7.2.4 Maksimaliseringestimation for ubalanceret regressionsmodel med tilfældige koefficienter**

Lad modellen være som i sætning 7.2.3.

Da fås maksimaliseringestimatorerne for parametrene \( \beta_0, \sigma^2 \) og \( \Gamma \) ved at maksimere

\[ l(\beta_0, \sigma^2, \Gamma; y_1, \ldots, y_k) = -k \times n \ln(\sigma^2) - \sum_{i=1}^{k} \ln[\det(D_i(\Gamma))] \]

\[ - \frac{1}{\sigma^2} \sum_{i=1}^{k} (y_i - X_i \beta_0)^T (D_i(\Gamma))^{-1} (y_i - X_i \beta_0) \] (7.2.14)

hvor

\[ D_i(\Gamma) = X_i \Gamma X_i^T + I_{n_i} \] (7.2.15)

med hensyn til \( \sigma^2 > 0, \beta_0 \in \mathbb{R}^p \) og \( \Gamma \) positiv semidefinit.

**Bevis:**
Sætningen vises ved at bemærke, at den marginale fordeling af \( Y_i \) er en \( N_{n_i}(X_i \beta_0, \sigma^2, [I_{n_i} + X_i \Gamma X_i^T]) \)-fordeling, og at (7.2.14) netop er logaritmen til likelihoodfunktionen.

\[ \square \]

**Bemærkning 1** Likelihoodligningerne ved maksimum i et indre punkt
7.2 Ubalancerede regressionsmodeller

Såfremt maksimumværdien findes i et indre punkt, tilfresstiller estima-
terne udtrykkene

$$\hat{\beta}_0 = \left[ \sum_{i=1}^{k} X_i^T (D_i(\Gamma))^{-1} X_i \right]^{-1} \sum_{i=1}^{k} X_i^T (D_i(\Gamma))^{-1} y_i \quad (7.2.16)$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{k} (y_i - X_i \hat{\beta}_0)^T (D_i(\hat{\Gamma}))^{-1} (y_i - X_i \hat{\beta}_0) \quad (7.2.17)$$

hvor $\hat{\Gamma}$ er bestemt som den positive semidefinite $p \times p$-matrix, der maksimører udtrykket

$$l(\Gamma) = -N \times \ln(sak^*(\Gamma)) - \sum_{i=1}^{k} \ln[\det(D_i(\Gamma))] \quad (7.2.18)$$

med

$$sak^*(\Gamma) = \sum_{i=1}^{k} y_i^T (D_i(\Gamma))^{-1} y_i$$

$$- \left[ \sum_{i=1}^{k} X_i^T (D_i(\Gamma))^{-1} y_i \right]^T \left[ \sum_{i=1}^{k} X_i^T (D_i(\Gamma))^{-1} X_i \right]^{-1} \left[ \sum_{i=1}^{k} X_i^T (D_i(\Gamma))^{-1} y_i \right]$$

 bespoke

Bemærkning 2 Estimation ved observationer med vilkårlig, kendt
dispersionsmatrix

Sætningen og bemærkningen gælder også i den generelle situation, hvor
dispersionsmatricen for observationsfejlene er $\sigma^2 V_i$. Man skal blot erstatte
detrykket (7.2.15) for $D_i(\Gamma)$ med

$$D_i(\Gamma) = X_i \Gamma X_i^T + V_i, \quad (7.2.19)$$

idet den marginale fordeling af $Y_i$ i dette tilfælde er en

$$N_{n_i}(X_i \beta_0, \sigma^2; [V_i + X_i \Gamma X_i^T]) - fordeling$$
Sætning 7.2.5 Momentestimation for ubalanceret regressionsmodel med tilfældige koefficienter

Under antagelserne fra sætning 7.2.3 fås momentestimaterne for \( \beta_0, \sigma^2 \) og \( \Gamma \) som

\[
\tilde{\beta}_0 = \frac{1}{k} \sum_{i=1}^{k} \tilde{\beta}_i \quad (7.2.20)
\]

\[
\tilde{\sigma}^2 = \frac{1}{N - kp} \text{sak}_1. \quad (7.2.21)
\]

\[
\tilde{\Gamma} = \frac{\text{sak}_\beta}{(k - 1)\tilde{\sigma}^2} - \frac{1}{k} \left( \sum_{i=1}^{k} X_i^T X_i \right)^{-1} \quad (7.2.22)
\]

hvor \( \tilde{\beta}_i \) og \( \text{sak}_1 \) er givet ved (7.2.1) og (7.2.4), og hvor

\[
\text{sak}_\beta = \sum_{i=1}^{k} (\tilde{\beta}_i - \tilde{\beta}_0)(\tilde{\beta}_i - \tilde{\beta}_0)^T
\]

**Bevis:**

Sætningen vises ved at bemærke, at \( \tilde{\beta}_0 \) er gennemsnittet, og \( \text{sak}_\beta/(k - 1) \) er den empiriske dispersionsmatrix for \( \tilde{\beta}_i \), hvor \( \tilde{\beta}_i \) er indbyrdes uafhængige med

\[
E[\tilde{\beta}] = \beta_0
\]

og

\[
D[\tilde{\beta}_i] = \sigma^2[(X^T X)^{-1} + \Gamma]
\]

ifølge sætning 7.2.3.

Vi har derfor

\[
E[\tilde{\beta}_0] = \beta_0, \quad \text{og}
\]

\[
E[\text{sak}_\beta] = \sum_{i=1}^{k} \left( 1 - \frac{1}{k} \right) D[\tilde{\beta}_i] = \sigma^2 \left( 1 - \frac{1}{k} \right) \sum_{i=1}^{k} [(X_i^T X_i)^{-1} + \Gamma]
\]
Bemærkning 1 Sammenligning mellem maksimaliseringsestimatet og momentestimater

Sammenligner vi momentestimatet (7.2.20) for \( \beta_0 \) med maksimaliseringsestimatenerne (7.2.16), ser vi, at i modsætning til momentestimatet inddrager maksimaliseringsestimatet såvel kendskabet til forskellene mellem modellmatricerne for de enkelte grupper samt kendskabet til dispensationsmatricen \( \Gamma \) ved den indbyrdes vægtning af de enkelte komponenter af \( \hat{\beta}_i \).

Man kunne forbedre skønnnet over \( \beta_0 \) ved at vægte \( \hat{\beta}_i \) med designpræcisionen ved benyttelse af

\[
\hat{\beta}_0 = \left( \sum_{i=1}^{k} X_i^T X_i \right)^{-1} \sum_{i=1}^{k} X_i^T y_i = \left( \sum_{i=1}^{k} X_i^T X_i \right)^{-1} \sum_{i=1}^{k} X_i^T X_i \hat{\beta}_i ,
\]

men anvendelsen af dette estimat vanskeliggør beskrivelsen af forholdene for \( \text{sak}_\beta \)-matricen.


Bemærkning 2 Momentestimation ved observationer med vilkårlig, kendet dispensationsmatrix

I det generelle tilfælde, hvor dispensationsmatricen for observationsfejlene er \( \sigma^2 V_i \), kan man ligeledes benytte estimaterne (7.2.20) og (7.2.21) for \( \beta_0 \) og \( \sigma^2 \) med \( \hat{\beta}_i \) bestemt ved (7.2.9).

Momentestimatet for \( \Gamma \) bestemmes i det generelle tilfælde ved

\[
\hat{\Gamma} = \frac{\text{sak}_\beta}{(k-1)\tilde{\sigma}^2} - \frac{1}{k} \left( \sum_{i=1}^{k} X_i^T V_i^{-1} X_i \right)^{-1} \tag{7.2.23}
\]

I dette tilfælde gælder, at \( \hat{\beta}_i \) er indbyrdes uafhængige med

\[
E [\hat{\beta}] = \beta_0
\]

og

\[
D [\hat{\beta}_i] = \sigma^2 [(X^T V^{-1} X)^{-1} + \Gamma] \]

Et forbedret skøn over $\beta_0$ fås i dette tilfælde som

$$\tilde{\beta}_0 = \left( \sum_{i=1}^{k} x_i^T v_i^{-1} x_i \right)^{-1} \sum_{i=1}^{k} x_i^T v_i^{-1} y_i$$

$$= \left( \sum_{i=1}^{k} x_i^T v_i^{-1} x_i \right)^{-1} \sum_{i=1}^{k} x_i^T v_i^{-1} x_i \tilde{\beta}_i ,$$

men skønnet vanskelig gør beskrivelsen af forholdene for sak-matricen.

\[ \square \]

### 7.3 Tidsrækkekmodeller

#### 7.3.1 Den endimensionale autoregressive proces af første orden.

Ved en tidsrække med diskret tid forstår vi en række af observationer $y_1, y_2, \ldots, y_n$, hvor indexmængden $1, 2, \ldots, n$ angiver ækvidistante punkter på tidsaksen.

Ofte vil sådanne observationer være autokorrelerede, dvs. $\text{COV}[Y_t, Y_{t-k}] \neq 0$ for nogle $k \geq 0$.

Vi vil her betrægde den simple autoregressive model af første orden. Modellen er givet ved at fordelingen af $Y_t$ kun afhænger af værdien af $Y_{t-1}$:

$$Y_t | \{ Y_{t-1} = y_{t-1} \} \in \mathcal{N}(\beta y_{t-1}, \sigma^2) \quad (7.3.1)$$

Ofte udtrykkes modellen rekursivt

$$Y_t = \beta Y_{t-1} + \epsilon_t \quad (7.3.2)$$

hvor $\epsilon_1, \epsilon_2, \ldots, \epsilon_t$ er uafhængige $\mathcal{N}(0, \sigma^2)$-fordelte størrelser.

Modellen kaldes autoregressiv, fordi fordelingen af $Y_t$ netop kan beskrives ved en lineær regression på den foregående værdi.

**Sætning 7.3.1** Maksimaliseringsestimation for $AR(1)$-model
Maksimaliseringsestimatorerne for parametrene $\beta$ og $\sigma$ svarende til modellen (7.3.2) for observationerne $y_0, y_1, \ldots, y_n$ er bestemt ved

$$
\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{\beta}y_{j-1})^2
$$

hvor $\hat{\beta}$ er løsning til

$$
-\frac{\hat{\beta}}{1 - \hat{\beta}^2} \hat{\sigma}^2 + \sum_{j=0}^{n-1} y_j y_{j+1} - \hat{\beta} \sum_{j=0}^{n-1} y_j^2 = 0 \quad (7.3.3)
$$

**Bevis:**
Overspringes

Indfører vi observationsvektoren $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ og $\mathbf{y}_{-1} = (y_0, y_1, \ldots, y_{n-1})^T$, ser vi, at vi kan udtrykke estimaterne som

$$
\hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - \hat{\beta}\mathbf{y}_{-1})^T (\mathbf{y} - \hat{\beta}\mathbf{y}_{-1})
$$

hvor $\hat{\beta}$ er løsning til

$$
-\frac{\hat{\beta}}{1 - \hat{\beta}^2} \hat{\sigma}^2 + \mathbf{y}_{-1}^T \mathbf{y}_{-1} - \hat{\beta} \mathbf{y}_{-1}^T \mathbf{y} = 0
$$

Ligningerne må løses rekursivt.

I stedet benyttes derfor ofte mindste kvadraters estimator for $\beta$,

$$
\hat{\beta} = (\mathbf{y}_{-1}^T \mathbf{y}_{-1})^{-1} \mathbf{y}_{-1}^T \mathbf{y} = \sum_{j=0}^{n-1} y_j y_{j+1} / \sum_{j=0}^{n-1} y_j^2
$$

og den tilsvarende minimale værdi af den gennemsnitlige residualkvadrat-afvigelse

$$
\hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - \hat{\beta}\mathbf{y}_{-1})^T (\mathbf{y} - \hat{\beta}\mathbf{y}_{-1})
$$

som skøn over $\sigma^2$.

Vi ser, at $\hat{\beta}$ fremkommer ved at se bort fra størrelsen $\hat{\beta}\hat{\sigma}^2/(1 - \hat{\beta}^2)$ i udtrykket for maksimaliseringsestimatorerne.
Den tilfældige model

Har vi nu $k$ tidsrækker fra $k$ forskellige grupper $y^i = (y^i_1, y^i_2, \ldots, y^i_n)^T$; $i = 1, 2, \ldots, k$, kan vi modellere variationen imellem disse tidsrækker ved en tilfældig model:

Vi antager

$$Y^i_t = \beta^i_t Y^i_{t-1} + e^i_t, \quad t = 1, 2, \ldots, n; \quad i = 1, 2, \ldots, k,$$

hvor $\beta^i_t \in N(\beta^i_0, \sigma^2_0)$ er indbyrdes uafhængige, og $e^i_t \in N(0, \sigma^2)$ ligeledes er indbyrdes uafhængige og uafhængige af $\beta^i_t$.

Sætning 7.3.2 *Momentestimation for AR(1)-model, tilfældig model*

Under den tilfældige model har man momentestimaterne:

$$\tilde{\beta}_0 = \frac{1}{k} \sum_{i=1}^{k} \tilde{\beta}_i \quad \text{og} \quad \tilde{\sigma}_0^2 = \frac{1}{k} \sum_{i=1}^{k} (\tilde{\beta}_i - \tilde{\beta}_0)^2$$

med

$$\tilde{\beta}_i = [(y^i_{-1})^{T} y_{-1}]^{-1} (y^i_{-1})^{T} y^i = \sum_{j=0}^{n-1} y^i_j y^i_{j+1} / \sum_{j=0}^{n-1} (y^i_j)^2$$

og det tilsvarende estimat for residualvariansen

$$\tilde{\sigma}^2 = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} (y^i - y^i_{-1} \tilde{\beta}_i)^T (y^i - y^i_{-1} \tilde{\beta}_i)$$

Bevis:
Overspringes

7.3.2 *Flerdimensionale tidsrækkekmodeller.*

Vi vil ganske kort angive resultaterne for flerdimensionale tidsrækkekmodeller.
Lad \( y_t \) angive en \( p \)-dimensional søjlevektor af observationer, og lad \( z_t \) angive en (kendt) \( q \)-dimensional vektor af kontrolvariable (regressionsvariable) til tidspunktet \( t, t = 0, 1, 2, \ldots, n \).

Antag at der gælder følgende model

\[
y_t = Ay_{t-1} + Cz_t + \epsilon_t, \quad t = 1, 2, \ldots, n \tag{7.3.4}
\]

hvor \( \epsilon_t \) angiver en \( p \)-dimensional stokastisk vektor med forventningsværdi nul, og \( A \) og \( C \) er ukendte paramettermatricer af dimensioner henholdsvis \( p \times p \) og \( p \times q \).

Vi kan opskrive (7.3.4) på matrixform

\[
Y^T = AY_{t-1}^T + CZ^T + \epsilon^T \tag{7.3.5}
\]

hvor de \( n \times p \)-dimensionale matricer \( Y, Y_{t-1} \) og \( \epsilon \) er givet ved

\[
Y = (y_1, y_2, \ldots, y_n)^T; \quad Y_{t-1} = (y_0, y_1, \ldots, y_{n-1})^T
\]

og \( \epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T \), og den \( n \times q \)-dimensionale matrix \( Z \) tilsvarende er givet ved \( Z = (z_1, z_2, \ldots, z_n)^T \).

Transponerer vi nu (7.3.5) fås

\[
Y = MB + \epsilon \tag{7.3.6}
\]

hvor den \( n \times (p + q) \)-dimensionale matrix \( M \) er givet ved

\[
M = (Y_{t-1}, Z)
\]

og den \( (p + q) \times p \)-dimensionale koefficientmatrix \( B = (A, C)^T \).

Vi ordner nu observationerne \( Y \) i en søjlevektor \( y \)

\[
y = \text{vec}(Y) = \begin{pmatrix} y_{11} \\ \vdots \\ y_{n1} \\ \vdots \\ y_{np} \end{pmatrix}
\]

og tilsvarende sættes \( \beta = \text{vec}(A) \) og \( \epsilon = \text{vec}(\epsilon) \). Endelig sætter vi \( X = I_p \otimes M \), hvor \( I_p \) angiver den \( p \times p \)-dimensionale enhedsmatrix.

Antages nu, at \( \epsilon \in N_{np \times np}(0, \Sigma) \), har vi den generaliserede mindste kvadraters estimator for \( \beta \)

\[
\hat{\beta} = (X^T \Sigma X)^{-1} X^T \Sigma^{-1} y \tag{7.3.7}
\]
Den tilfældige model

Såfremt vi har $k$ tidsrækker fra $k$ forskellige grupper (dvs med forskellige værdier af $\beta$), kan vi modellere variationen imellem disse tidsrækker ved en tilfældig model.

Vi betragter derfor modellen

$$Y^i = M_i B_i + \epsilon^i$$  \hspace{1cm} (7.3.8)

hvor den $n_i \times (p + q)$-dimensionale matrix $M_i$ er givet ved

$$M_i = (Y_{i-1}^i, Z^i)$$

og den $(p + q) \times p$-dimensionale koeﬃcientmatrix $B_i = (A_i, C_i)^T$.

Vi sætter som før

$$y^i = \text{vec}(Y^i); \beta_i = \text{vec}(B_i) \quad \text{og} \quad \epsilon^i = \text{vec}(\epsilon^i)$$

Endelig sætter vi $X_i = I_p \otimes M_i$

Vi antager da,

$$\epsilon^i \in N_{n_ip \times n_ip}(0, \Sigma_i) \quad \text{og} \quad \beta_i \in N_{(p+q) \times q}(\beta_0, \Sigma_0)$$  \hspace{1cm} (7.3.9)

Vi har da estimaterne under denne model:

a) $\Sigma_i$ kendt:

$$\hat{\beta}_0 = \frac{1}{k} \sum_{i=1}^k \hat{\beta}_i \quad \text{med} \quad \hat{\beta}_i = \hat{\beta}_i = (X_i^T \Sigma_i X_i)^{-1} X_i^T \Sigma_i^{-1} y^i$$ \hspace{1cm} (7.3.10)

og

$$\hat{\Sigma}_0 = \frac{1}{k} \sum_{i=1}^k (\beta_i - \hat{\beta}_0)(\beta_i - \hat{\beta}_0)^T$$ \hspace{1cm} (7.3.11)

b) $\Sigma_i = \sigma^2 V_i$, hvor $V_i$ er en kendt symmetrisk, positiv deﬁnit $n_ip \times n_ip$-dimensional matrix:

$$\hat{\beta}_0 = \frac{1}{k} \sum_{i=1}^k \hat{\beta}_i \quad \text{med} \quad \hat{\beta}_i = \hat{\beta}_i = (X_i^T V_i X_i)^{-1} X_i^T V_i^{-1} y^i$$ \hspace{1cm} (7.3.12)
7.4 Blandede modeller

og

\[ \hat{\sigma}^2 = \frac{1}{pk} \sum_{i=1}^{k} \frac{1}{n_i} (y^i - X_i \hat{\beta}_i)^T V_i^{-1} (y^i - X_i \hat{\beta}_i) \]  

(7.3.13)

goget \( \hat{\Sigma}_0 \) bestemmes ved (7.3.11) med \( \hat{\beta}_0 \) og \( \hat{\beta}_i \) givet ved (7.3.12).

c) \( \Sigma_i = \Sigma \otimes V_i \), hvor \( \Sigma \) er ukendt symmetrisk, positiv definit \( p \times p \)-matrix, og \( V_i \) er en kendt symmetrisk, positiv definit \( n_i \times n_i \)-dimensional matrix:

\[ \hat{B}_0 = \frac{1}{k} \sum_{i=1}^{k} \hat{B}_i \quad \text{med} \quad \hat{B}_i = (M_i^T V_i^{-1} M_i)^{-1} M_i^T V_i^{-1} Y^i \]  

(7.3.14)

\[ \hat{\beta}_i = \text{vec}(\hat{B}_i). \]

\[ \hat{\Sigma} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} (Y^i - M_i \hat{B}_i)^T V_i^{-1} (Y^i - M_i \hat{B}_i) \]  

(7.3.15)

og \( \hat{\Sigma}_0 \) bestemt ved (7.3.11) med \( \hat{\beta}_0 \) og \( \hat{\beta}_i \) givet ved (7.3.14).

7.4 Blandede modeller

De modeller, vi har betragtet i dette afsnit, er eksempler på såkaldte blandede modeller (eng: mixed models) for normalfordelte data.

Termen "blandet" refererer her til, at strukturen omfatter både en systematisk og en tilfældig komponent.

Idet vi som sædvanligt opstiller samtlige observationer i en vektor \( y \), modelleres den systematiske komponent ved

\[ y = X\beta + \epsilon \]

hvor \( X \) angiver modelmatricen for den systematiske effekt, \( \beta \) er vekten af effekter, og \( \epsilon \) angiver en vektor af tilfældige modelfagigler.
Bidraget fra den tilfældige komponent modelleres ved en modelmatrix $Z$, og en vektor $\mathbf{v}$ af tilfældige effekter. Den blandede model udtrykkes da som

$$y = X\beta + Z\mathbf{v} + \epsilon$$

Det antages, at $\mathbf{v}$ og $\epsilon$ er uafhængige og har forventningsværdien 0. Dispersionsmatricerne for $\mathbf{v}$ og $\epsilon$ betegnes hhv. $G$ og $R$, dvs

$$E[\mathbf{v}] = 0, \quad D[\mathbf{v}] = G$$

og

$$E[\epsilon] = 0, \quad D[\epsilon] = R.$$

Under disse antagelser har man

$$E[y] = X\beta, \quad D[y] = V = ZGZ^T + R. \quad (7.4.1)$$

Denne generelle formulering giver mulighed for at vælge mellem at placere tilfældige effekter i $\mathbf{v}$ eller i $\epsilon$.

Formuleringen giver endvidere mulighed for at påtrykke en bestemt struktur af dispersionsmatricerne $G$ og $R$, som f.eks. $R = \sigma^2 I$

Sædvanligvis vil man placere effekter svarende til en lille dimensionalitet (få søjler i $Z$) i $\mathbf{v}$, sådan at $G$ let kan inverteres, og placere resten i $\epsilon$.

Den generelle teori for blandede modeller er behandlet af Searle et al. (1992).

### 7.5 Referencer


Afsnit 8

Aposteriorifordelinger

I dette afsnit indfører vi begreberne apriorifordeling, aposteriorifordeling samt prædiktiv fordeling.

Vi vil tage udgangspunkt i dekomponeringen af variationen i en model med hierarkisk variation, sådan som disse modeller blev behandlet i afsnit 5, 6 og 7.

I dette afsnit vil vi vise, hvordan parametrene opdateres ved brug af Bayes’ sætning efter observation af et stikprøveresultat. I afsnit 8.3 giver vi en generel beskrivelse af opdateringen for eksponentielle dispersionsmodeller, og viser resultaterne for de fordelinger, der blev betragtet i afsnit 6.2 til 6.5.

Resultaterne er sammenfattet i tabel 8.1 på side 709, der viser sammenhængen mellem stikprøvefordeling og aposteriorifordeling udtrykt ved stikprøveresultat og parametrene α og β i apriorifordelingen af middelværdiparametere n. Tabel 8.2 på side 710 giver tilsvarende en oversigt over opdateringen af momentparametrene og

I Eksempel 8.3.2 er principperne illustreret.

Endelig beskriver vi kort opdateringen svarende til de lineære normalfordelingsmodeller, der blev introduceret i afsnit 7.
Vi gør opmærksom på at vi kun betragter modeller, hvor fordelingen af den tilfældige variation mellem grupper har en unimodal tæthed. Selv om Bayes’ sætning også finder udbredt anvendelse til behandling af klassifikationsproblemer, bliver sådanne problemer ikke behandlet i dette afsnit.

8.1 Betingede fordelinger, Bayes’ sætning

8.1.1 Bayes’ sætning

Vi minder om

Sætning 8.1.1 Bayes’ formel for hændelser

Lad $A_1, A_2, \ldots, A_n$ være en følge af disjunkte hændelser som tilsammen udgør hele udfaldsrummet $\Omega$, og lad $B$ være en hændelse.

Der gælder da

$$P[A_i|B] = \frac{P[A_i \cap B]}{\sum_j P[B|A_j]P[A_j]}$$ (8.1.1)

Bevis:
Se Jørsboe Sætning 2.1

For stokastiske variable $X$ og $Y$ har man tilsvarende

Sætning 8.1.2 Bayes’ formel for stokastiske variable

Lad $X$ og $Y$ være stokastiske variable, og lad $w(y)$ angive den marginale tæthed for $Y$.

Lad endvidere $g(x|y)$ angive den betingede tæthed af $X$ for givet $Y = y$.

Da er den betingede fordeling af $Y$ for givet $X = x$ en fordeling med tæthed

$$h(y|x) = \frac{g(x|y)w(y)}{k(x)}$$ (8.1.2)

hvor

$$k(x) = \int g(x|y)w(y)\nu\{dy\}$$
angiver den marginale tæthed for $X$.

**Bevis:**
Følger af resultater vedrørende produktmål fra mål- og integralteorien. 

**Bemærkning 1** *Udtale af Thomas Bayes’ navn*
Vi bemærker, at Bayes’ sætning er opkaldt efter den engelske præst, Thomas Bayes. Hans efternavn udtales [Bæ'-js].

## 8.2 Apriori- og aposteriorifordelinger

**Definition 8.2.1** *Apriori- aposteriorifordeling*

Betragt en statistisk model for observationssættet $X_1, X_2, \ldots, X_n$ givet ved familien af tætheder

$$\{f(x_1, x_2, \ldots, x_n | \theta)\}_{\theta \in \Theta} \quad (8.2.1)$$

hvor $\Theta \subset \mathbb{R}^k$.

Antag at fordelingen af $X_1, X_2, \ldots, X_n$ for en fastholdt værdi af $\theta$ har tætheden

$$f(x_1, x_2, \ldots, x_n | \theta)$$

og antag yderligere, at parameteren $\theta$ kan opfattes som en stokastisk variabel med tætheden

$$w(\theta)$$

Som tidligere nævnt kaldes fordelingen af $\theta$ med tætheden $w(\cdot)$ for strukturfordelingen, eller apriorifordelingen af $\theta$.

Den betingede fordeling af $\theta$ givet $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ kaldes aposteriorifordelingen af $\theta$ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. 

I de oftest forekommende situationer vil det være sådan, at at $X_1, X_2, \ldots, X_n$ er betinget uafhængige af $\theta$ i den simultane fordeling af $X_1, X_2, \ldots, X_n$ og $\theta$. Den simultane fordeling af $X_1, X_2, \ldots, X_n$ for fastholdt $\theta$ har da tætheden

$$f(x_1, x_2, \ldots, x_n | \theta) = \prod_{i=1}^{n} g(x_i | \theta) \quad (8.2.2)$$
hvor \( g(\cdot | \theta) \) angiver tætheden for en enkelt observation.

Sådanne tilfælde kan f.eks. være situationer, hvor parameteren \( \theta \) varierer i en population i overensstemmelse med apriorifordelingen \( w(\theta) \), og hvor en måling, \( X \) af \( \theta \) er behæftet med en målefejl svarende til fordelingen \( g(x|\theta) \). Modellen (8.2.2) svarer da til at der foretages \( n \) uafhængige målinger \( X_1, X_2, \ldots, X_n \) af en forelagt population.

Modellen (8.2.2) kan også beskrive en situation, hvor en proces genererer en serie af produktenheder hvor den \( i^{te} \) produktenhed karakteriseres ved \( X_i \). De enkelte produktenheder genereres uafhængigt af hinanden i overensstemmelse med en fordeling med tæthed \( g(x|\theta) \), hvor parameteren \( \theta \) varierer fra produktion til produktion i overensstemmelse med aprioritætheden \( w(\theta) \).

I begge disse situationer kan det være af interesse - ikke blot at beskrive aposteriorifordelingen af \( \theta \) efter observation af en stikprøve \( x_1, x_2, \ldots, x_n \), men også at beskrive fordelingen af fremtidige observationer fra den betragtede population.

**Definition 8.2.2 Prædiktiv fordeling**

Lad den simultane fordeling af \( \theta, X_1, X_2, \ldots, X_n \) og \( X'_1, X'_2, \ldots, X'_r \) være sådan, at for givet \( \theta \) er

\[
X_1, X_2, \ldots, X_n, X'_1, X'_2, \ldots, X'_r
\]

er indbyrdes uafhængige og identisk fordelte med en tæthed \( g(\cdot | \theta) \), dvs. de variable \( X'_1, X'_2, \ldots, X'_r \) er frembragt af det samme \( \theta \) som de variable \( X_1, X_2, \ldots, X_n \).

Antag at \( \theta \) er en stokastisk variabel med tætheden \( w(\theta) \).

Den prædiktive fordeling af \( X'_j \) for givet \( X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \), er da fordelingen med tæthed

\[
g_1(x'|x_1, x_2, \ldots, x_n) = \int_{\Theta} g(x'|\theta) w_1(\theta|x_1, x_2, \ldots, x_n) \nu\{d\theta\} \tag{8.2.3}
\]

hvor \( w_1(\theta|x_1, x_2, \ldots, x_n) \) angiver aposterioritætheden af \( \theta \) efter observation af \( X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \).
Lad $Z' = \sum_{j=1}^{r} X'_j$, og lad den betingede fordeling af $Z'$ for fastholdt $\theta$ have tætheden $g^{(r)}(z'|\theta)$. Den prædictive fordeling af $Z' = \sum_{j=1}^{r} X'_j$ for givet $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ er da fordelingen med tæthed

$$g^{(r)}_1(z'|x_1, x_2, \ldots, x_n) = \int \Theta g^{(r)}(z'|\theta) w_1(\theta|x_1, x_2, \ldots, x_n) \, d\nu\{\theta\}$$

(8.2.4)

hvor $w_1(\theta|x_1, x_2, \ldots, x_n)$ som før angiver aposterioritætheden af $\theta$ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

På tilsvarende måde defineres den prædictive fordeling af $Y' = \overline{X'} = Z'/r$ for givet $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ som fordelingen med tæthed

$$g^{[r]}_1(y'|x_1, x_2, \ldots, x_n) = \int \Theta g^{[r]}(y'|\theta) w_1(\theta|x_1, x_2, \ldots, x_n) \, d\nu\{d\theta\}$$

(8.2.5)

hvor $g^{[r]}(y'|\theta)$ angiver tætheden i den den betingede fordeling af $Y' = \sum_{j=1}^{r} X'_j/r$ for fastholdt $\theta$.

Den prædictive fordeling af $X'$ for givet $x_1, x_2, \ldots, x_n$ er således den marginale fordeling af $X'$ svarende til at man bruger aposteriorifordelingen af $\theta$ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, som strukturfordeling ("apriorifordeling") af $\theta$.

Tilsvarende er den prædictive fordeling af $Z'$ eller af $\overline{X'}$, den marginale fordeling af hhv $Z'$ eller $\overline{X'}$ svarende til at man bruger aposteriorifordelingen af $\theta$ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, som strukturfordeling ("apriorifordeling") af $\theta$.

Den prædictive fordeling for et sæt nye observatoner $X'_1, \ldots, X'_r$ vil sædvanligvis afhænge af det givne observationssæt $x_1, \ldots, x_n$.

I en planlægningsfase kan man imidlertid have interesse i - før indsamling af stikprøven $x_1, \ldots, x_n$ - at vurdere hvilken usikkerhed, der kan forventes i den prædictive fordeling.
Definition 8.2.3 Præposteriorimiddelværdi- og varians

Betragt situationen svarende til definition 8.2.1.

Betragt en funktion, $h(\theta)$ af $\theta$.

Aposteriorimiddelværdien $E[h(\theta)|x_1,\ldots,x_n]$ er da middelværdien af $h(\theta)$ i aposteriorifordelingen af $\theta$ efter observation af $x_1,\ldots,x_n$ og aposteriori-variansen

$$V[h(\theta)|x_1,\ldots,x_n]$$

er variansen af $h(\theta)$ i denne fordeling.

Aposteriorimiddelværdien $E[h(\theta)|x_1,\ldots,x_n]$ er en stokastisk variabel,

$$E[h(\theta)|x_1,\ldots,x_n] = t(X_1,\ldots,X_n)$$

med en fordeling, der kan beskrives ved fordelingen af $(X_1,\ldots,X_n)$.

Middelværdien (mht fordelingen af $(X_1,\ldots,X_n)$) af denne størrelse kaldes præposteriorimiddelværdien af $h(\theta)$ og betegnes

$$E_{X_1,\ldots,X_n}[E[h(\theta)|x_1,\ldots,x_n]].$$

Vi har altså

$$E_{X_1,\ldots,X_n}[E[h(\theta)|x_1,\ldots,x_n]] = E[t(X_1,\ldots,X_n)],$$

hvor $t(X_1,\ldots,X_n)$ er den størrelse, der fremkommer ved at opfatte

$$E[h(\theta)|x_1,\ldots,x_n]$$

som en stokastisk variabel.

Tilsvarende angiver præposteriorirvarianseren af $h(\theta)$ middelværdien (mht fordelingen af $(X_1,\ldots,X_n)$) af den stokastiske variable $u(X_1,\ldots,X_n)$, der fremkommer ved at opfatte $V[h(\theta)|x_1,\ldots,x_n]$ som en stokastisk variabel. Præposteriorirvarianseren af $h(\theta)$ betegnes med

$$E_{X_1,\ldots,X_n}[V[h(\theta)|x_1,\ldots,x_n]] = E[u(X_1,\ldots,X_n)].$$

Præposteriorimiddelværdien og -variansen kan altså opfattes som gennemsnitsværdien af aposteriorimiddelværdien $E[h(\theta)|x_1,\ldots,x_n]$ svarende til et stort antal forskellige stikprøver $x_1,\ldots,x_n$, og præposteriorirvarianseren angiver tilsvarende gennemsnitsværdien af aposteriorirvarianseren, $V[h(\theta)|x_1,\ldots,x_n]$, svarende til et stort antal forskellige stikprøver.
8.3 Aposteriorifordelinger for eksponentielle dispersionsmodeller

Vi vil i dette afsnit betragte de simple eksponentielle dispersionsmodeller med de naturlige konjugerede apriorifordelinger, der blev diskuteret i afsnit 6.2 til 6.5.

8.3.1 Resume af afsnit 6

I lighed med situationen i afsnit 6.2 til 6.5 vil vi operere med to parallele parametratiseringer af apriorifordelingerne.

Vi erindrer derfor om oversigtstabellerne, Tabel 6.1 og Tabel 6.2, der beskriver oversættelsen mellem momentparametrering og parametreringen ved $\alpha$ og $\beta$.

I lighed med afsnittene 6.2 til 6.5 vil vi også her betragte situationer, hvor $X_1, X_2, \ldots, X_n$ er betinget uafhængige af $\vartheta$ i den simultane fordeling af $X_1, X_2, \ldots, X_n$ og $\vartheta$, og hvor fordelingen af $X_i$ for givet $\vartheta$ har en tæthed af formen

$$g(x|\vartheta) = d(x) \exp\{[\vartheta x - \kappa(\vartheta)]/\sigma^2\},$$

og hvor fordelingen af $\vartheta$ er den konjugerede fordeling med tæthed

$$w(\vartheta; m, \gamma) = \frac{1}{C(m, \gamma)} \exp\{[\vartheta m - \kappa(\vartheta)]/\gamma\},$$

svarende til (6.1.1) og (6.1.7).

For de sædvanlige parametratiseringer ved $\mu = E[X|\vartheta]$ gælder da (jvf sætning 6.1.2) for den marginale fordeling af en enkelt måling, $X_i$

$$E[X_i] = m$$
$$V[X_i] = E[V(\mu)](\sigma^2 + \gamma)$$

hvor

$$\gamma = \frac{V[\mu]}{E[V(\mu)]}$$
For den marginale fordeling af summen, \( Z = \sum_{i=1}^{n} X_i \) af \( n \) målinger, der er frembragt med samme - tilfældige værdi af \( \mu \) - gælder da

\[
\begin{align*}
E[Z] &= n m \\
V[Z] &= n E[V(\mu)](\sigma^2 + n\gamma)
\end{align*}
\]

og endelig gælder for den marginale fordeling af gennemsnittet, \( \bar{X}_+ = \sum_{i=1}^{n} X_i/n \) af \( n \) målinger, der er frembragt med samme - tilfældige værdi af \( \mu \) - at

\[
\begin{align*}
E[\bar{X}_+] &= m \\
V[\bar{X}_+] &= E[V(\mu)] \left( \gamma + \frac{\sigma^2}{n} \right)
\end{align*}
\]

Såfremt \( \bar{X}_{1+}, \bar{X}_{2+}, \ldots, \bar{X}_{k+} \) angiver gennemsnittene i \( k \) uafhængige sæt af stikprøver (dvs med hver sin tilfældige værdi af \( \phi \)), og

\[
SAK_2 = \sum_{i=1}^{k} n_i(\bar{X}_{i+} - \bar{X}_{++})^2
\]

angiver den sædvanlige kvadratafvigelsessum, da gælder (jvf sætning 6.1.3), at

\[
E[SAK_2/(k-1)] = \frac{E[V(\mu)]}{\sigma^2} (1 + n_0\gamma)
\]

hvor den vægtede gennemsnitlige stikprøvestørrelse \( n_0 \) er givet ved (5.1.9).

Vi minder om, at størrelserne \( \gamma \) og \( E[V(\mu)] \) svarende til de sædvanlige fordelinger er resumeret i tabel 6.1.

### 8.3.2 Generelle resultater vedrørende aposteriorifordelinger

**Sætning 8.3.1 Aposteriorifordelning svarende til konjugeret apriorifordelning for sædvanlige dispersionsmodeller**

Lad \( X_1, X_2, \ldots, X_n \) være uafhængige, frembragt med samme \( \phi \) og med tæthed

\[
g(x|\phi) = d(x) \exp\left\{[\phi x - \kappa(\phi)]/\sigma^2\right\}, \tag{8.3.1}
\]
og lad apriorifordelingen af \( \vartheta \) være den konjugerede fordeling med tæthed

\[
w(\vartheta; m, \gamma) = \frac{1}{C(m, \gamma)} \exp\{[\vartheta m - \kappa(\vartheta)]/\gamma}\]  \hspace{1cm} (8.3.2)

Da afhænger aposteriorifordelingen af \( \vartheta \) efter observation af \( X_1 = x_1, \ldots, X_n = x_n \) alene af \( n \) og \( \bar{x} \).

Tætheden i aposteriorifordelingen af \( \vartheta \) efter observation af \( X_1 = x_1, \ldots, X_n = x_n \) er

\[
w\left( \vartheta \mid \sum X_i/n = \bar{x} \right) = \frac{1}{C(m_1, \gamma_1)} \exp\{[\vartheta m_1 - \kappa(\vartheta)]/\gamma_1}\]  \hspace{1cm} (8.3.3)

med

\[
m_1 = m_{\text{apost}} = \frac{m/\gamma + n\bar{x}/\sigma^2}{1/\gamma + n/\sigma^2} \hspace{1cm} (8.3.4)
\]
\[
\gamma_1 = \gamma_{\text{apost}} = \frac{1}{1/\gamma + n/\sigma^2} \hspace{1cm} (8.3.5)
\]

Aposteriorifordelingen for \( \vartheta \) er således af samme form som apriorifordelingen. Der er blot foretaget en opdatering af parametrene \( m \) og \( \gamma \).

**Bevis:**

Følger umiddelbart ved opskrivning af udtrykket for aposteriorifordelingen, og benyttelse af at \( Z = \sum X_i \) er sufficient for \( \vartheta \) i den betingede fordeling af \( X_1, \ldots, X_n \) for givet \( \vartheta \).

\[\square\]

Vi bemærker at tæthederne (8.3.1) og (8.3.2) svarer netop til (6.1.1) og (6.1.7).

**Bemærkning 1 Opdatering af apriorifordelingens parametre**

Vi bemærker, at aposteriorifordelingen af \( \vartheta \) er af samme form som apriorifordelingen. Der er blot foretaget en opdatering af parametrene \( m \) og \( \gamma \). Opdateringen er af formen:
\[
\frac{1}{\gamma_{\text{apost}}} = \frac{1}{\gamma_{\text{apriori}}} + \frac{n}{\sigma^2} \\
\frac{m_{\text{apost}}}{\gamma_{\text{apost}}} = \frac{m_{\text{apriori}}}{\gamma_{\text{apriori}}} + \frac{n}{\sigma^2} \bar{x}
\]

hvor \( \bar{x} = \frac{z}{n} = \frac{\sum x_i}{n} \).

Idet vi erindrer (sætning 6.1.1), at parameteren \( 1/\gamma = E[V(\mu)]/V[\mu] \) er et udtryk for den relative præcision i fordelingen af \( \theta \) i forhold til præcisionen i fordelingen af målestøjen, finder vi således, idet

\[
\gamma_{\text{apost}} = \frac{V[\mu|\bar{x}]}{E[V(\mu)|\bar{x}]} ,
\]

at

\[
\text{den relative præcision i aposteriorifordelingen er summen af den relative aprioripræcision og den relative stikprøvepræcision,}
\]

hvor den relative præcision af stikprøven måles som antallet af stikprøveenheder divideret med dispersionsparameteren \( \sigma^2 \).

Groft sagt, kan man altså for tolke parameteren \( 1/\gamma \) i apriorifordelingen som den aprioripræcision, der svarer til en stikprøve af størrelsen \( \sigma^2/\gamma \) fra en given gruppe. For "\( \gamma = \infty \)", d.v.s. \( 1/\gamma = 0 \) har man en "ikke-informativ" apriorifordeling.

Sammenholder vi (8.3.6) og (8.3.7) ser vi, at

\[
\text{forholdet mellem aposteriorivarianansen } V[\mu|\bar{x}] \text{ af } \mu \text{ og aposteriorimiddelværdien}
\]

\[
E[V(\mu)|\bar{x}] \text{ af variansfunktionen } V(\mu) \text{ afhænger alene af stikprøvestørrelsen}
\]

\( n \) (og aprioriforholdet \( \gamma_{\text{apriori}} \)), men ikke af det aktuelle stikprøveresultat \( \bar{x} \).

Vi kan således bestemme aposteriorivarianansen \( V[\mu|\bar{x}] \) ud fra aposteriorimid-
Aposteriorifordelinger

delværdien $E[V(\mu)|\bar{x}]$ af variansfunktionen $V(\mu)$ ved

$$V[\mu|\bar{x}] = \gamma_{apost} E[V(\mu)|\bar{x}] = \frac{E[V(\mu)|\bar{x}]}{1/\gamma_{apriori} + n/\sigma^2}.$$  (8.3.8)

Tabel 8.1 angiver aposteriorimiddelværdien $m_{apost} = E[\mu|\bar{x}]$ af $\mu$ for de sædvanlige fordelinger. Tabellen angiver desuden aposteriorimiddelværdien $E[V(\mu)|\bar{x}]$ af variansfunktionen. Aposteriorivariansen $V[\mu|\bar{x}]$ for $\mu$ kan da bestemmes af relationen (8.3.8).

Bemærkning 2: Aposteriorifordelingen af $\mu$ nærmer sig en etpunktsfordeling

Vi ser, at

$$\gamma_{apost} \to 0 \quad \text{for } n \to \infty$$

Endvidere finder vi, at

$$m_{apost} = \frac{m_{apriori}/\gamma_{apriori} + n\bar{x}/\sigma^2}{1/\gamma_{apriori} + n/\sigma^2} \to \bar{x} \quad \text{for } n \to \infty$$

For store stikprøvestørrelser vil aposteriorifordelingen af $\mu$ således nærme sig en etpunktsfordeling omkring stikprøvegennemsnittet $\bar{x}$.

Bemærkning 3: Den prædiktive fordeling af nye observationer fra samme gruppe

Lad $X_1, X_2, \ldots, X_n$ og $X'$ være sådan, at for fastholdt $\vartheta$ er $X_1, X_2, \ldots, X_n, X'$ uafhængige og identisk fordelte med en tæthed $g(\cdot|\vartheta)$ givet ved (8.3.1), dvs. $X'$ er frembragt af det samme $\vartheta$ som $X'$erne, og lad apriorifordelingen af $\vartheta$ være den konjugerede med tætheden $w(\vartheta;m,\gamma)$ givet ved (8.3.2).

Den prædiktive fordeling af $X'$ er af samme form som den marginale fordeling af $X$. Der er blot foretaget en opdatering af parametrene $m$ og $\gamma$. 
Lad $X'_1, \ldots, X'_r$ angive et sæt nye observationer, der er uafhængige og identisk fordelt, og med samme fordeling som $X_1, \ldots, X_n$, dvs specielt frembragt med samme værdi af $\vartheta$.

Lad $Z' = \sum_{j=1}^r X'_j$ angive summen af disse nye observationer, og lad

$$\overline{X'} = Z'/r = \sum_{j=1}^r X'_j/r$$

angive gennemsnittet.

Den prædiktive fordeling af $Z'$ og af $\overline{X'}$ er af samme form som fordelingen af hhv $Z = \sum X_i$ og $\overline{X}$. Der er blot foretaget en opdatering af parametrene $m$ og $\gamma$.

Vi kan derfor benytte sætning 6.1.2 til beskrivelse af momenterne i den prædiktive fordeling af $Z'$ og $\overline{X'}$, og vi finder at momenterne i den prædiktive fordeling af gennemsnittet $\overline{X'}$ af $r$ fremtidige observationer er:

$$E\left[\overline{X'} + \bar{x}\right] = m_{\text{apost}} \quad (8.3.9)$$

$$V\left[\overline{X'} + \bar{x}\right] = E\left[V(\mu) + \bar{x}\right] \left(\gamma_{\text{apost}} + \frac{\sigma^2}{r}\right) \quad (8.3.10)$$

Tilsvarende har vi for den prædiktive fordeling af summen $Z'$ af $r$ fremtidige observationer

$$E\left[Z'\mid \bar{x}\right] = r m_{\text{apost}} \quad (8.3.11)$$

$$V\left[Z'\mid \bar{x}\right] = r E\left[V(\mu)\mid \bar{x}\right](\sigma^2 + r \gamma_{\text{apost}}) \quad (8.3.12)$$
Bemærkning 4: Forventningsværdien i den prædiktive fordeling er et vejet gennemsnit

Forventningsværdien i den prædiktive fordeling af \( X' \) er

\[
m_{\text{apost}} = E[X' | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \frac{m/\gamma + n\bar{x}/\sigma^2}{1/\gamma + n/\sigma^2}
\]

Der gælder

\[
m_{\text{apost}} = \frac{E[V(\mu)] \times m + nV[\mu] \times \bar{x}/\sigma^2}{E[V(\mu)] + nV[\mu]/\sigma^2}
\]

Aposterioriforventningsværdien af en fremtidig observation fra den givne gruppe er således et vejet gennemsnit mellem aprioriforventningsværdien, \( m \), og stikprøvegennemsnittet \( \bar{x} \), hvor vægtene er de tilsvarende komponenter i opspaltung af den totale variation.

\[
m_{\text{apost}} = w_n m_{\text{apriori}} + (1 - w_n) \bar{x} \quad (8.3.13)
\]

med

\[
w_n = \frac{\sigma^2}{\sigma^2 + n\gamma} \quad \text{og} \quad 1 - w_n = \frac{n\gamma}{\sigma^2 + n\gamma}
\]

Parameteren \( \gamma \) i den prædiktive fordeling af \( X' \) er

\[
\gamma_{\text{apost}} = \frac{\sigma^2}{n} (1 - w_n) \quad (8.3.14)
\]

Bemærkning 5: Fortolkning af den prædiktive forventningsværdi som lineær prædiktor

Lad situationen være som i den foregående bemærkning. Da kan forventningsværdien i den prædiktive fordeling af \( X' \) fortolkes som regressionen af \( X' \) på \( \bar{x} \) ved udtrykket:

\[
m_{\text{apost}} = m + (1 - w_n)(\bar{x} - m) \quad (8.3.15)
\]
med

\[ w_n = \frac{\sigma^2}{\sigma^2 + n\gamma} \quad \text{og} \quad 1 - w_n = \frac{n\gamma}{\sigma^2 + n\gamma} \]

Vi kan fortolke vægten

\[ 1 - w_n = \frac{n\gamma}{\sigma^2 + n\gamma} = \frac{\text{COV}[\bar{X}, X']}{\text{V}[\bar{X}]} \]

som koefficienten til \( \bar{x} \) i regressionen af \( X' \) på \( \bar{x} \) i den simultane fordeling af \( \bar{X} \) og \( X' \).

Jo større værdi af \( n\gamma \), d.v.s jo mindre aprioriprecision, \( 1/\gamma \), (eller jo større stikprøvestørrelse, \( n \)), desto større vægt får stikprøvekorrektionen, \( (\bar{x} - m) \) til aprioriforventningsværdien, \( m \), ved fastsættelsen af aposterioriforventningsværdien.

\[ \square \]

**Bemærkning 6: Præaposteriorimiddelværdi og -varians**

Det følger af det generelle resultat vedrørende marginale og betingede middelværdier (Sætning 0.1.1 i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998), at præaposteriorimiddelværdien af \( \mu \) er

\[ E_{X_1, \ldots, X_n}[E[\mu|\bar{x}]] = E_{X_1, \ldots, X_n}[\mu] = m_{apriori} \]

Dette kunne også let verificeres ved at betragte udtrykket (8.3.13) for aposteriorimiddelværdien \( E[\mu|\bar{x}] = m_{apost} \).

Middelværdien (med hensyn til apriorifordelingen) af udtrykket (8.3.13) er

\[ E_{X_1, \ldots, X_n}[E[\mu|\bar{x}]] = E_{X_1, \ldots, X_n}[m_{apost}] \]

\[ = w_n m_{apriori} + (1 - w_n)E_{X_1, \ldots, X_n}[\bar{x}] \quad (8.3.16) \]

\[ = w_n m_{apriori} + (1 - w_n)m_{apriori} = m_{apriori} \quad (8.3.17) \]

Tilsvarende har vi for præaposteriorimiddelværdien af \( V(\mu) \),

\[ E_{X_1, \ldots, X_n}[V(\mu)|\bar{x}] = E_{X_1, \ldots, X_n}[V(\mu)] \]
Endelig følger det af (8.3.8) i Bemærkning 1, at præposteriorvariansen af \( \mu \) kan bestemmes ud fra præposteriorimiddelværdien af \( V(\mu) \) ved

\[
E_{X_1,\ldots,X_n}[V(\mu|\bar{x})] = \gamma_{\text{post}}E_{X_1,\ldots,X_n}[E[V(\mu)|\bar{x}]],
\]

hvorfor vi altså har

\[
E_{X_1,\ldots,X_n}[V(\mu|\bar{x})] = E[V(\mu)] \frac{1}{1/\gamma_{\text{apriori}} + n/\sigma^2},
\]

(8.3.18)

hvor \( E[V(\mu)] \) bestemmes i apriorifordelingen af \( \mu \).

I tabel 8.1 er apriorimiddelværdien \( E[V(\mu)] \) derfor anført i kolonnen for præposteriorimiddelværdien \( E_{X_1,\ldots,X_n}[E[V(\mu)|\bar{x}]] \) af \( V(\mu) \).

\[\square\]

For de almindeligt anvendte endimensionale fordelinger er sammenhængen mellem stikprøvefordeling, konjugeret apriorifordeling, tilsvarende marginal fordeling af observationer, og aposteriorifordeling for parameteren angivet i tabel 6.1 og 8.1. For normalfordelingen med varierende varians er en tilsvarende tabel anført i tabelform i tabel 8.3 på side 712.

I det følgende vil vi betragte en række eksempler på bestemmelse af aposteriorifordeling og prædiktiv fordeling svarende til disse standardsituationer. Vi vil specielt lægge vægt på at belyse sammenhængen mellem apriori- og aposteriorifordelingsparametre ved korrelations- og regressionsbetragtninger. Vi indleder med

### 8.3.3 Binomial-beta sampling

Lad \( X_i | p \in B(1,p), p \in \text{Be}(\alpha, \beta) \) og \( Z = X_1 + X_2 + \ldots + X_n \). Det følger da af sætning 6.2.2 at \( Z | p \in B(n,p) \), og endvidere har vi aposteriorifordelingen af \( p \):

\[
p | Z = z \in \text{Be}(\alpha', \beta')
\]

med

\[
\alpha' = \alpha + z; \quad \beta' = \beta + n - z,
\]

hvorfor

\[
E[p | Z = z] = \frac{\alpha'}{\alpha' + \beta'} = \frac{\alpha + z}{\alpha + \beta + n}
\]
og

\[ V[p | Z = z] = \frac{\alpha' \beta'}{(\alpha' + \beta')^2(\alpha' + \beta' + 1)} = \frac{(\alpha + z)(\beta + n - z)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \]

Den prædictive fordeling af \( Z' = X'_1 + \cdots + X'_r \) er en \( \mathrm{Pl}(r, \alpha', \alpha' + \beta') \)-fordeling.

Den sædvænligste parametrinisering af binomialfordelingen er netop middelvæRDIParametriniseringen

\[ \mu(p) = E[X_i | p] = p \]

og variansfunktionen er

\[ V_{Bin}(p) = V[X_i | p] = p(1 - p) \]

Indfører vi i overensstemmelse med betegnelserne i afsnit 6.2 den tilsvarende parametrinisering af strukturfordelingen

\[ \pi = E[p] = \frac{\alpha}{\alpha + \beta} \]

og

\[ \frac{1}{\gamma} = \frac{E[V_{Bin}(p)]}{V[p]} = \alpha + \beta, \]

ser vi i overensstemmelse med (8.3.13), at vi kan udtrykke aposterioriforventningsværdien af \( p \) (forventningsværdien i den prædictive fordeling af en ny observation, \( X' \) fra samme gruppe) som et vejet gennemsnit af apriori-forventningsværdi og stikprøveresultat

\[ E[p | Z/n = \bar{x}] = \frac{\alpha + n\bar{x}}{\alpha + \beta + n} = \frac{\pi/\gamma + n\bar{x}}{1/\gamma + n} = w_n \pi + (1 - w_n) \bar{x} \]

hvor vægten \( w_n = 1/(1 + n\gamma) \).

Sammenfattende har vi altså:

\[
\begin{align*}
\pi_{\text{aposteriori}} &= w_n \pi_{\text{apriori}} + (1 - w_n) \bar{x} \\
\frac{1}{\gamma_{\text{aposteriori}}} &= \frac{1}{\gamma_{\text{apriori}}} + n
\end{align*}
\]
Aposteriorisfordelinger

med

\[ w_n = \frac{1}{1 + n \gamma_{apriori}} \]

Aposteriorimiddelværdien \( E \left[ V_{\text{Bin}}(p) | \bar{x} \right] \) af \( V_{\text{Bin}}(p) \) er

\[ E \left[ V_{\text{Bin}}(p) | \bar{x} \right] = \frac{V_{\text{Bin}}(\pi_{apost})}{1 + \gamma_{apost}} = \frac{V_{\text{Bin}}(\pi_{apost})}{1/\gamma_{apriori} + n} \frac{1/\gamma_{apriori} + n + 1}{1/\gamma_{apriori} + n + 1} \]

Variansen i aposteriorifordelingen af \( p \) er derfor

\[ V \left[ p | \bar{x} \right] = \frac{E \left[ V_{\text{Bin}}(p) | \bar{x} \right]}{1/\gamma_{apriori} + n} = \frac{V_{\text{Bin}}(\pi_{apost})}{1/\gamma_{apriori} + n + 1} \]

Præposteriorimiddelværdien af \( V(p) \) er

\[ E_{X_1, \ldots, X_n} [V_{\text{Bin}}(p) | \bar{x}] = E \left[ V_{\text{Bin}}(p) \right] = \frac{V_{\text{Bin}}(\pi)}{1 + \gamma} \]

og præposteriorivariansen af \( p \) er

\[ E_{X_1, \ldots, X_n} [V(p)] = E \left[ V_{\text{Bin}}(p) \right] \frac{1}{1/\gamma_{apriori} + n} = \frac{V_{\text{Bin}}(\pi)}{1 + \gamma} \frac{1}{1/\gamma_{apriori} + n} \]

Den prædiktive fordeling

Momenterne i den prædiktive fordeling af \( \bar{X}' \) bliver jvf (8.3.9) og (8.3.10)

\[ E \left[ \bar{X}' + | \bar{x} \right] = \pi_{apost} \]

og

\[ V \left[ \bar{X}' + | \bar{x} \right] = E \left[ V_{\text{Bin}}(p) | \bar{x} \right] \left( \gamma_{apost} + \frac{1}{r} \right) \]

\[ = V_{\text{Bin}}(\pi_{apost}) \frac{1/\gamma_{apriori} + n}{1/\gamma_{apriori} + n + 1} \left( \frac{1}{1/\gamma_{apriori} + n + 1} + \frac{1}{r} \right) \]

Bemærkning 7 Uddybende bemærkninger

Vi vil uddybe ovenstående resultat med en nærmere redegørelse for samvariationen mellem tre sæt variable, der normalt vil indgå i betragtninger omkring bestemmelse af aposteriorifordelingen. De tre sæt variable, vi betragter, er
- stikprøven \( x_1, x_2, \ldots, x_n \) fra en given gruppe
- gruppeparameteren \( \theta \)
- Værdierne \( X'_1, X'_2, \ldots, X'_r \) af en fremtidig stikprøve fra den samme gruppe.

Vi antager, at \( X_1, X_2, \ldots, X_n \) og \( X'_1, X'_2, \ldots, X'_r \) for fastholdt \( p \) er indbyrdes uafhængige \( B(1,p) \)-fordelt med samme \( p \), og at \( p \in \text{Be}(\alpha, \beta) \)

Vi har altså de betingede sandsynligheder \( P[X_i = 1 \mid p] = p \), og de marginale sandsynligheder

\[
P[X_i = 1] = E[p] = \frac{\alpha}{\alpha + \beta} = \pi
\]

Endvidere har vi de marginale forventningsværdier og varianser (f.eks. fra Polyafordelingen med \( n = 1 \)):

\[
E[X_i] = \frac{\alpha}{\alpha + \beta} = \pi
\]

\[
V[X_i] = \frac{\alpha \beta}{(\alpha + \beta)^2} = \pi(1 - \pi)
\]

**Samvariationen mellem \( X_i \) og \( X_j \)**
Samvariationen kan f.eks. illustreres ved den simultane (marginale) frekvensfunktion

\[
P[X_i = x, X_j = z], \ (x,z) \in \{0,1\} \times \{0,1\}
\]

Vi har

\[
P[X_i = x, X_j = z \mid p] = p^{x+z}(1-p)^{2-x-z}
\]

således at den marginale sandsynlighed bliver

\[
P[X_i = x, X_j = z] = E[p^{x+z}(1-p)^{2-x-z}]
\]

De marginale sandsynligheder er anført i nedenstående skema:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( z )</th>
<th>( 0 )</th>
<th>( 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>( \frac{\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)} )</td>
<td>( \frac{\alpha \beta}{(\alpha+\beta)(\alpha+\beta+1)} )</td>
<td></td>
</tr>
<tr>
<td>( 1 )</td>
<td>( \frac{\alpha \beta}{(\alpha+\beta)(\alpha+\beta+1)} )</td>
<td>( \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} )</td>
<td></td>
</tr>
</tbody>
</table>
Samvariationen udtrykkes sædvanligvis ved kovariansen,

\[
\text{COV}[X_i, X_j] = E[\text{COV}[X_i, X_j | p]] + \text{COV}[E[X_i | p], E[X_j | p]]
\]

\[
= 0 + V[p] = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}
\]

\[
= \pi(1 - \pi) \frac{\gamma}{1 + \gamma} = \pi(1 - \pi)(1 - w_1)
\]

således at korrelationskoeficienten mellem \( X_i \) og \( X_j \) bliver

\[
\rho_{X_i, X_j} = \frac{1}{\alpha + \beta + 1} = \frac{\gamma}{1 + \gamma} = 1 - w_1
\]

Jo større værdi af \( \gamma \), d.v.s. jo mindre præcis apriorifordeling, desto større er korrelationen mellem enkeltobservationer fra samme gruppe (intraklyngekorrelationen).

**Samvariationen mellem \( \overline{X} \) og \( p \)**

Idet vi sætter \( \overline{X} = (X_1 + X_2 + \ldots + X_n) / n \), har vi den betingede forventningsværdi og varians

\[
E[\overline{X} | p] = p,
\quad V[\overline{X} | p] = \frac{V_{\text{Bin}}(p)}{n} = \frac{p(1 - p)}{n}
\]

og den marginale forventningsværdi og varians fås fra Polyafordelingen

\[
E[\overline{X}] = \frac{\alpha}{\alpha + \beta} = \pi
\]

\[
V[\overline{X}] = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \left(1 + \frac{\alpha + \beta}{n}\right)
\]

\[
= \pi(1 - \pi) \frac{\gamma}{1 + \gamma} \left(1 + \frac{1}{n \gamma}\right) = \pi(1 - \pi) \frac{1 - w_1}{1 - w_n}
\]

Endvidere har vi kovariansen mellem \( \overline{X} \) og \( p \)

\[
\text{COV}[\overline{X}, p] = V[p] = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}
\]

\[
= \pi(1 - \pi)(1 - w_1)
\]

hvorfor korrelationskoeficienten mellem gruppestikprøvegennemsnit, \( \overline{X} \), og gruppeparameter, \( p \), er

\[
\rho_{\overline{X}, p} = \frac{1}{\sqrt{1 + (\alpha + \beta) / n}} = \sqrt{\frac{n \gamma}{1 + n \gamma}} = \sqrt{1 - w_n}
\]
d.v.s. jo større stikprøvelængde, \( n \), (eller jo mindre aprioripræcision), desto større er korrelationen mellem stikprøvegennemsnit og gruppeparameter.

**Samvariationen mellem \( \bar{X} \) og \( \bar{X}' \).**

I modellen symboliserer \( \bar{X}' \) gennemsnittet af \( r \) fremtidige observationer, hidrørende fra den samme gruppe, som de allerede kendte \( X \)-værdier, (d.v.s. med den samme ukendte værdi af \( p \)). Efter observation af \( X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \) beskrives vores viden om \( p \) ved aposteriorifordelingen af \( p \), d.v.s. en \( \text{Be}(\alpha + n\bar{x}, \beta + n(1 - \bar{x})) \) fordeling, hvorfor den prædiktive fordeling af \( Z' = X'_1 + X'_2 + \cdots + X'_r \) er en \( \text{Pl}(r, \alpha + n\bar{x}, \alpha + \beta + n) \)-fordeling, nemlig den marginale fordeling af summen af \( r \) observationer under hensyntagen til den opdaterede viden om \( p \).

Den prædiktive fordeling af \( Z' \) kan imidlertid også betrages som den betingede fordeling af \( Z' \) i den simultane fordeling af \( \bar{X} \) og \( Z' \) (hvor vi har bortintegriteret fordelingen af \( p \)).

I stedet for at betragte fordelingen af totalen, \( Z' \), vil vi betragte gennemsnittet af de fremtidige observationer,

\[
\bar{X}' = \frac{Z'}{r}
\]

Korrelationen mellem \( \bar{X} \) og \( \bar{X}' \) i den simultane fordeling af \( \bar{X} \) og \( \bar{X}' \) er

\[
\rho_{\bar{X}, \bar{X}'} = \frac{1}{\sqrt{1 + (\alpha + \beta)/n}} \frac{1}{\sqrt{1 + (\alpha + \beta)/r}} = \frac{1}{\sqrt{1 + 1/n(\gamma)}} \frac{1}{\sqrt{1 + 1/r(\gamma)}}
\]

\[
= \sqrt{1 - w_n} \sqrt{1 - w_r}
\]

Den prædiktive fordeling af \( \bar{X}' \) kan altså fortolkes som regressionen af \( \bar{X}' \) på \( \bar{x} \) i denne fordeling. Vi har således:

\[
E[\bar{X}' | \bar{X} = \bar{x}] = \pi + (1 - w_n)(\bar{x} - \pi)
\]

med

\[
1 - w_n = \frac{\text{COV}[\bar{X}, \bar{X}']}\sqrt{\text{V}[\bar{X}]} = \frac{n\gamma}{1 + n\gamma}
\]

Endvidere har vi

\[
\text{V}[\bar{X}' | \bar{X} = \bar{x}] = \frac{(\alpha + n\bar{x})(\beta + n(1 - \bar{x}))}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} \left( 1 + \frac{\alpha + \beta + n}{r} \right)
\]

\[
\Box
\]

**Eksempel 8.3.1 Andelen af afvigende enheder i en produktion**
Vi betragter en situation svarende til den, der blev behandlet i eksempel 6.2.1.

Antag, at andelen \( p \) af afvigende enheder i en produktion varierer fra produktion til produktion i overensstemmelse med en \( \text{Be}(\alpha, \beta) \)-fordeling med

\[
\pi = E[p] = 0.1
\]

og

\[
\gamma = \frac{V[p]}{E[V_{\text{Bin}}(p)]} = 0.05 = \frac{1}{20}
\]

svarende til

\[
V[p] = \gamma \frac{V_{\text{Bin}}(\pi)}{1 + \gamma} = 0.05 \times \frac{0.1 \times 0.9}{1.05} = 0.004286 = (0.0655)^2
\]

Antag, at der udtages en stikprøve på \( n = 15 \) enheder fra en given produktion, og at man finder \( Z = 1 \) afvигende enheder i stikprøven.

Man ønsker nu at udtale sig om andelen af afvigende enheder blandt 100 tilfældigt udtagne enheder fra denne produktion.

Idet \( n = 15 \) og \( \gamma = 0.05 \) finder man

\[
w = \frac{1}{1 + 15 \times 0.05} = \frac{1}{1.75} = 0.571
\]

og

\[
1 - w = 0.429
\]

Idet \( \bar{x} = 1/15 = 0.06667 \) og \( \pi = 0.10 \) finder man aposteriorimiddelværdien

\[
\pi_{\text{apost}} = 0.571 \times 0.10 + 0.429 \times 0.06667 = 0.0857
\]

Den forventede andel afvигende enheder i denne produktion er således \( \pi_{\text{apost}} = 0.0857 \), og specielt er den forventede andel afvигende enheder blandt 100 tilfældigt udtagne

\[
E[\bar{X}|\bar{x} = 0.06667] = \pi_{\text{apost}} = 0.0857
\]

Til beskrivelse af variansen bestemmer man aposteriorimiddelværdien af \( V_{\text{Bin}}(p) \),

\[
E[V_{\text{Bin}}(p)|\bar{x}] = 0.06667 = V_{\text{Bin}}(0.0857) \frac{1/0.05 + 15}{1/0.05 + 15 + 1}
\]

\[
= 0.0857 \times 0.9143 \times \frac{35}{36}
\]
Man får derfor

\[ E[V_{Bin}(p)|\bar{x} = 0.06667] = 0.0784 \times 0.9722 = 0.0762, \]

og endvidere bestemmer man

\[ \gamma_{apost} = \frac{1}{\gamma_{apriori} + 15} = \frac{1}{20 + 15} = 0.028571 \]

Variansen i den prædiktive fordeling for \( \overline{X}_+ \) er således

\[
\begin{align*}
V[\overline{X}_+|\bar{x}] &= 0.06667 = E[V_{Bin}(p)|\bar{x} = 0.06667] \left( \gamma_{apost} + \frac{1}{\bar{x}} \right) \\
&= 0.0762 \times (0.028571 + 0.01)
\end{align*}
\]

Vi bemærker, at da aprioriusikkerheden \( \gamma = 1/20 \) så nogenlunde modsvarer stikprøvestørrelsen \( n = 15 \), bliver aposteriorimiddelværdien \( \pi_{apost} \) stort set et simpelt gennemsnit mellem apriorimiddelværdien \( \pi = 0.1 \) og stikprøveresultatet \( \bar{x} = 0.06667 \).

Aposteriorifordelingen for \( p \) er mere koncentreret, end apriorifordelingen. Vi har nemlig:

\[
\begin{align*}
V[p|\bar{x} = 0.06667] &= \gamma_{apost} E[V_{Bin}(p)|\bar{x} = 0.06667] \\
&= 0.028571 \times 0.0762 = 0.0002176 = (0.0467)^2
\end{align*}
\]

Aposteriorimiddelværdien \( E[V_{Bin}(p)|\bar{x} = 0.06667] \) af \( V_{Bin}(p) \) er ændret noget i forhold til aprioriværdien

\[ E[V_{Bin}(p)] = V_{Bin}(\pi)/1.05 = 0.0857 \]

Ændringen skyldes hovedsageligt, at fordelingen af \( p \) har ændret middelværdi. Endelig bemærker vi ved at sammenholde udtrykket for variansen i den prædiktive fordeling af \( \overline{X}_+ \)

\[ V[\overline{X}_+|\bar{x} = 0.06667] = 0.0762 \times (0.028571 + 0.01), \]

med den tilsvarende "prædiktionsvarians" såfremt vi ikke havde undersøgt en stikprøve fra partiet

\[ V[\bar{X}_+] = E[V_{Bin}(p)] \left( \gamma_{apriori} + \frac{1}{100} \right) = 0.0857 \times (0.05 + 0.01), \]

at usikkerheden på prædiktionen er formindsket svarende til reduktionen i usikkerheden vedrørende \( p \).
8.3.4 Negativ binomial- beta sampling

Lad $X_i | p \in \text{Geo}(p)$, $p \in \text{Be}(\alpha, \beta)$ og $Z = X_1 + X_2 + \cdots + X_n$. Det følger da af Sætning 6.3.2, at $Z | p \in \text{NB}(n, p)$ og man finder $p | Z = z \in \text{Be}(\alpha + n, \beta + z)$ med

$$E[p | Z = z] = \frac{\alpha + n}{\alpha + \beta + n + z}$$

$$V[p | Z = z] = \frac{(\alpha + n)(\beta + z)}{(\alpha + \beta + n + z)^2(\alpha + \beta + n + 1)}$$

Indfører vi i overensstemmelse med betegnelserne i afsnit 6.3 middelværdi-parametreriseringen ved

$$\mu(p) = E[X_i | p] = \frac{1 - p}{p}$$

og variansfunktionen

$$V_{NB}(\mu) = V[X_i | p] = \mu(1 + \mu)$$

med den tilsvarende parametrerisering af strukturforselingen

$$\psi = E[\mu] = \frac{\beta}{\alpha - 1}$$

og

$$\frac{1}{\gamma} = \frac{E[V_{NB}(\mu)]}{V[\mu]} = \alpha - 1$$

i overensstemmelse med betegnelserne i Lemma 6.3.1, finder man, at aposterioriforventningsværdien af $\mu$ (forventningsværdien i den prædiktive fordeling af en ny observation, $X'$ fra samme gruppe) er

$$E[X' | Z/n = \bar{x}] = E[\mu | Z/n = \bar{x}] = \frac{\psi/\gamma + n\bar{x}}{1/\gamma + n}$$

Vi har således

$$\psi_{aposteriori} = w_n \psi_{apriori} + (1 - w_n) \bar{x}$$

$$\frac{1}{\gamma_{aposteriori}} = \frac{1}{\gamma_{apriori}} + n$$
med

\[ w_n = \frac{1}{1 + n \gamma_{apriori}} \]

Udtrykker vi aposterioriforventningsværdien som regressionen af \( Y \) på \( \bar{x} \) har vi

\[ E [X' | Z/n = \bar{x}] = \psi + (1 - w_n)(\bar{x} - \psi) \]

### 8.3.5 Poisson-Gamma sampling

Lad \( X_i | \mu \in P(\mu), \mu \in G(\alpha, 1/\beta) \) og \( Z = X_1 + X_2 + \ldots + X_n \). Det gælder da, at \( Z | \mu \in P(n\mu) \), hvorfor det følger af sætning 6.4.2, at den marginale fordeling af \( Z \) er en \( NB(\alpha, \beta/(\beta + n)) \)-fordeling.

Det gælder da, at aposteriorifordelingen \( \mu | Z = z \in G(\alpha + z, 1/(\beta + n)) \) med

\[
E [\mu | Z = z] = \frac{\alpha + z}{\beta + n} \\
V [\mu | Z = z] = \frac{\alpha + z}{(\beta + n)^2}
\]

Den sædvanlige parametrsering af Poissonfordelingen er netop middelværdiparametrseringen \( \mu = E[X_i | \mu] \) og variansfunktionen er

\[ V_P(\mu) = \mu \]

Indfører vi i overensstemmelse med betegnelserne i afsnit 6.4 den tilsvarende parametrsering af strukturfordelingen

\[ m = E[\mu] = \frac{\alpha}{\beta}, \quad \text{og} \quad \frac{1}{\gamma} = \frac{E[V_P(\mu)]}{V[\mu]} = \beta, \]

kan vi udtrykke aposterioriforventningen af \( \mu \) (forventningsværdien i den prædikttive fordeling af en ny observation, \( X' \) fra samme gruppe) som:

\[ E [X' | Z/n = \bar{x}] = E [\mu | Z/n = \bar{x}] = \frac{\alpha + n\bar{x}}{\beta + n} = \frac{m/\gamma + n\bar{x}}{1/\gamma + n} \]

Vi har således
Aposteriorifordelinger

\[
m_{\text{aposteriori}} = w_n m_{\text{apriori}} + (1 - w_n) \bar{x} \quad (8.3.19)
\]

\[
\frac{1}{\gamma_{\text{aposteriori}}} = \frac{1}{\gamma_{\text{apriori}}} + n \quad (8.3.20)
\]

med

\[
w_n = \frac{1}{1 + n \gamma_{\text{apriori}}}
\]

Udtrykt ved regressionen af \(X'\) på \(\bar{x}\) har vi:

\[
E \left[ X' \mid Z/n = \bar{x} \right] = m + (1 - w_n)(\bar{x} - m)
\]

**Eksempel 8.3.2 Vævefejl i stof**

Vi betragter en produktion af klæde. Fra tidligere produktioner har man erfaring for, at antallet \(X\) af vævefejl i stikprøver på 1 \([m^2]\) udtaget tilfældigt af en produktion varierer henover produktionen i overensstemmelse med en \(P(\mu)\)-fordeling, hvor det gennemsnitlige antal fejl pr \(m^2\), \(\mu\), varierer fra produktion til produktion i overensstemmelse med en \(G(\alpha, 1/\beta)\)-fordeling med \(\alpha = 3\) og \(\beta = 1.2\).

Dette kunne eksempelvis svare til at man ved stikprøver på 10 \([m^2]\) fra hver af \(k\) produktioner har observeret \(s k_2/(k-1) = 23.325\) og et gennemsnitligt antal fejl på \(\bar{x}_{++} = 2.5\) [fejl/m\(^2\)].

Vi har nu af tabel 6.1, at

\[
m = E[\mu] = \frac{\alpha}{\beta} = 2.5 \text{ [fejl/m}^2\text{]}
\]

og

\[
\gamma = \frac{V[\mu]}{E[V_P(\mu)]} = \frac{1}{\beta} = 0.833
\]

Det følger af Sætning 6.4.2 at den marginale fordeling af \(Z = \sum_{i=1}^{n} X_i\) har momenterne

\[
E[Z] = nm \quad (8.3.21)
\]

\[
V[Z] = nm(1 + n\gamma) \quad (8.3.22)
\]
Antag nu, at der udføres en ny vævning, og der udtages en stikprøve bestående af \( n = 10 \) stykker på hver 1 [m²] fra produktionen, og antallet af fejl \( x_1, x_2, \ldots, x_{10} \) optælles. Antag, at man fandt det gennemsnitlige antal fejl,

\[
\bar{x} = 0.1 \text{ [fejl/m}^2]\]

i stikprøven.

Man finder nu af (8.3.20) at

\[
\frac{1}{\gamma_{\text{aposteriori}}} = \frac{1}{0.833} + 10
\]

hvoraf \( \gamma_{\text{aposteriori}} = 0.0893 \).

Idet \( n\gamma_{\text{apriori}} = 8.33 \) har man

\[
w_{10} = \frac{1}{1 + 8.33} = 0.1072 ,
\]

hvorfor (8.3.19) fører til

\[
\mu_{\text{aposteriori}} = 0.1072 \times 2.5 + 0.8928 \times 0.1 = 0.36 \text{ [fejl/m}^2]\]

Vi kan nu udtale os fremtidige prøver fra denne produktion. Lad \( Z' = \sum_{j=1}^{r} X_j' \) angive det totale antal fejl i \( r \) prøver á 1 [m²]. Der gælder da jvf. (8.3.21) og (8.3.22), at

\[
\begin{align*}
\mathbb{E} [Z'] &= r \mu_{\text{aposteriori}} \\
\mathbb{V} [Z'] &= r \mu_{\text{aposteriori}} (1 + r \gamma_{\text{aposteriori}})
\end{align*}
\]

For en prøve bestående af 20 [m²] vil vi altså forvente det totale antal fejl

\[
\mathbb{E} [Z'] = 20 \times 0.36 = 7.2 \text{ [fejl/m}^2]\]

Variansen på denne prædiktion er

\[
\mathbb{V} [Z'] = 20 \times 0.36(1 + 20 \times 0.0893) = 20.06 \text{ [fejl/m}^2]^2
\]

Vi kunne naturligvis også have fået disse resultater ved brug af Tabel 8.1, hvorefter den prædiktive fordeling af \( Z' \) bestemmes som den marginale fordeling af \( Z' \) svarende til \( n = 20 \) og de opdaterede værdier af \( \alpha \) og \( \beta \).

Nedenstående tabel illustrierer opdateringen svarende til forskellige stikprøvestørrelser og forskellige stikprøveresultater:
Eksempel på bestemmelse af aposteriorifordeling ved Poissonfordelt målestøj

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prøv</th>
<th>Efter observation af</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$n = 10$</td>
<td>$n = 10$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$z = 1$</td>
<td>$z = 60$</td>
</tr>
<tr>
<td>$m = E[\mu]$</td>
<td>2.5</td>
<td>0.36</td>
<td>5.63</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>0.833</td>
<td>0.0893</td>
<td>0.0893</td>
</tr>
<tr>
<td>$\sqrt{\mu} = m \times \gamma$</td>
<td>2.0825</td>
<td>0.03</td>
<td>0.50</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>3</td>
<td>4</td>
<td>63</td>
</tr>
<tr>
<td>$\beta$</td>
<td>1.2</td>
<td>11.2</td>
<td>11.2</td>
</tr>
</tbody>
</table>

Fordeling af fejl i én ny m

$X' \in \text{NB}(\alpha, \beta/(1 + \beta))$

| $E[X'] = m$ | 2.5 | 0.36 | 5.63 | 5.80 |
| $V[X'] = m(1 + \gamma)$ | 4.58 | 0.39 | 6.13 | 6.07 |

Fordeling af fejl i 10 nye m

$Z' \in \text{NB}(\alpha, \beta/(10 + \beta))$

| $E[Z'] = 10m$ | 25  | 3.6  | 56.3 | 58.0 |
| $V[Z'] = 10m(1 + 10\gamma)$ | 233.32 | 6.81 | 106.57 | 85.38 |

Fordeling af fejl i 20 nye m

$Z' \in \text{NB}(\alpha, \beta/(20 + \beta))$

| $E[Z'] = 20m$ | 50  | 7.2  | 112.6 | 116.0 |
| $V[Z'] = 20m(1 + 20\gamma)$ | 883 | 20.06 | 313.7 | 225.50 |

8.3.6 Exponential reciprok gamma sampling

Lad $X_i | \mu \in \text{Ex}(\mu)$, $\mu \in \text{RGam}(\alpha, \beta)$ og $Z = X_1 + X_2 + \cdots + X_n$. Det gælder da, at $Z | \mu \in G(n, \mu)$ og det følger nu af sætning 6.5.2, at den marginale fordeling af $Z$ er en $\text{RBet}(\alpha, n, \beta)$-fordeling, og endvidere har vi,
at \( \mu \mid Z = z \in \text{RGam}(\alpha + n, \beta + z) \) med

\[
E[\mu \mid Z = z] = \frac{\beta + z}{\alpha - 1 + n}
\]

\[
V[\mu \mid Z = z] = \frac{\beta + z}{(\alpha + n - 1)^2(\alpha + n - 2)}.
\]

I afsnit 6.5 betragtede i middelværdiparametrizeringen af gamafordelingen ved

\[E[X_i \mid \mu] = \mu\]

og variansfunktionen

\[V_G(\mu) = \mu^2\]

svarende til dispersionsparameter 1, og vægt \(1/n\). Indfører vi den tilsvarende parametrizering af strukturfordeelingen (jvf lemma 6.5.1)

\[m = E[\mu] = \frac{\beta}{\alpha - 1}\]

og

\[\frac{1}{\gamma} = \frac{E[V_G(\mu)]}{V[\mu]} = \alpha \quad 1,\]

can vi utrykke forventningsværdien i den prædiktive fordeling af en ny observation, \(X'\) fra samme gruppe (aposterioriforventningsværdien af \(\mu\)) som

\[
E[X' \mid Z/n = \bar{x}] = E[\mu \mid Z/n = \bar{x}] = \frac{m/\gamma + n\bar{x}}{1/\gamma + n}
\]

\[
m_{\text{aposteriori}} = w_n m_{\text{apriori}} + (1 - w_n)\bar{x}
\]

\[
\frac{1}{\gamma_{\text{aposteriori}}} = \frac{1}{\gamma_{\text{apriori}}} + n
\]

med

\[w_n = \frac{1}{1 + n\gamma_{\text{apriori}}}\]

Udtrykt ved regressionen af \(X'\) på \(\bar{x}\) har vi

\[
E[X' \mid Z/n = \bar{x}] = m + (1 - w_n)(\bar{x} - m)
\]
8.3.7 Normalfordeling med samme varians

Såfremt $X_i | \mu \in N(\mu, \sigma^2)$, $\mu \in N(m, \sigma_0^2)$ og $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$, da vil $\bar{X} | \mu \in N(\mu, \sigma^2/n)$ og den marginale fordeling af $\bar{X}$ er (jvf bemærkning 3 til sætning 5.3.2) $\bar{X} \in N(m, \sigma^2(\gamma + 1/n))$, hvor

$$\gamma = \frac{\sigma_0^2}{\sigma^2}.$$

Aposteriorifordelingen af $\mu$ efter observation af $\bar{X} = \bar{x}$ er en normalfordeling med forventningsværdi

$$E[\mu | \bar{X} = \bar{x}] = \frac{m/\sigma_0^2 + n\bar{x}/\sigma^2}{1/\sigma_0^2 + n/\sigma^2} = \frac{m/\gamma + n\bar{x}}{1/\gamma + n} = w_nm + (1 - w_n)\bar{x}$$

med $w = 1/(1 + n\gamma)$, og varians

$$\sigma_1^2 = V[\mu | \bar{X} = \bar{x}] = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}.$$

Vi ser, at også her gælder, at aposterioriforventningsværdien $E[\mu | \bar{X} = \bar{x}]$ er et vejet gennemsnit af aprioriforventningsværdien $m$ og stikprøvegennemsnittet $\bar{x}$ med de respektive præcisioner (reciproke varianser)

$$\frac{1}{V[\mu]} = \frac{1}{\sigma_0^2} \text{ og } \frac{n}{E[V(\mu)]} = \frac{n}{\sigma^2}$$

som vægte. Da det kun er de relative vægte, der er af betydning, har vi som tidligere valgt at parametrisere ved parameteren $\gamma$.

Sammenfattende har vi:

$$\mu_{\text{aposteriori}} = w_n\mu_{\text{apriori}} + (1 - w_n)\bar{x}$$

$$\frac{1}{\gamma_{\text{aposteriori}}} = \frac{1}{\gamma_{\text{apriori}}} + n$$

med

$$w_n = \frac{1}{1 + n\gamma_{\text{apriori}}}$$
Bemærkning 1 Uddybende bemærkninger

På grund af normalfordelingens udbredte anvendelse, vil vi i lighed med binomialfordelingssituationen uddybe fortolkningen af aposteriorifordelingen og den prædiktive fordeling.

Vi lader

- $x_1, x_2, \ldots, x_n$ betegne stikprøveresultatet fra en given gruppe
- $(\mu, \sigma^2)$ betegne middelværdi og varians for observationer fra den pågældende gruppe
- $X'_1, X'_2, \ldots, X'_r$ betegne observationer fra en fremtidig stikprøve fra denne gruppe

Vi antager, at $X_1, X_2, \ldots, X_n$ og $X'_1, X'_2, \ldots, X'_r$ for fastholdt $\mu$ er indbyrdes uafhængige $N(\mu, \sigma^2)$-fordelt med samme $\mu$, og at $\mu \in N(m, \sigma_0^2)$ med $\sigma_0^2 = \gamma \sigma^2$.

**Samvariationen mellem $X_i$ og $X_j$**

Vi minder om, at der gælder (Sætning 5.3.2)

$$\text{COV}[X_i, X_j] = \text{E}[\text{COV}[X_i, X_j | \mu]] + \text{COV}[\text{E}[X_i | \mu], \text{E}[X_j | \mu]]$$

$$= 0 + \text{V}[\mu] = \sigma^2_0$$

hvorfor vi har intraklyngekorrelationen

$$\rho_{X_i, X_j} = \frac{1}{1 + \sigma^2 / \sigma_0^2} = \frac{\gamma}{1 + \gamma}$$

**Samvariationen mellem $\bar{X}$ og $\mu$**

Idet $\bar{X} = (X_1 + X_2 + \cdots + X_n)/n$, har vi den betingede forventningsværdi og varians

$$\text{E}[\bar{X} | \mu] = \mu; \quad \text{V}[\bar{X} | \mu] = \sigma^2 / n$$

Den marginale forventningsværdi og varians er

$$\text{E}[\bar{X}] = \mu_0; \quad \text{V}[\bar{X}] = \sigma_0^2 + \sigma^2 / n$$

og kovariansen mellem $\bar{X}$ og $\mu$ er

$$\text{COV}[\bar{X}, \mu] = \text{E}[\text{COV}[\bar{X}, \mu] | \mu] + \text{COV}[\text{E}[\bar{X} | \mu], \mu] = 0 + \text{COV}[\mu, \mu] = \sigma_0^2$$

hvorfor vi har korrelationskoefficienten mellem stikprøvegennemsnit, $\bar{X}$ og gruppeparameter, $\mu$

$$\rho_{\bar{X}, \mu} = \frac{1}{\sqrt{1 + \sigma^2/(n\sigma_0^2)}} = \frac{1}{\sqrt{1 + 1/(n\gamma)}} = \sqrt{1 - w_n}$$
med \( w_n = 1/(1 + n\gamma) \)

For \( n \to \infty \) vil \( \rho_{\bar{X}, \mu} \to 1 \).

Aposteriorifordelingen af \( \mu \) givet \( \bar{X} = \bar{x} \) karakteriseres ved regressionen

\[
E[\mu | \bar{X} = \bar{x}] = m + (1 - w_n)(\bar{x} - m)
\]

\[
V[\mu | \bar{X} = \bar{x}] = \frac{1}{n(1/\sigma_0^2 + n/\sigma^2)} = \frac{\sigma^2}{n} (1 - w_n)
\]

For \( n \to \infty \) vil \( E[\mu | \bar{X} = \bar{x}] \to \bar{x} \), og \( V[\mu | \bar{X} = \bar{x}] \to 0 \).

**Samvariationen mellem \( \bar{X} \) og \( \bar{X}' \)**

I modellen symboliserer \( \bar{X}' \) gennemsnittet af \( r \) fremtidige observationer, hidrørende fra den samme gruppe, som de allerede kendte værdier af \( X \) (d.v.s. med den samme ukendte værdi af \( \mu \)). Efter observation af \( X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \) beskrives vores viden om \( \mu \) ved aposteriorifordelingen af \( \mu \) som beskrevet ovenfor, hvorfor den prædiktive fordeling af totalen \( Z'_+ = X'_1 + X'_2 + \cdots + Z'_r \) er en \( N(rm_1, r\sigma_0^2 + r^2\sigma_1^2) \)-fordeling, nemlig den marginale fordeling af summen af \( r \) observationer under hensyntag til den opdaterede viden om \( \mu \).

Korrelationen mellem \( \bar{X} \) og \( \bar{X}' \) i den simultane fordeling af \( \bar{X} \) og \( \bar{X}' \) er

\[
\rho_{\bar{X}, \bar{X}'} = \frac{1}{\sqrt{1 + \sigma^2/(n\sigma_0^2)}} \cdot \frac{1}{\sqrt{1 + \sigma^2/(r\sigma_0^2)}} = \frac{1}{\sqrt{1 + 1/(n\gamma)}} \cdot \sqrt{1 + 1/(r\gamma)}
\]

\[
= \sqrt{1 - wn} \sqrt{1 - w_r}
\]

Den prædiktive fordeling af \( \bar{X}' \) svarende til \( \bar{X} = \bar{x} \) kan da fortolkes som regressionen af \( \bar{X}' \) på \( \bar{x} \) i den simultane fordeling af \( \bar{X} \) og \( \bar{Y} \) (hvor vi har bortintegreret \( \mu \)).

\[
E[\bar{X}' | \bar{X} = \bar{x}] = m + (1 - w_n)(\bar{x} - m)
\]

\[
V[\bar{X}' | \bar{X} = \bar{x}] = \frac{\sigma^2}{p} + \frac{1}{1/\sigma_0^2 + n/\sigma^2} = \frac{\sigma^2}{p} + \frac{\sigma^2}{n} (1 - w_n)
\]

\[\Box\]

### 8.3.8 Empiriske varianser fra normalfordelte observationer

Såfremt \( S^2 | \sigma^2 \in \sigma^2 \chi^2(f) \) og

\[
1/\sigma^2 \in G(\nu/2, 2/[(\nu - 2)\sigma_0^2])
\]
da vil aposteriorifordelingen af $1/\sigma^2$ efter observation af $S^2 = s^2$ være en

$$G(\nu/2, 2/[(\nu - 2)\sigma_1^2])$$

fordeling

hvor

$$\nu_1 = \nu + f \quad (8.3.23)$$

og

$$\sigma_1^2 = \frac{(\nu - 2)\sigma_0^2 + fs^2}{\nu - 2 + f} \quad (8.3.24)$$

Aposterioriforventningsværdien af $\sigma^2$ er

$$E[\sigma^2|s^2] = \sigma_1^2$$

Det ses, at aposteriorimiddelværdien er et vejel gennemsnit mellem apriorimiddelværdien

$$E[\sigma^2] = \sigma_0^2$$

og stikprøveresultatet $s^2$ med vægtene henholdsvis $\nu - 2$ og $f$.

Den prædiktive fordeling for $(S')^2$, hvor $(S')^2|\sigma^2 \in \sigma_0^2 \chi^2(r)/r$ og $1/\sigma^2$ følger en $G(\nu_1/2, 2/[(\nu_1 - 2)\sigma_1^2])$-fordeling bliver en

$$RBet\left(\frac{\nu_1}{2}, \frac{r}{2}, \frac{\nu_1 - 2}{r - \sigma_1^2}\right)$$

fordeling

jvf afsnit 6.6.
| Stikprøvefordeling af $Z|\mu$ | Strukturfordeling $w(\cdot)$ | Aposteriorifordeling af $\mu$ efter observation af $Z = z$ | Prædiktivfordeling af $Z'$ |
|-----------------------------|-----------------------------|-------------------------------------------------|-----------------|
| $Z|p \in B(n,p)$ | $p \in \text{Be}(\alpha,\beta)$ | $\text{Be}(\alpha + z, \beta + n - z)$ | $\text{Pl}(r, \alpha + z, \alpha + \beta + n)$ | Afsn. 8.3.1 |
| $Z|p \in \text{NB}(n,p)$ | $p \in \text{Be}(\alpha,\beta)$ | $\text{Be}(\alpha + n, \beta + z)$ | $\text{NPl}(r, \beta + z, \alpha + \beta + n + z)$ | Afsn. 8.3.2 |
| $Z|\mu \in P(n\mu)$ | $\mu \in G(\alpha, 1/\beta)$ | $G(\alpha + z, 1/(\beta + n))$ | $\text{NB}(\alpha + z, (\beta + n)/(\beta + n + r))$ | Afsn. 8.3.5 |
| $Z|\mu \in \text{G}(n\mu)$ | $\mu \in \text{RGam}(\alpha,\beta)$ | $\text{RGam}(\alpha + n, \beta + z)$ | $\text{RBet}(\alpha + n, r, \beta + z)$ | Afsn. 8.3.6 |
| $Z|\mu \in \text{N}(n\mu, n\sigma^2)$ | $\mu \in \text{N}(m, \sigma_0^2)$ | $\mu \in \text{N}(m_1, \sigma_1^2)$ | $\text{N}(nm_1, r\sigma^2 + r^2\sigma_1^2)$ | Afsn. 8.3.7 |
| $m_1 = \frac{m/\gamma + z}{1/\gamma + n}$ | $1/\sigma_1^2 = 1/\sigma_0^2 + n/\sigma^2$ |
| $S^2|\sigma^2 \in \chi^2(f)/f$ | $1/\sigma^2 \in G(\nu/2, 2/\beta_0^*)$ | $1/\sigma^2 \in G(\nu_1/2, 2/\beta_1^*)$ | $(S')^2 \in \text{RBet}(\nu_1/2, r/2, (\nu_1 - 2)\sigma_1^2/r)$ | Afsn. 8.3.8 |

Tabel 8.1. Aposteriorifordelinger for endimensionale eksponentielle familier med naturlige konjugerede strukturfordelinger

1. Linien for $\sigma^2\chi^2$-fordelingen angiver $(S')^2$ en empirisk varians med $r$ frihedsgrader. $\beta_0^* = (\nu - 2)\sigma_0^2$; $\beta_1^* = (\nu_1 - 2)\sigma_1^2$ med $\sigma_1^2$ bestemt ved (8.3.24).
Momenter i prædiktiv fordeling af $X' = \sum_{j=1}^{r} X'_j / r$ efter observation af $\bar{x} = \sum_{i=1}^{n} x_i / n$.

$$E [X' | \bar{x}] = m_1; \quad V [X' | \bar{x}] = E [V(\mu) | \bar{x}] \left( \gamma_1 + \frac{1}{r} \right)$$

Middelværdi, $m_1$ og varians $V[\mu | \bar{x}]$ i aposteriorifordelingen af $\mu$ efter observation af $\bar{x} = \sum_{i=1}^{n} x_i / n$.

$$m_1 = (m / \gamma + n \bar{x}) / (1 / \gamma + n); \quad \gamma_1 = 1 / (1 / \gamma + n); \quad E [\mu | \bar{x}] = m_1; \quad V [\mu | \bar{x}] = E [V(\mu) | \bar{x}] / (1 / \gamma + n)$$

Præposteriorivarians af $\mu$:

$$E_{x_1, \ldots, x_n} [V(\mu | \bar{x})] = E [E [V(\mu) | \bar{x}] / (1 / \gamma + n)]$$

| Stikprøvestruktur- | Stikprøve- | Struktur- | Præposteriori- | Præposteriori- | Reference |
| fordeling af $X_i | \theta$ | fordeling $w(\cdot)$ | $m_1 = E [\mu | \bar{x}]$ | $E [V(\mu) | \bar{x}]$ | middelværdi af $V(\mu)$ | |
| B$(1, p)$ | $p \in \text{Be}(\alpha, \beta)$ | $\pi_1 = \frac{\pi / \gamma + n \bar{x}}{1 / \gamma + n}$ | $\pi_1 (1 - \pi_1) \frac{1 / \gamma + n}{1 / \gamma + n + 1}$ | $\frac{\pi (1 - \pi)}{1 + \gamma}$ | Afsn 8.3.3 |
| Geo$(1, p)$ | $p \in \text{Be}(\alpha, \beta)$ | $\psi_1 = \frac{\psi / \gamma + n \bar{x}}{1 / \gamma + n}$ | $\psi_1 (1 + \psi_1) \frac{1 / \gamma + n}{1 / \gamma + n - 1}$ | $\frac{\psi (1 + \psi)}{1 - \gamma}$ | Afsn 8.3.4 |
| P$(\mu)$ | $\mu \in \text{G}(\alpha, 1/\beta)$ | $m_1 = \frac{m / \gamma + n \bar{x}}{1 / \gamma + n}$ | $m_1$ | $m_1$ | $m$ | Afsn 8.3.5 |
| Ex$(\mu)$ | $\mu \in \text{RGam}(\alpha, 1/\beta)$ | $m_1 = \frac{m / \gamma + n \bar{x}}{1 / \gamma + n}$ | $m_1^2 \frac{1 / \gamma + n}{1 / \gamma + n - 1}$ | $m_1^2 \frac{1}{1 - \gamma}$ | Afsn 8.3.6 |
| N$(\mu, \sigma^2)$ | N$(m, \sigma_0^2)$ | $m_1 = \frac{m / \gamma + n \bar{x}}{1 / \gamma + n}$ | $\sigma^2$ | $\sigma^2$ | Afsn 8.3.7 |

Tabel 8.2. Aposterriorimiddelværdier og gennemsnitlig aposteriorivarians for endimensionale eksponentielle familier med naturlige konjugerede strukturfordelinger.
8.3.9 Normalfordelingsmodeller med tilfældigt varierende varians:

Vi afslutter med en summarisk oversigt over sammenhængen mellem stikprøvefordeling, apriorifordeling, marginalfordeling og aposteriorifordeling for normalfordelingsmodeller med tilfældigt varierende varians, svarende til den model, der er betragtet i afsnit 5.7.

For givet gruppe antages $X_1, X_2, \ldots, X_n$ at være uafhængige $N(\mu, \sigma^2)$-fordelte, hvor $\sigma^2$ antages at variere fra gruppe til gruppe i overensstemmelse med en $RG\alpha(\alpha, \beta)$-fordeling. Parameteren $\mu$ varierer ligeledes fra gruppe til gruppe, og det antages, at for givet gruppevarians, $\sigma^2$, vælges $\mu$ i overensstemmelse med en $N(\mu_0, \sigma^2/m)$-fordeling.
<table>
<thead>
<tr>
<th>Betinget ford. af ((T, Z)) for givet ((\mu, \sigma^2))</th>
<th>Konjugeret apriorifordeling af ((\mu, \sigma^2))</th>
<th>Aposteriorifordeling af ((\mu, \sigma^2)) efter observation. af ((t, z))</th>
<th>Marginalfordeling af ((T, Z))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford af ((T</td>
<td>\mu, \sigma^2))</td>
<td>Ford. af ((\mu</td>
<td>\sigma^2))</td>
</tr>
<tr>
<td>(N(n\mu, n\sigma^2))</td>
<td>(N(\mu_0, \sigma^2/m))</td>
<td>(N(\mu_1, \sigma^2/(m + r)))</td>
<td>(N(n\mu_0, n\sigma^2 + n^2\sigma^2/m))</td>
</tr>
<tr>
<td>Ubet. ford. af (\mu)</td>
<td>(T\left(2\alpha, \mu_0, \sqrt{\frac{\beta}{m\alpha}}\right))</td>
<td>(T\left(2\alpha + n, \mu_1, \sqrt{\frac{\beta_1}{(2\alpha + n)(m + n)}}\right))</td>
<td>(T\left(2\alpha, n\mu_0, \sqrt{\frac{\beta(n^2 + mn)}{m\alpha}}\right))</td>
</tr>
<tr>
<td>Ford. af ((Z</td>
<td>\sigma^2))</td>
<td>Ubet. ford. af (\sigma^2)</td>
<td>Ubet. ford. af (\sigma^2)</td>
</tr>
<tr>
<td>(G((n - 1)/2, 2\sigma^2))</td>
<td>(RGam(\alpha, \beta))</td>
<td>(RGam(\alpha + n/2, \beta_1/2))</td>
<td>(RBet(\alpha, (n - 1)/2, 2\beta))</td>
</tr>
</tbody>
</table>

Tabel 8.3. Aposteriorifordelinger for normalfordelte observationer med varierende middelværdi og varians. Stikprøvestørrelse \(n\).

\[
T = \sum X_i; \quad \overline{X} = \frac{\sum X_i}{n}; \quad Z = \sum (X_i - \overline{X})^2.
\]
8.4 Filtrering af en tidsrække

Vi betræt modellen:

\[
\begin{align*}
\theta_t &= \alpha \theta_{t-1} + \delta_t \quad (8.4.1) \\
X_t &= \theta_t + \epsilon_t, \ t = 1, 2, 3, \ldots \quad (8.4.2)
\end{align*}
\]

hvor \(\delta_1, \delta_2, \ldots\) er indbyrdes uafhængige, \(\delta_t \in \mathcal{N}(0, \sigma_\delta^2)\) og \(\epsilon_1, \epsilon_2, \ldots\) ligeledes er indbyrdes uafhængige, \(\epsilon_t \in \mathcal{N}(0, \sigma^2)\), og hvor \(\epsilon_t\) og \(\delta_t\) er uafhængige, \(t = 1, 2, 3, \ldots\).

Modellen kan beskrive en proces, hvor processens tilstand, \(\theta\), varierer med tiden i overensstemmelse med en AR(1) proces. Tilstanden, \(\theta_t\), kan imidlertid ikke observeres direkte, men observationen, \(X_t\), er behæftet med målestøj, \(\sigma^2\).

Vi antager, at parametrene \(\alpha, \sigma^2\) og \(\sigma_\delta^2\) er kendte.

Sætter vi startværdien \(\theta_0 = 0\) har vi apriorifordelingen \(\theta_1 \in \mathcal{N}(0, \sigma_\delta^2)\) og stikprøvefordelingen

\[
X_1 \mid \theta_1 \in \mathcal{N}(\theta_1, \sigma^2)
\]

Vi får derfor, at aposteriorifordelingen af \(\theta_1\) er en normalfordeling med forventningsværdi og varians

\[
\begin{align*}
\theta_1^f &= \mathbb{E} [\theta \mid X_1 = x_1] = (1 - w_1)x \\
\mathbb{V} [\theta \mid X_1 = x_1] &= \sigma^2(1 - w_1),
\end{align*}
\]

hvor

\[
(1 - w_1) = \frac{\gamma_0}{1 + \gamma_0}
\]

med \(\gamma_0 = \sigma_\delta^2 / \sigma^2\).

Den betingede fordeling af \(\theta_2\), givet observationen \(X_1 = x_1\) er da en normalfordeling med forventningsværdi og varians

\[
\begin{align*}
\hat{\theta}_2^f &= \mathbb{E} [\theta_2 \mid X_1 = x_1] = \alpha \theta_1^f \\
\mathbb{V} [\theta_2 \mid X_1 = x_1] &= \alpha^2 \mathbb{V} [\theta_1 \mid X_1 = x_1] + \sigma_\delta^2 = \alpha^2 \sigma^2 (1 - w_1) + \sigma_\delta^2
\end{align*}
\]

Ved observation af \(X_2\) er det da denne fordeling af \(\theta_2 \mid X_1 = x_1\), der fungerer som apriorifordeling. Stikprøvefordelingen af \(X_2\) givet \(\theta_2\) og givet \(X_1 = x_1\)
er en \(N(\theta_2, \sigma^2)\)-fordeling, og vi kan derefter bestemme aposteriorifordelingen for \(\theta_2\) etc.

Ved observation af \(X_t\) har vi således at apriorifordelingen for \(\theta_t\) (nemlig den betingede fordeling af \(\theta_t\) givet \(X_1 = x_1, X_2 = x_2, \ldots, X_{t-1} = x_{t-1}\)) er en normalfordeling med forventningsværdi \(\theta_t^p\) og varians \(\gamma_t \sigma^2\) mens stikprøvefordelingen er en \(N(\theta_t, \sigma^2)\)-fordeling. Aposteriorifordelingen af \(\theta_t\) efter observation af \(X_t = x_t\) (den betingede fordeling af \(\theta_t\) givet \(X_1 = x_1, X_2 = x_2, \ldots, X_{t-1} = x_{t-1}\) samt \(X_t = x_t\)) er da en normalfordeling med forventningsværdi (den filtrerede værdi):

\[
\hat{\theta}_t^f = E[\theta | X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t] = \hat{\theta}_t^p + (1 - w_t)(x_t - \hat{\theta}_t^p)
\]

(8.4.3)

og med variansen

\[
V[\theta | X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t] = \sigma^2(1 - w_t)
\]

(8.4.4)

hvor

\[
w_t = 1/(1 + \gamma_t)
\]

(8.4.5)

Prædiktionen \(\hat{\theta}_t^p\) bestemmes ved

\[
\hat{\theta}_t^p = \alpha \hat{\theta}_t^f
\]

(8.4.6)

Parameteren \(\gamma_t\) opdateres ved

\[
\gamma_{t+1} = \alpha^2(1 - w_t) + \gamma_0
\]

(8.4.7)

Opdateringen (8.4.7) af parameteren \(\gamma_t\) følger ved at bemærke, at inden observation af \(X_t\) har man

\[
V[\theta_t | x_1, x_2, \ldots, x_{t-1}] = \sigma^2 \gamma_t
\]

Efter observation af \(X_t = x_t\) er aposteriorivariansen derfor

\[
V[\theta_t | x_1, x_2, \ldots, x_t] = \sigma^2 \frac{\gamma_t}{1 + \gamma_t} \sigma^2(1 - w_t)
\]

Variansen i fordelingen af \(\theta_{t+1}\) (der er apriorifordeling ved observation af \(X_{t+1}\)) bliver da

\[
V[\theta_{t+1} | x_1, x_2, \ldots, x_t] = \alpha^2 V[\theta_t | x_1, x_2, \ldots, x_t] + \sigma^2 = \sigma^2\{\alpha^2(1 - w_t) + \gamma_0\}
\]
Opdateringen ved (8.4.3) til (8.4.7) benævnes et Kalman-filter. I filtrerings-
sammenhænge benævnes størrelsen \((1 - w_t)\) ofte Kalman-forstærkningen.

Vi ser, at for \(t \to \infty\) vil størrelserne \(\gamma_t\) og \(w_t\) konvergere mod værdier \(\gamma^*\) og \(w^*\), der kun afhænger af de givne parametre \(\alpha\) og \(\gamma_0\). Dette indebærer, at for store værdier af \(t\) vil aposteriorimiddelværdien \(\hat{\theta}_t^i\) stort set være et eksponentielt vægtet gennemsnit af samtlige foregående observationer.

Vi bemærker endelig, at til forskel fra de øvrige problemer, vi har betragtet, vil et voksende antal observationer \((t \to \infty)\) ikke betyde en mere præcis bestemmelse af parameteren \(\theta_t\). Dette skyldes naturligvis, at til ethvert nyt observationstidspunkt er der også en ny parameter, der skal bestemmes, i modsætning til de øvrige situationer, vi har betragtet, hvor gruppeparameteren har været fast. Ønsker man en nøjagtigere bestemmelse af den aktuelle parameter, \(\theta_t\), i filtreringsproblemet, må man mindske observationsstøjnen, \(\sigma^2\), for eksempel ved at tage flere observationer til hvert tidspunkt, såfremt dette er muligt.

8.5 Den flerdimensionale normalfordeling

Vi indleder med at angive aposteriorifordelingen svarende til en enkelt ob-
ervation, idet udledningerne i dette tilfælde er lidt enklere, end i det ge-
nerelle tilfælde.

**Lemma 8.5.1** *Aposteriorifordeling svarende til en enkelt ob-
ervation*

Lad \(X | \mu \in N_p(\mu, \Sigma)\) og lad \(\mu \in N_p(m, \Sigma_0)\), hvor \(\Sigma\) og \(\Sigma_0\) har fuld rang, \(p\).

Da er aposteriorifordelingen af \(\mu\) efter observation af \(X = x\) givet ved

\[
\mu | X = x \in N_p(Wm + (I - W)x, (I - W)\Sigma) \tag{8.5.1}
\]

med

\[W = \Sigma(\Sigma_0 + \Sigma)^{-1}\quad \text{og}\quad I - W = \Sigma_0(\Sigma_0 + \Sigma)^{-1}\]
Bevis:
Vi har
\[ f(x \mid \mu) = (\sqrt{2\pi})^{-p} \det(\Sigma)^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]
\[ \propto \exp \left\{ -\frac{1}{2} (\mu^T \Sigma^{-1} \mu - 2\mu^T \Sigma^{-1} x) \right\} \]
og
\[ w(\mu) = (\sqrt{2\pi})^{-p} \det(\Sigma_0)^{-1/2} \exp \left\{ -\frac{1}{2} (\mu - m)^T \Sigma_0^{-1} (\mu - m) \right\} \]
\[ \propto \exp \left\{ -\frac{1}{2} (\mu^T \Sigma_0^{-1} \mu - 2\mu^T \Sigma_0^{-1} m) \right\} \]

hvorfor
\[ h(\mu \mid x) \propto \exp \left[ -\frac{1}{2} \{\mu^T (\Sigma^{-1} + \Sigma_0^{-1}) \mu - 2\mu^T (\Sigma^{-1} x + \Sigma_0^{-1} m)\} \right] \]
\[ \propto \exp \left\{ -\frac{1}{2} (\mu - m_1)^T \Sigma_1^{-1} (\mu - m_1) \right\} \]

Sætter vi nu \( W = \Sigma_1 \Sigma_0^{-1} \), ser vi, at der gælder
\[ W = (\Sigma^{-1} + \Sigma_0^{-1})^{-1} \Sigma_0^{-1} = (\Sigma_0^{-1} \Sigma_0 \Sigma^{-1} + \Sigma_0^{-1})^{-1} \Sigma_0^{-1} \]
\[ = (\Sigma_0 \Sigma^{-1} + I)^{-1} \Sigma_0 \Sigma_0^{-1} = (\Sigma_0 \Sigma^{-1} + \Sigma \Sigma^{-1})^{-1} \]
\[ = \Sigma (\Sigma_0 + \Sigma)^{-1} \]

Endvidere bemærker vi, at der gælder
\[ \Sigma_1 \Sigma_0^{-1} + \Sigma_1 \Sigma^{-1} = \Sigma_1 (\Sigma_0^{-1} + \Sigma^{-1}) = (\Sigma_0^{-1} + \Sigma^{-1})^{-1} (\Sigma_0^{-1} + \Sigma^{-1}) = I \]
således at koefficientmatricen til \( x \) er
\[ \Sigma_1 \Sigma^{-1} = I - \Sigma_1 \Sigma_0^{-1} = I - W \]

Ved multiplikation til højre med \( \Sigma \) finder vi da aposteriorivariansten \( \Sigma_1 \) som
\[ \Sigma_1 = (I - W) \Sigma \]
(8.5.2)
Udtrykket for $I - W$ findes ved at bemærke, at

$$I - W = (\Sigma_0 + \Sigma)(\Sigma_0 + \Sigma)^{-1} - \Sigma(\Sigma_0 + \Sigma)^{-1}\Sigma_0(\Sigma_0 + \Sigma)^{-1}$$

**Bemærkning 1 : Vægtmatricerne udtrykt ved den generaliserede varianskvotient**

Lader vi $\Gamma = \Sigma_0 \Sigma^{-1}$ betegne den generaliserede kvotient mellem variansen mellem grupper og variansen inden for grupper, kan vi udtrykke matricerne $W$ og $I - W$ ved

$$W = (I + \Gamma)^{-1} \quad \text{og} \quad I - W = (I + \Gamma)^{-1}\Gamma$$

Man har nemlig

$$W = \Sigma(\Sigma_0 + \Sigma)^{-1} = \Sigma(\Sigma_0 \Sigma^{-1}\Sigma + \Sigma)^{-1} = \Sigma\Sigma^{-1}(I + \Gamma)^{-1} = (I + \Gamma)^{-1}$$

Udtrykket gælder også, selv om $\Sigma$ ikke har fuld rang.

Vi bemærker specielt, at $(I + \Sigma_0 \Sigma)$ er invertibel, også selv om $\Sigma_0$ ikke har fuld rang. Der gælder nemlig for vilkårlige matricer $A$ og $B$, at $\det(I + AB) = \det(I + BA)$ (Zellner 1971, p.231).

Sættes $A = \Sigma_0$ og $B = \Sigma^{-1}$ har vi da, at

$$\det(I + \Sigma_0 \Sigma^{-1}) = \det(\Sigma^{-1}\Sigma_0 + I) = \det(\Sigma^{-1})\det(\Sigma_0 + \Sigma)$$

Da nu $\Sigma$ er positiv definit, er også $\Sigma_0 + \Sigma$ samt $\Sigma^{-1}$ positiv definit, hvorfor begge determinanter på højre side er positive.

**Sætning 8.5.1** *Aposteriorifordeling efter observation af et gennemsnit*

Lad $X_1, X_2, \ldots, X_n$ angive et sæt variable for hvilke det gælder, at for givet $\mu$ er $X_1, X_2, \ldots, X_n$ uafhængige og identisk fordelte med $X_i \in N_p(\mu, \Sigma)$, og antag endvidere, at $\mu \in N_p(m, \Sigma_0)$
Da er aposteriorifordelingen af $\mu$ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ givet ved

$$
\mu \mid x_1, x_2, \ldots, x_n \in N_p(Wm + (I - W)x_+, \frac{1}{n}(I - W)\Sigma)
$$

hvor

$$
W = \Sigma(n\Sigma_0 + \Sigma)^{-1} = (I + n\Gamma)^{-1}
$$

$$
I - W = n\Sigma_0(n\Sigma_0 + \Sigma)^{-1} = n(I + n\Gamma)^{-1}\Gamma
$$

$$
\Gamma = \Sigma_0\Sigma^{-1}
$$

og

$$
x_+ = \frac{1}{n}\sum_{i=1}^{n} x_i
$$

Bevis:

Resultatet følger af ovenstående lemma ved at bemærke, at $X_+$ er sufficient for $\theta$, og

$$
\overline{X}_+ \mid \mu \in N_p(\mu, \frac{1}{n}\Sigma)
$$

\[\square\]

Eksempel 8.5.1 *Målefejl for flowmålere*

Vi betragter atter den situation, der blev behandlet i eksempel 6.7.1.

Antag, at fejlvurderingen for en måler karakteriseres ved fejlvurderingen $\mu = (\mu_1, \mu_2)^T$ ved de to flow, 0.1 [m$^3$/h] og 0.5 [m$^3$/h].

Antag, at fordelingen af fejlvurderingen for målerne i en målerpopulation kan beskrives ved en todimensionel normalfordeling, $\mu \in N_2(m, \Sigma_0)$, hvor

$$
m = \begin{pmatrix} 2.0 \\ 3.0 \end{pmatrix}; \quad \text{og} \quad \Sigma_0 = \begin{pmatrix} 3 & 5 \\ 5 & 10 \end{pmatrix}
$$

Antag endelig, at usikkerheden (kalibreringsfejlen) ved kalibrering af en måler ved de to flow kan beskrives ved en todimensionel normalfordeling, hvis middelværdi er målerens sande fejlvurdering, og med dispersionsmatricen

$$
\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}
$$
og antag at kalibreringen foretages på en sådan måde, at fejlene ved gentagne kalibreringer er uafhængige.

Figur 8.1 viser apriorifordelingen af målernes fejlvisning, og figur 8.2 viser fordelingen af kalibreringsfejlen for en måler med fejlvisning \((\mu_1, \mu_2) = (0, 0)\).

En måler blev sendt til kalibrering, og der blev foretaget \(n = 2\) bestemmelser af målerens fejlvisning. De to sæt kalibreringsresultater er anført nedenfor

\[
x_1 = \begin{pmatrix} 5.0 \\ 7.0 \end{pmatrix}; \quad x_2 = \begin{pmatrix} 1.0 \\ 5.0 \end{pmatrix}
\]

Gennemsnittet af de to kalibreringer er således

\[
\bar{x}_+ = \begin{pmatrix} 3.0 \\ 6.0 \end{pmatrix}
\]
Figur 8.1. Niveauproduktør i fordelingen af samhørende værdier af målerfejl ved to flow for population af målere (apriorifordeling)
Samhørende værdier af kalibreringsfejl ved to flowfordeling af fejl ved en kalibrering (n=1)

Σ = [4.0 0.0; 0.0 1.0]

---

Figur 8.2. Niveaukurver i fordelingen af samhørende værdier af kalibreringsfejl ved to flow for kalibrering af målere (stikprøvefordeling)
Vi antager nu, at man ønsker at inddrage kendskabet til variationen af fejlsvisningen i målerpopulationen ved fastlæggelsen af fejlsvisningen for den betragtede måler.

Man kan da bestemme aposteriorifordelingen for målerens fejlsvisning \( \mu \).

Man finder

\[
\Gamma = \Sigma_0 \Sigma^{-1} = \begin{pmatrix} 3 & 5 \\ 5 & 10 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}^{-1}
\]

\[
= \begin{pmatrix} 0.75 & 5.00 \\ 1.25 & 10.00 \end{pmatrix},
\]

hvorfor

\[
W = (I + 2\Gamma)^{-1} = \begin{pmatrix} 2.5 & 10.0 \\ 2.5 & 21.0 \end{pmatrix}^{-1}
\]

\[
= \begin{pmatrix} 0.7636 & -0.3636 \\ -0.0909 & 0.0909 \end{pmatrix}
\]

og

\[
I - W = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0.7636 & -0.3636 \\ -0.0909 & 0.0909 \end{pmatrix}
\]

\[
= \begin{pmatrix} 0.2364 & 0.3636 \\ 0.0909 & 0.9091 \end{pmatrix}.
\]

Aposteriorimiddelværdien for målerens fejlsvisning er derfor

\[
m_{\text{apost}} = W \begin{pmatrix} 2.0 \\ 3.0 \end{pmatrix} + (I - W) \begin{pmatrix} 3.0 \\ 6.0 \end{pmatrix}
\]

\[
= \begin{pmatrix} 0.7636 & -0.3536 \\ -0.0909 & 0.0909 \end{pmatrix} \begin{pmatrix} 2.0 \\ 3.0 \end{pmatrix} + \begin{pmatrix} 0.2364 & 0.3636 \\ 0.0909 & 0.9091 \end{pmatrix} \begin{pmatrix} 3.0 \\ 6.0 \end{pmatrix}
\]

\[
= \begin{pmatrix} 0.4364 \\ 0.0909 \end{pmatrix} + \begin{pmatrix} 2.8909 \\ 5.7273 \end{pmatrix} = \begin{pmatrix} 3.3273 \\ 5.8182 \end{pmatrix}
\]
8.5 Flerdimensional normalfordeling

Vi bemærker, at selv om aposteriorimiddelværdien kan opfattes som et vejet gennemsnit mellem apriorimiddelværdien og stikprøvevægennemsnittet, ligger den resulterende værdi ikke nødvendigvis på liniestykket, der forbinder apriorimiddelværdi og stikprøveresultat. Det vejede gennemsnit tilgodeser også samvariationen mellem fejlene ved de to flow. (Principielt også mellem kalibreringsfejlene ved de to flow, men denne samvariation er nul i den her betragtede situation).

Sammenvejningen af stikprøveresultat og apriorimiddelværdi er således mere kompliceret end for endimensionale fordelinger.

Selv om man også her kan fortolke aposteriorimiddelværdien som en lineær prædiktor (jvf bemærkning 5 til sætning 8.3.1), skal samvariationen mellem stikprøveresultat og prædiktion her udtrykkes ved kovarianser mellem par i en 4-dimensional fordeling.

I det aktuelle tilfælde bemærker vi først, at da kalibreringsusikkerheden ved flow 2 er væsentligt mindre end apriorivariansen svarende til dette flow, ligger aposteriorimiddelværdien svarende til flow 2 ganske tæt ved stikprøveresultatet. Den stærke samvariation mellem fejlene ved flow 1 og flow 2 i apriorifordelingen (korrelationen \( \rho = 0.91 \)) indebærer da, at aposteriorimiddelværdien af fejlen ved flow 1 skal være i overensstemmelse med denne værdi. Kalibreringsusikkerheden ved flow 1 er imidlertid så rummelig, at den tillader en vis afvigelse mellem stikprøvevægennemsnit og aposteriorimiddelværdi. Den resulterende aposteriorimiddelværdi af fejlen ved flow 1 bliver tilmed større end såvel stikprøvevægennemsnit som apriorimiddelværdi.

Såfremt vi ønsker et udtryk for usikkerheden på denne angivelse af målerens fejlvisning, kan vi benytte variansen i apriorifordelingen af \( \mu \). Idet \( n = 2 \) har vi

\[
D[\mu|\bar{x}] = \frac{1}{2} (I - W) \Sigma = \frac{1}{2} \begin{pmatrix} 0.2364 & 0.3636 \\ 0.0909 & 0.9091 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}
\]

\[
= \frac{1}{2} \begin{pmatrix} 0.9455 & 0.3636 \\ 0.3636 & 0.9091 \end{pmatrix} = \begin{pmatrix} 0.4727 & 0.1818 \\ 0.1818 & 0.4545 \end{pmatrix}
\]

Ved sammenligning med \( \Sigma_0 \) bemærker vi specielt, at på grund af den store stikprøvepræcision ved flow 2, har vi opnået en kraftig reduktion af usikkerheden vedrørende fejlen ved flow 2 i forhold til aprioriusikkerheden.
Ved sammenligning med usikkerheden $0.5 \Sigma$ på stikprøvegennemsnittet ser vi specielt, at vi har opnået en væsentlig reduktion af usikkerheden vedrørende fejlen ved flow 1 i forhold til variansen på stikprøvegennemsnittet.

Aposteriorifordelingen for målerens fejlvisning er illustreret i figur 8.3

Endelig viser figur 8.4 marginalfordelingen af kalibreringsresultatet ved en enkelt kalibrering af en tilfældig udvalgt måler, og figur 8.5 viser den prædiktive fordeling svarende til eventuelle andre kalibreringer af den udvalgte måler. Fordelingen i figur 8.4 illustrerer den prædiktion, man kan foretage om en kalibreringsresultater for en tilfældigt udvalgt måler før man har kalibreret den, mens fordelingen i figur 8.5 illustrerer den tilsvarende prædiktion efter at man har foretaget den beskrevne kalibrering af måleren.
Samhørende værdier af fejl ved to flow 
aposteriorisfordeling af måleverdier efter obs af (3,6) 
snit af 2 obs, $\Sigma_1 = \begin{bmatrix} 0.4727 & 0.1818; & 0.1818 & 0.4545 \end{bmatrix}$

Figur 8.3. Niveaukurver for posteriorisfordelingen af samhørende værdier af måleverl ved to flow for en måler med kalibreringsresultat (3.0;6.0) (snit af to kalibreringer).
Figur 8.4. Niveaukurver i fordelingen af samhørende værdier af kalibreringsresultat ved to flow for population af målere (marginalfordeling svarende til apriorifordelingen)
Figur 8.5. Niveaukurver i fordelingen af samhørende værdier af kalibreringsresultat ved to flow for den undersågte måler (prædiktiv fordeling) efter obs. af fejlene (3;6)
8.6 Regressionsmodeller

Sætning 8.6.1 Aposteriorifordeling i regressionsmodel for normalfordelte observationer

Lad $Y$ angive en $n \times 1$ dimensional vektor af observationer, og lad $X$ angive en $n \times p$ dimensional matrix af kendte koefficienter. Antag at $Y \mid \beta \in N_n(X\beta, \sigma^2V)$ og at apriorifordelingen af $\beta$ er

$$
\beta \in N_p(\beta_0, \sigma^2\Lambda)
$$

hvor $\Lambda$ har fuld rang.

Da er aposteriorifordelingen af $\beta$ efter observation af $Y = y$ givet ved

$$
\beta \mid Y = y \in N_p(\beta_1, \sigma^2\Lambda_1)
$$

hvor

$$
\beta_1 = W\beta_0 + W\Lambda X^TV^{-1}y
$$

med

$$
W = (I + \Gamma)^{-1}
$$

og

$$
\Gamma = \Lambda X^TV^{-1}X
$$

og hvor

$$
\Lambda_1 = (I + \Gamma)^{-1}\Lambda = WA
$$

Bevis:

Vi har

$$
f(y \mid \beta) = (\sqrt{2\pi\sigma^2})^{-n} \det(V)^{-1/2} \exp \left\{- \frac{1}{2\sigma^2} (y - X\beta)^T V^{-1} (y - X\beta) \right\}$$

$$
\propto \exp \left\{- \frac{1}{2\sigma^2} (\beta^T X^TV^{-1}X\beta - 2\beta^T X^TV^{-1}y) \right\}
$$
og

\[ w(\beta) = (\sqrt{2\pi\sigma^2})^{-p} \text{det}(\Lambda)^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (\beta - \beta_0)^T \Lambda^{-1} (\beta - \beta_0) \right\} \]

\[ \propto \exp \left\{ -\frac{1}{2\sigma^2} (\beta^T \Lambda^{-1} \beta - 2\beta^T \Lambda^{-1} \beta_0) \right\} \]

hvorfor

\[ h(\beta \mid y) \propto \exp \left\{ -\frac{1}{2\sigma^2} [\beta^T (X^T V^{-1} X + \Lambda^{-1}) \beta - 2\beta^T (X^T V^{-1} y + \Lambda^{-1} \beta_0)] \right\} \]

\[ \propto \exp \left\{ -\frac{1}{2\sigma^2} (\beta - \beta_1)^T \Lambda_1^{-1} (\beta - \beta_1) \right\} \]

hvor

\[ \Lambda_1 = (X^T V^{-1} X + \Lambda^{-1})^{-1} \]

og

\[ \beta_1 = \Lambda_1 \Lambda^{-1} \beta_0 + \Lambda_1 X^T V^{-1} y \]

Da \( \Lambda \) har fuld rang, gælder der

\[ \Lambda_1 = (\Lambda^{-1} \Lambda X^T V^{-1} X + \Lambda^{-1})^{-1} = (\Lambda X^T V^{-1} X + I)^{-1} \Lambda \]

\[ = (I + \Gamma)^{-1} \Lambda \]

med \( \Gamma = \Lambda X^T V^{-1} X \),

dvs vi har

\[ \beta_1 = (I + \Gamma)^{-1} \beta_0 + (I + \Gamma)^{-1} \Lambda X^T V^{-1} y \]

\[ \square \]

**Bemærkning 1: Invertering af de indgående matricer**

Såfremt \( n > p \) er det ofte lettere at benytte

\[ \beta_1 = \beta_0 + \Lambda X^T (X \Lambda X^T + V)^{-1} (y - X \beta_0) \]

\[ \Lambda_1 = \Lambda - \Lambda X^T (X \Lambda X^T + V)^{-1} \Lambda X \]

idet disse udtryk kun kræver inversion af \( p \times p \) matricer.

\[ \square \]
Bemærkning 2: Aposterioriværdien udtrykt som et vægtet gennemsnit

Såfremt $X$ har fuld rang, kan $X^T V^{-1} X$ inverteres, og vi får

$$
\beta_1 = W\beta_0 + (I - W)\hat{\beta}
$$

hvor $\hat{\beta}$ angiver den sædvanlige mindste kvadraters estimator (7.2.9) for $\beta$,

$$
\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y
$$

Bevis:

Vi betragter udtrykket

$$
\beta_1 = (I + \Gamma)^{-1} \beta_0 + (I + \Gamma)^{-1} \Lambda X^T V^{-1} y
$$

$$
= W\beta_0 + W\Lambda X^T V^{-1} y
$$

$$
= W\beta_0 + W\Lambda (X^T V^{-1} X)(X^T V^{-1} X)^{-1} X^T V^{-1} y
$$

$$
= W\beta_0 + (X^T V^{-1} X) W\Lambda \hat{\beta}
$$

men idet

$$
I - W = I - (I + \Gamma)^{-1} = (I + \Gamma)^{-1} \Gamma = (\Lambda X^T V^{-1} X + I)^{-1} \Lambda X^T V^{-1} X
$$

ser vi, at vi har

$$
W\Lambda (X^T V^{-1} X) = I - W
$$

Vi bemærker i øvrigt, at $\Gamma$ netop udtrykker $E [D [\hat{\beta} | \beta]] [D [E [\hat{\beta} | \beta]]]^{-1}$, og at

$$
I + \Gamma = \text{COV} [\hat{\beta}, \beta] D [\beta]^{-1}
$$

Sætning 8.6.2 Prædiktiv fordeling i regressionsmodel for normalfordelte observationer

Lad situationen være som i sætning 8.6.1, og antag, at der udover de $n$ observationer $Y_1, \ldots, Y_n$ foretages $\tau$ yderligere observationer $Y'_1, \ldots, Y'_{\tau}$ sådan
at sættet \( Y_1, \ldots, Y_n \) og sættet \( Y'_1, \ldots, Y'_r \) er betinget uafhængige af \( \beta \) i den simultane fordeling af \( Y_1, \ldots, Y_n, Y'_1, \ldots, Y'_r \) og \( \beta \).

Antag, at \( Y'|\beta \in N_r (X'|\beta, \sigma^2 V') \), hvor den \( r \times p \) dimensionale modeilmatrix \( X' \) er kendt, og hvor den \( r \times r \)-dimensionale symmetriske, positiv definitte matrix \( V' \) ligeledes er kendt.

Da vil fordelingen af \( Y'|Y \) (den prædiktive fordeling) være en \( r \)-dimensional normalfordeling med forventningsværdi

\[
E [Y'|Y] = X' \beta_1
\]  
(8.6.3)

og dispersionsmatrix

\[
D [Y'|Y] = \sigma^2 [V' + X'\Lambda_1 (X')^T]
\]  
(8.6.4)

hvor \( \beta_1 \) og \( \Lambda_1 \) er givet ved (8.6.1) og (8.6.2)

**Bevis:**
Følger ved at bemærke, at den prædiktive fordeling er en normalfordeling, og at momenterne bestemmes ved

\[
E [Y'|Y] = E \beta [E [Y|\beta]|Y] \Rightarrow E \beta|Y [X'|\beta] = X'/\beta_1
\]

og

\[
= E \beta|Y [\sigma^2 V'] + D \beta|Y [X'|\beta]
= \sigma^2 V' + X' D \beta|Y[\beta](X')^T
\]

**Eksempel 8.6.1 Ramushøjder**
Vi betragter atter den situation, der blev betragtet i eksempel 7.1.1.

Antag, at væksten af ramushøjderne i den betragtede population af 8-10 årige drenge kan beskrives ved følgende model:

En drengs ramushøjde i 8-10 års alderen udvikler sig som

\[
Y = \beta_1 + \beta_2 (x - 8.75) + \epsilon
\]
hvor $x$ angiver alderen målt i år og $\beta_1$ og $\beta_2$ er fordelt i populationen i
overensstemmelse med en $N(\beta_0, \sigma^2 \Lambda)$-fordeling hvor

$$\beta_0 = \begin{pmatrix} 52.7 \\ 1.4 \end{pmatrix}$$

og

$$\Lambda = \begin{pmatrix} 8 & 4 \\ 4 & 6 \end{pmatrix}$$

og hvor $\epsilon \in N(0, \sigma^2)$ med $\sigma^2 = 0.32$ [mm]$^2$ og $\epsilon$-størrelserne er indbyrdes
uafhængige såfremt tidsafstandene er et halvt år eller mere.

Vi betragter først en række udsagn baseret alene på denne apriorifordeling:

Ramushøjden $Y_{0.25}$ ved 9-års alderen for en tilfældigt udvalgt dreng kan
beskrives ved en normalfordelt variabel med middelværdi

$$E[Y_{0.25}] = 52.7 + 0.25 \times 1.4 = 52.7 + 0.35 = 53.05 \text{ [mm]}$$

og varians

$$V[Y] = E[V[Y|\beta]] + V[E[Y|\beta]]$$

$$= \sigma^2 (1 + x\Lambda x^T)$$

$$= \sigma^2 (1 + 10.3750) = 1.0238 = (1.0118)^2 \text{ [mm]}^2$$

idet vi har benyttet at

$$D[\delta \beta] = x \ D[\beta] x^T$$

hvor modelmatricen

$$x = (1 \ 0.25)$$

og $D[\beta] = \sigma^2 \Lambda$.

Ramushøjderne

$$Y = \begin{pmatrix} Y_{0.25} \\ Y_{0.75} \end{pmatrix}$$

ved alderen 9 år og 9 1/2 år for en tilfældigt udvalgt dreng kan beskrives
ved en todimensional normalfordeling med forventningsværdi

$$E[Y] = X\beta_0 = \begin{pmatrix} 52.7 + 0.35 \\ 52.7 + 1.05 \end{pmatrix} = \begin{pmatrix} 53.05 \\ 53.75 \end{pmatrix}$$
hvor modellmatricen $X$ er

$$X = \begin{pmatrix} 1 & 0.25 \\ 1 & 0.75 \end{pmatrix}$$

Dispersionsmatricen for $Y$ er

$$\mathbf{D}[Y] = \sigma^2(I_2 + X\Lambda X^T)$$

$$= \sigma^2 \begin{pmatrix} 11.375 & 13.125 \\ 13.125 & 18.375 \end{pmatrix}$$

$$= \begin{pmatrix} 1.0238 & 1.1813 \\ 1.1813 & 1.6538 \end{pmatrix}$$

Endelig finder man for tilvæksten $T$ fra alderen 9 år til 9 1/2 år:

$$T = Y_{0.75} - Y_{0.25} = (-1 \ 1)Y$$

at

$$E[T] = (-1 \ 1)X\beta_0 = 0.50\beta_2 = 0.70 \text{ [mm]}$$

og

$$V[T] = (-1 \ 1)\mathbf{D}[Y](-1 \ 1)^T$$

$$= \sigma^2 \times 1.5 = 0.135 = (0.367)^2 \text{ [mm]}^2$$

Man udvælger nu tilfældigt en dreng, som er 8 1/2 år gammel.

Hans aktuelle ramushøjde er 53.20 [mm] og ramushøjden på hans 8 års fødselsdag var 52.60 [mm].

Man har således observationen

$$Y = \begin{pmatrix} 52.6 \\ 53.2 \end{pmatrix}$$

svarende til modellmatricen

$$X = \begin{pmatrix} 1 & -0.75 \\ 1 & -0.25 \end{pmatrix}$$

Idet

$$X^T X = \begin{pmatrix} 2 & -1 \\ -1 & 0.6250 \end{pmatrix}$$
har man estimatet \( \hat{\beta} \) for regressionskoefficienterne for denne dreng:

\[
\hat{\beta} = (X^T X)^{-1} X^T Y
\]

\[
= \begin{pmatrix} 2 & -1 \\ -1 & 0.6250 \end{pmatrix}^{-1} \begin{pmatrix} 105.0 \ \\ -52.75 \end{pmatrix} = \begin{pmatrix} 2.5 & 4.00 \\ 4.00 & 8.00 \end{pmatrix}
\]

\[
= \begin{pmatrix} 53.50 \\ 1.20 \end{pmatrix}
\]

Endvidere har man

\[
\Gamma = \Lambda X^T X = \begin{pmatrix} 12 & -5.50 \\ 2 & -0.25 \end{pmatrix}
\]

og

\[
W = (I_2 + \Gamma)^{-1}
\]

\[
= \begin{pmatrix} 13 & -5.50 \\ 2 & 0.75 \end{pmatrix}^{-1} = \begin{pmatrix} 0.0361 & 0.2651 \\ -0.0964 & 0.6265 \end{pmatrix}
\]

således at dispersionsmatricen for aposteriorifordelingen af \( \beta \) er

\[
\sigma^2 \Lambda_1 = \sigma^2 W \Lambda = \sigma^2 \begin{pmatrix} 1.3494 & 1.7349 \\ 1.7349 & 3.3733 \end{pmatrix}
\]

Man får nu aposteriorimiddelværdien af \( \beta \)

\[
\beta_1 = W \beta_0 + (I - W) \hat{\beta}
\]

\[
= \begin{pmatrix} 2.2759 \\ -4.2024 \end{pmatrix} + \begin{pmatrix} 51.2482 \\ 5.6048 \end{pmatrix} = \begin{pmatrix} 53.5241 \\ 1.4024 \end{pmatrix}
\]

Prædiktionen (den prædiktive fordeling)

\[
Y' = \begin{pmatrix} Y_{0.25}' \\ Y_{0.75}' \end{pmatrix}
\]

for ramushøjderne ved alderen 9 år og 9 1/2 år for denne dreng har da forventningsværdien

\[
E [Y'|Y] = X \beta_1 = \begin{pmatrix} 53.5241 + 0.3506 \\ 53.5241 + 1.0518 \end{pmatrix} = \begin{pmatrix} 53.8747 \\ 54.5759 \end{pmatrix}
\]
8.7 Tidsrække modeller

Indet modellmatricen \( \mathbf{X} \) er

\[
\mathbf{X} = \begin{pmatrix}
1 & 0.25 \\
1 & 0.75
\end{pmatrix}
\]

Dispersionsmatricen for \( Y' \) er

\[
\mathbf{D} [Y'|Y] = \sigma^2 (\mathbf{I}_2 + \mathbf{X} \Lambda_1 \mathbf{X}^T) = \sigma^2 \begin{pmatrix}
3.4277 & 3.7169 \\
3.7169 & 6.8494
\end{pmatrix} = \begin{pmatrix}
0.3085 & 0.3345 \\
0.3345 & 0.6164
\end{pmatrix}
\]

Man finder derfor specielt, at prædiktionen af ramushøjden ved 9-års alderen har forventningsværdien

\[
\mathbb{E} [Y_{0.25}'|Y] = 53.8747 \text{ [mm]}
\]

og variansen

\[
\mathbb{V} [Y_{0.25}'|Y] = 0.3085 = (0.555)^2 \text{ [mm]^2}
\]

Endvidere har den prædiktive værdi \( T' \) for for tilvæksten fra alderen 9 år til 9 1/2 år forventningsværdien

\[
\mathbb{E} [T'|Y] = (-1 \quad 1) \mathbf{X} \beta_1 = 0.50 \times 1.4024 = 0.7012 \text{ [mm]}
\]

og variansen

\[
\mathbb{V} [T'|Y] = (-1 \quad 1) \mathbf{D} [Y'|Y] (-1 \quad 1)^T = \sigma^2 \times 0.8434 = 0.0759 = (0.2755)^2 \text{ [mm]^2}
\]

Da aposteriorimiddelværdien af hældningen ikke afviger væsentligt fra apriorimiddelværdien er prædiktionen af tilvæksten ikke væsentligt anderledes efter observation af højderne ved 8 og 8 1/2 år.

\[\square\]

8.7 Tidsrække modeller

Bayesestimationen for parametrene i den endimensionale autoregressive proces af første orden under den tilfældige model fremgår af
Sætning 8.7.1 Bayesestimation for endimensional autoregressiv tidsrække

Lad

\[ Y_t = \beta Y_{t-1} + \epsilon_t, \ t = 1, 2, \ldots n \]

hvor \( \beta \in N(\beta_0, \sigma_0^2) \) og \( \epsilon_t \in N(0, \sigma^2) \) er indbyrdes uafhængige. Da er aposteriorifordelingen af \( \beta \) efter observation af \( y = (y_0, y_1, \ldots, y_n)^T \) en normalfordeling med

\[
E [\beta \mid y] = w\beta_0 + (1 - w)\hat{\beta} \\
V [\beta \mid y] = 1 - w
\]

hvor

\[
w = \frac{1}{1 + (y_{-1}^T y_{-1}) \gamma}
\]

med \( \gamma = \sigma_0^2 / \sigma^2 \) og

\[ \hat{\beta} = (y_{-1}^T y_{-1})^{-1} y_{-1}^T y
\]

Bevis:
Resultatet vises ved at betragte likelihoodfunktionen for \( \beta \).

Bayesløsningen for den generelle flerdimensionale tidsrække-model, der er beskrevet i afsnit 7.3.2 er givet ved:

Sætning 8.7.2 Bayesløsning for den flerdimensionale tidsrække-model

For modellen givet ved (7.3.8) og (7.3.9) med \( \beta \in N_{(p+q) \times q}(\beta_0, \Sigma_0) \) og \( \epsilon \in N_{(p+q) \times q}(0, \Sigma) \) gælder

\[
E [\beta \mid y] = W\beta_0 + (I - W)\hat{\beta}
\]
hvor den $(p + q)p \times (p + q)p$ dimensionale matrix $W$ er bestemt ved a): for $\Sigma = \sigma^2 V$:

\[
W = (\Sigma^{-1}_0 + \sigma^{-2} X^T V^{-1} X)^{-1} \Sigma^{-1}_0 \quad (8.7.2)
\]

med $X = I_p \otimes M$, $M = (Y_{-1}, Z)$ og $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$

b) for $\Sigma$ af formen $\Sigma \otimes V$

\[
W = (\Sigma^{-1}_0 + \Sigma^{-1} M^T V^{-1} M)^{-1} \Sigma^{-1}_0 \quad (8.7.3)
\]

med

\[
M = (Y_{-1}, Z)
\]

og

\[
\hat{\beta} = \text{vec}(\hat{B})
\]

hvor

\[
\hat{B} = (M^T V^{-1} M)^{-1} M^T V^{-1} Y
\]

**Bevis:**
Indeks

\( \Gamma \), signal/støj forhold, 612
\( \gamma \), signal/støj forhold, 497, 554, 573, 581, 591
\( n_0 \), vægtet gennemsnitlig gruppestørrelse, 488
2 \times 2-tabeller
marginal symmetri, 412
Sammenligning af hyppigheder, konfidentsinterval for differens, 392
Sammenligning af hyppigheder, konfidentsinterval for odds ratio, 394
aliasing mellem parametre, 269
alternativ variation, 7, 30, 32
analyseenhed, 5
alternativt varierende, 7, 25, 30, 32
aposteriorifordeling, 679
for empiriske varianser fra normalfordelte obs, 707
ved binomial-beta fordeling, 691, 709
ved gamma-reciprok gamma fordeling, 703, 709
ved negativ binomial-beta fordeling, 699, 709
ved normal-normal fordeling, 705, 709
ved normalfordeling med tilfældig varians, 711
ved Poisson-gammafordeling, 700, 701, 709
apriorifordeling, 549, 679
konjugeret, 551
arbejdsresidual, 203
arbejdsrespons, working response, 203
B(1, p)-fordeling, 122, 125
B(n, p) fordeling, 132, 138
Bartlett’s test, 241
Bartlett-korrektionen, 241
baseline
logit, 439
odds, 439
odds ratio, 439
Bayes, Thomas, 679
Bernoullifordeling, 122, 125
betinget uafhængige variable, 467
binomialfordeling, 132, 138
linkfunktioner, 153
blandede modeller, 675
Bradley-Terry model, 419
bulk sampling, 530
buskunder, 64
case-control studier, 403
collapse over en variabel, 476
conjoint analyse, 442
Cook’s $D$, 343
cutpoint, 448
devians
  skaleret, 146
devians for naturlig eksponentiel familie, 126
devians mellem observationer og model, 146
devians, momenter for, 143
deviansanalyse, 290
  proc INSIGHT, 292
deviansresidual, 200
  studentiseret, 212
Dfbetas, 344
Dffits, 344
differentiel effekt, 312
dimension
  af antalstabel, 427
dimension af generaliseret lineær model, 165
discrete choice models, 441
diskret valg
  modeller for, 441
dispersionsparameter, 133
  estimation, 222
  estimation under successiv testning, 311
  maksimum likelihood estimat, 225
effektiv stikprøvestørrelse
  Poisson-gamma fordeling, 583
  ved binomial-beta fordeling, 561
eksponentiel dispersionsmodel
  additiv, 132
  indeksparameter, 132
  kanonisk parameter, 132
  middelværdiafbildning, 136
reproduktiv, 133
eksponentiel dispersionsmodel, enhedsdevians, 140
eksponentiel familie
devians, 126
  middelværdiparametrerisering, 124
naturlig, 122
eksponentiel familie, middelværdiafbildning, 123
empiriske varianser for normalfordelte obs.
estimation
  marginal fordeling, 605
tilfældig model, 600
empiriske varianser for normalfordelte
  posteriorifordelning, 707
empiriske varianser fra normalfordelte obs., 134, 599
endelig population
  indeksmængde, 5
  korrektionsfaktor, 24
  målgruppe, 17
  stikprøve fra, 17
  stikprøveramme, 17
tilfældig stikprøve, 17
enhedsdevians, 126
  Taylorudvikling, 129
enhedsdevians for eksponentiel dispersionsmodel, 140
enhedsvariansfunktion, 136
equikorrelationsmatrix, 499
equikorrelationsmatrix, 22
estimable kontraster, 259
estimation af dispersionsparameter, 222
  maksimum likelihood estimat, 225
estimation af populationsmiddelværdi, 84
oversigtstabel, 84
f, udvalgsbrøk, 24
faktor
ordnet, 248
faktorniveauer, 248
faktorniveauer, formelle, 248
faktorniveauer, labels, 248
faktorvariable, 248
Fisher information, 115
Fisher's scoringsmetode, 191
Fishers eksakte test, 399
fittede værdier, 181
forskellige hældninger, parametrisk fremstilling, 254
forsøg
kontrolleret, 400
fortsættelses logit, 446
frembringer for log-lineær model, 467, 473
fuld model, 165

generaliserede lineære modeller
homogenitetstest, 230
regressionsmodeller, 223
generaliseret lineær model
fuld model, 165
modelvektør, 166
mættet model, 165
generaliseret lineær model
fittede værdier, 181
hat-matrix, 210
linkfunktion, 167
lokal design matrix, 168
generaliseret lineær model, 164
dimension, 165
modelmatrix, 166
generaliseret lineær model, konfidensinterval for enkelte parametre, 188
generaliseret lineær model, test for modeltilpasning, 218
gennemsnitlig relativ værdi af interessevariabel pr analyseenhed, 10
gennemsnitlig relativ værdi pr analyseenhed, 36
gennemsnitlig værdi pr analyseenhed
kvotientskøn, 88
populationsværdi, 5
grafisk model, 476
Gumbel-regression, 366
Hartley-Ross estimator, 44
hat-matrix, 210
Helmert-transformation, 259
hierarkisk model, 497
homogenitetstest, 230
binomialfordeling, 556
empiriske varianser, 239
gamma fordeling, 590
negativ binomialfordeling, 572
normal fordeling, 491, 505
Poissonfordeling, 579
Horvitz-Thompson estimator, 72
hændelsesrate (hazard rate), 596
incidensmatrix, 256
indeksmængde for eksponentiel dispersionsmodel, 132
indeksparameter, 132
indeksparameter for eksponentiel dispersionsmodel, 132
information, forventet, 115
information, observeret, 114
information, forventet, 114
information, observeret, 114
information, ved transformationer, 117
informationsmatrix, 115
interaction, 312
intercept led, 247
intervalskala, 244
intraklassekorrelation
   ved beta-binomial sampling, 559
intraklassekorrelation, 499
   binomial-beta, 554
   Gamma-reciprok gamma, 592
   negativ binomial-beta, 573
   Poisson-gamma, 581
intraklyngekorrelation, 88, 706
iterative metoder
   Fisher’s scoringsmetode, 191
ITPRINT-option i procedure GENMOD, 286

Kalman filter, 715
Kalman forstærkning, 715
kanonisk form for eksponentiel familie, 122
kanonisk link, 151
kanonisk parameter, 122, 132
klassifikation, 255, 427
   ordnet, 431
klyngedvælgelse
   oversigtstabel, 100
klyngedvælgelse, 85
   brug af gennemsnitlig klyngetotal, 86
   kvotientskøn over gennemsnitlig værdi pr analyseenhed, 88
   størrelseskorrigerede klyngetotaler, 89

udvælgelse proportional med estimeret størrelse, 96
udvælgelse proportional med størrelse, 97
kohortestudier, 400, 403
kollinearitet, 274
komplementær log-log, 154
konfidensinterval
   for populationsandel afvigende enheder, 30
   for populationsmiddelværdi, 27, 28
   med fastlagt længde, 28
   konfidensinterval for parametre i generaliseret lineær model, 188
konjugeret klasse af fordelinger, 550
kontraster
   Helmert-transformation, 259
   sum-kodning, 259
   treatment-kodning, 260
kontraster, estimable, 259
kontrolleret forøg, 400
korrektion for effekter, 303
korrektion for endeligt population, 24
korrelationskoefficient
   partiel, 476
korrespondanceanalyse, 479
ekovariable
   kontinuerte, 244, 246
   kvalitative, 244, 248
   kumulantfrembringer, 122
   kumulativ
   logit, 447
   odds, 447
   kumulativ odds ratio, 448
   kvotient
relativ varians, 9
kvotientskøn, 38, 39, 46
korrigeret, 43
skævhed, 39-41
ved udvælgelse med vilkårlige ssh, 70
latent variabel, 442, 449, 460
LD₅₀, 358
leverage, 339
likelihood uafhængighed, 108
likelihood-sufficiens, 111
likelihoodfunktion, 107
likelihoodkvotient konfidensinterval, 188
eksempel, 284
Likert skala, 437
linkfunktion, 150, 151, 167
kanonisk, 151, 153
log-likelihoodfunktion, 107
log-lineær model, 325, 430, 473
for antalstabel, 466
logarithmisk normalfordeling
fordeling af produkt af, 9
todimensional, 14
logistisk regression, 170, 173, 193, 203, 358
deviansanalyse, 293
logit
baseline, 439
betinget, 442
fortsætelses-, 446
kumulativ, 447
multinomial, 441
nabo, 443
nested, 442
logit-transformation, 352
lokal design matrix, 168
maksimum likelihood estimat, 117
marginal symmetri, 412
marginal fordeling
for empiriske varianser fra normalfordelte obs., 602
for empiriske varianser udtrykt ved F-fordele, 604
ved binomial-beta fordeling, 557
ved eksponentielle familier, 553
ved gamma-Poisson fordeling, 582
ved gamma-reciprok gamma fordeling, 591
ved negativ binomial-beta fordeling, 572
ved normal-normal fordeling, 498
ved normalfordeling med tilfældig middelværdi og varians, 533
marginal tabel, 432
marginalisere, 476
marginalitet
af led i modelformel, 270
matrix
sammensat symmetrisk, 500
matrixeffekt, 312
McNemer’s test, 414
middelresidualdevians, 292
middelværdiafbildning, 123, 136
middelværddiligningen, 178
middelværdiparametrisering
af eksponentiel familie, 124
middelværdirum, 123
minimal model, modelmatrix, 248
ML-estimation af dispersionsparameter, 225
model I for normalfordelte obs., 489
model II for normalfordelte obs., 496
modelformel, 275
modelmatrix, 166
for kovariable, 247
modelvektor, 166
momentestimation
i binomial-beta fordeling, 562, 629
i gamma-reciprok gamma fordeling, 595, 629
i marginal fordeling af empiriske varianser, 607
i negativ binomial fordeling, 584, 629
i negativ binomial-beta fordeling, 578, 629
i negativ Polya fordeling, 629
i normal-normal fordeling, 503, 542
i Poisson-gamma fordeling, 584, 629
i Polya fordeling, 562, 629
i reciprok beta fordeling, 595, 629
negativ Polya fordeling, 578
ved normalfordeling med tilfældig middelværdi og varians, 543
ved normalfordeling med tilfældig middelværdi og varians, 536
momentfordeling, 64
Musefostre
bestemmelse af residualer, 203
deviansanalyseskema, 293
fittede værdier, 203
introduktion, 170
parameterestimation, 193
test for modeltilpasning, 219
Mål for influens, Dfbetas, 344
Mål for influens, Dffits, 344
målgruppe, 17
mættet model, 165
N(μ, σ²) fordeling, 133, 138
nabokategori
odds, 443
naturlig eksponentiel familie, 122
Neyman’s kriterium, 109
Neyman-allokering, 79
nominal skala, 244
normalfordeling, 133, 138
odds, 351
baseline, 439
fortsættelses, 445
kumulative, 447
nabokategori, 443
proportional, 448, 458
odds ratio
for baseline-odds, 439
kumulativ, 448
Odds ratio, fordeling af estimeret, 395
Odds-ratio, 354
odds-ratio
betinget test, 398
offset, 165
offset værdi, 234
operationskaracteristik for prøvningsteknik, 408
optimal allokering af stikprøveheder
ved stratifikation, 77
ordnede responskategorier, 450
ordnet klassifikation, 431
P(λ) fordeling, 127
parallele linier, parametrisk model, 254
partial leverage, 337
Parvise sammenligninger, 419
passagertilfredshed, 436
  baseline odds, 440
  fortsættelses odds, 446
  kumulative odds, 447
  nabokategori odds, 444
Pearson residual
  standardiseret, 212
Pearson residual, studentiseret, 212
Pearson-residual, 201
Pearson-teststørrelse for modeltilpasning, 220
Poisson-regression, 233
Poissonfordeling, 127
populationskovarians, 8
  korrigeret, 8
  relativ, 8
  udtrykt ved korrelationskoef
  ficient, 9
populationsmiddelværdi
  estimat for, 61
populationstotal, 5
populationsvarians, 6
  estimation, 25
  korrigeret, 6
potenstransformationer, 154
PPS-sampling, 68, 97
primære stikprøveenheder, 101
PROC GLM
  MANOVA, 620
  tilfældig model for middelvær
dier, 518, 520
PROC INSIGHT
  deviansanalyse, 292
PROC MIXED
tilfældig model for middelvær
dier, 521, 525
PROC VARCOMP
  tilfældig model for middelværdier, 526
produkt
  relativ varians, 9
profil-likelihood, 108
profil-log-likelihood, 108
profillikelihood estimat, fordeling,
  187
profilplot, 313
proportional allokering af stikprøveenheder, 75
proportional odds model, 448, 458
prospektive undersøgelser, 401
prædiktiv fordeling
  ved binomial-beta fordeling,
    709
  ved gamma-reciprok gamma fordeling, 709
  ved negativ binomial-beta fordeling, 709
  ved normal-normal fordeling,
    709
  ved Poisson-gamma fordeling, 709
prædiktin
  i tidsrække, 714
prædiktiv fordeling, 680, 687
  ved binomial-beta fordeling,
    693
  ved gamma-reciprok gamma fordeling, 704
  ved negativ binomial-beta fordeling, 699
  ved normal-normal fordeling,
    706
ved Poisson-gammafordeling, 702
prædiktor, 151
prædiktorrum, 151
præposteriorimiddelværdi, 682
af \(\mu\), 690
af \(V(\mu)\), 690
præposteriorivarians, 682
af \(\mu\), 691
quasi-devians, 148
quasi-likelihood, 148

Receiver Operating Characteristic
til klassifikationprocedure, 409
regressionsmodel, 223
aposteriorifordeling, 728
balanceret, 634
Poisson-fordeling, 234
regressionsmodeller, tilfældig modell, 646
momentestimation, 650
REML-estimation, 649
regressionsskøn for populationsgen-
nemsnit, 51, 60
relativ risiko, 353
relativ værdi af interessevariabel
for populationen, 11
pr analyseenhed, 10, 11
relativ værdi pr analyseenhed
gennemsnitlig, 36
stikprøvegennemsnit, 37
REML-estimat, 515
repeaterbarhedsbetingelser, 527
reproducerbarhedsbetingelser, 527
reproduktiv eksponentiel dispers-
sionsmodel, 133
residual
arbejds-, 203
working, 192
devians-, 200
Pearson-, 201
respons-, 200
standardiseret, 211
studentiseret, 211
Wald-, 202
working, 203
residualdevians, 217
residualdevians, skaleret, 217
response
working, 192
responsen
ordnede, 450
responskurve, 449
responsresidual, 200
standardiseret, 211
studentiseret, 212
responsvariabel, 428
Restricted maksimum likelihood
estimat, 515

Sammenligning af hyppigheder, for-
deling af odds ratio, 395
Sammenligning af hyppigheder, kon-
fidensinterval for differens, 392
Sammenligning af hyppigheder, kon-
fidensinterval for relativ
risiko, 393
Sammenligning af hyppigheder, ek-
sakt konfidensinterval for
odds ratio, 398
sandsynlighedskorrigeret værdi, 66
SAS GENMOD
konfidensintervaller for para-
metre, 285
SAS INSIGHT
konfidensintervaller for parametre, 284
scorefunktion, 112
ved binomial-beta fordeling, 563
ved Poisson-gamma fordeling, 586
selvvægtende estimator, 75
sensitivitet af klassifikationprocedure, 408
signal/støj forhold
binomial-beta, 554
flerdimensional normalfordeling, 612
Gamma-reciprok gamma, 591
negativ binomial-beta, 573
Poisson-gamma, 581
signal/støj-forhold
konfidensinterval, 506
signal/støjforhold
normalfordeling, 497
simpel tilfældig udvælgelse, 18
Simpson’s paradoks, 475
skaleret devians mellem observationer og model, 146
specificitet af klassifikationprocedure, 408
spredning
relativ, 7
standardform for eksponentiel familie, 122
statistisk model, 107
stikprøve
fra endelig population, 17
selvvægtende, 101
stikprøvegennemsnit
kovarians mellem, 33
momenter for, 23
som estimator for populationsgennemsnit, 23
stikprøvekovarians, 34
forventningsværdi, 35
stikprøveramme, 17
stikprøvevarians
momenter, 25
stratificeret udvælgelse
oversigtstabell, 84
stratifikation, 73
Neyman allokering, 78
optimal allokering, 77
oversigtstabell, 84
proportional allokering, 75
vilkårlig allokering, 74
strukturfordeling, 549, 679
konjugeret, 551
støtte, 122
sufficiens, 109
superpopulationsmodeller, 3
tabelform, 249
target population, 17
test af hypoteser vedrørende enkelte koeficients i generaliseret lineær model, 282
test for modelreduktion, 280
Wald teststørrelse, 295
test for modeltilpasning, 218
Pearson-teststørrelse, 220
Wald-teststørrelse, 221
Thurstone metode, 437
tilfældig stikprøve, 17
totrinsudvælgelse, 101
toxitet, 358
treatment-kodning, 260
tværsnitsundersøgelse, 400, 403, 432
udsnitsundersøgelse, 400
udvalgsbrøk, 19
udvalgsbrøk, $f$, 24
udvælgelse
  med tilbagelægning, 18
  proportional med interessevariabel, 63
  proportional med størrelse, 68
  simpel tilfældig, 18
  uden tilbagelægning, 18
udvælgelsessandsynligheder
  simpel tilfældig udvælgelse, 19
udvælgelsesvektor, 20
  momenter for, 21
undersøgelse
  prospektiv, 399
  retrospektiv, 400, 403
utility, 442

varians
  relativ, 7
variansestimat
  REML-estimation, 649
variansfunktion, 124
variansfunktion og devians, 129
varianshomogenitet
  test for, 239
varianskomponent, 497
varianskomponentmodel, 497
variansstabiliserende transformationer, 151
variationskoefficient, 7
  korrigeret, 7
vekselvirkning, 312
vægtet gennemsnitlig gruppestørrelse,
  $n_0$, 488
vægtet model, 143
vækst af Ramus-knogle, 226, 652

Wald-konfidensinterval, 188
  eksempel, 283
Wald-residual, 202
Wald-teststørrelse, 221
Wald-teststørrelse for fjernelse af led, 296
working residual, 203
working response, 192

Yule's krydsprodukt ratio, 387, 462