En Introduktion til Statistik

Bind 3C

Hierarkiske modeller

Poul Thyregod

LYNGBY 1998

IMM

Trykt af IMM - DTU
Bogbinder Hans Meyer

Indhold

Afsnit 4

Modeller for flerdimensionale antalstabeller

fil kont2.tex 98-04-12

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 klassifikationskriterium $(A, B, \ldots)$. 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 responsvariabel, 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 responsvariabel.

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

Forklarende variable Antal faktorer	Responsvariable Antal klassifikations-kriterier
Ingen	Flere
Én eller flere	Én eller flere
Flere	Ingen

Endelig kan der som forklarende variable også indgå kontinuerte kovariable.

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

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

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

OBSERVEREDE ANTAL, x_{ij}

		Klassifikation B			
		1	2	...	J	Ialt
Klassifikation A	1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,J}$	$x_{1,+}$
	2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,J}$	$x_{2,+}$
	⋮	⋮	⋮	⋱	⋮	⋮
	I	$x_{I,1}$	$x_{I,1}$	...	$x_{I,J}$	$x_{I,+}$
	Ialt	$x_{+,1}$	$x_{+,2}$	...	$x_{+,J}$	$x_{+,+}$

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

- Observation af $I \times J$ uafhængige $\mathrm{P}(\lambda_{i,j})$-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-fordelte observationer:

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

Modellen svarer til eksempel 2.13.2

Frekvensfunktionen svarende til denne model er

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

De forventede værdier af celleværdierne er $\mathsf{E}[X_{i,j}] = \lambda_{i,j}$. Vi bemærker at marginalsummerne ligeledes er Poisson-fordelt. Eksempelvis har vi for $X_{i,+} = \sum_{j=1}^{J}$, 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

$$l(\boldsymbol{\lambda}; \mathbf{x}) = \Sigma_{i,j} x_{i,j} \ln(\lambda_{i,j}) - \Sigma_{i,j} \lambda_{i,j} \tag{4.1.2}$$

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

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

$$\vartheta_{i,j} = \alpha_i + \beta_j \ ,$$

evt efterfulgt af en vurdering af rækkeeffekt og/eller søjleeffekt som diskuteret i afsnit 2.13. Modellen kaldes undertiden en log-lineær model, da den er lineær (additiv) i logaritmen til middelværdierne.

Modellen kan formelt formuleres ved kontraster mellem $\vartheta_{i,j}$ ved

$$\Delta^{A;B}_{i_1,i_2;j_1,j_2} = (\vartheta_{i_1,j_1} - \vartheta_{i_1,j_2}) - (\vartheta_{i_2,j_1} - \vartheta_{i_2,j_2}) = 0$$

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

Kvartal	Uheldskategori					Ialt
	Ene - uheld	Ind-hent-nings - uheld	Møde - uheld	Sving og kryds	Andre	
januar	105	67	54	310	29	565
april	146	76	85	485	37	829
juli	114	94	100	461	32	801
oktober	105	62	74	314	24	579
Ialt	470	299	313	1570	122	2774

Det vil naturligt at opfatte de 4×5 antal i tabellen som realisationer af 4×5 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:

Køn	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
Kvinder	234 2.3 %	559 5.4 %	1 157 11.2 %	5 826 56.4 %	2 553 24.7 %	10 329
Mænd	183 2.7 %	446 6.5 %	1 005 14.6 %	3 841 55.6 %	1 428 20.7 %	6 903
Ialt	417 2.4 %	1 005 5.8 %	2 162 12.5 %	9 667 56.1 %	3 981 23.1 %	17 232

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 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, $\text{Mult}(N, p_{1,1}, \ldots, p_{1,J}, \ldots, p_{I,1}, \ldots, p_{I,J})$

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

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

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

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

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

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

$$f(\mathbf{x}_i) = \frac{x_{i,+}!}{\prod_{j=1}^{J} x_{i,j}!} \prod_{j=1}^{J} p_{i,j}^{x_{i,j}} \tag{4.1.5}$$

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

$$f(\mathbf{x}) = \prod_{i=1}^{I} \frac{x_{i,+}!}{\prod_{j=1}^{J} x_{i,j}!} \prod_{j=1}^{J} p_{i,j}^{x_{i,j}} \tag{4.1.6}$$

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

Log-likelihoodfunktionen svarende til produkt-multinomialfordelingsmodellen er

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

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.

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

I overensstemmelse med betragtningerne i afsnit 2.8 vil vi forestille os afhængigheden af de forklarende variable beskrevet ved en modelmatrix, $\mathbf{X}$, hvis i'te række $\mathbf{x}_i^{*T}$ beskriver de forklarende variable svarende til den i'te multinomialfordeling.

Principielt kan man vælge et sæt parametre, $\boldsymbol{\beta}_j$ til hver af de lineære prædiktorer, ϑ_j, (og eventuelt også et sæt forklarende variable, $\mathbf{x}_{ij}^*$ til hver af disse). Man vil da opstille modeller af formen

$$\vartheta_{ij} = \mathbf{x}_{ij}^{*T} \boldsymbol{\beta}_j \tag{4.2.1}$$

for hver af de $r-1$ prædiktorer, ϑ_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

$$\vartheta_{ij} = \mathbf{x}_i^{*T} \boldsymbol{\beta}_j \ . \tag{4.2.2}$$

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.

Vi minder om, at familien af multinomialfordelinger, $\text{Mult}(n, p_1, p_2, \ldots, p_r)$ er en fuld eksponentiel 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

$$\tau_j(\vartheta) = \exp(\vartheta_j) \Big/ \sum_{\nu=1}^{r} \exp(\vartheta_\nu) = p_j , \tag{4.2.3}$$

og elementerne i enhedsdispersionsmatricen er

$$V_{i,j}(\mathbf{p}) = \begin{cases} p_i(1 - p_i) & \text{for } i = j \\ -p_i p_j & \text{for } i \neq j \end{cases} \tag{4.2.4}$$

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

Middelværdi og dispersion svarende til antalsparameteren n fås ved at multiplicere enhedsmiddelværdien og enhedsdispersionsmatricen med antalsparameteren n i analogi med de endimensionale eksponentielle dispersionsmodeller.

Parametriseringen ved ϑ er en overparametrisering. Således vil sættet ϑ_0, og sættet

$$\vartheta_1 = \vartheta_0 + c\mathbf{1} ,$$

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

I almindelighed fortæller den enkelte komponent af ϑ 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 busselskab stikprøveundersøgelser blandt passagerne, hvor passagererne bliver 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.

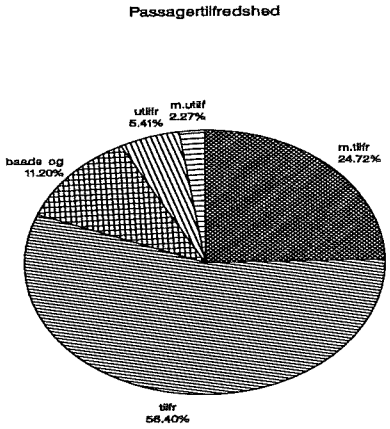
Et af spørgsmålene vedrører kundens tilfredshed med overholdelsen af køreplanen. Spørgsmålet har svarkategorierne "Meget utilfreds", "Utilfreds", "Både og", "Tilfreds", "Meget tilfreds".

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

	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
Antal	234	559	1 157	5 826	2 553	10 329
Andel (pct)	2.27	5.41	11.20	56.40	24.72	100.0

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, $p_1, p_2, \ldots, p_5$)-fordelt variabel, hvor $(p_1, p_2, \ldots, p_5)$ kan fortolkes som andelen af kvindelige passagerer i rettidige busser, som vil svare hhv "Meget utilfreds", "Utilfreds", etc. på dette spørgsmaal.

Fordelingen er illustreret grafisk i nedenstående figur.



Svarkategorierne 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***

Ovenstående svarskala er en variant af en såkaldt Likert skala benævnt efter den amerikanske statistiker, Rensis Likert (1903-1981), der introducerede denne metode i sin Ph.D. afhandling ved Columbia University's Department of Psychology, *A Technique for the Measurement of Attitudes* fra 1932. Likerts oprindelige ide var at lade respondenten angive sin grad af enighed/uenighed med en forelagt holdning på en skala med et neutralt midtpunkt. Likert viste i sin afhandling, at denne metode var lige så effektiv, som den mere komplicerede Thurstone metode, der bygger på en psykofysisk metode med intervaller, der fremtræder lige store. □

4.2.2 Odds- og oddsratioer, ét klassifikationskriterium

Betragt et endeligt udfaldsrum Ω for et forsøg.

Vi betragter en klassifikation $\Omega = \cup_{j=1}^{r} A_j$ i disjunkte kategorier med de

tilknyttede sandsynligheder

$$p_j = \mathsf{P}\,[A_j],\ j = 1, \ldots, r\,.$$

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 $p_1 = \mathsf{P}\,[A_1]$ som

$$\theta = \frac{p_1}{1 - p_1} \tag{4.2.5}$$

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 kategorierne $A_1, A_2, \ldots, A_r$ være en multinomialfordeling med den r-dimensionale kanoniske parameter $\boldsymbol{\vartheta}$ 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'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 parametrisering, i andre anvendelser er det en anden parametrisering, der muliggør en model, der er lineær i de forklarende variable. I praksis er det derfor fordelagtigt at have flere forskellige parametriseringer til rådighed.

Agresti (1990) beskriver en lang række modeller til analyse af flerdimensionalt 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 hændelsen A_j ved

$$\theta_j = \frac{\mathsf{P}[A_j]}{\mathsf{P}[A^0]} = \frac{p_j}{p^0} \tag{4.2.6}$$

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 referencekategorien netop er den sidste kategori, A_r, dvs sådan at

$$\theta_j = \frac{\mathsf{P}[A_j]}{\mathsf{P}[A_r]} = \frac{p_j}{p_r} \tag{4.2.7}$$

Det tilsvarende logaritmiske mål:

$$\vartheta_j = \ln(\theta_j),\ j = 1, 2, \ldots, r-1 \tag{4.2.8}$$

kaldes baseline logit'en. Forskellen mellem logit'erne, ϑ_j og ϑ_ν 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

$$\omega_{i,j} = \frac{\theta_i}{\theta_j} = \frac{p_i/p^0}{p_j/p^0} = \frac{\mathsf{P}[A_i]}{\mathsf{P}[A_j]}, \tag{4.2.9}$$

Det tilsvarende logaritmiske mål, differensen mellem baseline logit'erne,

$$\Delta_{i,j} = \ln(\omega_{i,j}) = \ln(p_i) - \ln(p_j) = \vartheta_i - \vartheta_j \qquad (4.2.10)$$

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 betragte odds-ratioerne, eller logaritmen til odds-ratioerne.

Hvis der er ialt r forskellige hændelser, er der ialt $\binom{r}{2}$ 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 betragter 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.

	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
Antal	234	559	1 157	5 826	2 553	10 329
Odds	0.2022	0.4831	1.0000	5.035	2.2066	
Logit	-1.5983	-0.7274	0.0000	1.6165	0.7914	

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

	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
Antal	234	559	1 157	5 826	2 553	10 329
Odds	0.0917	0.2190	0.4532	2.2821	-	
Logit	-2.3897	-1.5189	0.7914	0.8251	-	

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 vædierne $\mathbf{x}_i^*$ af de forklarende variable, og at man ønsker at beskrive sættet, $(p_{i1}, \ldots, p_{ir})$, af responssandsynligheder som funktion af disse forklarende variable.

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

$$p_j(\mathbf{x}_i^*) = \frac{\exp(\mathbf{x}_i^{*T}\boldsymbol{\beta}_j)}{\sum_{\nu=1}^{r}\exp(\mathbf{x}_i^{*T}\boldsymbol{\beta}_\nu)}, \tag{4.2.11}$$

hvor parametervektoren $\boldsymbol{\beta}_j$ svarende til baseline kategorien sættes til $\mathbf{0}$.

Såfremt man har valgt den sidste kategori, A_r, som referencekategori, dvs $\boldsymbol{\beta}_r = \mathbf{0}$, bliver modellen jvf. (4.2.8)

$$\vartheta_j(\mathbf{x}_i^*) = \ln\left(p_j(\mathbf{x}_i^*)/p_r(\mathbf{x}_i^*)\right) = \mathbf{x}_i^{*T}\boldsymbol{\beta}_j \tag{4.2.12}$$

Man modellerer således med en separat parametervektor, $\boldsymbol{\beta}_j$, for hver af kategorierne.

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

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

Modeller, hvor responsvariablen beskriver en diskret mængde af valgmuligheder, kaldes ofte for *discrete choice models*. Amemiya (1981) har givet en oversigt over modeller for diskret valg.

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 $\boldsymbol{\beta}$ er den samme for alle responskategorier. Modellen er på formen:

$$p_j(\mathbf{x}_i^*) = \frac{\exp(\mathbf{x}_i^{*T}\boldsymbol{\beta}_j)}{\sum_{\nu=1}^{r}\exp(\mathbf{x}_i^{*T}\boldsymbol{\beta}_\nu)}, \tag{4.2.13}$$

Idet $R_1, \ldots, R_r$ angiver r uafhængige stokastiske variable, der hver for sig følger en $\mathrm{Max}_1(\mathbf{x}_i^{*T}\boldsymbol{\beta}_j, 1)$-fordeling. Forbrugeren vælger nu kategorien

j, hvis R_j er større end de øvrige variable R_ν, $\nu \neq j$. Sandsynligheden for denne hændelse er netop p_j. De variable R_j kaldes undertiden for latente variable.

Størrelsen $v_{ij} = \mathbf{x}_i^{*T}\boldsymbol{\beta}_j$ 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

$$p_j(\mathbf{x}_{ij}^*) = \frac{\exp(\mathbf{x}_{ij}^{*T}\boldsymbol{\beta})}{\sum_{\nu \in C_i} \exp(\mathbf{x}_{i\nu}^{*T}\boldsymbol{\beta})}, \tag{4.2.14}$$

For ethvert par af muligheder A_j og A_ν finder man

$$\ln\left(p_j(\mathbf{x}_{ij}^*)/p_\nu(\mathbf{x}_{i\nu}^*)\right) = (\mathbf{x}_{ij}^* - \mathbf{x}_{i\nu}^*)^T\boldsymbol{\beta} \tag{4.2.15}$$

Betydningen af de enkelte forklarende variable for valget mellem A_j og A_ν 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_ν ikke afhænger af de øvrige valgmuligheder i mængden af alternativer, C_i. Denne egenskab kaldes undertiden for uafhængighed af irrelevante alternativer.

Modellen (4.2.14) kaldes undertiden en betinget logit model, eller en nested logit model (se Brownstone og Small (1989)). Modellerne har fundet udbredt anvendelse til beskrivelse af valg af transportmiddel, se fx Ben-Akiva og Lerman (1985).

Vi anfører 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 flerdimensional skaleringsmetode 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

tilknyttet en række (sædvanligvis binære) egenskaber, beskrevet ved vektoren $\mathbf{x}_j^*$. Den samlede nytte, som individet tillægger den j'te valgmulighed tænkes fremkommet som en linearkombination $\mathbf{x}_j^{*T}\beta$ af utiliteterne (partworth nytten) β_p for hver af egenskaberne p. Individet vælger den mulighed, der giver den største samlede nytte.

Conjoint analyse bruges ofte i forbindelse med forsøgsplanlægning til at bestemme den kombination af egenskaber, der giver den største nytte. Ofte bruges resultatet af en conjoint analyse (utiliteterne) som input til en simuleringsmodel 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,

$$\theta_j^N = \frac{\mathsf{P}\,[A_j]}{\mathsf{P}\,[A_{j+1}]} = \frac{p_j}{p_{j+1}}\,, \quad j = 1, \ldots, r-1 \tag{4.2.16}$$

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:

$$\vartheta_j^N = \ln(\theta_j^N),\ j = 1, 2, \ldots, r-1 \tag{4.2.17}$$

kaldes nabologit'en.

Relationen

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

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

$$\vartheta_j = \vartheta_j^N + \vartheta_{j+1}^N + \cdots + \vartheta_{r-1}^N \tag{4.2.18}$$

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.

	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
Antal	234	559	1 157	5 826	2 553	10 329
Odds	0.4186	0.4831	0.1986	2.2820	-	
Logit	-0.8708	-0.7274	-1.6165	0.8251	-	

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

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

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

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

der er på formen

$$\vartheta_j = \alpha_j^* + \mathbf{u}_j^{*T}\boldsymbol{\beta}$$

med $\mathbf{u}_j^* = (r-j)\mathbf{x}^*$. □

4.2.5 Fortsættelses-odds

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

$$\theta_j^F = \frac{\mathsf{P}\,[A_j]}{\mathsf{P}\,[A_{j+1}] + \cdots + \mathsf{P}\,[A_r]} = \frac{p_j}{p_{j+1} + \cdots + p_r}\,, \quad j = 1, \ldots, r-1 \tag{4.2.19}$$

eller, svarende til den omvendte orientering,

$$\theta_j^{*F} = \frac{\mathsf{P}\,[A_{j+1}]}{\mathsf{P}\,[A_1] + \cdots + \mathsf{P}\,[A_j]} = \frac{p_{j+1}}{p_1 + \cdots + p_j}\,, \quad j = 1, \ldots, r-1 \tag{4.2.20}$$

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

Lad nemlig π_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 π_j ved

$$\pi_j = \frac{p_j}{p_j + \cdots + p_r}\,, \; j = 1, \ldots, r-1\,, \tag{4.2.21}$$

og der gælder

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

hvorfor vi har

$$\theta_j^F = \frac{\pi_j}{1 - \pi_j}\,, \tag{4.2.22}$$

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

Tilsvarende kan θ_j^{*F} fortolkes som odds svarende til de betingede sandsynligheder, π_j^*, for at responset er A_{j+1} givet at responset er A_{j+1} eller lavere.

De logaritmerede værdier af fortsættelses odds,

$$\vartheta_j^F = \ln(\theta_j^F) = \ln\left(\frac{\pi_j}{1-\pi_j}\right), \quad j = 1, \ldots, r-1, \tag{4.2.23}$$

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) fortsætelses odds og logit'er for de forskellige svarkategorier.

	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
Antal	234	559	1 157	5 826	2 553	10 329
Antal forts.	10 095	9 536	8 379	2 553		
Odds	0.0232	0.0586	0.1381	2.2820	-	
Logit	-3.7645	-2.8367	-1.9799	0.8251	-	

□

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

$$\mathrm{b}(n, n_1, \pi_1) \times \mathrm{b}(n - n_1, n_2, \pi_2) \times \cdots \times \mathrm{b}(n - n_1 - \cdots - n_{r-2}, n_{r-1}, \pi_{r-1}) \tag{4.2.24}$$

Såfremt man har en model, hvor parametrene i modelspecifikationen for den i'te fortsættelseslogit er forskellige fra parametrene i modelspecifikationen for den j'te fortsættelseslogit ($i \neq j$), kan man altså bestemme maksimumlikelihood estimaterne ved at maksimere likelihoodfunktionen svarende til

hver fortsættelseslogit for sig. Ydermere fås kvotienttesttørrelsen for modeltilpasning for hele modellen ved at summere kvotienttesttørrelsen 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 parametriseringen, der fremkommer ved at aggregere kategorierne. Vi indfører de kumulerede sandsynligheder, Π_j ved

$$\Pi_j = p_1 + p_2 + \cdots + p_j \,, \quad j = 1, 2, \ldots, r-1 \tag{4.2.25}$$

De kumulative odds indføres da som

$$\theta_j^K = \frac{\Pi_j}{1 - \Pi_j} \,, \quad j = 1, 2, \ldots, r-1 \tag{4.2.26}$$

og de tilsvarende kumulative logit'er er bestemt ved

$$\vartheta_j^K = \ln(\theta_j^K) = \ln\left(\frac{\Pi_j}{1 - \Pi_j}\right) \,, \quad j = 1, 2, \ldots, r-1 \tag{4.2.27}$$

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.

	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
Antal	234	559	1 157	5 826	2 553	10 329
Kumuleret antal	234	793	1 950	7 776	10 329	
Odds	0.0232	0.0832	0.2327	3.0458	-	
Logit	-3.7645	-2.4870	-1.4579	1.1138	-	

□

Såfremt der optræder forklarende variable, $\mathbf{x}^*$, kan man betragte en lineær (affin) model for de kumulative logit'er,

$$\vartheta_j^K(\mathbf{x}) = \alpha_j + \mathbf{x}^{*T}\boldsymbol{\beta} \,, \tag{4.2.28}$$

hvor $\mathbf{x}^*$ 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 α_i er ikke-aftagende som funktion af i.

For modellen bestemt ved (4.2.28) gælder, at odds-ratioen svarende til to forskellige værdier, $\mathbf{x}_1^*$ og $\mathbf{x}_2^*$ af de forklarende variable er den samme for alle kategorier, j. Der gælder nemlig

$$\vartheta_j^K(\mathbf{x}_1^*) - \vartheta_j^K(\mathbf{x}_1^*) = \ln\left(\frac{\Pi_j(\mathbf{x}_1^*)/(1-\Pi_j(\mathbf{x}_1^*))}{\Pi_j(\mathbf{x}_2^*)/(1-\Pi_j(\mathbf{x}_2^*))}\right) = (\mathbf{x}_1^* - \mathbf{x}_2^*)^T\boldsymbol{\beta} \tag{4.2.29}$$

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

$$\vartheta_j^K = \alpha_j \,, j = 1, \ldots, r-1 \tag{4.2.30}$$

Modellen svarer til, at de kumulative sandsynligheder, Π_j er parametriseret som

$$\Pi_j = \frac{\exp(\alpha_j)}{1+\exp(\alpha_j)} \quad, j = 1, \ldots, r-1 \tag{4.2.31}$$

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

Definer nu den variable J ved

$$J = j \quad \text{hvis } \alpha_{j-1} < R \leq \alpha_j, \tag{4.2.32}$$

hvor $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_r = \infty$. Da gælder netop, at $\mathsf{P}\,[R \leq \alpha_j] = \Pi_j$, $j = 1, \ldots, r-1$, hvorfor vi har, at $\mathsf{P}\,[J = j] = \mathsf{P}\,[A_j] = p_j$ 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 responset. R betegnes undertiden den latente variable.

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

Specielt, hvis der kun er en enkelt, kontinuert, forklarende variabel, x, kan vi betragte responskurven svarende til det j'te kumulative respons,

$$F_j(x) \stackrel{\text{DEF}}{=} \mathsf{P}\,[A_1 \cup A_2 \cdots \cup A_j] = \mathsf{P}\,[R \leq \alpha_j] \tag{4.2.33}$$

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, kontinuert, forklarende variabel gælder altså

$$F_\nu(x) = F_j\big(x + (\alpha_\nu - \alpha_j)/\beta\big) \tag{4.2.34}$$

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. □

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 kontinuerte 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 kontinuerte forklarende variable kan man vælge at modellere de betingede sandsynligheder, π_j (4.2.21), eller de kumulerede sandsynligheder Π_j (4.2.25) ved en log-log, eller en komplementær log-log link, eller eventuelt ved en probit-link.

4.2.8 Regressionsmodeller

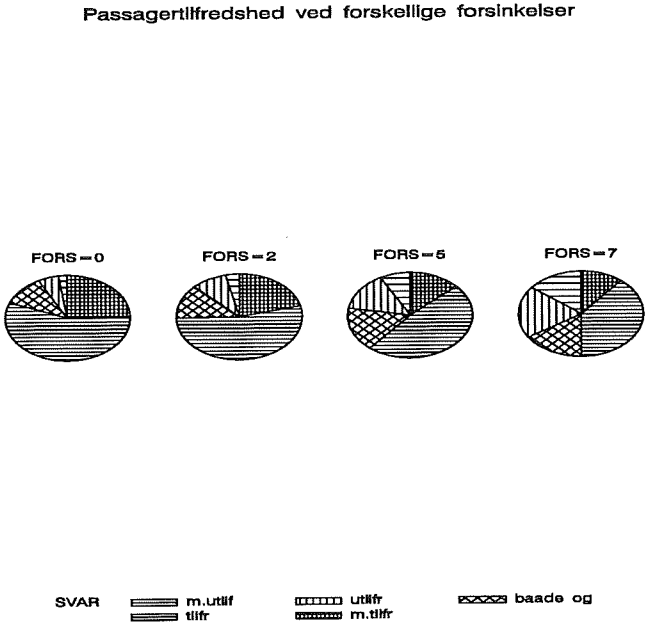
Eksempel 4.2.7 ***Tilfredshedsangivelser fra buspassagerer i afhængighed af bussens forsinkelse***

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 forsinkelse.

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

Forsinkelse min.	Meget utilfr.	Utilfr.	Både og	Tilfr.	Meget tilfr.	Ialt
0	234 2.3 %	559 5.4 %	1 157 11.2 %	5 826 56.4 %	2 553 24.7 %	10 329
2	41 3.6 %	100 8.9 %	145 12.9 %	602 53.5 %	237 21.1 %	1 125
5	42 7.9 %	76 14.3 %	89 16.7 %	254 47.7 %	72 13.5 %	533
7	35 14.3 %	48 19.7 %	39 16.0 %	95 38.9 %	27 11.1 %	244

Nedenstående figur illustrerer disse fordelinger grafisk.



Svarkategorierne er ordnede, gående fra "Meget utilfreds" til "Meget tilfreds". Ordningen er afspejlet 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:

For-sink-else min.	"Meget utilfr."	" Utilfr."	"Både og"	" Tilfr."	"Meget tilfr."
0	2.3 %	5.5 %	12.1 %	69.5 %	-
2	3.6 %	9.2 %	14.7 %	71.8 %	-
5	7.8 %	15.4 %	21.3 %	77.4 %	-
7	14.3 %	23.0 %	24.2 %	77.9 %	-

Modellen er

$$\ln\left(\pi_j(x_i)/(1-\pi_j(x_i))\right) = \alpha_j + \beta_j x_i \ , \tag{4.2.35}$$

hvor x_i angiver forsinkelsen i minutter.

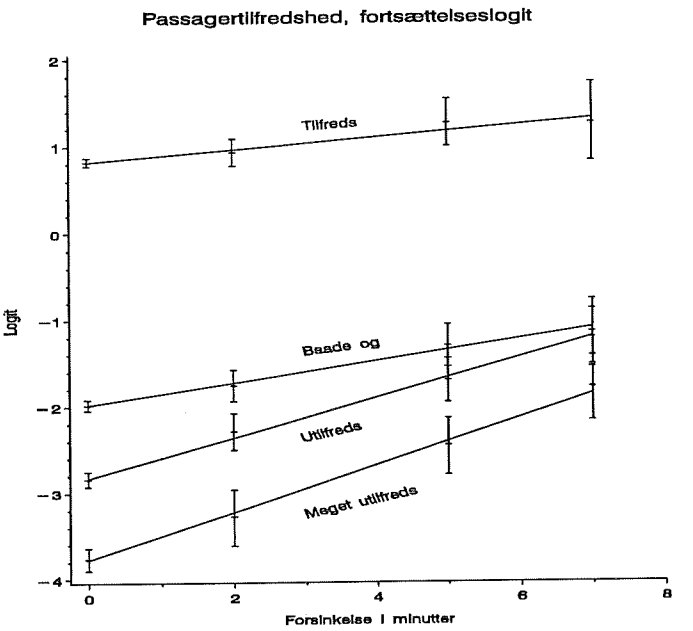
Estimationen kan udføres for hver svarkategori for sig, fx. ved SAS proceduren GENMOD.

Nedenstående tabel angiver for hver svarkategori estimaterne α_j og β_j samt de tilsvarende residualdevianser, $D(\mathbf{y}, \widehat{\boldsymbol{\pi}})$. Deviansbidragene er bestemt som deviansbidrag svarende til binomialfordelingen jvf udtrykket (4.2.24). Residualdeviansen, $D(\mathbf{y}, \widehat{\boldsymbol{\pi}})$ 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.

Svar-kategori	α_j	β_j	$D(\mathbf{y}, \widehat{\boldsymbol{\pi}})$
"Meget Utilfreds"	-3.7733	0.2736	0.3744
"Utilfreds"	-2.8287	0.2328	0.3152
"Både og"	-1.9814	0.1269	0.3297
"Tilfreds"	0.8234	0.0721	0.3271
"Meget Tilfreds"	-1.1101	-0.1367	1.0090
Ialt			2.3554

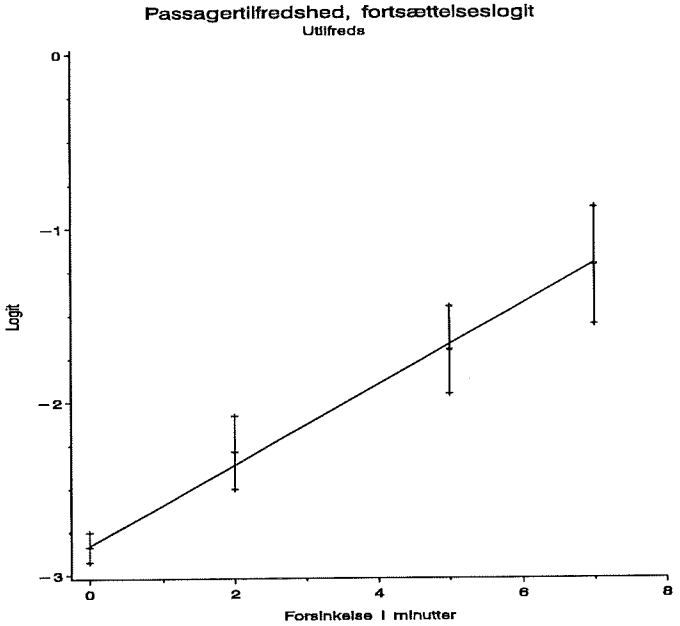
Residualdevianserne giver ikke anledning til afvisning af modellen.

De estimererede 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 π_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.

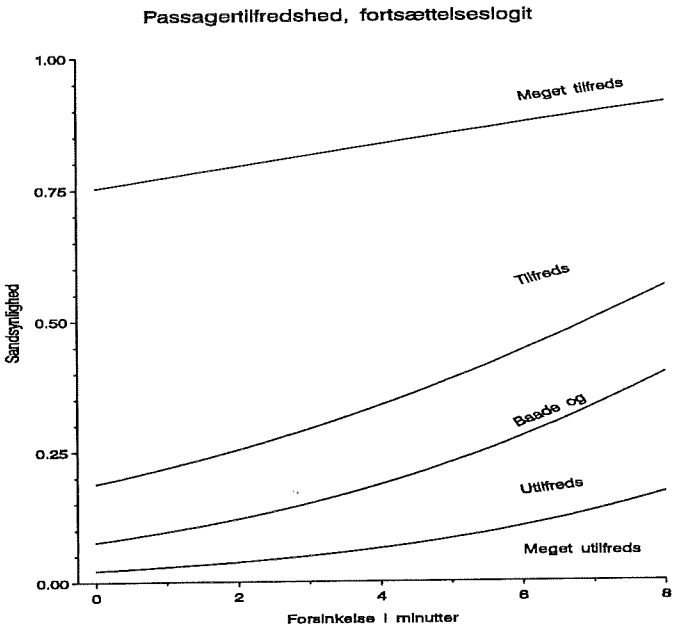


Betragter man deviansresidualerne

Forsink-else min.	Meget utilfr.	Utilfr.	Både og	Tilfr.
0	0.1335	-0.1829	0.0477	0.0712
2	-0.3088	0.7232	-0.3103	-0.4598
5	-0.3355	-0.2636	0.4061	0.5798
7	0.3856	-0.0669	-0.2573	-0.3186

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



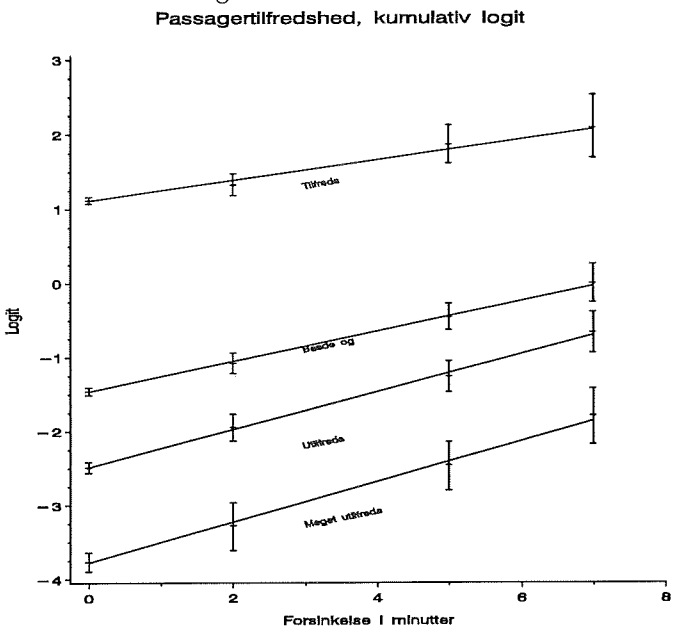
Modellering ved kumulative logit'er

Nedenstående tabel viser for hver svarkategori estimaterne α_j og β_j samt de tilsvarende residualdevianser, $D(\mathbf{y}, \widehat{\mathbf{\Pi}})$ 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.

Svar-kategori	α_j	β_j	$D(\mathbf{y}, \widehat{\mathbf{\Pi}})$
Meget Utilfreds	-3.7733	0.2736	0.3744
Utilfreds	-2.4849	0.2550	0.4157
Både og	-1.4603	0.2033	0.2157
Tilfreds	1.1101	0.1367	1.0090
Ialt			1.7991

Residualdevianserne giver ikke anledning til afvisning af modellen.

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

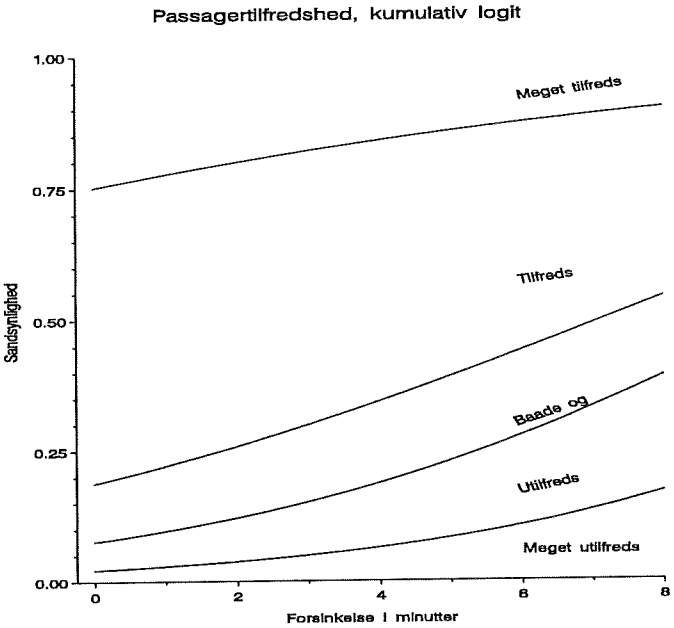


De tilsvarende deviansresidualer er angivet i nedenstående tabel

	Deviansresidualer			
Forsink-else min.	Meget utilfr.	Utilfr.	Både og	Tilfr.
0	0.1335	-0.0565	0.0956	0.1610
2	-0.3088	0.3551	-0.3280	-0.8505
5	-0.3355	-0.4577	-0.1140	0.5020
7	0.3856	-0.2773	0.2933	-0.0837

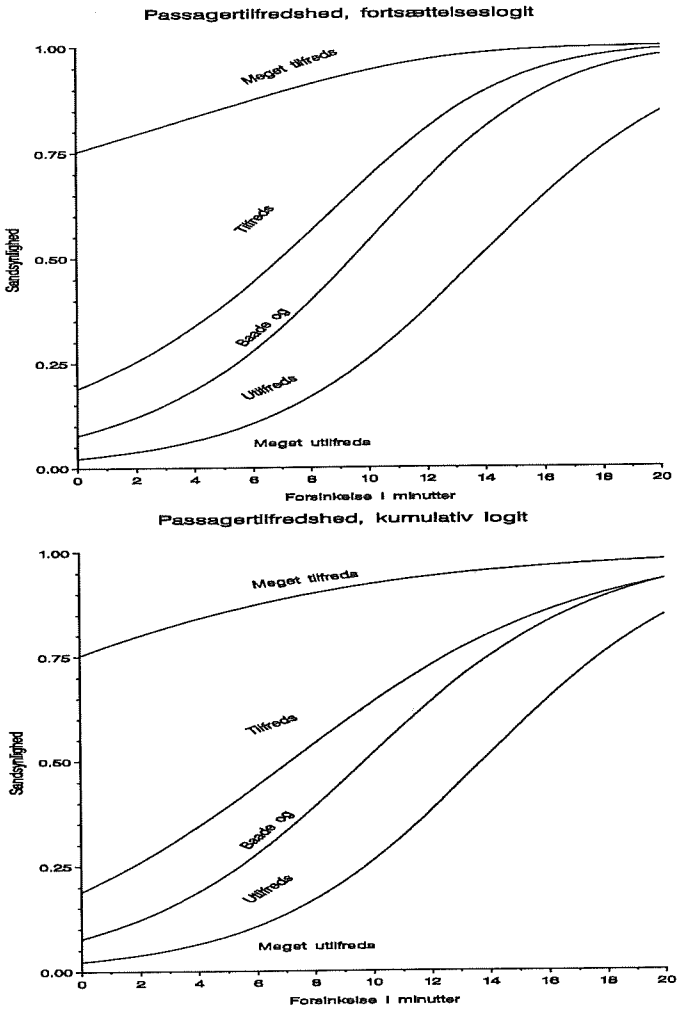
Også her ses at være en god tilpasning.

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

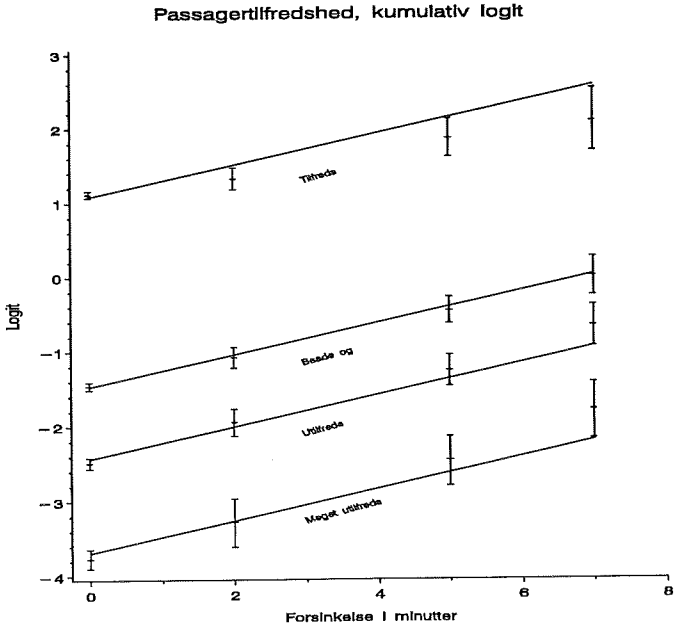
Vi vil her betragte modellen (4.2.28) svarende til at effekten af forsinkelsen på den kumulative logit er den samme for alle responskategorier, dvs modellen med proportionale odds.

Nedenstående tabel viser for hver svarkategori estimaterne, $\widehat{\alpha}_j$ for cut-points, estimatet, $\widehat{\beta}$ for den fælles hældning, samt residualdeviansen, $D(\mathbf{y}, \widehat{\mathbf{\Pi}})$.

Svar-kategori	α_j	β	$D(\mathbf{y}, \widehat{\mathbf{\Pi}})$
Meget Utilfreds	-3.6859	0.2129	6.9736
Utilfreds	-2.4390	0.2129	7.8757
Både og	-1.4675	0.2129	0.7657
Tilfreds	1.0804	0.2129	18.1811
Ialt			33.7961

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.

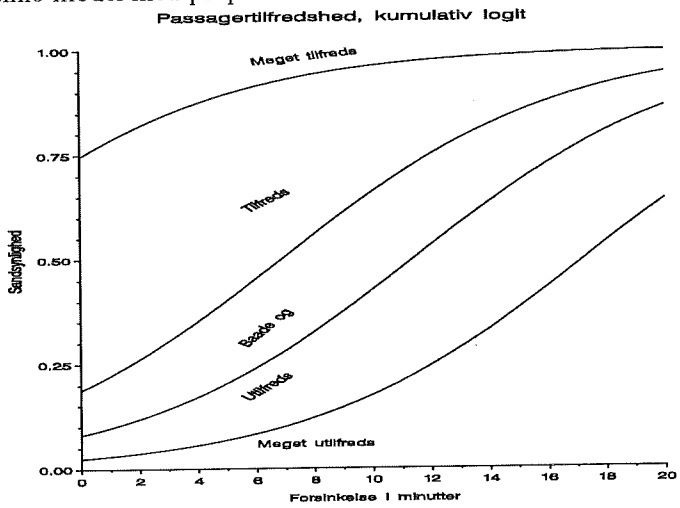
De estimerede relationer mellem de kumulative logit'er og forsinkelsen er illustreret i nedenstående figur.



De tilsvarende deviansresidualer er angivet i nedenstående tabel

	Deviansresidualer			
Forsinkelse min.	Meget utilfr.	Utilfr.	Både og	Tilfr.
0	-1.2033	-1.3078	0.3815	1.4672
2	-0.0929	0.7742	-0.5056	-2.4886
5	0.9888	1.1083	-0.5769	-2.1954
7	2.1306	2.0827	-0.1781	-2.2394

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.

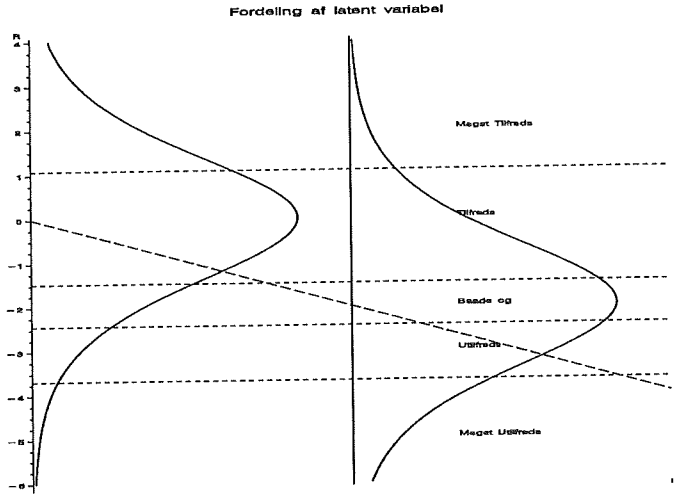


□

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 klassifikationskriterer, Yule's krydsprodukt-ratio

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

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

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

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

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

$$\omega^{A;B}_{i_1,i_2;j_1,j_2} = \frac{\omega^{A;B}_{i_1;j_1,j_2}}{\omega^{A;B}_{i_2;j_1,j_2}} = \frac{p_{i_1,j_1}}{p_{i_1,j_2}} \Big/ \frac{p_{i_2,j_1}}{p_{i_2,j_2}} = \frac{\mathsf{P}\,[A_{i_1,j_1}]}{\mathsf{P}\,[A_{i_1,j_2}]} \Big/ \frac{\mathsf{P}\,[A_{i_2,j_1}]}{\mathsf{P}\,[A_{i_2,j_2}]} \tag{4.3.1}$$

hvor vi har indført symbolet

$$\omega^{A;B}_{i_1;j_1,j_2} = \frac{p_{i_1,j_1}}{p_{i_1,j_2}}$$

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

$$\begin{aligned} \Delta^{A;B}_{i_1,i_2;j_1,j_2} &= \ln(\omega^{A;B}_{i_1,i_2;j_1,j_2}) = \frac{p_{i_1,j_1}}{p_{i_1,j_2}} \Big/ \frac{p_{i_2,j_1}}{p_{i_2,j_2}} \\ &= \frac{\mathsf{P}\,[A_{i_1,j_1}]}{\mathsf{P}\,[A_{i_1,j_2}]} \Big/ \frac{\mathsf{P}\,[A_{i_2,j_1}]}{\mathsf{P}\,[A_{i_2,j_2}]} \end{aligned} \tag{4.3.2}$$

Vi har

$$\Delta^{A;B}_{i_1,i_2;j_1,j_2} = \Delta^{B}_{i_1;j_1,j_2} - \Delta^{B}_{i_2;j_1,j_2} = \Delta^{A}_{i_1,i_2}\Delta^{B}_{i;j_1,j_2}$$

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

$$\omega^{*}_{i;j} = \omega_{i,i+1;j,j+1} = \frac{p_{i,j}}{p_{i,j+1}} \Big/ \frac{p_{i+1,j}}{p_{i+1,j+1}}\ ,\ i = 1,\ldots,r-1,\ j = 1,\ldots,s-1 \tag{4.3.3}$$

En anden minimal mængde fås som

$$\omega^{\diamond}_{i;j,} = \omega_{i,r;j,s} = \frac{p_{i,j}}{p_{i,s}} \Big/ \frac{p_{r,j}}{p_{r,s}}\ ,\ i = 1,\ldots,r-1,\ j = 1,\ldots,s-1 \tag{4.3.4}$$

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

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

og

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

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.

Omend den anvendte stikprøvemodel ikke altid fremgår af en given antalstabel $\{x_{ij}\}$, 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, $\mathrm{Mult}(N, p_{1,1},\ldots,p_{1,J},\ldots,p_{I,1},\cdots,p_{I,J})$

$$f(\mathbf{x}) = \frac{N!}{\Pi_{i,j} x_{i,j}!} \prod_{i,j} p_{i,j}^{x_{i,j}} \tag{4.3.5}$$

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

$$l(\mathbf{p};\mathbf{x}) = \frac{N!}{\Pi_{i,j} x_{i,j}!} \sum_{i,j} x_{i,j} \ln(p_{i,j}) \tag{4.3.6}$$

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

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

En naturlig reduktion af modellen er, at antage, at der er uafhængighed mellem de to klassifikationer, dvs. at

$$H_0 : p_{i,j} = p_{i,+} p_{+,j} \quad \text{for alle} \quad (i,j) \tag{4.3.7}$$

Der gælder

Sætning 4.3.1 ***Uafhænghedshypotese for todimensional tabel***

Under den multinomiale stikprøvemodel gælder, at $0 < p_{i,j} = p_{i,\cdot} p_{\cdot,j}$ for alle (i,j) hvis og kun hvis

$$\frac{p_{i_1 j_1} p_{i_2 j_2}}{p_{i_1 j_2} p_{i_2 j_1}} = \omega^{A;B}_{i_1,i_2;j_1,j_2} = 1 \tag{4.3.8}$$

for alle $i_1, i_2 \in (1,2,\ldots,I)^2$ 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

$$\omega^{A;B}_{1,2;2,3} = \frac{p_{1,2} p_{2,3}}{p_{1,3} p_{2,2}} = 1$$

og

$$\omega^{A;B}_{1,2;2,4} = \frac{p_{1,2} p_{2,4}}{p_{1,4} p_{2,2}} = 1 \,,$$

da vil også

$$\omega^{A;B}_{1,2;3,4} = \frac{p_{1,4}p_{2,3}}{p_{1,3}p_{2,4}} = \frac{p_{1,2}p_{2,3}}{p_{1,3}p_{2,2}} \Big/ \frac{p_{1,2}p_{2,4}}{p_{1,4}p_{2,2}} = 1$$

Generelt gælder, at

$$\omega^{A;B}_{i_1,i_2;j_1,j_2} = 1$$

for alle i_1, i_2 og j_1, j_2, hvis og kun hvis

$$\omega^{A;B}_{1,i;1,j} = \frac{p_{11}p_{i,j}}{p_{1j}p_{i1}} = 1$$

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 ***Uafhænghedshypotese for todimensional tabel ved kanoniske parametre***

Under den multinomiale stikprøvemodel gælder, at $0 < p_{i,j} = p_{i,\cdot}p_{\cdot,j}$ for alle (i,j) hvis og kun hvis log odds kan udtrykkes på formen

$$\vartheta_{i,j} = \alpha_i + \beta_j \tag{4.3.9}$$

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

$$p_{i,j} = \lambda_{i,j}/\lambda_{+,+}$$

svarende til

$$\vartheta_{i,j} = \ln(\lambda_{i,j}) - \ln(\lambda_{+,+})$$

Dette indebærer, at såfremt man blot er interesseret i relative forskelle under Poisson-modellen (4.1.1), dvs. ratioer

$$\omega_{(i,j),(i',j')} = \lambda_{i,j}/\lambda_{i',j'} = p_{i,j}/p_{i',j'} ,$$

svarende til kontraster

$$\Delta_{(i,j),(i',j')} = \ln(\lambda_{i,j}) - \ln(\lambda_{i',j'}) = \vartheta_{i,j} - \vartheta_{i',j'} ,$$

da vil multinomialmodellen være adækvat for den statistiske analyse. Det skal dog allerede her bemærkes, at for at de asymptotiske egenskaber for test skal være gyldige, kræves at den absolutte total $x_{+,+} \to \infty$.

□

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

$$\ln\left(\lambda_{i,j,k,\ldots}\right) = \mu + \alpha_i^A + \beta_j^B + \gamma_k^C + \ldots + (\alpha\beta)_{i,j}^{AB} + \ldots + (\alpha\beta\gamma)_{i,j,k}^{ABC} + \ldots \tag{4.4.1}$$

Modelstrukturen svarer til de sædvanlige modeller for faktorforsøg.

For at undgå overparametrisering kan man eksempelvis vælge at sætte $\alpha_1^A = \beta_1^B = \cdots = 0$ og $(\alpha\beta)_{1,j}^{AB} = (\alpha\beta)_{i,1}^{AB} = \cdots = 0$ etc. i overensstemmelse med betragtningerne i afsnit 2.9.3 □

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 α_i^A og β_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

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

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

$f_{X,Y,Z}(x,y,z) = g(x,y)h(x,y)$ for alle y og z og alle x 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 afknuder, 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.

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 antalstabel*

Betragt en antalstabel svarende til klassifikationerne $A, B, \ldots, H$. Ved uafhængighedsgrafen svarende til denne antalstabel forstås en graf med knuderne $1, 2, \ldots, r$ repræsenterende de r klassifikationer, og sådan, at hvis to klassifikationer er stokastisk uafhængige givet resten af tabellen, da er der 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 klasserne 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 beskrevej ved $I, JogK$.

Frekvensfunktionen svarende til denne model er multinomialfordelingen (Polynomialfordelingen)

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

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

De forventede værdier af celleværdierne er

$$\mathsf{E}\left[X_{i,j,k}\right] = N p_{i,j,k} \ .$$

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 $X_{\cdot,\cdot,\cdot} = N$ være en multinomial-fordeling med $p_{i,j,k} = \lambda_{i,j,k}/\lambda_{\cdot,\cdot,\cdot}$.

Log-likelihoodfunktionen svarende til multinomialfordelingsmodellen er

$$l(\mathbf{p}) = \frac{N!}{\Pi_{i,j,k} x_{i,j,k}!} \sum_{i,j,k} x_{i,j,k} \ln(p_{i,j,k}) \tag{4.6.2}$$

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

Vi har følgende principielt forskellige hypoteser:

$$H_0 : p_{ijk} = p_i^A p_j^B p_k^C \tag{4.6.3}$$

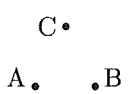
$$H_{1a} : p_{ijk} = p_i^A p_{jk}^{BC} \tag{4.6.4}$$

$$H_{2a} : p_{ijk} = \frac{p_{ij}^{AB} p_{ik}^{AC}}{p_i^A} \tag{4.6.5}$$

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

$$H_0 : A \perp B \perp C$$

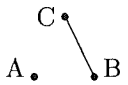
Uafhængighedsgrafen er:



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

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

Uafhængighedsgrafen er:



Ved permutation af indices får man de tilsvarende hypoteser:

$$H_{1b} : p_{ijk} = p_j^B p_{ik}^{AC}$$

$$H_{1c} : p_{ijk} = p_k^C p_{ij}^{AB}$$

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 hypotese 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

$$H_{2b}: \; p_{ijk} = \frac{p_{ij}^{AB} p_{jk}^{BC}}{p_j^B}$$

$$H_{2c}: \; p_{ijk} = \frac{p_{ik}^{AC} p_{jk}^{BC}}{p_k^C}$$

svarende til rækker og lag uafhængige for givet søjle ($H_{2b}: \; A \perp C|B$) og at rækker og søjler 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 tredie klassifikation.

Det skal bemærkes, at hypoteser af formen H_{2a} ikke indebærer, at B og C er uafhængige. Selv om hypotesen 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)

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

ikke afhænger af laget k.

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

$$\omega_{i_1,i_2;j;k_1,k_2}^{A;C} = \frac{p_{i_1,j,k_1}}{p_{i_1,,k_2}} \Big/ \frac{p_{i_2,j,k_1}}{p_{i_2,j,k_2}}$$

ikke afhænger af søjlen j.

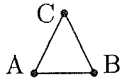
Ved i stedet at ombytte lag med rækker kan hypotesen endelig udtrykkes ved, at

$$\omega^{B;C}_{i;j_1,j_2;k_1,k_2} = \frac{p_{i,j_1,k_1}}{p_{i,j_2,k_2}} \Big/ \frac{p_{i,j_1,k_2}}{p_{i,j_2,k_2}}$$

ikke afhænger af rækken i.

Dvs. andenordens differenser $\Delta^A_{i_1,i_2}\Delta^B_{\cdot;j_1,j_2}$ er konstante, svarende til, at tredieordensdifferenser er nul.

Uafhængighedsgrafen svarende til denne hypotese er:



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

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

som et additivt udtryk.

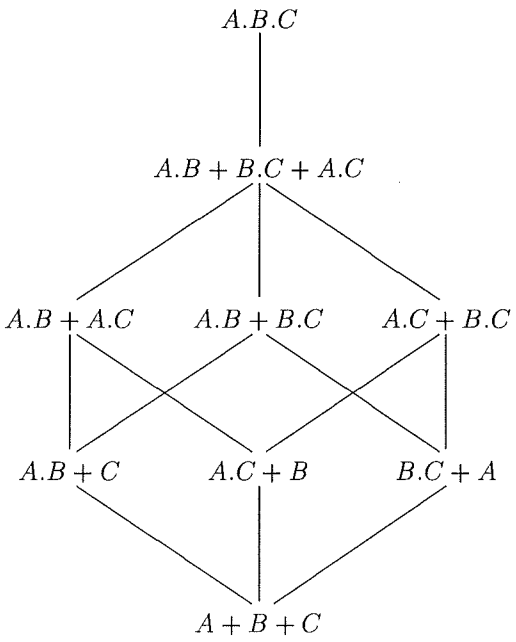
Vi skriver dem først på den mættede (saturated) form

$$\begin{aligned}
H_0 &: \eta_{i,j,k} = \mu + \alpha^A_i + \beta^B_j + \gamma^C_k \\
H_{1a} &: \eta_{i,j,k} = \mu + \alpha^A_i + \beta^B_j + \gamma^C_k + (\beta\gamma)^{BC}_{j,k} \\
H_{1b} &: \eta_{i,j,k} = \mu + \alpha^A_i + \beta^B_j + \gamma^C_k + (\alpha\gamma)^{AC}_{i,k} \\
H_{1c} &: \eta_{i,j,k} = \mu + \alpha^A_i + \beta^B_j + \gamma^C_k + (\alpha\beta)^{AB}_{i,j} \\
H_{2a} &: \eta_{i,j,k} = \mu + \alpha^A_i + \beta^B_j + \gamma^C_k + (\alpha\beta)^{AB}_{i,j} + (\alpha\gamma)^{AC}_{i,k} \\
H_{2b} &: \eta_{i,j,k} = \mu + \alpha^A_i + \beta^B_j + \gamma^C_k + (\alpha\beta)^{AB}_{i,j} + (\beta\gamma)^{BC}_{j,k} \\
H_{2c} &: \eta_{i,j,k} = \mu + \alpha^A_i + \beta^B_j + \gamma^C_k + (\alpha\gamma)^{AC}_{i,k} + (\beta\gamma)^{BC}_{j,k} \\
H_3 &: \eta_{i,j,k} = \mu + \alpha^A_i + \beta^B_j + \gamma^C_k + (\alpha\beta)^{AB}_{i,j} + (\beta\gamma)^{BC}_{j,k} + (\alpha\gamma)^{AC}_{i,k}
\end{aligned}$$

Vi kan fjerne overparametriseringen og opskrive modellerne ved frembringerne, angivet til højre i nedenstående oversigt.

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

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 model, $A.B.C$ symboliserer den fulde model, hvor hver celle har sin egen middelværdi:



Eksempel 4.6.1 ***Simpsons paradox, the National Halothane study***

Andel overlevende ved to behandlinger

	Behandling		
Køn	I	II	Ialt
Mænd	60/80	100/150	160/230
Kvinder	40/120	10/40	50/160
	100/200	110/190	210/390

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 aggregring 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 betragter den partielle korrelationskoefficient mellem to (blandt 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 ρ_{12}, og vi kan teste hypotesen $\rho_{12|3} = 0$ ved at teste $\rho_{12} = 0$.

I en tredimensional antalstabel kan 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 Grafiske modeller

Definition 4.7.1 *Grafisk model*

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

□

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

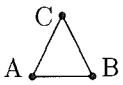
Ved analyse af flerdimensionale antalstabeller vil det ofte være ønskeligt at indskrænke sig til grafiske modeller, da sådanne modeller kan fortolkes ved relationer om betinget uafhængighed af visse variable, givet resten af de variable.

Eksempel 4.7.1 *Vekselvirkningsgraf for trefaktormodel*

Betragt 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



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 $(\alpha\beta\gamma)_{ijk}^{ABC}$.

□

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 grafen, 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 kliker. De irreducible komponenter er fuldstændige.

- 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 estimaterne 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 parametrer, 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.

Wermuth og Lauritzen (1990) giver en diskussion af strategier for modelvalg i antalstabeller med flerdimensionalt respons.

Vi anfører endelig en helt anden angrebsvinkel til modellering af antalstabeller med flerdimensionalt respons, nemlig den s

kaldte korrespondanceanalyse se fx Greenacre (1984). Korrespondanceanalysen 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 ϕ i F lader vi I_ϕ angive mængden af niveauer for faktoren ϕ.

Mængden af celler i antalstabellen er mængden

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

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

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

4.8.1 Relation til teorien for Markovfelter

Teorien for endelige Markovfelter opererer med en endelig mængde Γ af *sites*. Mængden af sites svarer her til mængden af faktorer.

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

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

kaldes mængden af konfigurationer. En given konfiguration betegnes med $\mathbf{i} = (\mathbf{i}_\gamma, \gamma \in \Gamma)$.

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

To knuder α og β siges at være naboer (adjacent), hvis og kun hvis $\{\alpha, \beta\} \in E(\mathbf{\Gamma})$. I dette tilfælde skriver vi $\alpha \sim \beta$.

Lad $a \subseteq \Gamma$. Randen af a, δa er mængden af kanter i $\Gamma \backslash a$, som er naboer til en knude i a. Afslutningen af a er mængden $\bar{a} = a \cup \delta a$.

En fuldstændig delmængde er en delmængde $a \subseteq \Gamma$ hvor alle elementer er indbyrdes naboer. En klike er en maksimal fuldstændig delmængde.

Betragt nu en sandsynlighed P på I sådan at $P(\mathbf{i}) > 0$ foralle $\mathbf{i} \in I$, og betragt de stokastiske variable defineret ved koordinatprojektionerne:

$$\begin{aligned} X_\gamma(\mathbf{i}) &= i_\gamma, \quad \text{for } \gamma \in \Gamma \\ X_a(\mathbf{i}) &= \mathbf{i}_a \quad \text{for } a \subseteq \Gamma,\ a \neq \emptyset \end{aligned}$$

Definition 4.8.1 ***Markov felt***

Det stokastiske felt $(X_\gamma, \gamma \in \Gamma)$ siges at være et Markov felt med hensyn til P og $\mathbf{\Gamma}$ (eller P er Markov med hensyn til $\mathbf{\Gamma}$) hvis en af nedensstående ækvivalente betingelser er opfyldt:

a) For ethvert $\gamma \in \Gamma$ er X_γ og $X_{\Gamma - \bar{\gamma}}$ betinget uafhængige givet $X_{\delta\gamma}$

b) For ethvert par $\alpha, \beta \in \Gamma$ med $\alpha \mathtt{nsim}\ \beta$, er X_α og X_β betinget uafhængige givet $X_{\Gamma \backslash \{\alpha,\beta\}}$

c) For ethvert $a \subseteq \Gamma$ er X_a og $X_{\Gamma \backslash \bar{a}}$ betinget uafhængige givet $X_{\delta a}$

d) Hvis to disjunkte delmængder $a \subseteq \Gamma$ og $b \subseteq \Gamma$ er separeret af en delmængde $d \subseteq \Gamma$, sådan at enhver sti fra a til b i Γ går gennem d, da er X_a og X_b betinget uafhængige givet X_d.

□

Bevis for ækvivalensen af de fire betingelser, se f.eks. Kemeny, Snell og Knapp (1976).

Et potential er en reel funktion, Φ på I af formen

$$\Phi(\mathbf{i}) = \sum_{a \subseteq \Gamma} \xi_a(\mathbf{i}_a)$$

hvor funktionerne ξ_a alene afhænger af $\mathbf{i}$ gennem $\mathbf{i}_a$. Funktionerne ξ_a kaldes vekselvirkningspotentialerne.

Et sandsynlighedsmål P på I kaldes en Gibbs tilstand med potential Φ, såfremt der gælder

$$P(\mathbf{i}) = \exp(\Phi(\mathbf{i}))$$

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

Φ 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 Γ.

Et sandsynlighedsmål P kaldes en nærmeste-nabo Gibbs tilstand hvis og kun hvis P er en Gibbs tilstand med potential Φ, hvor Φ 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 ***Ækvivalens mellem nærmeste nabo Gibbs-tilstand og Markovfelter***

Et sandsynlighedsmål 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) □

Sætningen knytter en forbindelse mellem lineære bånd på logaritmen til et sandsynlighedsmål (nemlig, at det er en nærmeste nabo Gibbs-tilstand) og en Markov-egenskab, (nemlig en fortolkning udtrykt ved betinget uafhængighed). Det er denne forbindelse, vi benytter, blandt andet ved formuleringer af modeller for antalstabeller.

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 $\mathbf{F}$ på faktorerne, specificeret ved mængden $V(\mathbf{F})$ af knuder, og mængden $E(\mathbf{F})$ af kanter.

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

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

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

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

Grafen $\mathbf{F}$ 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 $\mathbf{F}$.

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

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

Lad $\mathcal{F}$ være en frembringende klasse og sæt $F = \cup_{f \in \mathcal{F}}$. Vi kan da definere en graf $\mathbf{F} = (V(\mathbf{F}), E(\mathbf{F}))$ ved at sætte $V(\mathbf{F}) = F$ og definere en kant ved at der er førsteordensvekselvirkning mellem de indgående faktorer, dvs. $\{f_1, f_2\} \in E(\mathbf{F})$ 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 vekselvirkninger 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 vekselvirkningsstrukturen i modellen kan ikke beskrives fuldstændigt alene ved grafen.

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

4.9 Referencer

Agresti, A. (1990): *Categorical Data Analysis*, Wiley, New York

Amemiya, T. (1975): Qualitative Response Models. *Annals of Economic and Social Measurement*, **4**, pp 363-372

Amemiya, T. (1981): Qualitative Response Models: A Survey. *Journal of Economic Literature* **XIX**, pp. 1483-1536.

Andersen, E. B. (1990): *The Statistical Analysis of Categorial Data*, Springer Verlag, Heidelberg

Ben-Akiva, M. and Lerman, S. (1985): *Discrete Choice Analysis: Theory and Application to Travel Demand.* The MIT Press, Cambridge, Massachusetts.

Bhapker, V. P. and Koch, Gary, G. (1968): Hypothesis of "No Interaction" in Multidimensional Contingency Tables. *Technometrics*, **10**. pp 107-123.

Brownstone, D. and Small, K. A. (1989): Efficient Estimation of Nested Logit Models, *Journ. of Business & Economic Statistics*, **7**, pp. 67- 74.

Christensen, R. (1990): *Log-Linear Models*, Springer Verlag, New York

Green, P. E. and Srinivasan, V. (1990): Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. *Journal of Marketing*, **54**, pp 3-19.

Greenacre, M. J. (1984): *Theory and Applications of Correspondance Analysis*, Academic Press, London

Kemeny, J.G., Snell, J.L. and Knapp, A.W. (1976): *Denumerable Markov Chains*, nd ed. Springer-Verlag, Heidelberg, New York Berlin.

Luce, R. D. (1959): *Individual Choice Behavior: A theoretical Analysis*, Wiley, New York

Luce, R. D. (1977): The Choice Axiom After 20 Years. *Journal of Mathematical Psychology*, **15**, pp 215-233.

Manski, C. and McFadden, D. (1981): Alternative Estimates and Discrete Choice Analysis. pp 2-50 in *Structural Analysis of Discrete Data*, C. Manski and D. McFadden eds. The MIT Press, Cambridge, Massachusetts.

McCullagh, P. (1980): Regression models for ordinal data (with discussion). *J.Roy.Statist.Soc.* **B 42**, pp. 109-142

McFadden, D. (1984): Qualitative response models. Chapter 1 in *Advances in Econometrics*. Hildenbrand, W. editor. Cambridge University Press, Cambridge.

Simpson, E. H. (1951): The interpretation of interaction in contingency tables. *J.Roy.Statist.Soc.* **B 13**, pp. 238-241.

Speed, T.P. (1978): Relations between models for spatial data, contingency tables and Markov fields on graphs. *Suppl. Adv. Appl. Prob.* **10**, pp 111-122.

Wermuth, N. and Lauritzen, S. L. (1990): On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models. (with discussion) *J.R. Statist.Soc.B* **52**, pp 21-72

J.Whittaker (1990): *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester.

Yule, G. U. (1903): Notes on the theory of association of attributes in statistics. *Biometrika*, **2**, pp. 121-134.

Afsnit 5

Hierarkiske modeller for endimensionale normalfordelinger

fil: normhier.tex 1998-04-19

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 tilstrækbt rimeligt ensartede inden for hver af grupperne, men hvor omstændighederne antages at variere fra gruppe til gruppe.

Vi repræsenterer observationerne ved skemaet

Gruppe	Observationer
1	$X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$
2	$X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$
$\vdots$	$\vdots, \vdots$
k	$X_{k,1}, X_{k,2}, \ldots, X_{k,n_k}$

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,

$$\overline{X}_{i\cdot} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \tag{5.1.1}$$

og variabiliteten af disse observationer kan udtrykkes ved kvadratafvigelsessummen for den i'te række

$$\sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2,$$

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

$$SAK_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\cdot})^2 \tag{5.1.2}$$

som udtryk for variationen indenfor grupper.

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

$$\overline{\overline{X}}_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}/N = \sum_{i=1}^{k} n_i \overline{X}_{i\cdot} / \sum_{i=1}^{k} n_i \tag{5.1.3}$$

hvor N angiver det totale antal observationer,

$$N = \sum_{i=1}^{k} n_i \tag{5.1.4}$$

Vi bemærker, at det fælles gennemsnit kan beregnes som det vejede gennemsnit af gruppegennemsnitterne med de respektive observationsantal som vægte.

Gruppegennemsnittenes variabilitet udtrykkes naturligt ved kvadratafvigelsessummen

$$SAK_2 = \sum_{i=1}^{k} n_i(\overline{X}_{i\cdot} - \overline{\overline{X}}_{\cdot\cdot})^2 \tag{5.1.5}$$

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

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{\cdot\cdot})^2 = SAK_1 + SAK_2 \tag{5.1.6}$$

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

$$\begin{aligned} SAK_1 &= \sum_{i=1}^{k} SK_i - \sum_{i=1}^{k} S_i^2/n_i \\ SAK_2 &= \sum_{i=1}^{k} S_i^2/n_i - \left(\sum_{i=1}^{k} S_i\right)^2 /N \end{aligned}$$

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 $\overline{X}_{1\cdot}, \overline{X}_{2\cdot}, \ldots, \overline{X}_{k\cdot}$ er indbyrdes uafhængige med samme forventningsværdi $\mathsf{E}[\overline{X}_{i\cdot}] = \mu$, gælder

$$\mathsf{E}\,[SAK_2] = \sum_{i=1}^{k} n_i(1-w_i)\mathsf{V}\,[\overline{X}_{i\cdot}] \tag{5.1.7}$$

med stikprøveandelen w_i givet ved

$$w_i = n_i/N \tag{5.1.8}$$

hvor $N = \sum_i n_i$

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

$$n_0 = \frac{\sum_1^k n_i - (\sum_1^k n_i^2 / \sum_1^k n_i)}{k-1} = \left(N - \frac{\sum_i n_i^2}{N}\right)\Big/(k-1) \tag{5.1.9}$$

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 balancerede 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 Ensidet variansanalyse i den systematiske 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

$$X_{ij} \in \mathrm{N}(\mu_i, \sigma^2) \tag{5.2.1}$$

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,

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij} \ , \tag{5.2.2}$$

hvor

$$\epsilon_{ij} \in \mathrm{N}(0, \sigma^2) \tag{5.2.3}$$

er indbyrdes uafhængige.

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

$$\sum_{i=1}^{k} n_i \alpha_i = 0 \tag{5.2.4}$$

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

Sætning 5.2.1 ***Den systematiske model for ensidet variansanalyse***

For modellen givet ved (5.2.1) gælder:

$$\begin{aligned} \mathsf{E}\,[X_{ij}] &= \mu_i \\ \mathsf{COV}[X_{ij}, X_{hl}] &= \begin{cases} \sigma^2 & \text{for } (i,j) = (h,l) \\ 0 & \text{ellers} \end{cases} \end{aligned} \tag{5.2.5}$$

samt

$$\begin{aligned} \mathsf{E}\,[\overline{X}_{i\cdot}] &= \mu_i \\ \mathsf{COV}[\overline{X}_{i\cdot}, \overline{X}_{h\cdot}] &= \begin{cases} \sigma^2/n_i & \text{for } i = h \\ 0 & \text{ellers} \end{cases} \end{aligned} \tag{5.2.6}$$

Endelig har man

$$SAK_1 \in \sigma^2\chi^2(N-k) \tag{5.2.7}$$

og

$$SAK_2 \in \sigma^2\chi^2(k-1,\lambda) \tag{5.2.8}$$

hvor SAK_1 og SAK_2 er stokastisk uafhængige, og hvor ikke-centralitetsparameteren λ er givet ved

$$\lambda = \frac{\sum_i n_i(\mu_i - \mu_0)^2}{(k-1)\sigma^2} \tag{5.2.9}$$

med

$$\mu_0 = \sum_{i=1}^{k} n_i\mu_i \Big/ \sum_{i=1}^{k} n_i \tag{5.2.10}$$

Bevis:
Sætningen bevises direkte

□

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. Niveauforskellen mellem grupperne i og h er $\mu_i - \mu_h$, som estimeres ved $\overline{X}_{i\cdot} - \overline{X}_{h\cdot}$.

Der gælder:

Sætning 5.2.2 *Sammenligning af grupper i den systematiske model for ensidet variansanalyse*

For modellen givet ved (5.2.1) gælder:

$$\mathsf{E}\,[\overline{X}_{i\cdot} - \overline{X}_{h\cdot}] = \mu_i - \mu_h \tag{5.2.11}$$

(5.2.12)

$$\mathsf{V}\,[\overline{X}_{i\cdot} - \overline{X}_{h\cdot}] = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_h}\right) \quad \text{for } i \neq h \tag{5.2.13}$$

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

$$\mathrm{H}_I : \mu_1 = \cdots = \mu_k = 0 \tag{5.2.14}$$

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

Hypotesen er en homogenitetshypotese i en generaliseret lineær model (se nedenstående bemærkning 1). Det er velkendt, blandt andet fra Introduktion til Statistik, Bind 1, at denne hypotese kan testes ved at spalte 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*

$$\text{lad} \quad Z = \frac{SAK_2/(k-1)}{SAK_1/(N-k)} \tag{5.2.15}$$

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

$$Z \in \mathrm{F}(k-1, N-k, \lambda), \tag{5.2.16}$$

hvor ikke-centralitetsparameteren λ er bestemt ved (5.2.9)

Kvotienttesttørrelsen for hypotesen (5.2.14) har det kritiske område:

$$C = \{z | z > F(k-1, N-k)_{1-\alpha}\} \tag{5.2.17}$$

Bevis:
Se f.eks. Scheffé (1959) p. 55.

□

I praksis udføres testet ved at betragte variansanalyseskemaet svarende til opspaltningen (5.1.6)

Variation	SAK	f	E $[SAK/f]$
Mellem grupper	$\sum_i n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$	$k-1$	$\sigma^2 + \frac{1}{(k-1)} \sum_i n_i(\mu_i - \mu_0)^2$
Indenfor grupper	$\sum_i \sum_j (X_{ij} - \overline{X}_{i\cdot})^2$	$N-k$	σ^2
Total	$\sum_i \sum_j (X_{ij} - \overline{X}_{\cdot\cdot})^2$	$N-1$	

hvor μ_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 reservere symbolet $\mathbf{X}$ til modelmatricen.

Opfatter vi observationerne $\mathbf{Y}$ som en søjlevektor har vi, at den systematiske model svarer til en generaliseret lineær model for $\mathbf{Y}$, hvor $Y_{ij} \in \mathrm{N}(\mu_i, \sigma^2)$, dvs. at linkfunktionen er den identiske afbildning (den kanoniske link for normalfordelingen) og dispersionsparameteren er σ^2.

I analogi med formuleringen som en generaliseret lineær model vælger vi at benytte symbolet $\boldsymbol{\beta}$ for parameteren. Da linkfunktionen er identiten har vi altså $\eta \equiv \mu$, dvs. $\beta_i \equiv \mu_i$, dvs

$$H_0 : \mathsf{E}[Y_{ij}] = \beta_i \tag{5.2.18}$$

uden bånd på β-værdierne.

Dispersionsmatricen for søjlevektoren $\mathbf{Y}$ er

$$\mathbf{D}[\mathbf{Y}] = \sigma^2 \mathbf{I}\,, \tag{5.2.19}$$

hvor $\mathbf{I}$ angiver den $N \times N$ dimensionale enhedsmatrix.

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

Middelværdiligningen (2.5.6) på side 178 bliver derfor

$$\mathbf{U}^T \boldsymbol{\mu}(\boldsymbol{\beta}) = \mathbf{U}^T \mathbf{y}$$

der ved indsættelse af (5.2.18)

$$\boldsymbol{\mu}(\boldsymbol{\beta}) = \mathbf{U}\boldsymbol{\beta}$$

bliver

$$\mathbf{U}^T \mathbf{U}\boldsymbol{\beta} = \mathbf{U}^T \mathbf{y}\,. \tag{5.2.20}$$

Idet

$$\mathbf{U}^T \mathbf{U} = \mathrm{diag}\{n_i\}$$

har man

$$[\mathbf{U}^T \mathbf{U}]^{-1} = \mathrm{diag}\{n_i^{-1}\}$$

hvorfor

$$\widehat{\boldsymbol{\beta}} = [\mathbf{U}^T \mathbf{U}]^{-1} \mathbf{U}^T \mathbf{y} = \mathrm{diag}\{n_i^{-1}\} \mathbf{U}^T \mathbf{y}$$

Man har derfor parameterestimaterne

$$\widehat{\beta}_i = \overline{y}_{i+}$$

Man kan vise, at hat-matricen $\mathbf{H} = \mathbf{U}[\mathbf{U}^T\mathbf{U}]^{-1}\mathbf{U}^T$ kan udtrykkes som

$$\mathbf{H} = \mathbf{U}[\mathbf{U}^T\mathbf{U}]^{-1}\mathbf{U}^T = \text{Blok diag}\{n_i^{-1}\mathbf{J}_{n_i}\}\,, \tag{5.2.21}$$

hvor Blok diag betyder en blokdiagonal matrix, og hvor $\mathbf{J}_n$ angiver en $n \times n$ matrix med lutter ettaller.

De fittede værdier er således

$$\widehat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$$

Hypotesen (5.2.14) formuleres her som

$$H_M : \mu_1 = \mu_2 = \cdots = \mu_k \ (= \alpha) \tag{5.2.22}$$

dvs med modelmatricen svarende til den konstante faktor,

$$\mathbf{U}_M = \mathbf{1}_N$$

hvorfor middelværdiligningen bliver

$$\mathbf{1}_N^T\boldsymbol{\mu}(\alpha) = \mathbf{1}_N^T\mathbf{y}$$

dvs

$$\mathbf{1}_N^T\mathbf{1}_N\alpha = \mathbf{1}_N^T\mathbf{y}$$

eller

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

med løsningen

$$\hat{\alpha} = [\mathbf{1}_N^T\mathbf{1}_N]^{-1}\mathbf{1}_N^T\mathbf{y} = \overline{\overline{y}}_{++} \tag{5.2.23}$$

De fittede værdier bliver

$$\widehat{\widehat{\boldsymbol{\mu}}} = \mathbf{H}_0\mathbf{y} \tag{5.2.24}$$

med hat-matricen

$$\mathbf{H}_0 = \frac{1}{N}\mathbf{J}_N$$

Deviansbidragene er netop de kvadratiske afvigelser

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

Man får derfor

$$G^2(H_0) = \mathbf{y}^T(\mathbf{I}_N - \mathbf{H})\mathbf{y} = \sum_i \sum_j (y_{ij} - \overline{y}_{i+})^2$$

og

$$\begin{aligned} G^2(H_M|H_0) &= \mathbf{y}^T(\mathbf{I}_N - \mathbf{H})\mathbf{y} - \mathbf{y}^T(\mathbf{I}_N - \mathbf{H}_0)\mathbf{y} \\ &= \mathbf{y}^T(\mathbf{H} - \mathbf{H}_0)\mathbf{y} \\ &= \sum_{i=1}^{k} n_i(\overline{y}_{i+} - \overline{y}_{++})^2 , \end{aligned}$$

altså netop kvadratafvigelsesssummen mellem grupper.

Udtrykt ved disse matricer bliver variansanalyseskemaet

Variation	f	SAK matrixform	SAK analytisk
Mellem grupper	$k-1$	$\mathbf{y}^T(\mathbf{H} - \mathbf{H}_0)\mathbf{y}$	$\sum_i n_i(\overline{y}_{i+} - \overline{y}_{++})^2$
Indenfor grupper	$N-k$	$\mathbf{y}^T(\mathbf{I}_N - \mathbf{H})\mathbf{y}$	$\sum_i \sum_j (y_{ij} - \overline{y}_{i+})^2$
Total	$N-1$	$\mathbf{y}^T(\mathbf{I}_N - \mathbf{H}_0)\mathbf{y}$	$\sum_i \sum_j (y_{ij} - \overline{y}_{++})^2$

Maksimum-likelihood estimatoren for σ^2 under H_0 bliver

$$\begin{aligned} (\sigma^*)^2 &= \frac{1}{N}\mathbf{y}^T(\mathbf{I}_N - \mathbf{H})\mathbf{y} \\ &= \frac{1}{N}\sum_i \sum_j (y_{ij} - \overline{y}_{i+})^2 \end{aligned} \qquad (5.2.25)$$

Det gælder imidlertid, at størrelsen $\mathbf{y}^T(\mathbf{I}_N - \mathbf{H})\mathbf{y}$ er likelihood-sufficient (def. 2.1.5). for σ^2.

Man vælger derfor at estimere σ^2 ud fra fordelingen af den likelihoodsufficiente størrelse.

Idet

$$\mathbf{y}^T(\mathbf{I}_N - \mathbf{H})\mathbf{y} \in \sigma^2\chi^2(N-k)$$

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

$$\widehat{\sigma}^2 = \frac{1}{N-k}\mathbf{y}^T(\mathbf{I}_N - \mathbf{H})\mathbf{y} = \frac{SAK_1}{N-k}$$

□

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

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

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

Vi antager nemlig, at

$$\mu_i \in \mathrm{N}(\mu_0, \sigma_0^2), \tag{5.3.2}$$

hvor μ_i'erne antages indbyrdes uafhængige, og Y_{ij} antages at være indbyrdes uafhængige i den betingede fordeling af Y_{ij} for givet μ_i.

I lighed med den systematiske model kan man også formulere denne model som en additiv model for observationerne Y_{ij} ved

$$Y_{ij} = \mu_0 + \alpha_i + \epsilon_{ij} \ , \tag{5.3.3}$$

med $\epsilon_{ij} \in \mathrm{N}(0, \sigma^2)$ og $\alpha_i \in \mathrm{N}(0, \sigma_0^2)$, hvor ϵ_{ij} er indbyrdes uafhængige, og α_i ligeledes er indbyrdes uafhængige og endelig er α_i'erne uafhængige af ϵ_{ij}.

Da det ofte vil være af interesse at betragte forholdet mellem de to varianser, indfører vi symbolet γ for dette forhold,

$$\gamma = \sigma_0^2/\sigma^2 \tag{5.3.4}$$

Parameteren γ udtrykker således inhomogeniteten mellem grupper i forhold til gruppernes interne variation.

Vi vil ofte bruge betegnelsen signal/støj forholdet for parameteren γ.

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, μ_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, $\mathrm{V}[\mu_i] = \sigma_0^2$. Den tilfældige model er et specialtilfælde af den mere generelle varianskomponentmodel, og σ_0^2, 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

$$\begin{aligned} \mathsf{E}\,[Y_{ij}|\mu_i] &= \mu_i \\ \mathsf{COV}[Y_{ij}, Y_{il}|\mu_i] &= \begin{cases} 0 & \text{for } j \neq l \\ \sigma^2 & \text{for } j = l \end{cases} \end{aligned} \tag{5.3.5}$$

og endvidere er

$$\mathsf{E}\,[\overline{Y}_{i\cdot}|\mu_i] = \mu_i, \quad \text{og} \quad \mathsf{V}\,[\overline{Y}_{i\cdot}|\mu_i] = \sigma^2/n_i \tag{5.3.6}$$

Bevis:
Beviset følger direkte af antagelserne □

Betragter 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

$$\begin{aligned} \mathsf{E}\,[Y_{ij}] &= \mu_0 \\ \mathsf{COV}[Y_{ij}, Y_{hl}] &= \begin{cases} \sigma_0^2 + \sigma^2 & \text{for } (i,j) = (h,l) \\ \sigma_0^2 & \text{for } i = h,\ j \neq l \\ 0 & \text{for } i \neq h \end{cases} \end{aligned} \tag{5.3.7}$$

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 μ_0, nemlig i retning mod den pågældende gruppemiddelværdi.

Korrelationskoefficienten

$$\rho = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} = \frac{\gamma}{1+\gamma} \tag{5.3.8}$$

der beskriver korrelationen indenfor gruppe, benævnes ofte intraklassekorrelationen. □

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

$$\mathbf{E}_{n_i} = (1-\rho)\mathbf{I}_{n_i} + \rho\mathbf{J}_{n_i} = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix} \tag{5.3.9}$$

hvor $\mathbf{J}_{n_i}$ er en $n_i \times n_i$-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 $Y_i \in \mathrm{N}_{n_i}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_{n_i} + \sigma_0^2\mathbf{J}_{n_i})$, dvs at dispersionsmatricen for Y_i er

$$\begin{aligned} \mathbf{V}_i &= \mathbf{D}[Y_i] = \mathsf{E}[(Y_i - \boldsymbol{\mu})(Y_i - \boldsymbol{\mu})^T] \\ &= \begin{pmatrix} \sigma_0^2 + \sigma^2 & \sigma_0^2 & \ldots & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 + \sigma^2 & \ldots & \sigma_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_0^2 & \sigma_0^2 & \ldots & \sigma_0^2 + \sigma^2 \end{pmatrix} \end{aligned} \tag{5.3.10}$$

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

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

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

$$\mathbf{V} = \mathbf{D}[\mathbf{Y}] = \text{Blok diag}\{\mathbf{V}_i\} \tag{5.3.11}$$

hvor $\mathbf{V}_i$ er givet ved (5.3.10)

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

Bemærkning 4 ***Simultan fordeling af gruppegennemsnittene***

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

$$\mathsf{COV}[\overline{Y}_{i\cdot}, \overline{Y}_{h\cdot}] = \begin{cases} \sigma_0^2 + \sigma^2/n_i & \text{for } i = h \\ 0 & \text{ellers} \end{cases} \tag{5.3.12}$$

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

$$\mathsf{V}\,[\overline{Y}_{i\cdot}] = \sigma_0^2 + \sigma^2/n_i = \sigma^2(\gamma + 1/n_i)$$

omfatter både variansen på den tilfældige komponent, γ_i, og residualvariansen på gennemsnittet.

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

□

Bemærkning 5 ***Indlejring af modellen i model for kvadratafvigelssessummer***

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 opspaltning af kvadratafvigelsessummen.

Under modellen (5.3.1) og (5.3.2) har vi

$$\mathsf{E}\,[SAK_1/(N-k)] = \sigma^2 \quad \text{og} \quad \mathsf{E}\,[SAK_2/(k-1)] = \sigma_0^2 + n_0\sigma^2$$

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

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

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

kan τ^2 og σ^2 estimeres uafhængigt af hinanden ved de respektive kvadratafvigelsessummer.

Vi betragter nu den mere generelle normalfordelingsmodel givet ved

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

med $\nu^2 > 0$ og λ vilkårlig.

Dispersionsmatricen svarende til den n_i-dimensionale observationsvektor $\mathbf{Y}_i$ fra den i'te gruppe med middelværdivektoren μ (en vektor bestående af n_i μ'er) bliver således

$$\mathbf{D}[\mathbf{Y}_i] = \mathrm{E}[(\mathbf{Y}_i - \mu)(\mathbf{Y}_i - \mu)^T] = \begin{pmatrix} \nu^2 & \lambda & \dots & \lambda \\ \lambda & \nu^2 & \dots & \lambda \\ \vdots & \vdots & \ddots & \vdots \\ \lambda & \lambda & \dots & \nu^2 \end{pmatrix}$$

altså en equikorrelationsmatrix af formen

$$\mathbf{D}_i = (\nu^2 - \lambda)\mathbf{I}_{n_i} + \lambda \mathbf{J}_{n_i}$$

Idet

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

kan parametrene σ^2 og τ^2 udtrykkes ved ν^2 og λ 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 λ

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

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 *Momentestimater (variansanalyseestimater) i den tilfældige model*

Under modellen givet ved (5.3.1) og (5.3.2) finder man momentestimaterne for parametrene μ_0, σ^2 og σ_0^2 ved

$$\begin{aligned} \widetilde{\mu}_0 &= \overline{\overline{Y}}_{..} \\ \widetilde{\sigma^2} &= SAK_1/(N-k) \\ \widetilde{\sigma_0^2} &= \frac{SAK_2/(k-1) - SAK_1/(N-k)}{n_0} = \frac{SAK_2/(k-1) - \widetilde{\sigma^2}}{n_0} \end{aligned} \tag{5.3.14}$$

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

Bevis:

$$\begin{aligned} \mathrm{E}\left[SAK_1/(N-k)\right] &= \sigma^2 \\ \text{og} & \\ \mathrm{E}\left[SAK_2/(k-1)\right] &= \sigma^2 + n_0\sigma_0^2 \end{aligned}$$

□

Bemærkning 1 *Trunkering af variansestimat til ikke-negative værdier*

I stedet for den centrale estimator (5.3.14) benyttes ofte

$$\widehat{\sigma}_0^2 = \max\{\widetilde{\sigma_0^2}, 0\} \tag{5.3.15}$$

da vi jo har, at $\sigma_0^2 > 0$. □

Sætning 5.3.4 *Varianser for momentestimatorer i den tilfældige model*

Estimatorerne $\widetilde{\mu}, \widetilde{\sigma^2}$ og $\widetilde{\sigma_0^2}$ givet ved (5.3.14) er centrale, og varianserne for estimatorerne er givet ved

$$\begin{aligned} \mathsf{V}[\widetilde{\sigma^2}] &= \frac{2\sigma^4}{N-k} \\ \mathsf{V}[\widetilde{\sigma_0^2}] &= \frac{2\sigma^4}{n_0^2} A \end{aligned}$$

samt (5.3.16)

$$\mathsf{COV}[\widetilde{\sigma^2}, \widetilde{\sigma_0^2}] = -\frac{\mathsf{V}[\widetilde{\sigma^2}]}{n_0} \quad (5.3.17)$$

med

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

hvor

$$w_i(\gamma) = \frac{1}{1+n_i\gamma} \quad (5.3.18)$$

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

$$\begin{aligned} \mathsf{V}[\widetilde{\sigma^2}] &= \frac{2\sigma^4}{k(n-1)} \\ \mathsf{V}[\widetilde{\sigma_0^2}] &= \frac{2}{n^2}\left[\frac{\sigma^4}{k(n-1)} + \frac{(\sigma^2+n\sigma_0^2)^2}{k-1}\right] \end{aligned}$$

Bevis:
Sætningen bevises direkte ud fra udtrykkene (5.3.14) □

Bemærkning 1 *Centrale estimatorer for varianskvotient i balanceret tilfælde*

I det balancerede tilfælde, $n_1 = n_2 = \cdots = n_k = n$, kan vi angive eksplicitte centrale estimatorer for γ og $w(\gamma) = 1/(1+n\gamma)$. Der gælder

$$\widetilde{w} = \frac{SAK_1}{k(n-1)} \Big/ \frac{SAK_2}{k-3}$$

og (5.3.19)

$$\widetilde{\gamma} = \frac{1}{n}\left\{\frac{SAK_2}{k-1} \Big/ \frac{SAK_1}{k(n-1)-2} - 1\right\} \quad (5.3.20)$$

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

Bevis:
Følger ved at bemærke, at $SAK_1 \in \sigma^2\chi^2(k(n-1))$ og $SAK_2 \in \{\sigma^2/w(\gamma)\}\chi^2(k-1)$ er indbyrdes uafhængige med

$$\begin{aligned} \mathsf{E}\,[SAK_1/\{k(n-1)\}] &= \sigma^2 \\ \mathsf{E}\,[\{k(n-1)-2\}/SAK_1] &= 1/\sigma^2 \\ \mathsf{E}\,[SAK_2/(k-1)] &= \sigma^2/w(\gamma) \\ \mathsf{E}\,[(k-3)/SAK_2] &= w(\gamma)/\sigma^2 \end{aligned}$$

□

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

$$\mathrm{H}_{II} : \sigma_0^2 = 0. \quad (5.3.21)$$

Hypotesen testes ved at sammenligne varianskvotienten

$$Z = \frac{SAK_2/(k-1)}{SAK_1/(N-k)} \quad (5.3.22)$$

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

Sætning 5.3.5 *Test af homogenitetshypotese i den tilfældige model*

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

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

hvor z er givet ved (5.3.22)

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

$$Z \in (1 + n\gamma)\mathrm{F}(k-1, N-k) \tag{5.3.23}$$

Bevis:
Se f.eks. Scheffé (1959)

□

Testet er således det samme som i den systematiske model

Bemærkning 1 ***Konfidensinterval for varianskvotienten***

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

$$\mathsf{P}\left[\gamma_L < \gamma < \gamma_U\right]$$

fås ved at benytte

$$\begin{aligned} \gamma_L &= \frac{1}{n}\left(\frac{Z}{\mathrm{F}(k-1, N-k)_{1-\alpha/2}} - 1\right) \\ &\text{og} \\ \gamma_U &= \frac{1}{n}\left(\frac{Z}{\mathrm{F}(k-1, N-k)_{\alpha/2}} - 1\right) \end{aligned} \tag{5.3.24}$$

hvor Z er givet ved (5.3.22).

□

Eksempel 5.3.1 *Balancerede data, variansanalyseskema*

Data er fra J.M. Cameron: The use of Components of Variance in preparing schedules for sampling of baled wool. *Biometrics* **7**, (1951) pp. 83-96.

Råuld indeholder varierende mængder af fedt og andre urenheder, der må fjernes før den videre forarbejdning. Prisen - og værdien - af råulden afhænger af den mængde ren uld, der opnås efter omhyggelig rensning. Renheden af råulden udtrykkes som den rene uld i procent af vægten af råulden.

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.

Tabel 5.1. Renheden i % ren uld af 4 prøver fra hver af 7 baller uruguyansk uld.

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 variansanalyse, og får variansanalyseskemaet

Variation	SAK	f	$s^2 = SAK/f$	$\mathrm{E}[S^2]$
Mellem baller	65.9628	6	10.9938	$\sigma^2 + 4\sigma_0^2$
Indenfor baller	131.4726	21	6.2606	σ^2

Udfører vi et test for hypotesen $H_{II} : \sigma_0^2 = 0$, 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 $\widetilde{\sigma^2} = 6.261$ og variationen mellem baller estimeres ved

$$\widetilde{\sigma_0^2} = \frac{10.9938 - 6.2606}{4} = 1.183$$

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

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

og

$$\widetilde{V}[\widetilde{\sigma_0^2}] = \frac{2}{16}\left[\frac{(6.261)^2}{21} + \frac{(10.994)^2}{6}\right] = (1.659)^2$$

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

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 estimater for parametrene under den tilfældige model*

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

a) For $\sum_i n_i^2(\overline{y}_{i\cdot} - \overline{\overline{y}}_{\cdot\cdot})^2 < sak_1 + sak_2$ fås

$$\begin{aligned} \widehat{\mu} &= \overline{\overline{y}}_{\cdot\cdot} = \sum_i n_i \overline{y}_{i\cdot} \Big/ N \\ \widehat{\sigma}^2 &= (sak_1 + sak_2)/N \qquad (5.4.1) \end{aligned}$$

og

$$\widehat{\gamma} = 0 \qquad (5.4.2)$$

b) For $\sum_i n_i^2(\overline{y}_{i\cdot} - \overline{\overline{y}}_{\cdot\cdot})^2 > sak_1 + sak_2$ bestemmes estimaterne som løsning til

$$\widehat{\mu} = \sum_{i=1}^{k} n_i w_i(\widehat{\gamma}) \overline{y}_{i\cdot} \Big/ W(\widehat{\gamma}) \qquad (5.4.3)$$

$$\widehat{\sigma}^2 = \frac{1}{N}\{sak_1 + \sum_{i=1}^{k} n_i w_i(\widehat{\gamma})(\overline{y}_{i\cdot} - \mu^*)^2\} \qquad (5.4.4)$$

$$\begin{aligned} &\sum_{i=1}^{k} n_i^2 w_i(\widehat{\gamma})^2 (\overline{y}_{i\cdot} - \widehat{\mu})^2 \Big/ W(\widehat{\gamma}) \qquad (5.4.5) \\ = \; &\Big\{ sak_1 + \sum_{i=1}^{k} n_i w_i(\widehat{\gamma})(\overline{y}_{i\cdot} - \widehat{\mu})^2 \Big\} \Big/ N \end{aligned}$$

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

$$W(\gamma) = \sum_{i=1}^{k} n_i w_i(\gamma) \; . \qquad (5.4.6)$$

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

Bevis:

Logaritmen til likelihoodfunktionen bliver - på nær en additiv konstant -

$$\begin{aligned} l(\mu,\sigma^2,\gamma) &= -\frac{sak_1}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{k} n_i w_i(\gamma)(\overline{y}_{i\cdot}-\mu)^2 \\ &\quad -\frac{N}{2}\log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{k}\log(w_i(\gamma)) \end{aligned} \tag{5.4.7}$$

Vi ønsker at finde de værdier, $\widehat{\mu}, \sigma^{*2}, \widehat{\gamma}$, der maximerer l.

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

$$\begin{aligned} \frac{\partial l}{\partial \sigma^2} &= \frac{sak_1}{2\sigma^4} + \frac{1}{2\sigma^4}\sum_{i=1}^{k} n_i w_i(\gamma)(\overline{y}_{i\cdot}-\mu)^2 - \frac{N}{2\sigma^2} \\ \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{k} n_i w_i(\gamma)(\overline{y}_{i\cdot}-\mu) \\ \frac{\partial l}{\partial \gamma} &= \frac{1}{2\sigma^2}\sum_{i=1}^{k}\left\{n_i w_i(\gamma)(\overline{y}_{i\cdot}-\mu)\right\}^2 - \frac{1}{2}\sum_{i=1}^{k} n_i w_i(\gamma) \end{aligned}$$

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

Af $\dfrac{\partial l}{\partial \sigma^2} = 0$ fås

$$\sigma^2 = \left[sak_1 + \sum_i n_i w_i(\gamma)(\overline{y}_{i\cdot}-\mu)^2\right] \Big/ N$$

der netop er (5.4.4)

For at bestemme variansforholdet γ betragter vi

$$\frac{\partial l}{\partial \gamma} = \frac{1}{2\sigma^2}\sum_{i=1}^{k}\left\{n_i w_i(\gamma)(\overline{y}_{i\cdot}-\mu)\right\}^2 - \frac{1}{2}\sum_{i=1}^{k} n_i w_i(\gamma)$$

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

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

$$\widehat{\mu} = \overline{y}_{..} = \sum_{i=1}^{k} n_i \overline{y}_{i.} \Big/ \sum_{i=1}^{k} n_i$$

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

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{y}_{i.} - \overline{y}_{..})^2 \\ &= \Big[sak_1 + \sum_{i=1}^{k} n_i (\overline{y}_{i.} - \overline{y}_{..})^2 \Big] \Big/ N = (sak_1 + sak_2)/N \end{aligned}$$

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

$$\sum_{i=1}^{k} n_i^2 (\overline{y}_{i.} - \overline{y}_{..})^2 < sak_1 + sak_2$$

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

For $\sum_i n_i^2 (\overline{y}_{i.} - \overline{\overline{y}}_{..})^2 > sak_1 + sak_2$, bestemmes ML-estimatet $\widehat{\gamma}$ som løsning til $\frac{\partial l}{\partial\gamma} = 0$, hvilket netop fører til (5.4.5). □

Bemærkning 1 *Maksimum-likelihood estimatet $\widehat{\mu}$ er et vejet gennemsnit af gruppegennemsnittene*

Vi ser af (5.4.3), at $\widehat{\mu}$ er et vejet gennemsnit af gruppegennemsnittene, $\overline{y}_{i.}$, med de marginale præcisioner

$$\sigma^2 n_i w_i(\gamma) = \sigma^2 / \mathrm{V}\,[\overline{Y}_{i.}]$$

som vægte. Der gælder nemlig

$$\mathsf{V}\,[\overline{Y}_{i\cdot}] = \sigma_0^2 + \sigma^2/n_i = \frac{\sigma^2}{n_i}(1 + n_i\gamma) = \sigma^2/\{n_i w_i(\gamma)\}$$

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 $\widehat{\mu}$ blot er det simple gennemsnit af gruppegennemsnittene. □

Bemærkning 2 *Estimatet for σ^2 udnytter også variationen mellem grupper*

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

$$\mathsf{E}\,[(\overline{Y}_{i\cdot} - \mu)^2] = \mathsf{V}\,[\overline{Y}_{i\cdot}] = \sigma^2/\{n_i w_i(\gamma)\}$$

hvorfor leddene $(\overline{y}_{i\cdot} - \mu)^2$ indeholder information, såvel om σ^2, som information om γ. □

Bemærkning 3 *Estimatet for σ_0^2 er ikke nødvendigvis centralt*

Vi bemærker endvidere, at - som vanligt ved ML-estimatet- er estimatet for σ_0^2 ikke nødvendigvis centralt.

I stedet for maksimum-likelihood estimatet benytter man derfor undertiden 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 momentestimaterne for μ og σ^2 de samme som maksimum-likelihoodestimaterne.*

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

$$\widehat{\mu} = \sum_{i=1}^{k} \overline{y}_{i+}/k = \overline{\overline{y}}_{++} \,,$$

altså 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 σ^2 og $\gamma = \sigma_0^2/\sigma^2$ bliver

$$\begin{aligned} N\sigma^2 &= sak_1 + \frac{1}{1+n\gamma}\, sak_2 \\ N\,\frac{n}{1+n\gamma}\,\frac{sak_2}{k} &= sak_1 + \frac{1}{1+n\gamma}\, sak_2 \end{aligned}$$

med løsningen

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{sak_1}{N-k} \\ \widehat{\gamma} &= \frac{1}{n}\left[\frac{sak_2}{k\widehat{\sigma}^2} - 1\right] \end{aligned}$$

dvs.

$$\widehat{\sigma}_0^2 = \frac{sak_2/k - \widehat{\sigma}^2}{n}$$

Ved sammenligning med (5.3.14) ser man at maksimum-likelihood estimatet for σ^2 er den samme som momentestimatet, men maksimum-likelihood estimatet for σ_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 σ^2 forskelligt fra momentestimatet. □

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 μ fås ved at benytte (5.4.3) med

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

□

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

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

hvor

$$\begin{aligned} \mathbf{V}^* &= \text{Blok diag}\{\mathbf{I}_{n_i} + \gamma\mathbf{J}_{n_i}\} && (5.4.8)\\ \mathbf{r} &= \mathbf{y} - \widehat{\mu}(\gamma)\mathbf{1} && (5.4.9)\\ \widehat{\mu}(\gamma) &= \sum_{i=1}^{k} n_i w_i(\gamma)\overline{y}_{i\cdot} / \sum_{i=1}^{k} n_i w_i(\gamma) && (5.4.10) \end{aligned}$$

med

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

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

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

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

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

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

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

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 μ, svarende til fordelingen af residualerne bestemt ved

$$\begin{aligned} \widetilde{l}_r(\sigma^2, \gamma) &= -\frac{1}{2}\left[(N-1)\ln(\sigma^2) + \sum_{i=1}^{k} \ln(1 + n_i\gamma) + \ln\left(\sum_{i=1}^{k} n_i w_i(\gamma)\right)\right. \\ &\quad \left. + \frac{\mathbf{r}^T(\mathbf{V}^*)^{-1}\mathbf{r}}{\sigma^2} + N\ln(2\pi)\right] \ , \end{aligned} \tag{5.4.12}$$

hvor $\mathbf{V}^*$ og $\mathbf{r}$ er givet ved (5.4.8) og (5.4.9).
Estimatet $\widehat{\mu}, \widehat{\sigma}^2$ og $\widehat{\gamma}$, 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:
Overspringes, se f.eks. Ronald Christensen (1987), pp. 235 ff. □

Sætning 5.4.2 *Approximativ varians for maksimum-likelihood estimatorerne under den tilfældige model*

Den asymptotiske varians for maximaliseringsestimatorerne er

$$\begin{aligned} \mathsf{V}[\sigma^{*2}] &= 2\sigma^4 \sum_{i=1}^{k} \{n_i w_i(\gamma)\}^2 / D \\ \mathsf{V}[\sigma_0^{*2}] &= 2\sigma^4 (N - k + \sum_{i=1}^{k} n_i w_i(\gamma)^2) / D \end{aligned}$$

og

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

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 maximimumspunktet) □

Eksempel 5.4.1 ***Varianskvotienten γ's betydning for vægtningen af gruppegennemsnittene***

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

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

$$\mu^*_{\gamma=0} = \sum_{i=1}^{k} n_i \overline{y}_{i\cdot} / \sum_{i=1}^{k} n_i$$

d.v.s. alle N enkeltobservationer vægtes ens.

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

$$\mu^*_{\gamma=\infty} = \sum_{i=1}^{k} \overline{y}_{i\cdot} / k$$

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 balancerede data (lige mange observationer i alle grupper), som situationer med ubalancerede 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 *Balancerede 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, ialt 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                    p.1
        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
```

General Linear Models Procedure p.2

Dependent Variable: RENH

Source	DF	Sum of Squares	F Value	Pr > F
Model	6	65.96264286	1.76	0.1573
Error	21	131.47220000		
Corrected Total	27	197.43484286		

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

Source	DF	Type I SS	F Value	Pr > F
BALLE	6	65.96264286	1.76	0.1573

General Linear Models Procedure p.3

Source	Type I Expected Mean Square
BALLE	Var(Error) + 4 Var(BALLE)

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

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

Under overskriften `RENH Mean` er angivet gennemsnittet af renhedsmålingerne.

Testet for hypotesen $\sigma_0^2 = 0$ er det samme som testet for balleeffekt i den systematiske model. Denne teststørrelse er anført såvel i variansanalyseskemaet, 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.

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

forventningsværdien

$$\mathrm{E}\left[SAK_2/(k-1)\right] = \sigma^2 + 4\sigma_0^2$$

(jvf beviset for sætning 5.3.3).

Herved kan σ_0^2 estimeres som i eksempel 5.3.1. □

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

For sætttet af ubalancerede data vil programmet

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

give udskriften:

```
                General Linear Models Procedure                    p.2

Dependent Variable: RENH

Source                    DF   Sum of Squares   F Value      Pr > F

Model                      6      72.55028636      4.36      0.0096
Error                     15      41.573700000
Corrected Total           21     114.123986366

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

Source                    DF        Type I SS   F Value      Pr > F
BALLE                      6      72.55028636      4.36      0.0096
                General Linear Models Procedure                    p.3

Source   Type I Expected Mean Square

BALLE    Var(Error) + 3.0909 Var(BALLE)
```

Udskriftens struktur adskiller sig ikke fra udskriften fra de balancerede data.

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

$$\mathsf{E}\,[SAK_2/(k-1)] = \sigma^2 + 3.0903\sigma_0^2$$

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

Idet $\widetilde{\sigma^2} = 41.5737/15 = 2.77158$ får man derfor momentestimatet
$\widetilde{\sigma_0^2} = (12.091714 - 2.771580)/3.0909 = 3.0153$ □

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 variansanalyse i den tilfældige model.

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

Eksempel 5.5.3 ***Balancerede data, maksimum-likelihood estimation ved SAS®-proceduren MIXED***

Programmet:

```
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

Class     Levels  Values

BALLE          7  1 2 3 4 5 6 7
PROVE          4  1 2 3 4
```

til kontrol af de indlæste specifikationer.

Endvidere udskrives iterationsforløbet:

```
The MIXED Procedure

ML Estimation Iteration History

Iteration  Evaluations     Objective     Criterion

        0            1   82.68971508
        1            1   82.22198159    0.00000000

Convergence criteria met.
```

Da data er balancerede, behøves kun én iteration.

Derefter udskrives estimaterne for σ^2 og σ_0^2:

```
Covariance Parameter Estimates (MLE)

Cov Parm      Estimate
```

```
BALLE       0.79066344
Residual    6.26058095
```

De to varianser σ^2 og σ_0^2 kaldes kovariansparametre, da modellen jo er karakteriseret ved dispersionsmatricen, $\mathbf{V}$, (5.3.11)

Udskriftens første søjle Ratio angiver det estimerede forhold $\widehat{\gamma}$ mellem variansen svarende til den tilfældige effekt og residualvariansen.

I søjlen Estimate angives estimaterne $\widehat{\sigma}_0^2$ (svarende til BALLE) og $\widehat{\sigma}^2$ (svarende til residualvariansen). Søjlen Std Error angiver den estimerede spredning, hhv $\sqrt{\widehat{\mathrm{V}}\,[\widehat{\sigma}_0^2]}$ og $\sqrt{\widehat{\mathrm{V}}\,[\widehat{\sigma}^2]}$.

Søjlen Z angiver

$$\frac{\widehat{\sigma}_0^2}{\sqrt{\widehat{\mathrm{V}}\,[\widehat{\sigma}_0^2]}}$$

og 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

Cov Parm    Row        COVP1         COVP2

BALLE         1   1.81896982   -0.93321128
Residual      2   -0.93321128   3.73284513
```

Estimatet er den inverse observerede informationsmatrix. Den inverse observerede informationsmatrix beregnes som 2 gange den inverse Hessianmatrix i maksimumspunktet.

Endelig udskrives et resume af tilpasningen,

```
The MIXED Procedure
```

```
        Model Fitting Information for RENH

Description                          Value

Observations                       28.0000
Log Likelihood                    -66.8413
Akaike's Information Criterion    -68.8413
Schwarz's Bayesian Criterion      -70.1735
-2 Log Likelihood                 133.6825
```

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 ***Balancerede 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)

    Cov Parm       Estimate

    BALLE        1.18329821
    Residual     6.26058095
```

med de tilsvarende varians-kovariansestimater:

```
     Asymptotic Covariance Matrix of Estimates

    Cov Parm    Row          COVP1          COVP2
```

```
BALLE          1    2.75128329   -0.93321128
Residual       2   -0.93321128    3.73284513
```

Vi ser, at REML-estimatet for σ^2 er det samme, som ML-estimatet, men REML-estimatet for σ_0^2 er større end ML-estimatet, svarende til at ML-estimatet ikke korrigerer for frihedsgraderne i estimationen af σ_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

```
Covariance Parameter Estimates (MLE)

Cov Parm        Estimate

BALLE         2.62615956
Residual      2.80080883
```

Sammenigner vi med momentestimatet i eksempel 5.5.2 ser vi, at såvel estimatet for σ_0^2 som for σ^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.

```
Covariance Parameter Estimates (REML)
```

```
Cov Parm        Estimate

BALLE         3.26121525
Residual      2.79158363
```

Vi ser, at estimatet for σ_0^2 som ventet er større end det tilsvarende maksimumlikelihood estimat, og desuden bemærker vi, at der er en lille forskel på estimaterne for σ^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 normalfordelingssammenhænge.

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:

```
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

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 repeatibility and reproduceability.

Usikkerhedsmålene 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 μ angiver den vedtagne sande værdi plus en eventuel bias; σ^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 prøvningsmetode. 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)

	Laboratorium				
i	1	2	3	4	5
	97.4	92.6	96.2	95.2	93.0
	97.8	93.0	96.8	95.5	95.0
	98.8	92.5	96.8	95.7	94.6
	98.0	92.7	96.2	95.7	95.1
$\overline{y}_{i\cdot}$	98.000	92.700	96.500	95.525	94.425
s_i	0.589	0.216	0.346	0.236	0.974

	Laboratorium				
i	6	7	8	9	
	94.8	96.3	99.8	99.5	
	95.2	96.3	100.0	99.3	
	95.0	96.1	99.9	99.8	
	94.9	96.3	99.7	99.7	
$\overline{y}_{i\cdot}$	94.975	96.250	99.850	99.575	
s_i	0.171	0.100	0.129	0.222	

Rækken $\overline{y}_{i\cdot}$ angiver gennemsnittet af de fire prøvningsresultater på det i'te laboratorium og s_i angiver tilsvarende den empiriske spredning på dette laboratoriums resultater, $s_i^2 = \sum_j (y_{ij} - \overline{y}_{i\cdot})^2/3$.

Idet vi benytter modellen (5.3.1) og (5.3.2), hvor μ angiver den ved prøvningsmetoden bestemte værdi af prøvematerialets reflektans, σ^2 angiver metodens repeterbarhedsvarians, og σ_0^2 angiver variansen mellem laboratorier, finder vi variansanalyseskemaet:

Variation	SAK	f	$s^2 = SAK/f$	E $[S^2]$
Mellem laboratorier	179.8323	8	22.4790	$\sigma^2 + 4\sigma_0^2$
Indenfor laboratorier	4.8700	27	0.1804	σ^2

Man finder således estimatet $\widetilde{\sigma^2} = 0.1804 = (0.42\ [\%\text{reflektans}])^2$ for repeterbarhedsvariansen, og $\widetilde{\sigma_0^2} = 5.5747 = (2.36\ [\%\text{reflektans}])^2$ for variansen mellem laboratorier. Reproducerbarhedsvariansen, dvs. usikkerheden $\sigma_0^2 + \sigma^2$ på en reflektansbestemmelse estimeres da som $\widetilde{\sigma_0^2} + \widetilde{\sigma^2} = 5.5747 + 0.1804 = 5.7551 = (2.39\ [\%\text{reflektans}])^2$.

Usikkerheden på gennemsnittet af m gentagne prøvninger samme laboratorium kan således udtrykkes ved variansestimatet $\mathrm{V}[\overline{Y}] \simeq \widetilde{\sigma_0^2} + \widetilde{\sigma^2}/m$.

Selv ved vilkårligt mange gentagne prøvninger på samme laboratorium vil usikkerheden på gennemsnittet af gentagne prøvninger ikke kunne blive mindre end spredningen mellem laboratorier, $\sigma_0 \simeq 2.36$ [%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, hvorvidt kvaliteten er tilfredsstillende. Der udtages derfor tilfældigt m sække, hvorefter der udtages n prøver tilfældigt fra indholdet af hver sæk og kvalitetsegenskaben for den pågældende prøve bestemmes ved en laboratorieundersøgelse.

Lad $\overline{Y}_{i\cdot}$ betegne det gennemsnitlige indhold af de n prøver fra den i'te sæk, og lad tilsvarende $\overline{\overline{Y}}_{\cdot\cdot}$ betegne det totale gennemsnit af de ialt m n prøver.

Antag, at variansen på prøvningsresultatet ved gentagne prøver fra samme sæk er σ^2, og antag endvidere, at sækkenes middelkvalitet varierer fra sæk til sæk med variansen $\gamma\sigma^2$.

Man finder da variansen på gennemsnittet fra den i'te sæk,

$$\mathrm{V}[\overline{Y}_{i\cdot}] = \sigma^2(\gamma + 1/n)$$

og da prøverne fra de m sække er uafhængige, finder man variansen på det totale gennemsnit

$$\mathrm{V}[\overline{\overline{Y}}_{i\cdot}] = \frac{\sigma^2}{m}(\gamma + 1/n)$$

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 κ 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

$$n \simeq \sqrt{\kappa/\gamma} \qquad m = \frac{C}{\kappa + n}$$

Antallet af laboratorieprøver pr sæk bestemmes således af forholdet κ mellem omkostningen ved udvælgelse af en sæk og omkostningen ved undersøgelse af en prøve fra sækken samt af forholdet γ mellem variansen imellem sække og variansen indenfor en sæk.

Jo større værdi af κ, desto flere prøver vil man udtage pr. sæk; jo større værdi af γ, desto færre prøver vil man udtage pr sæk.

□

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 udsættes for n uafhængige prøvninger (oftest er $n = 1$), og det totale gennemsnit $\overline{\overline{Y}}_{..} = \sum_i \sum_j Y_{ij}/(nk)$ udregnes. Ved sammenligning af dette fundne gennemsnit med en specificeret værdi μ_0 for middelkvaliteten benyttes variansen

$$\mathsf{V}\,[\overline{\overline{Y}}_{..}] = \sigma_2^2/k + \sigma_3^2/(nk),$$

hvor σ_2^2 angiver variansen mellem kvalitetsmålet for de enkelte enheder og σ_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 $\sigma_3^2 = \sigma^2$ (repeterbarhedsvariansen), og at addere reproducerbarhedsbidraget σ_0^2 til udtrykket for $\mathsf{V}\,[\overline{\overline{Y}}_{..}]$.

□

Eksempel 5.6.4 *Bulk sampling*

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 (inkrementer) af samme størrelse fra partiet. Der foretages en fysisk gennemsnitsdannelse af disse inkrementer 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 $\overline{X}_\cdot = \sum_\nu X_\nu/m$ 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 μ angiver partiets middelindhold og α_i antages at kunne beskrives som uafhængige $\mathrm{N}(0, \sigma_\alpha^2)$-fordelte variable. (Den tilfældige udvælgelse af inkrementer tjener til at sikre, at eventuelle systematiske variationer af α_i elimineres.)

Middelindholdet $\overline{Y}_\cdot$ i basisbunken er da

$$\overline{Y}_\cdot = \mu + \overline{\alpha},$$

hvor $\overline{\alpha} \in \mathrm{N}(0, \sigma_\alpha^2/k)$, og middelindholdet Z i laboratorieprøven bliver

$$Z = \overline{Y}_\cdot + B,$$

hvor B antages at være uafhængig af α_i og $B \in \mathrm{N}(0, \sigma_B^2)$.Variansbidraget σ_B^2 beskriver variationen hidrørende fra forberedelsen af laboratorieprøven.

Endelig antages analyseresultatet X_ν at kunne beskrives som

$$X_\nu = Z + \epsilon_i, \quad \nu = 1, 2, \ldots, m$$

hvor ϵ_ν antages indbyrdes uafhængige og uafængige af α_i og B, og hvor $\epsilon_\nu \in \mathrm{N}(0, \sigma_3^2)$. Variansbidraget σ_3^2 beskriver variationen hidrørende fra forberedelse og prøvning af analyseprøven.

Under disse antagelser finder man endelig usikkerhedsvariansen på estimatet $\overline{X}_{\cdot}$ (variansen svarende til gentagelser af hele proceduren):

$$\mathsf{V}[\overline{X}_{\cdot}] = \sigma_\alpha^2/k + \sigma_B^2 + \sigma_3^2/m.$$

Bestemmelsen af de indgående varianser må foregå ved planlagte forsøg, der inddrager flere basisbunker etc., og deres størrelse må løbende estimeres med henblik på en overvågning af usikkerheden. Gy (1992) har formuleret en sammenhængende teori for bulksampling. Dele af teorien bygger imidlertid på nogle antagelser om partikelformen, og teorien anses derfor af 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 enkelte gruppe varierer tilfældigt fra gruppe til gruppe.

Modellen bygger på antagelsen

$$Y_{ij}|(\mu_i, \sigma_i^2) \in \mathrm{N}(\mu_i, \sigma_i^2) \tag{5.7.1}$$

hvor

$$\begin{aligned} \mu_i|\sigma_i^2 &\in \mathrm{N}(\mu_0, \sigma_i^2/m), \\ \text{og} & \\ \sigma_i^2 &\in \mathrm{RGam}(\alpha, \beta) \end{aligned} \tag{5.7.2}$$

og hvor Y_{ij} er betinget uafhængige (se definition 4.5.1 side 467) for givet (μ_i, σ_i^2), og endvidere er σ_i^2 indbyrdes uafængige.

Antagelsen indebærer en udvidelse i forhold til den tilfældige model, der blev beskrevet i afsnittene 5.3 til 5.6.

Den væsentligste udvidelse består i antagelsen om at gruppevariansen, σ_i^2 varierer tilfældigt fra gruppe til gruppe. Vi har valgt at antage, at variansen følger en såaldt 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 μ_i og σ_i^2 være en normalfordeling med

$$\begin{aligned} \mathsf{E}\,[Y_{ij}|\mu_i,\sigma_i^2] &= \mu_i \\ \mathsf{COV}[Y_{ij},Y_{il}|\mu_i,\sigma_i^2] &= \begin{cases} 0 & \text{for } j \neq l \\ \sigma_i^2 & \text{for } j = l \end{cases} \end{aligned}$$

Endvidere er de betingede fordelinger af $\overline{Y}_{i+}$ normale med

$$\mathsf{E}\,[\overline{Y}_{i+}|\mu_i,\sigma_i^2] = \mu_i \qquad \mathsf{V}\,[\overline{Y}_{i+}|\mu_i,\sigma_i^2] = \sigma_i^2/n_i$$

$$\mathsf{E}\,[\overline{Y}_{i+}|\sigma_i^2] = \mu_0 \qquad \mathsf{V}\,[\overline{Y}_{i+}|\sigma_i^2] = \sigma_i^2(\frac{1}{n_i}+\frac{1}{m})$$

Den marginale fordeling af gruppemiddelværdien, μ_i er givet ved

$$\mu_i \in \mathrm{T}\big(2\alpha,\mu_0,\sqrt{\beta/(m\alpha)}\big)\,, \tag{5.7.3}$$

hvor $T(\nu,\mu_0,\beta_1)$ angiver t-fordelingen med ν frihedsgrader og positionsparameter μ_0 og skalaparameter β_1 (se Oversigt over fordelinger med anvendelser i Statistik, IMM 1998).

Den marginale fordeling af $\overline{Y}_{i+}$ er givet ved

$$\mathrm{T}\left(2\alpha,\mu_0,\sqrt{\frac{\beta}{\alpha}\left(\frac{1}{n_i}+\frac{1}{m}\right)}\right) \tag{5.7.4}$$

For variansskønnet, S_i^2 for den i'te gruppe

$$S_i^2 = \sum_j (Y_{ij} - \overline{Y}_{i+})^2/(n_i - 1), \tag{5.7.5}$$

gælder

$$(n_i - 1)S_i^2|\sigma_i^2 \in \sigma_i^2\chi^2(n_i - 1),$$

og den marginale fordeling er givet ved

$$S_i^2 \in \mathrm{RBet}\Big(\alpha, \frac{1}{2}(n_i - 1), \frac{2\beta}{n_i - 1}\Big) \tag{5.7.6}$$

hvor RBet-fordelingen er den reciproke betafordeling.
Bevis:
Overspringes. □

Bemærkning 1 ***Fordelingen af gruppegennemsnittene har tykkere haler end normalfordelingen***

Vi bemærker, at den marginale fordeling af gruppemiddelværdierne, μ_i og af gruppegennemsnittene $\overline{Y}_{i+}$ er t-fordelinger, der har tykkere haler end normalfordelingen.

Dette skyldes, at det ikke kun er middelværdierne, μ_i, der varierer, men også de tilsvarende varianser. □

Bemærkning 2 ***Fortolkning af parametrene α og β i fordelingen af varianserne.***

Vi bemærker, at fordelingen af σ^2 kan udtrykkes ved at

$$\frac{1}{\sigma^2} \in \frac{1}{\sigma_0^2\,(\nu - 2)}\,\chi^2(\nu)\,, \tag{5.7.7}$$

med $\sigma_0^2 = \mathsf{E}\,[\sigma^2] = \beta/(\alpha - 1)$ og $\nu = 2\alpha$. se sætning 6.6.1 på side 601 □

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 gruppegennemsnittene har tykkere haler 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.

Bemærkning 3 *Variansanalyseskema ved tilfældigt varierende varians*

Lader vi S^2 betegne det sædvanlige vejede gennemsnit af de gruppevise variansskøn,

$$S^2 = SAK_1/(N-k) \tag{5.7.8}$$

har vi

$$\mathrm{E}\,[S^2] = \beta/(\alpha-1) = \sigma_0^2$$

således at variansanalyseskemaet bliver

Variation	SAK	f	E $[SAK/f]$
Mellem grupper	$\sum_i n_i(\overline{Y}_{i+} - \overline{Y}_{++})^2$	$k-1$	$\sigma_0^2(1+\frac{n_0}{m})$
Indenfor grupper	$\sum_i \sum_j (Y_{ij} - \overline{Y}_{i+})^2$	$N-k$	σ_0^2

Det ses, at det sædvanlige variansanalyseskema kun giver mulighed for estimation af middelvariansen $\mathrm{E}\,[\sigma^2] = \sigma_0^2$. For at bestemme parameteren ν er det nødvendigt yderligere at betragte variationen mellem de gruppevise varianser. □

Sætning 5.7.2 *Forventningsværdi af kvadratafvigelsessummen mellem empiriske varianser*

Lad

$$SAK_s = \sum_i (n_i - 1)(S_i^2 - S^2)^2 \tag{5.7.9}$$

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)

$$\begin{aligned}\mathrm{E}\,[SAK_s/(k-1)] &= (\sigma_0^2)^2\Big\{2\,\frac{\nu/2-1}{\nu/2-2} + \frac{1}{(\nu/2-2)(k-1)}\Big[N-k \\ &\quad - \frac{\sum_i (n_i-1)^2}{N-k}\Big]\Big\}\end{aligned} \tag{5.7.10}$$

Bevis:
Overspringes. □

Sætning 5.7.3 *Momentestimation i modellen med tilfældigt varierende varians*

Momentestimaterne for parametrene μ_0, σ_0^2, ν og m i modellen (5.7.1) og (5.7.2) er givet ved

$$\begin{aligned}
\widetilde{\mu}_0 &= \overline{Y}_{++} \\
\widetilde{\nu} &= 2\left[1 + \left\{ Q_1 + \frac{1}{k-1}\left[N - k - \frac{\sum_i (n_i - 1)^2}{N-k} \right] \right\} / (Q_1 - 2)\right] \\
\widetilde{\sigma_0^2} &= \frac{SAK_1}{N-k} \\
&\text{og} \\
\widetilde{m} &= \frac{1}{(k-1)(Q_2-1)} \left(N - \frac{\sum_i n_i^2}{N} \right) = \frac{n_0}{Q_2 - 1}
\end{aligned}$$

hvor

$$Q_1 = \frac{SAK_s/(k-1)}{[SAK_1/(N-k)]^2} \tag{5.7.11}$$

og (5.7.12)

$$Q_2 = \frac{SAK_2/(k-1)}{SAK_1/(N-k)} \tag{5.7.13}$$

og n_0 er givet ved (5.1.9).

Bevis:

Overspringes □

Bemærkning 1 *Momentestimation i det balancerede tilfælde*

I det balancerede tilfælde, $n_1 = n_2 = \cdots = n_k = n$, får vi

$$\begin{aligned}
\widetilde{\nu} &= 4 + 2(n+1)/(Q_1 - 2) \\
\widetilde{\sigma_0^2} &= SAK_1/[k(n-1)] \\
&\text{og} \\
\widetilde{m} &= n/(Q_2 - 1)
\end{aligned}$$

idet der i dette tilfælde gælder

$$\begin{aligned} \mathrm{E}\,[SAK_2/(k-1)] &= \sigma_0^2\,(1+\frac{n}{m}) \\ \mathrm{E}\,[SAK_s/(k-1)] &= \frac{(\sigma_0^2)^2}{(\nu/2-2)}[2(\nu/2-1)+n-1] \end{aligned}$$

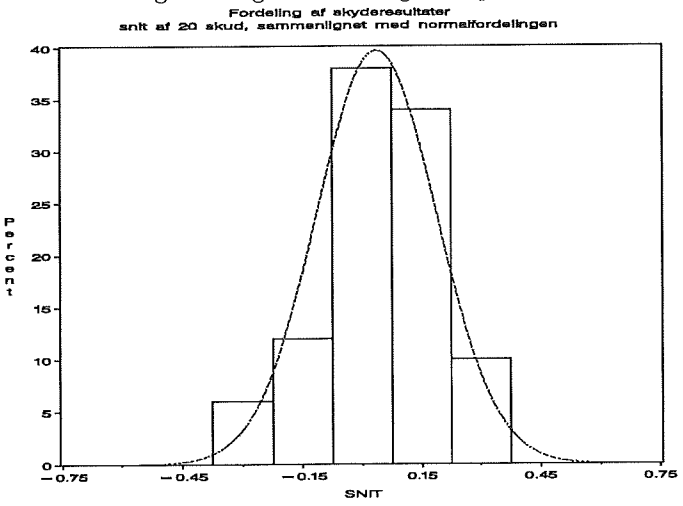
□

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 indtegnede kurve viser normalfordelingen med samme middelværdi og samme varians som de 35 resultater.

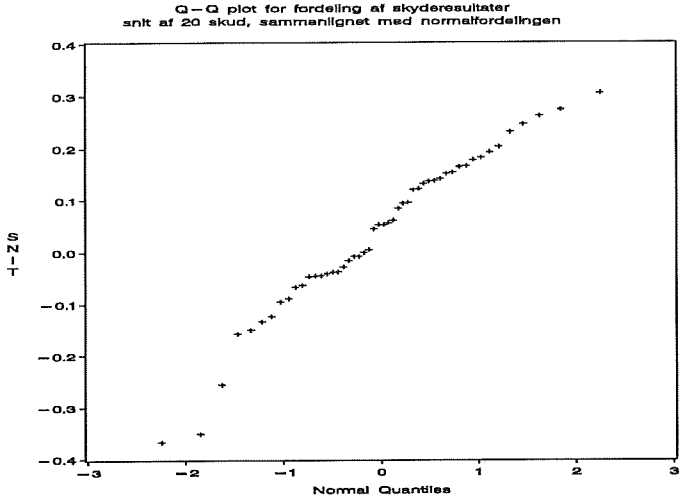
Det anes, at den observerede fordeling har noget tykkere haler, end normalfordelingen.

Den observerede fordeling af det gennemsnitlige træfpunkt for hver af 35 skytter

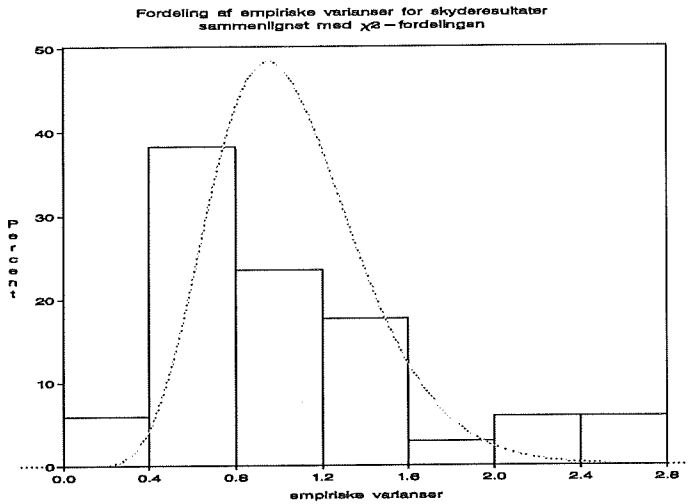


Fordelingen er sammenlignet med en normalfordeling med samme middelværdi og varians.

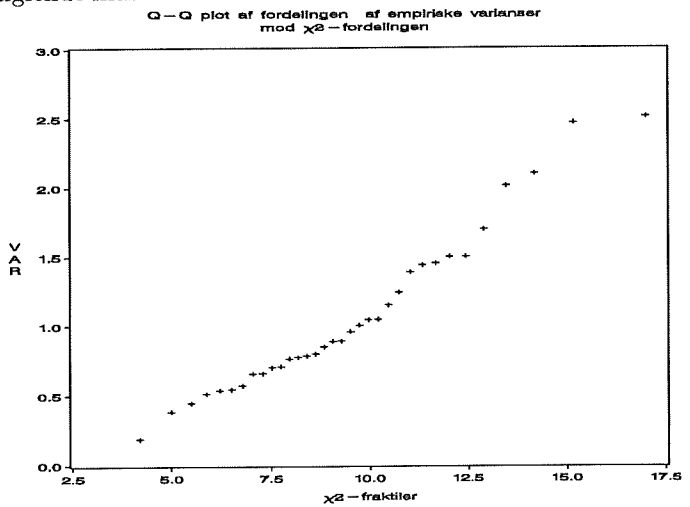
For nærmere at undersøge dette, viser nedenstående figur det tilsvarende fraktildiagram (Q-Q-plot).



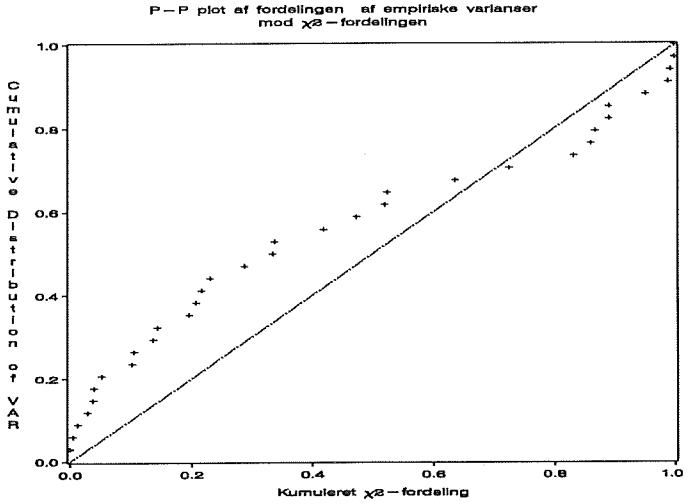
Man ser også her en antydning af, at fordelingen har tykkere haler, end normalfordelingen. For at belyse fænomenet yderligere, har man i den følgende figur tegnet et histogram over de empiriske varianser for hver skytte. I figuren er indtegnet tætheden for den tilsvarende $\sigma_0^2\chi^2(19)/19$-fordeling. Det ses, at den observerede fordeling har noget tykkere haler, end χ^2-fordelingen.



Det tilsvarende Q-Q plot giver også en antydning af de tykkere haler og den manglende masse i den centrale del:

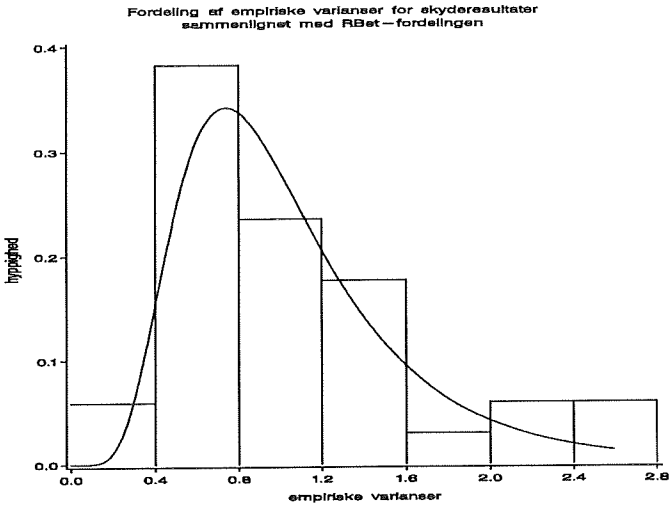


Og nedenstående P-P plot (kumulerede hyppigheder for de empiriske varianser mod tilsvarende kumulerede sandsynligheder i χ^2-fordelingen, se Shapiro og Wilk (1965) og Wilk og Gnanadesikan (1968)) viser en tydelig afvigelse fra χ^2-fordelingen.

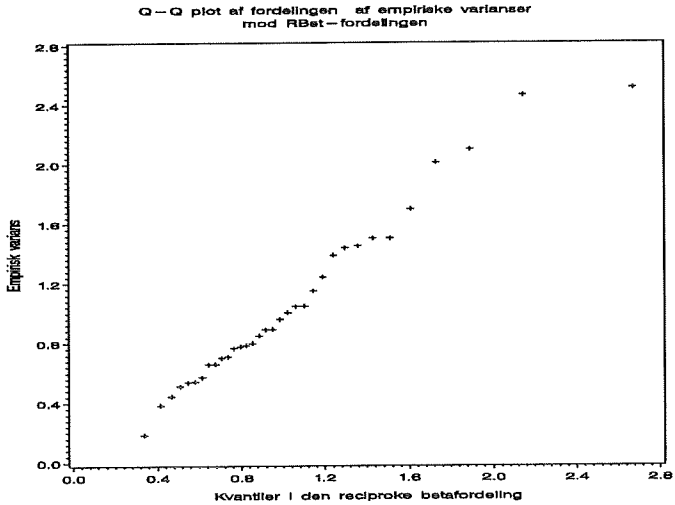


Man har derfor valgt at modellere resultaterne ved en normalfordelingsmodel med en tilfældig middelværdi, μ_i, og en tilfældig varians, σ^2, for hver skytte. I figuren er med en fuldt optrukket kurve indtegnet tætheden i den marginale fordeling (5.7.4) svarende til denne model.

I stedet er det derfor forsøgt at tilpasse en RBet-fordeling til fordelingen af empiriske varianser. Parametrene er estimeret ved maksimum-likelihood-metoden. Nedenstående figur viser fordelingen af de empiriske varianser sammenlignet med den estimerede RBet-fordeling. Man ser, at der er en rimelig god tilpasning til denne model.



Det tilsvarende Q-Q plot viser ligeledes en tilfredsstillende tilpasning



□

Tabel 5.2. Oversigt over marginale fordelinger i endimensionale varianskomponentfordelinger

Normaltfordelte observationer, samme varians i alle grupper.
k grupper og n_i gentagelser i hver gruppe.
Observationer: $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$;

$$\overline{Y}_{i+} = \sum_{j=1}^{n_i} Y_{ij}/n_i\,; \quad \overline{Y}_{++} = \sum_{i=1}^{k} n_i \overline{Y}_{i+}/N\,; \quad S_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2/(n_i - 1)$$

$$SAK_1 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2\,; \quad SAK_2 = \sum_{i=1}^{k} n_i (\overline{Y}_{i+} - \overline{Y}_{++})^2\,; s_2^2 = SAK_2/(k-1)$$

Model: $Y_{ij}|\mu_i \in \mathrm{N}(\mu_i, \sigma^2)$;

Parameteren μ_i er $\mathrm{N}(\mu_0, \sigma_0^2)$ fordelt, $(\sigma_0^2 = \gamma\sigma^2)$.

Stikprøve-ford. af $\overline{Y}_{i+}\|\mu_i$	Forde-ling af μ_i	Marginal ford. af $\overline{Y}_{i+}$	V $[\overline{Y}_{i+}]$	Fordeling af S_i^2	E $[S_i^2]$	V $[S_i^2]$
$\mathrm{N}(\mu_i, \sigma^2/n_i)$	$\mathrm{N}(\mu_0, \gamma\sigma^2)$	$\mathrm{N}\big(\mu_0, \sigma^2(\gamma + 1/n_i)\big)$	$\sigma^2(\gamma + 1/n_i)$	$\sigma^2\chi^2(n_i - 1)/(n_i - 1)$	σ^2	$2\sigma^4/(n_i - 1)$

Momentestimation af parametre i endimensionale varianskomponentmodeller:

$$\widetilde{\mu}_0 = \overline{Y}_{++}\,; \quad \widetilde{\sigma}^2 = SAK_1/(N-k)\,; \quad \widetilde{\sigma}_0^2 = \left\{ SAK_2 - \frac{k-1}{N-k} SAK_1 \right\} \Big/ \{(k-1)n_0\}$$

Den vægtede gennemsnitlige gruppestikprøvestørrelse, n_0 er bestemt ved (5.1.9).

Tabel 5.3. Oversigt over marginale fordelinger i endimensionale varianskomponentfordelinger

Normaltfordelte observationer, varierende varians i grupper.
k grupper og n_i gentagelser i hver gruppe.
Observationer: $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$;

$$\overline{Y}_{i+} = \sum_{j=1}^{n_i} Y_{ij}/n_i\,; \quad \overline{Y}_{++} = \sum_{i=1}^{k} n_i \overline{Y}_{i+}/N\,; \quad S_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2/(n_i - 1)$$

$$SAK_1 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i+})^2\,; \quad SAK_2 = \sum_{i=1}^{k} n_i (\overline{Y}_{i+} - \overline{Y}_{++})^2\,; \quad s_2^2 = SAK_2/(k-1)\,; \quad \overline{S}_+^2 = SAK_1/(N-k)$$

$$SAK_s = \sum_{i=1}^{k} (n_i - 1)(S_i^2 - \overline{S}_+^2)^2\,; \quad N = \sum_{i=1}^{k} n_i$$

Model: $Y_{ij}|(\mu_i, \sigma_i^2) \in \mathrm{N}(\mu_i, \sigma_i^2)$;

For givet σ_i^2 er $\mu_i|\sigma_i^2 \in \mathrm{N}(\mu_0, \sigma_i^2/m)$.
σ_i^2 er $\mathrm{RGam}(\alpha, \beta)$-fordelt

Marginal ford. af $\overline{Y}_{i+}\|\sigma_i^2$	Marginal fordeling af $\overline{Y}_{i+}$	Marginal fordeling af S_i^2	$\mathsf{E}\left[\frac{SAK_1}{N-k}\right]$	$\mathsf{E}\left[SAK_2/(k-1)\right]$
$\mathrm{N}\left(\mu, \sigma_i^2(\frac{1}{n_i} + \frac{1}{m})\right)$	$\mathrm{T}\left(2\alpha, \mu_0, \sqrt{\frac{\beta}{\alpha}\left(\frac{1}{n_i} + \frac{1}{m}\right)}\right)$	$\mathrm{RBet}\left(\alpha, \frac{n_i - 1}{2}, \frac{2\beta}{n_i - 1}\right)$	$\frac{\beta}{\alpha - 1}$	$\frac{\beta}{\alpha - 1}(1 + n_0/m)$

$$\sigma_0^2 = \beta/(\alpha - 1), \qquad \nu = 2\alpha$$

Momentestimation af parametre i endimensionale varianskomponentmodeller:
Normaltfordelte observationer, varierende varians i grupper.

$$
\begin{aligned}
\tilde{\mu} &= \overline{Y}_{++} \\
\tilde{\alpha} &= 1 + \left\{ Q_1 + \frac{1}{k-1}[N-k-\frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)^2] \right\} \Big/ \Big(Q_1 - 2 \Big) \\
\tilde{\beta} &= (\tilde{\alpha}-1)SAK_1/(N-k) \\
\tilde{m} &= n_0/(Q_2-1)
\end{aligned}
$$

med

$$
\begin{aligned}
Q_1 &= \{SAK_s/(k-1)\}/\{SAK_1/(N-k)\}^2 \\
Q_2 &= \{SAK_2/(k-1)\}/\{SAK_1/(N-k)\}
\end{aligned}
$$

Den vægtede gennemsnitlige gruppestikprøvestørrelse, n_0 er bestemt ved (5.1.9).

5.8 Referencer:

Christensen, R. (1987): *Plane Answers to Complex Problems*, Springer, New York

Gy, P.M. (1992): *Sampling of Heterogenous and Dynamic Material Systems.* Elsevier Scientific Publishing, Amsterdam.

Scheffé, H. (1959): *The Analysis of Variance*, John Wiley and Sons, New York

Shapiro, S. S. and Wilk, M. B. (1965): An analysis of variance test for normality (complete samples). *Biometrika*, **52**, pp. 591-611.

Wilk, M. B. and Gnanadesikan, R. (1968): Probability plotting methods for the analysis of data. *Biometrika*, **55**, pp 1-17.

Afsnit 6

Hierarkiske modeller for eksponentielle dispersionsmodeller

fil : hiergen1.tex 1998-04-19

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 ve 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

$$g(x;\vartheta_i) = d(x)\exp\{[\vartheta_i x - \kappa(\vartheta_i)]/\sigma^2\}\ , \tag{6.1.1}$$

der er fuldstændigt specificeret på nær den ukendte parameter ϑ_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 $\mathsf{E}[X_{ij}] = \mu_i = \tau(\vartheta_i)$, og med varians $\mathsf{V}[X_{ij}] = \sigma^2 V(\mu)$, hvor middelværdiafbildningen $\tau(\cdot)$ og enhedsvariansfunktionen $V(\cdot)$ som vanligt bestemmes ved

$$\begin{aligned} \mu &= \tau(\vartheta) = \kappa'(\vartheta) \\ V(\mu) &= \kappa''(\tau^{-1}(\mu)) \end{aligned} \tag{6.1.2}$$

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

$$Z_i = \sum_{j=1}^{n_i} X_{ij} = n_i \overline{X}_{i+}$$

sufficient for parameteren ϑ_i, og tætheden for Z_i er på formen

$$g_z(z;\vartheta_i) = h_z(z, n_i)\exp\{[\vartheta_i z - n_i\kappa(\vartheta_i)]/\sigma^2\}\ , \tag{6.1.3}$$

altså en eksponentiel dispersionsmodel med middelværdi $\mu_i = \mathsf{E}[Z_i] = n_i\tau(\vartheta_i)$ 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 σ^2 og variansfunktion $V(\mu)$) men med vægten $w_i = n_i$.

Tætheden for $\overline{X}_{i+}$ er således

$$g_{\overline{x}}(\overline{x}; \vartheta_i) = h_{\overline{x}}(\overline{x}, n_i) \exp\{n_i[\vartheta_i \overline{x} - \kappa(\vartheta_i)]/\sigma^2\} , \tag{6.1.4}$$

og der gælder

$$\begin{aligned} \mu_i &= \mathsf{E}\,[\overline{X}_{i+}] = \tau(\vartheta_i) \\ \mathsf{V}\,[\overline{X}_{i+}] &= \sigma^2 V(\mu_i)/n_i \ , \end{aligned}$$

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 middelværdiparametrisering. Specielt vil man således formulere homogenitetshypotesen ved middelværdiparameteren μ.

I afsnit 2.7.1 har vi i sætning 2.7.1 diskuteret test af homogenitetshypotesen

$$H_I : \mu_1 = \mu_2 = \cdots = \mu_k \tag{6.1.5}$$

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 ϑ_i, $i = 1, \ldots, k$ er realiserede værdier fra en fordeling af ϑ'er. Fordelingen af ϑ kaldes strukturfordelingen eller aprioriforde lingen af ϑ.

Under den tilfældige model vil vi som nævnt opfatte parameteren ϑ_i, der karakteriserer den betingede fordeling af observationerne $X_{i1}, X_{i2}, \ldots, X_{i_{n_i}}$ i den i'te gruppe, som en stokastisk variabel.

Vi vil yderligere antage, at $X_{i1}, X_{i2}, \ldots, X_{i_{n_i}}$ er betinget uafhængige (jvf definition 4.5.1 side 467) af ϑ_i i den simultane fordeling af $X_{i1}, X_{i2}, \ldots, X_{i_{n_i}}$ og ϑ_i.

For en udvalgt gruppe, i, er den betingede fordeling af $X_{i1}, X_{i2}, \ldots, X_{i_{n_i}}$ givet gruppeparameteren ϑ_i altså sådan at $X_{i1}, X_{i2}, \ldots, X_{i_{n_i}}$ er indbyrdes uafhængige med tæthed (6.1.1).

Den simultane fordeling af $X_{i1}, X_{i2}, \ldots, X_{i_{n_i}}$ for givet ϑ_i har da tætheden

$$f(x_{i1}, x_{i2}, \ldots, x_{i_{n_i}}|\vartheta) = \prod_{j=1}^{n_i} g(x_{ij}; \vartheta_i)$$

Definition 6.1.1 ***Konjugeret klasse af fordelinger***

Betragt en eksponentiel dispersionsmodel med tætheder af formen (6.1.1), dvs

$$g(x;\vartheta) = d(x)\exp\{[\vartheta x - \kappa(\vartheta)]/\sigma^2\} \tag{6.1.6}$$

hvor parameteren $\vartheta \in D$

Lad $\mathcal{M} = \tau(D)$ angive middelværdirummet svarende til denne fordeling.

For $m \in \mathcal{M}$ og $\gamma \in \Delta \subset \mathbb{R}_+$ vil

$$g(\vartheta; m, \gamma) = \frac{1}{C(m,\gamma)} \exp\{[\vartheta m - \kappa(\vartheta)]/\gamma\} \tag{6.1.7}$$

med

$$C(m,\gamma) = \int_D \exp\{[\vartheta m - \kappa(\vartheta)]/\gamma\}\, \mathrm{d}\vartheta$$

være en tæthed for fordelingen af ϑ.

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 kumulantfrembringende 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 ϑ, at for fastholdt γ vil familien af fordelinger af ϑ 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 ϑ 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 $\mu = \tau(\vartheta)$ svarende til den eksponentielle dispersionsmodel.

Det er klart, at når strukturfordelingen for ϑ er fastlagt, kan man blot indføre transformationen $\mu = \tau(\vartheta)$, hvorved strukturfordelingen for μ 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 ϑ være givet ved (6.1.7). Da gælder under visse regularitetsbetingelser

$$m = \mathsf{E}[\mu] \tag{6.1.8}$$

$$\gamma = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V(\mu)]}, \tag{6.1.9}$$

hvor $\mu = \mathsf{E}[X|\vartheta]$, og hvor $V(\mu)$ angiver variansfunktionen (6.1.2) i den betingede fordeling af observationerne.

BEMÆRK forskellen mellem symbolet $\mathsf{V}\,[\mu]$, der angiver variansen i aprio-rifordelingen af observationernes betingede middelværdier μ, 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

$$f(x|\vartheta) = d(x)\exp\{[\vartheta x - \kappa(\vartheta)]/\sigma^2\}$$

og tætheden for ϑ er

$$g(\vartheta; m, \gamma) = \frac{1}{C(m,\gamma)}\exp\{[\vartheta m - \kappa(\vartheta)]/\gamma\}$$

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 $g'_\vartheta(\vartheta; m, \gamma) \to 0$ for $\vartheta \to \pm\infty$.

Der gælder da

$$\int_{-\infty}^{\infty} g'_\vartheta(\vartheta; m, \gamma)d\vartheta = g(\vartheta; m, \gamma)\Big]_{\vartheta=-\infty}^{\infty} = 0\ ,$$

men

$$g'_\vartheta(\vartheta; m, \gamma) = \frac{m - \tau(\vartheta)}{\gamma}\, g(\vartheta; m, \gamma)\ ,$$

hvorfor

$$\int_{-\infty}^{\infty} g'_\vartheta(\vartheta; m, \gamma) = \int_{-\infty}^{\infty} \frac{m - \tau(\vartheta)}{\gamma}\, g(\vartheta; m, \gamma) = \frac{m - \mathsf{E}\,[\tau(\vartheta)]}{\gamma}\ ;$$

således at vi har

$$m = \mathsf{E}\,[\tau(\vartheta)]\ .$$

Tilsvarende får man

$$g''_{\vartheta^2}(\vartheta; m, \gamma) = \left[\frac{[m - \tau(\vartheta)]^2}{\gamma^2} - \frac{\kappa''(\vartheta)}{\gamma}\right] g(\vartheta; m, \gamma)\ ,$$

og idet

$$\int_{-\infty}^{\infty} g''_{\vartheta^2}(\vartheta; m, \gamma)d\vartheta = g'_\vartheta(\vartheta; m, \gamma)\Big]_{\vartheta=-\infty}^{\infty} = 0\ ,$$

får man da

$$\mathsf{V}\,[\tau(\vartheta)] = \gamma\mathsf{E}\,[\kappa''(\vartheta)]\ .$$

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 and F. Pukelsheim (1989). □

Bemærkning 1 ***Fortolkning ved varians indenfor grupper og varians mellem grupper***

Idet fordelingen af X antages at have dispersionsparameteren σ^2 (og vægten 1) har vi

$$\begin{aligned} \mathsf{E}[X|\vartheta] &= \mu = \tau(\vartheta) \\ \mathsf{V}[X|\vartheta] &= \sigma^2 V(\mu) = \sigma^2 \kappa''(\tau^{-1}(\mu)) \end{aligned}$$

og

$$m = \mathsf{E}[\tau(\vartheta)] = \mathsf{E}[\mathsf{E}[X|\vartheta]] = \mathsf{E}[\mu]$$

$$\gamma = \frac{\mathsf{V}[\tau(\vartheta)]}{\mathsf{E}[\kappa''(\vartheta)]} = \frac{\mathsf{V}[\mathsf{E}[X|\vartheta]]}{\mathsf{E}[\mathsf{V}[X|\vartheta]]/\sigma^2}$$

□

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 μ 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)

$$\mathsf{E}[X_{ij}] = m \tag{6.1.10}$$

$$\mathsf{COV}[X_{ij}, X_{hl}] = \begin{cases} \mathsf{E}[V(\mu)](\sigma^2 + \gamma) & \text{for } (i,j) = (h,l) \\ \mathsf{V}[\mu] = \gamma \mathsf{E}[V(\mu)] & \text{for } i = h,\ j \neq l \\ 0 & \text{for } i \neq h \end{cases}$$

Observationer X_{ij} og X_{il} i samme gruppe er således indbyrdes korrelerede med intraklassekorrelationen

$$\rho = \frac{\mathsf{V}[\mu]}{\sigma^2 \mathsf{E}[V(\mu)] + \mathsf{V}[\mu]} = \frac{\gamma \mathsf{E}[V(\mu)]}{\mathsf{E}[V(\mu)](\sigma^2+\gamma)} = \frac{\gamma}{\sigma^2+\gamma} \tag{6.1.11}$$

Omvendt kan signal/støj forholdet, γ, udtrykkes ved intragruppekorrelationen ρ som

$$\gamma = \frac{\rho}{1-\rho}\,\sigma^2 \tag{6.1.12}$$

Momenterne i fordelingen af gruppegennemsnittene, $\overline{X}_{i+}$, er

$$\mathsf{E}[\overline{X}_{i+}] = m \tag{6.1.13}$$

$$\mathsf{V}[\overline{X}_{i+}] = \mathsf{E}[V(\mu)](\gamma + \frac{\sigma^2}{n_i}) \tag{6.1.14}$$

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

$$\mathsf{E}[\overline{X}_{i+}] = \mathsf{E}[\,\mathsf{E}[\overline{X}_{i+}|\mu]\,] = \mathsf{E}[\mu] = m$$

Endvidere har man

$$\begin{aligned} \mathsf{V}[\overline{X}_{i+}] &= \mathsf{E}[\mathsf{V}[\overline{X}_{i+}|\mu]] + \mathsf{V}[\mathsf{E}[\overline{X}_{i+}|\mu]] \\ &= \frac{\sigma^2}{n_i}\,\mathsf{E}[V(\mu)] + \mathsf{V}[\mu] \\ &= \mathsf{E}[V(\mu)](\gamma + \frac{\sigma^2}{n_i}) \end{aligned} \tag{6.1.15}$$

□

Til brug for en simpel parameterestimation anfører vi

Sætning 6.1.3 *Forventningsværdi af kvadratafvigelssummen svarende til variationen mellem grupper under den tilfældige model*

Lad $\overline{X}_{i+}$ være som i ovenstående sætning, og sæt som vanligt

$$SAK_2 = \sum_{i=1}^{k} n_i(\overline{X}_{i+} - \overline{X}_{++})^2 \ . \tag{6.1.16}$$

Da gælder under den tilfældige model

$$\mathsf{E}\,[SAK_2] = (k-1)\mathsf{E}\,[V(\mu)](\sigma^2 + n_0\gamma) \ , \tag{6.1.17}$$

hvor n_0 er givet ved (5.1.9) og hvor γ 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

$$\begin{aligned} \mathsf{E}\,[SAK_2] &= \sum_i n_i(1 - \frac{n_i}{N})\mathsf{V}\,[\overline{X}_{i+}] \\ &= (k-1)\{\sigma^2\mathsf{E}\,[V(\mu)] + n_0\mathsf{V}\,[\mu]\} \end{aligned}$$

□

Relationen (6.1.17) sammen med

$$\mathsf{E}\,[\overline{X}_{++}] = \mathsf{E}\,[\mu] = m \tag{6.1.18}$$

kan benyttes til at beregne et simpelt estimat for parametrene i aprioriforfordelingen.

Vi bemærker, at momentmetoden anvendt på SAK_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 $\mathsf{V}[\overline{X}_{i+}]$. På grund af den simple opspaltning (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 simplest for de heltallige gruppetotaler, Z_i. Den praktiske interesse knytter sig imidlertid sædvanligvis til middelværdiparameteren μ svarende til en dispersionsmodel for $\overline{X}_{i+} = Z_i/n_i$.

Vi har derfor fundet det nødvendigt at operere med to parallelle parametriseringer af apriorifordelingen.

Den ene parametrisering, (sædvanligvis betegnet med symbolerne α og β) 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 μ og $V(\mu)$) af den betingede fordeling af observationerne, benytter en middelværdiparametrisering for fordelingen af μ (ved parametrene m og γ).

6.2 Bernoullifordelingen

Såfremt $X_{ij}|\mu_i \in \mathrm{B}(1, \mu_i)$, finder vi $Z_i|\mu_i \in \mathrm{B}(n_i, \mu_i)$

Sætning 6.2.1 ***Test for homogenitet***

Kvotienttesteststørrelsen for hypotesen (6.1.5) er

$$G^2(H_I) = \sum_i n_i d(h_i; \overline{h}_+) \tag{6.2.1}$$

hvor

$$h_i = \frac{z_i}{n_i} = \overline{x}_{i+} \quad \text{og} \quad \overline{h}_+ = \frac{\sum z_i}{\sum n_i} = \overline{x}_{++} \tag{6.2.2}$$

angiver de observerede relative hyppigheder i henholdsvis den i'te gruppe og i totalmaterialet og $d(h_i;\overline{h}_+)$ angiver enhedsdeviansen for binomialfordelingen,

$$d(h_i;\overline{h}_+) = 2\left[h_i \ln(h_i/\overline{h}_+) + (1-h_i)\ln[(1-h_i)/(1-\overline{h}_+)]\right]$$

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:
Den anførte teststørrelse er devianstesstørrelsen (Tabel 2.3) svarende til hypotesen H_I imod den fulde model $\overline{X}_{i+} \in \mathrm{B}(n_i,\mu_i)/n_i$. □

Bemærkning 1 ***Pearson teststørrelsen***

Pearson-teststørrelsen for hypotesen (6.1.5) er

$$X^2 = \sum_{i=1}^{k} n_i \frac{(h_i-\overline{h}_+)^2}{\overline{h}_+(1-\overline{h}_+)} = \sum_{i=1}^{k} \frac{(Z_i - n_i\overline{h}_+)^2}{n_i\overline{h}_+(1-\overline{h}_+)}\,, \quad f = k-1 \tag{6.2.3}$$

er asymptotisk ækvivalent med kvotienttestet. Dette test måler direkte variationen imellem grupper, SAK_2, i forhold til den estimerede binomialvarians, $\overline{h}_+(1-\overline{h}_+)$. □

Såfremt det findes begrundet at afvise hypotesen (6.1.5), kan man f.eks. vælge at modellere fordelingen af μ ved en $\mathrm{Be}(\alpha,\beta)$-fordeling.

Sætning 6.2.2 ***Den marginale fordeling af gruppetotal ved beta-binomial sampling***

Såfremt $Z|\mu \in \mathrm{B}(n,\mu)$ og $\mu \in \mathrm{Be}(\alpha,\beta)$, da er den marginale fordeling af Z en $\mathrm{Pl}(n,\alpha,\alpha+\beta)$-fordeling.

Der gælder

$$\begin{aligned} \mathsf{E}[Z] &= n\pi \\ \mathsf{V}[Z] &= n\,\frac{\pi(1-\pi)}{1+\gamma}\,[1+n\gamma] \end{aligned} \tag{6.2.4}$$

med

$$\pi = \alpha/(\alpha+\beta), \quad \gamma = \frac{1}{\alpha+\beta} \tag{6.2.5}$$

For den relative hyppighed $Y = Z/n$ gælder:

$$\begin{aligned} \mathsf{E}[Y] &= \pi \\ \mathsf{V}[Y] &= \frac{\pi(1-\pi)}{1+\gamma}\left(\gamma + \frac{1}{n}\right) \end{aligned} \tag{6.2.6}$$

Bevis:
For $\mu \in \mathrm{Be}(\alpha, \beta)$ har vi

$$\mathsf{E}[\mu] = \pi = \frac{\alpha}{\alpha+\beta} \tag{6.2.7}$$

$$\mathsf{V}[\mu] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\pi(1-\pi)}{\alpha+\beta+1} \tag{6.2.8}$$

(Familien af betafordelinger kan opfattes som en eksponentiel dispersionsmodel med middelværdiparameter $\pi = \alpha/(\alpha+\beta)$, variansfunktion $V_{Beta}(\pi) = \pi(1-\pi)$ og dispersionsparameter $\sigma^2 = 1/(\alpha+\beta+1)$).

Vi betragter nu enhedsvariansfunktionen, $V_{Bin}(\mu)$, for binomialfordelingen, bestemt ved $V_{Bin}(\mu) = \mu(1-\mu)$. Der gælder

$$\gamma = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V_{Bin}(\mu)]} = \frac{\mathsf{V}[\mu]}{\mathsf{E}[\mu(1-\mu)]} = \frac{1}{\alpha+\beta}, \tag{6.2.9}$$

idet

$$\mathsf{E}[V_{Bin}(\mu)] = \mathsf{E}[\mu(1-\mu)] = \pi(1-\pi)\frac{\alpha+\beta}{\alpha+\beta+1}$$

Vi har således

$$\mathsf{E}[V_{Bin}(\mu)] = \frac{\pi(1-\pi)}{1+\gamma} = \frac{V_{Bin}(\pi)}{1+\gamma}.$$

Sætningen følger nu ved indsætttelse i (6.1.14). □

Bemærkning 1 ***Overdispersion i forhold til binomialfordelingen***

Vi bemærker indledningsvis, at parameteren π netop angiver $\mathsf{E}[\mu]$.

Opspaltningen svarende til variationen indenfor grupper og variationen mellem grupper giver størrelserne

$$\begin{aligned}\mathsf{E}[\mathsf{V}[Z|\mu]] &= n\mathsf{E}[V_{Bin}(\mu)] = nV_{Bin}(\pi)\frac{1}{\gamma+1}\\ \mathsf{V}[\mathsf{E}[Z|\mu] &= n^2\mathsf{V}[\mu] = n^2\pi(1-\pi)\frac{\gamma}{\gamma+1} = n^2V_{Bin}(\pi)\frac{\gamma}{\gamma+1}\end{aligned}$$

Udtrykt ved intraklassekorrelationen

$$\rho = \frac{\gamma}{\gamma+1} = \frac{\alpha+\beta}{\alpha+\beta+1}$$

finder vi

$$\mathsf{E}[\mathsf{V}[Z|\mu]] = n\mathsf{E}[V_{Bin}(\mu)] = nV_{Bin}(\pi)(1-\rho)$$

$$\mathsf{V}[\mathsf{E}[Z|\mu] = n^2\mathsf{V}[\mu] = n^2V_{Bin}(\pi)\rho$$

og

$$\mathsf{V}[Z] = nV_{Bin}(\pi)\{1 + (n-1)\rho\}$$

Lader vi $Y = Z/n$ angive den observerede relative hyppighed finder vi tilsvarende

$$\begin{aligned} \mathsf{V}[Y] &= \frac{V_{Bin}(\pi)}{\gamma+1}\left(\gamma + \frac{1}{n}\right) \qquad (6.2.10) \\ \mathsf{E}[\mathsf{V}[Y|\mu]] &= \frac{V_{Bin}(\pi)}{n}\,\frac{1}{\gamma+1} \\ \mathsf{V}[\mathsf{E}[Y|\mu]] &= \mathsf{V}[\mu] = V_{Bin}(\pi)\frac{\gamma}{\gamma+1} \end{aligned}$$

dvs

$$\mathsf{V}[Y] = V_{Bin}(\pi)\{\rho + \frac{1-\rho}{n}\}$$

Vi bemærker, at for $n \to \infty$ vil $\mathsf{V}[Y]$ nærme sig variansen $\mathsf{V}[\mu] = V_\beta(\pi)$ i fordelingen af μ, hvilket er i overensstemmelse med at fordelingen af Y vil nærme sig fordelingen af μ.

For begrænsede værdier af n vil fordelingen af Z have en overdispersion i forhold til binomialfordelingen, $\mathrm{B}(n,\pi)$, med samme antalsparameter og med sandsynlighedsparameter π svarende til $\mathsf{E}[\mu]$. Overdispersionen er

$$\sigma^2 = \frac{\mathsf{V}[Z]}{nV_{Bin}(\pi)} = \frac{1+n\gamma}{1+\gamma} \qquad (6.2.11)$$

Tilsvarende vil fordelingen af $Y = Z/n$ have overdispersionen $\sigma^2 = (1 + n\gamma)/(1+\gamma)$ i forhold til $\mathrm{B}(n,\pi)/n$-fordelingen. □

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

$$\mathsf{V}[Y] = V_{Bin}(\pi)/n_{eff},$$

hvor den effektive stikprøvestørrelse, n_{eff} er bestemt ved

$$n_{eff} = \frac{n}{\sigma^2} = n\,\frac{\gamma+1}{n\gamma+1} \tag{6.2.12}$$

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 $\mathrm{B}(n_{eff},\pi)/n_{eff}$ fordelt variabel.

Jo større værdi af intraklassekorrelationen ρ, desto mindre bliver den effektive stikprøvestørrelse, n_{eff}.

Ved analyse af strukturer i middelværdiparameteren π kan man altså med rimelig tilnærmelse benytte generaliserede lineære modeller svarende til binomialfordelingen, idet man blot erstatter 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 π ved bestemmelse af konfidensintervaller for en tilsvarende binomialfordelt størrelse med stikprøvestørrelsen n_{eff}. □

Bemærkning 3 ***Fortolkning af parameteren*** γ

Parameteren $\gamma = \mathsf{V}[\mu]/\mathsf{E}[V_{Bin}(\mu)]$ er sammen med stikprøvestørrelsen bestemmende for overdispersionen.

Vi har

$$\gamma^{-1} = \left[\frac{V_{Bin}(\mathsf{E}[\mu])}{\mathsf{V}[\mu]} - 1\right]$$

Parameteren γ måler således afvigelsen fra den rene binomialfordeling af Z. For $\pi \in]0,1[$, vil $V[\mu] \to 0$ være enbetydende med $\gamma \to 0$, og fordelingen af Z vil nærme sig en $B(n,\pi)$-fordeling.

For $\gamma > \pi$ vil fordelingen af μ være J-formet med modus for $\mu = 0$. For $\gamma > 1-\pi$ vil fordelingen af μ være J-formet med modus for $\mu = 1$. For $\gamma < \min\{\pi, 1-\pi\}$ er fordelingen af μ 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 $Z_i \in Pl(n_i, \alpha, \alpha+\beta)$

Momentestimaterne for $\pi = \dfrac{\alpha}{\alpha+\beta}$ og $\gamma = \dfrac{1}{\alpha+\beta}$ er:

$$\begin{aligned} \tilde{\pi} &= \overline{h}_+ \\ \tilde{\gamma} &= \frac{s_2^2 - \overline{h}_+(1-\overline{h}_+)}{n_0\overline{h}_+(1-\overline{h}_+) - s_2^2} \end{aligned} \tag{6.2.13}$$

hvor

$$s_2^2 = sak_2/(k-1) = \sum_{i=1}^{k} n_i(h_i - \overline{h}_+)^2/(k-1) \tag{6.2.14}$$

$$h_i = z_i/n_i; \quad \overline{h}_+ = \sum_{i=1}^{k} z_i / \sum_{i=1}^{k} n_i \tag{6.2.15}$$

og hvor den vægtede gennemsnitlige stikprøvestørrelse, n_0, er givet ved (5.1.9).

Momentestimatorerne for α og β bliver derfor

$$\begin{aligned} \tilde{\alpha} &= \overline{h}_+ \frac{n_0\overline{h}_+(1-\overline{h}_+) - s_2^2}{s_2^2 - \overline{h}_+(1-\overline{h}_+)} \\ \tilde{\beta} &= (1-\overline{h}_+)\frac{n_0\overline{h}_+(1-\overline{h}_+) - s_2^2}{s_2^2 - \overline{h}_+(1-\overline{h}_+)} \end{aligned} \tag{6.2.16}$$

Bevis:
Overspringes □

Bemærkning 1 *Singulær løsning*

Hvis $s_2^2 < \overline{h}_+(1-\overline{h}_+)$, 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 μ estimeres til at være en étpunktfordeling, og fordelingen af Z_i bliver da en $B(n,\pi)$-fordeling. Parametriseringen ved (π,γ) tillader netop estimation af π, 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 $Z_i \in \mathrm{Pl}(n_i, \alpha, \alpha+\beta)$

Maksimum-likelihood estimaterne $(\hat{\pi},\hat{\gamma})$ for $\pi = \dfrac{\alpha}{\alpha+\beta}$ og $\gamma = \dfrac{1}{\alpha+\beta}$ findes da ved at maksimere

$$l(\pi,\gamma;z_1,z_2,\ldots,z_k) = = \sum_{i=1}^{k}\Big[\sum_{\nu=0}^{z_i-1}\ln(\pi+\nu\gamma) + \sum_{\nu=0}^{n_i-z_i-1}\ln(1-\pi+\nu\gamma) - \sum_{\nu=0}^{n_i-1}\ln(1+\nu\gamma)\Big] \tag{6.2.17}$$

med hensyn til π og γ.

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 estimaterne 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 $\tilde{\pi}$ og $\tilde{\gamma}$ bestemt ved (6.2.13).

Såfremt $s_2^2 < \overline{h}_+(1-\overline{h}_+)$, må man formode, at maksimum ligger på randen af området svarende til $\hat{\gamma} = 0$ (omend dette ikke er alment bevist). For $\gamma = 0$ er maksimum-likelihood estimatoren for π bestemt ved $\hat{\pi} = \overline{h}_+$. □

Bemærkning 2 ***Kvotienttest for homogenitet***
Kvotienteststørrelsen for hypotesen

$$\mathrm{H}_{II} : \gamma = 0 \quad \text{med alternativet} \quad \overline{\mathrm{H}}_{II} : \gamma > 0$$

er

$$Z = 2\sum_{i=1}^{k}\left[\sum_{\nu=0}^{z_i-1}\ln\left(\frac{\hat{\pi}+\nu\hat{\gamma}}{\overline{h}_+}\right) + \sum_{\nu=0}^{n_i-z_i-1}\ln\left(\frac{1-\hat{\pi}+\nu\hat{\gamma}}{1-\overline{h}_+}\right) - \sum_{\nu=0}^{n_i-1}\ln(1+\nu\hat{\gamma})\right] \tag{6.2.18}$$

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 $\breve{\pi} = \overline{h}_+$. Ved indsættelse af $\gamma = 0$ og $\breve{\pi} = \overline{h}_+$ i (6.2.17) fås

$$l(\breve{\pi}, 0; z_1, z_2, \ldots, z_k) = \sum_{i=1}^{k}\left[z_i \ln(\overline{h}_+) + (n_i - z_i)\ln(1-\overline{h}_+)\right]$$

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 á 770 låg fra hvert af 229 produktionspartier
(Kilde: J.H Ford (1951))

Antal afvigende z_i	Antal stikprøver .	B(770, 0.0014) Forv.	$\frac{(obs-forv)^2}{forv}$	Pl(770, 0.460, 319.1) Forv.	$\frac{(obs-forv)^2}{forv}$
0	131	75.47	40.85	130.64	0.00
1	38	83.83	25.06	42.52	0.48
2	28	46.50	7.36	21.96	1.66
3	11	17.17	2.22	12.74	0.24
4	4	4.75		7.79	1.85
5	5	1.05		4.91	0.00
6	5	0.19	37.19	3.16	
7	2	0.03		2.06	1.91
8	3	-		1.36	
9	2	-		0.90	
Ialt	229		112.68		6.14

Man finder ialt $z_+ = \sum_{i=1}^{229} z_i = 254$ afvigende låg i de 229 prøver, dvs $\overline{z}_+ = 254/229 = 1.11$ afvigende/pr prøve, hvorfor man har $\overline{h}_+ = 1.11/770 = 0.0014$ afvigende/pr låg. Endvidere finder man

$$s_2^2 = \sum_{i=1}^{229} 770(h_i - \overline{h}_+)^2/228 = 0.00449.$$

Under den systematiske model finder man Pearson-teststørrelsen for samme andel afvigende i de 229 partier $X^2 = 711.71$ med $f = 228$. Idet $\mathrm{E}\,[\chi^2(f)] = f$, og $\mathrm{V}\,[\chi^2(f)] = 2f$, finder man, at den observerede værdi af X^2 er væsentligt større, end

$$\chi^2(228)_{0.95} \approx \mathrm{E}\,[\chi^2(228)] + 1.64\sqrt{\mathrm{V}\,[\chi^2(228)]} = 228 + 35 = 263$$

Der synes således ikke at være belæg for at modellere observationerne som realisationer af samme binomialfordeling.

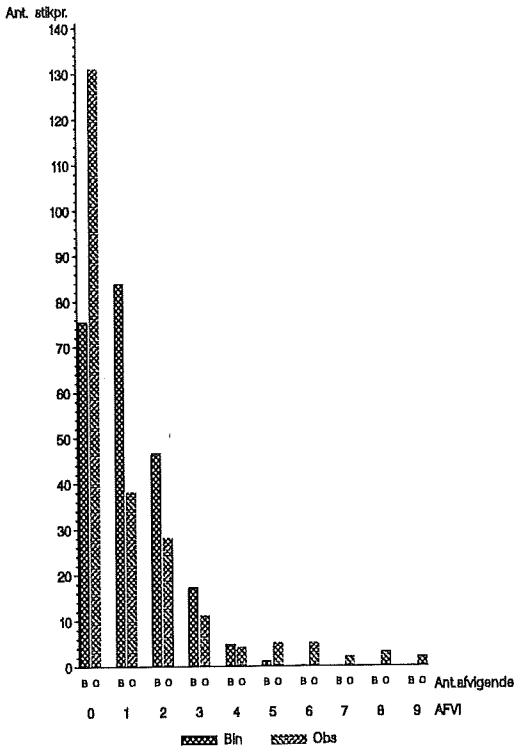
Vi bemærker dog, at approximationen ved χ^2-fordelingen er ganske ringe. X^2-teststørrelsen består af 229 led af formen $(z_i - 1.11)^2/1.11$ og χ^2-approximationen fremkommer essentielt ved at approximere hvert led med en $\mathrm{N}(0,1)^2$-størrelse. Men de mulige værdier af z_i er $0, 1, \ldots, 770$ 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

$(0 - 1.11)^2/1.11$, $(1 - 1.11)^2/1.11$ og $(2 - 1.11)^2/1.11$. Det er således en grov tilnærmelse at approximere denne fordeling med den kontinuerte χ^2-fordeling.

Havde vi i stedet bestemt kvotienttesstørrelsen (6.2.1), havde vi fundet størrelsen $G^2(H_I) = 273.4$. 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 á 770 låg sammenlignet med en B(770, 0.0014) fordeling.



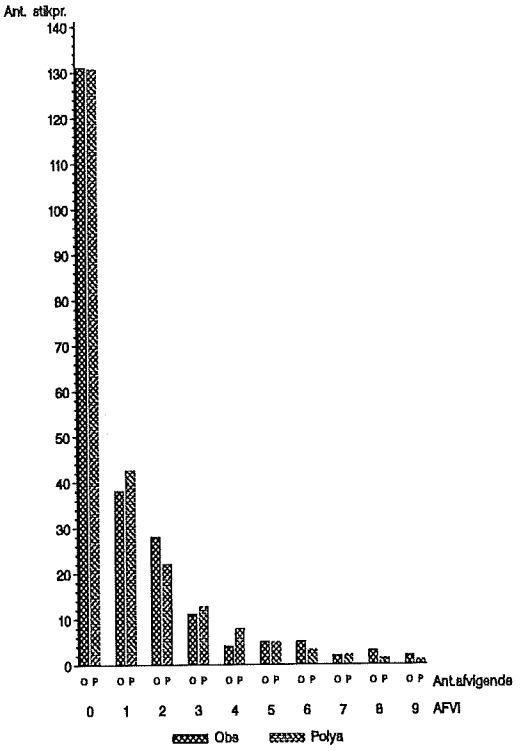
I tabellen over observationerne er endvidere angivet de forventede antal stikprøver med henholdsvis $0, 1, 2, \ldots$ afvigende låg, beregnet under binomialfordelingsforudsætningen. Det ses, at det sædvanlige χ^2 - test for fordelingstype (Statistik 1, afsnit 4.2.2) fører til $\chi^2 = 112.68$ med $f = 3$. 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 $\mathrm{Be}(\alpha, \beta)$- fordeling, vil den marginale fordeling af Z (jvf. sætn. 6.2.2) kunne beskrives ved en $\mathrm{Pl}(770, \alpha, \alpha + \beta)$ fordeling.

Momentestimaterne for α og β 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 maksimeringsrutine finder man maksimumlikelihood 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 frekvensfunktionen for en Pl(770,0.459,319.1)-fordeling. Der ses at være en væsentlig 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 χ^2- test for fordelingstype 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 Polyafordelte 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

$$\mu = \mathsf{E}\,[X_{ij}|p] = \frac{p}{1-p}\,, \tag{6.3.1}$$

dvs

$$p = \frac{\mu}{1+\mu}\,. \tag{6.3.2}$$

Familien har enhedsvariansfunktionen

$$V_{NB}(\mu) = \mu(1+\mu) = \frac{p}{(1-p)^2} \tag{6.3.3}$$

Den kanoniske linkfunktion er

$$\eta(\mu) = \ln\left(\frac{\mu}{1+\mu}\right) \tag{6.3.4}$$

For $Z \in \text{NB}^*(n,p)$ gælder

$$\begin{aligned} \mathsf{E}\,[Z] &= n\mu \\ \mathsf{V}\,[Z] &= nV_{NB}(\mu) = n\mu(1+\mu) \end{aligned}$$

For $Y = Z/n$ har vi derfor:

$$\begin{aligned} \mathsf{E}\,[Y] &= \mu = \frac{p}{1-p} \\ \mathsf{V}\,[Y] &= \frac{V_{NB}(\mu)}{n} = \frac{\mu(1+\mu)}{n} \end{aligned}$$

med μ og $V_{NB}(\mu)$ givet ved (6.3.1) og (6.3.3)

Sætning 6.3.1 Test for homogenitet

Kvotienttesteststørrelsen for hypotesen (6.1.5) er

$$G^2(H_I) = \sum_{i=1}^{k} n_i d(y_i; \overline{y}_+) \tag{6.3.5}$$

med $y_i = z_i/n_i$ $(= \overline{x}_{i+})$ og med enhedsdeviansen $d(y_i; \overline{y}_+)$ givet ved

$$d(y_i; \overline{y}_+) = 2\big[y_i \ln(y_i/\overline{y}_+) - (y_i + 1)\ln[(1 + y_i)/(1 + \overline{y}_+)]\big]$$

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

$$X^2 = \sum_{i=1}^{k} n_i \frac{(y_i - \overline{y}_+)^2}{\overline{y}_+(1 + \overline{y}_+)} \tag{6.3.6}$$

Testet er asymptotisk ækvivalent med kvotienttestet. Testet måler variationen imellem gruppegennemsnittene, SAK_2 , i forhold til den estimerede varians, $V_{NB}(\overline{y}_+)$ i den negative binomialfordeling.

□

Såfremt man ønsker at modellere eventuelle forskelle mellem gruppperne ved en tilfældig model, kan man vælge at modellere fordelingen af p ved en $\mathrm{Be}(\alpha, \beta)$- fordeling. Den marginale fordeling af Z_i bliver da en $\mathrm{NPl}(n_i, \beta, \alpha+\beta)$- fordeling.

Lemma 6.3.1 Momenter i fordelingen af $\mu = \mathrm{p}/(1-\mathrm{p})$ ved beta-negativ binomial sampling

Såfremt $Z|p \in \mathrm{NB}^*(n, p)$ og $p \in \mathrm{Be}(\alpha, \beta)$, da gælder for fordelingen af $\mu = p/(1-p)$:

For $\beta \leq 1$ har fordelingen ikke nogen middelværdi. Såfremt $1 < \beta$, har fordelingen af μ middelværdien

$$\psi = \mathsf{E}[\mu] = \mathsf{E}\left[\frac{p}{1-p}\right] = \frac{\alpha}{\beta - 1} \tag{6.3.7}$$

For $\beta \leq 2$ har fordelingen ingen varians. Såfremt $2 < \beta$, har fordelingen af μ variansen

$$\mathsf{V}[\mu] = \frac{V_{NB}(\psi)}{\beta - 2} = \frac{\gamma}{1-\gamma} V_{NB}(\psi) \tag{6.3.8}$$

Den marginale middelværdi af variansfunktionen $V_{NB}(\mu)$ er

$$\mathsf{E}[V_{NB}(\mu)] = \frac{\beta - 1}{\beta - 2} V_{NB}(\psi) = \frac{1}{1-\gamma} V_{NB}(\psi) \ , \tag{6.3.9}$$

hvor "signal/støj forholdet," γ, eksisterer såfremt $2 < \beta$. Signal/støjforholdet γ er givet ved

$$\gamma = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V_{NB}(\mu)]} = \frac{1}{\beta - 1} \tag{6.3.10}$$

med $0 < \gamma < 1$.

Endelig er intraklassekorrelationen, ρ, givet ved

$$\rho = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V_{NB}(\mu)] + \mathsf{E}[\mathsf{V}[\mu]]} = \frac{1}{\beta} \tag{6.3.11}$$

for $2 < \beta$.

Bevis:

Ved integration i betafordelingen finder man

$$\mathsf{V}[\mu] = \mathsf{V}\left[\frac{p}{1-p}\right] = \frac{\alpha}{\beta-1}\left(1+\frac{\alpha}{\beta-1}\right)\frac{1}{\beta-2} = \frac{\psi(1+\psi)}{\beta-2},$$

der ved indsættelse af $\psi(1+\psi) = V_{NB}(\psi)$ (jvf(6.3.3)), netop er (6.3.8).

Tilsvarende finder man

$$\mathsf{E}[V_{NB}(\mu)] = \mathsf{E}[\mu(1+\mu)] = \frac{\alpha}{\beta-2}\left(1+\frac{\alpha}{\beta-1}\right) = \frac{\beta-1}{\beta-2}\psi(1+\psi).$$

Idet vi atter udnytter relationen $\psi(1+\psi) = V_{NB}(\psi)$ (jvf (6.3.3)), fås (6.3.9).

Endelig fås udtrykket for intraklassekorrelationen ρ ved indsættelse af udtrykket for γ i (6.1.11).

□

Sætning 6.3.2 Den marginale fordeling af gruppetotal og gruppegennemsnit ved negativ binomial-beta sampling

Såfremt $Z|p \in \mathrm{NB}^*(n,p)$, og $p \in \mathrm{Be}(\alpha,\beta)$, da er den marginale fordeling af Z en $\mathrm{NPl}^*(n,\alpha,\alpha+\beta)$-fordeling.

Den negative polyafordeling, $\mathrm{NPl}(n,\beta,\alpha+\beta)$-fordelingen, er beskrevet i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

Såfremt $\beta > 2$ gælder:

$$\mathsf{E}[Z] = n\psi \tag{6.3.12}$$

(6.3.13)

$$\mathsf{V}[Z] = n\,\frac{\psi(1+\psi)}{1-\gamma}[1+n\gamma] \tag{6.3.14}$$

med ψ og γ givet ved (6.3.7) og (6.3.10).

Middelværdien $\mathsf{E}[Z]$ eksisterer dog blot $\beta > 1$.

For den relative hyppighed $Y = Z/n$ gælder tilsvarende for $\beta > 2$:

$$\mathsf{E}[Y] = \psi \tag{6.3.15}$$

$$\mathsf{V}[Y] = \frac{\psi(1+\psi)}{1-\gamma}\left[\gamma + \frac{1}{n}\right] \tag{6.3.16}$$

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.

□

Bemærkning 1 ***Opspaltning af den totale varians***

Opspaltningen af den totale varians (6.3.14)

$$\mathsf{V}[Z] = n\,\frac{V_{NB}(\psi)}{1-\gamma}\,[1+n\gamma]$$

i komponenter svarende til den gennemsnitlige varians indenfor grupper og variansen mellem gruppemiddelværdier, giver størrelserne

$$\mathsf{E}[\mathsf{V}[Z|\mu]] = n\,\mathsf{E}[V_{NB}(\mu)] = n\,\frac{V_{NB}(\psi)}{1-\gamma}$$

$$\mathsf{V}[\mathsf{E}[Z|\mu] = n^2\,\mathsf{V}[\mu] = n^2\gamma\,\frac{V_{NB}(\psi)}{1-\gamma}$$

□

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 $p = 1/(1 + \psi - \gamma)$).

Parametriseringen ved ψ og γ svarer til området $\{2 < \alpha\} \times \{2 < \beta\}$. Parametriseringen 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 parametriseringen 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 γ som en kvotient mellem varianser.

□

Bemærkning 3 ***Parametrisering ved momenter i betafordelingen***

I lighed med det parametriseringen ved beta-binomial fordelingen kunne man betragte parametriseringen

$$\pi = \alpha/(\alpha + \beta), \quad \gamma^* = 1/(\alpha + \beta) \tag{6.3.17}$$

hvor parameteren π angiver $\mathsf{E}[p]$, og γ^* er bestemt ved

$$\gamma^* = \frac{\mathsf{V}[p]}{\mathsf{E}[p(1-p)]}$$

Der gælder

$$\psi = \frac{1-\pi}{1-\pi-\gamma^*} \qquad \gamma = \frac{\gamma^*}{1-\pi-\gamma^*}$$
$$\pi = \frac{1+\psi}{1+\psi+\gamma} \qquad \gamma^* = \frac{1}{1+\psi+\gamma}$$

□

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

$$\sigma^2 = \frac{\mathsf{V}[Z]}{nV_{NB}(\psi)} = \frac{1+n\gamma}{1-\gamma} \tag{6.3.18}$$

Tilsvarende vil fordelingen af $Y = Z/n$ have overdispersionen $\sigma^2 = (1 + n\gamma)/(1-\gamma)$ i forhold til NB*$(n, \psi/(1+\psi))/n$-fordelingen.

Den marginale varians af den gennemsnitlige rate Y kan altså udtrykkes som

$$\mathsf{V}[Y] = V_{NB}(\psi)/n_{eff} \ ,$$

hvor den effektive stikprøvestørrelse, n_{eff} er bestemt ved

$$n_{eff} = \frac{n}{\sigma^2} = n\,\frac{1-\gamma}{1+n\gamma} \tag{6.3.19}$$

For ψ fastholdt, $0 < \psi < \infty$ og $\gamma \to 0$ vil $\mathsf{V}[Z]$ nærme sig $n\psi(1+\psi)$, der netop er variansen i en NB*$(n, \psi/(1+\psi))$-fordeling, svarende til at fordelingen af Z nærmer sig en NB*$(n, \psi/(1+\psi))$-fordeling.

□

Sætning 6.3.3 Maksimum-likelihood estimation i den negative Polyafordeling

Er endnu ikke udarbejdet

□

Bemærkning 1 ***Bestemmelse af maksimum-likelihood estimaterne***

Er endnu ikke udarbejdet

□

Bemærkning 2 ***Kvotienttesstørrelsen for homogenitetstest***

Er endnu ikke udarbejdet

□

Sætning 6.3.4 Momentestimation i den negative Polyafordeling

Lad $Z_1, Z_2, \ldots, Z_k$ være uafhængige variable, hvor $Z_i \in \mathrm{NPI}^*(n_i, \alpha, \alpha+\beta)$
Momentestimaterne for $\psi = \alpha/(\beta-1)$ og $\gamma = 1/(\beta-1)$ er da

$$\tilde{\psi} = \overline{y}_+ \tag{6.3.20}$$

$$\tilde{\gamma} = \frac{s_2^2 - \overline{y}_+(1+\overline{y}_+)}{s_2^2 + n_0\overline{y}_+(1+\overline{y}_+)} \tag{6.3.21}$$

med

$$s_2^2 = \sum_{i=1}^{k} n_i(y_i - \overline{y}_+)^2/(k-1)\ ,$$

hvor $y_i = z_i/n_i$, og hvor den vægtede gennemsnitlige stikprøvestørrelse, n_0, er bestemt ved (5.1.9).

De tilsvarende estimater for α og β er

$$\tilde{\alpha} = \overline{y}_+\left(1 + \frac{n_0+1}{s_2^2 - \overline{y}_+(1+\overline{y}_+)}\right) \tag{6.3.22}$$

$$\tilde{\beta} = 2 + \frac{n_0+1}{s_2^2 - \overline{y}_+(1+\overline{y}_+)} \tag{6.3.23}$$

Bevis:
Følger af sætning 6.1.3.

□

Bemærkning 1 *Singularitet ved momentestimation*

Hvis $s_2^2 < \overline{y}_+(1+\overline{y}_+)$ bliver $\tilde{\gamma} < 0$. Dette kan tages som udtryk for enten, at $\gamma = 0$ svarende til en etpunktsfordeling af μ (og p), eller at $\tilde{\alpha} < 2$ svarende til at fordelingen af μ 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 $X_{ij}|\mu_i \in \mathrm{P}(\mu_i)$ finder vi for $Z_i = X_{i1} + \cdots + X_{in_i}$, at $Z_i|\mu_i \in \mathrm{P}(n_i\mu_i)$.

Familien af Poissonfordelinger er en additit eksponentiel dispersionsmodel med middelværdi

$$\mu = \mathsf{E}\,[X_{ij}|\mu] \tag{6.4.1}$$

Familien har enhedsvariansfunktionen

$$V_P(\mu) = \mu \tag{6.4.2}$$

Den kanoniske linkfunktion er

$$\eta(\mu) = \ln(\mu) \tag{6.4.3}$$

For $Z \in \mathrm{P}(n\mu)$ gælder

$$\begin{aligned} \mathsf{E}\,[Z] &= n\mu \\ \mathsf{V}\,[Z] &= nV_P(\mu) = n\mu \end{aligned}$$

For $Y = Z/n$ har vi derfor:

$$\mathsf{E}\,[Y] = \mu$$

$$\mathsf{V}\,[Y] = \frac{V_P(\mu)}{n} = \frac{\mu}{n}$$

Sætning 6.4.1 Test for homogenitet

Kvotienttesteststørrelsen for hypotesen (6.1.5) er

$$G^2(H_I) = 2\sum_{i=1}^{k} n_i \left[y_i \ln(y_i/\overline{y}_+) - (y_i - \overline{y}_+)\right] \tag{6.4.4}$$

med $y_i = z_i/n_i$ $(= \overline{x}_{i+})$.

Kvotienttestst\o rrelsen svarer til de vægtede enhedsdevianser (med vægtene n_i)

$$d(y_i, \overline{y}_+) = 2n_i \left[y_i \ln(y_i/\overline{y}_+) - (y_i - \overline{y}_+) \right]$$

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

$$X^2 = \sum_{i=1}^{k} n_i \frac{(y_i - \overline{y}_+)^2}{\overline{y}_+} \qquad f = k - 1 \tag{6.4.5}$$

Testet er asymptotisk ækvivalent med kvotienttestet. Dette test måler direkte variationen imellem grupper, SAK_2 , i forhold til den estimerede varians, $V_P(\overline{y}_+) = \overline{y}_+$. (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

$$\frac{a-b}{a} \leq \ln(\frac{a}{b}) \leq \frac{a-b}{b}$$

med lighedstegn gyldigt for $a = b$. Approximerer man $\ln(a/b)$ med gennemsnittet af øvre og nedre grænse finder man

$$\ln\left(\frac{a}{b}\right) \approx \frac{a^2 - b^2}{2ab} \tag{6.4.6}$$

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 μ ved en $G(\alpha, 1/\beta)$-fordeling.

Lemma 6.4.1 Momenter i fordelingen af μ ved gamma-Poisson sampling

Såfremt $Z|\mu \in \mathrm{P}(n\mu)$ og $\mu \in \mathrm{G}(\alpha, 1/\beta)$, da gælder for fordelingen af μ:

$$m = \mathsf{E}[\mu] = \frac{\alpha}{\beta} \tag{6.4.7}$$

$$\mathsf{V}[\mu] = \frac{V_P(m)}{\beta} = \gamma\, V_P(m) \tag{6.4.8}$$

Den marginale middelværdi af variansfunktionen $V_P(\mu)$ er

$$\mathsf{E}[V_P(\mu)] = V_P(m) = m \tag{6.4.9}$$

hvor "signal/støj forholdet," γ, er givet ved

$$\gamma = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V_P(\mu)]} = \frac{1}{\beta} \tag{6.4.10}$$

med $0 < \gamma < \infty$.

Endelig er intraklassekorrelationen, ρ, givet ved

$$\rho = \frac{\mathsf{V}[\mu]}{\mathsf{E}[\mathsf{V}[\mu]]} = \frac{1}{1+\beta} \tag{6.4.11}$$

Bevis:

Beviset følger ved at bemærke, at der for $\mu \in G(\alpha, 1/\beta)$ gælder

$$\begin{aligned} E[\mu] &= \frac{\alpha}{\beta} \\ V[\mu] &= \frac{\alpha}{\beta^2} \end{aligned}$$

□

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

$$E[Z] = nm \tag{6.4.12}$$
$$V[Z] = nm(1+n\gamma) \tag{6.4.13}$$

med m og γ givet ved (6.4.7) og (6.4.10).

For den relative hyppighed $Y = Z/n$ gælder tilsvarende

$$E[Y] = m \tag{6.4.14}$$
$$V[Y] = m\left[\gamma + \frac{1}{n}\right] \tag{6.4.15}$$

Bevis:
Udtrykket for momenterne fås ved indsættelse af resultaterne fra lemma

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)

$$\mathsf{V}\,[Z] = n\,V_P(m)\,[1+n\gamma]$$

i komponenter svarende til den gennemsnitlige varians indenfor grupper og variansen mellem gruppemiddelværdier, giver størrelserne

$$\begin{aligned}\mathsf{E}\,[\mathsf{V}\,[Z|\mu]] &= n\,\mathsf{E}\,[V_P(\mu)] = n\,m\\ \mathsf{V}\,[\mathsf{E}\,[Z|\mu] &= n^2\,\mathsf{V}\,[\mu] = n^2\gamma\,V_P(m)\end{aligned}$$

□

Bemærkning 2 *Overdispersion i forhold til Poissonfordelingen*

For begrænsede værdier af n vil fordelingen af Z have en overdispersion i forhold til Poissonfordelingen $\mathrm{P}(n\,m)$ med samme antalsparameter n og med middelværdiparameteren m.

Overdispersionen er

$$\delta = \frac{\mathsf{V}\,[Z]}{nV_P(m)} = 1+n\gamma \tag{6.4.16}$$

Tilsvarende vil fordelingen af $Y = Z/n$ have overdispersionen $\delta = (1+n\gamma)$ i forhold til $\mathrm{P}(n\,m)/n$-fordelingen.

Den marginale varians af den gennemsnitlige rate Y kan altså udtrykkes som

$$\mathsf{V}\,[Y] = V_P(m)/n_{eff}\ ,$$

hvor den effektive stikprøvestørrelse, n_{eff} er bestemt ved

$$n_{eff} = \frac{n}{\delta} = \frac{n}{1+n\gamma}$$

svarende til

$$\frac{1}{n_{eff}} = \frac{1}{n} + \gamma \tag{6.4.17}$$

For $Y \in \mathrm{NB}(\alpha, \beta/(\beta+n))/n$ gælder altsaa

$$\mathsf{V}[Y] = \frac{m}{n_{eff}}$$

med m og γ bestemt ved (6.4.7) og(6.4.10).

Variansen for Y er således den samme som variansen for gennemsnittet af n_{eff} uafhængige målinger af en $\mathrm{P}(m)$ fordelt variabel. Såfremt man kender parameteren γ, kan eksempelvis approximative konfidensintervaller for m altså bestemmes som konfidensintervaller for en tilsvarende Poissonfordelt størrelse med stikprøvestørrelsen n_{eff}.

Parameteren γ udtrykker afvigelsen fra den rene Poissonfordeling af Z. For $m > 0$ og $\gamma \to 0$ vil $\mathsf{V}[\mu] \to 0$, og fordelingen af Z vil nærme sig en $\mathrm{P}(nm)$-fordeling.

□

Sætning 6.4.3 Momentestimation i den negative binomialfordeling

Lad $Z_1, Z_2, \ldots, Z_k$ være uafhængige variable, hvor $Z_i \in \mathrm{NB}(\alpha, \beta/(\beta+n_i))$

Momentestimaterne for $m = \alpha/\beta$ og $\gamma = 1/\beta$ er da

$$\tilde{m} = \overline{y}_+ \tag{6.4.18}$$

$$\tilde{\gamma} = \frac{s_2^2/\overline{y}_+ - 1}{n_0} \tag{6.4.19}$$

med

$$s_2^2 = \sum_{i=1}^{k} n_i(y_i - \overline{y}_+)^2/(k-1)\ ,$$

hvor $y_i = z_i/n_i$, og hvor den vægtede gennemsnitlige stikprøvestørrelse, n_0, er bestemt ved (5.1.9).

Momentestimatorerne for α og β bliver derfor

$$\tilde{\alpha} = n_0 \frac{\overline{y}_+^2}{s_2^2 - \overline{y}_+} \tag{6.4.20}$$

$$\tilde{\beta} = n_0 \frac{\overline{y}_+}{s_2^2 - \overline{y}_+} \tag{6.4.21}$$

Bevis:
Resultatet følger umiddelbart ved identifikation af momenterne.

□

Bemærkning 1 ***Singulariteter ved momentestimationen***

Hvis $s_2^2 < \overline{y}_+$ bliver $\tilde{\gamma}$, $\tilde{\alpha}$ og $\tilde{\beta}$ negative. I dette tilfælde vil det være naturligt at sætte disse størrelser til nul, d.v.s at fordelingen af μ estimeres til at være en etpunktfordeling, og fordelingen af Z_i bliver da en P($n_i m$)- fordeling. Parametriseringen ved (m, γ) tillader netop estimation af m, selv i tilfældet $\gamma = 0$.

□

Sætning 6.4.4 Maksimum-likelihood estimation i den negative binomialfordeling

Lad $Z_1, Z_2, \ldots, Z_k$ være uafhængige variable, hvor $Z_i \in \mathrm{NB}(\alpha, \beta/(\beta+n_i))$.

Maksimum-likelihood estimaterne $(\hat{m}, \hat{\gamma})$ for $m = \alpha/\beta$ og $\gamma = 1/\beta$ findes da ved at maksimere

$$l(m, \gamma; z_1, z_2, \ldots, z_k) = \sum_{i=1}^{k} \sum_{\nu=0}^{z_i - 1} \ln(m + \nu\gamma) - \sum_{i=1}^{k} \left(\frac{m}{\gamma} + z_i \right) \ln(1 + n_i\gamma) \tag{6.4.22}$$

med hensyn til m og γ.

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).

□

Bemærkning 1 *Bestemmelse af maksimum-likelihood estimaterne*

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 estimaterne ved at sætte scorefunktionen lig nul og løse ligningerne med hensyn til parametrene m og γ.

Ligningerne må løses iterativt. Som udgangsværdier for iterationen kan benyttes momentestimaterne $\tilde{m}$ og $\tilde{\gamma}$ bestemt ved (6.4.18) og (6.4.19).

Såfremt $s_2^2 < \overline{y}_+$, må man formode, at maksimum ligger på randen af området, svarende til $\gamma = 0$. I det balancerede tilfælde kan det vises, at $s_2^2 < \overline{y}_+ \Rightarrow \hat{\gamma} = 0$. (Se W.Simonsen (1976,1980) og J.Bonitzer (1978)).

For $\gamma = 0$ er maksimum-likelihood estimatet for m bestemt ved $\hat{m} = \overline{y}_+$.

□

Bemærkning 2 *Kvotienttest af hypotesen* $\gamma = 0$

Kvotienttestet for hypotesen

$$\mathrm{H}_{II} : \gamma = 0 \quad \text{med alternativet} \quad \overline{\mathrm{H}}_{II} : \gamma > 0$$

under den tilfældige model, har teststørrelsen

$$Z = 2 \left[\sum_{i=1}^{k} \sum_{\nu=0}^{z_i-1} \ln\left(\frac{\hat{m}+\nu\hat{\gamma}}{\overline{y}_+} \right) - \sum_{i=1}^{k} \left(\frac{\hat{m}}{\hat{\gamma}} + t_i \right) \ln(1+n_i\hat{\gamma}) \right] \quad (6.4.23)$$

hvor $\hat{m}$ og $\hat{\gamma}$ angiver maksimum-likelihood estimaterne 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).

Nedenstående tabel viser fordelingen af antal daglige tordenvejrsepisoder ved Cape Kennedy, Florida i månederne juni, juli og august for 10-års perioden 1957-1966, ialt 920 døgn.

Fordelingen af antal dage med 0,1,2 eller flere episoder af tordenvejr ved Cape Kennedy

Antal episoder z_i	Antal dage $\# i$	Poisson forventet	Negativ binomial forventet
0	803	791.85	802.92
1	100	118.78	100.15
2	14	8.91	14.38
3+	3	0.46	2.55

Alle observationsperioder er på $n_i = 1$ dag.

Man finder gennemsnittet, $\overline{y}_+ = 0.15$ tordenvejr/dag. Det fremgår klart af tabellen, at den observerede fordeling har en større spredning (tykkere haler), 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 $\overline{y}_+ = 0.15$ og $s_2^2 = 0.18$, finder man momentestimaterne $\tilde{m} = 0.15$ [tordenvejr/dag] og $\tilde{\gamma} = 0.2$. Disse estimater afviger ikke meget fra maksimum-likelihood estimaterne, der bestemmes ved en numerisk maksimeringsrutine 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 χ^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 jævnt ud over tidsaksen, men at der er en vis træghed i disse fænonemer, 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

Nedenstående tabel angiver antallet af fejl, z_i , ved airconditioneringsanlægget i løbet af 1000 flyvetimer for hver af 10 fly. (Data fra F.Proschan: Theoretical Explanation of Observed Decreasing Failure Rate. *Technometrics* **5**, 1963, pp 375-383. Se også L.J.Bain og F.T.Wright: The Negative Binomial Process with Applications to Reliability. *Journ. Qual.Techn.* **14**, 1982, pp. 60-66.)

Antal fejl ved airconditioneringsanlæg i 1000 flyvetimer for 10 fly.

	Fly nr									
	7908	7909	7910	7911	7912	7913	7914	7915	8044	8045
z_i	8	16	9	6	10	13	16	4	9	12

Man finder $\overline{y}_+ = 10.30$ [fejl/1000 timer] og $s_2^2 = s^2 = 15.79$. Benyttes Pearson-størrelssen til sammenligning af de 10 observerede fejlintensiteter under den systematiske model finder man

$$(n-1)s^2/\overline{y}_+ = 13.80 > \chi^2_{0.95}(9) ,$$

og man vælger derfor at modellere variationen med en tilfældig model for intensiteten.

Idet det antages, at $\mu \in \mathrm{G}(m/\gamma, \gamma)$ med $m = \alpha/\beta$ og $\gamma = 1/\beta$ finder man estimaterne

$$\tilde{m} = 10.30 \text{ [fejl/1000 timer]}$$

og

$$\tilde{\gamma} = 15.79/10.30 - 1 = 0.533 \text{ [1000 timer]}$$

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

$$\text{E}[X_{ij}|\mu] = \mu \,, \tag{6.5.1}$$

og variansfunktionen

$$V_G(\mu) = \mu^2 \tag{6.5.2}$$

Den kanoniske linkfunktion er

$$\eta(\mu) = \frac{1}{\mu} \tag{6.5.3}$$

For $Z \in \text{G}(n, \mu)$ gælder

$$\begin{aligned} \text{E}[Z] &= n\mu \\ \text{V}[Z] &= nV_G(\mu) = n\mu^2 \end{aligned}$$

For $Y = Z/n$ har vi derfor:

$$\begin{aligned} \mathsf{E}[Y] &= \mu \\ \mathsf{V}[Y] &= \frac{V_G(\mu)}{n} = \frac{\mu^2}{n} \end{aligned}$$

med μ og $V_G(\mu)$ givet ved (6.5.1) og (6.5.2)

Sætning 6.5.1 Test for homogenitet

Kvotienttesstørrelsen for hypotesen (6.1.5) er

$$G^2(H_I) = 2\sum_{i=1}^{k} n_i \left[\ln(\overline{y}_+/y_i) + \frac{y_i - \overline{y}_+}{\overline{y}_+}\right] \tag{6.5.4}$$

med $y_i = z_i/n_i$ $(= \overline{x}_{i+})$ og $\overline{y}_+ = \sum z_i / \sum n_i$ $(= \overline{\overline{x}}_{++})$.

Kvotienttesstørrelsen svarer til de vægtede enhedsdevianser (med vægtene n_i)

$$d(y_i, \overline{y}_+) = 2n_i \left[\ln(\overline{y}_+/y_i) + \frac{y_i - \overline{y}_+}{\overline{y}_+}\right]$$

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

$$X^2 = \sum_{i=1}^{k} n_i \frac{(y_i - \overline{y}_+)^2}{\overline{y}_+^2} \tag{6.5.5}$$

Testet er asymptotisk ækvivalent med kvotienttestet. Testet måler variansen imellem gruppegennemsnittene, SAK_2 , i forhold til den estimerede varians, $V_G(\overline{y}_+) = \overline{y}_+^2$ i eksponentialfordelingen.

□

Såfremt man ønsker at modellere eventuelle forskelle imellem grupper ved en tilfældig model, kan man vælge at beskrive fordelingen af μ ved en RGam(α, β)- fordeling svarende til at fordelingen af $1/\mu$ modelleres ved en G$(\alpha, 1/\beta)$-fordeling.

Den marginale fordeling af Z_i bliver da en RBet(α, n_i, β)-fordeling.

Lemma 6.5.1 Momenter i fordelingen af μ ved reciprok gamma-gamma sampling

Såfremt $Z|\mu \in \mathrm{G}(n, \mu)$ og $\mu \in \mathrm{RGam}(\alpha, \beta)$, da gælder for fordelingen af μ:

For $\alpha \leq 1$ har fordelingen ikke nogen middelværdi. Såfremt $1 < \alpha$, har fordelingen af μ middelværdien

$$m = \mathsf{E}[\mu] = \frac{\beta}{\alpha - 1} \tag{6.5.6}$$

For $\alpha \leq 2$ har fordelingen ingen varians. Såfremt $2 < \alpha$, har fordelingen af μ variansen

$$\mathsf{V}[\mu] = \frac{V_G(m)}{\alpha - 2} = \frac{\gamma}{1 - \gamma} V_G(m) \ , \tag{6.5.7}$$

hvor "signal/støj forholdet," γ, eksisterer såfremt $2 < \alpha$. Signal/støjforholdet γ er givet ved

$$\gamma = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V_G(\mu)]} = \frac{1}{\alpha - 1} \tag{6.5.8}$$

med $0 < \gamma < 1$.

Den marginale middelværdi af variansfunktionen $V_G(\mu)$ er

$$\mathsf{E}\left[V_G(\mu)\right] = \frac{\alpha - 1}{\alpha - 2} V_G(m) = \frac{1}{1-\gamma} V_G(m) \ , \tag{6.5.9}$$

Endelig er intraklassekorrelationen, ρ, givet ved

$$\rho = \frac{\mathsf{V}\left[\mu\right]}{\mathsf{E}\left[\mathsf{V}\left[\mu\right]\right]} = \frac{1}{\alpha} \tag{6.5.10}$$

for $2 < \alpha$.

Bevis:

Momenterne findes direkte ud fra momenterne i den reciproke gammafordeling

□

Sætning 6.5.2 Den marginale fordeling af gruppetotal ved Reciprok gamma-eksponentiel sampling

Såfremt $Z|\mu \in \mathrm{G}(n,\mu)$ og $\mu \in \mathrm{RGam}(\alpha,\beta)$, da er den marginale fordeling af Z en $\mathrm{RBet}(\alpha,n,\beta)$- fordeling

For $Z \in \mathrm{RBet}(\alpha,n,\beta)$ vil $T = \beta/(\beta+Z)$ følge en $\mathrm{Be}(\alpha,n)$-fordeling.

Såfremt $\alpha > 2$ gælder:

For $\alpha > 1$ har fordelingen en forventningsværdi

$$\mathsf{E}\left[Z\right] = nm \tag{6.5.11}$$

(6.5.12)

$$\mathsf{V}\left[Z\right] = n\,\frac{m^2}{1-\gamma}[1+n\gamma] \tag{6.5.13}$$

med m og γ givet ved (6.5.6) og (6.5.8).

Middelværdien $\mathsf{E}[Z]$ eksisterer dog blot $\alpha > 1$.

For den gennemsnitlige værdi $Y = Z/n$ gælder tilsvarende for $\alpha > 2$:

$$\mathsf{E}[Y] = m \tag{6.5.14}$$

$$\mathsf{V}[Y] = \frac{m^2}{1-\gamma}\left[\gamma + \frac{1}{n}\right] \tag{6.5.15}$$

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)

$$\mathsf{V}[Z] = n\,\frac{V_G(m)}{1-\gamma}\,[1 + n\gamma]$$

i komponenter svarende til den gennemsnitlige varians indenfor grupper og variansen mellem gruppemiddelværdier, giver størrelserne

$$\mathsf{E}[\mathsf{V}[Z|\mu]] = n\,\mathsf{E}[V_G(\mu)] = n\,\frac{V_G(m)}{1-\gamma}$$

$$\mathsf{V}[\mathsf{E}[Z|\mu] = n^2\,\mathsf{V}[\mu] = n^2\gamma\,\frac{V_G(m)}{1-\gamma}$$

□

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

$$\delta = \frac{\mathsf{V}[Z]}{nV_G(m)} = \frac{1+n\gamma}{1-\gamma} \tag{6.5.16}$$

Tilsvarende vil fordelingen af $Y = Z/n$ have overdispersionen $\delta = (1 + n\gamma)/(1-\gamma)$ i forhold til $\mathrm{G}(n,m)/n$-fordelingen.

Den marginale varians af gennemsnittet Y kan altså udtrykkes som

$$\mathsf{V}[Y] = V_G(m)/n_{eff} \ ,$$

hvor den effektive stikprøvestørrelse, n_{eff} er bestemt ved

$$n_{eff} = \frac{n}{\delta} = n\,\frac{1-\gamma}{1+n\gamma} \tag{6.5.17}$$

For $0 < m < \infty$ vil $\gamma \to 0$ være ensbetydende med $\mathsf{V}[1/\theta] \to 0$, og fordelingen af Z vil nærme sig en $\mathrm{G}(n,m)$-fordeling.

□

Sætning 6.5.3 Maksimum-likelihood estimation i den reciproke Betafordeling

Lad $T_1, T_2, \ldots, T_k$ være uafhængige variable, hvor $T_i \in \mathrm{RBet}(\alpha, n_i, \beta)$

Maksimum-likelihood estimaterne $\hat{m}$ og $\hat{\gamma}$) for $m = \beta/(\alpha-1)$ og $\gamma = 1/(\alpha-1)$ findes da ved at maksimere

$$l(m,\gamma;t_1,t_2,\ldots,t_k) = \sum_{i=1}^{k}\left[-n_i \ln\left(\frac{m}{1+\gamma}\right) - \left(n_i + \frac{1+\gamma}{\gamma}\right)\ln(1+t_i\gamma/m) + \sum_{\nu=1}^{n_i-1}\ln\left(1+\nu\gamma/(1+\gamma)\right)\right]$$

med hensyn til m og γ.

Bevis:
Sætningen bevises ved at notere, at - på nær en konstant - er logaritmen til likelihoodfunktionen givet ved (6.5.18).

□

Bemærkning 1 *Bestemmelse af maksimum-likelihood estimaterne*
Er endnu ikke udarbejdet

□

Sætning 6.5.4 *Kvotienttesstørrelsen 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 $Z_i \in \mathrm{RBet}(\alpha, n_i, \beta)$
Momentestimaterne for $m = \beta/(\alpha - 1)$ og $\gamma = 1/(\alpha - 1)$ er da

$$\tilde{m} = \overline{y}_+ \tag{6.5.18}$$

$$\tilde{\gamma} = \frac{s_2^2 - \overline{y}_+^2}{s_2^2 + n_0 \overline{y}_+^2} \tag{6.5.19}$$

med

$$s_2^2 = \sum_{i=1}^{k} n_i (y_i - \overline{y}_+)^2/(k-1) \,,$$

hvor $y_i = z_i/n_i$, og hvor den vægtede gennemsnitlige stikprøvestørrelse, n_0, er bestemt ved (5.1.9).

De tilsvarende estimater for α og β er

$$\tilde{\alpha} = 1 + \frac{s_2^2 + n_0 \overline{y}_+^2}{s_2^2 - \overline{y}_+^2}$$

$$\tilde{\beta} = \overline{y}_+ \frac{s_2^2 + n_0 \overline{y}_+^2}{s_2^2 - \overline{y}_+^2}$$

Bevis:
Følger af sætning 6.1.3.

□

Bemærkning 1 *Singularitet ved momentestimation*

Såfremt $s_2^2 < \overline{y}_{\cdot+}$ bliver $\tilde{\gamma}$ negativ. I dette tilfælde vil det være naturligt, at sætte γ til nul, d.v.s. fordelingen af μ estimeres til at være en etpunktfordeling, og fordelingen af Z bliver da en $\mathrm{G}(n, m)$-fordeling, hvor m estimeres som $\tilde{m} = \overline{y}_{+}$.

□

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 $F(x) = \mathrm{P}[X \leq x]$ 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 infinitesimale dødssandsynlighed til tiden x, givet komponenten er i live til tiden x. Forløbet af hændelsesraten beskriver ældningsforholdene for populationen.

Levetiden for elektroniske komponenter modelleres ofte ved eksponentialfordelingen, i det mindste inden for den operationelle horisont. Hændelsesraten 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 forskellige stresspåvirkninger. Dette kan modelleres ved at lade middellevetiden, θ, 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 $X|\theta \in \mathrm{Ex}(\theta)$, og $\theta \in \mathrm{RGam}(\alpha, \beta)$, da vil den marginale fordeling af X være en $\mathrm{RBet}(\alpha, 1, \beta)$-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 komponenten 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 θ, og man finder da, at jo længere komponenten lever, des mindre bliver den umiddelbare dødsintensitet.

□

Eksempel 6.5.2 Ventetid til udskiftning

Betragter 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 $X_1, X_2, ..., X_n$ 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 $\mathrm{Ex}(\theta)$ fordeling, og at middeltiden mellem fejl $\mathrm{E}[X|\theta] = \theta$ varierer imellem apparaterne i overensstemmelse med en $\mathrm{RGam}(\alpha, \beta)$-fordeling. Lad T_n angive ventetiden til den n'te udskiftning. Der gælder da

$$\mathrm{P}[T_n \leq t|\theta] = \mathrm{P}[\mathrm{G}(n, \theta) \leq t] \tag{6.5.20}$$

og

$$\mathrm{P}[T_n \leq t|\theta] = \mathrm{P}[\mathrm{RBet}(\alpha, n, \beta) \leq t] = \mathrm{P}\left[\mathrm{Be}(n, \alpha) \leq \frac{t}{t+\beta}\right] \tag{6.5.21}$$

Betragter vi i stedet antallet N_t af udskiftninger i en periode af længden t, finder vi

$$\mathrm{P}[N_t \leq x|\theta] = \mathrm{P}[\mathrm{P}(t/\theta) \leq x] \tag{6.5.22}$$

og ifølge Sætning 6.4.2 har vi da, idet $1/\theta \in \mathrm{G}(\alpha, 1/\beta)$, at

$$\mathrm{P}[N_t \leq x] = \mathrm{P}[\mathrm{NB}(\alpha, \beta/(\beta+t)) \leq x] \tag{6.5.23}$$

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)

$$\begin{aligned} \mathrm{P}[\mathrm{RBet}(\alpha, n, \beta) \leq t] &= \mathrm{P}\left[\mathrm{Be}(n, \alpha) \leq \frac{t}{t+\beta}\right] \\ &= 1 - \mathrm{P}[\mathrm{NB}(\alpha, \beta/(\beta+t)) \leq n-1] \end{aligned} \tag{6.5.24}$$

svarende til den velkendte relation

$$\mathsf{P}\,[\mathrm{G}(n,\theta) \leq t] = 1 - \mathsf{P}\,[\mathrm{P}(t/\theta) \leq n-1]$$

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.

Ventetiden til n_i'te fejl ved airconditioneringsanlæg for 10 fly.

	Fly nr									
	7908	7909	7910	7911	7912	7913	7914	7915	8044	8045
n_i	8	16	9	6	10	13	16	4	9	12
t_i	865	983	842	944	917	812	991	650	934	921

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 $\mathrm{Ex}(\mu)$-fordelte variable, hvor middeltiden mellem fejl, μ, varierer mellem fly i overensstemmelse med en $\mathrm{RGam}(\alpha,\beta)$-fordeling, svarende til at fejlintensiteten $\lambda = 1/\mu$ varierer i overensstemmelse med en $\mathrm{G}(\alpha,\beta)$-fordeling.

Idet $\overline{y}_+ = 86.01$ [timer], $s_2^2 = 9853$ og $n_0 = 10.146$, finder man momentestimaterne $\tilde{m} = 86.01$ [timer] og $\tilde{\gamma} = 0.0289$ svarende til $\tilde{\alpha} = 35.58$ og $\tilde{\beta} = 2974$ [timer]. Sammenlignes med estimaterne $\tilde{\alpha}^* = 19.32$ og $\tilde{\beta}^* = 1876$, der blev fundet fra de samme data i eksempel 6.4.2 synes forskellen umiddelbart ganske stor, men betragter man f.eks. skønnet over $\mathsf{E}\,[\mu] = \alpha/\beta$ og $\mathsf{V}\,[\mu] = \alpha/\beta^2$ finder man under ventetidsmodellen $\tilde{\alpha}/\tilde{\beta} = 0.012$ og $\tilde{\alpha}/(\tilde{\beta})^2 = (0.002)^2$, mens estimation under Poisson-Gamma modellen fører til $\tilde{\alpha}^*/\tilde{\beta}^* = 0.010$ og $\tilde{\alpha}^*/(\tilde{\beta}^*)^2 = (0.002)^2$, hvilket viser at den reelle forskel ikke er så stor.

□

6.6 Fordeling af empiriske varianser for normaltfordelte 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

$$SAK_i = \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i+})^2$$

med

$$\overline{x}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

Antag, at $X_{ij}|(\mu_i, \sigma_i^2) \in \mathrm{N}(\mu_i, \sigma_i^2)$, og X_{ij}, $j = 1, \ldots, n_i$ er betinget uafhængige givet sættet af (μ_i, σ_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 μ_i.

Betragter vi specielt de empiriske varianser

$$S_i^2 = \frac{SAK_i}{f_i} \tag{6.6.1}$$

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:

$$S_i^2|\sigma_i^2 \in \mathrm{G}\big(f_i/2, \ \sigma_i^2/(f_i/2)\big)$$

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 normaltfordelte observationer, vil vi alligevel her give en beskrivelse af fordelingsforholdene i dette specialtilfælde.

Homogenitetshypotesen

$$H_I : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

svarende til den systematiske model (Bartletts test) blev behandlet i eksempel 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

$$\sigma_i^2 \in \mathrm{RGam}(\alpha, \beta) \,, \tag{6.6.2}$$

hvilket er det samme som

$$\frac{1}{\sigma_i^2} \in \mathrm{G}(\alpha, 1/\beta) \,, \tag{6.6.3}$$

Det følger af betragtningerne i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998, at

$$\mathsf{E}\,[\sigma^2] = \frac{\beta}{\alpha - 1}$$

og

$$\mathsf{V}\,[\sigma^2] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} = \frac{\mathsf{E}\,[\sigma^2]^2}{\alpha - 2}$$

Dvs den relative spredning i fordelingen af σ^2 er

$$\sqrt{\mathsf{V}\,[\sigma^2]}/\mathsf{E}\,[\sigma^2] = 1/\sqrt{\alpha - 2} \tag{6.6.4}$$

6.6.3 Fortolkning af parametre i strukturfordelingen af σ^2

I stedet for parametriseringen af fordelingen af σ^2 ved α og β vil vi indføre en parametrisering, der er relateret til χ^2-fordelingen.

Der gælder

Sætning 6.6.1 ***Fortolkning af strukturfordelingen af*** $1/\sigma^2$ ***som en*** χ^2***-fordeling***

Antag, at strukturfordelingen af σ^2 er som i (6.6.3), dvs $\sigma^2 \in \mathrm{RGam}(\alpha, \beta)$. Da gælder

$$\frac{1}{\sigma^2} \in \frac{1}{\sigma_0^2\,(\nu-2)}\,\chi^2(\nu)\,, \tag{6.6.5}$$

med $\sigma_0^2 = \mathsf{E}\,[\sigma^2]$ og $\nu = 2\alpha$.

Bevis:
Indfører vi $\nu = 2\alpha$ i (6.6.2), har vi

$$\frac{1}{\sigma^2} \in \mathrm{G}(\nu/2, 1/\beta)\,, \tag{6.6.6}$$

dvs formparameteren er $\nu/2$ og skalaparameteren er $1/\beta$. Der gælder

$$\mathsf{E}\,[\sigma^2] = \frac{\beta}{\nu/2-1}\,, \tag{6.6.7}$$

hvorfor vi kan udtrykke parameteren β som

$$\beta = (\nu-2)\;\sigma_0^2\;/2\,, \tag{6.6.8}$$

hvor vi har sat

$$\sigma_0^2 = \mathsf{E}\,[\sigma_i^2] \tag{6.6.9}$$

Vi kan udtrykke gammafordelingen (6.6.2) som en χ^2-fordeling ved $\chi^2(f) \equiv \mathrm{G}(f/2, 2)$, dvs

$$\frac{1}{\sigma^2} \in \frac{1}{2\beta}\,\chi^2(\nu)\,, \tag{6.6.10}$$

eller, idet vi udtrykker skalaparameteren β ved forventningsværdien, σ_0^2 af σ^2 (jvf (6.6.8)), har vi at

$$\frac{1}{\sigma^2} \in \mathrm{G}(\nu/2,\ 2/[(\nu-2)\sigma_0^2])\ , \tag{6.6.11}$$

hvilket er det samme som (6.6.5). □

Bemærkning 1 *Parameteren ν betegnes undertiden "frihedsgraderne" i fordelingen af $1/\sigma^2$*

På grund af relationen (6.6.5) betegner man undertiden parameteren ν som "frihedsgraderne i fordelingen af $1/\sigma^2$ ". Vi bemærker dog, at parameteren ν ikke behøver være heltallig.

Det følger af (6.6.4), og af relationen $\nu = 2\alpha$, at parameteren ν er bestemt ved den relative spredning i fordelingen af σ^2. □

Figur 6.3 viser et eksempel på fordelingen af σ^2.

6.6.4 Marginal fordeling af stikprøvevariansen

Sætning 6.6.2 Den marginale fordeling af S^2 ved reciprok gamma strukturfordeling

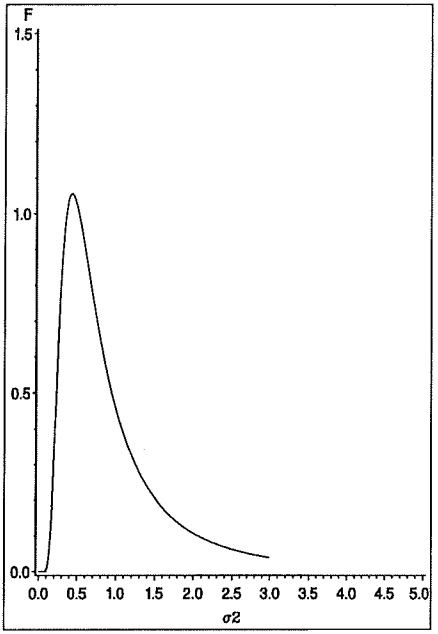
Såfremt $S^2|\sigma^2 \in \sigma^2\chi^2(f)/f$ og $1/\sigma^2 \in \frac{1}{\sigma_0^2(\nu-2)}\ \chi^2(\nu)$, da er den marginale fordeling af S^2 givet ved

$$S^2 \in \mathrm{RBet}\Big(\nu/2,\ f/2,\ \frac{\nu-2}{f}\ \sigma_0^2\Big) \tag{6.6.12}$$

Såfremt $\nu \leq 2$ har fordelingen af S^2 ingen middelværdi. For $2 < \nu$ har fordelingen af S^2 middelværdien

$$\mathrm{E}\,[S^2] = \mathrm{E}\,[\sigma^2] = \sigma_0^2 \tag{6.6.13}$$

Figur 6.3. Strukturfordeling af sand varians, σ^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

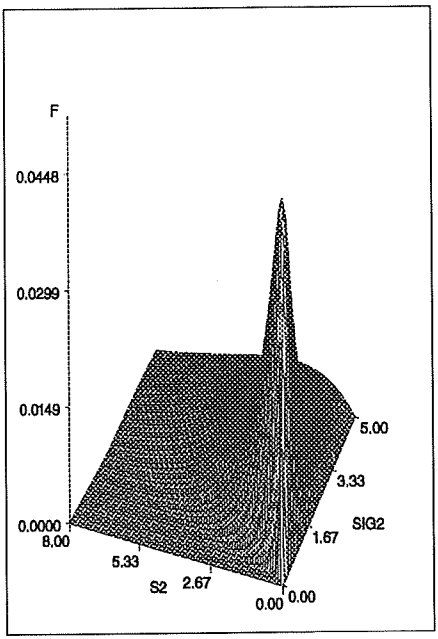
$$\mathrm{V}\left[S^2\right] = \frac{(\sigma_0^2)^2}{\nu/2 - 2}\left[1 + \frac{2(\nu/2 - 1)}{f}\right] \tag{6.6.14}$$

Bevis:
Fås af (6.5.6) og (6.5.7)

□

Den simultane fordeling af variansen, σ^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 σ^2
(Strukturfordeling af σ^2 som i figur 6.3.)

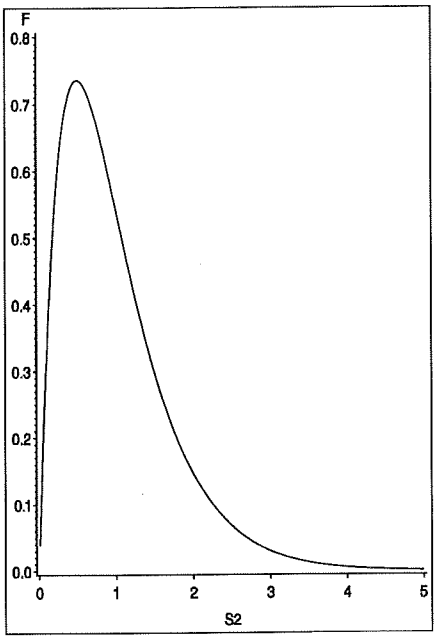


Figur 6.5 viser den betingede fordeling af den empiriske varians, S^2, svarende til en givet værdi af σ^2, ($\sigma^2 = 1$), og figur 6.6 viser den marginale fordeling af den empiriske varians, s^2, svarende til fordelingen i figur 6.4. Det ses, at den marginale fordeling af S^2 har tykkere haler, end 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, $\sigma^2 = 1$.



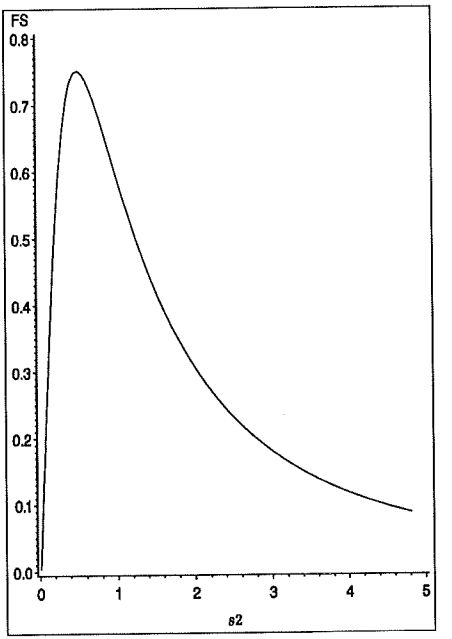
man, at der gælder

$$\frac{1}{S^2} \in \frac{\nu}{(\nu-2)\sigma_0^2}\,\mathrm{F}(\nu, f)\,, \tag{6.6.15}$$

hvor $\mathrm{F}(\nu, f)$ angiver en stokastisk variabel, der følger en F-fordeling med frihedsgraderne (ν, 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 σ^2 som i figur 6.3.)



Antag, at situationen er som beskrevet i sætning 6.6.2, hvor strukturfordelingen af σ^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 $f = n - 1 = 7$.

Vi ønsker således at bestemme sandsynligheden

$$\mathsf{P}\left[S^2 > 16.67\right] = \mathsf{P}\left[\frac{1}{S^2} \leq \frac{1}{16.67}\right] = \mathsf{P}\left[\frac{\nu}{(\nu-2)\sigma_0^2}\,\mathrm{F}(\nu, f) \leq \frac{1}{16.67}\right],$$

dvs

$$\mathsf{P}\left[\mathrm{F}(\nu, f) \leq \frac{(\nu-2)\sigma_0^2}{16.67\nu}\right], = \mathsf{P}\left[\mathrm{F}(4.5, 7) \leq \frac{2.5 \times 9}{4.5 \times 16.67}\right]$$

eller

$$\mathsf{P}\left[\mathrm{F}(4.5, 7) \leq 0.30\right],$$

Da tabellen over fraktiler i F-fordelingen kun angiver fraktiler svarende til sandsynligheder, der er større end 50 %, benytter vi relationen

$$\mathsf{P}\left[\mathrm{F}(f_1, f_2) \leq x\right] = 1 - \mathsf{P}\left[\mathrm{F}(f_2, f_1) \leq 1/x\right]$$

dvs den søgte sandsynlighed fås som

$$\mathsf{P}\left[\mathrm{F}(4.5, 7) \leq 0.30\right] = 1 - \mathsf{P}\left[\mathrm{F}(7, 4.5) \leq 3.33\right]$$

Ved opslag i tabellen over fraktiler i F-fordelingen finder vi, at $\mathsf{P}\left[\mathrm{F}(7,4) \leq 3.98\right] = 0.90$ og $\mathsf{P}\left[\mathrm{F}(7,5) \leq 3.37\right] = 0.90$. Den søgte sandsynlighed, $\mathsf{P}\left[\mathrm{F}(4.5,7) \leq 0.30\right]$ 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 estimaterne må bestemmes ved iteration.

Momentestimaterne 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

$$S_i^2 \in \mathrm{RBet}\Big(\nu/2,\ f_i/2,\ \frac{\nu-2}{f_i}\,\sigma_0^2\Big)$$

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 σ_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 σ_0^2 og ν bestemt ved

$$\begin{aligned} \widetilde{\sigma}_0^2 &= \overline{s^2} \\ \widetilde{\nu} &= 2\Big[1+\Big\{Q_1+\frac{1}{k-1}\Big[\sum_i f_i - \frac{\sum_i f_i^2}{\sum_i f_i}\Big]\Big\}/(Q_1-2)\Big], \end{aligned} \tag{6.6.16}$$

hvor

$$Q_1 = \frac{SAK_s/(k-1)}{\left(\overline{s^2}\right)^2},$$

med

$$SAK_s = \sum_i (n_i-1)(S_i^2-\overline{S^2})^2 \tag{6.6.17}$$

og

$$\overline{s^2} = \Big(\sum_{i=1}^k f_i s_i^2\Big)\Big/\Big(\sum_{i=1}^k f_i\Big) \tag{6.6.18}$$

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

$$\tilde{\nu} = 2\big(2+(f+2)/(Q_1-2)\big)$$

□

6.7 Den flerdimensionale normalfordeling

6.7.1 Den systematiske model

Vi vil betragte den sædvanlige flerdimensionale variansanalysemodel. Lad observationerne X_{ij} være p-dimensionale vektorer, hvor

$$X_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{ij},\ i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i \tag{6.7.1}$$

med

$$\sum_{i=1}^{k} n_i \boldsymbol{\alpha}_i = 0$$

hvor $\boldsymbol{\mu}, \boldsymbol{\alpha}_i$ og $\boldsymbol{\epsilon}_{ij}$ angiver p-dimensionale vektorer med $\boldsymbol{\epsilon}_{ij}$ indbyrdes uafhængige, $\boldsymbol{\epsilon}_{ij} \in \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, og hvor $\boldsymbol{\Sigma}$ angiver den fælles $p \times p$-dimensionale kovariansmatrix. For simpelheds skyld antager vi at $\boldsymbol{\Sigma}$ har fuld rang.

Under disse antagelser finder vi at $Z_i \sum_j X_{ij} \in \mathrm{N}_p(n_i(\boldsymbol{\mu} + \boldsymbol{\alpha}_i), n_i \boldsymbol{\Sigma})$.

I dette tilfælde beskrives variationen ved $p \times p$-dimensionale SAK-matricer. Vi vil derfor i lighed med tidligere indføre betegnelserne

$$\overline{X}_{i+} = \sum_{j=1}^{n_i} X_{ij}/n_i \tag{6.7.2}$$

$$\overline{X}_{i+} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}/N = \sum_{i=1}^{k} n_i \overline{X}_{i+} / \sum_{i=1}^{k} n_i \tag{6.7.3}$$

til beskrivelse af gruppegennemsnittene og det fælles gennemsnit, ligesom vi indfører

$$\mathbf{SAK}_1 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i+})(X_{ij} - \overline{X}_{i+})^T \tag{6.7.4}$$

$$\mathbf{SAK}_2 = \sum_{i=1}^{k} n_i (\overline{X}_{i+} - \overline{X}_{++})(\overline{X}_{i+} - \overline{X}_{++})^T \tag{6.7.5}$$

$$\mathbf{SAK}_0 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{++})(X_{ij} - \overline{X}_{++})^T \tag{6.7.6}$$

til beskrivelse af henholdsvis variationen indenfor grupper ($\mathbf{SAK}_1$), imellem grupper ($\mathbf{SAK}_2$) og den totale variation ($\mathbf{SAK}_0$).

Vi bemærker, at også her gælder den sædvanlige pythagoræiske relation

$$\mathbf{SAK}_0 = \mathbf{SAK}_1 + \mathbf{SAK}_2 \tag{6.7.7}$$

Sætning 6.7.1 Test for homogenitet

Kvotienttesteststørrelsen for hypotesen

$$\mathrm{H}_I : \alpha_1 = \alpha_2 = \cdots = \alpha_k = \mathbf{0} \tag{6.7.8}$$

mod alternativet

$$\overline{\mathrm{H}}_I : \alpha_i \neq 0 \text{ for mindst ét } i$$

har teststørrelsen

$$Z = \det(\mathbf{SAK}_1)/\det(\mathbf{SAK}_0)$$

hvor $\det(\mathbf{A})$ angiver determinanten af matricen $\mathbf{A}$.

Under hypotesen H_I er Z fordelt som Wilk's Λ med parametrene $(p, k-1, N-k)$

Bevis:
Se f.eks. Rao p. 556.

□

Bemærkning 1 ***Approximativ fordeling af teststørrelsen***

Wilk's Λ 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 $\mathrm{F}(p(k-1), ms - 2\lambda)$-fordeling.

□

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

$$\chi'^2 = \sum_{i=1}^{k} n_i(\overline{X}_{i+} - \overline{X}_{++})^T \mathbf{S}_1^{-1}(\overline{X}_{i+} - \overline{X}_{++}) \tag{6.7.9}$$

hvor

$$\mathbf{S}_1 = \mathbf{SAK}_1/(N-k)$$

angiver det sædvanlige skøn over kovariansmatricen $\mathbf{\Sigma}$.

Den generaliserede Mahalanobis afstand χ'^2 er en skalar. Under hypotesen H_I er χ'^2 approximativt $\chi^2(p(k-1))$-fordelt. Testet forkaster for store værdier af χ'^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 α_i med en $N_p(\mathbf{0}, \mathbf{\Sigma}_0)$-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 α_i er uafhængige, $\alpha_i \in N_p(\mathbf{0}, \mathbf{\Sigma}_0), i = 1, 2, \ldots, k$ og hvor ϵ_{ij} er indbyrdes uafhængige, $\epsilon_{ij} \in N_p(\mathbf{0}, \mathbf{\Sigma})$,

da er den marginale fordeling af $Z_i = \sum_j X_{ij}$ en

$$\mathrm{N}_p(n_i\boldsymbol{\mu}, n_i\boldsymbol{\Sigma} + n_i^2\boldsymbol{\Sigma}_0)\text{-fordeling}$$

og

den marginale fordeling af $\overline{X}_{i+}$ er en

$$\mathrm{N}_p(\boldsymbol{\mu}, \frac{1}{n_i}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_0)\text{-fordeling}$$

Endvidere gælder, at

$$\mathbf{SAK}_1 \in \mathrm{Wis}_p(N-k, \boldsymbol{\Sigma})$$

og $\mathbf{SAK}_1$ 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.

$$\mathsf{D}\,[Z] = \mathsf{D}\,[\mathsf{E}\,[Z|\boldsymbol{\mu}]] + \mathsf{E}\,[\,\mathsf{D}\,[Z|\boldsymbol{\mu}]] \tag{6.7.10}$$

□

Bemærkning 1 *Det generaliserede signal/støj-forhold*

Indfører vi den $p \times p$-dimensionale matrix $\boldsymbol{\Gamma}$ for forholdet mellem variationen mellem grupper og variansen inden for grupper ("signal/støj-forholdet"), $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}$, finder vi :

$$\mathsf{D}\,[\overline{X}_{i+}] = \left(\frac{1}{n_i}\mathbf{I} + \boldsymbol{\Gamma}\right)\boldsymbol{\Sigma}$$

□

Sætning 6.7.3 Momentestimation i den flerdimensionale normalfordelingsmodel

Under antagelserne fra sætning 6.7.2 findes momentestimatorerne for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ og $\boldsymbol{\Sigma}_0$ som

$$\begin{aligned} \tilde{\boldsymbol{\mu}} &= \overline{x}_{++} \\ \tilde{\boldsymbol{\Sigma}} &= \frac{1}{N-k}\,\mathbf{sak}_1 \\ \tilde{\boldsymbol{\Sigma}}_0 &= \frac{1}{n_0}\left(\frac{\mathbf{sak}_2}{k-1} - \tilde{\boldsymbol{\Sigma}}\right) \end{aligned} \tag{6.7.11}$$

Bevis:
Sætningen bevises ved at bemærke, at

$$\mathsf{E}\,[\overline{X}_{++}] = \boldsymbol{\mu}, \quad \mathsf{E}\,[\mathbf{SAK}_1] = (N-k)\boldsymbol{\Sigma}, \quad \text{og} \quad \mathsf{E}\,[\mathbf{SAK}_2] = (k-1)(\boldsymbol{\Sigma}+n_0\boldsymbol{\Sigma}_0)$$

hvor n_0 er givet ved (5.1.9).

□

Bemærkning 1 *Momentestimatet er ikke nødvendigvis ikke-negativ definit*

Skønnet $\tilde{\boldsymbol{\Sigma}}_0$ er centralt for $\boldsymbol{\Sigma}_0$, men skønnet er ikke nødvendigvis en ikke-negativ definit matrix. Såfremt $\tilde{\boldsymbol{\Sigma}}_0$ ikke har fuld rang, indikerer det, at fordelingen af $\boldsymbol{\alpha}$ kan være udartet. □

Sætning 6.7.4 Maksimum-likelihood estimation under den flerdimensionale normalfordelingsmodel

Under antagelserne fra sætning 6.7.2 findes maksimum-likelihood estimaterne for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ og $\boldsymbol{\Sigma}_0$ ved at maksimere

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_0; \overline{x}_{1+}, \ldots, \overline{x}_{k+}) = -\frac{N-k}{2}\ln(\det(\boldsymbol{\Sigma})) - \frac{1}{2}\mathrm{tr}(\mathbf{sak}_1\boldsymbol{\Sigma}^{-1})$$

$$-\sum_{i=1}^{k}\left[\ln\left(\det\left(\frac{\boldsymbol{\Sigma}}{n_i}+\boldsymbol{\Sigma}_0\right)\right) + \frac{1}{2}(\overline{x}_{i+}-\boldsymbol{\mu})^T\left(\frac{\boldsymbol{\Sigma}}{n_i}+\boldsymbol{\Sigma}_0\right)^{-1}(\overline{x}_{i+}-\boldsymbol{\mu})\right] \tag{6.7.12}$$

med hensyn til $\boldsymbol{\mu} \in \mathbb{R}^p$ og $\boldsymbol{\Sigma}$ og $\boldsymbol{\Sigma}_0$ i mængden af ikke-negativ definite symmetriske $p \times p$- matricer.

Bevis:
Beviset følger ved at bemærke, at $\mathbf{SAK}_1$ følger en $\mathrm{Wis}_p(N-k, \boldsymbol{\Sigma})$-fordeling

og at $\mathbf{SAK}_1$ er uafhængig af $\overline{X}_{i+}$, $i = 1, 2, \ldots, k$, samt at $\overline{X}_{i+} \in \mathrm{N}_p(\mu, \boldsymbol{\Sigma}/n_i + \boldsymbol{\Sigma}_0)$ er indbyrdes uafhængige, $i = 1, 2, \ldots, k$.

□

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 parametrisere varians-kovariansmatricerne ved de enkelte elementer. Afgrænsningen af søgeområdet foretages da ved at teste på de resulterende matricer, og indlægge en passende straffunktion, såfremt determinanten er negativ.

□

Eksempel 6.7.1 Variation mellem målefejl for flowmålere

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 [m^3 /h] og 0.5 [m^3 /h].

	Gentagelse					
	1		2		3	
	flow		flow		flow	
Maaler	0.1	0.5	0.1	0.5	0.1	0.5
41	-2.0	1.0	2.0	3.0	2.0	2.0
42	5.0	3.0	1.0	1.0	2.0	2.0
43	2.0	1.0	-3.0	-1.0	1.0	0.0
44	4.0	4.0	-1.0	2.0	3.0	5.0
45	4.0	2.0	0.0	1.0	-1.0	0.0
46	5.0	9.0	4.0	8.0	6.0	10.0

Med henblik på beskrivelse af variationen af fejlvisningen i målerpopulationen opstiller man følgende model:

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \; i = 41, 42, ..., 46; \quad j = 1, 2, 3.$$

hvor α_i er uafhængige, $\alpha_i \in \mathrm{N}_2(0, \boldsymbol{\Sigma}_0)$, og hvor ϵ_{ij} er indbyrdes uafhængige, $\epsilon \in \mathrm{N}_2(0, \boldsymbol{\Sigma})$.

Man finder

$$\overline{x}_{41+} = \begin{pmatrix} -1.00 \\ 1.67 \end{pmatrix}; \quad \overline{x}_{42+} = \begin{pmatrix} 2.67 \\ 2.00 \end{pmatrix}; \quad \overline{x}_{43+} = \begin{pmatrix} 0.00 \\ 0.00 \end{pmatrix};$$

$$\overline{x}_{44+} = \begin{pmatrix} 2.00 \\ 3.67 \end{pmatrix}; \quad \overline{x}_{45+} = \begin{pmatrix} 1.00 \\ 1.00 \end{pmatrix}; \quad \overline{x}_{46+} = \begin{pmatrix} 5.00 \\ 9.00 \end{pmatrix};$$

$$\mathbf{sak}_1 = \begin{pmatrix} 66.67 & 29.00 \\ 29.00 & 15.33 \end{pmatrix}; \quad \mathbf{sak}_0 = \begin{pmatrix} 134.28 & 116.22 \\ 116.22 & 171.78 \end{pmatrix}$$

og

$$\mathbf{sak}_2 = \begin{pmatrix} 67.61 & 87.22 \\ 87.22 & 156.44 \end{pmatrix}$$

Under den systematiske model finder man kvotienttesttørrelsen, Wilks Λ for hypotesen $\alpha_{41} = \cdots = \alpha_{46} = 0$ imod alternativet, at mindst to α-værdier er forskellige er $\Lambda = 0.018$.

Gentagelsesvariatonen udspænder således kun omkring 2 % af variationen i hele materialet.

Ved indsættelse i (6.7.11) finder man momentestimaterne

$$\widetilde{\boldsymbol{\mu}} = \begin{pmatrix} 1.61 \\ 2.89 \end{pmatrix}$$

kalibreringsusikkerheden

$$\widetilde{\boldsymbol{\Sigma}} = \mathbf{sak}_1/12 = \begin{pmatrix} 5.56 & 2.42 \\ 2.42 & 1.28 \end{pmatrix}$$

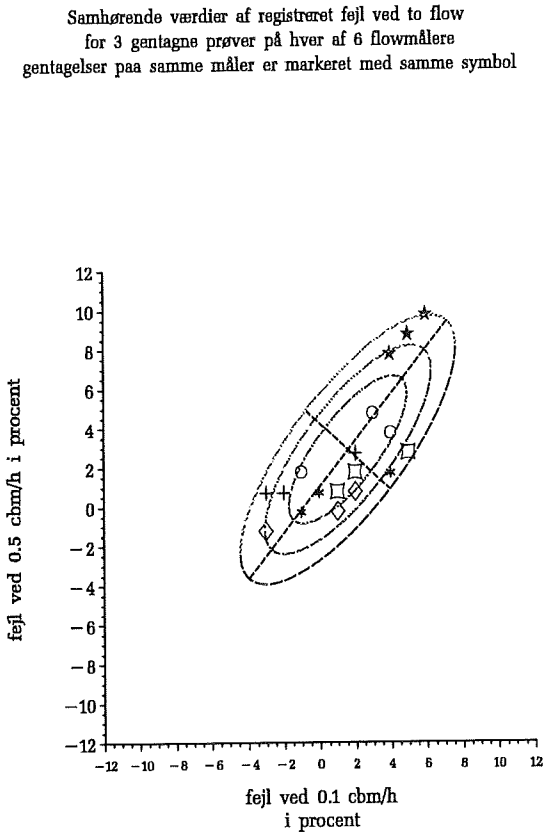
og dispersionsmatricen for målerpopulationen

$$\widetilde{\boldsymbol{\Sigma}}_0 = \frac{1}{3}[\mathbf{sak}_2/5 - \mathbf{sak}_1/12] = \begin{pmatrix} 2.65 & 5.01 \\ 5.01 & 10.00 \end{pmatrix}$$

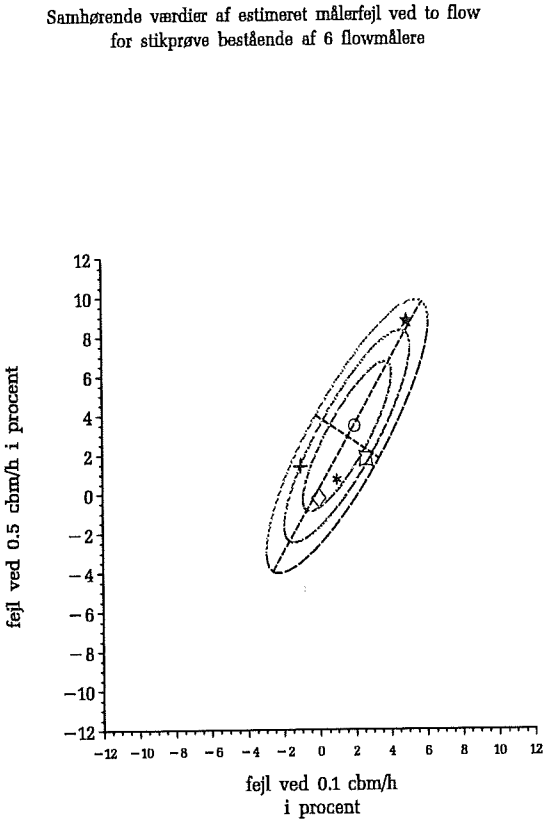
Hosstående figurer illustrerer dekomponeringen af variationen.

Vi bemærker af figur 6.8 at målerens "fejlniveau", $\alpha_{i1} + \alpha_{i2}$ varierer i populationen stort set uafhængigt af målerens "fejldifferens", $\alpha_{i1} - \alpha_{i2}$.

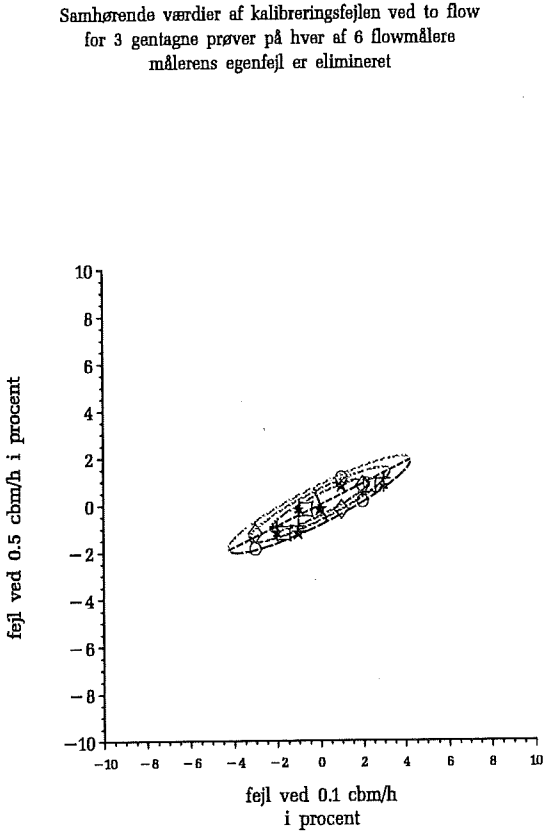
Man kan forestille sig, at man for hver måler bestemmer målerens kalibreringskurve, nemlig visningsfejlen som funktion af flow'et. I området mellem de to betragtede flow vil kalibreringskurven erfaringsmæssigt kunne tilnærmes med en ret linie. Man har da, at "fejlniveauet" vil være udtryk for kalibreringskurvens 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.



Figur 6.7. Illustration af den totale variation af fejlvisningen ved to flow for en række målere



Figur 6.8. Illustration af variationen mellem grupper (målerfejlen)



Figur 6.9. Illustration af variationen indenfor grupper (gentagelsesvariationen)

En væsentlig del af beregningerne i eksemplet kunne udføres i SAS®-programsystemet ved proceduren GLM under benyttelse af ordren MANOVA, der bevirker 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 PRINTE HTYPE=1 ETYPE=1;
RANDOM lbnr;
RUN;
```

definerer i ordren `CLASS lbnr ;`, at den variable `lbnr` som klassifikationsvariabel. 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 PRINTE ;
```

angiver nøgleordet `MANOVA`, at man ønsker at opfatte de to variable `fejl1` og `fejl2` på venstre side i modelformlen som en todimensional observation.

Ordren `H=lbnr` angiver, at man ønsker at teste effekten svarende til `lbnr`. Ordren er efterfulgt af en række options, nemlig `PRINTH`, der bevirker, at **SAK**-matricen svarende til `lbnr` effekten udskrives (dvs. $\mathbf{sak}_2$ matricen), og optionen `PRINTE`, der bevirker, at **SAK**-matricen svarende til residualerne udskrives (dvs. $\mathbf{sak}_1$-matricen).

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
Class Level Information

Class	Levels	Values
LBNR	6	241 242 243 244 245 246

Number of observations in data set = 18

General Linear Models Procedure

Dependent Variable: FEJL1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	67.6111	13.5222	2.43	0.0960
Error	12	66.6667	5.5556		
Corrected Total	17	134.2778			

R-Square	C.V.	Root MSE	FEJL1Mean
0.503517	146.2980	2.3570	1.6111

General Linear Models Procedure

Dependent Variable: FEJL2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	156.4444	31.2889	24.49	0.0001
Error	12	15.3333	1.2778		
Corrected Total	17	171.7778			

R-Square	C.V.	Root MSE	FEJL2 Mean
0.910737	39.12883	1.1304	2.8889

De endimensionale analyser viser, at der er signifikant forskel på fejlvisningen `fejl2` (ved flow 0.5 [m^3 /h]), $F(5,12) = 24.49$, mens man ikke kan påvise forskel på fejlvisningen `fejl1` ved flow 0.1 [m^3 /h] $F(5,12) = 2.43$.

$\mathbf{sak}_1$ matricen udskrives under overskriften `E = Error SS&CP Matrix`

```
E = Error SS&CP Matrix

                FEJL1          FEJL2

FEJL1         66.6667             29
FEJL2              29        15.3333
```

og endvidere udskrives korrelationsmatricen svarende til $\mathbf{sak}_1$

```
                General Linear Models Procedure
               Multivariate Analysis of Variance

Partial Correlation Coefficients from the Error SS&CP Matrix/Prob > |r|

     DF = 12        FEJL1     FEJL2

     FEJL1       1.000000  0.907038
                    0.0001    0.0001

     FEJL2       0.907038  1.000000
                    0.0001    0.0001
```

Det ses, at der er en stærk positiv korrelation $\hat{\rho} = 0.91$ mellem kalibreringsfejlen ved de to flow.

$\mathbf{sak}_2$-matricen udskrives under overskriften `H = Type I SS&CP Matrix for LBNR` som:

```
          General Linear Models Procedure
         Multivariate Analysis of Variance

          H = Type I SS&CP Matrix for LBNR

                        FEJL1              FEJL2
```

```
FEJL1      67.611111111      87.222222222
FEJL2      87.222222222      156.44444444
```

og endelig udskrives en række teststørrelser for test af hypotesen om forsvindende effekt af `lbnr` i den systematiske model. Vi bemærker, at kvotienttesstørrelsen 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 = Type I SS&CP Matrix for LBNR   E = Error SS&CP Matrix

                        S=2    M=1    N=4.5

Statistic                  Value        F  Num DF  Den DF  Pr > F

Wilks' Lambda          0.01895960  13.7775      10      22  0.0001
Pillai's Trace         1.29172915   4.3771      10      24  0.0015
Hotelling-Lawley Trace 35.35683630 35.3568      10      20  0.0001
Roy's Greatest Root    34.88712493 83.7291       5      12  0.0001

     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

$$\mathrm{E}\left[\mathrm{sak}_2/(k-1)\right] = \boldsymbol{\Sigma} + n_0\boldsymbol{\Sigma}_0$$

jvf. sætning 6.7.3:

```
Source      Type I Expected Mean Square

LBNR        Var(Error) + 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 ($\mathbf{S}_1 = \mathbf{sak}_1/(N-k)$), 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.

Gruppe	i	n_i	x_{1i}	x_{2i}	x_{3i}	x_{4i}
Basti	1	86	191.92	139.88	133.36	51.24
Brahmin	2	92	191.35	139.50	132.68	50.40
Chattri	3	139	191.92	139.88	133.36	51.24
Muslim	4	167	190.78	137.40	131.52	51.38
Bhatu	5	148	186.10	138.58	133.55	52.06
Habru	6	124	186.94	137.40	131.16	50.30
Bhil	7	187	181.87	137.62	131.18	48.60
Dom	8	113	186.40	137.52	132.64	50.34
Ahir	9	68	187.45	138.12	131.70	48.98
Kurmi	10	94	188.86	137.86	131.82	49.22
Artisan	11	173	187.69	136.84	131.30	48.72
Kahar	12	57	188.83	136.28	130.70	48.62
Ialt	$N = 1448$					
Vejet gennemsnit, $\overline{x}_{++}$			188.03	137.25	131.91	50.30

$$\mathbf{S}_1 = \begin{pmatrix} 43.6500 & 5.8865 & 8.4396 & 4.0610 \\ & 20.2500 & 11.1438 & 2.7326 \\ & & 20.9764 & 2.9688 \\ & & & 12.2500 \end{pmatrix}$$

Af tabellen over gruppegennemsnit finder man da

$$\mathbf{sak}_2 = \begin{pmatrix} 14695 & -2744 & 730 & 3804 \\ & 5783 & 1012 & -1305 \\ & & 1046 & 956 \\ & & & 2906 \end{pmatrix}$$

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 $\mathbf{sak}_1$ finder man $\mathbf{sak}_1 = (N-k)\mathbf{S}_1$, således at

$$\mathbf{sak}_1 = \begin{pmatrix} 62681 & 8453 & 12119 & 5832 \\ & 29079 & 16002 & 3924 \\ & & 30122 & 4263 \\ & & & 17591 \end{pmatrix}$$

hvorved $\mathbf{sak}_0$ kan findes som $\mathbf{sak}_0 = \mathbf{sak}_1 + \mathbf{sak}_2$.

Da $\det(\mathbf{sak}_1) = 5.9212 \times 10^{17}$, og $\det(\mathbf{sak}_2) = 10.82 \times 10^{17}$ finder man kvotienttesstørrelsen 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 momentestimaterne

$$\tilde{\mu} = \begin{pmatrix} 188.03 \\ 137.25 \\ 131.91 \\ 50.30 \end{pmatrix}$$

$$\tilde{\mathbf{\Sigma}} = \mathbf{S}_1 = \begin{pmatrix} 43.6500 & 5.8865 & 8.4396 & 4.0610 \\ & 20.2500 & 11.1438 & 2.7326 \\ & & 20.9764 & 2.9688 \\ & & & 12.2500 \end{pmatrix}$$

og

$$\tilde{\Sigma}_0 = \begin{pmatrix} 10.8246 & -2.1383 & 0.4854 & 2.8622 \\ & 4.2341 & 0.6780 & -1.0161 \\ & & 0.6221 & 0.7035 \\ & & & 2.1103 \end{pmatrix}$$

For at vurdere, hvorvidt variationen i middelværdier udfylder hele det 4-dimensionale rum, har vi beregnet determinanten $\det(\tilde{\Sigma}_0) = 5.44$ og endvidere har vi bestemt egenværdierne for $\tilde{\Sigma}_0$. 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.

□

6.8 Oversigtstabeller

Stikprøve-fordeling af $X_i\|\theta$	$\mu = \mathsf{E}[X\|\theta]$	$V(\mu)$	Struktur-fordeling $w(\cdot)$	$m = \mathsf{E}[\mu]$	$\mathsf{E}[V(\mu)]$	$\gamma = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V(\mu)]}$	Reference
$\mathrm{B}(1,p)$	p	$\mu(1-\mu)$	$p \in \mathrm{Be}(\alpha,\beta)$	$\pi = \frac{\alpha}{\alpha+\beta}$	$\frac{\pi(1-\pi)}{1+\gamma}$	$\frac{1}{\alpha+\beta}$	Afsn. 6.2
$\mathrm{Geo}(1,p)$	$\frac{1-p}{p}$	$\mu(1+\mu)$	$p \in \mathrm{Be}(\alpha,\beta)$	$\psi = \frac{\beta}{\alpha-1}$	$\frac{\psi(1+\psi)}{1-\gamma}$	$\frac{1}{\alpha-1}$	Afsn. 6.3
$\mathrm{P}(\mu)$	μ	μ	$\mu \in \mathrm{G}(\alpha, 1/\beta)$	$m = \frac{\alpha}{\beta}$	m	$\frac{1}{\beta}$	Afsn. 6.4
$\mathrm{Ex}(\mu)$	μ	μ^2	$\mu \in \mathrm{RGam}(\alpha, 1/\beta)$	$m = \frac{\beta}{\alpha-1}$	$\frac{m^2}{1-\gamma}$	$\frac{1}{\alpha-1}$	Afsn. 6.5
$\mathrm{N}(\mu,\sigma^2)$	μ	σ^2	$\mathrm{N}(m,\sigma_0^2)$	m	σ^2	σ_0^2/σ^2	Afsn. 5.3

$$\mathsf{E}[\overline{X}_+] = m\,; \qquad \mathsf{V}[\overline{X}_+] = \mathsf{E}[V(\mu)]\left(\gamma + \frac{1}{n}\right)$$

Tabel 6.1. Hierarkiske modeller for endimensionale eksponentielle familier med naturlige konjugerede aprio-rifordelinger

Stikprøve-fordeling af Z	Struktur-fordeling $w(\cdot)$	Marginal fordeling af Z	$\mu = \mathsf{E}[X\|\theta]$	$V(\mu)$	$m = \mathsf{E}[\mu]$	$\gamma = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V(\mu)]}$	$\mathsf{E}[V(\mu)]$
$\mathrm{B}(n,p)$	$p \in \mathrm{Be}(\alpha,\beta)$	$\mathrm{Pl}(n,\alpha,\alpha+\beta)$	p	$\mu(1-\mu)$	$\pi = \frac{\alpha}{\alpha+\beta}$	$\frac{1}{\alpha+\beta}$	$\frac{V(\pi)}{1+\gamma}$
$\mathrm{NB}^*(n,p)$	$p \in \mathrm{Be}(\alpha,\beta)$	$\mathrm{NPl}^*(n,\beta,\alpha+\beta)$	$\frac{p}{1-p}$	$\mu(1+\mu)$	$\psi = \frac{\alpha}{\beta-1}$	$\frac{1}{\beta-1}$	$\frac{V(\psi)}{1-\gamma}$
$\mathrm{P}(n\mu)$	$\mu \in \mathrm{G}(\alpha,1/\beta)$	$\mathrm{NB}(\alpha,\beta/(\beta+n))$	μ	μ	$m = \frac{\alpha}{\beta}$	$\frac{1}{\beta}$	$V(m)$
$\mathrm{G}(n,\mu)$	$\mu \in \mathrm{RGam}(\alpha,1/\beta)$	$\mathrm{RBet}(\alpha,n,\beta)$	μ	μ^2	$m = \frac{\beta}{\alpha-1}$	$\frac{1}{\alpha-1}$	$\frac{V(m)}{1-\gamma}$

$$Z = X_1 + X_2 + \cdots + X_n; \quad Y = Z/n$$

$$\mathsf{E}[Y] = m \qquad \mathsf{V}[Y] = \mathsf{E}[V(\mu)]\left(\gamma + \frac{1}{n}\right)$$

Tabel 6.2. Momenter i de marginale fordelinger ved hierarkisk variation

Marginal fordeling af Z	$m = \mathsf{E}[Y_i]$	$\gamma = \dfrac{\mathsf{V}[\mu]}{\mathsf{E}[V(\mu)]}$	$\tilde{m}$	$\tilde{\gamma}$
$\mathrm{Pl}(n, \alpha, \alpha+\beta)$	$\pi = \dfrac{\alpha}{\alpha+\beta}$	$\dfrac{1}{\alpha+\beta}$	$\tilde{\pi} = \overline{y}_+$	$\dfrac{s_2^2 - \overline{y}_+(1-\overline{y}_+)}{n_0\overline{y}_+(1-\overline{y}_+) - s_2^2}$
$\mathrm{NPl}^*(n, \beta, \alpha+\beta)$	$\psi = \dfrac{\alpha}{\beta-1}$	$\dfrac{1}{\beta-1}$	$\tilde{\psi} = \overline{y}_+$	$\dfrac{s_2^2 - \overline{y}_+(1+\overline{y}_+)}{s_2^2 + n_0\overline{y}_+(1+\overline{y}_+)}$
$\mathrm{NB}(\alpha, \beta/(\beta+n))$	$m = \dfrac{\alpha}{\beta}$	$\dfrac{1}{\beta}$	$\tilde{m} = \overline{y}_+$	$\dfrac{s_2^2/\overline{y}_+ - 1}{n_0}$
$\mathrm{RBet}(\alpha, n, \beta)$	$m = \dfrac{\beta}{\alpha-1}$	$\dfrac{1}{\alpha-1}$	$\tilde{m} = \overline{y}_+$	$\dfrac{s_2^2 - \overline{y}_+^2}{s_2^2 + n_0\overline{y}_+^2}$

$Z_i = X_{i1} + X_{i2} + \cdots + X_{in_i}$; $\quad Y_i = Z_i/n_i$

$$SAK_2 = \sum_{i=1}^{k} n_i(y_i - \overline{y}_+)^2 \; ; \quad s_2^2 = SAK_2/(k-1)$$

$$\overline{y}_+ = \sum_{i=1}^{k} n_i y_i / \sum_{i=1}^{k} n_i$$

n_0 bestemt ved (5.1.9)

Tabel 6.3. Momentestimation af m og γ ved hierarkisk variation

Marginal fordeling af Z	$\tilde{\alpha}$	$\tilde{\beta}$
$\mathrm{Pl}(n,\alpha,\alpha+\beta)$	$\overline{y}_+ \dfrac{n_0\overline{y}_+(1-\overline{y}_+)-s_2^2}{s_2^2-\overline{y}_+(1-\overline{y}_+)}$	$(1-\overline{y}_+)\dfrac{n_0\overline{y}_+(1-\overline{y}_+)-s_2^2}{s_2^2-\overline{y}_+(1-\overline{y}_+)}$
$\mathrm{NPl}^*(n,\beta,\alpha+\beta)$	$\overline{y}_+\left(1+\dfrac{n_0+1}{s_2^2-\overline{y}_+(1+\overline{y}_+)}\right)$	$2+\dfrac{n_0+1}{s_2^2-\overline{y}_+(1+\overline{y}_+)}$
$\mathrm{NB}(\alpha,\beta/(\beta+n))$	$n_0\dfrac{\overline{y}_+^2}{s_2^2-\overline{y}_+}$	$n_0\dfrac{\overline{y}_+}{s_2^2-\overline{y}_+}$
$\mathrm{RBet}(\alpha,n,\beta)$	$1+\dfrac{s_2^2+n_0\overline{y}_+^2}{s_2^2-\overline{y}_+^2}$	$\overline{y}_+\dfrac{s_2^2+n_0\overline{y}_+^2}{s_2^2-\overline{y}_+^2}$

$Z_i = X_{i1}+X_{i2}+\cdots+X_{in_i}; \qquad Y_i = Z_i/n_i$

$$SAK_2 = \sum_{i=1}^{k} n_i(y_i-\overline{y}_+)^2\,; \quad s_2^2 = SAK_2/(k-1)$$

$$\overline{y}_+ = \sum_{i=1}^{k} n_i y_i \Big/ \sum_{i=1}^{k} n_i$$

n_0 bestemt ved (5.1.9)

Tabel 6.4. Momentestimation af α og β ved hierarkisk variation

6.9 Referencer

L.J.Bain og F.T.Wright: The Negative Binomial Process with Applications to Reliability. *Journ. Qual.Techn.* **14**, 1982, pp. 60-66.

J.Bonitzer: Unicité de la solution de l'equation de vraisemblance de la loi binomiale négative, *Revue de Statistique Appliquée*, vol XXVI (1978) n^0 4, pp 55-59)

Consonni, G. and Veronese, P. (1992): Conjugate Priors for Exponential Families Having Quadratic Variance Functions. *Journ. Amer. Statist. Assoc.* **87**, pp 1123-1127.

Diaconis, P. and Ylvisaker, D. (1979): Conjugate Priors for Exponential Families *The Annals of Statistics* **7**, pp 269-281

J.H Ford : *Examples of the process curve*, M.Sc.Thesis, Imperial College, London 1951

Gutiérez-Peña and Smith, A. F. M. (1995): Conjugate Parametrizations for Natural Exponential Families, *Journ. Amer. Statist. Assoc.* **90**, pp 1347-1356.

U. Müller-Funk and F. Pukelsheim: How Regular are conjugate exponential families ? *Statistics & Probability Letters* **7** (1989), pp 327-333

F.Proschan: Theoretical Explanation of Observed Decreasing Failure Rate. *Technometrics* **5**, 1963, pp 375-383.

C.R. Rao: *Linear Statistical Inference and its Applications*, Wiley (1973)

C.R. Rao: The utilization of multiple measurements in problems of biological classification. *Journ. Roy.Statist.Soc. B* **10** (1948), pp. 159-203.

Simonsen, W. (1976): On the solution of a maximum-likelihood equation of the negative binomial distribution, *Scand.Actuarial J.* pp. 220-231, corrigenda *Scand.Actuarial J.* 1980, pp. 41-42 og pp 227-228

P.V.Sukhatme : On the distribution of χ^2 in samples of the Poisson series. *Journ. Roy. Statist. Soc. Suppl.* **5**, (1938) pp 75-79)

Williford, W. O., Carter, M. C. and Hsieh, P. (1974): A Bayesian analysis of two probability models describing thunderstorm activity at Cape Kennedy, Florida. *Journal of Applied Meteorology* **13**, pp. 718 - 725.

Afsnit 7

Lineære normalfordelingsmodeller

7.1 Balancerede regressionsmodeller med varierende koefficienter

fil: linnorm.tex 1997-04-25

7.1.1 Indledning

I eksempel 2.7.1 betragtede vi en regressionsmodel for et sæt normalt fordelte observationer med èn 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 regressionskoefficienternes 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 regressionskoefficienter.

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 forklarende variable er de samme i alle k grupper, dvs. at de k observationssæt har tilknyttet samme, kendte $(n \times p)$-matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \tag{7.1.1}$$

af værdier x_{rj} af de p forklarende variable, hvor x_{rj}, $r = 1, 2, \dots, n$, $j = 1, 2, \dots, p$ angiver den værdi af den j'te forklarende variabel, der svarer til den r'te underobservation. Den j'te søjle i $\mathbf{X}$-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 $\mathbf{X}$ har fuld rang, dvs at $(\mathbf{X}^T\mathbf{X})^{-1}$ eksisterer.

Disse begrænsende forudsætninger indebærer, at notationen og estimationen 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

$$H_0: \quad Y_i = \mathbf{X}\beta_i + \epsilon_i\,,\ i = 1, 2, \dots, k\,; \quad \epsilon_i \in \mathrm{N}_n(0, \sigma^2\mathbf{I}_n)\,, \tag{7.1.2}$$

hvor Y_i angiver de n observationer fra den i'te gruppe skrevet som en n-dimensional søjlevektor; den fælles modelmatrix $\mathbf{X}$ er en kendt $n \times p$ matrix, og hvor β_i, $i = 1, \dots, k$ angiver en p-dimensional vektor af ukendte

regressionskoefficienter, og hvor de n-dimensionale vektorer ϵ_i af "observationsfejl" er indbyrdes uafhængige, $i = 1, \ldots, k$.

Antagelsen $\epsilon_i \in \mathrm{N}_n(0, \sigma^2 \mathbf{I}_n)$ indebærer, at observationsfejlene ikke kun er uafhængige imellem grupper, men at også de enkelte observationsfejl ϵ_{ir}, $r = 1, \ldots, n$ indenfor samme gruppe er indbyrdes uafhængige.

Skrevet fuldt ud er modellen svarende til den i'te gruppe:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_{i1} \\ \vdots \\ \beta_{ip} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in} \end{pmatrix}$$

Modellen er en generaliseret lineær model for normaltfordelte observationer.

Modellen svarer til variansfunktionen $V(\mu) = 1$; dispersionsparameteren er σ^2, og linkfunktionen $\eta(\mu) = \mu$ er identiteten, dvs netop den kanoniske linkfunktion.

Modellen (7.1.2) kan udtrykkes ved modelformlen

$$Y = A.X_1 + A.X_2 + \cdots + A.X_p \ , \tag{7.1.3}$$

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 gruppper.

I modelformlen har vi ikke eksplicit 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 ettaller 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 $\mathbf{X}$-matricen bestå af ettaller.

Modellen (7.1.2) har den parametriske fremstilling

$$\mu_{ir} = \beta_{i1} x_{r1} + \beta_{i2} x_{r2} + \cdots + \beta_{ip} x_{rp} \ , \tag{7.1.4}$$

for $i = 1, \ldots, k$, $r = 1, \ldots, n$.

Modellen udtrykker netop, at regressionskoefficienterne $\boldsymbol{\beta}_i$ for de k grupper tillades at være forskellige.

Modelmatricen for de kn observationer er den $kn \times kp$-dimensionale matrix $\mathbf{X}_0$ givet ved

$$\mathbf{X}_0 = \begin{pmatrix} \mathbf{X} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{X} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{X} \end{pmatrix}. \tag{7.1.5}$$

Modelmatricen (7.1.5) kan udtrykkes mere kompakt som

$$\mathbf{X}_0 = \mathbf{X} \otimes \mathbf{I}_k \tag{7.1.6}$$

hvor $\mathbf{X} \otimes \mathbf{I}_k$ angiver tensorproduktet mellem den $n \times p$-dimensionale matrix $\mathbf{X}$ og den $k \times k$-dimensionale enhedsmatrix $\mathbf{I}_k$.

Sætning 7.1.1 *Maksimum likelihood estimation i den balancerede regressionsmodel*

Under modellen (7.1.2) er maksimum likelihood estimatoren for $\boldsymbol{\beta}_i$ givet ved

$$\widehat{\boldsymbol{\beta}}_i = \mathbf{P}\mathbf{y}_i \tag{7.1.7}$$

med

$$\mathbf{P} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{7.1.8}$$

og den marginale maksimum likelihood estimator for σ^2 er

$$\widehat{\sigma}^2 = \frac{sak_{1\cdot}}{k(n-p)} \tag{7.1.9}$$

hvor

$$sak_{1\cdot} = \sum_{i=1}^{k} sak_{1,i} \tag{7.1.10}$$

med

$$sak_{1,i} = (\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i)^T(\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i) \ .$$

Der gælder

$$\widehat{\boldsymbol{\beta}}_i \in \mathrm{N}_p(\boldsymbol{\beta}_i, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \ ; \quad SAK_{1\cdot} \in \sigma^2\chi^2(k(n-p)) \tag{7.1.11}$$

og $\widehat{\boldsymbol{\beta}}_i$ og $SAK_{1\cdot}$ er indbyrdes uafhængige.

Bevis:

Loglikelihoodfunktionen er

$$l(\boldsymbol{\beta}; \mathbf{y}_1, \ldots, \mathbf{y}_k) = -\frac{1}{2\sigma^2}\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i)^T(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i) - (k/2)\ln[n\sigma^2]$$

Modellen (7.1.2) specificerer relationen

$$\boldsymbol{\mu} = \mathbf{X}_0\boldsymbol{\beta} \ , \tag{7.1.12}$$

mellem middelværdierne, hvor modelmatricen $\mathbf{X}_0$ er givet ved (7.1.5).

Middelværdiligningen (2.7.1) svarende til modellen (7.1.12) er da

$$\mathbf{X}_0^T\mathbf{y} = \mathbf{X}_0^T\mathbf{X}_0\boldsymbol{\beta} \ .$$

Ligningen har løsningen

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{y} \,. \tag{7.1.13}$$

Indsætter vi nu $\mathbf{X}_0 = \mathbf{X} \otimes \mathbf{I}_k$ fra (7.1.6), får vi

$$(\mathbf{X}_0^T\mathbf{X}_0)^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} \otimes \mathbf{I}_k \,,$$

således at parameterprojektionsmatricen $\mathbf{P}_0$, der projicerer den kn-dimensionale observationsvektor Y ned på det kp-dimensionale parameterrum, kan udtrykkes som

$$\begin{aligned} \mathbf{P}_0 &= [(\mathbf{X}^T\mathbf{X})^{-1} \otimes \mathbf{I}_k][\mathbf{I}_k \otimes \mathbf{X}^T] \\ &= [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] \otimes \mathbf{I}_k = \mathbf{P} \otimes \mathbf{I}_k \,, \end{aligned}$$

hvor vi har indført de individuelle parameterprojektionsmatricer $\mathbf{P}$ givet ved (7.1.8).

Udtrykket (7.1.13) for estimatet $\widehat{\boldsymbol{\beta}}$ bliver da

$$\widehat{\boldsymbol{\beta}} = \mathbf{P}_0 Y = [\mathbf{P} \otimes \mathbf{I}_k]\mathbf{y} \,,$$

der netop spalter op i de k individuelle estimater (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 $-(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i)^T(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i)$, der kan maksimeres enkeltvis med hensyn til $\boldsymbol{\beta}_i$).

Tilsvarende finder man, at hat-matricen, der fører observationsvektoren $\mathbf{y}$ over i de fittede værdier, $\widehat{\boldsymbol{\mu}}$, bliver

$$\begin{aligned} \mathbf{H}_0 &= \mathbf{X}_0\mathbf{P}_0 = \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T \\ &= \mathbf{H} \otimes \mathbf{I}_k \,, \end{aligned}$$

hvor vi har indført hat-matricen $\mathbf{H}$ svarende til de individuelle regressionsplaner:

$$\mathbf{H} = \mathbf{X}\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \,. \tag{7.1.14}$$

Vektoren $\mathbf{R}_0$ af residualer bliver

$$\begin{aligned} \mathbf{R}_0 &= \mathbf{y} - \mathbf{X}_0\widehat{\boldsymbol{\beta}} = (\mathbf{I}_{nk} - \mathbf{H}_0)\mathbf{y} = (\mathbf{I}_n \otimes \mathbf{I}_k - \mathbf{H} \otimes \mathbf{I}_k)\mathbf{y} \\ &= [(\mathbf{I}_n - \mathbf{H}) \otimes I_k]\mathbf{y} \,. \end{aligned}$$

Den kvadratiske form $(\mathbf{I}_{nk} - \mathbf{H}_0)^T(\mathbf{I}_{nk} - \mathbf{H}_0)$ svarende til residualkvadratsummen (deviansen),

$$sak_{1\cdot} = D(\mathbf{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{X}_0\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}_0\widehat{\boldsymbol{\beta}}) = \mathbf{R}_0^T\mathbf{R}_0$$

spalter i en sum af ens former

$$(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H}) ,$$

svarende til residualkvadratsummerne

$$sak_{1,i} = \mathbf{y}_i^T(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H})\mathbf{y}_i = (\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i)^T(\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i) .$$

Profilloglikelihood'en for σ^2 (med hensyn til $\boldsymbol{\beta}$) er

$$\tilde{l}(\sigma^2; \mathbf{y}) = -\frac{1}{\sigma^2}\sum_{i=1}^{k} sak_{1,i} - (k/2)\ln(n\sigma^2) ,$$

hvorfor $sak_{1\cdot} = \sum sak_{1,i}$ er likelihood-sufficient for σ^2. Man vil derfor benytte den marginale likelihood svarende til $sak_{1\cdot}$ og bestemme den marginale likelihood estimator for σ^2.

For at betemme fordelingen af $sak_{1\cdot}$ bemærker vi, at da $\mathbf{H}$ er en projektionsmatrix, er den idempotent. Endvidere er $\mathbf{H}$ symmetrisk, og der gælder derfor

$$\begin{aligned}(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H}) &= \mathbf{I}_n - \mathbf{H}^T - \mathbf{H} + \mathbf{H}\mathbf{H}^T \\ &= \mathbf{I}_n - \mathbf{H} .\end{aligned}$$

Da $\mathbf{X}^T\mathbf{X}$ er antaget at have fuld rang p, har den kvadratiske form bestemt ved $(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H})$ rangen $n - p$, og residualkvadratsummen $sak_{1,i} =$ følger derfor en $\sigma^2\chi^2(n-p)$-fordeling, og fordelingen af $sak_{1,i}$ er uafhængig af $\widehat{\boldsymbol{\beta}}_i = \mathbf{P}Y_i$, svarende til at residualerne $(\mathbf{I}_n - \mathbf{H})\mathbf{y}_i = \mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i$ er ortogonale på underrummet udspændt af søjlerne i $\mathbf{X}$.

Da observationssættet $\mathbf{y}_i$ fra én gruppe er uafhængigt af observationssættet $\mathbf{y}_j$ fra en anden gruppe, er også residualkvadratsummerne $sak_{1,i}$ indbyrdes uafhængige. Det følger da, at fordelingen af summen $sak_{1\cdot}$ er en $\sigma^2\chi^2(k(n-p))$-fordeling. Den marginale maksimum likelihood for σ^2 bliver derfor (7.1.9).

Da $\widehat{\boldsymbol{\beta}}_i = \mathbf{P}\mathbf{y}_i$ er en lineær transformation af den $\mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}_i, \sigma^2\mathbf{I}_n)$-fordelte størrelse $\mathbf{y}_i$, fås fordelingen af $\widehat{\boldsymbol{\beta}}_i$ umiddelbart ved brug af transformationssætningen for normalfordelingen.

□

Bemærkning 1 *Estimation ved observationer med vilkårlig, kendt dispersonsmatrix*

I formuleringenen af ovenstående sætning antog vi, at observationsfejlene ϵ_{ir} havde samme varians σ^2 for alle værdier af de forklarende variable, og desuden, at observationsfejlene indenfor samme gruppe var indbyrdes uafhængige, nemlig at pånær dispersionsparameteren σ^2 kunne dispersionsmatricen for vektoren $\boldsymbol{\epsilon}_i$ 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 fremkomer ved at man i (7.1.2) erstatter antagelsen $\boldsymbol{\epsilon}_i \in \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ med

$$H_0: \quad Y_i = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i\,,\ i = 1, 2, \ldots, k\,; \quad \boldsymbol{\epsilon}_i \in \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{V})\,, \tag{7.1.15}$$

hvor $\mathbf{V}$ er en kendt symmetrisk positiv definit $n \times n$ matrix.

Modellen (7.1.15) svarer til en vægtet model. Hvis $\mathbf{V}$ er en diagonalmatrix, har observation Y_{ir} tilknyttet vægten $w_r = V_{rr}^{-1}$. I det generelle tilfælde er vægtmatricen $\mathbf{W} = \mathbf{V}^{-1}$.

Under modellen (7.1.15) får man loglikelihoodfunktionen

$$l(\boldsymbol{\beta}; \mathbf{y}_1, \ldots, \mathbf{y}_k) = -\frac{1}{2\sigma^2}\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i)^T\mathbf{V}^{-1}(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_i) - (k/2)\ln[\sigma^2 \det(\mathbf{V})]$$

Maksimaliseringsestimatoren er

$$\widehat{\boldsymbol{\beta}}_i = \mathbf{P}_w \mathbf{y}_i \tag{7.1.16}$$

med

$$\mathbf{P}_w = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \tag{7.1.17}$$

og den marginale maksimaliseringsestimator for σ^2 er givet ved (7.1.9), hvor

$$sak_{1,i} = (\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i)^T \mathbf{V}^{-1} (\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i) \; . \tag{7.1.18}$$

Der gælder

$$\widehat{\boldsymbol{\beta}}_i \in \mathrm{N}_p\big(\boldsymbol{\beta}_i, \sigma^2 (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}\big) \; ; \quad SAK_{1\cdot} \in \sigma^2 \chi^2\big(k(n-p)\big)$$

□

En nærliggende reduktion af modellen (7.1.2) er en hypotese om at regressionskoefficienterne er de samme i alle k grupper, dvs.

$$H_1 : \quad \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_k \; (= \boldsymbol{\beta}_0) \; ,$$

imod alternativet

$$H_1^{\mathrm{c}} : \; \boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j \;\; \text{for mindst eet sæt } (i,j) \text{ med } i \neq j$$

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

$$H_1: \quad \beta_1 = \beta_2 = \cdots = \beta_k \ (= \beta_0)\,, \tag{7.1.19}$$

teststørrelsen

$$Z = \frac{SAK_2/[p(k-1)]}{SAK_1/(k(n-p))} \tag{7.1.20}$$

hvor

$$SAK_2 = \sum_{i=1}^{k} (\widehat{\beta}_i - \widehat{\beta}_0)^T \mathbf{X}^T \mathbf{X} (\widehat{\beta}_i - \widehat{\beta}_0) \tag{7.1.21}$$

med

$$\widehat{\beta}_0 = \mathbf{P} \left(\sum_{i=1}^{k} \mathbf{y}_i / k \right)\,, \tag{7.1.22}$$

hvor parameterprojektionsmatricen $\mathbf{P}$ er givet ved (7.1.8).

Under H_1 følger Z en $\mathrm{F}\big(p(k-1), k(n-p)\big)$-fordeling.

Testet forkaster for store værdier af z.

Bevis:
Hypotesen H_1 svarer til at observationerne kan beskrives som k identisk fordelte gentagelser af $Y_i \in \mathrm{N}_n(\mathbf{X}\beta_0, \sigma^2 \mathbf{I}_n)$.

Under H_1 er modelmatricen den $(kn \times p)$-dimensionale matrix

$$\mathbf{M}_1 = \begin{pmatrix} \mathbf{X} \\ \mathbf{X} \\ \cdots \\ \mathbf{X} \end{pmatrix} = \mathbf{X} \otimes \mathbf{1}_k\,,$$

hvor $\mathbf{1}_k$ angiver en k-dimensional søjlevektor med ettaller.

Der gælder

$$\mathbf{M}_1^T\mathbf{M}_1 = (\mathbf{X}^T \otimes \mathbf{1}_k^T)(\mathbf{X} \otimes \mathbf{1}_k) = (\mathbf{X}^T\mathbf{X}) \otimes (\mathbf{1}_k^T\mathbf{1}_k) = k(\mathbf{X}^T\mathbf{X}) ,$$

hvorfor man finder parameterprojektionsmatricen

$$\mathbf{P}_1 = (\mathbf{M}_1^T\mathbf{M}_1)^{-1}\mathbf{M}_1^T = \frac{1}{k}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \otimes \mathbf{1}_k = \frac{1}{k}\mathbf{P} \otimes \mathbf{1}_k ,$$

der netop fører til at estimatet $\widehat{\boldsymbol{\beta}}_0$ bestemmes som regressionen på de observerede gennemsnit

$$\bar{y}_{\cdot r} = \sum_{i=1}^{k} y_{ir}/k$$

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

$$\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_0)^T(\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_0) = \mathrm{sak}_{1\cdot} + \mathrm{sak}_2 \qquad (7.1.23)$$

af residualkvadratsummen $G^2(H_1)$

$$G^2(H_1) = \mathrm{SAK}_0 = (\mathbf{y} - \mathbf{M}_1\widehat{\boldsymbol{\beta}}_0)^T(\mathbf{y} - \mathbf{M}_1\widehat{\boldsymbol{\beta}}_0)$$

svarende til H_1.

Det følger af spaltningssætningen, at under H_1 vil $\mathrm{SAK}_2 \in \sigma^2\chi^2(p(k-1))$ 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.

Variationskilde	f	Devians	middeldevians	Test
Mellem H_1 og H_0	$p(k-1)$	$D(\mu(\widehat{\beta});\mu(\widehat{\beta}_0))$	$\dfrac{D(\mu(\widehat{\beta});\mu(\widehat{\beta}_0))}{p(k-1)}$	$\dfrac{D(\mu(\widehat{\beta});\mu(\widehat{\beta}_0))/[p(k-1)]}{D(\mathbf{y};\mu(\widehat{\beta}_0))/[k(n-p)]}$
Afvigelse fra H_0	$k(n-p)$	$D(\mathbf{y};\mu(\widehat{\beta}))$	$\hat{\sigma}^2 = \dfrac{D(\mathbf{y};\mu(\widehat{\beta}))}{k(n-p)}$	
Total	$kn-p$	$D(\mathbf{y};\mu(\widehat{\beta}_0))$	$\dfrac{D(\mathbf{y};\mu(\widehat{\beta}_0))}{kn-p}$	

Tabel 7.1. Deviansopspaltning svarende til hierarkisk modelreduktion af regressionsmodellen i sætning 7.1.2

Variation	SAK	f
Mellem regressionsplaner	$sak_2 = \sum_{i=1}^{k} (\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)^T \mathbf{X}^T \mathbf{X} (\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)$	$p(k-1)$
Omkring individuelle planer	$sak_{1\cdot} = \sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i)^T (\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_i)$	$k(n-p)$
Omkring fælles plan	$\sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_0)^T (\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_0)$	$kn-p$

Det fremgår ved sammenligning af udtrykket (7.1.21) for SAK_2 med udtrykket (7.1.11) for fordelingen af $\widehat{\boldsymbol{\beta}}_i$, at afvigelserne $(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)$ vægtes med deres præcisioner $\mathbf{X}^T\mathbf{X}$ (den inverse dispersionsmatrix).

□

Bemærkning 2 *Test for samme indflydelse af enkelte af de forklarende variable*

Ovenstående test sammenligner hele parametervektoren $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}_0$ under hypotesen (7.1.19)

$$\widehat{\boldsymbol{\beta}}_0 = \mathbf{P}_w[\mathbf{y}_i/k]\,, \tag{7.1.24}$$

hvor parameterprojektionsmatricen, $\mathbf{P}_w$, er givet ved (7.1.17), og teststørrelsen for hypotesen (7.1.19)

$$Z = \frac{SAK_2/[p(k-1)]}{SAK_{1\cdot}/(N-pk)}$$

hvor

$$SAK_2 = \sum_{i=1}^{k} (\widehat{\beta}_i - \widehat{\beta}_0)^T \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} (\widehat{\beta}_i - \widehat{\beta}_0) \tag{7.1.25}$$

med sak_1. bestemt ved brug af (7.1.18).

Under H_1 følger Z en $\mathrm{F}(p(k-1), k(n-p))$-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 $\boldsymbol{\beta}_i \in \mathrm{N}_p(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Gamma})$, hvor $\boldsymbol{\beta}_0$ angiver den ukendte middel-parametervektor, og den $p \times p$-dimensionale symmetriske, positiv definite matrix $\sigma^2\boldsymbol{\Gamma}$ angiver dispersionsmatricen for fordelingen af $\boldsymbol{\beta}_i$ omkring $\boldsymbol{\beta}_0$. 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 σ^2.

Fordelingsforholdene under denne model fremgår af

Sætning 7.1.3 ***Marginal fordeling af observationer og estimater under regressionsmodel med tilfældige koeffficienter***

Lad

$$Y_i = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \ i = 1, 2, \ldots, k$$

hvor

$$\boldsymbol{\beta}_i \in \mathrm{N}_p(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Gamma}), \ \ \boldsymbol{\epsilon}_i \in \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \ , \tag{7.1.26}$$

og hvor $\boldsymbol{\beta}_i, \boldsymbol{\beta}_j$ er indbyrdes uafhængige for $i \neq j$, og $\boldsymbol{\epsilon}_i$ og $\boldsymbol{\epsilon}_j$ er indbyrdes uafhængige for $i \neq j$, og endvidere $\boldsymbol{\beta}_i$ og $\boldsymbol{\epsilon}_j$ er uafhængige.

Da er den marginale fordeling af Y_i givet ved

$$Y_i \in \mathrm{N}_n\big(\mathbf{X}\boldsymbol{\beta}_0, \sigma^2\{\mathbf{I}_n + \mathbf{X}\boldsymbol{\Gamma}\mathbf{X}^T\}\big) \ ,$$

og den marginale fordeling af estimaterne, $\widehat{\boldsymbol{\beta}}_i$ (7.1.7), for regressionskoefficienterne er

$$\widehat{\boldsymbol{\beta}}_i \in \mathrm{N}_p\left(\boldsymbol{\beta}_0, \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1} + \boldsymbol{\Gamma}]\right) \tag{7.1.27}$$

Bevis:
Udtrykket for den marginale dispersionsmatrix fås ved at bemærke, at

$$\mathbf{D}\,[Y_i] = \mathrm{E}\,[\,\mathbf{D}\,[Y_i|\boldsymbol{\beta}]\,] + \mathbf{D}\,[\mathrm{E}\,[Y_i|\boldsymbol{\beta}]\,] = \sigma^2\mathbf{I}_n + \mathbf{D}\,[\mathbf{X}\boldsymbol{\beta}]$$

jvf Sætning 0.1.1, formel (0.1.2) i Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

□

Bemærkning 1 ***Fordelingen af estimaterne for regressionskoefficienterne omfattter både estimationsusikkerheden og variansen mellem grupper***

Fordelingen af estimaterne $\widehat{\boldsymbol{\beta}}_i$ er analog til fordelingen af gruppegennemsnit i de modeller, vi betragtede i afsnit 5 og 6. Fordelingen indeholder et bidrag, $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ svarende til estimationsusikkerheden, og et bidrag, $\sigma^2\boldsymbol{\Gamma}$ svarende til dispersionen mellem gruppernes regressionskoefficienter. Da estimatet $\widehat{\boldsymbol{\beta}}_i$, 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.

□

Bemærkning 2 ***Fordeling af estimater ved observationer med vilkårlig, kendt dispersionsmatrix***

Såfremt man i (7.1.26) antager den mere generelle fordeling af observationsfejlene,

$$\boldsymbol{\epsilon}_i \in \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{V})\,, \tag{7.1.28}$$

hvor $\mathbf{V}$ er en kendt symmetrisk positiv definit $n \times n$ matrix, da gælder at den marginale fordeling af Y_i er

$$Y_i \in \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}_0, \sigma^2\{\mathbf{V} + \mathbf{X}\boldsymbol{\Gamma}\mathbf{X}^T\})\,,$$

og den marginale fordeling af estimaterne, $\widehat{\boldsymbol{\beta}}_i$ (7.1.16), for regressionskoefficienterne er

$$\widehat{\boldsymbol{\beta}}_i \in \mathrm{N}_p(\boldsymbol{\beta}_0, \sigma^2\{(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} + \boldsymbol{\Gamma}\}) \tag{7.1.29}$$

□

Sætning 7.1.4 *Maksimaliseringsestimation for regressionsmodel med tilfældige koeffficienter*

Under antagelserne fra Sætning 7.1.3 fås maksimaliseringsestimatorerne for parametrene $\boldsymbol{\beta}_0$, σ^2 og $\boldsymbol{\Gamma}$ ved at maksimere

$$\begin{aligned} l(\boldsymbol{\beta}_0, \sigma^2, \boldsymbol{\Gamma};\, \mathbf{y}_1, \ldots, \mathbf{y}_k) &= -\frac{k}{2}\{n\ln(\sigma^2) + \ln[\det(\mathbf{D}(\boldsymbol{\Gamma}))] \\ &\quad + \frac{1}{\sigma^2}\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_0)^T(\mathbf{D}(\boldsymbol{\Gamma}))^{-1}(\mathbf{y}_i - \mathbf{X}\boldsymbol{\beta}_0) \end{aligned} \tag{7.1.30}$$

hvor

$$\mathbf{D}(\boldsymbol{\Gamma}) = \mathbf{X}\boldsymbol{\Gamma}\mathbf{X}^T + \mathbf{I}_n \tag{7.1.31}$$

med hensyn til $\sigma^2 > 0$, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ og $\boldsymbol{\Gamma}$ positiv semidefinit.

Bevis:

Sætningen vises ved at bemærke, at (7.1.30) netop er logaritmen til likelihoodfunktionen.

□

Bemærkning 1 *Likelihoodligningerne ved maksimum i et indre punkt*

Såfremt maksimumværdien findes i et indre punkt, tilfredsstiller estimaterne udtrykkene

$$\widehat{\boldsymbol{\beta}}_0 = \frac{1}{k}[\mathbf{X}^T(\mathbf{D}(\boldsymbol{\Gamma}))^{-1}\mathbf{X}]^{-1}\mathbf{X}^T(\mathbf{D}(\boldsymbol{\Gamma}))^{-1}\left(\sum_{i=1}^{k}\mathbf{y}_i\right) \tag{7.1.32}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_0)^T(\mathbf{D}(\widehat{\boldsymbol{\Gamma}}))^{-1}(\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_0) \tag{7.1.33}$$

hvor $\widehat{\mathbf{\Gamma}}$ er bestemt som den positive semidefinitte $p \times p$-matrix, der maksimerer udtrykket

$$l(\mathbf{\Gamma}) = -N \times \ln(sak^*(\mathbf{\Gamma})) - k \times \ln[\det(\mathbf{D}(\mathbf{\Gamma}))] \tag{7.1.34}$$

hvor

$$sak^*(\mathbf{\Gamma}) = \sum_{i=1}^{k} \mathbf{y}_i^T \left(\mathbf{D}(\mathbf{\Gamma})\right)^{-1} \mathbf{y}_i$$
$$\frac{1}{k}\left(\sum_{i=1}^{k} \mathbf{y}_i^T\right)\left(\mathbf{D}(\mathbf{\Gamma})\right)^{-1}\mathbf{X}[\mathbf{X}^T\left(\mathbf{D}(\mathbf{\Gamma})\right)^{-1}\mathbf{X}]^{-1}\mathbf{X}^T\left(\mathbf{D}(\mathbf{\Gamma})\right)^{-1}\left(\sum_{i=1}^{k} \mathbf{y}_i\right)$$

Maksimaliseringsestimatet for $\mathbf{\Gamma}$ kan da bestemmes ved en numerisk bestemmelse af maksimum for (7.1.34) over rummet af symmetriske positiv semidefinite $p \times p$-matricer $\mathbf{\Gamma}$. Maksimaliseringsestimaterne for de øvrige parametre kan herefter bestemmes ved at indsætte den fundne værdi i (7.1.32) og (7.1.33). (Se f.eks. Thyregod (1983))

□

Bemærkning 2 ***Central estimator, REML-estimation*** Det er velkendt, at maksimaliseringsestimatoren ikke nødvendigvis er central. Såfremt $\mathbf{\Gamma}$ var kendt, ville estimatoren

$$\breve{\sigma}^2 = \frac{1}{N-p}\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_0)^T\left(\mathbf{D}(\widehat{\mathbf{\Gamma}})\right)^{-1}(\mathbf{y}_i - \mathbf{X}\widehat{\boldsymbol{\beta}}_0)$$

være central for σ^2. Man benytter derfor ofte estimatet baseret på maksimering af likelihoodfunktionen svarende til fordelingen af residualerne (svarende til $\mathbf{X}\widehat{\boldsymbol{\beta}}$), det såkaldte REML-estimat (se bemærkning 3 til sætning 5.4.1)

□

Bemærkning 3 ***Estimation ved observationer med vilkårlig, kendt dispersionsmatrix***

Sætningen og bemærkningerne gælder også i den generelle situation, hvor dispersionsmatricen for observationsfejlene ϵ_i er $\sigma^2\mathbf{V}$. Man skal blot erstatte udtrykket (7.1.31) for $\mathbf{D}(\mathbf{\Gamma})$ med

$$\mathbf{D}(\mathbf{\Gamma}) = \mathbf{X}\mathbf{\Gamma}\mathbf{X}^T + \mathbf{V} \tag{7.1.35}$$

□

Sætning 7.1.5 *Momentestimation for regressionsmodel med tilfældige koefficienter*

Under antagelserne fra sætning 7.1.3 bestemmes momentestimatorerne for parametrene $\boldsymbol{\beta}_0$, σ^2 og $\boldsymbol{\Gamma}$ ved

$$\widetilde{\boldsymbol{\beta}}_0 = \frac{1}{k}\sum_{i=1}^{k}\widehat{\boldsymbol{\beta}}_i = \mathbf{P}\left(\sum_{i=1}^{k}\mathbf{y}_i/k\right) \tag{7.1.36}$$

$$\tilde{\sigma}^2 = \frac{1}{k(n-p)}\,\mathrm{sak}_1. \tag{7.1.37}$$

$$\widetilde{\boldsymbol{\Gamma}} = \frac{\mathbf{sak}_{\boldsymbol{\beta}}}{(k-1)\tilde{\sigma}^2} - (\mathbf{X}^T\mathbf{X})^{-1}, \tag{7.1.38}$$

hvor $\widehat{\boldsymbol{\beta}}_i$, $\mathbf{P}$ og sak_1. er givet ved (7.1.8), (7.1.7) og (7.1.10), og hvor

$$\mathbf{sak}_{\boldsymbol{\beta}} = \sum_{i=1}^{k}(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)^T$$

Der gælder, at

$$\mathbf{sak}_{\boldsymbol{\beta}} \in \mathrm{Wis}_p\left(k-1, \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1} + \boldsymbol{\Gamma}]\right)$$

og at

$$\widetilde{\boldsymbol{\beta}}_0 \in \mathrm{N}_p\left(\boldsymbol{\beta}_0, (\sigma^2/k)[(\mathbf{X}^T\mathbf{X})^{-1} + \boldsymbol{\Gamma}]\right)$$

er central og variansminimal.

Bevis:

Sætningen vises ved at bemærke, at $\widetilde{\boldsymbol{\beta}}_0$ er gennemsnittet, og $\mathbf{sak}_{\boldsymbol{\beta}}/(k-1)$ er den empiriske dispersionsmatrix for $\widehat{\boldsymbol{\beta}}_i$.

Idet $\widehat{\boldsymbol{\beta}}_i$ er indbyrdes uafhængige med

$$\mathsf{E}\,[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_0$$

og

$$\mathbf{D}\,[\widehat{\boldsymbol{\beta}}_i] = \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1} + \boldsymbol{\Gamma}])$$

ifølge (7.1.27), har vi

$$\begin{aligned} \mathsf{E}\,[\widetilde{\beta}_0] &= \beta_0, \quad \text{og} \\ \mathsf{E}\,[\mathbf{sak}_\beta] &= \left(1-\frac{1}{k}\right)\sum_{i=1}^{k} \mathbf{D}\,[\widehat{\beta}_i] = (k-1)\sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}+\boldsymbol{\Gamma}] \end{aligned}$$

□

Bemærkning 1 *Estimaterne for σ^2 og β udnytter også variationen mellem grupper*

Sammenligner vi momentestimaterne (7.1.36) og (7.1.37) for β_0 og σ^2 med maksimaliseringsestimaterne (7.1.32) og (7.1.33) ser vi, at i modsætning til momentestimaterne tilgodeser maksimaliseringsestimaterne for σ^2 og $\boldsymbol{\Gamma}$ kendskabet til dispersionsmatricen β ved den indbyrdes vægtning af de enkelte komponenter af β_i.

Vi vil her ikke komme nærmere ind på test af hypoteser vedrørende $\boldsymbol{\Gamma}$. I Bondeson (1989) er en række forskellige tests diskuteret.

□

Bemærkning 2 *Estimation ved observationer med vilkårlig, kendt dispersionsmatrix*

Sætningen udvides let til at dække også i den generelle situation, hvor dispersionsmatricen for observationsfejlene ϵ_i er $\sigma^2\mathbf{V}$.

Lad modellen være som i sætning 7.1.3, men lad fordelingen af ϵ_i være givet ved (7.1.28).

Da fås momentestimaterne for β_0, σ^2 og $\boldsymbol{\Gamma}$ ved

$$\widetilde{\boldsymbol{\beta}}_0 = \frac{1}{k}\sum_{i=1}^{k}\widehat{\boldsymbol{\beta}}_i = \mathbf{P}_w\left(\sum_{i=1}^{k}\mathbf{y}_i/k\right) \tag{7.1.39}$$

$$\tilde{\sigma}^2 = \frac{1}{k(n-p)}\,\mathrm{sak}_1. \tag{7.1.40}$$

$$\widetilde{\boldsymbol{\Gamma}} = \frac{\mathbf{sak}_\beta}{(k-1)\tilde{\sigma}^2} - (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}, \tag{7.1.41}$$

hvor $\widehat{\boldsymbol{\beta}}_i$, $\mathbf{P}_w$ og sak_1. er givet ved (7.1.17), (7.1.16) og (7.1.18), og hvor

$$\mathbf{sak}_\beta = \sum_{i=1}^{k}(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)^T$$

Der gælder, at

$$\mathbf{sak}_\beta \in \mathrm{Wis}_p(k-1, \sigma^2[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} + \boldsymbol{\Gamma}])$$

og at

$$\widetilde{\boldsymbol{\beta}}_0 \in \mathrm{N}_p(\boldsymbol{\beta}_0, (\sigma^2/k)[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} + \boldsymbol{\Gamma}])$$

er central og variansminimal.

□

Eksempel 7.1.1 ***Vækst af Ramus-knoglen for 5 drenge, momentestimation***

Til illustration af beregningerne betragter vi samhørende værdier af alder og ramushøjde for fem drenge. Data er udvalgte data fra Elston og Grizzle (1962).

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 normaltfordelt 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.

	Alder i år					
	8	8 1/2	9	9 1/2		
Dreng	reduceret alder x_i = alder -8.75				β_{i1}	β_{i2}
	-0.75	-0.25	0.25	0.75		
A	52.5	53.2	53.3	53.7	53.175	0.74
B	51.2	53.0	54.3	54.5	53.250	2.24
C	51.2	51.4	51.6	51.9	51.525	0.46
D	52.1	52.8	53.7	55.0	53.400	1.92
E	50.7	51.7	52.7	53.3	52.100	1.76
snit	51.54	52.42	53.12	53.68	52.690	1.424

Tilsvarende vil vi her modellere relationen mellem ramushøjde og alder ved en lineær regressionsmodel, der tillader en individuel vækstrelation for hver dreng. Vi formulerer 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 ϵ_{ij} antages uafhængige identisk normaltfordelte $\mathrm{N}(0, \sigma^2)$.

Vi har altså modelmatricen

$$\mathbf{X} = \begin{pmatrix} 1.0 & -0.75 \\ 1.0 & -0.25 \\ 1.0 & 0.25 \\ 1.0 & 0.75 \end{pmatrix} \quad \text{med} \quad \mathbf{X}^T\mathbf{X} = \begin{pmatrix} 4.00 & 0.00 \\ 0.00 & 1.25 \end{pmatrix}$$

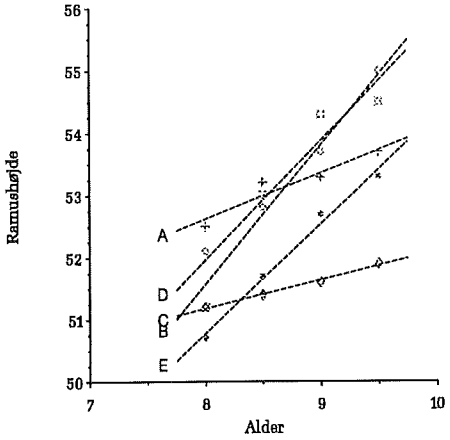
hvorfor

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.25 & 0.00 \\ 0.00 & 0.80 \end{pmatrix}$$

Skønnene over de individuelle regressionkoefficienter $\widehat{\boldsymbol{\beta}}_i$ beregnet ved $\widehat{\boldsymbol{\beta}}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_i$ 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 ramushøide 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

$$\hat{\sigma}^2 = \text{sak}_1./10 = 0.0864 = (0.2939 \text{ [mm]})^2$$

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

$$sak_2 = sak_0 - sak_{1\cdot} = 14.0604 \text{ [mm]}^2$$

Resultatet kan samles i variansanalyseskemaet

Variation	sak	f	sak/f
Mellem individuelle regressionslinier	14.0604	8	1.7575
Omkring individuelle regressionslinier	0.8640	10	0.0864
Omkring fælles linie	14.9244	18	

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 $\boldsymbol{\beta}_i \in N(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Gamma})$

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 $\boldsymbol{\beta}_i$. Man får

$$\mathbf{sak}_\beta = \begin{pmatrix} 2.4219 & 1.4022 \\ 1.4022 & 2.7583 \end{pmatrix}$$

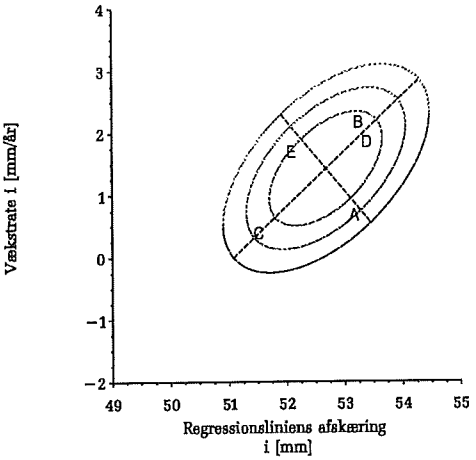
Niveaukurverne svarende til denne fordeling er indtegnet i figur 7.2.

Vi finder nu momentskønnet over $\boldsymbol{\Gamma}$

$$\begin{aligned} \widetilde{\boldsymbol{\Gamma}} &= \frac{\mathbf{sak}_\beta}{4 \times 0.0864} - (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 7.9812 & 4.0573 \\ 4.0573 & 7.0072 \end{pmatrix} - \begin{pmatrix} 0.25 & 0.00 \\ 0.00 & 0.80 \end{pmatrix} \\ &= \begin{pmatrix} 7.7312 & 4.0573 \\ 4.0573 & 6.2072 \end{pmatrix} \end{aligned}$$

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

$$\hat{\rho} = 4.0573/\sqrt{7.7312 \times 6.2072} = 0.54$$

Vi kan altså beskrive ramushøjderne for 8-10 årige 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. Residualvariansen 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

$$\hat{\sigma}^2\widetilde{\Gamma} = 0.2939^2 \begin{pmatrix} 7.7312 & 4.0573 \\ 4.0573 & 6.2072 \end{pmatrix} = \begin{pmatrix} 0.8173^2 & 0.3506 \\ 0.3506 & 0.7323^2 \end{pmatrix}$$

□

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:

```
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, β_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. estimater for de systematiske effekter udskrevet.

Sætningen `RANDOM INTERCEPT ald /TYPE=UN SUB=dreng ;` angiver, at vi vil modellere såvel interceptparameteren som koefficienten til alder som tilfældige effekter. Optionen `TYPE=UN` specificerer at dispersionsmatricen svarende til hele observationssættet

$$\mathbf{D}[\mathbf{Y}] = \sigma^2[\mathbf{I}_n + \mathbf{X}\Gamma\mathbf{X}^T] \otimes \mathbf{I}_k$$

ikke er en diagonalmatrix, og endelig angiver optionen `SUB=dreng`, at observationer svarende til forskellige værdier af `dreng` er uafhængige, dvs. at $\mathbf{D}[\mathbf{Y}]$ er en blokdiagonalmatrix.

Proceduren giver anledning til følgende udskrift

```
The MIXED Procedure

Class Level Information

Class      Levels  Values

DRENG           5  A B C D E
```

til kontrol af de indlæste specifikationer

Endvidere udskrives iterationsforløbet

```
The MIXED Procedure

ML Estimation Iteration History

Iteration  Evaluations      Objective      Criterion

        0            1    14.14530368
        1            1    -4.78207975     0.00000000

Convergence criteria met.
```

Da data er balancerede behøves kun én iteration.

Derefter udskrives estimaterne for σ^2 og $\mathbf{\Gamma}$

```
Covariance Parameter Estimates (MLE)

Cov Parm    Subject       Estimate

UN(1,1)     DRENG       0.53005000
UN(2,1)     DRENG       0.28044000
UN(2,2)     DRENG       0.41526400
Residual                0.08640000
```

Estimaterne står i søjlen betegnet `estimate`.

Estimatet $\widehat{\sigma}^2$ står i linien `Residual`, og elementerne i $\widehat{\sigma^2\mathbf{\Gamma}}$ står i rubrikkerne svarende til hhv. `UN(1,1)`, `UN(2,1)` og `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:

Asymptotic Covariance Matrix of Estimates

Cov Parm	Row	COVP1	COVP2	COVP3	COVP4
UN(1,1)	1	0.12182040	0.06188189	0.03175724	-0.00037325
UN(2,1)	2	0.06188189	0.06917141	0.05433626	0.00000000
UN(2,2)	3	0.03175724	0.05433626	0.09480666	-0.00119439
Residual	4	-0.00037325	0.00000000	-0.00119439	0.00149299

Elementerne i den udskrevne matrix identificeres som elementerne i matricen

$$\mathbf{D}\begin{bmatrix} \widehat{\sigma^2\gamma_{11}} \\ \widehat{\sigma^2\gamma_{12}} \\ \widehat{\sigma^2\gamma_{22}} \\ \widehat{\sigma}^2 \end{bmatrix}$$

Vi ser, at ingen af komponenterne i $\sigma^2\mathbf{\Gamma}$ synes at afvige signifikant fra nul (ingen af dem ligger mindre end 2 spredninger væk fra 0; således er estimatet for $\sigma^2\gamma_{1,1} = 0.53005$ med den estimerede spredning $\sqrt{0.12182040} = 0.349$).

Endelig udskrives et resume af tilpasningen

Model Fitting Information for RAMUS

Description	Value
Observations	20.0000
Log Likelihood	-15.9877
Akaike's Information Criterion	-19.9877
Schwarz's Bayesian Criterion	-21.9792
-2 Log Likelihood	31.9755

Null Model LRT Chi-Square	18.9274
Null Model LRT DF	3.0000
Null Model LRT P-Value	0.0003

samt de ønskede estimater 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

Source	NDF	DDF	Type III F	Pr > F
ALD	1	4	20.93	0.0102

hvor `Std Error` angiver den estimerede spredning for den pågældende koefficient, 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 ;
```

ville vi få estimaterne

Covariance Parameter Estimates (REML)

Cov Parm	Subject	Estimate
UN(1,1)	DRENG	0.66796250

```
UN(2,1)     DRENG       0.35055000
UN(2,2)     DRENG       0.53636000
Residual                0.08640000
```

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:

MIXED REML

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

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
ALD	1	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 formelapparatet 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 $\mathbf{X}_i$ en kendt $n_i \times p$ dimensional matrix således at $(\mathbf{X}_i{}^T\mathbf{X}_i)^{-1}$ eksisterer. Antag, at

$$Y_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,\ i = 1, 2, \ldots, k$$

hvor

$$\boldsymbol{\epsilon}_i \in \mathrm{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$$

og $\boldsymbol{\epsilon}_i$ og $\boldsymbol{\epsilon}_j$ er indbyrdes uafhængige for $i \neq j$.

Da er maksimaliseringsestimatoren for $\boldsymbol{\beta}_i$ givet ved

$$\widehat{\boldsymbol{\beta}}_i = \mathbf{P}_i\mathbf{y}_i\,, \tag{7.2.1}$$

hvor parameterprojektionsmatricen for den i'te gruppe er er givet ved

$$\mathbf{P}_i = (\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T \tag{7.2.2}$$

og den centrale estimator for σ^2 er

$$\hat{\sigma}^2 = sak_{1\cdot}/(N - kp) \tag{7.2.3}$$

hvor

$$sak_{1\cdot} = \sum_{i=1}^{k} sak_{1,i} \tag{7.2.4}$$

med

$$sak_{1,i} = (\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i)^T(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i)$$

og $N = \sum_i n_i$

Der gælder

$$\widehat{\boldsymbol{\beta}}_i \in \mathrm{N}_p(\mathbf{0}, \sigma^2(\mathbf{X}_i^T\mathbf{X}_i)^{-1});\ \ SAK_{1\cdot} \in \sigma^2\chi^2(N - kp)$$

og $\widehat{\boldsymbol{\beta}}_i$ og $SAK_{1\cdot}$ er indbyrdes uafhængige.

Bevis:

Beviset følger i analogi med beviset for sætning 7.1.1 ved at bemærke, at maksimaliseringsestimatoren fås ved at minimerere kvadratafvigelsessummen

$$l(\boldsymbol{\beta}; \mathbf{y}_1, \ldots, \mathbf{y}_k) = -\sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)^T (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)$$

Uafhængigheden følger af spaltningssætningen.

□

Sætning 7.2.2 *Test for fælles regressionskoefficienter*

Kvotienttestet for hypotesen

$$H_1: \quad \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_k \ (= \boldsymbol{\beta}_0)\,, \tag{7.2.5}$$

imod alternativet

$$H_1^c: \ \boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j \ \text{ for mindst eet sæt } (i,j) \text{ med } i \neq j$$

har teststørrelsen

$$Z = \frac{\mathrm{SAK}_2/[p(k-1)]}{\mathrm{SAK}_1./(N-k)} \tag{7.2.6}$$

hvor

$$\mathrm{SAK}_2 = \sum_{i=1}^{k} (\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)^T \mathbf{X}_i^T \mathbf{X}_i (\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0) \tag{7.2.7}$$

med

$$\widehat{\boldsymbol{\beta}}_0 = \frac{1}{k} \sum_{i=1}^{k} \mathbf{P}_i \mathbf{y}_i \tag{7.2.8}$$

Under H_1 følger Z en $\mathrm{F}(p(k-1), N-kp)$-fordeling. Testet forkaster for store værdier af z.

Bevis:

Sætningen vises analogt med sætning 7.1.2 ved at bemærke, at der gælder opspaltningen

$$\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0)^T(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_0) = sak_{1\cdot} + sak_2$$

og benytte spaltningssætningen.

□

Bemærkning 1 *Estimation og test ved observationer med vilkårlig, kendt dispersionsmatrix*

Såfremt

$$\boldsymbol{\epsilon}_i \in \mathrm{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{V}_i)$$

fås maksimaliseringsestimatoren ved at minimerere kvadratafvigelsessummen

$$l(\boldsymbol{\beta}; \mathbf{y}_1, \ldots, \mathbf{y}_k) = -\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)$$

Man finder estimatet

$$\widehat{\boldsymbol{\beta}}_i = \mathbf{P}_{wi}\mathbf{y}_i \tag{7.2.9}$$

med parameterprojektionsmatricen

$$\mathbf{P}_{wi} = (\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{V}_i^{-1}\,. \tag{7.2.10}$$

Den centrale estimator for σ^2 er givet ved (7.2.3) med

$$sak_{1,i} = (\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i)$$

Der gælder

$$\widehat{\boldsymbol{\beta}}_i \in \mathrm{N}_p(\mathbf{0}, \sigma^2(\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1});\quad SAK_{1\cdot} \in \sigma^2\chi^2(N - kp)$$

Kvotienttestet for hypotesen (7.2.5) har teststørrelsen (7.2.6), hvor

$$\mathrm{SAK}_2 = \sum_{i=1}^{k}(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)^T\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0) \tag{7.2.11}$$

med

$$\widehat{\beta}_0 = \frac{1}{k}\,(\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X})_i^{-1}\mathbf{X}_i^T\sum_{i=1}^{k}\mathbf{y}_i \tag{7.2.12}$$

svarende til opspaltningen

$$\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{X}_i\widehat{\beta}_0)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\beta}_0) = sak_1. + sak_2$$

□

7.2.2 Den tilfældige model

I lighed med det foregående afsnit kan man vælge at udbygge modellen med en tilfældig komponent ved antagelsen $\boldsymbol{\beta}_i \in \mathrm{N}_p(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Gamma})$. For denne model gælder

Sætning 7.2.3 ***Marginal fordeling af observationer og estimater under regressionsmodel med tilfældige koeffficienter***

Lad

$$Y_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,\ i = 1, 2, \ldots, k$$

hvor

$$\boldsymbol{\beta}_i \in \mathrm{N}_p(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Gamma})$$

og hvor $\boldsymbol{\beta}_i$, $\boldsymbol{\beta}_j$ er indbyrdes uafhængige for $i \neq j$, og $\boldsymbol{\epsilon}_i$ og $\boldsymbol{\epsilon}_j$ er indbyrdes uafhængige for $i \neq j$ og endvidere $\boldsymbol{\beta}_i$ og $\boldsymbol{\epsilon}_j$ er uafhængige.

Da er den marginale fordeling af Y_i en

$$\mathrm{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}_0, \sigma^2, [\mathbf{I}_{n_i} + \mathbf{X}_i\boldsymbol{\Gamma}\mathbf{X}_i^T]) - \text{fordeling},$$

og den marginale fordeling af estimatet $\widehat{\boldsymbol{\beta}}_i$ (7.2.1) for regressionskoefficienterne svarende til den i'te gruppe er

$$\widehat{\boldsymbol{\beta}}_i \in \mathrm{N}_p(\boldsymbol{\beta}_0, \sigma^2[(\mathbf{X}_i^T\mathbf{X}_i)^{-1} + \boldsymbol{\Lambda}]) \tag{7.2.13}$$

Bevis:
Beviset forløber i analogi med beviset for sætning 7.1.3.

□

Sætning 7.2.4 *Maksimaliseringsestimation for ubalanceret regressionsmodel med tilfældige koeffficienter*

Lad modellen være som i sætning 7.2.3.

Da fås maksimaliseringsestimatorerne for parametrene $\boldsymbol{\beta}_0$, σ^2 og $\boldsymbol{\Gamma}$ ved at maksimere

$$l(\boldsymbol{\beta}_0, \sigma^2, \boldsymbol{\Gamma};\, y_1, \ldots, y_k) = -k \times n \ln(\sigma^2) - \sum_{i=1}^{k} \ln[\det(\mathbf{D}_i(\boldsymbol{\Gamma}))]$$

$$- \frac{1}{\sigma^2} \sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_0)^T (\mathbf{D}_i(\boldsymbol{\Gamma}))^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}_0) \tag{7.2.14}$$

hvor

$$\mathbf{D}_i(\boldsymbol{\Gamma}) = \mathbf{X}_i\boldsymbol{\Gamma}\mathbf{X}_i^T + \mathbf{I}_{n_i} \tag{7.2.15}$$

med hensyn til $\sigma^2 > 0$, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ og $\boldsymbol{\Gamma}$ positiv semidefinit.
Bevis:
Sætningen vises ved at bemærke, at den marginale fordeling af Y_i er en $\mathrm{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}_0, \sigma^2, [\mathbf{I}_{n_i} + \mathbf{X}_i\boldsymbol{\Gamma}\mathbf{X}_i^T])$- fordeling, og at (7.2.14) netop er logaritmen til likelihoodfunktionen. □

Bemærkning 1 *Likelihoodligningerne ved maksimum i et indre punkt*

Såfremt maksimumværdien findes i et indre punkt, tilfredsstiller estimaterne udtrykkene

$$\widehat{\boldsymbol{\beta}}_0 = \Big[\sum_{i=1}^{k} \mathbf{X}_i^T (\mathbf{D}_i(\boldsymbol{\Gamma}))^{-1} \mathbf{X}_i\Big]^{-1} \sum_{i=1}^{k} \mathbf{X}_i^T (\mathbf{D}_i(\boldsymbol{\Gamma}))^{-1} \mathbf{y}_i \quad (7.2.16)$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}}_0)^T (\mathbf{D}_i(\hat{\boldsymbol{\Gamma}}))^{-1} (\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}}_0) \quad (7.2.17)$$

hvor $\widehat{\boldsymbol{\Gamma}}$ er bestemt som den positive semidefinitte $p \times p$-matrix, der maksimerer udtrykket

$$l(\boldsymbol{\Gamma}) = -N \times \ln(sak^*(\boldsymbol{\Gamma})) - \sum_{i=1}^{k} \ln[\det(\mathbf{D}_i(\boldsymbol{\Gamma}))] \quad (7.2.18)$$

med

$$sak^*(\boldsymbol{\Gamma}) = \sum_{i=1}^{k} \mathbf{y}_i^T (\mathbf{D}_i(\boldsymbol{\Gamma}))^{-1} \mathbf{y}_i$$
$$- \Big[\sum_{i=1}^{k} \mathbf{X}_i^T (\mathbf{D}_i(\boldsymbol{\Gamma}))^{-1} \mathbf{y}_i\Big]^T \Big[\sum_{i=1}^{k} \mathbf{X}_i^T (\mathbf{D}_i(\boldsymbol{\Gamma}))^{-1} \mathbf{X}_i\Big]^{-1} \Big[\sum_{i=1}^{k} \mathbf{X}_i^T (\mathbf{D}_i(\boldsymbol{\Gamma}))^{-1} \mathbf{y}_i\Big]$$

□

Bemærkning 2 ***Estimation ved observationer med vilkårlig, kendt dispersionsmatrix***

Sætningen og bemærkningen gælder også i den generelle situation, hvor dispersionsmatricen for observationsfejlene er $\sigma^2 \mathbf{V}_i$. Man skal blot erstatte udtrykket (7.2.15) for $\mathbf{D}_i(\boldsymbol{\Gamma})$ med

$$\mathbf{D}_i(\boldsymbol{\Gamma}) = \mathbf{X}_i \boldsymbol{\Gamma} \mathbf{X}_i^T + \mathbf{V}_i \,, \quad (7.2.19)$$

idet den marginale fordeling af Y_i i dette tilfælde er en

$$\mathrm{N}_{n_i}\big(\mathbf{X}_i \boldsymbol{\beta}_0, \sigma^2, [\mathbf{V}_i + \mathbf{X}_i \boldsymbol{\Gamma} \mathbf{X}_i^T]\big) - \text{fordeling}$$

□

Sætning 7.2.5 *Momentestimation for ubalanceret regressionsmodel med tilfældige koeffficienter*

Under antagelserne fra sætning 7.2.3 fås momentestimaterne for β_0, σ^2 og $\mathbf{\Gamma}$ som

$$\widetilde{\beta}_0 = \frac{1}{k}\sum_{i=1}^{k}\widehat{\beta}_i \tag{7.2.20}$$

$$\tilde{\sigma}^2 = \frac{1}{N-kp}\,\mathrm{sak}_1. \tag{7.2.21}$$

$$\widetilde{\mathbf{\Gamma}} = \frac{\mathbf{sak}_\beta}{(k-1)\tilde{\sigma}^2} - \frac{1}{k}\left(\sum_{i=1}^{k}\mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \tag{7.2.22}$$

hvor $\widehat{\beta}_i$ og sak_1. er givet ved (7.2.1) og (7.2.4), og hvor

$$\mathbf{sak}_\beta = \sum_{i=1}^{k}(\widehat{\beta}_i - \widehat{\beta}_0)(\widehat{\beta}_i - \widehat{\beta}_0)^T$$

Bevis:
Sætningen vises ved at bemærke, at $\widetilde{\beta}_0$ er gennemsnittet, og $\mathbf{sak}_\beta/(k-1)$ er den empiriske dispersionsmatrix for $\widehat{\beta}_i$, hvor $\widehat{\beta}_i$ er indbyrdes uafhængige med

$$\mathsf{E}\,[\widehat{\beta}] = \beta_0$$

og

$$\mathbf{D}\,[\widehat{\beta}_i] = \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{\Gamma}])$$

ifølge sætning 7.2.3.

Vi har derfor

$$\mathsf{E}\,[\widetilde{\beta}_0] = \beta_0, \quad \text{og}$$

$$\mathsf{E}\,[\mathbf{sak}_\beta] = \sum_{i=1}^{k}\left(1-\frac{1}{k}\right)\mathbf{D}\,[\widehat{\beta}_i] = \sigma^2\left(1-\frac{1}{k}\right)\sum_{i=1}^{k}[(\mathbf{X}_i^T\mathbf{X}_i)^{-1} + \mathbf{\Gamma}]$$

□

Bemærkning 1 *Sammenligning mellem maksimaliseringsestimater og momentestimater*

Sammenligner vi momentestimatet (7.2.20) for β_0 med maksimaliseringsestimaterne (7.2.16), ser vi, at i modsætning til momentestimatet inddrager maksimaliseringsestimatet såvel kendskabet til forskellene mellem modelmatricerne for de enkelte grupper samt kendskabet til dispersionsmatricen $\mathbf{\Gamma}$ ved den indbyrdes vægtning af de enkelte komponenter af $\widehat{\beta}_i$.

Man kunne forbedre skønnet over β_0 ved at vægte $\widehat{\beta}_i$ med designpræcisionen ved benyttelse af

$$\breve{\beta}_0 = \left(\sum_{i=1}^{k} \mathbf{X}_i^T \mathbf{X}_i\right)^{-1} \sum_{i=1}^{k} \mathbf{X}_i^T y_i = \left(\sum_{i=1}^{k} \mathbf{X}_i^T \mathbf{X}_i\right)^{-1} \sum_{i=1}^{k} \mathbf{X}_i^T \mathbf{X}_i \widehat{\beta}_i \ ,$$

men anvendelsen af dette estimat vanskeliggør beskrivelsen af forholdene for $\mathbf{sak}_\beta$-matricen.

□

Bemærkning 2 *Momentestimation ved observationer med vilkårlig, kendt dispersionsmatrix*

I det generelle tilfælde, hvor dispersionsmatricen for observationsfejlene er $\sigma^2 \mathbf{V}_i$, kan man ligeledes benytte estimaterne (7.2.20) og (7.2.21) for β_0 og σ^2 med $\widehat{\beta}_i$ bestemt ved (7.2.9).

Momentestimatet for $\mathbf{\Gamma}$ bestemmes i det generelle tilfælde ved

$$\widetilde{\mathbf{\Gamma}} = \frac{\mathbf{sak}_\beta}{(k-1)\tilde{\sigma}^2} - \frac{1}{k}\left(\sum_{i=1}^{k} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \tag{7.2.23}$$

I dette tilfælde gælder, at $\widehat{\beta}_i$ er indbyrdes uafhængige med

$$\mathsf{E}\,[\widehat{\beta}] = \beta_0$$

og

$$\mathbf{D}\,[\widehat{\beta}_i] = \sigma^2[(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} + \mathbf{\Gamma}])$$

Et forbedret skøn over β_0 fås i dette tilfælde som

$$\begin{aligned}\breve{\beta}_0 &= \left(\sum_{i=1}^{k} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{k} \mathbf{X}_i^T \mathbf{V}_i^{-1} y_i \\ &= \left(\sum_{i=1}^{k} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{k} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \widehat{\beta}_i \quad ,\end{aligned}$$

men skønnet vanskeliggør beskrivelsen af forholdene for sak-matricen.

□

7.3 Tidsrækkemodeller

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 $1, 2, ..., n$ angiver ækvidistante punkter på tidsaksen.

Ofte vil sådanne observationer være autokorrelerede, dvs. $\mathsf{COV}[Y_t, Y_{t-k}] \neq 0$ for nogle $k \geq 0$.

Vi vil her betragte den simple autoregressive model af første orden. Modellen er givet ved at fordelingen af Y_t kun afhænger af værdien af Y_{t-1} :

$$Y_t | \{Y_{t-1} = y_{t-1}\} \in \mathrm{N}(\beta y_{t-1}, \sigma^2) \tag{7.3.1}$$

Ofte udtrykkes modellen rekursivt

$$Y_t = \beta Y_{t-1} + \epsilon_t \tag{7.3.2}$$

hvor $\epsilon_1, \epsilon_2, \ldots, \epsilon_t$ er uafhængige $\mathrm{N}(0, \sigma^2)$-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 β og σ 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

$$-\frac{\hat{\beta}}{1-\hat{\beta}^2}\hat{\sigma}^2 + \sum_{j=0}^{n-1} y_j y_{j+1} - \hat{\beta}\sum_{j=0}^{n-1} y_j^2 = 0 \tag{7.3.3}$$

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 β,

$$\tilde{\beta} = (\mathbf{y}_{-1}^T\mathbf{y}_{-1})^{-1}\mathbf{y}_{-1}^T\mathbf{y}_{-1}^T\mathbf{y}y = \sum_{j=0}^{n-1} y_j y_{j+1} \Big/ \sum_{j=0}^{n-1} y_j^2$$

og den tilsvarende minimale værdi af den gennemsnitlige residualkvadratafvigelse

$$\tilde{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \tilde{\beta}\mathbf{y}_{-1})^T(\mathbf{y} - \tilde{\beta}\mathbf{y}_{-1})$$

som skøn over σ^2.

Vi ser, at $\tilde{\beta}$ 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 $y^i = (y_1^i, y_2^i, \ldots, y_n^i)^T$; $i = 1, 2, \ldots, k$, kan vi modellere variationen imellem disse tidsrækker ved en tilfældig model:

Vi antager

$$Y_t^i = \beta_i Y_{t-1}^i + \epsilon_t^i, t = 1, 2, \ldots, n;\ i = 1, 2, \ldots, k,$$

hvor $\beta_i \in \mathrm{N}(\beta_0, \sigma_0^2)$ er indbyrdes uafhængige, og $\epsilon_t^i \in \mathrm{N}(0, \sigma^2)$ ligeledes er indbyrdes uafhængige og uafhængige af β_i.

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

$$\tilde{\beta}_i = [(\mathbf{y}_{-1}^i)^T \mathbf{y}_{-1}]^{-1}(\mathbf{y}_{-1}^i)^T \mathbf{y}^i = \sum_{j=0}^{n-1} y_j^i y_{j+1}^i \Big/ \sum_{j=0}^{n-1}(y_j^i)^2$$

og det tilsvarende estimat for residualvariansen

$$\tilde{\sigma}^2 = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}(\mathbf{y}^i - \mathbf{y}_{-1}^i\tilde{\beta}_i)^T(\mathbf{y}^i - \mathbf{y}_{-1}^i\tilde{\beta}_i)$$

Bevis:
Overspringes

□

7.3.2 Flerdimensionale tidsrækkemodeller.

Vi vil ganske kort angive resultaterne for flerdimensionale tidsrækkemodeller.

Lad $\mathbf{y}_t$ angive en p-dimensional søjlevektor af observationer, og lad $\mathbf{z}_t$ angive en (kendt) q-dimensional vektor af kontrolvariable (regressionsvariable) til tidspunktet t, $t = 0, 1, 2, \ldots, n$.

Antag at der gælder følgende model

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{C}\mathbf{z}_t + \boldsymbol{\epsilon}_t, \quad t = 1, 2, ..., n \tag{7.3.4}$$

hvor $\boldsymbol{\epsilon}_t$ angiver en p-dimensional stokastisk vektor med forventningsværdi nul, og $\mathbf{A}$ og $\mathbf{C}$ er ukendte parametermatricer af dimensioner henholdsvis $p \times p$ og $p \times q$.

Vi kan opskrive (7.3.4) på matrixform

$$\mathbf{Y}^T = \mathbf{A}\mathbf{Y}_{-1}^T + \mathbf{C}\mathbf{Z}^T + \boldsymbol{\epsilon}^T \tag{7.3.5}$$

hvor de $n \times p$-dimensionale matricer $\mathbf{Y}, \mathbf{Y}_{-1}$ og $\boldsymbol{\epsilon}$ er givet ved

$$\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)^T; \ \mathbf{Y}_{-1} = (\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{n-1})^T$$

og $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_n)^T$, og den $n \times q$-dimensionale matrix $\mathbf{Z}$ tilsvarende er givet ved $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)^T$.

Transponerer vi nu (7.3.5) fås

$$\mathbf{Y} = \mathbf{M}\mathbf{B} + \boldsymbol{\epsilon} \tag{7.3.6}$$

hvor den $n \times (p + q)$ dimensionale matrix $\mathbf{M}$ er givet ved

$$\mathbf{M} = (\mathbf{Y}_{-1}, \mathbf{Z})$$

og den $(p + q) \times p$-dimensionale koefficientmatrix $\mathbf{B} = (\mathbf{A}, \mathbf{C})^T$.

Vi ordner nu observationerne $\mathbf{Y}$ i en søjlevektor $\mathbf{y}$

$$\mathbf{y} = \mathrm{vec}(\mathbf{Y}) = \begin{pmatrix} y_{11} \\ \ldots \\ y_{n1} \\ \ldots \\ y_{np} \end{pmatrix}$$

og tilsvarende sættes $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{A})$ og $\boldsymbol{\epsilon} = \mathrm{vec}(\boldsymbol{\epsilon})$. Endelig sætter vi $\mathbf{X} = \mathbf{I}_p \otimes \mathbf{M}$, hvor $\mathbf{I}_p$ angiver den $p \times p$-dimensionale enhedsmatrix.

Antages nu, at $\boldsymbol{\epsilon} \in \mathrm{N}_{np \times np}(0, \boldsymbol{\Sigma})$, har vi den generaliserede mindste kvadraters estimator for $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} \tag{7.3.7}$$

Den tilfældige model

Såfremt vi har k tidsrækker fra k forskellige grupper (dvs med forskellige værdier af β), kan vi modellere variationen imellem disse tidsrækker ved en tilfældig model.

Vi betragter derfor modellen

$$\mathbf{Y}^i = \mathbf{M}_i\mathbf{B}_i + \boldsymbol{\epsilon}^i \tag{7.3.8}$$

hvor den $n_i \times (p+q)$- dimensionale matrix $\mathbf{M}_i$ er givet ved

$$\mathbf{M}_i = (\mathbf{Y}^i_{-1}, \mathbf{Z}^i)$$

og den $(p+q) \times p$-dimensionale koefficientmatrix $\mathbf{B}_i = (\mathbf{A}_i, \mathbf{C}_i)^T$.

Vi sætter som før

$$\mathbf{y}^i = \mathrm{vec}(\mathbf{Y}^i);\ \boldsymbol{\beta}_i = \mathrm{vec}(\mathbf{B}_i) \quad \text{og} \quad \boldsymbol{\epsilon}^i = \mathrm{vec}(\boldsymbol{\epsilon}^i)$$

Endelig sætter vi $\mathbf{X}_i = \mathbf{I}_p \otimes \mathbf{M}_i$

Vi antager da,

$$\boldsymbol{\epsilon}^i \in \mathrm{N}_{n_ip \times n_ip}(0, \boldsymbol{\Sigma}_i) \quad \text{og} \quad \boldsymbol{\beta}_i \in \mathrm{N}_{(p+q)\times q}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0) \tag{7.3.9}$$

Vi har da estimaterne under denne model:

a) $\boldsymbol{\Sigma}_i$ kendt :

$$\widehat{\boldsymbol{\beta}}_0 = \frac{1}{k}\sum_{i=1}^{k}\widehat{\boldsymbol{\beta}}_i \quad \text{med} \quad \widehat{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\beta}}_i = (\mathbf{X}_i^T\boldsymbol{\Sigma}_i\mathbf{X}_i)^{-1}\mathbf{X}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{y}^i \tag{7.3.10}$$

og

$$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{k}\sum_{i=1}^{k}(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_0)^T \tag{7.3.11}$$

b) $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{V}_i$, hvor $\mathbf{V}_i$ er en kendt symmetrisk, positiv definit $n_ip \times n_ip$-dimensional matrix:

$$\widehat{\boldsymbol{\beta}}_0 = \frac{1}{k}\sum_{i=1}^{k}\widehat{\boldsymbol{\beta}}_i \quad \text{med} \quad \widehat{\boldsymbol{\beta}}_i = \widehat{\boldsymbol{\beta}}_i = (\mathbf{X}_i^T\mathbf{V}_i\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{y}^i \tag{7.3.12}$$

og

$$\hat{\sigma}^2 = \frac{1}{pk}\sum_{i=1}^{k}\frac{1}{n_i}(\mathbf{y}^i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i)^T\mathbf{V}_i^{-1}(\mathbf{y}^i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}_i) \tag{7.3.13}$$

og $\widehat{\boldsymbol{\Sigma}}_0$ bestemmes ved (7.3.11) med $\widehat{\boldsymbol{\beta}}_0$ og $\widehat{\boldsymbol{\beta}}_i$ givet ved (7.3.12).

c) $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} \otimes \mathbf{V}_i$, hvor $\boldsymbol{\Sigma}$ er ukendt symmetrisk, positiv definit $p \times p$-matrix, og $\mathbf{V}_i$ er en kendt symmetrisk, positiv definit $n_i \times n_i$-dimensional matrix:

$$\widehat{\mathbf{B}}_0 = \frac{1}{k}\sum_{i=1}^{k}\widehat{\mathbf{B}}_i \quad \text{med} \quad \widehat{\mathbf{B}}_i = (\mathbf{M}_i^T\mathbf{V}_i^{-1}\mathbf{M}_i)^{-1}\mathbf{M}_i^T\mathbf{V}_i^{-1}\mathbf{Y}^i \tag{7.3.14}$$

med $\widehat{\boldsymbol{\beta}}_i = \mathrm{vec}(\widehat{\mathbf{B}}_i)$.

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}(\mathbf{Y}^i - \mathbf{M}_i\widehat{\mathbf{B}}_i)^T\mathbf{V}_i^{-1}(\mathbf{Y}^i - \mathbf{M}_i\widehat{\mathbf{B}}_i) \tag{7.3.15}$$

og $\widehat{\boldsymbol{\Sigma}}_0$ bestemt ved (7.3.11) med $\widehat{\boldsymbol{\beta}}_0$ og $\widehat{\boldsymbol{\beta}}_i$ 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 $\mathbf{y}$, modelleres den systematiske komponent ved

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

hvor $\mathbf{X}$ angiver modelmatricen for den systematiske effekt, $\boldsymbol{\beta}$ er vektoren af effekter, og $\boldsymbol{\epsilon}$ angiver en vektor af tilfældige modelafvigelser.

Bidraget fra den tilfældige komponent modelleres ved en modelmatrix $\mathbf{Z}$, og en vektor $\mathbf{v}$ af tilfældige effekter. Den blandede model udtrykkes da som

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{v} + \epsilon$$

Det antages, at $\mathbf{v}$ og ϵ er uafhængige og har forventningsværdien $\mathbf{0}$. Dispersionsmatricerne for $\mathbf{v}$ og ϵ betegnes hhv. $\mathbf{G}$ og $\mathbf{R}$, dvs

$$\mathsf{E}[\mathbf{v}] = \mathbf{0}\,, \qquad \mathbf{D}[\mathbf{v}] = \mathbf{G}$$

og

$$\mathsf{E}[\epsilon] = 0\,, \qquad \mathbf{D}[\epsilon] = \mathbf{R}$$

Under disse antagelser har man

$$\mathsf{E}[\mathbf{y}] = \mathbf{X}\beta\,, \qquad \mathbf{D}[\mathbf{y}] = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R} \tag{7.4.1}$$

Denne generelle formulering giver mulighed for at vælge mellem at placere tilfældige effekter i $\mathbf{v}$ eller i ϵ.

Formuleringen giver endvidere mulighed for at påtrykke en bestemt struktur af dispersionsmatricerne $\mathbf{G}$ og $\mathbf{R}$, som f.eks. $\mathbf{R} = \sigma^2\mathbf{I}$

Sædvanligvis vil man placere effekter svarende til en lille dimensionalitet (få søjler i $\mathbf{Z}$) i $\mathbf{v}$, sådan at $\mathbf{G}$ let kan inverteres, og placere resten i ϵ.

Den generelle teori for blandede modeller er behandlet af Searle et al. (1992).

7.5 Referencer

J. Bondeson: *Random coefficient regression models in biostatistics*, Inst. för Matematisk Statistik, Lunds Universitet, 1989

R.C. Elston og J.E.Grizzle : Estimation of time-response curves and their confidence bands. *Biometrics* **18** (1962), pp 148-159.

R.I. Jennrich and M.D. Schluchter (1986): Unbalanced Repeated Measures Models with Structured Covariance Matrices. *Biometrics* **42**, pp 805-820

A.Linder (1960): *Statistische Methoden*, 3.udg., Birkhäuser Verlag)

S.R.Searle, G. Casella and C.E. McCullogh (1992): *Variance Components*, John Wiley & Sons, New York

P. Thyregod: Comments on S. Johansens paper: Regression. *Scand.Journ. Statist.* **10**, 1983 pp. 190- 191.

Afsnit 8

Aposteriorifordelinger

fil: apost.tex 1998-04-21

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 middelværdiparameteren μ. Tabel 8.2 på side 710 giver tilsvarende en oversigt over opdateringen af momentparametrene og

I Eksempel 8.3.2 er princippene 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 Ω, og lad B være en hændelse.

Der gælder da

$$\mathsf{P}\,[A_i|B] = \frac{\mathsf{P}\,[A_i \cap B]}{\sum_j \mathsf{P}\,[B|A_j]\mathsf{P}\,[A_j]} \tag{8.1.1}$$

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

$$h(y|x) = \frac{g(x|y)w(y)}{k(x)} \tag{8.1.2}$$

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

$$\{f(x_1, x_2, \ldots, x_n|\theta)\}_{\theta\in\Theta} \tag{8.2.1}$$

hvor $\Theta \subset \mathbb{R}^k$.

Antag at fordelingen af $X_1, X_2, \ldots, X_n$ for en fastholdt værdi af θ har tætheden

$$f(x_1, x_2, \ldots, x_n|\theta)$$

og antag yderligere, at parameteren θ kan opfattes som en stokastisk variabel med tætheden

$$w(\theta)$$

Som tidligere nævnt kaldes fordelingen af θ med tætheden $w(\cdot)$ for strukturfordelingen, eller aprioriordelingen af θ.

Den betingede fordeling af θ givet $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ kaldes aposteriorifordelingen af θ 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 θ i den simultane fordeling af $X_1, X_2, \ldots, X_n$ og θ. Den simultane fordeling af $X_1, X_2, \ldots, X_n$ for fastholdt θ har da tætheden

$$f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} g(x_i|\theta) \tag{8.2.2}$$

hvor $g(\cdot|\theta)$ angiver tætheden for en enkelt observation.

Sådanne tilfælde kan f.eks. være situationer, hvor parameteren θ varierer i en population i overensstemmelse med apriorifordelingen $w(\theta)$, og hvor en måling, X af θ 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 θ varierer fra produktion til produktion i overensstemmelse med aprioritætheden $w(\theta)$.

I begge disse situationer kan det være af interesse - ikke blot at beskrive aposterierifordelingen af θ 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 θ, $X_1, X_2, \ldots, X_n$ og $X_1', X_2', \ldots, X_r'$ være sådan, at for givet θ 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 θ som de variable $X_1, X_2, \ldots, X_n$.

Antag at θ 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 θ 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 θ have tætheden $g^{(r)}(z'|\theta)$. Den prædiktive 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

$$g_1^{(r)}(z'|x_1, x_2, \ldots, x_n) = \int_\Theta g^{(r)}(z'|\theta) w_1(\theta|x_1, x_2, \ldots, x_n) d\nu\{\theta\} \tag{8.2.4}$$

hvor $w_1(\theta|x_1, x_2, \ldots, x_n)$ som før angiver aposterioritætheden af θ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$.

På tilsvarende måde defineres den prædiktive 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

$$g_1^{[r]}(y'|x_1, x_2, \ldots, x_n) = \int_\Theta g^{[r]}(y'|\theta) w_1(\theta|x_1, x_2, \ldots, x_n) \nu\{d\theta\} \tag{8.2.5}$$

hvor $g^{[r]}(y'|\theta)$ angiver tætheden i den den betingede fordeling af $Y' = \sum_{j=1}^{r} X_j'/r$ for fastholdt θ.

□

Den prædiktive 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 θ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. som strukturfordeling ("aprioriforde ling") af θ .

Tilsvarende er den prædiktive fordeling af Z' eller af $\overline{X'}$. den marginale fordeling af hhv Z' eller $\overline{X'}$. svarende til at man bruger aposteriorifordelingen af θ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. som strukturfordeling ("aprioriforde ling") af θ .

□

Den prædiktive 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 prædiktive fordeling.

Definition 8.2.3 *Præposteriorimiddelværdi- og varians*

Betragt situationen svarende til definition 8.2.1.

Betragt en funktion, $h(\theta)$ af θ.

Aposteriorimiddelværdien $\mathsf{E}\,[h(\theta)|x_1,\ldots,x_n]$ er da middelværdien af $h(\theta)$ i aposteriorifordelingen af θ efter observation af $x_1,\ldots,x_n$ og aposteriorivariansen

$$\mathsf{V}\,[h(\theta)|x_1,\ldots,x_n]$$

er variansen af $h(\theta)$ i denne fordeling.

Aposteriorimiddelværdien $\mathsf{E}\,[h(\theta)|x_1,\ldots,x_n]$ er en stokastisk variabel,

$$\mathsf{E}\,[h(\theta)|x_1,\ldots,x_n] = t(X_1,\ldots,X_n)$$

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
$\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{E}\,[h(\theta)|x_1,\ldots,x_n]]$.

Vi har altså

$$\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{E}\,[h(\theta)|x_1,\ldots,x_n]] = \mathsf{E}\,[t(X_1,\ldots,X_n)]\,,$$

hvor $t(X_1,\ldots,X_n)$ er den størrelse, der fremkommer ved at opfatte

$$\mathsf{E}\,[h(\theta)|x_1,\ldots,x_n]$$

som en stokastisk variabel.

Tilsvarende angiver præposteriorivariansen 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 $\mathsf{V}\,[h(\theta)|x_1,\ldots,x_n]$ som en stokastisk variabel. Præposteriorivariansen af $h(\theta)$ betegnes med

$$\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{V}\,[h(\theta)|x_1,\ldots,x_n]] = \mathsf{E}\,[u(X_1,\ldots,X_n)]$$

Præposteriorimiddelværdien og -variansen kan altså opfattes som gennemsnitsværdien af aposteriorimiddelværdien $\mathsf{E}\,[h(\theta)|x_1,\ldots,x_n]$ svarende til et stort antal forskellige stikprøver $x_1,\ldots,x_n$, og præposteriorivariansen angiver tilsvarende gennemsnitsværdien af aposteriorivariansen, $\mathsf{V}\,[h(\theta)|x_1,\ldots,x_n]$, 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 aprioriforde linger, 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 parallelle parametriseringer af aprioriforde lingerne.

Vi erindrer derfor om oversigtstabellerne, Tabel 6.1 og Tabel 6.2, der beskriver oversættelsen mellem momentparametrisering og parametriseringen ved α og β.

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 ϑ i den simultane fordeling af $X_1, X_2, \ldots, X_n$ og ϑ, og hvor fordelingen af X_i for givet ϑ har en tæthed af formen

$$g(x|\vartheta) = d(x)\exp\{[\vartheta x - \kappa(\vartheta)]/\sigma^2\} ,$$

og hvor fordelingen af ϑ 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 parametriseringer ved $\mu = \mathsf{E}[X|\vartheta]$ gælder da (jvf sætning 6.1.2) for den marginale fordeling af en enkelt måling, X_i

$$\begin{aligned} \mathsf{E}[X_i] &= m \\ \mathsf{V}[X_i] &= \mathsf{E}[V(\mu)](\sigma^2 + \gamma) \end{aligned}$$

hvor

$$\gamma = \frac{\mathsf{V}[\mu]}{\mathsf{E}[V(\mu)]}$$

For den marginale fordeling af summen, $Z = \sum_{i=1}^{n} X_i$ af n målinger, der er frembragt med samme - tilfældige værdi af μ - gælder da

$$\begin{aligned} \mathsf{E}[Z] &= n\,m \\ \mathsf{V}[Z] &= n\mathsf{E}[V(\mu)](\sigma^2 + n\gamma) , \end{aligned}$$

og endelig gælder for den marginale fordeling af gennemsnittet, $\overline{X}_+ = \sum_{i=1}^{n} X_i/n$ af n målinger, der er frembragt med samme - tilfældige værdi af μ - at

$$\begin{aligned} \mathsf{E}[\overline{X}_+] &= m \\ \mathsf{V}[\overline{X}_+] &= \mathsf{E}[V(\mu)]\left(\gamma + \frac{\sigma^2}{n}\right) \end{aligned}$$

Såfremt $\overline{X}_{1+}, \overline{X}_{2+}, \ldots, \overline{X}_{k+}$ angiver gennemsnittene i k uafhængige sæt af stikprøver (dvs med hver sin tilfældige værdi af ϑ), og

$$SAK_2 = \sum_{i=1}^{k} n_i(\overline{X}_{i+} - \overline{X}_{++})^2$$

angiver den sædvanlige kvadratafvigelssum, da gælder (jvf sætning 6.1.3), at

$$\mathsf{E}[SAK_2/(k-1)] = \frac{\mathsf{E}[V(\mu)]}{\sigma^2}(1 + n_0\gamma)$$

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

Vi minder om, at størrelserne γ og $\mathsf{E}[V(\mu)]$ svarende til de sædvanlige fordelinger er resumeret i tabel 6.1.

8.3.2 Generelle resultater vedrørende aposteriorifordelinger

Sætning 8.3.1 *Aposteriorifordeling svarende til konjugeret aprioriordeling for sædvanlige dispersionsmodeller*

Lad $X_1, X_2, \ldots, X_n$ være uafhængige, frembragt med samme ϑ og med tæthed

$$g(x|\vartheta) = d(x)\exp\{[\vartheta x - \kappa(\vartheta)]/\sigma^2\} , \tag{8.3.1}$$

og lad apriorifordelingen af ϑ være den konjugerede fordeling med tæthed

$$w(\vartheta; m, \gamma) = \frac{1}{C(m,\gamma)} \exp\{[\vartheta m - \kappa(\vartheta)]/\gamma\} \tag{8.3.2}$$

Da afhænger aposteriorifordelingen af ϑ efter observation af $X_1 = x_1, \ldots, X_n = x_n$ alene af n og $\bar{x}$.

Tætheden i aposteriorifordelingen af ϑ efter observation af $X_1 = x_1, \ldots, X_n = x_n$ er

$$w\Big(\vartheta \mid \sum X_i/n = \bar{x}\Big) = \frac{1}{C(m_1,\gamma_1)} \exp\{[\vartheta m_1 - \kappa(\vartheta)]/\gamma_1\} \tag{8.3.3}$$

med

$$m_1 = m_{\text{apost}} = \frac{m/\gamma + n\bar{x}/\sigma^2}{1/\gamma + n/\sigma^2} \tag{8.3.4}$$

$$\gamma_1 = \gamma_{\text{apost}} = \frac{1}{1/\gamma + n/\sigma^2} \tag{8.3.5}$$

Aposteriorifordelingen for ϑ er således af samme form som apriorifordelingen. Der er blot foretaget en opdatering af parametrene m og γ.

Bevis:
Følger umiddelbart ved opskrivning af udtrykket for aposteriorifordelingen, og benyttelse af at $Z = \sum X_i$ er sufficient for ϑ i den betingede fordeling af $X_1, \ldots, X_n$ for givet ϑ.

□

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 ϑ er af samme form som apriorifordelingen. Der er blot foretaget en opdatering af parametrene m og γ. Opdateringen er af formen:

$$\frac{1}{\gamma_{\text{apost}}} = \frac{1}{\gamma_{\text{apriori}}} + n/\sigma^2$$

$$\frac{m_{\text{apost}}}{\gamma_{\text{apost}}} = \frac{m_{\text{apriori}}}{\gamma_{\text{apriori}}} + n\,\bar{x}/\sigma^2 \qquad (8.3.6)$$

hvor $\bar{x} = z/n = \sum x_i/n$.

Idet vi erindrer (sætning 6.1.1), at parameteren $1/\gamma = \mathsf{E}\,[V(\mu)]/\mathsf{V}\,[\mu]$ er et udtryk for den relative præcision i fordelingen af ϑ i forhold til præcisionen i fordelingen af målestøjen, finder vi således, idet

$$\gamma_{\text{apost}} = \frac{\mathsf{V}\,[\mu|\bar{x}]}{\mathsf{E}\,[V(\mu)|\bar{x}]}\,, \qquad (8.3.7)$$

at

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 σ^2.

Groft sagt, kan man altså fortolke parameteren $1/\gamma$ i aposteriorifordelingen som den aprioripræcision, der svarer til en stikprøve af størrelsen σ^2/γ fra en given gruppe. For "$\gamma = \infty$", d.v.s. $1/\gamma = 0$ har man en "ikke-informativ" aposteriorifordeling.

Sammenholder vi (8.3.6) og (8.3.7) ser vi, at

forholdet mellem aposteriorivariansen $\mathsf{V}\,[\mu|\bar{x}]$ af μ og aposteriorimiddelværdien $\mathsf{E}\,[V(\mu)|\bar{x}]$ af variansfunktionen $V(\mu)$ afhænger alene af stikprøvestørrelsen n (og aprioriforholdet γ_{apriori}), men ikke af det aktuelle stikprøveresultat $\bar{x}$.

Vi kan således bestemme aposteriorivariansen $\mathsf{V}\,[\mu|\bar{x}]$ ud fra aposteriorimid-

delværdien $\mathsf{E}\,[V(\mu)|\bar{x}]$ af variansfunktionen $V(\mu)$ ved

$$\mathsf{V}\,[\mu|\bar{x}] = \gamma_{\text{apost}}\mathsf{E}\,[V(\mu)|\bar{x}] = \frac{\mathsf{E}\,[V(\mu)|\bar{x}]}{1/\gamma_{\text{apriori}} + n/\sigma^2} \tag{8.3.8}$$

Tabel 8.1 angiver aposteriorimiddelværdien $m_{\text{apost}} = \mathsf{E}\,[\mu|\bar{x}]$ af μ for de sædvanlige fordelinger. Tabellen angiver desuden aposteriorimiddelværdien $\mathsf{E}\,[V(\mu)|\bar{x}]$ af variansfunktionen. Aposteriorivariansen $\mathsf{V}\,[\mu|\bar{x}]$ for μ kan da bestemmes af relationen (8.3.8).

□

Bemærkning 2 : ***Aposteriorifordelingen af μ nærmer sig en etpunktsfordeling***

Vi ser, at

$$\gamma_{\text{apost}} \to 0 \qquad \text{for } n \to \infty$$

Endvidere finder vi, at

$$m_{\text{apost}} = \frac{m_{\text{apriori}}/\gamma_{\text{apriori}} + n\bar{x}/\sigma^2}{1/\gamma_{\text{apriori}} + n/\sigma^2} \to \bar{x} \qquad \text{for } n \to \infty$$

For store stikprøvestørrelser vil aposteriorifordelingen af μ således nærme sig en étpunktsfordeling 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 ϑ 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 ϑ som X'erne, og lad aposteriorifordelingen af ϑ 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 γ.

Lad $X_1', \ldots, X_r'$ angive et sæt nye observationer, der er uafhængige og identisk fordelte, og med samme fordeling som $X_1, \ldots, X_n$, dvs specielt frembragt med samme værdi af ϑ.

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 $\bar{X}$. Der er blot foretaget en opdatering af parametrene m og γ.

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:

$$\mathsf{E}\,[\overline{X'}_+|\bar{x}] = m_{\text{apost}} \tag{8.3.9}$$

$$\mathsf{V}\,[\overline{X'}_+|\bar{x}] = \mathsf{E}\,[V(\mu)|\bar{x}]\left(\gamma_{\text{apost}} + \frac{\sigma^2}{r}\right) \tag{8.3.10}$$

Tilsvarende har vi for den prædiktive fordeling af summen Z' af r fremtidige observationer

$$\mathsf{E}\,[Z'|\bar{x}] = r\, m_{\text{apost}} \tag{8.3.11}$$

$$\mathsf{V}\,[Z'|\bar{x}] = r\,\mathsf{E}\,[V(\mu)|\bar{x}](\sigma^2 + r\,\gamma_{\text{apost}}) \tag{8.3.12}$$

□

Bemærkning 4 : ***Forventningsværdien i den prædiktive fordeling er et vejet gennemsnit***

Forventningsværdien i den prædiktive fordeling af X' er

$$m_{\text{apost}} = \mathsf{E}\left[X' \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] = \frac{m/\gamma + n\bar{x}/\sigma^2}{1/\gamma + n/\sigma^2}$$

Der gælder

$$m_{\text{apost}} = \frac{\mathsf{E}\left[V(\mu)\right] \times m + n\mathsf{V}\left[\mu\right] \times \bar{x}/\sigma^2}{\mathsf{E}\left[V(\mu)\right] + n\mathsf{V}\left[\mu\right]/\sigma^2}$$

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 opspaltningen af den totale variation.

$$m_{\text{apost}} = w_n m_{\text{apriori}} + (1 - w_n)\bar{x} \tag{8.3.13}$$

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 γ i den prædiktive fordeling af X' er

$$\gamma_{\text{apost}} = \frac{\sigma^2}{n}(1 - w_n) \tag{8.3.14}$$

□

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:

$$m_{\text{apost}} = m + (1 - w_n)(\bar{x} - m) \tag{8.3.15}$$

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

$$1 - w_n = \frac{n\gamma}{\sigma^2 + n\gamma} = \frac{\mathsf{COV}[\overline{X}, X']}{\mathsf{V}[\overline{X}]}$$

som koefficienten til $\bar{x}$ i regressionen af X' på $\bar{x}$ i den simultane fordeling af $\overline{X}$ og X'.

Jo større værdi af $n\gamma$, d.v.s jo mindre aprioripræcision, $1/\gamma$, (eller jo større stikprøvestørrelse, n), desto større vægt får stikprøvekorrektionen, $(\bar{x} - m)$ til aprioriforventningsværdien, m, ved fastsættelsen af aposterioriforventningsværdien.

□

Bemærkning 6 : ***Præposteriorimiddelværdi 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 præposteriorimiddelværdien af μ er

$$\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{E}[\mu|\bar{x}]] = \mathsf{E}_{X_1,\ldots,X_n}[\mu] = m_{\text{apriori}}$$

Dette kunne også let verificeres ved at betragte udtrykket (8.3.13) for aposteriorimiddelværdien $\mathsf{E}[\mu|\bar{x}] = m_{\text{apost}}$.

Middelværdien (med hensyn til aprioriforde lingen) af udtrykket (8.3.13) er

$$\begin{aligned} \mathsf{E}_{X_1,\ldots,X_n}[\mathsf{E}[\mu|\bar{x}]] &= \mathsf{E}_{X_1,\ldots,X_n}[m_{\text{apost}}] \\ &= w_n m_{\text{apriori}} + (1 - w_n)\mathsf{E}_{X_1,\ldots,X_n}[\bar{x}] \quad (8.3.16) \\ &= w_n m_{\text{apriori}} + (1 - w_n) m_{\text{apriori}} = m_{\text{apriori}} \quad (8.3.17) \end{aligned}$$

Tilsvarende har vi for præposteriorimiddelværdien af $V(\mu)$,

$$\mathsf{E}_{X_1,\ldots,X_n}[V(\mu)|\bar{x}] = \mathsf{E}_{X_1,\ldots,X_n}[V(\mu)]$$

Endelig følger det af (8.3.8) i Bemærkning 1, at præposteriorivariansen af μ kan bestemmes ud fra præposteriorimiddelværdien af $V(\mu)$ ved

$$\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{V}[\mu|\bar{x}]] = \gamma_{\text{apost}}\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{E}[V(\mu)|\bar{x}]] ,$$

hvorfor vi altså har

$$\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{V}[\mu|\bar{x}]] = \mathsf{E}[V(\mu)] \frac{1}{1/\gamma_{\text{apriori}} + n/\sigma^2} , \tag{8.3.18}$$

hvor $\mathsf{E}[V(\mu)]$ bestemmes i apriorifordelingen af μ.

I tabel 8.1 er apriorimiddelværdien $\mathsf{E}[V(\mu)]$ derfor anført i kolonnen for præposteriorimiddelværdien $\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{E}[V(\mu)|\bar{x}]]$ af $V(\mu)$.

□

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 $X_i \,|\, p \in \mathrm{B}(1,p)$, $p \in \mathrm{Be}(\alpha,\beta)$ og $Z = X_1 + X_2 + \ldots + X_n$. Det følger da af sætning 6.2.2 at $Z|p \in \mathrm{B}(n,p)$, og endvidere har vi aposteriorifordelingen af p:

$$p \,|\, Z = z \in \mathrm{Be}(\alpha',\beta')$$

med

$$\alpha' = \alpha + z\,; \qquad \beta' = \beta + n - z\,,$$

hvorfor

$$\mathsf{E}[p \,|\, Z = z] = \frac{\alpha'}{\alpha'+\beta'} = \frac{\alpha+z}{\alpha+\beta+n}$$

og

$$\mathsf{V}\,[p \mid Z = z] = \frac{\alpha'\beta'}{(\alpha'+\beta')^2(\alpha'+\beta'+1)} = \frac{(\alpha+z)(\beta+n-z)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

Den prædiktive fordeling af $Z' = X_1' + \cdots + X_r'$ er en $\mathrm{Pl}(r, \alpha', \alpha'+\beta')$-fordeling.

Den sædvanlige parametrisering af binomialfordelingen er netop middelværdiparametriseringen

$$\mu(p) = \mathsf{E}\,[X_i|p] = p$$

og variansfunktionen er

$$V_{Bin}(p) = \mathsf{V}\,[X_i|p] = p(1-p)$$

Indfører vi i overensstemmelse med betegnelserne i afsnit 6.2 den tilsvarende parametrisering af strukturfordelingen

$$\pi = \mathsf{E}\,[p] = \frac{\alpha}{\alpha+\beta}$$

og

$$\frac{1}{\gamma} = \frac{\mathsf{E}\,[V_{Bin}(p)]}{\mathsf{V}\,[p]} = \alpha+\beta\,,$$

ser vi i overensstemmelse med (8.3.13), at vi kan udtrykke aposteriorifor-ventningsværdien af p (forventingsværdien i den prædiktive fordeling af en ny observation, X' fra samme gruppe) som et vejet gennemsnit af apriori-forventningsværdi og stikprøveresultat

$$\mathsf{E}\,[p \mid Z/n = \bar{x}] = \frac{\alpha+n\bar{x}}{\alpha+\beta+n} = \frac{\pi/\gamma+n\bar{x}}{1/\gamma+n} = w_n\pi + (1-w_n)\bar{x}$$

hvor vægten $w_n = 1/(1+n\gamma)$.

Sammenfattende har vi altså:

$$\begin{aligned} \pi_{\text{aposteriori}} &= w_n\pi_{\text{apriori}} + (1-w_n)\bar{x} \\ \frac{1}{\gamma_{\text{aposteriori}}} &= \frac{1}{\gamma_{\text{apriori}}} + n \end{aligned}$$

med

$$w_n = \frac{1}{1 + n\gamma_{\text{apriori}}}$$

Aposteriorimiddelværdien $\mathsf{E}\,[V_{Bin}(p)|\bar{x}]$ af $V_{Bin}(p)$ er

$$\mathsf{E}\,[V_{Bin}(p)|\bar{x}] = \frac{V_{Bin}(\pi_{\text{apost}})}{1 + \gamma_{\text{apost}}} = V_{Bin}(\pi_{\text{apost}})\,\frac{1/\gamma_{\text{apriori}} + n}{1/\gamma_{\text{apriori}} + n + 1}$$

Variansen i aposteriorifordelingen af p er derfor

$$\mathsf{V}\,[p|\bar{x}] = \frac{\mathsf{E}\,[V_{Bin}(p)|\bar{x}]}{1/\gamma_{\text{apriori}} + n} = \frac{V_{Bin}(\pi_{\text{apost}})}{1/\gamma_{\text{apriori}} + n + 1}$$

Præposteriorimiddelværdien af $V(p)$ er

$$\mathsf{E}_{X_1,\ldots,X_n}[V_{Bin}(p)|\bar{x}] = \mathsf{E}\,[V_{Bin}(p)] = \frac{V_{Bin}(\pi)}{1 + \gamma}$$

og præposteriorivariansen af p er

$$\mathsf{E}_{X_1,\ldots,X_n}[\mathsf{V}\,[p|\bar{x}]] = \mathsf{E}\,[V_{Bin}(p)]\,\frac{1}{1/\gamma_{\text{apriori}} + n} = \frac{V_{Bin}(\pi)}{1 + \gamma}\,\frac{1}{1/\gamma_{\text{apriori}} + n}$$

Den prædiktive fordeling

Momenterne i den prædiktive fordeling af $\overline{X'}$ bliver jvf (8.3.9) og (8.3.10)

$$\mathsf{E}\,[\overline{X'}_+|\bar{x}] = \pi_{\text{apost}}$$

og

$$\begin{aligned}\mathsf{V}\,[\overline{X'}_+|\bar{x}] &= \mathsf{E}\,[V_{Bin}(p)|\bar{x}]\left(\gamma_{\text{apost}} + \frac{1}{r}\right)\\ &= V_{Bin}(\pi_{\text{apost}})\frac{1/\gamma_{\text{apriori}} + n}{1/\gamma_{\text{apriori}} + n + 1}\left(\frac{1}{1/\gamma_{\text{apriori}} + n} + \frac{1}{r}\right)\end{aligned}$$

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 ϑ
- 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 $\mathrm{B}(1,p)$-fordelt med samme p, og at $p \in \mathrm{Be}(\alpha,\beta)$

Vi har altså de betingede sandsynligheder $\mathsf{P}[X_i = 1 \mid p] = p$, og de marginale sandsynligheder

$$\mathsf{P}[X_i = 1] = \mathsf{E}[p] = \frac{\alpha}{\alpha+\beta} = \pi$$

Endvidere har vi de marginale forventningsværdier og varianser (f.eks. fra Polyafordelingen med $n = 1$):

$$\mathsf{E}[X_i] = \frac{\alpha}{\alpha+\beta} = \pi$$

$$\mathsf{V}[X_i] = \frac{\alpha\beta}{(\alpha+\beta)^2} = \pi(1-\pi)$$

Samvariationen mellem X_i og X_j
Samvariationen kan f.eks. illustreres ved den simultane (marginale) frekvensfunktion

$$\mathsf{P}[X_i = x, X_j = z],\ (x,z) \in \{0,1\} \times \{0,1\}$$

Vi har

$$\mathsf{P}[X_i = x, X_j = z \mid p] = p^{x+z}(1-p)^{2-x-z}$$

således at den marginale sandsynlighed bliver

$$\mathsf{P}[X_i = x, X_j = z] = \mathsf{E}[p^{x+z}(1-p)^{2-x-z}]$$

De marginale sandsynligheder er anført i nedenstående skema:

$\mathsf{P}[X_i = x, X_j = z]$

x	z = 0	z = 1
0	$\frac{\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)}$	$\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}$
1	$\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}$	$\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$

Samvariationen udtrykkes sædvanligvis ved kovariansen,

$$\begin{aligned}\mathsf{COV}[X_i, X_j] &= \mathsf{E}\,[\mathsf{COV}[X_i, X_j|p]] + \mathsf{COV}[\mathsf{E}\,[X_i|p], \mathsf{E}\,[X_j|p]] \\ &= 0 + \mathsf{V}\,[p] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \pi(1-\pi)\frac{\gamma}{1+\gamma} = \pi(1-\pi)(1-w_1)\end{aligned}$$

således at korrelationskoefficienten 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 γ, d.v.s. jo mindre præcis aprioriordeling, desto større er korrelationen mellem enkeltobservationer fra samme gruppe (intraklyngekorrelationen).

Samvariationen mellem $\overline{X}$ og p

Idet vi sætter $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$, har vi den betingede forventningsværdi og varians

$$\mathsf{E}\,[\overline{X}|p] = p, \quad \mathsf{V}\,[\overline{X}|p] = \frac{V_{Bin}(p)}{n} = \frac{p(1-p)}{n}$$

og den marginale forventningsværdi og varians fås fra Polyafordelingen

$$\begin{aligned}\mathsf{E}\,[\overline{X}] &= \frac{\alpha}{\alpha+\beta} = \pi \\ \mathsf{V}\,[\overline{X}] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\left(1 + \frac{\alpha+\beta}{n}\right) \\ &= \pi(1-\pi)\,\frac{\gamma}{1+\gamma}\left(1 + \frac{1}{n\gamma}\right) = \pi(1-\pi)\,\frac{1-w_1}{1-w_n}\end{aligned}$$

Endvidere har vi kovariansen mellem $\overline{X}$ og p

$$\begin{aligned}\mathsf{COV}[\overline{X}, p] &= \mathsf{V}\,[p] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \pi(1-\pi)(1-w_1)\end{aligned}$$

hvorfor korrelationskoefficienten 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 stikprøvelsestørrelse, n, (eller jo mindre aprioripræcision), desto større er korrelationen mellem stikprøvegennemsnit og gruppeparameter.

Samvariationen mellem $\overline{X}$ og $\overline{X'}$.

I modellen symboliserer $\overline{X'}$ 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 $\mathrm{Be}(\alpha + n\bar{x}, \beta + n(1-\bar{x}))$ fordeling, hvorfor den prædiktive fordeling af $Z' = X_1' + X_2' + \cdots + X_r'$ er en $\mathrm{Pl}(r, \alpha+n\bar{x}, \alpha+\beta+n)$-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 betragtes som den betingede fordeling af Z' i den simultane fordeling af $\overline{X}$ og Z' (hvor vi har bortintegreret fordelingen af p).

I stedet for at betragte fordelingen af totalen, Z', vil vi betragte gennemsnittet af de fremtidige observationer,

$$\overline{X'} = \frac{Z'}{r}$$

Korrelationen mellem $\overline{X}$ og $\overline{X'}$ i den simultane fordeling af $\overline{X}$ og $\overline{X'}$ er

$$\begin{aligned} \rho_{\overline{X},\overline{X'}} &= \frac{1}{\sqrt{1+(\alpha+\beta)/n}\,\sqrt{1+(\alpha+\beta)/r}} = \frac{1}{\sqrt{1+1/(n\gamma)}\,\sqrt{1+1/(r\gamma)}} \\ &= \sqrt{1-w_n}\sqrt{1-w_r} \end{aligned}$$

Den prædiktive fordeling af $\overline{X'}$ kan altså fortolkes som regressionen af $\overline{X'}$ på $\bar{x}$ i denne fordeling. Vi har således:

$$\mathsf{E}\,[\overline{X'} \mid \overline{X} = \bar{x}] = \pi + (1-w_n)(\bar{x} - \pi)$$

med

$$1 - w_n = \frac{\mathsf{COV}[\overline{X}, \overline{X'}]}{\mathsf{V}\,[\overline{X}]} = \frac{n\gamma}{1+n\gamma}$$

Endvidere har vi

$$\mathsf{V}\,[\overline{X'} \mid \overline{X} = \bar{x}] = \frac{(\alpha+n\bar{x})[\beta+n(1-\bar{x})]}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \left(1 + \frac{\alpha+\beta+n}{r}\right)$$

□

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 $\mathrm{Be}(\alpha, \beta)$-fordeling med

$$\pi = \mathsf{E}\,[p] = 0.1$$

og

$$\gamma = \frac{\mathsf{V}\,[p]}{\mathsf{E}\,[V_{Bin}(p)]} = 0.05 = \frac{1}{20}$$

svarende til

$$\mathsf{V}\,[p] = \gamma\,\frac{V_{Bin}(\pi)}{1+\gamma} = 0.05 \times \frac{0.1 \times 0.9}{1.05} = 0.004286 = (0.0655)^2$$

Antag, at der udtages en stikprøve på $n = 15$ enheder fra en given produktion, og at man finder $Z = 1$ afvigende 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 afvigende enheder i denne produktion er således $\pi_{\text{apost}} = 0.0857$, og specielt er den forventede andel afvigende enheder blandt 100 tilfældigt udtagne

$$\mathsf{E}\,[\overline{X'}|\bar{x} = 0.06667] = \pi_{\text{apost}} = 0.0857$$

Til beskrivelse af variansen bestemmer man aposteriorimiddelværdien af $V_{Bin}(p)$,

$$\begin{aligned}\mathsf{E}\,[V_{Bin}(p)|\bar{x} &= 0.06667] = V_{Bin}(0.0857)\,\frac{1/0.05 + 15}{1/0.05 + 15 + 1} \\ &= 0.0857 \times 0.9143 \times \frac{35}{36}\end{aligned}$$

Man får derfor

$$\mathsf{E}\,[V_{Bin}(p)|\bar{x} = 0.06667] = 0.0784 \times 0.9722 = 0.0762\,,$$

og endvidere bestemmer man

$$\gamma_{\text{apost}} = \frac{1}{1/\gamma_{\text{apriori}} + 15} = \frac{1}{20 + 15} = 0.028571$$

Variansen i den prædiktive fordeling for $\overline{X'}_+$ er således

$$\begin{aligned} \mathsf{V}\,[\overline{X'}_+|\bar{x} &= 0.06667] = \mathsf{E}\,[V_{Bin}(p)|\bar{x} = 0.06667]\left(\gamma_{\text{apost}} + \frac{1}{r}\right) \\ &= 0.0762 \times (0.028571 + 0.01) \end{aligned}$$

Vi bemærker, at da aprioriusikkerheden $\gamma = 1/20$ så nogenlunde modsvarer stikprøvestørrelsen $n = 15$, bliver aposteriorimiddelværdien π_{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:

$$\begin{aligned} \mathsf{V}\,[p|\bar{x} = 0.06667] &= \gamma_{\text{apost}}\mathsf{E}\,[V_{Bin}(p)|\bar{x} = 0.06667] \\ &= 0.028571 \times 0.0762 = 0.0002176 = (0.0467)^2 \end{aligned}$$

Aposteriorimiddelværdien $\mathsf{E}\,[V_{Bin}(p)|\bar{x} = 0.06667]$ af $V_{Bin}(p)$ er ændret noget i forhold til aprioriværdien

$$\mathsf{E}\,[V_{Bin}(p)] = V_{Bin}(\pi)/1.05 = 0.0857$$

Ændringen skyldes hovedsageligt, at fordelingen af p har ændret middelværdi.

Endelig bemærker vi ved at sammenholde udtrykket for variansen i den prædiktive fordeling af $\overline{X'}_+$

$$\mathsf{V}\,[\overline{X'}_+|\bar{x} = 0.06667] = 0.0762 \times (0.028571 + 0.01)\,,$$

med den tilsvarende "prædiktionsvarians" såfremt vi ikke havde undersøgt en stikprøve fra partiet

$$\mathsf{V}\,[\overline{X}_+] = \mathsf{E}\,[V_{Bin}(p)]\left(\gamma_{\text{apriori}} + \frac{1}{100}\right) = 0.0857 \times (0.05 + 0.01)\,,$$

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

$$\begin{aligned} \mathsf{E}\left[p \mid Z = z\right] &= \frac{\alpha + n}{\alpha + \beta + n + z} \\ \mathsf{V}\left[p \mid Z = z\right] &= \frac{(\alpha + n)(\beta + z)}{(\alpha + \beta + n + z)^2(\alpha + \beta + n + 1)} \end{aligned}$$

Indfører vi i overensstemmelse med betegnelserne i afsnit 6.3 middelværdiparametriseringen ved

$$\mu(p) = \mathsf{E}\left[X_i|p\right] = \frac{1-p}{p}$$

og variansfunktionen

$$V_{NB}(\mu) = \mathsf{V}\left[X_i|p\right] = \mu(1+\mu)$$

med den tilsvarende parametrisering af strukturfordelingen

$$\psi = \mathsf{E}\left[\mu\right] = \frac{\beta}{\alpha - 1}$$

og

$$\frac{1}{\gamma} = \frac{\mathsf{E}\left[V_{NB}(\mu)\right]}{\mathsf{V}\left[\mu\right]} = \alpha - 1$$

i overensstemmelse med betegnelserne i Lemma 6.3.1, finder man, at aposterioriforventingsværdien af μ (forventningsværdien i den prædiktive fordeling af en ny observation, X' fra samme gruppe) er

$$\mathsf{E}\left[X' \mid Z/n = \bar{x}\right] = \mathsf{E}\left[\mu \mid Z/n = \bar{x}\right] = \frac{\psi/\gamma + n\bar{x}}{1/\gamma + n}$$

Vi har således

$$\begin{aligned} \psi_{\text{aposteriori}} &= w_n \psi_{\text{apriori}} + (1 - w_n)\bar{x} \\ \frac{1}{\gamma_{\text{aposteriori}}} &= \frac{1}{\gamma_{\text{apriori}}} + n \end{aligned}$$

med

$$w_n = \frac{1}{1 + n\gamma_{\text{apriori}}}$$

Udtrykker vi aposterioriforventningsværdien som regressionen af Y på $\bar{x}$ har vi

$$\mathsf{E}\left[X' \mid Z/n = \bar{x}\right] = \psi + (1 - w_n)(\bar{x} - \psi)$$

8.3.5 Poisson-Gamma sampling

Lad $X_i|\mu \in \mathrm{P}(\mu)$, $\mu \in \mathrm{G}(\alpha, 1/\beta)$ og $Z = X_1 + X_2 + \ldots + X_n$. Det gælder da, at $Z|\mu \in \mathrm{P}(n\mu)$, hvorfor det følger af sætning 6.4.2, at den marginale fordeling af Z er en $\mathrm{NB}(\alpha, \beta/(\beta + n))$-fordeling.

Det gælder da, at aposteriorifordelingen $\mu|Z = z \in \mathrm{G}(\alpha+z, 1/(\beta+n))$ med

$$\begin{aligned} \mathsf{E}\left[\mu \mid Z = z\right] &= \frac{\alpha + z}{\beta + n} \\ \mathsf{V}\left[\mu \mid Z = z\right] &= \frac{\alpha + z}{(\beta + n)^2} \end{aligned}$$

Den sædvanlige parametrisering af Poissonfordelingen er netop middelværdiparametriseringen $\mu = \mathsf{E}\left[X_i|\mu\right]$ og variansfunktionen er

$$V_P(\mu) = \mu$$

Indfører vi i overensstemmelse med betegnelserne i afsnit 6.4 den tilsvarende parametrisering af strukturfordelingen

$$m = \mathsf{E}\left[\mu\right] = \frac{\alpha}{\beta}, \quad \text{og} \quad \frac{1}{\gamma} = \frac{\mathsf{E}\left[V_P(\mu)\right]}{\mathsf{V}\left[\mu\right]} = \beta\,,$$

kan vi udtrykke aposterioriforventningen af μ (forventingsværdien i den prædiktive fordeling af en ny observation, X' fra samme gruppe) som:

$$\mathsf{E}\left[X' \mid Z/n = \bar{x}\right] = \mathsf{E}\left[\mu \mid Z/n = \bar{x}\right] = \frac{\alpha + n\bar{x}}{\beta + n} = \frac{m/\gamma + n\bar{x}}{1/\gamma + n}$$

Vi har således

$$m_{\text{aposteriori}} = w_n m_{\text{apriori}} + (1 - w_n)\bar{x} \tag{8.3.19}$$

$$\frac{1}{\gamma_{\text{aposteriori}}} = \frac{1}{\gamma_{\text{apriori}}} + n \tag{8.3.20}$$

med

$$w_n = \frac{1}{1 + n\gamma_{\text{apriori}}}$$

Udtrykt ved regressionen af X' på $\bar{x}$ har vi:

$$\mathsf{E}\,[X' \mid Z/n = \bar{x}] = m + (1 - w_n)(\bar{x} - m)$$

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(μ)-fordeling, hvor det gennemsnitlige antal fejl pr m^2, μ, 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 $sak_2/(k-1) = 23.325$ og et gennemsnitligt antal fejl på $\bar{x}_{++} = 2.5$ [fejl/m^2].

Vi har nu af tabel 6.1, at

$$m = \mathsf{E}\,[\mu] = \frac{\alpha}{\beta} = 2.5 \text{ [fejl/m}^2\text{]}$$

og

$$\gamma = \frac{\mathsf{V}\,[\mu]}{\mathsf{E}\,[V_P(\mu)]} = \frac{1}{\beta} = 0.833$$

Det følger af Sætning 6.4.2 at den marginale fordeling af $Z = \sum_{i=1}^{n} X_i$ har momenterne

$$\mathsf{E}\,[Z] = nm \tag{8.3.21}$$

$$\mathsf{V}\,[Z] = nm(1 + n\gamma) \tag{8.3.22}$$

Antag nu, at der udføres en ny vævning, og der udtages en stikprøve bestående af $n = 10$ stykker på hver 1 [m^2] fra produktionen, og antallet af fejl $x_1, x_2, \ldots, x_{10}$ optælles. Antag, at man fandt det gennemsnitlige antal fejl,

$$\bar{x} = 0.1 \text{ [fejl/m}^2\text{]}$$

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

$$\mu_{\text{aposteriori}} = 0.1072 \times 2.5 + 0.8928 \times 0.1 = 0.36 \text{ [fejl/m}^2\text{]}$$

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 [m^2]. Der gælder da jvf. (8.3.21) og (8.3.22), at

$$\begin{aligned} \mathsf{E}[Z'] &= r\,\mu_{\text{aposteriori}} \\ \mathsf{V}[Z'] &= r\,\mu_{\text{aposteriori}}(1 + r\,\gamma_{\text{aposteriori}}) \end{aligned}$$

For en prøve bestående af 20 [m^2] vil vi altså forvente det totale antal fejl

$$\mathsf{E}[Z'] = 20 \times 0.36 = 7.2 \text{ [fejl/m}^2\text{]}$$

Variansen på denne prædiktion er

$$\mathsf{V}[Z'] = 20 \times 0.36(1 + 20 \times 0.0893) = 20.06 \text{ [fejl/m}^2\text{]}^2$$

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 α og β.

Nedenstående tabel illustrerer opdateringen svarende til forskellige stikprøvestørrelser og forskellige stikprøveresultater:

Eksempel på bestemmelse af aposteriorifordeling ved Poissonfordelt målestøj

Parameter	Før	Efter observation af		
		$n = 10$ $z = 1$ $\bar{x} = 0.1$	$n = 10$ $z = 60$ $\bar{x} = 6$	$n = 20$ $z = 120$ $\bar{x} = 6$
$m = \mathsf{E}[\mu]$	2.5	0.36	5.63	5.80
γ	0.833	0.0893	0.0893	0.0472
$\mathsf{V}[\mu] = m \times \gamma$	2.0825	0.03	0.50	0.27
α	3	4	63	123
β	1.2	11.2	11.2	21.2
Fordeling af fejl i én ny m $X' \in \mathrm{NB}(\alpha, \beta/(1+\beta))$				
$\mathsf{E}[X'] = m$	2.5	0.36	5.63	5.80
$\mathsf{V}[X'] = m(1+\gamma)$	4.58	0.39	6.13	6.07
Fordeling af fejl i 10 nye m $Z' \in \mathrm{NB}(\alpha, \beta/(10+\beta))$				
$\mathsf{E}[Z'] = 10m$	25	3.6	56.3	58.0
$\mathsf{V}[Z'] = 10m(1+10\gamma)$	233.32	6.81	106.57	85.38
Fordeling af fejl i 20 nye m $Z' \in \mathrm{NB}(\alpha, \beta/(20+\beta))$				
$\mathsf{E}[Z'] = 20m$	50	7.2	112.6	116.0
$\mathsf{V}[Z'] = 20m(1+20\gamma)$	883	20.06	313.7	225.50

□

8.3.6 Exponential reciprok gamma sampling

Lad $X_i|\mu \in \mathrm{Ex}(\mu)$, $\mu \in \mathrm{RGam}(\alpha, \beta)$ og $Z = X_1 + X_2 + \cdots + X_n$. Det gælder da, at $Z|\mu \in \mathrm{G}(n, \mu)$ og det følger nu af sætning 6.5.2, at den marginale fordeling af Z er en $\mathrm{RBet}(\alpha, n, \beta)$-fordeling, og endvidere har vi,

at $\mu \mid Z = z \in \mathrm{RGam}(\alpha + n, \beta + z)$ med

$$\begin{aligned} \mathsf{E}\,[\mu \mid Z = z] &= \frac{\beta + z}{\alpha - 1 + n} \\ \mathsf{V}\,[\mu \mid Z = z] &= \frac{\beta + z}{(\alpha + n - 1)^2(\alpha + n - 2)} \,. \end{aligned}$$

I afsnit 6.5 betragtede i middelværdiparametriseringen af gammafordelingen ved

$$\mathsf{E}\,[X_i|\mu] = \mu$$

og variansfunktionen

$$V_G(\mu) = \mu^2$$

svarende til dispersionsparameter 1, og vægt $1/n$. Indfører vi den tilsvarende parametrisering af strukturfordelingen (jvf lemma 6.5.1)

$$m = \mathsf{E}\,[\mu] = \frac{\beta}{\alpha - 1}$$

og

$$\frac{1}{\gamma} = \frac{\mathsf{E}\,[V_G(\mu)]}{\mathsf{V}\,[\mu]} = \alpha - 1 \,,$$

kan vi udtrykke forventingsværdien i den prædiktive fordeling af en ny observation, X' fra samme gruppe (aposterioriforventningsværdien af μ) som

$$\mathsf{E}\,[X' \mid Z/n = \bar{x}] = \mathsf{E}\,[\mu \mid Z/n = \bar{x}] = \frac{m/\gamma + n\bar{x}}{1/\gamma + n}$$

$$\begin{aligned} m_{\text{aposteriori}} &= w_n m_{\text{apriori}} + (1 - w_n)\bar{x} \\ \frac{1}{\gamma_{\text{aposteriori}}} &= \frac{1}{\gamma_{\text{apriori}}} + n \end{aligned}$$

med

$$w_n = \frac{1}{1 + n\gamma_{\text{apriori}}}$$

Udtrykt ved regressionen af X' på $\bar{x}$ har vi

$$\mathsf{E}\,[X' \mid Z/n = \bar{x}] = m + (1 - w_n)(\bar{x} - m)$$

8.3.7 Normalfordeling med samme varians

Såfremt $X_i \,|\, \mu \in \mathrm{N}(\mu, \sigma^2)$, $\mu \in \mathrm{N}(m, \sigma_0^2)$ og $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$, da vil $\overline{X} \,|\, \mu \in \mathrm{N}(\mu, \sigma^2/n)$ og den marginale fordeling af $\overline{X}$ er (jvf bemærkning 3 til sætning 5.3.2) $\overline{X} \in \mathrm{N}(m, \sigma^2(\gamma + 1/n))$, hvor

$$\gamma = \frac{\sigma_0^2}{\sigma^2} \,.$$

Aposteriorifordelingen af μ efter observation af $\overline{X} = \bar{x}$ er en normalfordeling med forventningsværdi

$$\begin{aligned} \mathrm{E}\,[\mu \,|\, \overline{X} = \bar{x}] &= \frac{m/\sigma_0^2 + n\bar{x}/\sigma^2}{1/\sigma_0^2 + n/\sigma^2} \\ &= \frac{m/\gamma + n\bar{x}}{1/\gamma + n} = w_n m + (1 - w_n)\bar{x} \end{aligned}$$

med $w = 1/(1 + n\gamma)$, og varians

$$\sigma_1^2 = \mathrm{V}\,[\mu \,|\, \overline{X} = \bar{x}] = \frac{1}{\dfrac{1}{\sigma_0^2} + \dfrac{n}{\sigma^2}}$$

Vi ser, at også her gælder, at aposterioriforventningsværdien $\mathrm{E}\,[\mu | \overline{X} = \bar{x}]$ er et vejet gennemsnit af aprioriforventningsværdien m og stikprøvegennemsnittet $\bar{x}$ med de respektive præcisioner (reciproke varianser)

$$\frac{1}{\mathrm{V}\,[\mu]} = \frac{1}{\sigma_0^2} \quad \text{og} \quad \frac{n}{\mathrm{E}\,[V(\mu)]} = \frac{n}{\sigma^2}$$

som vægte. Da det kun er de relative vægte, der er af betydning, har vi som tidligere valgt at parametrisere ved parameteren γ.

Sammenfattende har vi:

$$\begin{aligned} \mu_{\text{aposteriori}} &= w_n \mu_{\text{apriori}} + (1 - w_n)\bar{x} \\ \frac{1}{\gamma_{\text{aposteriori}}} &= \frac{1}{\gamma_{\text{apriori}}} + n \end{aligned}$$

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
- (μ, σ^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 μ er indbyrdes uafhængige $\mathrm{N}(\mu, \sigma^2)$-fordelt med samme μ, og at $\mu \in \mathrm{N}(m, \sigma_0^2)$ med $\sigma_0^2 = \gamma\sigma^2$.

Samvariationen mellem X_i og X_j
Vi minder om, at der gælder (Sætning 5.3.2)

$$\begin{aligned} \mathsf{COV}[X_i, X_j] &= \mathsf{E}\,[\mathsf{COV}[X_i, X_j \mid \mu]] + \mathsf{COV}[\mathsf{E}\,[X_i \mid \mu], \mathsf{E}\,[X_j \mid \mu]] \\ &= 0 + \mathsf{V}\,[\mu] = \sigma_0^2 \end{aligned}$$

hvorfor vi har intraklyngekorrelationen

$$\rho_{X_i, X_j} = \frac{1}{1 + \sigma^2/\sigma_0^2} = \frac{\gamma}{1+\gamma}$$

Samvariationen mellem $\overline{X}$ og μ
Idet $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$, har vi den betingede forventningsværdi og varians

$$\mathsf{E}\,[\overline{X} \mid \mu] = \mu; \quad \mathsf{V}\,[\overline{X} \mid \mu] = \sigma^2/n$$

Den marginale forventningsværdi og varians er

$$\mathsf{E}\,[\overline{X}] = \mu_0; \quad \mathsf{V}\,[\overline{X}] = \sigma_0^2 + \sigma^2/n$$

og kovariansen mellem $\overline{X}$ og μ er

$$\mathsf{COV}[\overline{X}, \mu] = \mathsf{E}\,[\mathsf{COV}[\overline{X}, \mu] \mid \mu] + \mathsf{COV}[\mathsf{E}\,[\overline{X} \mid \mu], \mu] = 0 + \mathsf{COV}[\mu, \mu] = \sigma_0^2$$

hvorfor vi har korrelationskoefficienten mellem stikprøvegennemsnit, $\overline{X}$ og gruppeparameter, μ

$$\rho_{\overline{X}, \mu} = \frac{1}{\sqrt{1 + \sigma^2/(n\sigma_0^2)}} = \frac{1}{\sqrt{1 + 1/(n\gamma)}} = \sqrt{1 - w_n}$$

med $w_n = 1/(1+n\gamma)$

For $n \to \infty$ vil $\rho_{\overline{X},\mu} \to 1$.

Aposteriorifordelingen af μ givet $\overline{X} = \bar{x}$ karakteriseres ved regressionen

$$\begin{aligned} \mathsf{E}\,[\mu \mid \overline{X} = \bar{x}] &= m + (1 - w_n)(\bar{x} - m) \\ \mathsf{V}\,[\mu \mid \overline{X} = \bar{x}] &= \frac{1}{1/\sigma_0^2 + n/\sigma^2} = \frac{\sigma^2}{n}(1 - w_n) \end{aligned}$$

For $n \to \infty$ vil $\mathsf{E}\,[\mu \mid \overline{X} = \bar{x}] \to \bar{x},$, og $\mathsf{V}\,[\mu \mid \overline{X} = \bar{x}] \to 0$.

Samvariationen mellem $\overline{X}$ og $\overline{X'}$

I modellen symboliserer $\overline{X'}$ 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 μ). Efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ beskrives vores viden om μ ved aposteriorifordelingen af μ som beskrevet ovenfor, hvorfor den prædiktive fordeling af totalen $Z'_+ = X'_1 + X'_2 + \cdots + Z'_r$ er en $\mathrm{N}(rm_1, r\sigma^2 + r^2\sigma_1^2)$-fordeling, nemlig den marginale fordeling af summen af r observationer under hensyntagen til den opdaterede viden om μ.

Korrelationen mellem $\overline{X}$ og $\overline{X'}$ i den simultane fordeling af $\overline{X}$ og $\overline{X'}$ er

$$\begin{aligned} \rho_{\overline{X},\overline{X'}} &= \frac{1}{\sqrt{1+\sigma^2/(n\sigma_0^2)}\,\sqrt{1+\sigma^2/(r\sigma_0^2)}} = \frac{1}{\sqrt{1+1/(n\gamma)}\,\sqrt{1+1/(r\gamma)}} \\ &= \sqrt{1-wn}\sqrt{1-w_r} \end{aligned}$$

Den prædiktive fordeling af $\overline{X'}$ svarende til $\overline{X} = \bar{x}$ kan da fortolkes som regressionen af $\overline{X'}$ på $\bar{x}$ i den simultane fordeling af $\overline{X}$ og $\overline{Y}$ (hvor vi har bortintegreret μ).

$$\begin{aligned} \mathsf{E}\,[\overline{X'} \mid \overline{X} = \bar{x}] &= m + (1 - w_n)(\bar{x} - m) \\ \mathsf{V}\,[\overline{X'} \mid \overline{X} = \bar{x}] &= \frac{\sigma^2}{p} + \frac{1}{1/\sigma_0^2 + n/\sigma^2} = \frac{\sigma^2}{p} + \frac{\sigma^2}{n}(1 - w_n) \end{aligned}$$

□

8.3.8 Empiriske varianser fra normalfordelte observationer

Såfremt $S^2|\sigma^2 \in \sigma^2\chi^2(f)$ og

$$1/\sigma^2 \in \mathrm{G}\big(\nu/2,\ 2/[(\nu-2)\sigma_0^2]\big)\ ,$$

da vil aposteriorifordelingen af $1/\sigma^2$ efter observation af $S^2 = s^2$ være en

$$\mathrm{G}\big(\nu_1/2,\ 2/[(\nu_1-2)\sigma_1^2]\big) - \text{fordeling}$$

hvor

$$\nu_1 = \nu + f \tag{8.3.23}$$

og

$$\sigma_1^2 = \frac{(\nu-2)\sigma_0^2 + fs^2}{\nu - 2 + f} \tag{8.3.24}$$

Aposterioriforventningsværdien af σ^2 er

$$\mathsf{E}\,[\sigma^2|s^2] = \sigma_1^2$$

Det ses, at aposteriorimiddelværdien er et vejet gennemsnit mellem apriorimiddelværdien

$$\mathsf{E}\,[\sigma^2] = \sigma_0^2$$

og stikprøveresultatet s^2 med vægtene henholdsvis $\nu - 2$ og f.

Den prædiktive fordeling for $(S')^2$, hvor $(S')^2|\sigma^2 \in \sigma^2\chi^2(r)/r$ og $1/\sigma^2$ følger en $\mathrm{G}\big(\nu_1/2,\ 2/[(\nu_1-2)\sigma_1^2]\big)$-fordeling bliver en

$$\mathrm{RBet}\Big(\nu_1/2,\ r/2,\ \frac{\nu_1-2}{r}\sigma_1^2\Big) - \text{fordeling}$$

jvf afsnit 6.6.

Stikprøve-fordeling af $Z\|\mu$	Struktur-fordeling $w(\cdot)$	Aposteriori-fordeling af μ efter observation af $Z=z$	Prædiktiv fordeling af Z'	Referen
$Z\|p \in \mathrm{B}(n,p)$	$p \in \mathrm{Be}(\alpha,\beta)$	$\mathrm{Be}(\alpha+z,\beta+n-z)$	$\mathrm{Pl}(r,\alpha+z,\alpha+\beta+n)$	Afsn 8.3
$Z\|p \in \mathrm{NB}(n,p)$	$p \in \mathrm{Be}(\alpha,\beta)$	$\mathrm{Be}(\alpha+n,\beta+z)$	$\mathrm{NPl}(r,\beta+z,\alpha+\beta+n+z)$	Afsn 8.3
$Z\|\mu \in \mathrm{P}(n\mu)$	$\mu \in \mathrm{G}(\alpha,1/\beta)$	$\mathrm{G}(\alpha+z,1/(\beta+n))$	$\mathrm{NB}(\alpha+z,(\beta+n)/(\beta+n+r))$	Afsn 8.3.5
$Z\|\mu \in \mathrm{G}(n\mu)$	$\mu \in \mathrm{RGam}(\alpha,\beta)$	$\mathrm{RGam}(\alpha+n,\beta+z)$	$\mathrm{RBet}(\alpha+n,r,\beta+z)$	Afsn 8.3.6
$Z\|\mu \in \mathrm{N}(n\mu,n\sigma^2)$	$\mu \in \mathrm{N}(m,\sigma_0^2)$	$\mu \in \mathrm{N}(m_1,\sigma_1^2)$ $m_1=\dfrac{m/\gamma+z}{1/\gamma+n}$ $1/\sigma_1^2=1/\sigma_0^2+n/\sigma^2$	$\mathrm{N}(nm_1,r\sigma^2+r^2\sigma_1^2)$	Afsn 8.3.7
$S^2\|\sigma^2 \in \sigma^2\chi^2(f)/f$	$1/\sigma^2 \in \mathrm{G}(\nu/2,2/\beta_0^*)$	$1/\sigma^2 \in \mathrm{G}(\nu_1/2,2/beta_1^*)$	$(S')^2 \in \mathrm{RBet}(\nu_1/2,r/2,(\nu_1-2)\sigma_1^2/r)$	Afsn. 8.3.8

Tabel 8.1. Aposteriorifordelinger for endimensionale eksponentielle familier med naturlige konjugerede strukturfordelinger
I 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 σ_1^2 bestemt ved (8.3.24).

Momenter i prædiktiv fordeling af $\overline{X}' = \sum_{j=1}^{r} X_j'/r$ efter observation af $\bar{x} = \sum_{i=1}^{n} x_i/n$.

$$\mathsf{E}\,[\overline{X}'|\bar{x}] = m_1\,; \qquad \mathsf{V}\,[\overline{X}'|\bar{x}] = \mathsf{E}\,[V(\mu)|\bar{x}]\left(\gamma_1 + \frac{1}{r}\right)$$

Middelværdi, m_1 og varians $\mathsf{V}\,[\mu|\bar{x}]$ i aposteriorifordelingen af μ efter observation af $\bar{x} = \sum_{i=1}^{n} x_i/n$.

$$m_1 = (m/\gamma + n\bar{x})/(1/\gamma + n)\,; \qquad \gamma_1 = 1/(1/\gamma + n)\,; \qquad \mathsf{E}\,[\mu|\bar{x}] = m_1\,; \qquad \mathsf{V}\,[\mu|\bar{x}] = \mathsf{E}\,[V(\mu)|\bar{x}]/(1/\gamma + n)$$

Præposteriorivarians af μ : $\quad \mathsf{E}_{X_1,\ldots,X_n}[\mathsf{V}\,[\mu|\bar{x}] = \mathsf{E}\,[\mathsf{E}\,[V(\mu)|\bar{x}]/(1/\gamma + n)$

Stikprøve-fordeling af $X_i\|\theta$	Struktur-fordeling $w(\cdot)$	$m_1 = \mathsf{E}\,[\mu\|\bar{x}]$	$\mathsf{E}\,[V(\mu)\|\bar{x}]$	Præposteriori-middelværdi af $V(\mu)$	Reference
$\mathrm{B}(1,p)$	$p \in \mathrm{Be}(\alpha,\beta)$	$\pi_1 = \dfrac{\pi/\gamma + n\bar{x}}{1/\gamma + n}$	$\pi_1(1-\pi_1)\,\dfrac{1/\gamma + n}{1/\gamma + n + 1}$	$\dfrac{\pi(1-\pi)}{1+\gamma}$	Afsn 8.3.3
$\mathrm{Geo}(1,p)$	$p \in \mathrm{Be}(\alpha,\beta)$	$\psi_1 = \dfrac{\psi/\gamma + n\bar{x}}{1/\gamma + n}$	$\psi_1(1+\psi_1)\,\dfrac{1/\gamma + n}{1/\gamma + n - 1}$	$\dfrac{\psi(1+\psi)}{1-\gamma}$	Afsn 8.3.4
$\mathrm{P}(\mu)$	$\mu \in \mathrm{G}(\alpha, 1/\beta)$	$m_1 = \dfrac{m/\gamma + n\bar{x}}{1/\gamma + n}$	m_1	m	Afsn 8.3.5
$\mathrm{Ex}(\mu)$	$\mu \in \mathrm{RGam}(\alpha, 1/\beta)$	$m_1 = \dfrac{m/\gamma + n\bar{x}}{1/\gamma + n}$	$m_1^2\,\dfrac{1/\gamma + n}{1/\gamma + n - 1}$	$\dfrac{m^2}{1-\gamma}$	Afsn 8.3.6
$\mathrm{N}(\mu,\sigma^2)$	$\mathrm{N}(m,\sigma_0^2)$	$m_1 = \dfrac{m/\gamma + n\bar{x}}{1/\gamma + n}$	σ^2	σ^2	Afsn 8.3.7

Tabel 8.2. Aposteriorimiddelværdier 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 $\mathrm{N}(\mu, \sigma^2)$- fordelte, hvor σ^2 antages at variere fra gruppe til gruppe i overensstemmelse med en $\mathrm{RGam}(\alpha, \beta)$-fordeling. Parameteren μ varierer ligeledes fra gruppe til gruppe, og det antages, at for givet gruppevarians, σ^2 , vælges μ i overensstemmelse med en $\mathrm{N}(\mu_0, \sigma^2/m)$- fordeling.

Betinget ford. af (T, Z) for givet (μ, σ^2)	Konjugeret aprio-rifordeling af (μ, σ^2)	Aposteriorifordeling af (μ, σ^2) efter observation af (t, z)	Marginalfordeling af (T, Z)
Ford af $(T\|\mu, \sigma^2)$	Ford. af $(\mu\|\sigma^2)$	Ford. af $(\mu\|\sigma^2)$	Ford. af $(T\|\sigma^2)$
$\mathrm{N}(n\mu, n\sigma^2)$	$\mathrm{N}(\mu_0, \sigma^2/m)$	$\mathrm{N}(\mu_1, \sigma^2/(m+n))$ $\mu_1 = \frac{m\mu_0 + n\bar{x}}{m+n}$	$\mathrm{N}(n\mu_0, n\sigma^2 + n^2\sigma^2/m)$
	Ubet.ford. af μ $\mathrm{T}\left(2\alpha, \mu_0, \sqrt{\frac{\beta}{m\alpha}}\right)$	Ubet. ford. af μ $\mathrm{T}\left(2\alpha + n, \mu_1, \sqrt{\frac{\beta_1}{(2\alpha+n)(m+n)}}\right)$ $\beta_1 = z + 2\beta + \frac{nm}{n+m}(\bar{x} - \mu_0)^2$	Ubet. ford. af T $\mathrm{T}\left(2\alpha, n\mu_0, \sqrt{\frac{\beta(n^2+mn)}{m\alpha}}\right)$
Ford. af $(Z\|\sigma^2)$ $\mathrm{G}((n-1)/2, 2\sigma^2)$	Ubet.ford. af σ^2 $\mathrm{RGam}(\alpha, \beta)$	Ubet. ford. af σ^2 $\mathrm{RGam}(\alpha + n/2, \beta_1/2)$	Ubet. ford. af Z $\mathrm{RBet}(\alpha, (n-1)/2, 2\beta)$

Tabel 8.3. Aposteriorifordelinger for normaltfordelte observationer med varierende middelværdi og varians. Stikprøvestørrelse n.

$T = \sum X_i; \quad \overline{X} = \sum X_i/n; \quad Z = \sum (X_i - \overline{X})^2.$

8.4 Filtrering af en tidsrække

Vi betragter modellen:

$$\theta_t = \alpha\theta_{t-1} + \delta_t \tag{8.4.1}$$
$$X_t = \theta_t + \epsilon_t,\ t = 1, 2, 3, \ldots \tag{8.4.2}$$

hvor $\delta_1, \delta_2, \ldots$ er indbyrdes uafhængige, $\delta_t \in \mathrm{N}(0, \sigma_\delta^2)$ og $\epsilon_1, \epsilon_2, \ldots$ ligeledes er indbyrdes uafhængige, $\epsilon_t \in \mathrm{N}(0, \sigma^2)$, og hvor ϵ_t og δ_t er uafhængige, $t = 1, 2, 3, \ldots$.

Modellen kan beskrive en proces, hvor processens tilstand, θ, varierer med tiden i overensstemmelse med en AR(1) proces. Tilstanden, θ_t, kan imidlertid ikke observeres direkte, men observationen, X_t, er behæftet med målestøj, σ^2.

Vi antager, at parametrene α, σ^2 og σ_δ^2 er kendte.

Sætter vi startværdien $\theta_0 = 0$ har vi apriorifordelingen $\theta_1 \in \mathrm{N}(0, \sigma_\delta^2)$ og stikprøvefordelingen

$$X_1 \mid \theta_1 \in \mathrm{N}(\theta_1, \sigma^2)$$

Vi får derfor, at aposteriorifordelingen af θ_1 er en normalfordeling med forventningsværdi og varians

$$\theta_1^f = \mathrm{E}[\theta \mid X_1 = x_1] = (1 - w_1)x$$
$$\mathrm{V}[\theta \mid X_1 = x_1] = \sigma^2(1 - w_1),$$

hvor

$$(1 - w_1) = \frac{\gamma_0}{1 + \gamma_0}$$

med $\gamma_0 = \sigma_\delta^2/\sigma^2$.

Den betingede fordeling af θ_2, givet observationen $X_1 = x_1$ er da en normalfordeling med forventningsværdi og varians

$$\hat{\theta}_2^p = \mathrm{E}[\theta_2 \mid X_1 = x_1] = \alpha\theta_1^f$$
$$\mathrm{V}[\theta_2 \mid X_1 = x_1] = \alpha^2\mathrm{V}[\theta_1 \mid X_1 = x_1] + \sigma_\delta^2 = \alpha^2\sigma^2(1 - w_1) + \sigma_\delta^2$$

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 θ_2 og givet $X_1 = x_1$

er en $N(\theta_2, \sigma^2)$-fordeling, og vi kan derefter bestemme aposteriorifordelingen for θ_2 etc.

Ved observation af X_t har vi således at apriorifordelingen for θ_t (nemlig den betingede fordeling af θ_t givet $X_1 = x_1, X_2 = x_2, \ldots, X_{t-1} = x_{t-1}$) er en normalfordeling med forventningsværdi θ_t^p og varians $\gamma_t\sigma^2$ mens stikprøvefordelingen er en $N(\theta_t, \sigma^2)$-fordeling. Aposteriorifordelingen af θ_t efter observation af $X_t = x_t$ (den betingede fordeling af θ_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):

$$\hat{\theta}_t^f = \mathsf{E}\,[\theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t] = \hat{\theta}_t^p + (1 - w_t)(x_t - \hat{\theta}_t^p) \tag{8.4.3}$$

og med variansen

$$\mathsf{V}\,[\theta \mid X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t] = \sigma^2(1 - w_t) \tag{8.4.4}$$

hvor

$$w_t = 1/(1 + \gamma_t) \tag{8.4.5}$$

Prædiktionen $\hat{\theta}_t^p$ bestemmes ved

$$\hat{\theta}_t^p = \alpha\hat{\theta}_t^f \tag{8.4.6}$$

Parameteren γ_t opdateres ved

$$\gamma_{t+1} = \alpha^2(1 - w_t) + \gamma_0 \tag{8.4.7}$$

Opdateringen (8.4.7) af parameteren γ_t følger ved at bemærke, at inden observation af X_t har man

$$\mathsf{V}\,[\theta_t \mid x_1, x_2, \ldots, x_{t-1}] = \sigma^2\gamma_t$$

Efter observation af $X_t = x_t$ er aposteriorivariansen derfor

$$\mathsf{V}\,[\theta_t \mid x_1, x_2, \ldots, x_t] = \sigma^2\frac{\gamma_t}{1 + \gamma_t}\sigma^2(1 - w_t)$$

Variansen i fordelingen af θ_{t+1} (der er apriorifordeling ved observation af X_{t+1}) bliver da

$$\mathsf{V}\,[\theta_{t+1} \mid x_1, x_2, \ldots, x_t] = \alpha^2\mathsf{V}\,[\theta_t \mid x_1, x_2, \ldots, x_t] + \sigma_\delta^2 = \sigma^2\{\alpha^2(1 - w_t) + \gamma_0\}$$

Opdateringen ved (8.4.3) til (8.4.7) benævnes et Kalman-filter. I filtreringssammenhænge benævnes størrelsen $(1 - w_t)$ ofte Kalman-forstærkningen.

Vi ser, at for $t \to \infty$ vil størrelserne γ_t og w_t konvergere mod værdier γ^* og w^*, der kun afhænger af de givne parametre α og γ_0. Dette indebærer, at for store værdier af t vil aposteriorimiddelværdien $\hat{\theta}_t^f$ 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 θ_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, θ_t, i filtreringsproblemet, må man mindske observationsstøjen, σ^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 observation, idet udledningerne i dette tilfælde er lidt enklere, end i det generelle tilfælde.

Lemma 8.5.1 ***Aposteriorifordeling svarende til en enkelt observation***

Lad $X \mid \boldsymbol{\mu} \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ og lad $\boldsymbol{\mu} \in \mathrm{N}_p(\mathbf{m}, \boldsymbol{\Sigma}_0)$, hvor $\boldsymbol{\Sigma}$ og $\boldsymbol{\Sigma}_0$ har fuld rang, p.

Da er aposteriorifordelingen af $\boldsymbol{\mu}$ efter observation af $X = \mathbf{x}$ givet ved

$$\boldsymbol{\mu} \mid X = \mathbf{x} \in \mathrm{N}_p(\mathbf{W}\mathbf{m} + (\mathbf{I} - \mathbf{W})\mathbf{x}, (\mathbf{I} - \mathbf{W})\boldsymbol{\Sigma}) \tag{8.5.1}$$

med

$$\mathbf{W} = \boldsymbol{\Sigma}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma})^{-1} \quad \text{og} \quad \mathbf{I} - \mathbf{W} = \boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma})^{-1}$$

Bevis:
Vi har

$$\begin{aligned} f(\mathbf{x} \mid \boldsymbol{\mu}) &= (\sqrt{2\pi})^{-p} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\right\} \end{aligned}$$

og

$$\begin{aligned} w(\boldsymbol{\mu}) &= (\sqrt{2\pi})^{-p} \det(\boldsymbol{\Sigma}_0)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m})^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\mathbf{m})\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}_0^{-1}\mathbf{m})\right\}, \end{aligned}$$

hvorfor

$$\begin{aligned} h(\boldsymbol{\mu} \mid \mathbf{x}) &\propto \exp\left[-\frac{1}{2}\{\boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\Sigma}_0^{-1}\mathbf{m})\}\right] \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\mathbf{m}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}-\mathbf{m}_1)\right\} \end{aligned}$$

Sætter vi nu $\mathbf{W} = \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1}$, ser vi, at der gælder

$$\begin{aligned} \mathbf{W} &= (\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{\Sigma}_0^{-1} = (\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_0^{-1})^{-1}\boldsymbol{\Sigma}_0^{-1} \\ &= (\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}+\mathbf{I})^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_0^{-1} = (\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1})^{-1} \\ &= \boldsymbol{\Sigma}(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma})^{-1} \end{aligned}$$

Endvidere bemærker vi, at der gælder

$$\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}^{-1}) = (\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}^{-1}) = \mathbf{I}$$

således at koefficientmatricen til $\mathbf{x}$ er

$$\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}^{-1} = \mathbf{I} - \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1} = \mathbf{I} - \mathbf{W}$$

Ved multiplikation til højre med $\boldsymbol{\Sigma}$ finder vi da aposteriorivariansen $\boldsymbol{\Sigma}_1$ som

$$\boldsymbol{\Sigma}_1 = (\mathbf{I}-\mathbf{W})\boldsymbol{\Sigma} \tag{8.5.2}$$

Udtrykket for $\mathbf{I}-\mathbf{W}$ findes ved at bemærke, at

$$\mathbf{I}-\mathbf{W}=(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma})(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma})^{-1}-\boldsymbol{\Sigma}(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma})^{-1}$$

□

Bemærkning 1 : *Vægtmatricerne udtrykt ved den generaliserede varianskvotient*

Lader vi $\boldsymbol{\Gamma}=\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}$ betegne den generaliserede kvotient mellem variansen mellem grupper og variansen inden for grupper, kan vi udtrykke matricerne $\mathbf{W}$ og $\mathbf{I}-\mathbf{W}$ ved

$$\mathbf{W}=(\mathbf{I}+\boldsymbol{\Gamma})^{-1} \quad \text{og} \quad \mathbf{I}-\mathbf{W}=(\mathbf{I}+\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}$$

Man har nemlig

$$\mathbf{W}=\boldsymbol{\Sigma}(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma})^{-1}=\boldsymbol{\Sigma}(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}+\boldsymbol{\Sigma})^{-1}=\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\mathbf{I}+\boldsymbol{\Gamma})^{-1}=(\mathbf{I}+\boldsymbol{\Gamma})^{-1}$$

Udtrykket gælder også, selv om $\boldsymbol{\Sigma}$ ikke har fuld rang.

Vi bemærker specielt, at $(\mathbf{I}+\boldsymbol{\Sigma}_0\boldsymbol{\Sigma})$ er invertibel, også selv om $\boldsymbol{\Sigma}_0$ ikke har fuld rang. Der gælder nemlig for vilkårlige matricer $\mathbf{A}$ og $\mathbf{B}$, at $\det(\mathbf{I}+\mathbf{AB})=\det(\mathbf{I}+\mathbf{BA})$ (Zellner 1971, p.231).

Sættes $\mathbf{A}=\boldsymbol{\Sigma}_0$ og $\mathbf{B}=\boldsymbol{\Sigma}^{-1}$ har vi da, at

$$\det(\mathbf{I}+\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1})=\det(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0+\mathbf{I})=\det(\boldsymbol{\Sigma}^{-1})\det(\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma})$$

Da nu $\boldsymbol{\Sigma}$ er positiv definit, er også $\boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}$ samt $\boldsymbol{\Sigma}^{-1}$ 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 $\boldsymbol{\mu}$ er $X_1, X_2, \ldots, X_n$ uafhængige og identisk fordelte med $X_i \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, og antag endvidere, at $\boldsymbol{\mu} \in \mathrm{N}_p(\mathbf{m}, \boldsymbol{\Sigma}_0)$

Da er aposteriorifordelingen af $\boldsymbol{\mu}$ efter observation af $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ givet ved

$$
\begin{aligned}
\boldsymbol{\mu} \mid x_1, x_2, \ldots, x_n &\in \mathrm{N}_p(\mathbf{W}\mathbf{m} + (\mathbf{I} - \mathbf{W})\overline{\mathbf{x}}_+, \frac{1}{n}(\mathbf{I} - \mathbf{W})\boldsymbol{\Sigma}) \\
\text{hvor} & \\
\mathbf{W} &= \boldsymbol{\Sigma}(n\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma})^{-1} = (\mathbf{I} + n\boldsymbol{\Gamma})^{-1} \\
\mathbf{I} - \mathbf{W} &= n\boldsymbol{\Sigma}_0(n\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma})^{-1} = n(\mathbf{I} + n\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma} \\
\boldsymbol{\Gamma} &= \boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1} \\
\text{og} & \\
\overline{\mathbf{x}}_+ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i
\end{aligned}
$$

Bevis:
Resultatet følger af ovenstående lemma ved at bemærke, at $\overline{X}_+$ er sufficient for $\boldsymbol{\theta}$, og

$$\overline{\mathbf{X}}_+ \mid \boldsymbol{\mu} \in \mathrm{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

□

Eksempel 8.5.1 ***Målefejl for flowmålere***

Vi betragter atter den situation, der blev behandlet i eksempel 6.7.1.

Antag, at fejlvisningen for en måler karakteriseres ved fejlvisningen $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ ved de to flow, 0.1 [m^3 /h] og 0.5 [m^3 /h].

Antag, at fordelingen af fejlvisningen for målerne i en målerpopulation kan beskrives ved en todimensional normalfordeling, $\boldsymbol{\mu} \in \mathrm{N}_2(\mathbf{m}, \boldsymbol{\Sigma}_0)$, hvor

$$\mathbf{m} = \begin{pmatrix} 2.0 \\ 3.0 \end{pmatrix}; \qquad \text{og} \qquad \boldsymbol{\Sigma}_0 = \begin{pmatrix} 3 & 5 \\ 5 & 10 \end{pmatrix}$$

Antag endelig, at usikkerheden (kalibreringsfejlen) ved kalibrering af en måler ved de to flow kan beskrives ved en todimensional normalfordeling, hvis middelværdi er målerens sande fejlvisning, og med dispersionsmatricen

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix},$$

og antag at kalibreringen foretages på en sådan måde, at fejlene ved gentagne kalibreringer er uafhængige.

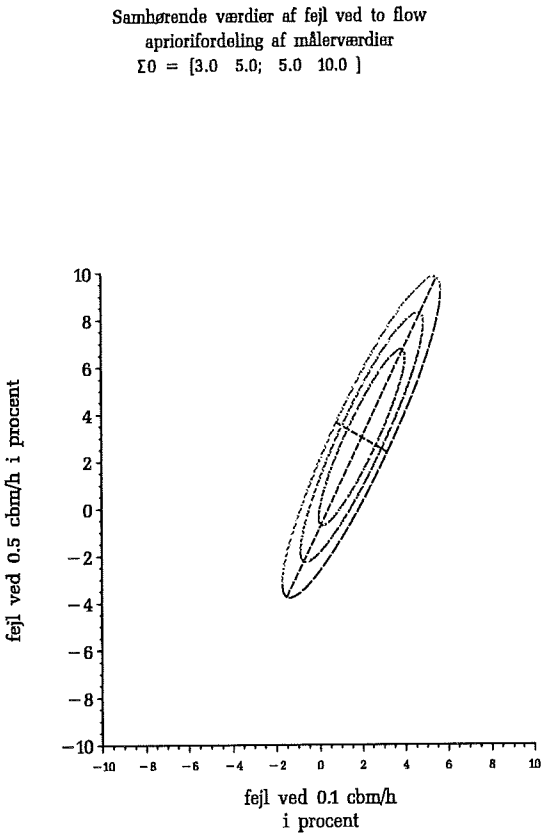
Figur 8.1 viser aprioriforde lingen af målernes fejlvisning, og figur 8.2 viser fordelingen af kalibreringsfejlen for en måler med fejlvisning $(\mu_1, \mu_2) = (0, 0)$.

En måler blev sendt til kalibrering, og der blev foretaget $n = 2$ bestemmelser af målerens fejlvisning. De to sæt kalibreringsresultater er anført nedenfor

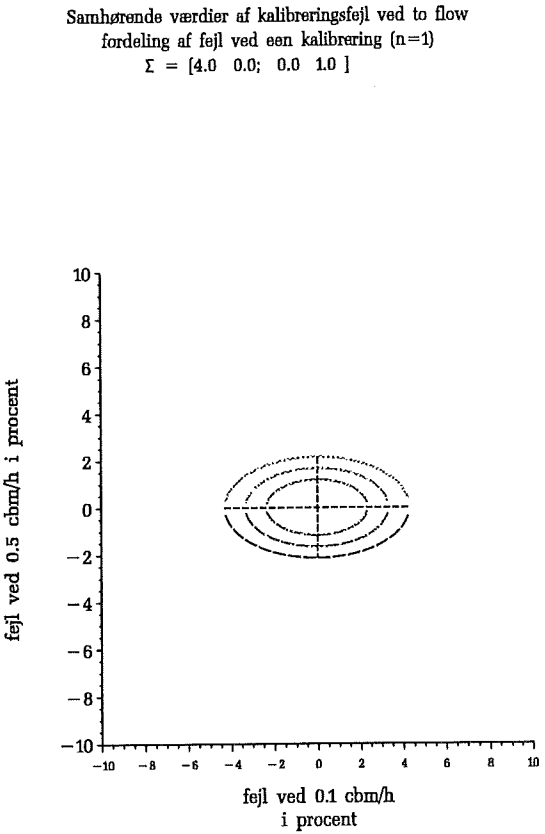
$$\mathbf{x}_1 = \begin{pmatrix} 5.0 \\ 7.0 \end{pmatrix}; \qquad \mathbf{x}_2 = \begin{pmatrix} 1.0 \\ 5.0 \end{pmatrix}$$

Gennemsnittet af de to kalibreringer er således

$$\overline{\mathbf{x}}_+ = \begin{pmatrix} 3.0 \\ 6.0 \end{pmatrix}$$



Figur 8.1. Niveaukurver i fordelingen af samhørende værdier af målerfejl ved to flow for population af målere (apriorifordeling)



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 fejlvisningen i målerpopulationen ved fastlæggelsen af fejlvisningen for den betragtede måler.

Man kan da bestemme aposteriorifordelingen for målerens fejlvisning $\boldsymbol{\mu}$.

Man finder

$$\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 3 & 5 \\ 5 & 10 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} 0.75 & 5.00 \\ 1.25 & 10.00 \end{pmatrix},$$

hvorfor

$$\mathbf{W} = (\mathbf{I} + 2\boldsymbol{\Gamma})^{-1} = \begin{pmatrix} 2,5 & 10.0 \\ 2.5 & 21.0 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} 0.7636 & -0.3636 \\ -0.0909 & 0.0909 \end{pmatrix}$$

og

$$\mathbf{I} - \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0.7636 & -0.3636 \\ -0.0909 & 0.0909 \end{pmatrix}$$

$$= \begin{pmatrix} 0.2364 & 0.3636 \\ 0.0909 & 0.9091 \end{pmatrix}.$$

Aposteriorimiddelværdien for målerens fejlvisning er derfor

$$\mathbf{m}_{\text{apost}} = \mathbf{W} \begin{pmatrix} 2.0 \\ 3.0 \end{pmatrix} + (\mathbf{I} - \mathbf{W}) \begin{pmatrix} 3.0 \\ 6.0 \end{pmatrix}$$

$$= \begin{pmatrix} 0.7636 & -0.3636 \\ -0.0909 & 0.0909 \end{pmatrix} \begin{pmatrix} 2.0 \\ 3.0 \end{pmatrix} + \begin{pmatrix} 0.2364 & 0.3636 \\ 0.0909 & 0.9091 \end{pmatrix} \begin{pmatrix} 3.0 \\ 6.0 \end{pmatrix}$$

$$= \begin{pmatrix} 0.4364 \\ 0.0909 \end{pmatrix} + \begin{pmatrix} 2.8909 \\ 5.7273 \end{pmatrix} = \begin{pmatrix} 3.3273 \\ 5.8182 \end{pmatrix}$$

Vi bemærker, at selv om aposteriorimiddelværdien kan opfattes som et vejet gennemsnit mellem apriorimiddelværdien og stikprøvegennemsnittet, ligger den resulterende værdi ikke nødvendigvis på liniestykket, der forbinder apriorimiddelværdi og stikprøveresultat. Det vejede gennemnsnit tilgodeser også samvariationen mellem fejlene ved de to flow. (Principielt også mellem kalibreringsfejlen 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 aprioriforde lingen (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 stikprøvegennemsnit og aposteriorimiddelværdi. Den resulterende aposteriorimiddelværdi af fejlen ved flow 1 bliver tilmed større end såvel stikprøvegennemsnit som apriorimiddelværdi.

Såfremt vi ønsker et udtryk for usikkerheden på denne angivelse af målerens fejlvisning, kan vi benytte variansen i aposteriorifordelingen af $\boldsymbol{\mu}$. Idet $n = 2$ har vi

$$\mathbf{D}\,[\boldsymbol{\mu}|\bar{\mathbf{x}}] = \frac{1}{2}(\mathbf{I} - \mathbf{W})\boldsymbol{\Sigma} = \frac{1}{2}\begin{pmatrix} 0.2364 & 0.3636 \\ 0.0909 & 0.9091 \end{pmatrix}\begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \frac{1}{2}\begin{pmatrix} 0.9455 & 0.3636 \\ 0.3636 & 0.9091 \end{pmatrix} = \begin{pmatrix} 0.4727 & 0.1818 \\ 0.1818 & 0.4545 \end{pmatrix}$$

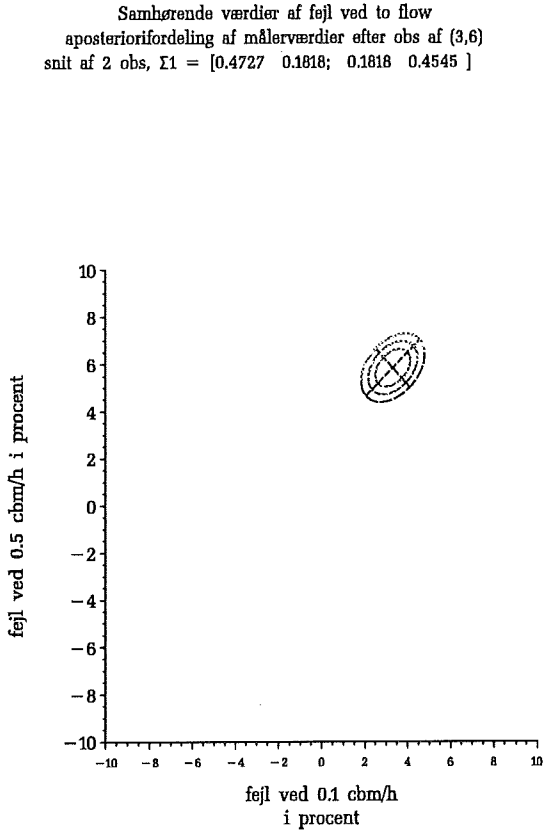
Ved sammenligning med $\boldsymbol{\Sigma}_0$ bemærker vi specielt, at på grund af den store stikprøvepræcision ved flow 2, har vi opnået en kraftig reduktion af usikkerheden vedrørende fejlen ved flow 2 i forhold til aprioriusikkerheden.

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.

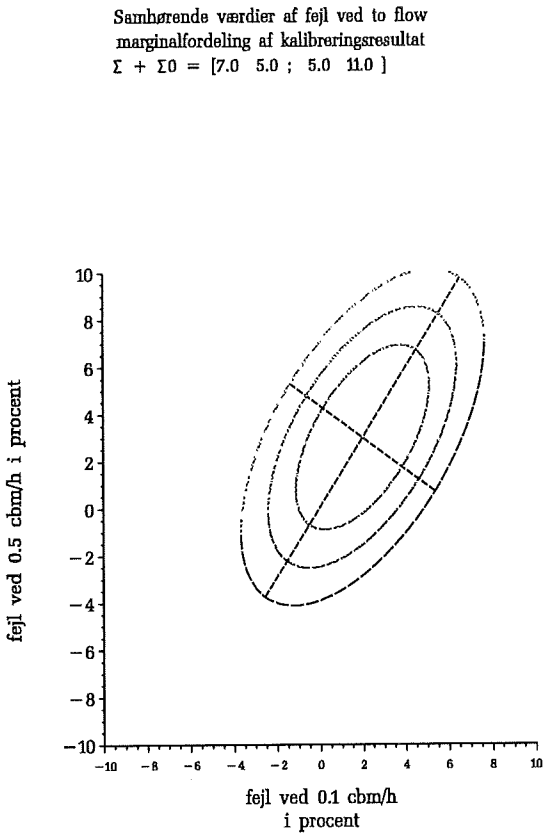
Aposteriorifordelingen for målerens fejlvisning er illustreret i figur 8.3

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.

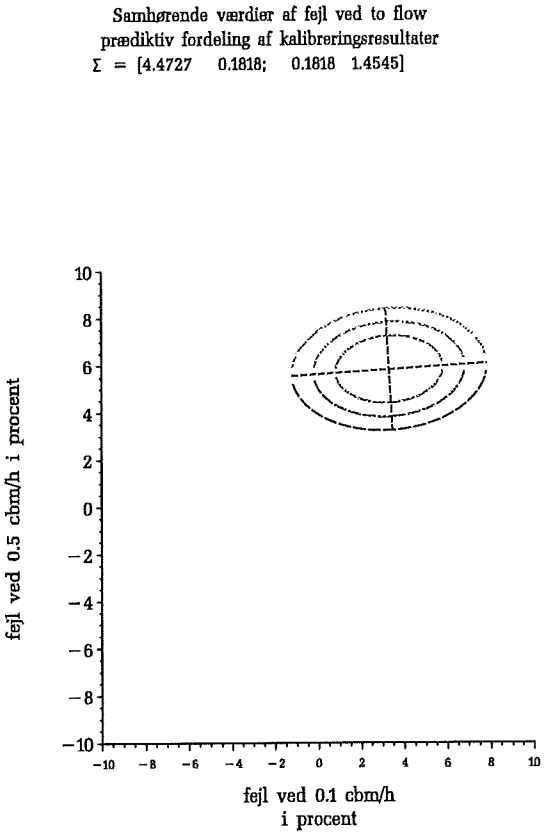
□



Figur 8.3. Niveaukurver for aposteriorifordelingen af samhørende værdier af målerfejl 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 aprioriforde lingen)



Figur 8.5. Niveaukurver i fordelingen af samhørende værdier af kalibreringsresultat ved to flow for den undersøgte 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 $Y \mid \boldsymbol{\beta} \in \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$ og at aprioriforde lingen af $\boldsymbol{\beta}$ er

$$\boldsymbol{\beta} \in \mathrm{N}_p(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Lambda})$$

hvor $\boldsymbol{\Lambda}$ har fuld rang.

Da er aposteriorifordelingen af $\boldsymbol{\beta}$ efter observation af $Y = \mathbf{y}$ givet ved

$$\boldsymbol{\beta} \mid Y = \mathbf{y} \in \mathrm{N}_p(\boldsymbol{\beta}_1, \sigma^2\boldsymbol{\Lambda}_1)$$

hvor

$$\boldsymbol{\beta}_1 = \mathbf{W}\boldsymbol{\beta}_0 + \mathbf{W}\boldsymbol{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{8.6.1}$$

med

$$\mathbf{W} = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}$$

og

$$\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$$

og hvor

$$\boldsymbol{\Lambda}_1 = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda} = \mathbf{W}\boldsymbol{\Lambda} \tag{8.6.2}$$

Bevis:
Vi har

$$f(\mathbf{y} \mid \boldsymbol{\beta}) = (\sqrt{2\pi\sigma^2})^{-n} \det(\mathbf{V})^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y})\right\}$$

og

$$\begin{aligned} w(\beta) &= (\sqrt{2\pi\sigma^2})^{-p}\det(\mathbf{\Lambda})^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\beta-\beta_0)^T\mathbf{\Lambda}^{-1}(\beta-\beta_0)\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\beta^T\mathbf{\Lambda}^{-1}\beta - 2\beta^T\mathbf{\Lambda}^{-1}\beta_0)\right\} \end{aligned}$$

hvorfor

$$\begin{aligned} &h(\beta \mid \mathbf{y}) \propto \\ &\exp\left\{-\frac{1}{2\sigma^2}[\beta^T({}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\mathbf{\Lambda}^{-1})\beta - 2\beta^T({}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}+\mathbf{\Lambda}^{-1}\beta_0)]\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\beta-\beta_1)^T\mathbf{\Lambda}_1^{-1}(\beta-\beta_1)\right\} \end{aligned}$$

hvor

$$\mathbf{\Lambda}_1 = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\mathbf{\Lambda}^{-1})^{-1}$$

og

$$\beta_1 = \mathbf{\Lambda}_1\mathbf{\Lambda}^{-1}\beta_0 + \mathbf{\Lambda}_1\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

Da $\mathbf{\Lambda}$ har fuld rang, gælder der

$$\begin{aligned} \mathbf{\Lambda}_1 &= (\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\mathbf{\Lambda}^{-1})^{-1} = (\mathbf{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\mathbf{I})^{-1}\mathbf{\Lambda} \\ &= (\mathbf{I}+\mathbf{\Gamma})^{-1}\mathbf{\Lambda} \end{aligned}$$

med $\mathbf{\Gamma} = \mathbf{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$,

dvs vi har

$$\beta_1 = (\mathbf{I}+\mathbf{\Gamma})^{-1}\beta_0 + (\mathbf{I}+\mathbf{\Gamma})^{-1}\mathbf{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

□

Bemærkning 1 : ***Invertering af de indgående matricer***

Såfremt $n > p$ er det ofte lettere at benytte

$$\begin{aligned} \beta_1 &= \beta_0 + \mathbf{\Lambda}\mathbf{X}^T(\mathbf{X}\mathbf{\Lambda}\mathbf{X}^T+\mathbf{V})^{-1}(\mathbf{y}-\mathbf{X}\beta_0) \\ \mathbf{\Lambda}_1 &= \mathbf{\Lambda} - \mathbf{\Lambda}\mathbf{X}^T(\mathbf{X}\mathbf{\Lambda}\mathbf{X}^T+\mathbf{V})^{-1}\mathbf{X}\mathbf{\Lambda} \end{aligned}$$

idet disse udtryk kun kræver inversion af $p \times p$ matricer.

□

Bemærkning 2 : *Aposterioriværdien udtrykt som et vægtet gennemsnit*

Såfremt $\mathbf{X}$ har fuld rang, kan $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$ inverteres, og vi får

$$\beta_1 = \mathbf{W}\beta_0 + (\mathbf{I} - \mathbf{W})\hat{\beta}$$

hvor $\hat{\beta}$ angiver den sædvanlige mindste kvadraters estimator (7.2.9) for β,

$$\hat{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

Bevis:
Vi betragter udtrykket

$$\begin{aligned}
\beta_1 &= \mathbf{I} + \mathbf{\Gamma})^{-1}\beta_0 + (\mathbf{I} + \mathbf{\Gamma})^{-1}\mathbf{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\
&= \mathbf{W}\beta_0 + \mathbf{W}\mathbf{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\
&= \mathbf{W}\beta_0 + \mathbf{W}\mathbf{\Lambda}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\
&= \mathbf{W}\beta_0 + (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\mathbf{W}\mathbf{\Lambda}\hat{\beta}
\end{aligned}$$

men idet

$$\mathbf{I} - \mathbf{W} = \mathbf{I} - (\mathbf{I} + \mathbf{\Gamma})^{-1} = (\mathbf{I} + \mathbf{\Gamma})^{-1}\mathbf{\Gamma} = (\mathbf{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \mathbf{I})^{-1}\mathbf{\Lambda}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$$

ser vi, at vi har

$$\mathbf{W}\mathbf{\Lambda}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}) = \mathbf{I} - \mathbf{W}$$

Vi bemærker i øvrigt, at $\mathbf{\Gamma}$ netop udtrykker $\mathsf{E}\,[\,\mathbf{D}\,[\hat{\beta} \mid \beta]](\mathbf{D}\,[\mathsf{E}\,[\hat{\beta} \mid \beta]])^{-1}$, og at

$$\mathbf{I} + \mathbf{\Gamma} = \mathsf{COV}[\hat{\beta}, \beta]\ \mathbf{D}\,[\beta]^{-1}$$

□

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 r yderligere observationer $Y_1', \ldots, Y_r'$ sådan

at sættet $Y_1, \ldots, Y_n$ og sættet $Y_1', \ldots, Y_r'$ er betinget uafhængige af $\boldsymbol{\beta}$ i den simultane fordeling af $Y_1, \ldots, Y_n$, $Y_1', \ldots, Y_r'$ og $\boldsymbol{\beta}$.

Antag, at $Y'|\boldsymbol{\beta} \in \mathrm{N}_r(\mathbf{X}'\boldsymbol{\beta}, \sigma^2\mathbf{V}')$, hvor den $r \times p$ dimensionale modelmatrix $\mathbf{X}'$ er kendt, og hvor den $r \times r$-dimensionale symmetriske, positiv definitte matrix $\mathbf{V}'$ ligeledes er kendt.

Da vil fordelingen af $Y'|Y$ (den prædiktive fordeling) være en $r$dimensional normalfordeling med forventningsværdi

$$\mathsf{E}\,[Y'|Y] = \mathbf{X}'\boldsymbol{\beta}_1 \tag{8.6.3}$$

og dispersionsmatrix

$$\mathbf{D}\,[Y'|Y] = \sigma^2[\mathbf{V}' + \mathbf{X}'\boldsymbol{\Lambda}_1(\mathbf{X}')^T] \tag{8.6.4}$$

hvor $\boldsymbol{\beta}_1$ og $\boldsymbol{\Lambda}_1$ 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

$$\mathsf{E}\,[Y'|Y] = \mathsf{E}_{\,\beta}[\mathsf{E}\,[Y|\boldsymbol{\beta}]|Y] == \mathsf{E}_{\,\beta|Y}[\mathbf{X}'\boldsymbol{\beta}] = \mathbf{X}'\boldsymbol{\beta}_1$$

og

$$\begin{aligned}
\mathbf{D}\,[Y'|Y] &= \mathsf{E}_{\,\beta}[\,\mathbf{D}\,[Y|\boldsymbol{\beta}]|Y] + \mathbf{D}_{\,\beta}[\mathsf{E}\,[Y|\boldsymbol{\beta}]|Y] \\
&= \mathsf{E}_{\,\beta|Y}[\sigma^2\mathbf{V}'] + \mathbf{D}_{\,\beta|Y}[\mathbf{X}'\boldsymbol{\beta}] \\
&= \sigma^2\mathbf{V}' + \mathbf{X}'\,\mathbf{D}\,\beta|Y[\boldsymbol{\beta}](\mathbf{X}')^T
\end{aligned}$$

□

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 β_1 og β_2 er fordelt i populationen i overensstemmelse med en $\mathrm{N}(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{\Lambda})$-fordeling hvor

$$\boldsymbol{\beta}_0 = \begin{pmatrix} 52.7 \\ 1.4 \end{pmatrix}$$

og

$$\boldsymbol{\Lambda} = \begin{pmatrix} 8 & 4 \\ 4 & 6 \end{pmatrix}$$

og hvor $\epsilon \in \mathrm{N}(0, \sigma^2)$ med $\sigma^2 = 0.3^2$ [mm]^2 og ϵ-størrelserne er indbyrdes uafhængige såfremt tidsafstandene er et halvt år eller mere.

Vi betragter først en række udsagn baseret alene på denne aprioriforde ling:

Ramushøjden $Y_{0.25}$ ved 9-års alderen for en tilfældigt udvalgt dreng kan beskrives ved en normalfordelt variabel med middelværdi

$$\mathsf{E}\,[Y_{0.25}] = 52.7 + 0.25 \times 1.4 = 52.7 + 0.35 = 53.05 \text{ [mm]}$$

og varians

$$\begin{aligned} \mathsf{V}\,[Y] &= \mathsf{E}\,[\mathsf{V}\,[Y|\boldsymbol{\beta}]] + \mathsf{V}\,[\mathsf{E}\,[Y|\boldsymbol{\beta}]] \\ &= \sigma^2(1 + \mathbf{x}\boldsymbol{\Lambda}\mathbf{x}^T) \\ &= \sigma^2(1 + 10.3750) = 1.0238 = (1.0118)^2 \text{ [mm]}^2 \end{aligned}$$

idet vi har benyttet at

$$\mathbf{D}\,[\mathbf{x}\boldsymbol{\beta}] = \mathbf{x}\;\mathbf{D}\,[\boldsymbol{\beta}]\mathbf{x}^T$$

hvor modelmatricen

$$\mathbf{x} = (1 \quad 0.25)$$

og $\mathbf{D}\,[\boldsymbol{\beta}] = \sigma^2\boldsymbol{\Lambda}$.

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

$$\mathsf{E}\,[Y] = \mathbf{X}\boldsymbol{\beta}_0 = \begin{pmatrix} 52.7 + 0.35 \\ 52.7 + 1.05 \end{pmatrix} = \begin{pmatrix} 53.05 \\ 53.75 \end{pmatrix}$$

hvor modelmatricen $\mathbf{X}$ er

$$\mathbf{X} = \begin{pmatrix} 1 & 0.25 \\ 1 & 0.75 \end{pmatrix}$$

Dispersionsmatricen for Y er

$$\begin{aligned} \mathsf{D}[Y] &= \sigma^2(\mathbf{I}_2 + \mathbf{X}\Lambda\mathbf{X}^T) \\ &= \sigma^2 \begin{pmatrix} 11.375 & 13.125 \\ 13.125 & 18.375 \end{pmatrix} \\ &= \begin{pmatrix} 1.0238 & 1.1813 \\ 1.1813 & 1.6538 \end{pmatrix} \end{aligned}$$

Endelig finder man for tilvæksten T fra alderen 9 år til 9 1/2 år:

$$T = Y_{0.75} - Y_{0.25} = (-1 \quad 1)Y$$

at

$$\mathsf{E}[T] = (-1 \quad 1)\mathbf{X}\boldsymbol{\beta}_0 = 0.50\beta_2 = 0.70 \text{ [mm]}$$

og

$$\begin{aligned} \mathsf{V}[T] &= (-1 \quad 1)\, \mathsf{D}[Y](-1 \quad 1)^T \\ &= \sigma^2 \times 1.5 = 0.135 = (0.367)^2 \text{ [mm]}^2 \end{aligned}$$

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 modelmatricen

$$\mathbf{X} = \begin{pmatrix} 1 & -0.75 \\ 1 & -0.25 \end{pmatrix}$$

Idet

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 2 & -1 \\ -1 & 0.6250 \end{pmatrix}$$

har man estimatet $\widehat{\boldsymbol{\beta}}$ for regressionskoefficienterne for denne dreng:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y \\ &= \begin{pmatrix} 2 & -1 \\ -1 & 0.6250 \end{pmatrix}^{-1} \begin{pmatrix} 105.80 \\ -52.75 \end{pmatrix} = \begin{pmatrix} 2.5 & 4.00 \\ 4.00 & 8.00 \end{pmatrix} \\ &= \begin{pmatrix} 53.50 \\ 1.20 \end{pmatrix}\end{aligned}$$

Endvidere har man

$$\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 12 & -5.50 \\ 2 & -0.25 \end{pmatrix}$$

og

$$\begin{aligned}\mathbf{W} &= (\mathbf{I}_2 + \boldsymbol{\Gamma})^{-1} \\ &= \begin{pmatrix} 13 & -5.50 \\ 2 & 0.75 \end{pmatrix}^{-1} = \begin{pmatrix} 0.0361 & 0.2651 \\ -0.0964 & 0.6265 \end{pmatrix}\end{aligned}$$

således at dispersionsmatricen for aposteriorifordelingen af $\boldsymbol{\beta}$ er

$$\sigma^2\boldsymbol{\Lambda}_1 = \sigma^2\mathbf{W}\boldsymbol{\Lambda} = \sigma^2 \begin{pmatrix} 1.3494 & 1.7349 \\ 1.7349 & 3.3733 \end{pmatrix}$$

Man får nu aposteriorimiddelværdien af $\boldsymbol{\beta}$

$$\begin{aligned}\boldsymbol{\beta}_1 &= \mathbf{W}\boldsymbol{\beta}_0 + (\mathbf{I} - \mathbf{W})\widehat{\boldsymbol{\beta}} \\ &= \begin{pmatrix} 2.2759 \\ -4.2024 \end{pmatrix} + \begin{pmatrix} 51.2482 \\ 5.6048 \end{pmatrix} = \begin{pmatrix} 53.5241 \\ 1.4024 \end{pmatrix}\end{aligned}$$

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

$$\mathsf{E}\,[Y'|Y] = \mathbf{X}\boldsymbol{\beta}_1 = \begin{pmatrix} 53.5241 + 0.3506 \\ 53.5241 + 1.0518 \end{pmatrix} = \begin{pmatrix} 53.8747 \\ 54.5759 \end{pmatrix}$$

idet modelmatricen $\mathbf{X}$ er

$$\mathbf{X} = \begin{pmatrix} 1 & 0.25 \\ 1 & 0.75 \end{pmatrix}$$

Dispersionsmatricen for Y' er

$$\begin{aligned} \mathbf{D}[Y'|Y] &= \sigma^2(\mathbf{I}_2 + \mathbf{X}\boldsymbol{\Lambda}_1\mathbf{X}^T) \\ &= \sigma^2 \begin{pmatrix} 3.4277 & 3.7169 \\ 3.7169 & 6.8494 \end{pmatrix} \\ &= \begin{pmatrix} 0.3085 & 0.3345 \\ 0.3345 & 0.6164 \end{pmatrix} \end{aligned}$$

Man finder derfor specielt, at prædiktionen af ramushøjden ved 9-års alderen har forventningsværdien

$$\mathsf{E}[Y'_{0.25}|Y] = 53.8747 \text{ [mm]}$$

og variansen

$$\mathsf{V}[Y'_{0.25}|Y] = 0.3085 = (0.555)^2 \text{ [mm]}^2$$

Endvidere har den prædiktive værdi T' for for tilvæksten fra alderen 9 år til 9 1/2 år forventningsværdien

$$\mathsf{E}[T'|Y] = (-1 \quad 1)\mathbf{X}\boldsymbol{\beta}_1 = 0.50 \times 1.4024 = 0.7012 \text{ [mm]}$$

og variansen

$$\begin{aligned} \mathsf{V}[T'|Y] &= (-1 \quad 1)\,\mathbf{D}[Y'|Y](-1 \quad 1)^T \\ &= \sigma^2 \times 0.8434 = 0.0759 = (0.2755)^2 \text{ [mm]}^2 \end{aligned}$$

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.

□

8.7 Tidsrækkemodeller

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 $\beta \in \mathrm{N}(\beta_0, \sigma_0^2)$ og $\epsilon_t \in \mathrm{N}(0, \sigma^2)$ er indbyrdes uafhængige. Da er aposteriorifordelingen af β efter observation af $\mathbf{y} = (y_0, y_1, \ldots, y_n)^T$ en normalfordeling med

$$\begin{aligned} \mathsf{E}\,[\beta \mid \mathbf{y}] &= w\beta_0 + (1-w)\hat{\beta} \\ \mathsf{V}\,[\beta \mid \mathbf{y}] &= 1 - w \end{aligned}$$

hvor

$$w = \frac{1}{1 + (\mathbf{y}_{-1}^T \mathbf{y}_{-1})\gamma}$$

med $\gamma = \sigma_0^2/\sigma^2$ og

$$\hat{\beta} = (\mathbf{y}_{-1}^T \mathbf{y}_{-1})^{-1} \mathbf{y}_{-1}^T \mathbf{y}$$

Bevis:
Resultatet vises ved at betragte likelihoodfunktionen for β.

□

Bayesløsningen for den generelle flerdimensionale tidsrækkemodel, der er beskrevet i afsnit 7.3.2 er givet ved:

Sætning 8.7.2 *Bayesløsning for den flerdimensionale tidsrækkemodel*

For modellen givet ved (7.3.8) og (7.3.9) med $\boldsymbol{\beta} \in \mathrm{N}_{(p+q)\times q}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ og $\boldsymbol{\epsilon} \in \mathrm{N}_{(p+q)\times q}(\mathbf{0}, \boldsymbol{\Sigma})$ gælder

$$\mathsf{E}\,[\boldsymbol{\beta} \mid \mathbf{y}] = \mathbf{W}\boldsymbol{\beta}_0 + (\mathbf{I} - \mathbf{W})\hat{\boldsymbol{\beta}} \tag{8.7.1}$$

hvor den $(p+q)p \times (p+q)p$ dimensionale matrix $\mathbf{W}$ er bestemt ved

a): for $\boldsymbol{\Sigma} = \sigma^2\mathbf{V}$:

$$\mathbf{W} = (\boldsymbol{\Sigma}_0^{-1} + \sigma^{-2}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\boldsymbol{\Sigma}_0^{-1} \tag{8.7.2}$$

med $\mathbf{X} = \mathbf{I}_p \otimes \mathbf{M}$, $\mathbf{M} = (\mathbf{Y}_{-1}, \mathbf{Z})$ og $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$

b) for $\boldsymbol{\Sigma}$ af formen $\boldsymbol{\Sigma} \otimes \mathbf{V}$

$$\mathbf{W} = (\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}^{-1}\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\boldsymbol{\Sigma}_0^{-1} \tag{8.7.3}$$

med

$$\mathbf{M} = (\mathbf{Y}_{-1}, \mathbf{Z})$$

og

$$\widehat{\boldsymbol{\beta}} = \mathrm{vec}(\widehat{\mathbf{B}})$$

hvor

$$\hat{\mathbf{B}} = (\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{V}^{-1}\mathbf{Y}$$

Bevis:
Se f.ex. J.Bellach : Bayessche Schätzungen und Vorhersagen bei stochastischen linearen Differenzengleichungssystemen. *Math Operationsforsch. Statist.* 5 (1974) pp 599-623.

□

Indeks