FORDELINGER MED ANVENDELSER I STATISTIK

Poul Thyregod

LYNGBY 1998

IMM

Forord

Nærværende oversigt er udarbejdet som et supplement til noterne i faget Statistik 3 ved IMM.

Det har været hensigten at supplere den gennemgang af fundamentale fordelinger i statistikken, der er i Introduktion til Statistik, Bind 1, med en række af de øvrige fordelinger, man også møder i de statistiske anvendelser, herunder de forskellige compound fordelinger, der indgår i Statistik 3 kurset. Endvidere behandles en række forskellige levetidsfordelinger, der benyttes i pålidelighedsteorien.

Der er sket en udvikling indenfor de matematisk-statistiske teoridannelser siden udarbejdelsen af Introduktion til Statistik, Bind 1. Specielt er teorierne omkring de eksponentielle familier og de eksponentielle dispersionsparameterfamilier blevet væsentligt mere afklarede. Nærværende oversigt medtager derfor også de fordelinger, der er beskrevet i Introduktion til Statistik, Bind 1. Ved beskrivelsen af disse fordelinger er det tilstræbt ikke kun at opremse egenskaberne ved fordelingerne, men også at placere dem i en større sammenhæng, specielt som medlemmer af eksponentielle familier og eksponentielle dispersionsmodeller.

Oversigten er desværre endnu ganske ufuldkommen, således er de senere års resultater omkring approksimationer ikke medtaget, men det er mit håb at selv denne foreløbige udgave kan være til nytte som opslagsværk. Jeg modtager selvsagt meget gerne rettelser og andre kommentarer til denne præmature udgave.

Lyngby januar 1998
Poul Thyregod

Indhold

Afsnit 0

Lidt om momenter og flerdimensionale variable

0.1 Momenter og flerdimensionale stokastiske variable

Fil: /tex/stat3/fordbog/fordafsn0a.tex 1998-01-12

0.1.1 Momenter

Definition 0.1.1 ***Betingede og marginale momenter***
Lad X og Y være stokastiske variable med den simultane tæthed $f(x,y)$ (med hensyn til målet $\mu \times \nu$ på $\mathcal{X} \times \mathcal{Y}$)

Ved den marginale middelværdi af $\phi(X)$ vil vi forstå værdien af integralet

$$\mathsf{E}[\phi(X)] \stackrel{\text{DEF}}{=} \int_{\mathcal{X}\times\mathcal{Y}} \phi(x) f(x,y) \mu\{dx\}\nu\{dy\} = \int_{\mathcal{X}} \phi(x) k_X(x) \mu\{dx\}$$

hvor $k_X(x)$ angiver den marginale tæthed for X.

Ved den betingede middelværdi af $\phi(X)$ for givet $Y = y$ vil vi forstå den stokastiske variabel, der for $Y = y$ antager værdien

$$\int_{\mathcal{X}} \phi(x)h(x|y)\mu\{dx\}$$

Kort skriver vi

$$\mathsf{E}[\phi(X)|Y = y] \stackrel{\text{DEF}}{=} \int_{\mathcal{X}} \phi(x)h(x|y)\mu\{dx\} \tag{0.1.1}$$

□

Sætning 0.1.1 ***Relation mellem marginale og betingede momenter***

Lad X, Y og Z være endimensionale stokastiske variable. Da gælder:

$$\begin{aligned} \mathsf{E}[X] &= \mathsf{E}_Y[\,\mathsf{E}[\,X|Y]\,] \\ \mathsf{V}[X] &= \mathsf{E}_Y[\,\mathsf{V}[\,X|Y]\,] + \mathsf{V}_Y[\,\mathsf{E}[\,X|Y]\,] \\ \mathsf{COV}[X,Z] &= \mathsf{E}_Y[\,\mathsf{COV}[X,Z]\,|Y\,] + \mathsf{COV}_Y[\,\mathsf{E}[\,X|Y], \mathsf{E}[\,Z|Y]\,] \end{aligned} \tag{0.1.2}$$

Bevis:
Sætningen vises ved opstilling af integralerne og brug af en variant af Fubini's sætning. □

Bemærkning 1 ***Fortolkning ved tyngdepunkter og inertimomenter***

Opfatter man forventningsværdien af en stokastisk variabel som *tyngdepunktet* i en massefordeling svarende til fordelingen af sandsynlighedsmassen ser man, at sætningen vedrørende den marginale forventningsværdi er analog til den fra mekaniken kendte sætning om bestemmelse af tyngdepunktet for et sammensat legeme. Opfatter man tilsvarende variansen for en stokastisk variabel som *inertimomentet* omkring (en akse gennem) tyngdepunktet ser man, at sætningen vedrørende den marginale varians svarer til at inertimomentet for at sammensat legeme bestemmes som summen

af inertimomenterne omkring de lokale tyngdepunkter plus inertimomentet om det fælles tyngdepunkt med de lokale masser samlet i deres tyngdepunkt (Steiner's sætning). □

0.1.2 Momenter for flerdimensionale variable

Vi indleder med et resultat fra den lineære algebra:

Sætning 0.1.2 ***Invertering af matrix opdelt i blokmatricer***

Lad $\mathbf{\Sigma}$ være en symmetrisk matrix med opspaltningen i blokmatricer:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} .$$

Da gælder

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{A}^T \\ -\mathbf{A}\mathbf{B}^{-1} & \mathbf{\Sigma}_{22}{}^{-1} + \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^T \end{pmatrix} ,$$

hvor

$$\mathbf{A} = \mathbf{\Sigma}_{22}{}^{-1}\mathbf{\Sigma}_{21}{}^{-1} ,$$

og $\mathbf{B}$ er Schur komplementet til $\mathbf{\Sigma}_{22}$,

$$\mathbf{B} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}{}^{-1}\mathbf{\Sigma}_{21}^T .$$

Bevis:

Resultatet følger umiddelbart ved multiplikation af $\mathbf{\Sigma}$ og $\mathbf{\Sigma}^{-1}$. □

For en p-dimensional stokastisk variabel

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

definerer vi forventningsværdien $\mathsf{E}[\,X]$ som vektoren

$$\mathsf{E}[\,X] = \begin{pmatrix} \mathsf{E}[X_1] \\ \mathsf{E}[X_2] \\ \vdots \\ \mathsf{E}[X_p] \end{pmatrix} \tag{0.1.3}$$

Forventningsværdien er lineær, dvs. såfremt X og Y er p-dimensionale stokastiske vektorer, da er

$$\begin{aligned} \mathsf{E}[X+Y] &= \mathsf{E}[X] + \mathsf{E}[Y] \\ \mathsf{E}[aX] &= a\mathsf{E}[X] \end{aligned}$$

Såfremt $\mathbf{A}$ er en $n \times p$ matrix af konstanter, da gælder

$$\mathsf{E}[\mathbf{A}X] = \mathbf{A}\,\mathsf{E}[X]$$

Definition 0.1.2 ***Kovarians mellem flerdimensionale observationer***
Kovariansen $\mathbf{COV}[X,Y]$ mellem den p-dimensionale stokastiske vektor X og den q-dimensionale stokastiske vektor Y defineres som

$$\mathbf{COV}[X,Y] \stackrel{\text{DEF}}{=} \mathsf{E}[\,(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])^T\,] \tag{0.1.4}$$

Udtrykt ved kovarianserne for de endimensionale variable $X_1, \ldots, X_p$ og $Y_1, \ldots, Y_q$ har vi

$$\mathbf{COV}[X,Y] = \begin{pmatrix} \mathsf{COV}[X_1,Y_1] & \ldots & \mathsf{COV}[X_1,Y_q] \\ \mathsf{COV}[X_2,Y_1] & \ldots & \mathsf{COV}[X_2,Y_q] \\ \vdots & \ddots & \vdots \\ \mathsf{COV}[X_p,Y_1] & \ldots & \mathsf{COV}[X_p,Y_q] \end{pmatrix}$$

□

Der gælder

Sætning 0.1.3 ***Kovariansen er en symmetrisk, bilinear form***

Kovariansen er lineær i hvert af sine to argumenter: Lad X og Y være p-dimensionale og lad U og V være q-dimensionale stokastiske vektorer. Da gælder

$$\begin{aligned} \mathbf{COV}[X+Y,U] &= \mathbf{COV}[X,U] + \mathbf{COV}[Y,U] \\ \mathbf{COV}[X,U+V] &= \mathbf{COV}[X,U] + \mathbf{COV}[X,V] \\ \mathbf{COV}[aX,U] &= a\,\mathbf{COV}[X,U] \end{aligned}$$

Såfremt $\mathbf{A}$ er en $n \times p$ og $\mathbf{B}$ en $m \times q$ matrix af konstanter gælder

$$\mathbf{COV}[\mathbf{A}X, \mathbf{B}U] = \mathbf{A}\ \mathbf{COV}[X,U]\mathbf{B}^T \tag{0.1.5}$$

Ydermere gælder symmetrirelationen

$$\mathbf{COV}[X,U] = (\ \mathbf{COV}[U,X])^T$$

Bevis:
Bevises direkte ved brug af definitionen □

Definition 0.1.3 ***Dispersionsmatricen for en flerdimensional stokastisk variabel***

For en p-dimensional stokastisk variabel X indfører vi dispersionsmatricen $\mathbf{D}[X]$ ved

$$\mathbf{D}[X] \stackrel{\text{DEF}}{=} \mathbf{COV}[X,X] = \mathsf{E}[\,(X - \mathsf{E}[X])(X - \mathsf{E}[X])^T\,] \tag{0.1.6}$$

Udtrykt ved momenterne for de endimensionale variable $X_1, X_2, \ldots, X_n$ har vi

$$\mathbf{D}[X] = \begin{pmatrix} \mathsf{V}[X_1] & \mathsf{COV}[X_1,X_2] & \ldots & \mathsf{COV}[X_1,X_p] \\ \mathsf{COV}[X_2,X_1] & \mathsf{V}[X_2] & \ldots & \mathsf{COV}[X_2,X_p] \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{COV}[X_p,X_1] & \mathsf{COV}[X_p,X_2] & \ldots & \mathsf{V}[X_p] \end{pmatrix}$$

Såfremt $\mathbf{A}$ er en $n \times p$ matrix af konstanter følger det af egenskaberne for kovariansen at

$$\mathbf{D}[\mathbf{A}X] \;=\; \mathbf{A}\;\mathbf{D}[X]\,\mathbf{A}^T \tag{0.1.7}$$

endvidere har vi, såfremt X og Y begge er p-dimensionale stokastiske variable:

$$\mathbf{D}[X+Y] = \mathbf{D}[X] + \mathbf{D}[Y] + \mathbf{COV}[X,Y] + \mathbf{COV}[Y,X]$$

□

Sætning 0.1.4 ***Dispersionsmatricen er ikke-negativ definit***

Lad X angive en p-dimensional stokastisk variabel. Da gælder at dispersionsmatricen $\mathbf{D}[X]$ er en en ikke-negativ definit symmetrisk matrix.

For flerdimensionale variable X, Y og Z gælder de analoge relationer til (0.1.2):

$$\begin{aligned} \mathsf{E}[X] &= \mathsf{E}_Y[\,\mathsf{E}[X|Y]\,] \\ \mathbf{D}[X] &= \mathsf{E}_Y[\,\mathbf{D}[X|Y]\,] + \mathbf{D}_Y[\,\mathsf{E}[X|Y]\,] \\ \mathbf{COV}[X,Z] &= \mathsf{E}_Y[\,\mathbf{COV}[X,Z|Y]\,] + \mathbf{COV}_Y[\,\mathsf{E}[X|Y],\,\mathsf{E}[Z|Y]\,] \end{aligned} \tag{0.1.8}$$

Bevis:
Følger i lighed med beviset for sætning 0.1.1. □

Definition 0.1.4 ***Partiel varians og kovarians***

Lad X, Y og Z angive (evt flerdimensionale) stokastiske variable sådan at $\mathbf{D}[X]$ har fuld rang.

Ved den partielle kovarians af Y og Z givet X vil vi forstå den bilineære form (bilineær i Y og Z)

$$\mathbf{COV}[Y,Z:X] \stackrel{\text{DEF}}{=} \mathbf{COV}[Y,Z] - \mathbf{COV}[Y,X]\;\mathbf{D}[X]^{-1}\;\mathbf{COV}[X,Z] \tag{0.1.9}$$

Ved den partielle varians af Y givet X vil vi specielt forstå

$$\mathbf{D}[Y:X] \stackrel{\text{DEF}}{=} \mathbf{COV}[Y,Y:X] = \mathbf{D}[Y] - \mathbf{COV}[Y,X]\,\mathbf{D}[X]^{-1}\,\mathbf{COV}[X,Y] \tag{0.1.10}$$

For endimensionale variable X og Y finder vi altså specielt

$$\mathsf{V}[Y:X] = \mathsf{V}[Y] - \mathsf{COV}[Y,X]^2/\mathsf{V}[X]$$

Den partielle varians af Y givet X er Schur komplementet til $\mathbf{D}[X]$ i

$$\mathbf{D}\begin{bmatrix} X \\ Y \end{bmatrix}$$

□

Vi bemærker, at mens den betingede kovarians $\mathbf{COV}[Y,Z|X=x]$ afhænger af den aktuelle værdi x af X, da afhænger den partielle kovarians $\mathbf{COV}[Y,Z:X]$ kun af momenterne i fordelingen af X m.v.

Sætning 0.1.5 ***Den partielle kovarians*** $\mathbf{COV}[Y,Z:X]$ ***er invariant overfor lineære transformationer af*** $\mathbf{X}$.

Lad X være en p-dimensional variabel, og lad $\mathbf{A}$ være en $p \times p$ matrix af konstanter sådan at $\mathbf{A}$ har fuld rang.

Da gælder

$$\mathbf{COV}[Y,Z:\mathbf{A}X] = \mathbf{COV}[Y,Z:X] \tag{0.1.11}$$

Bevis:
Følger direkte □

0.2 Equikorrelerede observationer

0.2.1 Equikorrelationsmatricer

Lad $\mathbf{E}$ være defineret ved

$$\mathbf{E} = (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p$$

hvor $\mathbf{I}_p$ angiver den $p \times p$-dimensionale enhedsmatrix, og $\mathbf{J}_p = \mathbf{1}_p \mathbf{1}_p{}^T$ angiver den $p \times p$-dimensionale matrix med lutter ettaller.

Der gælder $e_{ii} = 1$ og $e_{ij} = \rho$ for $i \neq j$.

For $-(1-p)^{-1} < \rho < 1$ gælder:

$$\mathbf{E}^{-1} = (1-\rho)^{-1}[\mathbf{I}_p - \rho\{1+(p-1)\rho\}^{-1}\mathbf{J}_p]$$

og

$$\det(\mathbf{E}) = (1-\rho)^{p-1}\{1+\rho(p-1)\}$$

0.2.2 Centrerende matricer

Lad $\mathbf{H}$ være den $n \times n$-dimensionale matrix defineret ved

$$\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$$

hvor $\mathbf{I}_n$ og $\mathbf{J}_n$ er som ovenfor, og lad x være en n-dimensional søjlevektor. Da gælder:

$$\begin{aligned} \mathbf{H}^T &= \mathbf{H}; \quad \mathbf{H}^2 = \mathbf{H} \\ \mathbf{H1} &= \mathbf{0}; \quad \mathbf{HJ}_n = \mathbf{J}_n\mathbf{H} = \mathbf{0} \\ \mathbf{Hx} &= \mathbf{x} - \bar{x}\mathbf{1} \quad \text{med} \quad \bar{x} = \sum_1^n x_i/n \end{aligned}$$

$$\mathbf{x}^T\mathbf{H}\mathbf{x} = \sum_1^n (x_i - \bar{x})^2 \tag{0.2.1}$$

0.2.3 Fordelingen af gennemsnit af equikorrelerede observationer

Sætning 0.2.1 *Momenter for gennemsnit af equikorrelerede observationer*

Lad $\mathbf{X}$ angive en n-dimensional vektor af equikorrelerede observationer, dvs.

$$\mathsf{E}[\mathbf{X}] = \mu\, 1_n; \qquad \mathbf{D}[\mathbf{X}] = \sigma^2[(1-\rho)\mathbf{I}_n + \rho\mathbf{J}_n]$$

med $0 \leq \rho \leq 1$, og lad $\overline{X}_+$ angive gennemsnittet af disse observationer, $\overline{X}_+ = \sum_1^n X_i$.

Da gælder

$$\mathsf{E}[\overline{X}_+] = \mu; \qquad \text{og} \qquad \mathsf{V}[\overline{X}_+] = \sigma^2 \left[\rho + \frac{1}{n}(1-\rho)\right] \tag{0.2.2}$$

Bevis:
Betragt summen $S = \sum_1^n X_i$. Vi har

$$S = \mathbf{1}_n^T \mathbf{X}$$

Det følger da af 0.1.7 at

$$\begin{aligned} \mathsf{E}[S] &= \mathbf{1}_n^T \mathsf{E}[\mathbf{X}] = \mu \mathbf{1}_n^T \mathbf{1}_n = n\mu \\ \mathsf{V}[S] &= \mathbf{1}_n^T\, \mathbf{D}[\mathbf{X}] \mathbf{1}_n = \sigma^2[(1-\rho)\mathbf{1}_n^T \mathbf{I}_n \mathbf{1}_n + \rho \mathbf{1}_n^T \mathbf{E}_n \mathbf{1}_n] \\ &= \sigma^2[n(1-\rho) + n^2\rho] = \sigma^2[n + n(n-1)\rho] \end{aligned}$$

hvorved sætningen følger □

Lemma 0.2.1 ***Forventningsværdi for kvadratafvigelsessum omkring vægtet gennemsnit***

Lad $Y_1, Y_2, \ldots, Y_k$ være uafhængige observationer med $\mathsf{E}[Y_i] = \mu$, og lad

$$\overline{Y}_+ = \sum_1^k \alpha_i Y_i$$

hvor $0 \leq \alpha_i \leq 1$; $i = 1, 2, \ldots, k$ er givne vægte med $\sum \alpha_i = 1$.

Betragt den vægtede kvadratafvigelsessum

$$Z = \sum_1^k \alpha_i (Y_i - \overline{Y}_+)^2$$

Da gælder

$$\mathsf{E}[\,Z] = \sum_1^k \alpha_i(1-\alpha_i)\;\mathsf{V}[Y_i] \tag{0.2.3}$$

Bevis:
Beviset følger ved at betragte opspaltningen

$$\sum_1^k \alpha_i(Y_i-\mu)^2 = \sum_1^k \alpha_i(Y_i-\overline{Y}_+)^2 + (\overline{Y}_+-\mu)^2 \sum_1^k \alpha_i = Z + (\overline{Y}_+-\mu)^2$$

□

Betragter vi specielt k uafhængige stikprøver, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$, hvor $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})^T$ angiver en n_i-dimensional observationsvektor af equikorrelerede observationer,

$$\mathsf{E}[\mathbf{X}_i] = \mu\,\mathbf{1}_{n_i}; \qquad \mathbf{D}[\mathbf{X}] = \sigma^2[(1-\rho)\mathbf{I}_{n_i} + \rho\mathbf{J}_{n_i}]$$

med $0 \le \rho \le 1$, og lader vi $\overline{X}_{i\cdot}$ angive gennemsnittet af observationerne i den i'te stikprøve,

$$\overline{X}_{i\cdot} = \sum_1^{n_i} X_{ij}/n_i\ ,$$

og tilsvarende $\overline{X}_{\cdot\cdot}$ angive gennemsnittet af samtlige observationer,

$$\overline{X}_{\cdot\cdot} = \sum_i \sum_j X_{ij} / \sum_i n_i = \sum_i n_i \overline{X}_{i\cdot} / \sum_i n_i\ ,$$

da gælder for

$$\mathrm{SAK} = \sum_1^k n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

$$\mathsf{E}[\,\mathrm{SAK}]/(k-1) = \sigma^2[1+(n_0-1)\rho] \tag{0.2.4}$$

hvor

$$n_0 = \frac{\sum_1^k n_i - (\sum_1^k n_i^2 / \sum_1^k n_i)}{k-1} \tag{0.2.5}$$

Bevis:
Følger af Lemma 0.2.1 □

Bemærkning 1 ***Den vægtede gennemsnitlige gruppestikprøvestørrelse***

Størrelsen n_0 bestemt ved (0.2.5) betegnes ofte den vægtede gennemsnitlige gruppestikprøvestørrelse. Såfremt alle grupper er lige store, $n_1 = n_2 = \cdots = n_k = n$, finder man

$$n_0 = n \ ,$$

den fælles stikprøvestørrelse.

Almindeligt gælder

$$n_0 = \overline{n}\left[1 - \frac{k}{k-1}\sum_1^k \left(\frac{n_i}{n_+} - \frac{1}{k}\right)^2\right] \ ,$$

hvor $\overline{n} = n_+/k$ med $n_+ = \sum_i n_i$ angiver den gennemsnitlige stikprøvestørrelse for de k grupper.

Størrelsen,

$$\sum_1^k \left(\frac{n_i}{n_+} - \frac{1}{k}\right)^2 \ ,$$

i udtrykket for n_0 måler afvigelsen mellem den aktuelle fordeling af det totale stikprøveomfang n_+ på de k grupper og ligefordelingen på grupperne. Jo mere den aktuelle fordeling af stikprøveomfanget afviger fra en ligefordeling, desto mere skal det sædvanlige aritmetiske gennemsnit af stikprøvestørrelserne formindskes for at opnå den vægtede gennemsnitlige stikprøvestørrelse, n_0. □

0.3 Karakteristiske funktioner og kumulanter

Vi minder om definitionen på den karakteristiske funktion for en stokastisk variabel (jvf Conradsen (1997))

Definition 0.3.1 ***Karakteristisk funktion***

Ved den karakteristiske funktion for fordelingen for en stokastisk variabel X forstås funktionen

$$\phi_X(t) = \mathsf{E}[\exp(\mathrm{i}tX)] \quad (= \mathsf{E}[\cos(tX)] + \mathrm{i}\,\mathsf{E}[\sin(tX)]) \tag{0.3.1}$$

hvor i angiver den imaginære enhed i det komplekse tallegeme.

Såfremt X er k-dimensional, defineres den karakteristiske funktion som en funktion på $\mathbb{R}^k$, hvor produktet tX i (0.3.1) erstattes med det indre produkt $t^T X$. □

Der gælder

Sætning 0.3.1 ***Karakteristisk funktion for affin transformation*** Lad fordelingen for den stokastiske variable X have den karakteristiske funktion $\phi_X(\cdot)$. Da er den karakteristiske funktion for fordelingen af $Y = a + bX$ bestemt ved

$$\phi_Y(t) = \exp(\mathrm{i}at)\phi_X(bt)$$

Bevis:
Følger ved opskrivelse af definitionen.

□

Sætning 0.3.2 ***Karakteristisk funktion for summer af uafhængige variable*** Lad X og Y være uafhængige stokastiske variable med de karakteristiske funktioner hhv. $\phi_X(\cdot)$ og $\phi_Y(\cdot)$. Da er den karakteristiske funktion for $Z = X + Y$ bestemt ved

$$\phi_Z(t) = \phi_X(t)\phi_Y(t)$$

Bevis:
Følger ved opskrivelse af definitionen

□

Sætning 0.3.3 ***Differentiabilitet af karakteristisk funktion*** Lad den stokastiske variabel X have den karakteristiske funktion $\phi(\cdot)$. Såfremt det r'te moment for fordelingen af X eksisterer, da er $\phi(\cdot)$ differentiabel r gange, og der gælder

$$\phi^{(r)}(0) = \mathrm{i}^r\,\mathsf{E}[X^r]$$

Bevis:
Se fx Jørsboe (1988) p. 155.

□

I stedet for momenterne for en stokastisk variabel betragtes i en række sammenhænge kumulanterne

Definition 0.3.2 ***Kumulanter***

Lad den stokastiske variabel X have den karakteristiske funktion $\phi(\cdot)$ og antag at det r'te moment for X eksisterer. Betragt udviklingen

$$\ln \phi(t) = \sum_{j=0}^{r} \kappa_j \frac{(it)^j}{j!} + o(t^r)$$

for $t \to 0$. Koefficienten κ_j i udviklingen af $\ln \phi(\cdot)$ kaldes den j'te kumulant for fordelingen af X. □

Kumulanterne kaldes undertiden semiinvarianter, (specielt i ældre litteratur). Dette udtryk, der skyldes danskeren *Thiele*, benyttes, da kumulanterne (bortset fra κ_1) er invariante under translationer af den stokastiske variable.

Momenterne op til en vilkårlig orden, p, kan udtrykkes ved de p første kumulanter og omvendt.

Sætning 0.3.4 Relation mellem momenter og kumulanter

Antag at det r'te moment for fordelingen af X eksisterer, da gælder for $1 \leq j \leq r$

$$\mathsf{E}[X^j] = \sum_{\nu=0}^{j-1} \binom{j-1}{\nu} \mathsf{E}[X^\nu]\, \kappa_{j-\nu} \tag{0.3.2}$$

hvor κ_j angiver den j'te kumulant.
Sæt $\mu = \mathsf{E}[X]$. Da gælder

$$\mathsf{V}[X] = \mathsf{E}[(X-\mu)^2] = \kappa_2 \qquad \mathsf{E}[(X-\mu)^3] = \kappa_3$$

og for $4 \leq j$

$$\mathsf{E}[(X-\mu)^j] = \kappa_j + \sum_{\nu=0}^{j-2} \binom{j-1}{\nu} \mathsf{E}[(X-\mu)^\nu]\, \kappa_{j-\nu} \tag{0.3.3}$$

Bevis:
Ved induktion, se fx Morris (1982).

□

Specielt finder man for de første fire kumulanter

$$\begin{aligned}
\kappa_1 &= \mathsf{E}[X] \stackrel{\text{DEF}}{=} \mu \\
\kappa_2 &= \mathsf{E}[(X-\mu)^2] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 \\
\kappa_3 &= \mathsf{E}[(X-\mu)^3] = \mathsf{E}[X^3] - 3\mathsf{E}[X](\mathsf{E}[X])^2 + 3(\mathsf{E}[X])^3 \\
\kappa_4 &= \mathsf{E}[(X-\mu)^4] - 3\mathsf{E}[(X-\mu)^2] \\
&= \mathsf{E}[X^4] - 3(\mathsf{E}[X^2])^2 - 4\mathsf{E}[X]\mathsf{E}[X^3] + 12(\mathsf{E}[X])^2\mathsf{E}[X^2] - 6(\mathsf{E}[X])^4
\end{aligned}$$

og omvendt

$$\begin{aligned}
\mathsf{E}[X] &= \kappa_1 & \mathsf{E}[X^3] &= \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3 \\
\mathsf{E}[X^2] &= \kappa_2 + \kappa_1^2 & \mathsf{E}[X^4] &= \kappa_4 + 3\kappa_2^2 + 4\kappa_3\kappa_1 + 6\kappa_2\kappa_1^2 + \kappa_1^4
\end{aligned}$$

Eksempel 0.3.1 ***Kumulanter for normalfordelingen***

Lad $X \in \mathrm{N}(\mu, \sigma^2)$. Da har fordelingen af X kumulanterne $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, og $\kappa_i = 0$ for $i = 3, 4, \ldots$.

□

Sætning 0.3.5 ***Kumulanter for summer af uafhængige variable***

Lad X og Y være uafhængige variable sådan at fordelingen af X har kumulanterne $\kappa_{1,X}, \kappa_{2,X}, \ldots, \kappa_{r,X}$ og lad tilsvarende fordelingen af Y have kumulanterne $\kappa_{1,Y}, \kappa_{2,Y}, \ldots, \kappa_{s,Y}$ med $r \leq s$. Da har fordelingen for $Z = X + Y$ kumulanter op til r'te orden, og den i'te kumulant $\kappa_{i,Z}, i = 1, 2, \ldots, r$ for fordelingen af Z tilfredsstiller

$$\kappa_{i,Z} = \kappa_{i,X} + \kappa_{i,Y}$$

Bevis:
Følger af definitionen. □

Definition 0.3.3 ***Kumulantfrembringende funktion***

Lad X være en stokastisk variabel, og lad

$$D = \{t : \mathrm{E}[\exp(tX)] \text{ eksisterer}\} .$$

Mængden D er en konveks mængde. Såfremt X er k-dimensional, er $D \subset \mathbb{R}^k$.

Funktionen

$$K(t) = \ln(\mathrm{E}[\exp(tX)]) \quad \text{for } t \in D \tag{0.3.4}$$

kaldes den kumulantfrembringende funktion for fordelingen af X. □

Mens den karakteristiske funktion (0.3.1) er defineret for alle værdier af t, gælder dette ikke for den kumulantfrembringende funktion, hvorfor vi har været nødt til at indføre definitionsmængden D.

Sætning 0.3.6 ***Egenskaber ved den kumulantfrembringende funktion***

Lad $K(\cdot)$ være den kumulantfrembringende funktion for fordelingen af en stokastisk variabel X. Da er $K(\cdot)$ en konveks funktion på D. Funktionen er strengt konveks, med mindre støtten for X er koncentreret på et affint underrum af $\mathbb{R}^k$.

Såfremt nul er et indre punkt i D, findes der et $t_0 > 0$ sådan at $K(\cdot)$ kan udvikles i en potensrække om nul for

$$K(t) = \sum_{i=0}^{\infty} \frac{t^i}{i\,!}\,\kappa_i \quad \text{for } |t| < t_0$$

Koefficienterne κ_i er kumulanterne for fordelingen af X

Bevis:
Se fx Barndorff-Nielsen (1978). □

0.4 Laplacetransform

Stokastiske variable, der beskriver fænomener i tid, som fx ventetider eller levetider, er positive stokastiske variable. I modsætning til den almindelige situation i sandsynlighedsregningen, hvor man sædvanligvis benytter den karakteristiske funktion til beskrivelse af fordelingens egenskaber, er det for positive variable ofte mere bekvemt at benytte den Laplace-transformerede. Vi resumerer derfor kort de vigtigste resultater vedrørende Laplacetransformer.

Definition 0.4.1 ***Laplacetransform*** For en kontinuert funktion $f(\cdot)$, med støtte indeholdt i $\mathbb{R}_+$ defineres den Laplacetransformerede $\psi_f(\cdot)$ ved

$$\psi_f(s) = \int_0^\infty \exp(-st) f(t) dt \tag{0.4.1}$$

□

Bemærkning 1 ***Den Laplacetransformerede er en analytisk funktion***

Såfremt integralet konvergerer for $\mathrm{Re}(s) = a$, er det også konvergent for $\mathrm{Re}(s) > a$. Funktionen $\psi(\cdot)$ er således defineret i den komplekse halvplan $\mathrm{Re}(s) > a$, og $\psi(\cdot)$ er analytisk i denne halvplan. □

Sætning 0.4.1 ***Laplacetransform for positiv stokastisk variabel***

Lad T være en positiv stokastisk variabel med kontinuert tæthed $f(\cdot)$.

Da er den Laplacetransformerede for $f(\cdot)$ bestemt ved

$$\psi_f(s) = \mathrm{E}[\exp(-sT)] \tag{0.4.2}$$

Bevis:
Beviset følger ved at opskrive definitionen på ψ_f.

□

Vi anfører uden bevis en række egenskaber for Laplacetransformationen:

Lad $f(\cdot)$ og $g(\cdot)$ have støtten indeholdt i $\mathbb{R}_+$, og lad a og b være reelle tal. Da gælder

$$\psi_{af+bg}(s) = a\psi_f(s) + b\psi_g(s). \tag{0.4.3}$$

Såfremt $f'(t) = \mathrm{d}f(t)/\mathrm{d}t$ eksisterer, gælder

$$\psi_{f'}(s) = s\psi_f(s) - f(0+). \tag{0.4.4}$$

Lad $F(t) = \int_0^t f(u)du$, da gælder

$$\psi_F(s) = \frac{1}{s}\psi_f(s)\ . \tag{0.4.5}$$

Lad

$$g(t) = \begin{cases} 0 & \text{for } t < a \\ f(t-a) & \text{for } t \geq a\ , \end{cases}$$

da er

$$\psi_g(s) = \exp(-as)\psi_f(s)\ . \tag{0.4.6}$$

Lad

$$g(t) = \exp(-at)f(t)\ ,$$

da er

$$\psi_g(s) = \psi_f(a+s)\ . \tag{0.4.7}$$

Lad $f(\cdot)$ have den Laplacetransformerede $\psi(\cdot)$. Da gælder

$$\lim_{s\to 0} s\psi_f(s) = f(0) \quad \text{og} \lim_{s\to\infty} s\psi_f(s) = \lim_{t\to\infty} f(t)\ . \tag{0.4.8}$$

Lad $f(\cdot)$ være tætheden for en positiv stokastisk variabel T. Da gælder

$$\psi_f(0) = 1 \quad \psi_f'(0) = -\mathsf{E}[T] \quad \psi_f''(0) = \mathsf{E}[T^2]\ , \tag{0.4.9}$$

og almindeligt

$$\psi_f^{(n)}(0) = (-1)^n \mathsf{E}[T^n]\ . \tag{0.4.10}$$

Lad U og V være uafhængige positive stokastiske variable med kontinuert tæthed $f(\cdot)$ og $g(\cdot)$, og lad $T = aU + bV$, hvor $a \in \mathbb{R}_+$ og $b \in \mathbb{R}_+$. Lad $h(\cdot)$ angive tætheden for T. Da gælder

$$\psi_h(s) = \psi_f(as) + \psi_g(bs)\ . \tag{0.4.11}$$

□

Bemærkning 1 *Momentfrembringende funktion*

Funktionen defineret ved (0.4.2) kaldes undertiden også den momentfrembringende funktion for X. I nyere litteratur synes den fremherskende tendens dog at være, at man bruger betegnelsen momentfrembringende funktion om funktionen

$$M(s) = \mathrm{E}[\exp(sX)] \,. \tag{0.4.12}$$

Når vi i denne note bruger betegnelsen momentfrembringende funktion, vil vi mene funktionen $M(\cdot)$ bestemt ved (0.4.12). Det gælder da (jvf (0.3.4) og (0.4.12)), at den kumulantfrembringende funktion, $K(t)$ fås fra den momentfrembringende funktion som

$$K(t) = \ln(M(t)) \tag{0.4.13}$$

□

0.5 Projektioner og mindste kvadraters metode

Lad V være et k-dimensionalt vektorrum.

Et lineært underrum L_0 af dimension m kan specificeres på to måder:

Som billedet af en lineær afbildning

$$\eta \in L_0 \Leftrightarrow \eta = \mathbf{X}\beta \,, \beta \in \mathbb{R}^m$$

hvor $\mathbf{X}$ er en $k \times m$ matrix af rang m.

Eller som kernen for en lineær afbildning

$$\eta \in L_0 \Leftrightarrow \mathbf{M}^T\eta = \mathbf{0} \in V$$

hvor $\mathbf{M}^T$ er en $(k-m) \times m$ matrix af rang $k-m$ som tilfredsstiller

$$\mathbf{M}^T\mathbf{X} = \mathbf{0}$$

Den første metode er velegnet til bestemmelse af parameterestimater, idet

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{0.5.1}$$

angiver projektionsmatricen ned på L_0, og

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{0.5.2}$$

og

$$\widehat{\boldsymbol{\eta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\,. \tag{0.5.3}$$

Der gælder

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Matricen $\mathbf{H}$ bestemt ved (0.5.1) kaldes hatmatricen, da den overfører observationerne $\mathbf{y}$ til de tilsvarende fittede værdier, $\widehat{\boldsymbol{\eta}}$, der netop er kendetegnet ved en "hat" over symbolet for den teoretiske værdi.

Den anden metode er velegnet til bestemmelse af kvadratafvigelsessummer, idet residualerne direkte kan udtrykkes ved matricen $\mathbf{M}$.

Vi har

$$\mathbf{M}^T\mathbf{X} = \mathbf{0}\,,$$

og derfor gælder

$$\mathbf{I} - \mathbf{H} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$$

idet $\mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$ er projektionen på billedet af $\mathbf{M}$ og

$$\mathrm{im}(\mathbf{M}) = \ker(\mathbf{M}^T)^{\perp} = L_0^{\perp}$$

da $L_0 = \ker(\mathbf{M}^T)$.

Sætning 0.5.1 ***Den ortogonale projektionsmatrix på rummet udspændt af*** $\mathbf{X}$

Lad L_0 være rummet udspændt af søjlerne i $\mathbf{X}$. Da er den ortogonale projektionsmatrix på L_0 givet ved $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Bevis:

Vises direkte ved indsættelse i definitionen □

Hvis $\mathbf{X}^T\mathbf{X}$ ikke har fuld rang, gælder resultatet for en generaliseret invers til $\mathbf{X}^T\mathbf{X}$.

Sætning 0.5.2 *Mindste kvadraters estimat*

Lad $\mathbf{y}$ angive en k-dimensional vektor af observationer og lad $\mathbf{X}$ være en $k \times m$ dimensional matrix af koefficienter med fuld rang, $m < k$.

Betragt kvadratafvigelssummen

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{k} \left(y_i - \sum_{j=1}^{m} x_{ij}\beta_j \right)^2 = \sum_{i=1}^{k} \left(y_i - \mathbf{x}_i^* \boldsymbol{\beta} \right)^2 , \tag{0.5.4}$$

hvor $\mathbf{x}_i^*$ er vektoren bestående af i'te række i $\mathbf{X}$.

Den værdi af $\boldsymbol{\beta}$, der minimerer kvadratafvigelsessummen (0.5.4) kaldes mindste kvadraters estimatet for $\boldsymbol{\beta}$.

Lad $\boldsymbol{\beta}^{LS}$ betegne mindste kvadraters estimatoren for $\boldsymbol{\beta}$. Der gælder

$$\boldsymbol{\beta}^{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{0.5.5}$$

Endvidere gælder, at matricen

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{0.5.6}$$

er matricen for den ortogonale projektion ned på underrummet L udspændt af søjlerne i $\mathbf{X}$.

Vektoren $\mathbf{r}$ af residualer

$$\mathbf{r} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I}_k - \mathbf{H})\mathbf{y} \tag{0.5.7}$$

er således ortogonal på L, og derfor specielt ortogonal på vektoren

$$\widehat{\boldsymbol{\mu}} = \mathbf{X}\boldsymbol{\beta}^{LS} = \mathbf{H}\mathbf{y} \tag{0.5.8}$$

af fittede værdier.

Bevis:

Følger ved at opskrive kvadratafvigelssummen som

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

og betragte normalligningerne

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) ,$$

der netop fører til ortogonalitetsbetingelsen

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{LS}) = \mathbf{0} \tag{0.5.9}$$

□

Sætning 0.5.3 *Vægtet mindste kvadraters estimat*

Betragt kvadratafvigelssummen

$$S^*(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \,, \tag{0.5.10}$$

hvor $\mathbf{W}$ er en vilkårlig positiv definit, $k \times k$-dimensional symmetrisk matrix.

Den værdi af $\boldsymbol{\beta}$, der minimerer kvadratafvigelsessummen (0.5.10) kaldes det vægtede mindste kvadraters estimat for $\boldsymbol{\beta}$.

Lad $\boldsymbol{\beta}^{WLS}$ betegne det vægtede mindste kvadraters estimat for $\boldsymbol{\beta}$. Der gælder

$$\boldsymbol{\beta}^{WLS} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \tag{0.5.11}$$

Lad $\mathbf{Q}$ være en $k \times k$-dimensional matrix, der tilfredsstiller

$$\mathbf{W} = \mathbf{Q}^T\mathbf{Q} \tag{0.5.12}$$

Det gælder da, at matricen

$$\mathbf{H} = \mathbf{X}\left[\mathbf{X}^T\mathbf{W}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{W} \tag{0.5.13}$$

er matricen for en projektion ned på underrummet L udspændt af søjlerne i $\mathbf{X}$. Projektionen er ikke ortogonal med hensyn til det sædvanlige indre produkt, den er en ortogonal projektion svarende til et indre produkt på $\mathbb{R}^k$ defineret ved $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T\mathbf{W}\mathbf{v}$.

Der gælder

$$\mathbf{X}^T\mathbf{W}\left[\mathbf{I}_k - \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\right]\mathbf{y} = \mathbf{0} \tag{0.5.14}$$

Transformationen $\mathbf{y}_1 = \mathbf{Q}\mathbf{y}$, $\mathbf{X}_1 = \mathbf{Q}\mathbf{X}$ fører problemet over i et sædvanligt mindste kvadraters problem vedrørende $\mathbf{y}_1$ og underrrummet L_1 udspændt

af søjlerne i $\mathbf{X}_1$.

Bevis:

Beviset følger ved at bemærke, at det sædvanlige mindste kvadraters problem svarende til de transformerede variable $\mathbf{y}_1 = \mathbf{Q}\mathbf{y}$ $\mathbf{X}_1 = \mathbf{Q}\mathbf{X}$ netop har kvadratafvigelsessummen

$$S(\boldsymbol{\beta}) = (\mathbf{Q}\mathbf{y} - \mathbf{Q}\mathbf{X}\boldsymbol{\beta})^T (\mathbf{Q}\mathbf{y} - \mathbf{Q}\mathbf{X}\boldsymbol{\beta})$$

der kan udtrykkes

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{Q}^T \mathbf{Q}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

altså netop $S^*(\boldsymbol{\beta})$ □

Bemærkning 1 ***Bestemmelse af "kvadratrodsmatricen" Q***

Matricen $\mathbf{Q}$, der indgår i transformationen af det vægtede problem kan eksempelvis bestemmes ved hjælp af den såkaldte Cholesky dekomposition, der resulterer i en trekantsmatrix $\mathbf{Q}$.

En matrix $\mathbf{Q}$, der tilfredsstiller (0.5.12) betegnes med symbolet $(\mathbf{W})^{1/2}$.

Såfremt vægtmatricen W er en diagonalmatrix, kan $\mathbf{Q}$ ligeledes vælges som en diagonalmatrix med elementerne $\sqrt{w_i}$. Betragtes denne situation geometrisk ses, at vi blot har ændret enheden på koordinatakserne. □

0.6 Referencer

Barndorff-Nielsen, O. (1978): *Information and Exponential Families in Statistical Theory*, Wiley & Sons, Inc. New York

Conradsen, K.: (1997): *En Introduktion til Statistik, Bind 1A*, Institut for Matematisk Modellering, DTU.

Jørsboe, O.G. (1984): *Sandsynlighedsregning*, Matematisk Institut, DTH

Afsnit 1

Familier af fordelinger

fil: /tex/stat3/fordbog/forda.tex 1998-01-13

1.1 Lidt om familier af fordelinger

Ved formuleringen af en statistisk model, vil det ofte være naturligt at specificere formen af observationernes fordeling. Dette gøres ved at angive en samling (mængde) af fordelinger som mulige fordelinger for de indgående stokastiske variable.

Ofte bruger vi betegnelsen familie af fordelinger om en sådan samling af fordelinger.

En familie af fordelinger behøver ikke at være særligt subtilt struktureret. Det væsentlige er, at den betragtede mængde af fordelinger er indekseret, dvs at der er en enentydig sammenhæng mellem mængden af fordelinger og en indeksmængde, således at man kan udpege en bestemt fordeling ved at angive værdien af dens indeks.

Når man har en sådan indekseret mængde af fordelinger siger man ofte, at den er parametriseret. Parameterområdet er netop indeksmængden, og parameteren er det benyttede indeks.

I praksis vil der sædvanligvis være en naturlig afgrænsning af den betragtede familie af fordelinger, feks ved at tætheden for fordelingerne har samme analytiske udtryk. Tilsvarende vil man vælge en parametrisering (indeksering), som er meningsfuld i den betragtede sammenhæng.

I det følgende vil vi betragte nogle familier af fordelinger, som har speciel interesse i statistikken, de såkaldte eksponentielle familier, samt den lidt mere omfattende samling af eksponentielle dispersionsparameterfordelinger.

1.1.1 Uendeligt delbare fordelinger

Vi indleder med at indføre et begreb, der undertiden er af interesse ved karakterisering af en familie af fordelinger.

Definition 1.1.1 ***Uendeligt delbar fordeling***

En sandsynlighedsfordeling, $F(\cdot)$, siges at være uendeligt delbar, hvis det gælder for ethvert $n \in \mathbb{N}$, at fordelingen kan repræsenteres som fordelingen af en sum,

$$Z_n = X_{1,n} + X_{2,n} + \cdots + X_{n,n}$$

af uafhængige, identisk fordelte variable $X_{i,n}$. □

Uendeligt delbare familier spiller en særlig rolle i forbindelse med beskrivelse af additive stokastiske processer.

Eksempelvis er enhver normalfordeling og enhver Poissonfordeling uendeligt delbar.

Vi henviser til Feller (1966) for en diskussion af uendeligt delbare fordelinger.

1.2 Eksponentielle familier

De fleste af de fordelinger, der blev introduceret i Statistik I, har nogle væsentlige egenskaber fælles. Således har vi set i Statistik I, at ved uafhængige, identisk fordelte gentagelser, vil summen af observationerne, eller summen af en funktion af observationerne være sufficient. Det gælder endvidere, at

ved en passende parametrisering vil affine hypoteser vedrørende parameteren modsvares af tilsvarende affine transformationer af observationerne.

Sådanne egenskaber deles af en stor gruppe af fordelinger, der under ét betegnes eksponentielle familier. Vi skal i dette afsnit introducere de eksponentielle familier og resumere nogle vigtige resultater for uafhængige observationer fra endimensionale eksponentielle familier.

Vi indfører grundformen (den kanoniske form) for en eksponentiel familie og den tilsvarende parametrisering ved den kanoniske parameter. Ønsker man direkte at relatere observationer til parameterværdier er det imidlertid mere bekvemt at betragte den såkaldte middelværdiparametrisering af familierne. For de endimensionale familier gælder specielt, at familien er karakteriseret ved den funktion, der beskriver variansen som funktion af middelværdien.

1.2.1 Naturlig eksponentiel familie

I 1990'erne har udviklingen af teorier for de eksponentielle familier været koncentreret omkring beskrivelse af de såkaldte naturlige eksponentielle familier. Disse familier repræsenterer grundformen for eksponentielle familier og de fleste egenskaber for eksponentielle familier følger af egenskaberne for de naturlige eksponentielle familier. Vi vil derfor indlede med en beskrivelse af disse familier.

Definition 1.2.1 ***Naturlig eksponentiel familie, kanonisk stikprøvefunktion, kanonisk parameter***

Lad $\nu\{\cdot\}$ være et σ-endeligt mål[1] på $\mathbb{R}^k$, som ikke er koncentreret på et affint underrum af $\mathbb{R}^k$.

Betragt

$$\psi(\vartheta) \stackrel{\text{DEF}}{=} \int \exp\{\vartheta^T x\}\, \nu\{dx\}\,, \vartheta \in \mathbb{R}^k\,. \tag{1.2.1}$$

og sæt

$$D = \{\vartheta \in \mathbb{R}^k :\ \psi(\vartheta) < \infty\} \tag{1.2.2}$$

[1] Sædvanligvis vil $\nu\{\cdot\}$ være på formen $\nu\{dx\} = c(x)dx$, hvor dx er enten Lebesguemålet (kontinuert tæthed) eller tællemålet (diskret tæthed)

Mængden D er konveks. Lad $\text{int}(D)$ angive det indre af D.

Lad endelig

$$\kappa(\vartheta) \stackrel{\text{DEF}}{=} \ln(\psi(\vartheta)) \quad \text{for } \vartheta \in D. \tag{1.2.3}$$

Såfremt $\text{int}(D) \neq \emptyset$, kaldes familien af fordelinger med tætheder (mht. $\nu\{\cdot\}$)

$$f(x;\vartheta) = \exp\{\vartheta^T x - \kappa(\vartheta)\} \quad \text{for } \vartheta \in D \tag{1.2.4}$$

for den naturlige eksponentielle familie frembragt af målet $\nu\{\cdot\}$ på $\mathbb{R}^k$. Såfremt specielt målet $\nu\{\cdot\}$ er en sandsynlighedsfordeling, $F(\cdot)$, siger man, at familien er frembragt af fordelingen $F(\cdot)$. En naturlig eksponentiel familie kan frembringes af enhver af fordelingerne i familien.

Den stokastiske variable X kaldes den kanoniske stikprøvefunktion, parameteren ϑ kaldes den kanoniske parameter, og mængden D kaldes det kanoniske parameterområde for familien. Det positive heltal, k, kaldes for familiens orden. □

Bemærkning 1 ***Den eksponentielle familie er bestemt af det frembringende mål*** $\nu(\cdot)$

Det ses, at såvel $\psi(\cdot)$ (og dermed også $\kappa(\cdot)$) og parameterområdet D er bestemt af målet $\nu(\cdot)$.

Såfremt man har blot ét sandsynlighedsmål ν på $\mathbb{R}$, kan man altså benytte dette mål til at frembringe en eksponentiel familie ved (1.2.4). Nedenstående eksempel 1.2.1 viser et eksempel på frembringelsen af en eksponentiel familie udfra et enkelt sandsynlighedsmål. □

Eksempel 1.2.1 ***Frembringelse af en naturlig eksponentiel familie***

Betragt Bernoulli-fordelingen med $p = 1/2$. Fordelingen er karakteriseret ved

$$\nu\{0\} = \nu\{1\} = 1/2\,. \tag{1.2.5}$$

Funktionen $\psi(\vartheta)$ (1.2.1) for dette mål er

$$\psi(\vartheta) = \frac{1}{2}\,\mathrm{e}^{\vartheta\cdot 0} + \frac{1}{2}\,\mathrm{e}^{\vartheta\cdot 1} = \frac{1}{2}\,(1+\mathrm{e}^{\vartheta})$$

Den naturlige eksponentielle familie frembragt af (1.2.5) er karakteriseret ved frekvensfunktionen

$$f(x;\vartheta) = \frac{2}{1+\exp(\vartheta)}\mathrm{e}^{\vartheta\cdot x} = \begin{cases} \dfrac{1}{1+\exp(\vartheta)} & \text{for } x = 0 \\[2ex] \dfrac{\exp(\vartheta)}{1+\exp(\vartheta)} & \text{for } x = 1. \end{cases}$$

Havde vi i stedet taget udgangspunkt i Bernoulli-fordelingen med $p = 3/4$, dvs fordelingen bestemt ved

$$\nu_1\{0\} = 1/4\,;\ \nu_1\{1\} = 3/4\,, \tag{1.2.6}$$

havde vi fundet funktionen

$$\psi_1(\vartheta_1) = \frac{1}{4}\,\mathrm{e}^{\vartheta_1\cdot 0} + \frac{3}{4}\,\mathrm{e}^{\vartheta_1\cdot 1} = \frac{1}{4}[1+\exp(\vartheta_1+\ln 3)]\,;$$

dvs den naturlige eksponentielle familie frembragt af (1.2.6) er af formen

$$f_1(x;\vartheta_1) = \frac{4}{1+\exp(\vartheta_1+\ln 3)}\mathrm{e}^{\vartheta_1\cdot x} = \begin{cases} \dfrac{1}{1+\exp(\vartheta_1+\ln 3)} & \text{for } x = 0 \\[2ex] \dfrac{3\exp(\vartheta_1)}{1+\exp(\vartheta_1+\ln 3)} & \text{for } x = 1. \end{cases}$$

altså den samme familie som før, med $\vartheta = \vartheta_1 + \ln 3$. □

Eksempel 1.2.2 ***Binomialfordelingen som naturlig eksponentiel familie***

Betragt familien af binomialfordelinger, $\mathrm{B}(n,p)$, for et fast $n \in \mathbb{N}_+$ og med $p \in]0,1[$.

Indfører vi

$$\vartheta = \frac{p}{1-p} \quad \text{for } 0 < p < 1$$

og

$$\kappa(\vartheta) = n \ln\{1 + \exp(\vartheta)\}$$

ser vi, at frekvensfunktionen kan udtrykkes som tætheden

$$f(z;\vartheta) = \exp\{\vartheta z - \kappa(\vartheta)\}$$

med hensyn til målet

$$\nu_n\{dz\} = \binom{n}{z}$$

gange tællemålet.

Familien af $\mathrm{B}(n,p)$-fordelinger er således en naturlig eksponentiel familie med kanonisk parameter $\vartheta = p/(1-p)$, kumulantfrembringer $\kappa(\vartheta) = n \ln\{1 + \exp(\vartheta)\}$ og kanonisk parameterområde $D = \mathbb{R}$.

Vi bemærker, at selv om det er naturligt at betragte familien af binomialfordelinger for hele det lukkede interval $0 \leq p \leq 1$, omfatter den tilsvarende naturlige eksponentielle familie kun fordelingerne svarende til $0 < p < 1$. □

Bemærkning 2 ***Parametrisering ved kanonisk parameter giver en fælles grundform for tætheden***

I mange situationer parametriserer man ikke en given eksponentiel familie ved den kanoniske parametrisering, men man bruger en funktion, $\theta = \theta(\vartheta)$, fex middelværdifunktionen (se definition 1.2.5) af den kanoniske parameter.

Repræsentationen af familien på kanonisk form tjener til at beskrive den fælles grundform for disse familier, sådan at man kan beskrive de egenskaber, der kan afledes af denne fælles grundform. □

Bemærkning 3 ***Repræsentationen af en eksponentiel familie beskriver forskelle på familiens elementer***

I repræsentationen (1.2.4) af tæthederne er enhver faktor af formen $k(x)$, der alene afhænger af x, absorberet i målet ν. Funktionen $\psi(\cdot)$ givet ved (1.2.1) er bestemt af ν sådan at tætheden virkelig er en sandsynlighedstæthed, dvs så

$$\mathsf{P}[X \in B] = \int_B \exp\{\vartheta x - \kappa(\vartheta)\}\nu\{dx\}$$

udtrykker sandsynligheden for hændelsen $[X \in B]$ for enhver målelig mængde B, og altså specielt så $\mathsf{P}[X \in \mathbb{R}] = 1$. □

Definition 1.2.2 ***Konveks støtte for naturlig eksponentiel familie.***

Støtten, S, for den eksponentielle familie er mængden

$$S = \left\{ x \in \mathbb{R}^k : \int_{O(x)} \nu\{dx\} > 0 \quad \text{for enhver omegn, } O(x), \text{ af } x \right\}$$

Støtten afhænger således ikke af parameteren ϑ.

Ved den konvekse støtte, C, for familien vil vi forstå den den mindste konvekse mængde, der indeholder støtten S. □

Definition 1.2.3 ***Kumulantfrembringer for naturlig eksponentiel familie.***

Betragt en naturlig eksponentiel familie på formen (1.2.4).

Afbildningen $\kappa(\cdot)$ givet ved (1.2.3) kaldes kumulantfrembringeren eller undertiden kumulantfunktionen for familien. □

Det kan vises, at kumulantfrembringeren er en strengt konveks funktion af $\vartheta \in D$.

Sætning 1.2.1 ***Momentfrembringende og kumulantfrembringende funktion for en fordeling tilhørende en eksponentiel familie***

Såfremt fordelingen af X tilhører en naturlig eksponentiel familie med tætheder (1.2.4), da er den momentfrembringende funktion for fordelingen af X givet ved

$$M(t;\vartheta) = \exp\{\kappa(\vartheta+t) - \kappa(\vartheta)\} \quad \text{for } t \in D - \vartheta \ ,$$

og den kumulantfrembringende funktion for fordelingen af X er

$$K(t;\vartheta) = \kappa(\vartheta+t) - \kappa(\vartheta) \quad \text{for } t \in D - \vartheta \quad . \tag{1.2.7}$$

Bevis:
Den momentfrembringende funktion for X er bestemt ved

$$\begin{aligned} M(t) &= \mathrm{E}[\exp(t^T X)] = \int \exp(t^T x) f(x;\vartheta)\nu\{dx\} \\ &= \int \exp\left\{(\vartheta+t)^T x - \kappa(\vartheta)\right\}\nu\{dx\} \\ &= \exp\left\{\kappa(\vartheta+t) - \kappa(\vartheta)\right\} \int f(x;\vartheta+t)\nu\{dx\} \\ &= \exp\{\kappa(\vartheta+t) - \kappa(\vartheta)\} \ , \end{aligned}$$

idet det sidste integral jo er én.

Da den kumulantfrembringende funktion $K(t) = \ln\big(M(t)\big)$ følger (1.2.7) umiddelbart. □

Sætning 1.2.2 ***Entydighed af kumulantfrembringer***

Kumulantfrembringeren, $\kappa(\cdot)$, er entydigt bestemt af familien (på nær relationen $\kappa_1(\vartheta) = c + \kappa(\vartheta+\vartheta_0)$.
Bevis:
Se Jørgensen (1997). □

Sætning 1.2.3 ***Momenter for fordelinger i eksponentielle familier***

Lad fordelingen af X tilhøre en naturlig eksponentiel familie med tæthed (1.2.4).

Såfremt $\vartheta \in \mathrm{int}(D)$, eksisterer alle momenter af X, og den i'te kumulant for fordelingen af X er

$$\kappa_i(\vartheta) = \mathrm{D}^i\kappa(\vartheta) \ ,$$

hvor D angiver den sædvanlige differentialoperator, og i er et k-dimensionalt sæt af ikke-negative heltal, $i = (i_1, \ldots, i_k)$.

For $\vartheta \in \text{int}(D)$ gælder altså specielt, at

$$\mathsf{E}[X] = \frac{\partial}{\partial \vartheta} \kappa(\vartheta) \tag{1.2.8}$$

og

$$\mathbf{D}[X] = \frac{\partial^2}{\partial \vartheta^T \, \partial \vartheta} \kappa(\vartheta) \, , \tag{1.2.9}$$

hvor $\mathbf{D}[X]$ er positiv definit[1].

De lokale egenskaber for funktionen $\kappa(\cdot)$ i omegnen af ϑ beskriver således fordelingsforholdene svarende til fordelingen med parameterværdien ϑ.

Bevis:
Følger umiddelbart af udtrykket

$$\frac{\partial}{\partial \vartheta} \kappa(\vartheta) = \int x \exp\left\{\vartheta^T x - \kappa(\vartheta)\right\} \nu\{dx\} = \mathsf{E}[X]$$

På tilsvarende måde fås (1.2.9). □

Definition 1.2.4 ***Middelværdiafbildning og middelværdirum for naturlig eksponentiel familie***

Betragt en naturlig eksponentiel familie af fordelinger med tætheder på formen (1.2.4).

Afbildningen $\tau(\cdot)$ bestemt ved

$$\tau(\vartheta) \stackrel{\text{DEF}}{=} \frac{\partial}{\partial \vartheta} \kappa(\vartheta) = \mathsf{E}[X] \, , \tag{1.2.10}$$

der afbilder $\text{int}(D) \to \mathcal{M} \subset \mathbb{R}^k$, kaldes middelværdiafbildningen , og billedmængden $\mathcal{M} = \tau(\text{int}(D))$ kaldes middelværdirummet. □

[1]Vi har her benyttet det generelle symbol $\mathbf{D}[X]$ for dispersionsmatricen. For en endimensional eksponentiel familie er det bare variansen, $\mathsf{V}[X] = \kappa''(\vartheta)$

Middelværdiafbildningen er en reel, analytisk, bijektiv afbildning fra $\text{int}(D)$ ind på middelværdirummet $\mathcal{M}$.

Middelværdirummet $\mathcal{M}$ er en konveks delmængde af $\mathbb{R}^k$. Da $\text{int}(D)$ er åben, og da $\mathcal{M}$ er billedet af $\text{int}(D)$, er $\mathcal{M}$ ligeledes en åben mængde.

Bemærkning 1 ***Naturlig eksponentiel familie på vilkårligt vektorrum***

Teorien for eksponentielle familier knytter sig stærkt til teorien for vektorrum. I ovenstående definition antog vi, at der var valgt basis i vektorrummet således at det kunne identificere med $\mathbb{R}^k$. Dette er imidlertid ikke nødvendigt. Man kan definere en naturlig eksponentiel familie koordinatfrit på et abstrakt, endeligdimensionalt vektorrum. Vi vil her skitsere denne definition.

Lad E være et vilkårligt endeligdimensionalt vektorrum og lad E^* angive det duale rum (mængden af linearformer på E). Lad desuden $\langle \vartheta, x \rangle$ angive den kanoniske bilinearform på $E^* \times E$.

Lad endelig $\nu\{\cdot\}$ være et mål på E og betragt

$$\psi(\vartheta) \stackrel{\text{DEF}}{=} \int \exp\{\langle \vartheta, x \rangle\}\, \nu\{dx\}\,, \vartheta \in E^*\,.$$

Lad

$$D = \{\vartheta \in E^* : \psi(\vartheta) < \infty\}$$

og sæt

$$\kappa(\vartheta) \stackrel{\text{DEF}}{=} \ln(\psi(\vartheta)) \quad \text{for } \vartheta \in D.$$

Lad $\mathcal{P}$ angive samlingen af sandsynlighedsmål, der har tætheden

$$f(x;\vartheta) = \exp\Big\{\langle \vartheta, x \rangle - \kappa(\vartheta)\Big\} \quad \text{for } \vartheta \in D \tag{1.2.11}$$

med hensyn til målet ν.

Familien $\mathcal{P}$ med elementer $P \in \mathcal{P}$ kaldes den naturlige eksponentielle familie frembragt af målet $\nu\{\cdot\}$ og funktionen $\kappa(\cdot)$ kaldes kumulantfrembringeren for familien.

Det væsentlige indhold i denne koordinatfri fremstilling er, at den tydeliggør at de kanoniske parametre er elementer i det duale vektorrum til vektorrummet, E, for observationerne.

Middelværdien svarende til fordelingen $P \in \mathcal{P}$ er den linearform på E^*, der fører $\ell \in E^*$ over i $\int \ell(x) P\{dx\} = \mathsf{E}[\ell(X)]$, dvs middelværdien er et element i E^{**} (mængden af linearformer på E^*), der er isomorf med E.

Afbildningen

$$\tau(\vartheta) = \frac{\partial}{\partial \vartheta}\,\kappa(\vartheta) = \int x \exp\left\{\langle \vartheta, x\rangle - \kappa(\vartheta)\right\} \nu\{dx\}$$

er en reel analytisk, bijektiv afbildning, der afbilder $\mathrm{int}(D)$ ind på middelværdimængden, $\mathcal{M}$.

Kovariansen, Σ svarende til fordelingen $P \in \mathcal{P}$ er den symmetriske, positiv semidefinite bilinearform på $E^* \times E^*$, der fører (ℓ_1, ℓ_2) over i

$$\mathsf{E}[(\ell_1(X) - \mathsf{E}[\ell_1(X)])(\ell_2(X) - \mathsf{E}[\ell_2(X)])]$$

Nu er imidlertid mængden af bilinearformer på $E^* \times E^*$ isomorf med mængden af lineære afbildninger $E^* \to E$. Man kan således opfatte en kovarians Σ som en lineær afbildning $E^* \to E$. (Afbildningen er symmetrisk, dvs den er lig med sin transponerede).

De følgende resultater vedrørende middelværdiparametrisering og variansfunktion gælder også i denne koordinatfri repræsentation. □

Definition 1.2.5 ***Middelværdiparametrisering af naturlig eksponentiel familie.***

Betragt en naturlig eksponentiel familie på formen (1.2.4). Da afbildningen $\tau(\cdot) : \mathrm{int}(D) \to \mathcal{M}$ er injektiv, kan man parametrisere familien $\{f(\cdot;\vartheta)\}_{\vartheta \in \mathrm{int}(D)}$ ved middelværdien $\mu \in \mathcal{M}$.

Denne parametrisering kaldes middelværdiparametriseringen af den naturlige eksponentielle familie (1.2.4).

Sammenhængen mellem de to parametriseringer er bestemt ved

$$\begin{aligned} \mu &= \tau(\vartheta) \quad \text{for } \vartheta \in \mathrm{int}(D) \\ \vartheta &= \tau^{-1}(\mu) \quad \text{for } \mu \in \mathcal{M} \end{aligned} \tag{1.2.12}$$

Afbildningen $\tau^{-1}(\cdot)$, der fører middelværdien over i den kanoniske parameter, kaldes ofte den kanoniske link.

Hvis D bestemt ved (1.2.2) er en åben mængde, vil middelværdiparametriseringen (1.2.12) parametrisere hele den eksponentielle familie.

Hvis D indeholder nogle af sine randpunkter, kan middelværdiparametriseringen (1.2.12) udvides ved kontinuitet til at omfatte disse randpunkter (om nødvendigt ved at tillade uendelige værdier for middelværdien, μ). □

For en naturlig eksponentiel familie gælder, at $\mathcal{M} \subseteq \text{int}(C)$, hvor C angiver den konvekse støtte for familien.

Definition 1.2.6 ***Variansfunktion for naturlig eksponentiel familie***

Betragt en naturlig eksponentiel familie på formen (1.2.4), og lad middelværdiafbildningen $\tau(\cdot)$ være givet ved (1.2.10).

Funktionen

$$V(\mu) \stackrel{\text{DEF}}{=} \frac{\partial^2}{\partial \vartheta^T \, \partial \vartheta} \kappa(\tau^{-1}(\mu)) \quad \text{for } \mu \in \mathcal{M} \,, \tag{1.2.13}$$

der afbilder middelværdirummet ind i mængden af symmetriske, positiv definitte kvadratiske $(k \times k)$-matricer, kaldes variansfunktionen for familien.

For en endimensional familie er variansfunktionen blot

$$V(\mu) \stackrel{\text{DEF}}{=} \kappa''(\tau^{-1}(\mu)) = \tau'(\tau^{-1}(\mu)) \,. \tag{1.2.14}$$

□

Bemærkning 1 ***Variansfunktionen afhænger ikke af den valgte parametrisering af familien***

Vi bemærker, at variansfunktionen beskriver, hvorledes variansen svarende til en fordeling i familien afhænger af middelværdien for denne fordeling. Denne funktion har ikke noget med den valgte parametrisering at gøre. □

Sætning 1.2.4 ***En naturlig eksponentiel familie er fastlagt ved sin variansfunktion***

Bevis:

Vi skitserer beviset for en endimensional familie. Beviset i den generelle situation er anført af Letac (1992).

Det følger af (1.2.12) og relationen

$$\frac{\partial}{\partial \vartheta} \tau(\vartheta) = \frac{\partial}{\partial \vartheta^T} \left(\frac{\partial}{\partial \vartheta} \kappa(\vartheta) \right) = \frac{\partial^2}{\partial \vartheta^T \, \partial \vartheta} \kappa(\vartheta) \tag{1.2.15}$$

at

$$\frac{\partial}{\partial \mu} \tau^{-1}(\mu) = \{V(\mu)\}^{-1} , \tag{1.2.16}$$

og endvidere har vi af (1.2.10), at

$$\frac{\partial}{\partial \vartheta} \kappa(\vartheta) = \tau(\vartheta) . \tag{1.2.17}$$

For en given variansfunktion, $V(\cdot)$, kan man således bestemme κ ved at løse differentialligningen (1.2.16), invertere τ^{-1}, og derefter løse differentialligningen (1.2.17). Når κ er fastlagt, er familien (1.2.4) fastlagt. □

Bemærkning 1 ***Variansfunktionen udtrykt ved middelværdiafbildningen***

Det fremgår af (1.2.16), at variansfunktionen $V(\cdot)$ kan udtrykkes som

$$V(\mu) = \left\{ \frac{\partial}{\partial \mu} \tau^{-1}(\mu) \right\}^{-1} . \tag{1.2.18}$$

□

Eksempel 1.2.3 ***Bestemmelse af eksponentiel familie med*** $V(\mu) = \mu$

Antag, at vi ønsker at bestemme en naturlig eksponentiel familie af fordelinger sådan at $V(\mu) = \mu$, dvs sådan at

$$\mathsf{V}[Y] = \mathsf{E}[Y]$$

for enhver parameterværdi.

Det følger af (1.2.16), at den inverse afbildning til middelværdiafbildningen skal opfylde

$$\frac{d\vartheta}{d\mu} = \frac{1}{\mu} ,$$

dvs

$$\tau^{-1}(\mu) = \vartheta(\mu) = \int^{\mu} \frac{1}{t} dt = \ln(\mu) - \ln(c)$$

Middelværdiafbildningen er altså af formen

$$\tau(\vartheta) = c_1 \exp(\vartheta)$$

dvs at (1.2.17) bliver

$$\frac{d\kappa}{d\vartheta} = c_1 \exp(\vartheta)$$

hvorfor

$$\kappa(\vartheta) = \int^{\vartheta} c_1 \exp(t) dt = c_1 \exp(\vartheta) - c_1 \exp(c_2)$$

Dette er netop kumulantfrembringeren for Poissonfordelingen (se afsnit 4.24.2). □

Definition 1.2.7 ***Stejl og regulær eksponentiel familie***

Betragt en naturlig eksponentiel familie af fordelinger med tætheder på formen (1.2.4) og med den konvekse støtte C.

Såfremt middelvædirummet $\mathcal{M} = \text{int}(C)$, siges familien af være stejl (engelsk: *steep*).

Såfremt det kanoniske parameterområde, D, er en åben mængde, siges familien at være regulær. □

For en naturlig eksponentiel familie gælder, at såfremt den er regulær, er den også stejl.

Begrebet stejlhed er egentlig knyttet til kumulantfrembringeren $\kappa(\cdot)$. En differentiabel konveks funktion, $f(\cdot)$, siges at være stejl, hvis det gælder, at $|\mathrm{D}f(x_i)| \to \infty$ for enhver følge, $x_1, \ldots$, der konvergerer mod et randpunkt. Se Barndorff-Nielsen (1978) for en nærmere diskussion af stejle eksponentielle familier.

Regularitet og stejlhed spiller en vigtig rolle i forbindelse med maksimum-likelihood estimation. For en regulær eksponentiel familie er ligningen til bestemmelse af maksimum-likelihood estimatet netop middelværdiligningen

$$\tau(\vartheta) = x \,, \tag{1.2.19}$$

der udtrykker at maksimum-likelihood estimatet fås ved at sætte observationen lig sin middelværdi.

For en regulær eksponentiel familie gælder det, at maksimum-likelihood estimatet eksisterer, hvis og kun hvis (1.2.19) har en løsning, dvs hvis og kun hvis $x \in \text{int}(S)$.

Eksempel 1.2.4 *Maksimum-likelihood estimation i binomialfordelingen*

Vi så i eksempel 1.2.2, at familien af binomialfordelinger, $\text{B}(n,p)$, for ethvert fast n er en naturlig eksponentiel familie med kanonisk parameter $\vartheta = p/(1-p)$, kumulantfrembringer $\kappa(\vartheta) = n\ln\{1+\exp(\vartheta)\}$, middelværdiafbildning $\tau(\vartheta) = np(\vartheta)$. Det kanoniske parameterområde er $D = \mathbb{R}$, som er en åben mængde og familien er regulær. Middelværdimængden er $\mathcal{M} = \tau(D) =]0,n[$ mens den konvekse støtte for fordelingen er $C = [0,n]$. Da $\mathcal{M} = \text{int}(C)$ er familien stejl.

Middelværdiligningen er

$$np(\vartheta) = x\ ,$$

svarende til

$$\frac{\exp(\vartheta)}{1+\exp(\vartheta)} = \frac{x}{n}$$

For $x = 0$ og $x = n$ har ligningen ingen løsning. De tilsvarende værdier af den kanoniske parameter er jo $\vartheta = -\infty$ og $\vartheta = \infty$. Sådanne anomalier er prisen, der må betales for at opnå de fordele, der opnås ved behandlingen som en eksponentiel familie. □

Definition 1.2.8 *Deviansfunktion for endimensional naturlig eksponentiel familie*

Lad X følge en fordeling, der tilhører en endimensional naturlig eksponentiel familie med kumulantfrembringer $\kappa(\cdot)$, middelværdiafbildning $\tau(\vartheta) = \kappa'(\vartheta)$, kanonisk parameterområde D, konveks støtte C og middelværdimængde $\mathcal{M} = \tau\big(\text{int}(D)\big)$.

Funktionen

$$d(x;\mu) \stackrel{\text{DEF}}{=} 2\Big[\sup_{\vartheta\in D}\{\vartheta x - \kappa(\vartheta)\} - l_\mu(\mu;x)\Big] \qquad \text{for } x \in C \text{ og } \mu \in \mathcal{M}\,, \tag{1.2.20}$$

hvor

$$l_\mu(\mu;x) = \tau^{-1}(\mu)\, x - \kappa\big(\tau^{-1}(\mu)\big) \tag{1.2.21}$$

kaldes for deviansen svarende til x og μ. □

Bemærkning 1 ***Deviansen måler forskellen mellem x og μ udtrykt ved forskellen i loglikelihood***

Det følger af udtrykket (1.2.4), at logaritmen til likelihoodfunktionen svarende til observationen x er

$$l(\vartheta;x) = \vartheta x - \kappa(\vartheta) \tag{1.2.22}$$

(på nær et additivt bidrag, der kun afhænger af x).

Betragter vi nu i stedet loglikelihoodfunktionen som funktion af middelværdiparameteren, μ, finder vi

$$l_\mu(\mu;x) \stackrel{\text{DEF}}{=} l\big(\tau^{-1}(\mu);x\big) = \tau^{-1}(\mu)x - \kappa\big(\tau^{-1}(\mu)\big)\,,$$

der netop er (1.2.21).

Udtrykket $d(x;\mu)$ givet ved (1.2.20) er således to gange differensen mellem extremumsværdien for loglikelihoodfunktionen og loglikelihood'en svarende til værdien μ. Ved at betragte denne differens har vi elimineret indflydelsen fra det additive bidrag (der kun afhænger af x) til loglikelihood'en.

Sættes differentialkvotienten til $l(\vartheta;x)$ lig nul, får man netop (idet $\kappa'(\vartheta) = \tau(\vartheta)$) middelværdiligningen (1.2.19) til bestemmelse af maksimum-likelihood estimatet, $\widehat{\vartheta}$, for ϑ.

middelværdiligningen.

Såfremt $x \in \mathcal{M}$ har ligningen løsningen $\widehat{\vartheta} = \tau^{-1}(x)$. Maksimumværdien for l er da

$$\max_{\vartheta\in D} l(\vartheta;x) = x\tau^{-1}(x) - \kappa\big(\tau^{-1}(x)\big)$$

For $x \in \mathcal{M}$ kan vi derfor udtrykke deviansen (1.2.20) som

$$\begin{aligned} d(x;\mu) &= 2\{l_\mu(x;x) - l_\mu(\mu;x)\} \\ &= 2\left[x\{\tau^{-1}(x) - \tau^{-1}(\mu)\} - \{\kappa(\tau^{-1}(x)) - \kappa(\tau^{-1}(x))\}\right]. \end{aligned} \tag{1.2.23}$$

Såfremt $x \in C\backslash\mathcal{M}$, dvs. såfremt observationen ligger på randen af middelværdirummet, har (1.2.19) ingen løsning, og ekstremumværdien udtrykkes blot som

$$\sup_{\vartheta \in D} l(\vartheta;x) = \sup_{\vartheta \in D}\{\vartheta x - \kappa(\vartheta)\}$$

se fx eksempel 1.2.4. □

Bemærkning 2 ***Lokale egenskaber for deviansen***

For $x \in \mathcal{M}$ har man

$$d(x;x) = 0$$

Ved udnyttelse af (1.2.16) finder man for $x \in \mathcal{M}$ de første to afledede af deviansen

$$\frac{\partial}{\partial\mu} d(x;\mu) = -2\,\frac{x-\mu}{V(\mu)} \tag{1.2.24}$$

$$\frac{\partial^2}{\partial\mu^2} d(x;\mu) = 2\left\{\frac{1}{V(\mu)} + (x-\mu)\frac{V'(\mu)}{V(\mu)^2}\right\} \tag{1.2.25}$$

Tilsvarende finder man for $x \in \mathcal{M}$ den afledede af deviansen med hensyn til x som

$$\begin{aligned} \frac{\partial}{\partial x} d(x;\mu) &= 2\{\tau^{-1}(x) - \tau^{-1}(\mu)\} \\ \frac{\partial^2}{\partial x^2} d(x;\mu) &= \frac{2}{V(\mu)}, \end{aligned}$$

der viser, at $d(x;\mu)$ er en konveks funktion af $x \in \mathcal{M}$.

Specielt har man for $x = \mu$, at

$$\begin{aligned} d(\mu;\mu) &= 0\,, \\ \frac{\partial}{\partial x}\, d(\mu;\mu) &= \frac{\partial}{\partial \mu}\, d(\mu;\mu) = 0 \\ \frac{\partial^2}{\partial x^2}\, d(\mu;\mu) &= = \frac{\partial^2}{\partial \mu^2}\, d(\mu;\mu) = \frac{2}{V(\mu)}\,. \end{aligned}$$

□

Bemærkning 3 ***Taylorudvikling af deviansen***

Ved en Taylorudvikling af deviansen omkring $x = \mu$ finder man jvf ovenstående:

$$d(x;\mu) \approx \frac{(x-\mu)^2}{V(\mu)}\,. \tag{1.2.26}$$

□

Sætning 1.2.5 ***Relation mellem variansfunktion og devians***

Betragt en endimensional naturlig eksponentiel familie med kumulantfrembringer $\kappa(\cdot)$, variansfunktionen $V(\cdot)$, og med deviansen givet ved (1.2.20).

Der gælder da, at variansfunktionen kan udtrykkes ved deviansen som

$$V(\mu) = 2 \Big/ \left[\frac{\partial^2}{\partial \mu^2}\, d(x;\mu)\right]_{x=\mu}\,, \tag{1.2.27}$$

og omvendt kan deviansen udtrykkes som

$$d(x;\mu) = 2\int_{\mu}^{x} \frac{x-u}{V(u)}\, du\quad. \tag{1.2.28}$$

for $x \in \mathcal{M}$ og $\mu \in \mathcal{M}$

Bevis:

Udtrykket for variansen følger af (1.2.25), mens udtrykket for enhedsdeviansen følger af (1.2.24). Morris (1982) har givet en yderligere diskussion af disse relationer.

□

Der gælder endvidere

Sætning 1.2.6 ***Tætheden for en endimensional naturlig eksponentiel familie udtrykt ved deviansen***

Betragt en endimensional naturlig eksponentiel familie frembragt af et mål ν på $\mathbb{R}$.

Tætheden med hensyn til ν kan da udtrykkes ved hjælp af deviansen som

$$f(x;\mu) = a(x)\ \exp\left\{-\frac{1}{2}\,d(x;\mu)\right\}\ , \tag{1.2.29}$$

hvor normeringsfaktoren $a(x)$ er bestemt ved

$$a(x) = \exp\left[\sup_{\vartheta\in D}\{\vartheta x - \kappa(\vartheta)\}\right]\ .$$

Bevis:
Se Jørgensen (1997). □

Sætning 1.2.7 ***Fordelingen af summen af uafhængige variable***

Lad $X_1,\ldots,X_n$ være uafhængige identisk fordelte variable, hvis fordeling tilhører en naturlig eksponentiel familie med kumulantfrembringer $\kappa_X(\vartheta)$.

Da tilhører fordelingen af $Z = X_1 + \cdots + X_n$ ligeledes en eksponentiel familie. Kumulantfrembringeren for familien af fordelinger af Z er

$$\kappa_Z(\vartheta) = n\kappa_X(\vartheta)$$

Bevis:
Beviset følger ved at opskrive den simultane tæthed for $X_1,\ldots,X_n$. □

Vi bemærker dog, at støtten for fordelingen af Z ikke nødvendigvis er den samme som støtten for fordelingen af X_i. (Man kan blot tænke på binomialfordelingen som fordelingen af en sum af Bernoullivariable).

Sætning 1.2.8 ***Uendeligt delbarhed af fulde, naturlige eksponentielle familier***

Betragt en fuld, naturlig eksponentiel familie. Såfremt ét medlem af familien er uendeligt delbar, da er ethvert medlem uendeligt delbar.

Bevis:

Følger ved at betragte den naturlige eksponentielle familie frembragt af det uendeligt delbare medlem. Se Barndorff-Nielsen (1978). □

1.2.2 Oversigt over endimensionale naturlige eksponentielle familier

En række af de fordelinger, der blev behandlet i Introduktion til Statistik, Bind 1, kan formuleres som endimensionale naturlige eksponentielle familier. Den følgende tabel giver en oversigt over de vigtigste relationer ved en sådan fortolkning. Fordelingerne er behandlet mere indgående i afsnit 4.

Tabel 1.1. **Kanoniske parametre for nogle naturlige eksponentielle familier**

Fordeling	Normal $\mathrm{N}(\alpha,\sigma^2)$	Poisson $\mathrm{P}(\lambda)$	Gamma $\mathrm{G}(\alpha,\beta)$
Tæthed	$\frac{1}{\sigma\sqrt{2\pi}}\exp\{-(x-\alpha)^2/(2\sigma^2)\}$	$\frac{\lambda^x}{x!}\exp(-\lambda)$	$\frac{1}{\beta\Gamma(\alpha)}(x/\beta)^{\alpha-1}\exp(-x/\beta)$
Støtte	$x\in\mathbb{R}$	$x=0,1,2,\ldots$	$x\in\mathbb{R}_+$
Parametre	$\alpha\in\mathbb{R}$	$\lambda\in\mathbb{R}_+$	$\alpha\in\mathbb{R}_+$
			$\beta\in\mathbb{R}_+$
Elementar fordeling	$\mathrm{N}(\alpha,1)$	$\mathrm{P}(\lambda)$	$\mathrm{Ex}(\beta)$
Kan. parameter ϑ	α/σ^2	$\ln(\lambda)$	$-1/\beta$
Kan. parameteromr D	$\mathbb{R}$	$\mathbb{R}$	$(-\infty,0)$
$\kappa(\vartheta)$	$\sigma^2\vartheta^2/2=\alpha^2/(2\sigma^2)$	$\exp(\vartheta)=\lambda$	$-\alpha\ln(-\vartheta)=\alpha\ln(\beta)$
Middelværdi			
$\mu=\kappa'(\vartheta)$	$\alpha=\vartheta\sigma^2$	$\lambda=\exp(\vartheta)$	$\alpha\beta=-\alpha/\vartheta$
Ω	$\mathbb{R}$	$\mathbb{R}_+$	$\mathbb{R}_+$
Kanonisk link, g	identitet	log	reciprok
$\vartheta=a+b\,g(\mu)$			
$V(\mu)=\kappa''(\vartheta)$	σ^2	μ	μ^2/α
$\ln(\mathrm{E}[\exp(tX)])$			
$=\kappa(t+\vartheta)-\kappa(\vartheta)$	$t\alpha+t^2\sigma^2/2$	$\lambda(\exp(t)-1)$	$-\alpha\ln(1-\beta t)$
Ortogonal pol.	Hermite	Poisson-Charlier	Gen. Laguerre
Grænseford.	Normal	Normal	Normal

Tabel 1.1. **Kanoniske parametre for naturlige eksponentielle familier (fortsat)**

Fordeling	Binomial $B(n,p)$	Negativ Bin $NB^*(\alpha,p)$	Gen Hyp Secant $GHS(\alpha,\lambda)$
Tæthed	$\binom{n}{x} p^x(1-p)^{n-x}$	$\frac{\Gamma(\alpha+x)}{\Gamma(\alpha)\,x!}\, p^x(1-p)^\alpha$	
Støtte	$x = 0, 1, 2, \ldots, n$	$x = 0, 1, 2, \ldots$	$x \in \mathbb{R}$
Parametre	$0 < p < 1$	$0 < p < 1$	$\lambda \in \mathbb{R}$
	$n = 1, 2, \ldots$	$\alpha \in \mathbb{R}_+$	$\alpha \in \mathbb{R}_+$
Elementar fordeling	Bernoulli $B(1,p)$	Geometrisk $G(1,p)$	Gen.Hyp. secant $GHS(1,\lambda)$
Kan. parameter ϑ	$\ln\{p/(1-p)\}$	$\ln(p)$	$\arctan(\lambda)$
Kan. parameteromr D	$\mathbb{R}$	$(-\infty, 0)$	$(-\pi/2, \pi/2)$
$\kappa(\vartheta)$	$n\ln\{1+\exp(\vartheta)\}$	$-\alpha\ln\{1-\exp(\vartheta)\}$	$(\alpha/2)\ln(1+\lambda^2/2)$ $-\alpha\ln\{\cos(\vartheta)\}$
Middelværdi $\mu = \kappa'(\vartheta)$	np	$\alpha p/(1-p)$	$\alpha\lambda$
	$n/\{1+\exp(-\vartheta)\}$	$\alpha/\{\exp(-\vartheta)-1\}$	$\alpha\,\tan(\vartheta)$
Ω	$(0, n)$	$\mathbb{R}_+$	$\mathbb{R}$
Kanonisk link, g $\vartheta = a + b\,g(\mu)$	logit	$\ln(p/(1+p))$	arctan
$V(\mu) = \kappa''(\vartheta)$	$np(1-p)$	$\alpha p/(1-p)^2$	$\alpha(1+\lambda^2)$
	$n\exp(\vartheta)/\{1+\exp(\vartheta)\}^2$		
	$-\mu^2/n + \mu$	$\mu^2/\alpha + \mu$	$\mu^2/\alpha + \alpha$
$\ln(\mathrm{E}[\exp(tX)])$ $= \kappa(t+\vartheta) - \kappa(\vartheta)$	$n\ \ln\{p\exp(t) + 1 - p\}$	$\alpha\ \ln\left\{\frac{1-p}{1-p\exp(t)}\right\}$	$-\alpha\ln\{\cos(t) - \lambda\sin(t)\})$
Ortogonal pol.	Krawtchouk	Meixner	Pollaczek
Grænseford.	Normal	Normal	Normal
	Poisson	Poisson	Gamma
		Gamma	

Tabel 1.1. **Kanoniske parametre for naturlige eksponentielle familier (fortsat)**

Fordeling	Invers Gauss $\mathrm{IG}(\mu, \lambda)$
Tæthed	$\left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right\}$
Støtte	$x \in \mathbb{R}_+$
Parametre	$\mu \in \mathbb{R}_+$
	$\lambda \in \mathbb{R}_+$
Kan. parameter ϑ	$-\lambda/(2\mu^2)$
Kan. parameteromr D	$(-\infty, 0]$
$\kappa(\vartheta)$	$-\sqrt{-2\lambda\vartheta}$
Middelværdi	μ
$\mu = \kappa'(\vartheta)$	$\sqrt{\lambda/(-2\vartheta)}$
Ω	$\mathbb{R}_+$
Kanonisk link, g	$1/\mu^2$
$V(\mu) = \kappa''(\vartheta)$	μ^3/λ
$\ln(\mathsf{E}[\exp(tX)])$ $= \kappa(t+\vartheta) - \kappa(\vartheta)$	$-\frac{\lambda}{\mu^2}\{\sqrt{1-2t\mu^2/\lambda} - 1\}$

1.2.3 Generel eksponentiel familie

For fuldstændighedens skyld introducerer vi en generel eksponentiel familie.

En generel eksponentiel familie fremkommer af en naturlig eksponentiel familie ved transformation af de variable og/eller en reduktion af indeksmængden.

Definition 1.2.9 ***Generel eksponentiel familie, kanonisk stikprøvefunktion, kanonisk parameter***

Betragt en familie af fordelinger med tætheder $\{f(x;\omega)\}_{\omega\in\Omega}$, med hensyn til et mål $\nu\{\cdot\}$ på $\mathbb{R}^m$.

Såfremt tæthederne kan skrives på formen

$$f(x;\omega) = b(x)\exp\{\vartheta(\omega)^T t(x) - \kappa(\vartheta(\omega)) \quad \text{for } \omega \in \Omega\} \tag{1.2.30}$$

hvor ϑ og t er af dimension k, kaldes familien for en generel eksponentiel familie, eller blot en eksponentiel familie.

De k-dimensionale størrelser, $\vartheta(\omega)$ og $t(x)$ kaldes henholdsvis for den kanoniske parameter og den kanoniske stikprøvefunktion.

Det mindste heltal, k, for hvilket man kan udtrykke familien på formen (1.2.30), hvor $\vartheta(\omega)$ og $t(x)$ er k-dimensionale, kaldes for familiens orden.

Såfremt størrelserne $\vartheta(\omega)$ og $t(x)$ i repræsentationen (1.2.30) netop har samme dimension som familiens orden, siges repræsentationen at være minimal, og den kanoniske parameter og den kanoniske stikprøvefunktion kaldes henholdsvis for den minimale kanoniske parameter og den minimale kanoniske stikprøvefunktion.

Hvis den minimale kanoniske stikprøvefunktion netop er identiten, $t(x) \equiv x$ kaldes familien for lineær.

I det følgende vil vi antage, at funktionen $\vartheta(\omega)$ er en enentydig funktion af $\omega \in \Omega$. Hvis dette ikke var tilfældet, ville parametriseringen ved ω være udtryk for en overparametrisering, som er uden relevans for de generelle egenskaber.

Vi vil derfor alene betragte parametriseringen ved den kanoniske parameter $\vartheta = \vartheta(\omega)$. Parameterområdet for den kanoniske parameter er billedmængden $\Theta = \vartheta(\Omega)$ af variationsområdet for ϑ. □

En række af de begreber, der blev indført for de naturlige eksponentielle familier, overføres umiddelbart til generelle eksponentielle familier.

Således indfører vi det mulige variationsområde for den kanoniske parameter, Lad

$$D = \left\{ \vartheta \in \mathbb{R}^k : \int b(x) \exp\left(\vartheta^T t(x)\right) \nu\{dx\} < \infty \right\} \tag{1.2.31}$$

Også her gælder det, at mængden D er konveks.

Definitionen af støtten, S, og den konvekse støtte, C , på side (29) overføres umiddelbart til generelle eksponentielle familier.

Definition 1.2.10 ***Fuld og regulær eksponentiel familie***

Betragt den eksponentielle familie af fordelinger med tætheder (1.2.30) og med parameterområdet $\Theta = \vartheta(\Omega)$ for den kanoniske parameter.

Såfremt familien udnytter hele det mulige parameterområde, D, bestemt ved (1.2.31), dvs hvis $\Theta = D$, siges familien at være fuld (engelsk: *full*).

Familien siges at være regulær, hvis den er fuld, og hvis desuden Ω er en åben delmængde af $\mathbb{R}^k$.

Hvis Ω er en differentiabel flade i D, siges familien at være en krum eksponentiel familie. □

I lighed med definition (1.2.3) kaldes funktionen $\kappa(\cdot)$ for kumulantfrembringeren for familien (1.2.30).

I lighed med sætning 1.2.1 gælder

Sætning 1.2.9 ***Momentfrembringende og kumulantfrembringende funktion***

Betragt den eksponentielle familie af fordelinger med tætheder (1.2.30). Såfremt familien er fuld (dvs udfylder hele det mulige parameterområde, D), er den momentfrembringende funktion for fordelingen af T givet ved

$$M(s;\vartheta) = \exp\{\kappa(\vartheta + s) - \kappa(\vartheta)\} \quad \text{for } s \in D - \vartheta \ ,$$

og den kumulantfrembringende funktion for fordelingen af T er

$$K(s;\vartheta) = \kappa(\vartheta + s) - \kappa(\vartheta) \quad \text{for } s \in D - \vartheta \tag{1.2.32}$$

Såfremt $\vartheta \in \text{int}(D)$ eksisterer alle momenter i fordelingen af T, og den i'te kumulant for fordelingen af T er

$$\kappa_i(\vartheta) = \mathrm{D}^i \kappa(\vartheta) \ ,$$

hvor D angiver den sædvanlige differentialoperator, og i er et k-dimensionalt sæt af ikke-negative heltal, $i = (i_1, \ldots, i_k)$.

For $\vartheta \in \text{int}(D)$ gælder altså specielt, at

$$\mathsf{E}[T] = \frac{\partial}{\partial \vartheta}\,\kappa(\vartheta) \tag{1.2.33}$$

og

$$\mathbf{D}[T] = \frac{\partial^2}{\partial \vartheta^T\,\partial \vartheta}\,\kappa(\vartheta)\,, \tag{1.2.34}$$

hvor $\mathbf{D}[T]$ er positiv definit.

Bevis:
Følger i lighed med beviset for sætning 1.2.1 til 1.2.3. □

For en fuld eksponentiel familie definerer man middelværdiafbildningen i lighed med definition 1.2.4.

Sætning 1.2.10 ***Den kanoniske stikprøvefunktion er sufficient***

Betragt den eksponentielle familie af fordelinger med tætheder (1.2.30).

Den kanoniske stikprøvefunktion, $t = t(x)$ er sufficient for θ. Hvis repræsentationen (1.2.30) er minimal, er t minimal sufficient, dvs at der ikke findes nogen stikprøvefunktion $t_1(x)$ med dimension mindre end k, som er sufficient for θ.

Bevis:
Følger af definition på sufficiens (se Introduktion til Statistik, Bind 1) □

Hvis $\theta \in \mathrm{int}(\Theta)$, kan man parametrisere familien ved middelværdiparameteren

Ved middelværdimængden for familien vil vi forstå mængden

$$\mathcal{M} = \{\mu \in \mathbb{R}^k \,:\, \int x b(x) \exp\{\theta(\omega)^T t(x) - \kappa(\theta(\omega))\}\nu\{dx\}\}$$

For enhver eksponentiel familie gælder at

$$\mathcal{M} \subseteq \mathrm{int}(C)$$

hvor $\mathrm{int}(C)$ angiver det indre af C.

Definition 1.2.11 *Likelihood-uafhængighed*

Betragt en situation, hvor parameteren θ er flerdimensional, og opdelt i komponenterne $(\theta^{(1)}, \theta^{(2)})$.

Såfremt likelihoodfunktionen for $(\theta^{(1)}, \theta^{(2)})$ kan faktoriseres i et produkt

$$L(\theta^{(1)}, \theta^{(2)}; \mathbf{y}) = L_1(\theta^{(1)}; \mathbf{y})\, L_2(\theta^{(2)}; \mathbf{y})$$

af to funktioner, hvor den ene alene afhænger af $\theta^{(1)}$, og den anden kun afhænger af $\theta^{(2)}$, siges komponenterne $\theta^{(1)}$ og $\theta^{(2)}$ at være likelihood-uafhængige.

Det er klart, at såfremt log-likelihoodfunktionen for $(\theta^{(1)}, \theta^{(2)})$ kan skrives som en sum

$$l(\theta^{(1)}, \theta^{(2)}; \mathbf{y}) = l_1(\theta^{(1)}; \mathbf{y}) + l_2(\theta^{(2)}; \mathbf{y})$$

af to funktioner, hvor den ene alene afhænger af $\theta^{(1)}$ og den anden kun afhænger af $\theta^{(2)}$, da er $\theta^{(1)}$ og $\theta^{(2)}$ likelihood-uafhængige. □

1.3 Eksponentielle dispersionsmodeller

1.3.1 Indledning

En eksponentiel familie tillader en simpel behandling af lineære (evt. affine) hypoteser vedrørende den kanoniske parameter.

En væsentlig egenskab ved en naturlig eksponentiel familie er, at familien er fastlagt ved sin variansfunktion, dvs. ved den måde, hvorpå variansen i fordelingen ændrer sig med middelværdien.

Ved en række anvendelser (feks ved analyse af normalt fordelte observationer) vil man opleve, at en endimensional eksponentiel familie er fyldestgørende til at fastlægge den basale variansstruktur, (variansens afhængighed af middelværdien), men at man har behov for at indføre en vægtning af variansen med kendte vægte, der fx afspejler antallet af observationer, eller at man behov for at indføre en fri faktor, σ^2, som karakteriserer skalaen for den aktuelle variansfunktion, dvs $\mathsf{V}[Y] = \sigma^2 V(\mu)$ som fex ved normalfordelingen.

Det er derfor bekvemt at betragte familier, der er lidt rigere, end endimensionale naturlige eksponentielle familier, men som alligevel bevarer den fundamentale struktur i sammenhængen mellem observation y og parameter ϑ.

Vi tager udgangspunkt i en naturlig eksponentiel familie.

Betragt en endimensional naturlig eksponentiel familie, $\mathcal{P}$, med kumulantfrembringer

$$\kappa(\vartheta) = \ln\left(\int \exp(\vartheta z)\nu\{dz\}\right), \tag{1.3.1}$$

hvor $\nu\{\cdot\}$ er et mål[1] på $\mathbb{R}$.

Det kanoniske parameterområde for familien er

$$D = \{\vartheta \in \mathbb{R} \mid \kappa(\vartheta) < \infty\}$$

Såfremt hverken $\nu\{\cdot\}$ eller D er udartede, vil familien af tætheder (mht målet ν)

$$f(z;\vartheta) = \exp\{\vartheta z - \kappa(\vartheta)\} \quad \text{for } \vartheta \in D \tag{1.3.2}$$

være en naturlig eksponentiel familie.

Lad

$$\tau(\vartheta) = \kappa'(\vartheta) \tag{1.3.3}$$

angive middelværdiafbildningen, og lad

$$\mathcal{M} = \tau(\text{int}(D)) \tag{1.3.4}$$

angive middelværdirummet.

Lad nu λ være et reelt positivt tal og betragt udtrykket

$$\lambda\kappa(\vartheta) = \ln\left(\int \exp(\vartheta z)\nu_\lambda^*\{dz\}\right) \tag{1.3.5}$$

[1] Sædvanligvis vil $\nu\{\cdot\}$ være på formen $\nu\{dz\} = c(z)dz$, hvor dz er enten Lebesguemålet (kontinuert tæthed) eller tællemålet (diskret tæthed)

Såfremt der findes et mål ν_λ^*, der opfylder (1.3.5), definerer dette mål en naturlig eksponentiel familie med kumulantfrembringer $\lambda\kappa(\vartheta)$ og med tætheden (med hensyn til ν_λ^*)

$$f^*(z;\vartheta,\lambda) = \exp\{\vartheta z - \lambda\kappa(\vartheta)\} \quad \text{for } \vartheta \in D \tag{1.3.6}$$

Lad Λ angive mængden af positive tal, λ, for hvilke der eksisterer et mål, ν_λ^*, sådan at (1.3.5) er opfyldt.

Definition 1.3.1 ***Additiv og reproduktiv eksponentiel dispersionsmodel***

Familien af fordelinger defineret ved tæthederne (1.3.6) for $(\vartheta,\lambda) \in D \times \Lambda$ kaldes den additive eksponentielle dispersionsmodel frembragt af ν.

Den tilsvarende familie af fordelinger for $Y = Z/\lambda$ kaldes den reproduktive eksponentielle dispersionsmodel frembragt af ν. Sædvanligvis parametriseres en reproduktiv dispersionsmodel ved parametrene $\mu = \tau(\vartheta)$ og $\sigma^2 = 1/\lambda$. Parameteren μ defineres ved kontinuitet på randen af D, evt ved at tillade uendelige værdier.

Parameteren ϑ kaldes den kanoniske parameter, mængden D kaldes det kanoniske parameterområde, parameteren λ kaldes indeksparameteren, og mængden Λ kaldes indeksmængden.

Når den reproduktive dispersionsmodel parametriseres ved parametrene $\mu = \tau(\vartheta)$ og $\sigma^2 = 1/\lambda$, kaldes μ for middelværdiparameteren og σ^2 kaldes for dispersionsparameteren. Indeksmængden for dispersionsparameteren betegnes undertiden med Δ.

Kumulantfrembringeren $\kappa(\cdot)$ kaldes for enhedskumulantfrembringeren og funktionen

$$V(\mu) \stackrel{\text{DEF}}{=} \kappa''(\tau^{-1}(\mu)) \tag{1.3.7}$$

kaldes enhedsvariansfunktionen.

Enhedsvariansfunktionen kan udtrykkes alene ved middelværdiafbildningen, $\tau(\cdot)$, som

$$V(\mu) = \tau'(\tau^{-1}(\mu))$$

□

Vi bemærker, at indeksmængden Λ ikke er tom. Relationen (1.3.5) er jo opfyldt for den oprindelige familie svarende til $\lambda = 1$ og $\nu_\lambda = \nu$.

Eksempel 1.3.1 ***Familien af binomialfordelinger udgør en additiv eksponentiel dispersionsmodel***

Betragt familien af $\mathrm{B}(n,p)$-fordelinger med $n \in \{1,2,\ldots\}$ og $p \in]0,1[$.

Vi så i eksempel 1.2.2, at for ethvert n er familien af $\mathrm{B}(n,p)$-fordelinger en naturlig eksponentiel familie med kanonisk parameter $\vartheta = \ln(p/(1-p))$ og med kumulantfrembringer

$$\kappa_n(\vartheta) = n \ln\left(1 + \exp(\vartheta)\right) .$$

Familien af $\mathrm{B}(n,p)$-fordelinger for $n \in \{1,2,\ldots\}$ er derfor en additiv eksponentiel dispersionsmodel med indeksmængde $\Lambda = \{1,2,\ldots\}$. (Binomialfordelingen er diskuteret nærmere i 4.20. □

Eksempel 1.3.2 ***Familien af normalfordelinger udgør en reproduktiv eksponentiel dispersionsmodel***

Betragt familien af $\mathrm{N}(\mu,\sigma^2)$-fordelinger for $(\mu,\sigma^2) \in \mathbb{R} \times \mathbb{R}_+$.

For enhver fast værdi af σ^2 er familien af $\mathrm{N}(\mu,\sigma^2)$-fordelinger en naturlig eksponentiel familie med kanonisk parameter $\vartheta = \mu/\sigma^2$ og med kumulantfrembringer

$$\kappa_{\sigma^2}(\vartheta) = \frac{\sigma^2\vartheta^2}{2} ,$$

der netop er $(1/\sigma^2)$ gange kumulantfrembringeren for den naturlige eksponentielle familie af $\mathrm{N}(\mu,1)$-fordelinger (se afsnit 3.1). Man har altså indeksparameteren $\lambda = 1/\sigma^2$ og indeksmængden er $\Lambda = \mathbb{R}_+$ svarende til at værdimængden for dispersionsparameteren σ^2 er $\Delta = \mathbb{R}_+$ □

Bemærkning 1 ***Tætheden svarende til en eksponentiel dispersionsmodel***

Hvis målene ν_λ^*, der optræder i (1.3.5), kan skrives på formen

$$\nu_\lambda^*\{dz\} = c^*(z;\lambda)\, \nu^*\{dz\} ,$$

hvor målet ν^* (der sædvanligvis vil være Lebesgue-målet eller tællemålet) ikke afhænger af λ, kan tæthederne (mht ν^*) for den additive dispersionsmodel udtrykkes som

$$f^*(z;\vartheta,\lambda) = c^*(z;\lambda)\ \exp\{\vartheta z - \lambda\kappa(\vartheta)\} \quad \text{for } z \in \mathbb{R} \tag{1.3.8}$$

Betragter man udtrykket (1.3.6) for tætheden (mht målet ν_λ^*) for z, kan man bestemme et tilsvarende udtryk for tætheden for $Y = Z/\lambda$ (mht et mål ν_λ)

$$f(y;\vartheta,\lambda) = \exp\left[\lambda\{\vartheta y - \kappa(\vartheta)\}\right] \quad \text{for } \vartheta \in D$$

Hvis det nu gælder, at målene ν_λ på tilsvarende måde kan udtrykkes som

$$\nu_\lambda\{dy\} = c(y;\lambda)\ \nu\{dy\}\ ,$$

hvor målet ν ikke afhænger af λ, kan tætheden for Y (med hensyn til ν) udtrykkes som

$$f(y;\vartheta,\sigma^2) = c(y;\sigma^2)\ \exp\left[\frac{\vartheta y - \kappa(\vartheta)}{\sigma^2}\right] \quad \text{for } y \in \mathbb{R}\ , \tag{1.3.9}$$

hvor $\vartheta = \tau^{-1}(\mu)$ og $\lambda = 1/\sigma^2$. □

Bemærkning 2 ***Diskrete fordelinger kan ikke repræsenteres som reproduktive eksponentielle dispersionsmodeller***

Hvis ét medlem af en additiv eksponentiel dispersionsmodel er en diskret fordeling med ækvidistante værdier, da kan man vise, at alle medlemmer af modellen vil være diskrete fordelinger med samme afstand mellem værdierne, og så kan den ikke være reproduktiv. Eksempelvis vil $\mathrm{B}(10,p)/10$ antage værdierne $0, 0.1, 0.2, \ldots, 1.0$, mens $\mathrm{B}(11,p)/11$ antager værdierne $0, 1/11, \ldots, 1$.

Omvendt, hvis det for en værdi, λ_0, af indeksparameteren gælder at det tilsvarende medlem af en additiv eksponentiel dispersionsmodel er en kontinuert fordeling, da vil fordelingerne være kontinuerte for $\lambda > \lambda_0$.

En familie af kontinuerte fordeling kan repræsenteres både som en reproduktiv og som en additiv dispersionsmodel. Oftest vælger man kun at repræsentere den på den reproduktive form. □

Sætning 1.3.1 ***Moment- og kumulantfrembringende funktion for elementer i en eksponentiel dispersionsmodel***

Antag, at fordelingen af Z kan beskrives ved en additiv eksponentiel dispersionsmodel med enhedskumulantfrembringer $\kappa(\cdot)$ og det kanoniske parameterområde D.

Da er den momentfrembringende funktion for fordelingen svarende til parametrene (ϑ, λ) givet ved

$$M(t;\vartheta,\lambda) = \exp\left(\lambda\{\kappa(\vartheta+t) - \kappa(\vartheta)\}\right)$$

og den kumulantfrembringende funktion er

$$K(t;\vartheta,\lambda) = \lambda\{\kappa(\vartheta+t) - \kappa(\vartheta)\}$$

for $t \in D - \vartheta$

Tilsvarende, hvis fordelingen af Y kan beskrives ved en reproduktiv eksponentiel dispersionsmodel med enhedskumulantfrembringer $\kappa(\cdot)$ og det kanoniske parameterområde D, da er den momentfrembringende funktion for fordelingen svarende til parametrene (μ, σ^2) givet ved

$$M(t;\vartheta,\sigma^2) = \exp\left(\{\kappa(\vartheta+\sigma^2 t) - \kappa(\vartheta)\}/\sigma^2\right)$$

og den kumulantfrembringende funktion er

$$K(t;\vartheta,\sigma^2) = \{\kappa(\vartheta+\sigma^2 t) - \kappa(\vartheta)\}/\sigma^2$$

for $t \in (D-\vartheta)/\sigma^2$ og med $\vartheta = \tau^{-1}(\mu)$.

Bevis:

Beviset følger ved at bemærke, at for fastholdt λ er familien af fordelinger af Z en naturlig eksponentiel familie med kumulantfrembringer $\lambda\kappa(\cdot)$. Resultatet for den reproduktive model fås ved at bemærke, at modellen fremkommer ved at sætte $Y = Z/\lambda$ med $\lambda = 1/\sigma^2$. □

Sætning 1.3.2 ***Forventningsværdi og varians for variable fra eksponentiel dispersionsmodel.***

Såfremt fordelingen af Z kan beskrives ved en additiv eksponentiel dispersionsmodel med enhedskumulantfrembringer $\kappa(\cdot)$, middelværdiafbildning $\tau(\cdot) = \kappa(\cdot)$ og enhedsvariansfunktion $V(\mu) = \tau'(\tau^{-1}(\mu))$, da gælder

$$\mathsf{E}[Z] = \xi \quad \text{og} \quad \mathsf{V}[Z] = \lambda\, V(\xi/\lambda)\,, \tag{1.3.10}$$

hvor $\xi = \lambda\, \tau(\vartheta) \in \lambda\, \mathcal{M}$

Tilsvarende, hvis fordelingen af Y kan beskrives ved en reproduktiv eksponentiel dispersionparametermodel med enhedskumulantfrembringer $\kappa(\cdot)$, middelværdiafbildning $\tau(\cdot) = \kappa(\cdot)$ og enhedsvariansfunktion $V(\mu) = \tau'(\tau^{-1}(\mu))$, da gælder

$$\mathsf{E}[Y] = \mu \quad \text{og} \quad \mathsf{V}[Y] = \sigma^2\, V(\mu)\,, \tag{1.3.11}$$

hvor $\mu = \tau(\vartheta) \in \mathcal{M}$.
Bevis:
Følger direkte af sætning 1.3.1. □

Bemærkning 1 ***Alternativt udtryk for forventningsværdi og varians for additiv eksponentiel dispersionsmodel***

Indfører vi symbolet μ for billedet af middelværdiafbildningen, $\mu = \tau(\vartheta)$, kan vi udtrykke forventningsværdi og varians svarende til en additiv eksponentiel dispersionsmodel (1.3.10) som

$$\mathsf{E}[Z] = \lambda\, \mu \quad \text{og} \quad \mathsf{V}[Z] = \lambda\, V(\mu)\,, \tag{1.3.12}$$

hvor $\mu = \tau(\vartheta) \in \mathcal{M}$.

Denne brug af symbolet μ, der jo sædvanligvis er reserveret til at betegne forventningsværdien af den variable, kan forsvares, idet μ jo her angiver forventningsværdien af en "enhedsobservation" (svarende til $\lambda = 1$). □

Bemærkning 2 ***Dispersionsparameteren angiver en fri faktor til variansen***

Det fremgår af udtrykket (1.3.11), at dispersionsparameteren σ^2 i en reproduktiv eksponentiel dispersionsmodel angiver en fri faktor til variansen udover den del $V(\mu)$, der er bestemt af middelværdien μ. □

Bemærkning 3 ***En eksponentiel dispersionsmodel er fastlagt ved enhedsvariansfunktionen*** $V(\mu)$

Vi minder om, at en eksponentiel dispersionsmodel er frembragt ud fra enhedskumulantfrembringeren, $\kappa(\cdot)$ for en naturlig eksponentiel familie. Men den naturlige eksponentielle familie er jo netop fastlagt ved sin variansfunktion (sætning 1.2.4).

En eksponentiel dispersionsmodel er således fastlagt, når blot variansfunktionen $V(\mu)$ er specificeret.

Det enkelte medlem i modellen er fastlagt, når middelværdiparameteren μ og dispersionsparameteren σ^2 er fastlagt. □

1.3.2 Enhedsdevians

Definition 1.3.2 ***Enhedsdevians for eksponentiel dispersionsmodel***

Betragt en reproduktiv eksponentiel dispersionsmodel med enhedskumulantfrembringer $\kappa(\cdot)$ og med den konvekse støtte C.

For fastholdt dispersionsparameter, σ^2, er modellen en naturlig eksponentiel familie. Vi kan derfor derfor betragte deviansen (def. 1.2.8) svarende til en observation $y \in C$ og $\mu \in \mathcal{M}$. Vi definerer enhedsdeviansen, $d(y;\mu)$ som deviansen svarende til dispersionsparameteren $\sigma^2 = 1$, dvs

$$d(y;\mu) \stackrel{\text{DEF}}{=} 2\Big[\sup_{\vartheta \in D}\{\vartheta y - \kappa(\vartheta)\} - y\tau^{-1}(\mu) + \kappa\big(\tau^{-1}(\mu)\big)\Big] \tag{1.3.13}$$

for $y \in C$ og $\mu \in \mathcal{M}$.

I lighed med (1.2.23) finder vi, at for $y \in \mathcal{M}$ kan enhedsdeviansen udtrykkes som

$$d(y;\mu) = 2\big[y\{\tau^{-1}(y) - \tau^{-1}(\mu)\} - \{\kappa(\tau^{-1}(y)) - \kappa(\tau^{-1}(\mu))\}\big] \,. \tag{1.3.14}$$

Ved udnyttelse af relationen mellem den additive og den reproduktive model, $Y = Z/\lambda$ med $\lambda = 1/\sigma^2$, finder vi tilsvarende enhedsdeviansen for en additiv dispersionsmodel

$$d^*(z,\xi) \stackrel{\text{DEF}}{=} d(z/\lambda;\xi/\lambda) = d(y;\mu) \,, \tag{1.3.15}$$

hvor $y = z/\lambda$ og $\mu = \mathsf{E}[Y]$ jvf (1.3.12) og $d(y;\mu)$ er givet ved (1.3.13). □

Bemærkning 1 ***Momenter for enhedsdeviansen***

Ved benyttelse af bemærkning 2 på side 39 finder man, at såfremt fordelingen af Y tilhører en reproduktiv eksponentiel dispersionsmodel med enhedsdeviansen $d(y;\mu)$, enhedsvariansfunktion $V(\mu)$ og dispersionsparameter σ^2 gælder

$$\mathsf{E}\left[\frac{\partial}{\partial\mu}\, d(Y;\mu)\right] = 0 \tag{1.3.16}$$

$$\mathsf{V}\left[\frac{\partial}{\partial\mu}\, d(Y;\mu)\right] = \frac{4\sigma^2}{V(\mu)} \tag{1.3.17}$$

$$\mathsf{E}\left[\frac{\partial^2}{\partial\mu^2}\, d(Y;\mu)\right] = \frac{2}{V(\mu)} \tag{1.3.18}$$

Såfremt fordelingen af Z følger en additiv eksponentiel dispersionsmodel med indeksparameter λ, gælder udtrykkene (1.3.16) og (1.3.18) ligeledes for $Y = Z/\lambda$, mens (1.3.17) erstattes af

$$\mathsf{V}\left[\frac{\partial}{\partial\mu}\, d(Y;\mu)\right] = \frac{4}{\lambda V(\mu)} \tag{1.3.19}$$

med $Y = Z/\lambda$ □

Såfremt fordelingerne i den eksponentielle dispersionsmodel alle har en tæthed med hensyn til det samme mål som i bemærkningen på side 52, kan man i lighed med sætning 1.2.6 udtrykke tæthederne ved enhedsdeviansen (1.3.13).

Sætning 1.3.3 ***Tætheden svarende til en eksponentiel dispersionsmodel udtrykt ved deviansen***

Antag at fordelingen af Y kan beskrives ved en reproduktiv eksponentiel dispersionsmodel parametriseret ved middelværdiparameteren μ og dispersionsparameteren σ^2 og med enhedsdeviansen $d(y;\mu)$. Da kan tætheden udtrykkes som

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\frac{d(y;\mu)}{2\sigma^2}\right\}, \tag{1.3.20}$$

hvor

$$a(y;\sigma^2) = c(y;\sigma^{-2})\ \exp\left[\sigma^{-2}\sup_{\vartheta\in\Theta}\{\vartheta y - \kappa(\vartheta)\}\right]$$

For en additiv dispersionsparametermodel gælder

$$f(z;\xi,\lambda) = a^*(z;\lambda)\exp\left\{-\frac{\lambda}{2}\,d(z/\lambda;\xi/\lambda)\right\}, \tag{1.3.21}$$

hvor

$$a^*(z;\lambda) = c^*(z;\lambda)\ \exp\left[\lambda\sup_{\vartheta\in\Theta}\{\vartheta z - \kappa(\vartheta)\}\right]$$

Bevis:
Beviset for en reproduktiv familie følger af sætning 1.2.6.

Beviset for en additiv familie følger dernæst ved at indføre parameteren $\xi = \lambda\mu$. □

Bemærkning 1 ***For uafhængige observationer fra eksponentielle dispersionsmodeller med samme indeksparameter, λ eller samme dispersionsparameter, σ^2, afhænger logaritmen til likelihoodfunktionen for μ alene af summen af observationernes devianser.***

Bemærkningen følger af den foregående sætning. Konsekvensen heraf er, at maksimum-likelihood estimatet for μ fås ved at minimere summen af observationernes devianser. □

Bemærkning 2 *Middelværdiparametrisering af en eksponentiel dispersionsmodel*

Ved betragtning af en familie $\mathcal{P}$ af fordelinger for en stokastisk variabel Y som en eksponentiel dispersionsmodel, vil man sædvanligvis benytte en middelværdiparametrisering af familien.

Man siger således at familien $\mathcal{P}$ af fordelinger for Y er en eksponentiel dispersionsmodel med variansfunktion $V(\mu)$ og dispersionsparameter σ^2, hvis tætheden i fordelingen af Y kan skrives som

$$g_Y(y;\mu,\sigma^2) = h(y,\sigma^2)\exp\left[\frac{\tau^{-1}(\mu)y - \kappa(\tau^{-1}(\mu))}{\sigma^2}\right] \tag{1.3.22}$$

hvor

$$\mu = \mathsf{E}[Y]$$

og $\kappa(\cdot)$ og $\tau(\cdot)$ tilfredstiller (1.3.1) og (1.3.3). □

1.3.3 Reproduktions- og grænseegenskaber

Definition 1.3.3 *Vægtfaktor for præcisionen i en eksponentiel dispersionsmodel*

Hvis dispersionsparameteren σ^2 for en eksponentiel dispersionsmodel er udtrykt som

$$\sigma^2 = \frac{{\sigma^2}_0}{w}\,, \tag{1.3.23}$$

hvor w er en kendt størrelse, siger vi at vi har en vægtet model, med dispersionsparameter ${\sigma^2}_0$ og vægt w. □

Baggrunden for at indføre en vægtparameter er den, at man i situationer, hvor man betragter flere observationssæt $Y_1, Y_2, \ldots, Y_k$, foretrækker

at betragte modeller med samme dispersionsparameter σ^2 for alle observationssæt $Y_1, Y_2, \ldots, Y_k$. I nogle situationer er dispersionsparameteren imidlertid proportional med en kendt størrelse (feks. en stikprøvestørrelse). Ved at trække denne proportionalitetsfaktor ud i en såkaldt vægtning, kan man uden at ændre på middelværdistrukturen opnå, at $Y_1, Y_2, \ldots, Y_k$ har samme dispersionsparameter.

Bemærkning 1 ***En vægtning er ikke entydig***

En vægtet model er ikke entydig. Vi opnår samme model, hvis vi multiplicerer dispersionsparameteren med en konstant a, og samtidig multiplicerer vægtene med a. □

Sætning 1.3.4 ***Reproduktivitetsegenskaber for eksponentielle dispersionsparametermodeller***

Lad $Y_1, Y_2, \ldots, Y_k$ være uafhængige stokastiske variable, hvis fordelinger tilhører samme reproduktive eksponentielle dispersionsmodel, og antag, at parametrene for fordelingen af $Y_1, Y_2, \ldots, Y_k$ er hhv. $(\mu, \sigma^2/w_1), \ldots, (\mu, \sigma^2/w_k)$, dvs at de variable har samme middelværdi.

Lad

$$\overline{Y}_w = \frac{1}{w_+} \sum_{i=1}^{k} w_i Y_i \tag{1.3.24}$$

med

$$w_+ = \sum_{i=1}^{k} w_i$$

Da vil fordelingen af $\overline{Y}_w$. tilhøre den samme eksponentielle dispersionsmodel og middelværdien af $\overline{Y}_w$. vil ligeledes være μ. Dispersionsparameteren for fordelingen af $\overline{Y}_w$. er σ^2/w_+.

Bevis:

Følger ved betragtning af den kumulantfrembringende funktion. □

Bemærkning 1 ***Fordeling af gennemsnit af identisk fordelte variable fra eksponentiel dispersionsmodel***

Det følger specielt, at hvis $Y_1, Y_2, \ldots, Y_k$ er uafhængige og identisk fordelt, og fordelingen af Y_i tilhører en eksponentiel dispersionsmodel med variansfunktion V og med middelværdi μ og dispersionsparameter σ^2, da vil fordelingen af gennemsnittet

$$\overline{Y}_\cdot = \frac{1}{k}\sum_{i=1}^{k} Y_i$$

tilhøre den samme eksponentielle dispersionsmodel, og middelværdien af $\overline{Y}_\cdot$ vil ligeledes være μ. Dispersionsparameteren for fordelingen af $\overline{Y}_\cdot$ er σ^2/k, dvs præcisionen er k gange så stor.

Tætheden for $\overline{Y}_\cdot$ er

$$g_{\overline{y}}(\overline{y};\vartheta) = h_{\overline{y}}(\overline{y},k)\exp\{k[\vartheta\overline{y} - \kappa(\vartheta)]/\sigma^2\} \tag{1.3.25}$$

En familie af fordelinger, der er sådan, at fordelingen af stikprøvegennemsnittet af observationer fra en fordeling i familien også selv tilhører familien, kaldes reproduktiv. Ovenstående resultat begrunder således betegnelsen *reproduktiv* for denne klasse af eksponentielle dispersionsmodeller. □

Sætning 1.3.5 ***Additionsegenskaber for eksponentielle dispersionsparametermodeller***
Lad $Z_1, Z_2, \ldots, Z_k$ væe uafhængige stokastiske variable, hvis fordelinger tilhører samme additive eksponentielle dispersionsmodel, og antag, at parametrene for fordelingen af $Z_1, Z_2, \ldots, Z_k$ er hhv. $(\vartheta, \lambda_1), \ldots, (\vartheta, \lambda_k)$, hvor $\lambda_i \in \Lambda$.

Da gælder, at fordelingen af

$$Z_+ = Z_1 + Z_2 + \cdots + Z_k$$

tilhører samme exponentielle dispersionsmodel. Parametrene i fordelingen af Z_+ er ϑ og $\lambda_+ = \lambda_1 + \cdots + \lambda_k$.
Bevis:
Lad enhedskumulantfrembringeren være $\kappa(\cdot)$.

Den kumulantfrembringende funktion for fordelingen af Z_i er da

$$K_i(t;\vartheta) = \lambda_i\{\kappa(\vartheta + t) - \kappa(\vartheta)\} \tag{1.3.26}$$

og man har derfor den kumulantfrembringende funktion for fordelingen af Z_+

$$K(t;\vartheta) = \sum_{i=1}^{k} \lambda_i \{\kappa(\vartheta + t) - \kappa(\vartheta)\} ,$$

der netop er på formen (1.3.26) med λ_i erstattet af λ_+. □

En additiv model er således afsluttet overfor addition af medlemmer i familien med samme værdi af den kanoniske parameter, ϑ, hvilket begrunder, at familien kaldes *additiv*.

Sætning 1.3.6 ***Konvergens mod normalfordelingen***

Såfremt fordelingen af Y kan beskrives ved en reproduktiv eksponentiel dispersionsmodel med middelværdi $\mu_0 + \sigma\mu$, enhedsvariansfunktion $V(\cdot)$, og dispersionsparameter σ^2 for et $\mu_0 \in \mathcal{M}$ og $\mu \in \mathbb{R}$, da vil fordelingen af

$$U = \frac{Y - \mu_0}{\sigma}$$

konvergere mod en $\mathrm{N}(\mu, V(\mu))$-fordeling for $\sigma^2 \to 0$.
Bevis:
Se Jørgensen (1997). □

Sætning 1.3.7 ***Konvergens mod Gammafordelingen***
Antag, at fordelingen af Y_c kan beskrives ved en reproduktiv eksponentiel dispersionsmodel med middelværdi $c\mu$, dispersionsparameter σ^2 og middelværdirummet $\mathcal{M} = \mathbb{R}_+$

Lad enhedsvariansfunktionen være $V(\cdot)$ Såfremt der findes et $c_0 > 0$ sådan at der gælder

$$V(\mu) \approx c_0\mu^2$$

for $\mu \to 0$ eller $\mu \to \infty$, da vil fordelingen af

$$U = \frac{Y_c}{c}$$

konvergere mod en $\mathrm{G}(\alpha, \beta)$-fordeling for $c \to 0$ hhv $c \to \infty$, hvor parametrene α og β er

$$\alpha = \frac{1}{c_0\sigma^2} , \quad \text{og} \quad \beta = \mu \times c_0\sigma^2 .$$

Grænsefordelingen er element i en reproduktiv eksponentiel dispersionsmodel med middelværdi μ, variansfunktion $V(\mu) = \mu^2$, og dispersionsparameter $c_0\sigma^2$ (se 4.9.3).
Bevis:
Se Jørgensen (1997). □

Sætning 1.3.8 ***Konvergens mod Poissonfordelingen***

Lad Z_λ være en diskret stokastisk variabel, der kun kan antage ikke-negative heltallige værdier.

Antag, at fordelingen af Z_λ kan beskrives ved en additiv eksponentiel dispersionsmodel, der tillægger positiv sandsynlighed til værdierne 0 og 1.

Lad den kanoniske parameter være $\vartheta = \tau^{-1}(\xi/\lambda)$ med $\xi > 0$ og lad indeksparameteren være λ. Da vil fordelingen at Z_λ konvergere mod en $\mathrm{P}(\xi)$-fordeling for $\lambda \to \infty$.
Bevis:
Se Jørgensen (1997). □

1.3.4 Oversigt over enhedsvariansfunktioner, dispersionsparametre og enhedsdevianser for sædvanlige eksponentielle dispersionsmodeller

Tabel 1.2 og 1.3 resumerer enhedsvariansfunktion og dispersionsparametre for eksponentielle dispersionsmodeller:

Vi bemærker, at for disse familier gælder at variansen er et polynomium i højst tredie grad af middelværdien. Ses bort fra den inverse Gauss-fordeling, er variansen en (højst) kvadratisk funktion af middelværdien. De anførte seks familier med højst kvadratisk variansfunktion er de eneste endimensionale eksponentielle familier med denne egenskab. Morris (1982) resumerer en række egenskaber for disse seks familier.

For de sædvanlige modeller er enhedsdeviansen $d(y;\mu)$ angivet i tabel 1.4 på side 65.

Tabel 1.2. **Enhedsvariansfunktionen $V(\mu)$ for reproduktive eksponentielle dispersionsparametermodeller**

Fordeling af Y	E[Y] μ	Variansfunktion Betegnelse	Dispersionsparam. σ^2	Ref.
$\mathrm{N}(\mu, \sigma^2)$	μ	$V_N(\mu) = 1$	σ^2	3.1.2
$\mathrm{G}(\alpha, \beta/\alpha)$	β	$V_G(\mu) = \mu^2$	$1/\alpha$	4.9.3
$\mathrm{IG}(\mu, \lambda)$	μ	$V_{IG}(\mu) = \mu^3$	$1/\lambda$	4.2.6
$\mathrm{GHS}(\mu, \sigma^2)$	μ	$V_{GHS}(\mu) = 1 + \mu^2$	σ^2	4.7

Tabel 1.3. **Enhedsvariansfunktionen $V(\mu)$ for additive eksponentielle dispersionsparametermodeller**

Fordeling af Z	Y	E[Y] μ	Variansfunktion Betegnelse	Indeksparam. λ	Ref.
$\mathrm{P}(\mu)$	Z	μ	$V_P(\mu) = \mu$	†	4.24.2
$\mathrm{B}(n, p)$	Z/n	p	$V_{Bin}(\mu) = \mu(1-\mu)$	n	4.20.2
$\mathrm{NB}^*(\alpha, p)$	Z/α	$p/(1-p)$	$V_{NB}(\mu) = \mu(1+\mu)$	α	4.23.2
$\mathrm{G}(\alpha, \beta)$	Z/α	β	$V_G(\mu) = \mu^2$	α	4.9.3
$\mathrm{IG}(\lambda\mu, \lambda^2)$	Z/λ	μ	$V_{IG}(\mu) = \mu^3$	λ	4.2.6
† Middelværdi og indeksparameter kan ikke adskilles					

Tabel 1.4. **Enhedsdevians svarende til sædvanlige endimensionale fordelinger**

Fordeling af Y	$\mathsf{E}[Y]$ μ	Enhedsdevians $d(y;\mu)$
$\mathrm{N}(\mu,\sigma^2)$	μ	$(y-\mu)^2$
$\mathrm{P}(\mu)$	μ	$2\times\{y\ln(y/\mu)-(y-\mu)\}$
$\mathrm{B}(n,p)/n$	p	$2\times\{y\ln(y/\mu)+(1-y)\times\ln((1-y)/(1-\mu))\}$
$\mathrm{NB}^*(n,p)/n$	$p/(1-p)$	$2\times\left\{y\ln\left(\frac{y(1+\mu)}{(1+y)\mu}\right)+\ln\frac{1+\mu}{1+y}\right\}$
$\mathrm{G}(\alpha,\beta/\alpha)$	β	$2\times\{y/\mu-\ln(y/\mu)-1\}$
$\mathrm{IG}(\mu,\lambda)$	μ	$(y-\mu)^2/(y\times\mu^2)$
$\mathrm{GHS}(\mu,\sigma^2)$	μ	$2y\{\arctan(y)-\arctan(\mu)\}+\ln\left(\frac{1+\mu^2}{1+y^2}\right)$

Bemærk: Tabellinien svarende til binomialfordelingen og den negative binomialfordeling vedrører fordelingen af $Y=Z/n$, hvor $Z\in\mathrm{B}(n,p)$, hhv $Z\in\mathrm{NB}^*(n,p)$.

1.4 Referencer

Barndorff-Nielsen, O.E. (1978): *Information and Exponential Families.* Wiley, Chichester.

Feller, W. (1966): *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, New York.

Jørgensen, B. (1997): *The Theory of Dispersion Models.* Chapman & Hall, London.

Letac, G. (1992): *Lectures on Natural Exponential Families and Their Variance Functions.* Monografias de Matematica 50. Instituto de Matematica Pura e Aplicada. Rio de Janeiro.

Morris, C.N. (1982): Natural exponential families with quadratic variance functions. *Ann. Statist.* **10**, 65-80.

Afsnit 2

Beskrivelse af fordelinger af levetider og ventetider

2.1 Generelle begreber

En række fænomener modelleres som kvalitative hændelser, der indtræffer på en tidsakse. Blandt andet på grund af tidsaksens særlige ordningsstruktur er der udviklet et specielt begrebsapparat, knyttet til de specifikke egenskaber ved tidsdimensionen, til beskrivelse af de statistiske egenskaber for sådanne fænomener.

Som eksempler på fænomener, der modelleres som hændelser på tidsaksen, kan nævnes tidspunkt for sammenbrud af en komponent, tidspunkt for reparation af et system, etc.

En væsentlig del af det statistiske begrebsapparat er udviklet i forbindelse med analyse af levetider, dvs tiden, der forløber fra idriftsættelse af en komponent til komponentens sammenbrud, eller tiden, der forløber fra et menneske fødes til det dør. Traditionelt benyttes derfor ofte termer, knyttet til disse anvendelser, som f.eks. levetid, dødsrate etc. I de senere år har teorierne imidlertid fundet anvendelse i andre sammenhænge. I industrielle sammenhænge er det således naturligt at betragte tider for indtræffelse af fejl, eller for overskridelse af givne specifikationer. Den terminologi, der

benyttes i denne fremstilling, vil søge at tilgodese de generelle anvendelser af teorien idet det samtidigt forsøges at tage hensyn til etablerede termer.

Definition 2.1.1 ***Overlevelsesfunktion***

Lad T være en stokastisk variabel med fordelingsfunktionen $F(\cdot)$, hvor $F(t) = \mathsf{P}[T \leq t]$. Overlevelsesfunktionen, $\overline{F}(\cdot)$, for T defineres ved

$$\overline{F}(t) = \mathsf{P}[T > t] = 1 - F(t) \tag{2.1.1}$$

□

Bemærkning 1 ***Pålidelighedsfunktion***
Overlevelsesfunktionen (2.1.1) kaldes undertiden for pålidelighedsfunktionen (engelsk: *Reliability function*) og betegnes med symbolet

$$R(t) = \mathsf{P}[T > t] \, .$$

Betragt et system, der kun har to tilstande: virker/virker ikke. Systemets pålidelighed til tiden t defineres da som som sandsynligheden for at systemet virker til tiden t. Såfremt T beskriver levetiden for systemet vil pålideligheden netop være givet ved $R(\cdot)$. □

Sætning 2.1.1 ***Relation mellem tæthed og overlevelsesfunktion***

Lad fordelingen af T have tæthed $f(\cdot)$;

$$f(t) = F'(t) = \frac{\mathrm{d}}{\mathrm{dt}} F(t) \, .$$

Tætheden kan da udtrykkes ved overlevelsesfunktionen på følgende måde:

$$f(t) = -\frac{\mathrm{d}}{\mathrm{dt}}\overline{F}(t) = -\overline{F}'(t). \tag{2.1.2}$$

Bevis:
Beviset følger umiddelbart. □

Sætning 2.1.2 ***Forventningsværdi udtrykt ved overlevelsesfunktionen***

Lad T have overlevelsesfunktionen $\overline{F}(\cdot)$, og antag at forventningsværdien $\mathsf{E}[T]$ eksisterer. Da gælder:

$$\mathsf{E}[T] = \int_0^\infty \overline{F}(t)dt \tag{2.1.3}$$

Bevis:
Følger ved at betragte

$$\mathsf{E}[T] = \int_0^\infty xf(x)dx = \int_0^\infty f(x)\left[\int_{t=0}^x dt\right]dx$$

Ved ombytning af integrationsrækkefølgen fås da

$$\mathsf{E}[T] = \int_{t=0}^\infty \left[\int_{x=t}^\infty f(x)dx\right]dt = \int_{t=0}^\infty \overline{F}(t)dt$$

□

Sætningen udtrykker, at den gennemsnitlige levetid er arealet, der ligger mellem den vandrette akse og grafen af overlevelsesfunktionen $\overline{F}(\cdot)$.

Egenskaberne ved en almindelig kontinuert stokastisk variabel aflæses i almindelighed af tætheden for den variable, idet tætheden opfattes som en idealiseret hyppighedsfordeling. For en variabel, der beskriver en levetid, eller en tid til fejl, er det imidlertid ofte af større interesse at betragte et statistisk mål for *ældningen*, dvs en størrelse, der beskriver, hvorledes dødsrisikoen, eller fejlrisikoen ændres med tiden.

Et sådant simpelt mål er hændelsesraten.

Definition 2.1.2 ***Hændelsesrate (Hazard rate)***

Lad T være en stokastisk variabel med kontinuert tæthed $f(\cdot)$ og overlevelsesfunktion $\overline{F}(\cdot)$.

Hændelsesraten $\lambda(\cdot)$ defineres som

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathsf{P}[t \leq T < t + \Delta t | T > t]}{\Delta t} = \frac{f(t)}{\overline{F}(t)} \tag{2.1.4}$$

og den kumulerede hændelsesrate $\Lambda(\cdot)$ er givet ved

$$\Lambda(t) = \int_0^t \lambda(u) du \tag{2.1.5}$$

□

Bemærkning 1 ***Fortolkning af hændelsesraten***

Såfremt T udtrykker ventetiden til en hændelse indtræffer, angiver hændelsesraten den infinitesimale sandsynlighed for at hændelsen indtræffer til tiden t, betinget af at den ikke er indtruffet inden t. Såfremt T er en levetid, angiver hændelsesraten således den infinitesemale dødssandsynlighed til tiden t, betinget af at individet har opnået alderen t. Hændelsesraten benævnes også *Failure rate, hazard rate, fejlrate, dødsrate, Force of Mortality.* Vi har her valgt at benytte den neutrale betegnelse hændelsesrate for at kunne benytte en fælles terminologi for undersøgelse af levetider og undersøgelse af punktprocesser i almindelighed. □

Der gælder:

Sætning 2.1.3 ***Relation mellem tæthed, overlevelsesfunktion og hændelsesrate***

Lad T være en ikke-negativ stokastisk variabel med kontinuert tæthed $f(\cdot)$, overlevelsesfunktion $\overline{F}(\cdot)$, og hændelsesrate $\lambda(\cdot)$, da gælder

$$\Lambda(t) = \ln(\overline{F}(t)), \quad \lambda(t) = \frac{-\mathrm{d}\ln\overline{F}(t)}{\mathrm{d}t} \tag{2.1.6}$$

$$\overline{F}(t) = \exp\left(-\int_0^t \lambda(u)du\right) = \exp\big(-\Lambda(t)\big) \tag{2.1.7}$$

$$f(t) = \lambda(t)\exp\left(-\int_0^t \lambda(u)du\right) = \lambda(t)\exp\big(-\Lambda(t)\big) \tag{2.1.8}$$

Bevis:
Sætningen er en direkte følge af definitionerne. □

Sætning 2.1.4 ***Fordeling af realiseret kumuleret hændelsesrate***

Lad T være en positiv stokastisk variabel med kumuleret hændelsesrate $\Lambda(\cdot)$.
Da gælder for den realiserede kumulerede hændelsesrate $Z = \Lambda(T)$:

$$Z \in \mathrm{Ex}(1)$$

Bevis:
Der gælder

$$\mathsf{P}[Z \le z] = \mathsf{P}[\Lambda(T) \le z] = \mathsf{P}[T \le \Lambda^{-1}(z)] = F_T(\Lambda^{-1}(z))$$

$$= 1 - \overline{F}_T(\Lambda^{-1}(z)) = 1 - \exp(-z) = P[\mathrm{Ex}(1) \le z]$$

□

Bemærkning 1 Da $Z = -\ln(\overline{F}(T))$, gælder tilsvarende $-\ln(\overline{F}(T)) \in \mathrm{Ex}(1)$.
Endvidere gælder

$$\mathsf{E}[1/\lambda(T)] = \mathsf{E}[T] \tag{2.1.9}$$

Bevis:
Der gælder

$$\begin{aligned} \mathsf{E}[1/\lambda(T)] &= \int_0^\infty \frac{1}{\lambda(t)} f(t)dt = \int_0^\infty \frac{1}{\lambda(t)} \lambda(t) \exp(-\Lambda(t))dt \\ &= \int_0^\infty \overline{F}(t)dt = \mathsf{E}[T] \end{aligned}$$

□

Såfremt man har observeret, at individet er i live til tiden t, dvs at hændelsen $T > t$ er indtruffet, er det ofte af interesse at beskrive den betingede fordeling af restlevetiden.

Definition 2.1.3 ***Betinget overlevelsesfunktion***

Lad T have overlevelsesfunktionen $\overline{F}(\cdot)$. Den betingede overlevelsesfunktion, $\overline{F}(\cdot|t)$, svarende til at hændelsen ikke er indtruffet til tiden t defineres da ved:

$$\overline{F}(x|t) = \mathsf{P}[T > t + x | T > t] = \overline{F}(t + x)/\overline{F}(t) \tag{2.1.10}$$

□

Der gælder

Sætning 2.1.5 ***Bestemmelse af forventet restlevetid***

Lad T være en stokastisk variabel med overlevelsesfunktion $\overline{F}(\cdot)$ og hændelsesrate $\lambda(\cdot)$, og lad $r(t)$ betegne den forventede restventetid,

$$r(t) = \mathsf{E}[T - t | T > t], \quad \text{for} \quad 0 \leq t < \infty$$

Da gælder

$$r(t) = \int_0^t \overline{F}(u)du \Big/ \overline{F}(t),$$

og omvendt

$$\overline{F}(t) = \frac{r(0)}{r(t)} \exp\left(-\int_0^t \frac{\mathrm{d}u}{r(u)}\right)$$

samt

$$\lambda(t) = \frac{1}{r(t)} + \frac{r'(t)}{r(t)}$$

Bevis:
Beviset følger umiddelbart af definitionerne. □

Definition 2.1.4 ***Strukturfunktion for system*** Betragt et system bestående af n komponenter. Ved strukturfunktionen for systemet vil vi forstå funktionen,

$$\phi(x_1, x_2, \ldots, x_n) \; : \; \{0,1\}^n \to \{0,1\}$$

der til enhver kombination af tilstandene $x_i \in \{0,1\}$ for komponenterne ($x_i = 1$ hvis komponenten virker, $x_i = 0$ hvis komponenten er fejlet) angiver systemets tilstand, $\phi = 1$ hvis systemet virker, $\phi = 0$ hvis systemet er fejlet. □

Definition 2.1.5 ***Monoton strukturfunktion***

En strukturfunktion $\phi(x_1, x_2, \ldots, x_n)$ er *monoton*, hvis den er monotont voksende i ethvert af sine argumenter. □

Sætning 2.1.6 ***Modeller for konkurrerende risici***

Lad $T_1, T_2, \ldots, T_n$ være uafhængige levetider med overlevelsesfunktioner $\overline{F}_1(\cdot), \overline{F}_2(\cdot), \ldots, \overline{F}_n(\cdot)$, hændelsesrater $\lambda_1(\cdot), \lambda_2(\cdot), \ldots, \lambda_n(\cdot)$, og lad

$$T = \min(T_1, T_2, \ldots, T_n)$$

Da har fordelingen af T overlevelsesfunktionen

$$\overline{F}(t) = \prod_{i=1}^{n} \overline{F}_i(t) \tag{2.1.11}$$

og hændelsesraten.

$$\lambda(t) = \sum_{i=1}^{n} r_i(t) \tag{2.1.12}$$

Bevis:
Beviset følger af relationen $[T \geq t] = \bigcap_{i=1}^{n} [T_i \geq t]$ og anvendelse af sætning 2.1.3. □

Bemærkning 1 ***Fortolkning af konkurrerende risici***

Modeller for konkurrerende risici bruges blandt andet til beskrivelse af levetider for systemer, der består af en række serieforbundne komponenter. I en sådan situation er det naturligt at forestille sig, at systemet fejler, når blot én af komponenterne fejler. Systemets levetid er da tiden indtil den første komponent fejler.

Strukturfunktionen for et sådant seriesystem er

$$\phi(x_1, x_2, \ldots, x_n) = \min\{x_1, x_2, \ldots, x_n\} .$$

I epidemiologiske sammenhænge bruges modeller for konkurrerende risici til at modellere dødeligheden under hensyntagen til forskellige dødsårsager, der gensidigt udelukker hinanden. □

2.2 Monotoniegenskaber

Da hændelsesraten beskriver de statistiske egenskaber i forbindelse med ældningen af det betragtede objekt, er det ofte af interesse at forholde sig til monotoniegenskaberne for hændelsesraten, samt hvorvidt disse monotoniegenskaber overføres til systemer, sammensat af sådanne objekter. I den statistiske litteratur optræder en række forskellige karakteriseringer af ældningsegenskaber.

Vi indfører indledningsvist:

Definition 2.2.1 ***Monoton hændelsesrate***

Fordelingen af T siges at have monoton hændelsesrate, hvis $\lambda(t)$ er en monoton funktion af t. Såfremt $\lambda(t)$ er voksende i t, siges T at have voksende hændelsesrate. Tilsvarende siges T at have aftagende hændelsesrate, hvis $\lambda(t)$ er aftagende i t. □

Definition 2.2.2 ***Gennemsnitligt monoton hændelsesrate***

Fordelingen af T siges at have gennemsnitligt monoton hændelsesrate, hvis $\Lambda(t)/t$ er en monoton funktion af t. Hvis $\Lambda(t)/t$ er voksende, siges T at have gennemsnitligt voksende hændelsesrate. Tilsvarende siges T at have gennemsnitligt aftagende hændelsesrate, hvis $\Lambda(t)/t$ er aftagende i t. □

Bemærkning 1 ***Engelske betegnelser***

I den internationale litteratur benytter man betegnelsen IFR *(increasing failure rate)* og IFRA *(increasing failure rate average)* for fordelinger med voksende hændelsesrate og med gennemsnitligt voksende hændelsesrate. Tilsvarende benyttes betegnelsen DFR *(decreasing failure rate)* og DFRA *(decreasing failure rate average)* for fordelinger med aftagende hændelsesrate og med gennemsnitligt aftagende hændelsesrate. □

Definition 2.2.3 ***New better/worse than used***

Fordelingen af T siges at være NBU (engelsk: *new better than used*), såfremt der gælder:

$$\overline{F}(x+y) \leq \overline{F}(x)\overline{F}(y).$$

Fordelingen af T siges at være NWU (engelsk:*new worse than used*), såfremt der gælder:

$$\overline{F}(x+y) \geq \overline{F}(x)\overline{F}(y).$$

□

Definition 2.2.4 ***New better/worse than used in expectation***

Fordelingen af T siges at være NBUE (engelsk: *new better than used in expectation*), såfremt der gælder:

$$r(0) \geq r(t) \quad \text{for ethvert} \quad t \geq 0.$$

Fordelingen af T siges at være NWUE (engelsk: *new worse than used in expectation*), såfremt der gælder:

$$r(0) \leq r(t) \quad \text{for ethvert} \quad t \geq 0.$$

□

Sætning 2.2.1 ***Monotonicitet af betinget overlevelsesfunktion***

For en fordeling med voksende hændelsesrate vil den betingede overlevelsessandsynlighed $\overline{F}(x|t)$ være aftagende i t for ethvert x, og for en fordeling med aftagende hændelsesrate er $\overline{F}(x|t)$ voksende i t.

Bevis:

Sætningen bevises direkte ved at betragte udtrykket for den betingede overlevelsessandsynlighed,

$$\overline{F}(x|t) = \exp\left(-\int_t^{t+x} \lambda(u)du\right)$$

□

Der gælder følgende relation mellem de anførte ældningsbegreber:

Sætning 2.2.2 ***Relation mellem ældningsbegreber***

$$\begin{aligned} T \in \text{IFR} &\Rightarrow T \in \text{IFRA} \Rightarrow T \in \text{NBU} \Rightarrow T \in \text{NBUE} \\ T \in \text{DFR} &\Rightarrow T \in \text{DFRA} \Rightarrow T \in \text{NWU} \Rightarrow T \in \text{NWUE}. \end{aligned}$$

Bevis:

Se Launer (1984). □

Sætning 2.2.3 ***Relation mellem ældningsegenskaber og momenter***

$$\begin{aligned} T \in \text{IFRA} &\Rightarrow \mathsf{V}[T] \leq \mathsf{E}[T]^2 \\ T \in \text{DFRA} &\Rightarrow \mathsf{V}[T] \geq \mathsf{E}[T]^2 \end{aligned}$$

Bevis:

Se Barlow og Proschan (1975). □

Sætning 2.2.4 ***Bevarelse af ældningsegenskaber ved minimumsdannelser***

Lad $T_1, T_2, \ldots, T_n$ angive uafhængige variable, og lad $T = \min\{T_1, T_2, \ldots, T_n\}$. Da gælder:

$$\begin{aligned} T_1, T_2, \ldots, T_n \in \text{IFR} &\Rightarrow T \in \text{IFR} \\ T_1, T_2, \ldots, T_n \in \text{IFRA} &\Rightarrow T \in \text{IFRA} \\ T_1, T_2, \ldots, T_n \in \text{DFR} &\Rightarrow T \in \text{DFR} \\ T_1, T_2, \ldots, T_n \in \text{DFRA} &\Rightarrow T \in \text{DFRA} \end{aligned}$$

Bevis:

Se Barlow and Proschan (1975), side 104. □

Sætning 2.2.5 ***Bevarelse af ældningsegenskaber ved miksturer***

Lad Y være en stokastisk variabel med fordelingen $G(\cdot)$, og lad den betingede fordeling af T givet $Y = y$ have DFR (DFRA) egenskaben for ethvert y, da vil den marginale fordeling af T også have DFR (DFRA) egenskaben.
Bevis:
Se Barlow og Proschan (1975), sætning 4.7. □

Bemærkning 1 ***Fortolkning af miksturer***

Sætningen udsiger at en mikstur af DFR (DFRA) fordelinger igen er en DFR (DFRA) fordeling. En sådan mikstur opleves for eksempel, hvis levetiden af en komponent har en DFR fordeling, når den placeres i bestemte omgivelser. De aggregerede levetidsdata for komponenter placeret i en række forskellige omgivelser vil da ligeledes have en DFR fordeling. □

2.3 Totaltesttidstransform

Sætning 2.3.1 ***Den forventede testtid ved tidsafsluttet prøvning***

Antag, at en enhed sættes til levetidsafprøvning, og at prøvningen afsluttes, når enheden fejler, eller senest ved levetiden t_0. Den forventede prøvningstid er da:

$$\mathsf{E}[\min(T, t_0)] = \int_{t=0}^{t_0} \overline{F}(t)dt \tag{2.3.1}$$

Bevis:
Den forventede prøvetid bestemmes som

$$\begin{aligned}
\mathsf{E}[\min(T, t_0)] &= \int_{x=0}^{t_0} x f(x)dx + t_0 \int_{x=t_0}^{\infty} f(x)dx \\
&= \int_{x=0}^{t_0} f(x) \left[\int_{t=0}^{x} dt\right] dx + \int_{x=t_0}^{\infty} f(x) \left[\int_{t=0}^{t_0} dt\right] dx \\
&= \int_{x=0}^{\infty} f(x) \left[\int_{t=0}^{\min(x,t_0)} dt\right] dx
\end{aligned}$$

Ved ombytning af integrationsrækkefølgen finder vi da

$$\mathsf{E}[\min(T, t_0)] = \int_{t=0}^{t_0} \left[\int_{x=t}^{\infty} f(x)dx \right] dt$$

der netop er (2.3.1). □

Såfremt t_0 er valgt som p-fraktilen i fordelingen af T, finder vi den totale testtid svarende til andelen p af fejlende komponenter som

Definition 2.3.1 ***Total testtidstransform***

Lad T være en ikke-negativ stokastisk variabel med den kumulerede fordelingsfunktion $F(\cdot)$.

Ved den totale testtidstransform svarende til fordelingen $F(\cdot)$ forstås funktionen $H_F^{-1}(p)$ bestemt ved:

$$H_F^{-1}(p) \stackrel{\text{DEF}}{=} \mathsf{E}[\min(T, F^{-1}(p)] = \int_{t=0}^{F^{-1}(p)} \overline{F}(t)dt \qquad (2.3.2)$$

Tilsvarende defineres den skalerede totale testtidstranssform som

$$\phi_F(p) \stackrel{\text{DEF}}{=} \frac{1}{\mu} \; H_F^{-1}(p) \qquad (2.3.3)$$

□

Bemærkning 1 ***Den totale testtidstransform angiver den forventede (dvs den gennemsnitlige) prøvningstid svarende til andelen p af fejlende komponenter.*** □

Bemærkning 2 ***Den skalerede totaltidstesttransform er Lorenz-koncentrationskurven for den reciproke hændelsesrate***

Det følger af (2.1.7) og (2.1.8), at den totale testtidstransform (2.3.2) kan udtrykkes som

$$H_F^{-1}(p) = \int_{t=0}^{F^{-1}(p)} \overline{F}(t)dt = \int_{t=0}^{F^{-1}(p)} \frac{f(t)}{\lambda(t)} \, dt$$

hvilket i forbindelse med (2.1.9) viser, at den skalerede totaltidstesttransform er koncentrationskurven (jvf def. 5.2.4) for den reciproke hændelsesrate $1/\lambda(t)$. □

Den empiriske analog til totaltesttidtransformen (definition 2.3.1) er

Definition 2.3.2 ***Totaltesttid stikprøvefunktionen***

Lad $T_1, T_2, \ldots, T_n$ angive en række uafhængige, levetider, og lad $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ angive de tilsvarende ordnede observationer, og lad

$$Z_k = \sum_{j=1}^{k} (n - j + 1)(T_{(j)} - T_{(j-1)}) \tag{2.3.4}$$

hvor $T_{(0)} = 0$.
De stokastiske variable, $Z_1, Z_2, \ldots, Z_n$ kaldes for totaltesttid stikprøvefunktionen. □

Det følger af Glivenko-Cantelli's lemma i forbindelse med de store tals stærke lov, at såfremt $F(\cdot)$ er strengt voksende, da vil $Z_k \to H_F^{-1}(p)$ ligeligt i $0 \leq p \leq 1$ for $n \to \infty$ og $k/n \to p$.

Totaltesttidstikprøvefunktionen blev introduceret af Epstein og Sobel (1953) i forbindelse med undersøgelser af eksponentialfordelte levetider. Totaltesttidtransformen blev introduceret af Barlow og Campo i (1975). En nyere oversigt over egenskaberne ved totaltesttidstransformen er givet af Bergman og Klefsjö (1984).

Såfremt n apparater med levetider $T_1, T_2, \ldots, T_n$ alle er sat i drift til tiden $t = 0$, da vil Z_k angive den totale levetid på det tidspunkt, hvor det k-te apparat fejler. Størrelsen Z_k kaldes derfor ofte *Total time on test*, eller total testtid.

Såfremt specielt $T_1, T_2, \ldots, T_n$ er uafhængige $\mathrm{Ex}(\beta)$-fordelte, har vi at

$$Z_k \in \mathrm{G}(k, \beta)$$

2.4 Ordning efter hændelsesrate

Vi indleder med at betragte

Eksempel 2.4.1 ***Ordning af Weibull-fordelinger***

Lad $X \in \mathrm{We}(k_1, \beta)$ og $Y \in \mathrm{We}(k_2, \beta)$. Hændelsesraterne for X og Y er da henholdsvis

$$\lambda_X(t) = k_1(t/\beta)^{k_1-1} \quad \text{og} \quad \lambda_Y(t) = k_2(t/\beta)^{k_2-1}$$

Såfremt $1 \leq k_1 < k_2$ vil $\lambda_Y(\cdot)$ vokse hurtigere, end $\lambda_X(\cdot)$.
Vi bemærker også, at

$$F_X^{-1}(F_Y(t)) = (t/\beta)^{k_2/k_1}$$

er en konveks funktion af t. □

Definition 2.4.1 ***Konveks ordning af fordelinger***

Lad X og Y være positive stokastiske variable med de tilsvarende kumulerede fordelingsfunktioner $F_X(\cdot)$ og $F_Y(\cdot)$.

Vi siger da, at Y er konvekst underordnet X, såfremt der gælder

$$F_X^{-1}(F_Y(t))$$

er en konveks funktion af t.
Kort skriver vi $F_Y \overset{c}{<} F_X$.
Tilsvarende indfører vi

Definition 2.4.2 ***Stjerneordning af fordelinger***

Lad X og Y være positive stokastiske variable med de tilsvarende kumulerede fordelingsfunktioner $F_X(\cdot)$ og $F_Y(\cdot)$.

Vi siger da, at Y er stjerneunderordnet X, såfremt der gælder

$$F_X^{-1}(F_Y(t))$$

er en ikke-aftagende funktion af t.
Kort skriver vi $F_Y \overset{*}{<} F_X$. □

Stjerneordningen er åbenbart en svagere ordning end den konvekse ordning. Der gælder:

Sætning 2.4.1 ***Relation mellem konveks ordning og stjerneordning***

Lad X og Y angive stokastiske variable med de tilhørende fordelingsfunktioner $F_X(\cdot)$ og $F_Y(\cdot)$.

1) Såfremt $F_Y \overset{<}{c} F_X$, da gælder $F_Y \overset{<}{*} F_X$.

2) Såfremt $F_Y \overset{<}{*} F_X$ og $\mathsf{E}[X] = \mathsf{E}[Y]$, da gælder

$$L_Y(p) \geq L_X(p) \quad \text{og} \quad G_Y \leq G_X$$

hvor $L_X(\cdot)$ og $L_Y(\cdot)$ betegner Lorenzkurverne (jvf def. 5.2.1)svarende til X og Y, og G_X og G_Y betegner Giniindekset (jvf. def. 5.2.2) svarende til X og Y.

3) Såfremt $X \in \mathrm{Ex}(\beta)$ og

   a) såfremt $F_Y \overset{<}{c} F_X$, da har Y en voksende hændelsesrate (IFR).

   b) såfremt $F_Y \overset{<}{*} F_X$, da har Y en gennemsnitligt voksende hændelsesrate (IFR).

Bevis:
se Barlow og Proschan (1975). □

Sætning 2.4.2 ***Bevarelse af monotoniegenskaber ved monoton strukturfunktion***

Antag at et system med n komponenter har en monoton strukturfunktion. Såfremt komponenterne har uafhængige levetidsfordelinger, der hver for sig er IFRA, da er systemets levetidsfordeling også IFRA.
Bevis:
Barlow og Proschan Sætning 4.2.6 □

2.5 Modeller med proportionale hændelsesrater

Der er en speciel gruppe af levetidsmodeller, der giver mulighed for rimeligt simple analyse, også når data er censurerede. Vi vil her kort introducere disse modeller, der blev beskrevet af Cox (1972).

Definition 2.5.1 ***Proportionale hændelsesrater***

Lad T_1 og T_2 have levetidsfordelingen givet ved $F_1(\cdot)$ og $F_2(\cdot)$, og tilhørende hændelsesrater $\lambda_1(\cdot)$ og $\lambda_2(\cdot)$. T_1 og T_2 siges at have proportionale hændelsesrater (engelsk: *proportional hazard*), såfremt $\lambda_2(t) = C\lambda_1(t)$, hvor C er en positiv konstant.

I det almindelige tilfælde, hvor $T_1, T_2, \ldots, T_n$ har levetidsfordelingerne $F_1(\cdot), F_2(\cdot), \ldots, F_n(\cdot)$ med hændelsesraterne $\lambda_1(\cdot), \lambda_2(\cdot), \ldots, \lambda_n(\cdot)$ siges sættet $T_1, T_2, \ldots, T_n$ at have proportionale hændelsesrater, såfremt der findes en ikke-negativ funktion $\lambda_0(\cdot)$ og et sæt konstanter $\alpha_1, \alpha_2, \ldots, \alpha_n$ således at der gælder

$$\lambda_i(t) = \lambda_0(t) \times \exp(\alpha_i) \quad i = 1, 2, \ldots, n$$

Hændelsesraten for den i'te levetid er baseline hændelsesraten, $\lambda_0(\cdot)$, multipliceret med en konstant, $\exp(\alpha_i)$, der karakteriserer fordelingen af den i'te levetid. □

Sætning 2.5.1 ***Eksponentialfordelte levetider***

Lad $T_1, T_2, \ldots, T_n$ være uafhængige levetider, og lad $T_i \in \mathrm{Ex}(\beta_i)$. Da gælder, at sættet $\{T_i\}_{i=1,2,\ldots,n}$ har proportionale hændelsesrater.
Bevis:
Idet vi har $\lambda_i(t) = 1/\beta_i$, ser vi nemlig, at der gælder

$$\lambda_i(t) = \lambda_0(t) \exp(-\ln \beta_i).$$

□

Sætning 2.5.2 ***Weibull fordelinger med samme formparameter***

Lad $T_1, T_2, \ldots, T_n$ være uafhængige levetider, og antag, at $T_i \in \mathrm{We}(k, \beta_i)$. Da gælder, at sættet $\{T_i\}_{i=1,2,\ldots,n}$ har proportionale hændelsesrater.
Bevis:
Idet vi har $\lambda_i(t) = k \times t^{k-1}/\beta_i^k$, ser vi, at der gælder

$$\lambda_i(t) = \lambda_0(t) \exp(-k \ln \beta_i).$$

med

$$\lambda_0(t) = k \times t^{k-1}$$

□

Sætning 2.5.3 ***For modeller med proportionale hændelsesrater vil graferne for dobbeltlogaritmen til overlevelsesfunktionen være parallelle***

Lad $T_1, T_2, \ldots, T_n$ være uafhængige positive stokastiske variable med overlevelsesfunktioner $\overline{F}_i(\cdot)$ og med proportionale hændelsesrater $\lambda_i(\cdot)$, $i = 1, 2, \ldots, n$. Da vil graferne af $\ln(-\ln \overline{F}_i(t))$ have samme indbyrdes afstand.
Bevis:
Følger umiddelbart □

Bemærkning 1 ***Analogi med den ensidede variansanalysemodel***

Betragt et sæt af positive stokastiske variable,

$$\begin{array}{cccc} T_{11}, & T_{12}, & \ldots, & T_{1n_1} \\ T_{21}, & T_{22}, & \ldots, & T_{2n_2} \\ \vdots & \vdots & \ddots & \vdots \\ T_{21}, & T_{22}, & \ldots, & T_{2n_2} \end{array}$$

hvor T_{ij} har hændelsesraten $\lambda_i(\cdot)$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$.

En model med proportionale hændelsesrater er karakteriseret ved:

$$H_0 : \lambda_i(t) = \alpha_i \times \lambda_0(t) ,$$

dvs at hændelsesraterne fremkommer af hinanden ved multiplikationer med konstant. (I den ensidede variansanalysemodel er det middelværdierne, der fremkommer af hinanden ved addition af en konstant). □

Bemærkning 2 ***Modeller med kovariater***

Ofte har man til hver variabel T_i knyttet et sæt kovariater $\mathbf{z}_i^T = (z_{i1}, z_{i2}, \ldots, z_{ip})$, der "forklarer"forskellene på de variable, nemlig

$$\lambda_i(t) = \lambda_0(t) \exp(\mathbf{z}_i^T \boldsymbol{\beta}) \tag{2.5.1}$$

hvor

$$\boldsymbol{\beta} = \left\{ \begin{array}{c} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_p \end{array} \right\}$$

er en parametervektor. Størrelsen $\exp(\beta_j)$ angiver da den relative ændring i hændelsesraten, når kovariaten z_j ændres én enhed.

Den kumulerede hændelsesrate for T_i er:

$$\Lambda_i(t) = \exp(\mathbf{z}_i^T \boldsymbol{\beta}) \Lambda_0(t),$$

hvor

$$\Lambda_0(t) = \int_0^t \lambda(u) du,$$

Overlevelsesfunktionen svarende til T_i bliver

$$\overline{F}_i(t) = \exp(-\Lambda_i(t)) = [\overline{F}_0(t)]^{\exp(\mathbf{z}_i^T \boldsymbol{\beta})}$$

□

Sætning 2.5.4 ***Ved lineært uafhængige kovariater vil overlevelsesfunktionerne ikke krydse hinanden***

Lad $T_1, \ldots, T_k$ være et sæt stokastisk uafhængige levetider, og antag at hændelsesraten for T_i er givet ved (2.5.1), hvor kovariaterne $\mathbf{z}_i$ og $\mathbf{z}_j$ er lineært uafhængige for $i \neq j$.

Da gælder for de kumulerede hændelsesrater $\Lambda_i(\cdot)$ og $\Lambda_j(\cdot)$ med $i \neq j$:

enten	$\Lambda_i(t) < \Lambda_j(t)$	for alle t
eller	$\Lambda_i(t) > \Lambda_j(t)$	for alle t

For overlevelsesfunktionerne, $\overline{F}_i(\cdot)$ og $\overline{F}_j(\cdot)$ gælder tilsvarende:

enten	$\overline{F}_i(t) < \overline{F}_j(t)$	for alle t
eller	$\overline{F}_i(t) > \overline{F}_j(t)$	for alle t

Bevis:
For to sæt lineært uafhængige vektorer $\mathbf{z}_i$ og $\mathbf{z}_j$ gælder

$$\begin{array}{ll} \text{enten} & \exp(\mathbf{z}_i^T\boldsymbol{\beta}) < \exp(\mathbf{z}_j^T\boldsymbol{\beta}) \\ \text{eller} & \exp(\mathbf{z}_i^T\boldsymbol{\beta}) > \exp(\mathbf{z}_j^T\boldsymbol{\beta}) \end{array}$$

□

2.6 Log-lineære modeller

Antag atter, at der er tilknyttet en række kovariable $z_1, z_2, \ldots, z_p$, og at T_i har tilknyttet værdierne $\mathbf{z}_i' = (z_{i1}, z_{i2}, \ldots, z_{ip})$.

Der knytter sig en særlig interesse til modeller, hvor de kovariate optræder som positions- og skalaparametre i fordelinger knyttet til levetider. Da levetider og ventetider modelleres som positive variable, betragtes oftest modeller, hvor de kovariable optræder som positionsparametre i fordelingen af logaritmen til ventetiden, og hvor de betragtede fordelinger har samme skalaparameter.

Vi indleder med at erindre om den lineære regressionsanalysemodel for normaltfordelte variable:
Antag, at vi har en række stokastiske variable $Y_1, Y_2, \ldots, Y_n$, og at der til hver af disse er knyttet et sæt værdier $\mathbf{z}_i^T = (z_{i1}, z_{i2}, \ldots, z_{ip})$ af de forklarende variable. Den sædvanlige lineære regressionsanalysemodel udtrykkes da som $Y_i \in \mathrm{N}(\mathbf{z}_i^T\boldsymbol{\beta}, \sigma^2)$, svarende til en affin relation for positionsparameteren, og en fælles skalaparameter.

De log-lineære modeller svarer til, at de forklarende variable $\mathbf{z}$ virker på tidsaksen, nemlig at

$$T_i^* = T_i \exp(\mathbf{z}_i^T\boldsymbol{\beta})$$

er uafhængige og identisk fordelte med den fælles hændelsesrate $\lambda_0(\cdot)$, med andre ord, at T_i har hændelsesraten:

$$\lambda(t; \mathbf{z}_i) = \lambda_0\big(t \times \exp(\mathbf{z}_i^T\boldsymbol{\beta})\big) \times \exp(\mathbf{z}_i^T\boldsymbol{\beta})$$

Modellen kaldes også en *accelereret levetidsmodel.*

Såfremt $\lambda_0(t) = \lambda k(\lambda t)^{k-1}$, finder vi

$$\lambda(t; \mathbf{z}_i) = \lambda k(\lambda t)^{k-1} \exp(k\mathbf{z}_i^T\boldsymbol{\beta})$$

og $T \in \mathrm{We}\big(k, \lambda \exp(\mathbf{z}_i^T \beta)\big)$.

Lad nu $Y_i = \ln T_i$; $\alpha = -\ln \lambda$; $\sigma = 1/k$; $\beta^\star = \beta/k$.

Da vil

$$W_i^\star = \frac{Y_i + \mathbf{z}_i' \beta^\star - \alpha}{\sigma} \in \mathrm{Min}_1(0,1)$$

Formuleret på en anden måde:

Såfremt $T_i \in \mathrm{We}(k, \lambda \exp(-\mathbf{z}_i'\beta))$, da gælder at

$$W_i = \ln T_i \in \mathrm{Min}_1(\mathbf{z}_i'\beta + \ln \lambda, 1/k),$$

dvs at W_i følger en Min_1-fordeling med positionsparameter $\mu_1 = \mathbf{z}_i'\beta + \ln \lambda$ og med skalaparameter $\sigma = 1/k$:

$$\frac{\ln T_i + \mathbf{z}_i'\beta + \ln \lambda}{1/k} \in \mathrm{Min}_1(0,1)$$

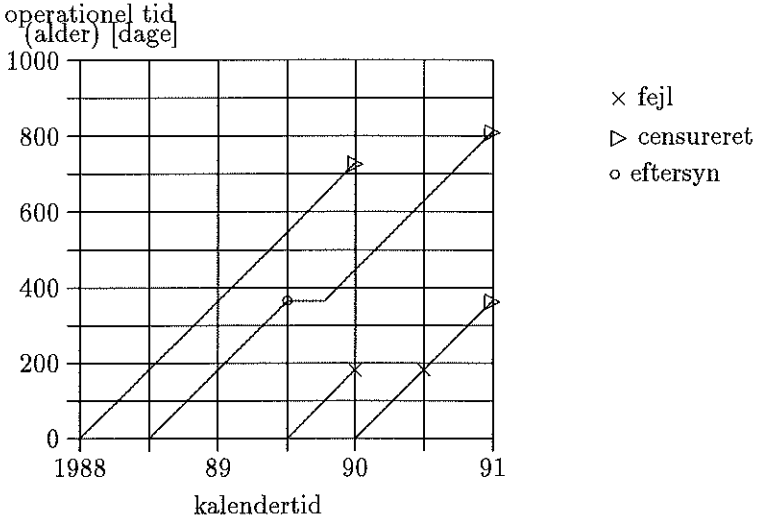
2.7 Censurering

Et problem, der ofte dukker op i forbindelse med analyse af levetider eller af brudstyrker, er, at de praktiske omstændigheder omkring dataindsamlingen bevirker, at en række observationer ikke angives med deres numeriske værdi, men man ved kun, at levetiden T er større end en vis værdi, t_0, eller man har kun observeret de k mindste levetider. Man siger, at disse observationer er censurerede ved tiden t_0, eller ved den k'te hændelse.

2.8 Lexis diagram

Ved analyse af punktprocesser, dvs. hændelser på tidsaksen, kan det ofte være nyttigt at indtegne punkterne i et såkaldt Lexis-diagram, dvs. en afbildning af observationerne med kalendertid ud ad den vandrette akse, og individalder op ad den lodrette akse.

I et sådant diagram vil livshistorien for det enkelte individ repræsenteres ved en linie, der begynder på den vandrette akse (alder = 0) og har hældningen 45°. De hændelser, der indtræffer for individidet, markeres på diagrammet. Linien standser ved individets død, eller ved censurering. Et observationssæt vil ofte være begrænset til at omfatte en kalendertidsperiode, dvs forløbene mellem to lodrette linier i diagrammet.



Figur A.1 - Eksempel på Lexis diagram
(konstruerede data)

System	Dato	Hændelse	Aktion	reparationstid
# 1	01 01 88	Idriftsat	-	-
# 1	31 12 89	Udgået	-	-
# 2	01 07 88	Idriftsat	-	-
# 2	01 07 89	Eftersyn	-	100 dage
# 3	01 07 89	Idriftsat	-	-
# 3	31 12 89	Fejl i XX	Totalskade	-
# 4	01 01 90	Idriftsat	-	-
# 4	01 07 90	Fejl i YY	YY udskiftet	0 dage

2.9 Ikke-parametrisk estimation af overlevelsesfunktion.

Selv om data er censurerede, er det sædvanligvis muligt at foretage en ikke-parametrisk estimation af overlevelsesfunktionen.

Vi henviser indtil videre til litteraturen vedrørende life-table estimation, Kaplan-Meyer (product limit) estimater og Nelson-Aalen estimater.

2.10 Referencer

Barlow, R.E. and Proschan, F. (1975): *Statistical Theory of Reliability and Life Testing,* Holt, Rinehart and Winston, New York

Barlow, R. E. and Campo, R. (1975) Total Time on Test Processes and Applications to Failure Data Analysis, *Reliability and Fault Tree Analysis.* SIAM, Philadelphia.

Bergman, B. and Klefsjö, B. (1984): The Total Time on Test Concept and Its Use in Reliability Theory, *Oper. Res.*, **32**, pp. 596-606

Bryson, M.C. and Siddiqui, M.M. (1969): Some Criteria for Ageing, *Journ. Amer. Statist. Assoc.* pp. 1472-1483.

Chhikara, R.S. and Folks, J.L. (1977): The Inverse Gaussian Distribution as a Lifetime Model, *Technometrics* **18**, pp. 189-193.

Cox, D.R. (1972): Regression Models and Life Tables (with discussion), *Journ. Roy. Statist. Soc. B*, **34**, pp 187-220.

Epstein, L. and Sobel, M. (1953): Life testing, *J. Amer. Statist. Assoc.* **48**, pp. 486-502

Launer, R.L. (1984): Inequalities for NBUE and NWUE Life Distributions, *Operations Research*, **32**, pp 660-667.

Afsnit 3

Normalfordelingen og afledte fordelinger

Vi vil i dette og det følgende afsnit supplere den oversigt over naturlige fordelinger, der er givet i Statistik I, med lidt flere detaljer vedrørende de enkelte fordelinger, og desuden vil vi betragte forskellige levetidsfordelinger samt en række fordelinger, der finder anvendelse ved betragtningen af miksturer.

3.1 Den endimensionale normalfordeling

Vi indleder med at repetere definitionen af den endimensionale normalfordeling:

Definition 3.1.1 *Den endimensionale normalfordeling*

En kontinuert fordelt stokastisk variabel, X, der kan antage alle reelle værdier, siges at være normalfordelt med parametrene μ og σ^2, hvis tætheden for X er på formen

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad \text{for } x \in \mathbb{R} \tag{3.1.1}$$

hvor $\sigma > 0$.

Kort skriver vi $X \in \mathrm{N}(\mu, \sigma^2)$.

□

Lad $X \in \mathrm{N}(\mu, \sigma^2)$. Da gælder

$$\mathrm{E}[X] = \mu; \qquad \mathrm{V}[X] = \sigma^2 \tag{3.1.2}$$

Den karakteristiske funktion for $\mathrm{N}(\mu, \sigma^2)$-fordelingen er

$$\phi(t) = \exp\left(\mathrm{i}\mu t - \frac{\sigma^2 t^2}{2}\right)$$

Sætning 3.1.1 ***Familien af normalfordelinger er afsluttet overfor lineære transformationer***

Lad $X \in \mathrm{N}(\mu, \sigma^2)$ og lad $Y = a + bX$. Da gælder $Y \in \mathrm{N}(a + b\mu, b^2\sigma^2)$.
Bevis:
Følger ved at betragte den karakteristiske funktion. □

Sætning 3.1.2 ***Familien af normalfordelinger er afsluttet overfor addition***

Lad $X \in \mathrm{N}(\mu_1, \sigma_1^2)$ og $Y \in \mathrm{N}(\mu_2, \sigma_2^2)$ være uafhængige. Da gælder

$$X + Y \in \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Bevis:
Følger ved at betragte den karakteristiske funktion. □

3.1.1 Normalfordelingen som eksponentiel familie

Sætning 3.1.3 ***Familien af $N(\mu, \sigma^2)$-fordelinger er en eksponentiel familie af orden 2***

Familien af $N(\mu, \sigma^2)$-fordelinger for $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ udgør en eksponentiel familie af orden 2 med kanonisk parameter

$$\vartheta = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \end{pmatrix} = \begin{pmatrix} \dfrac{\mu}{\sigma^2} \\ \\ -\dfrac{1}{2\sigma^2} \end{pmatrix}, \tag{3.1.3}$$

kanonisk stikprøvefunktion

$$\mathbf{t} = \begin{pmatrix} t_1(x) \\ t_2(x) \end{pmatrix} = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

og kumulantfrembringer

$$\kappa(\vartheta) = \ln\left(\sqrt{\pi}\right) - \frac{1}{2}\ln(-\vartheta_2) - \frac{\vartheta_1^2}{4\vartheta_2}.$$

Det kanoniske parameterområde er $D = \mathbb{R} \times]-\infty, 0[$, som er åbent. Familien er regulær.

Bevis:

Omskrivningen

$$f(x; \mu, \sigma^2) = \frac{\exp\left(-\mu^2/(2\sigma^2)\right)}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu}{\sigma^2}\, x - \frac{1}{2\sigma^2}\, x^2\right) \tag{3.1.4}$$

viser, at tætheden kan udtrykkes på formen (1.2.30). □

Sætning 3.1.4 ***For fastholdt σ^2 er familien af $N(\mu, \sigma^2)$-fordelinger en naturlig eksponentiel familie***

Bevis:

Følger ved at betragte tætheden (3.1.4). □

Kumulantfrembringer mv er angivet i tabel 1.1 på side 43.

3.1.2 Normalfordelingen som eksponentiel dispersionsmodel

Lad $Y \in N(\mu, \sigma^2)$. Vi så i eksempel 1.3.2 på side 52 at familien af $N(\mu, \sigma^2)$-fordelinger for $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ udgør en reproduktiv eksponentiel dispersionsfamilie med kanonisk parameter $\vartheta = \mu$ og med enhedskumulantfrembringer $\kappa(\vartheta) = \vartheta^2/2$. Parameterområdet er $D = \mathbb{R}$ og $\Delta = \mathbb{R}_+$.

Middelværdiafbildningen er Da normalfordelingen netop i forvejen er parametriseret ved sin middelværdi, μ, og da den kanoniske parameter netop er μ, bibringer middelværdiafbildningen

$$\tau(\vartheta) = \kappa'(\vartheta) = \vartheta$$

os ikke noget nyt. Vi bemærker dog, at den kanoniske link,

$$\tau^{-1}(\mu) = \mu$$

er den identiske afbildning.

Enhedsvariansfunktionen er

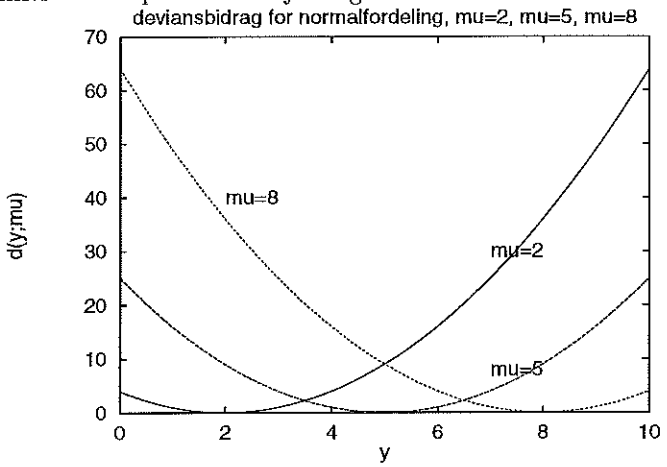
$$V_n(\mu) = \tau'(\tau^{-1}(\mu)) = 1 \, .$$

Variansen afhænger således ikke af middelværdien, men er alene udtrykt gennem dispersionsparameteren σ^2.

Enhedsdeviansen for $\mathrm{N}(\mu, \sigma^2)$ fordelingen er

$$d(y;\mu) = (y-\mu)^2 \tag{3.1.5}$$

Nedenstående figur viser enhedsdeviansen for normalfordelingen svarende til $\mu = 2$, $\mu = 5$ og $\mu = 8$. Det ses at grafen af enhedsdeviansen er symmetrisk omkring $y = \mu$ og at grafen svarende til forskellige vædier af μ fremkommer ved en parallelforskydning.



Den sædvanlige repræsentation af tætheden for en $N(\mu, \sigma^2)$-fordeling er netop udtrykt ved enhedsdeviansen $d(y;\mu)$ på formen (1.3.20) med

$$a(y;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}$$

De vigtigste størrelser ved fortolkningen af familien af $N(\mu, \sigma^2)$-fordelinger som en eksponentiel dispersionsmodel er angivet i nedenstående oversigt:

$N(\mu, \sigma^2)$-fordelingen som reproduktiv eksponentiel dispersionsmodel				
Kanonisk parameter ϑ	Kumulant-frembringer $\kappa(\vartheta)$	Middelværdi-afb. $\mu = \tau(\vartheta)$	Enheds-variansfunktion $V_N(\mu)$	Dispersionsparameter σ^2
μ	$\vartheta^2/2$	ϑ	1	σ^2

3.1.3 Estimation af μ og σ^2

Vi erindrer om resultatet fra Introduktion til Statistik, Bind 1:

Lad $X_1, X_2, \ldots, X_n$ være uafhængige stokastiske variable med $X_i \in N(\mu, \sigma^2)$.

Lad som vanligt $\overline{X}_\cdot = \sum X_i/n$ og $S^2 = \sum(X_i - \overline{X}_\cdot)^2/(n-1)$. Da gælder

$$\overline{X}_\cdot \in N(\mu, \sigma^2/n) \ , \quad S^2 \in \sigma^2\chi^2(n-1)/(n-1)$$

og $\overline{X}_\cdot$ og S^2 er stokastisk uafhængige.

I afsnit 3.7 er fordelingen af S^2 mere indgående behandlet.

3.1.4 Foldet normalfordeling

Definition 3.1.2 ***Foldet normalfordeling***

Lad $X \in N(\mu, \sigma^2)$. Fordelingen af $|X|$ kaldes undertiden en foldet normalfordeling.

Fordelingen af $|X|$ er identisk med en $\sigma\sqrt{\chi^2\left(1,(\mu/\sigma)^2\right)}$-fordeling, hvor $\chi^2\left(1,(\mu/\sigma)^2\right)$ angiver en ikke-central χ^2-fordeling med 1 frihedsgrad og med ikke-centralitetsparameteren $(\mu/\sigma)^2$. □

3.2 Den p-dimensionale normale fordeling

3.2.1 Den todimensionale normalfordeling

Vi indleder med at beskrive den todimensionale normalfordeling

Definition 3.2.1 *Todimensional normalfordeling*

En todimensional stokastisk variabel (X,Y) siges at følge en todimensional normal fordeling med parametrene $(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$, såfremt (X,Y) har tætheden

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}Q(x,y)\right\} \tag{3.2.1}$$

med

$$Q(x,y)=\frac{1}{1-\rho^2}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{x-\mu_y}{\sigma_y}\right)^2\right\}$$

□

Lige som i den endimensionale normalfordeling har parametrene simple fortolkninger. Der gælder:

Sætning 3.2.1 *Momenter i todimensional normalfordeling*

Såfremt (X,Y) følger en todimensional normalfordeling med parametre $(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$ gælder

$$\begin{aligned}\mathsf{E}[X]=\mu_x;\qquad & \mathsf{V}[X]=\sigma_x^2\\ \mathsf{E}[Y]=\mu_y\qquad & \mathsf{V}[Y]=\sigma_y^2\\ \mathsf{COV}[X,Y]\ =\ & \rho\sigma_x\sigma_y\end{aligned} \tag{3.2.2}$$

Endvidere gælder:

$$\begin{aligned} \mathsf{E}[Y|X=x] &= \mu_y + \rho(\sigma_y/\sigma_x)(x-\mu_x) \\ \mathsf{V}[Y|X=x] &= \sigma_y^2(1-\rho^2) \end{aligned} \tag{3.2.3}$$

samt at

$$Y - \mathsf{E}[Y|X] \quad \text{og} \quad X \quad \text{er stokastisk uafhængige}$$

Såfremt (X, Y) følger en todimensional normal fordeling med tætheden (3.2.1) gælder for den kvadratiske form $Z = Q(X, Y)$, at $Z \in \chi^2(2)$.
Bevis:
Følger direkte □

Sætning 3.2.2 ***Familien af todimensionale normale fordelinger er en eksponentiel familie***

Familien af tætheder af formen (3.2.1) med $(\mu_x, \mu_y) \in \mathbb{R}^2$, $-1 < \rho < 1$, og $(\sigma_x, \sigma_y) \in \mathbb{R}_+ \times \mathbb{R}_+$ er en eksponentiel familie af orden 5. Den kanoniske stikprøvefunktion er

$$t = \begin{pmatrix} \sum x_i \\ \sum x_i^2 \\ \sum y_i \\ \sum y_i^2 \\ \sum x_i y_i \end{pmatrix}$$

og den kanoniske parameter er

$$\vartheta = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vartheta_3 \\ \vartheta_4 \\ \vartheta_5 \end{pmatrix} = \begin{pmatrix} \dfrac{\mu_x}{(1-\rho^2)\sigma_x^2} - \dfrac{\rho\mu_y}{(1-\rho^2)\sigma_x\sigma_y} \\ \dfrac{-1}{2(1-\rho^2)\sigma_x^2} \\ \dfrac{\mu_y}{(1-\rho^2)\sigma_y^2} - \dfrac{\rho\mu_x}{(1-\rho^2)\sigma_x\sigma_y} \\ \dfrac{-1}{2(1-\rho^2)\sigma_y^2} \\ \dfrac{\rho}{(1-\rho^2)\sigma_x\sigma_y} \end{pmatrix}$$

Bevis:
Følger direkte □

3.2.2 Den p-dimensionale normalfordeling

Vi betragter en p-dimensional stokastisk variabel $\mathbf{X}$ med komponenterne $X_1, X_2, \ldots, X_p$. Vi opskriver $\mathbf{X}$ som en søjlevektor:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

Definition 3.2.2 *p-dimensional normal fordeling*

Lad $\boldsymbol{\mu}$ angive en p-dimensional vektor af reelle tal, og lad $\boldsymbol{\Sigma}$ angive en $p \times p$-dimensional symmetrisk, positiv definit matrix af reelle tal,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad \text{og} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}$$

Vi siger at den p-dimensionale stokastiske vektor $\mathbf{X}$ er normalt fordelt med parametre $\boldsymbol{\mu}$ og $\boldsymbol{\Sigma}$, hvis tætheden for $\mathbf{X}$ har formen

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\boldsymbol{\Sigma})}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})) \tag{3.2.4}$$

Vi skriver $\mathbf{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Såfremt der ikke kan være tvivl om dimensionen, skriver vi ofte blot $\mathbf{X} \in \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Såfremt $\mathbf{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ gælder

$$\mathsf{E}[\mathbf{X}] = \boldsymbol{\mu} \qquad \mathbf{D}[\mathbf{X}] = \boldsymbol{\Sigma}$$

Parametrene $\boldsymbol{\mu}$ og $\boldsymbol{\Sigma}$ angiver altså middelværdivektor og dispersionsmatrix for $\mathbf{X}$.

Vi bemærker, at specielt er de en- og todimensionale normalfordelinger omfattet af ovenstående definition for henholdsvis $p = 1$ og $p = 2$.

Familien $\mathrm{N}_p(\cdot, \cdot)$ af p-dimensionale normale fordelinger er en fuld exponential familie på $\mathbb{R}^p$ frembragt af standard-normalfordelingen $\mathrm{N}_p(0, \mathbf{I})$ med tæthedsfunktion

$$f(\mathbf{x}) = (2\pi)^{-p/2} \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right)$$

og stikprøvefunktionen

$$t(\mathbf{x}) = (\mathbf{x}, -\mathbf{x}^T\mathbf{x}/2)$$

Den kanoniske parameter er

$$\boldsymbol{\vartheta} = (\boldsymbol{\zeta}\boldsymbol{\Delta}, \boldsymbol{\Delta})$$

hvor $\boldsymbol{\zeta} \in \mathbb{R}^p$ og $\boldsymbol{\Delta} \in \boldsymbol{\Gamma}_p$, hvor $\boldsymbol{\Gamma}_p$ angiver mængden af positiv definitte $p \times p$-matricer, $\boldsymbol{\Gamma}_p \subset \mathbb{R}^m$ med $m = \binom{p+1}{2}$. Da $\boldsymbol{\Gamma}_p$ er åben, er familien regulær. □

Sætning 3.2.3 *χ^2-fordeling af kvadratiske former*

Såfremt $\mathbf{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ og $Z = (\mathbf{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$, da vil $Z \in \chi^2(p)$.

Bevis:

Overspringes □

Sætning 3.2.4 *Betingede fordelinger i $N_p(\mu, \Sigma)$-fordelingen*

Lad $\mathbf{X} \in \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ og betragt spaltningen af $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \; ; \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \; ; \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

Der gælder da de analoge relationer til (3.2.3) :

$$\begin{aligned} \mathsf{E}[\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2] &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \mathsf{D}[\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2] &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{aligned} \tag{3.2.5}$$

Endvidere er $\mathsf{E}[\mathbf{X}_1|\mathbf{X}_2]$ og $\mathbf{X}_2$ stokastisk uafhængige.

Bevis:

Følger ved opskrivning af de betingede fordelinger. □

Bemærkning 1 *Fortolkning af parametrene som regressionskoefficienter*

Vi ser, at parametrene i de betingede fordelinger netop svarer til parametrene i den lineære regression af $\mathbf{X}_1$ på $\mathbf{X}_2$. □

3.3 t-fordelingen

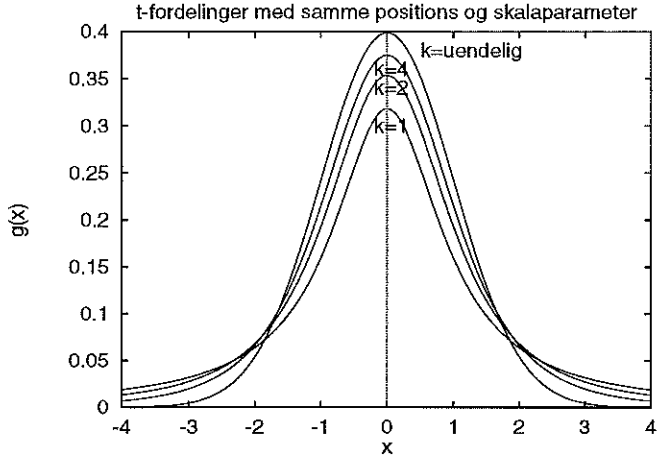
En kontinuert fordelt stokastisk variabel X, der kan antage alle reelle værdier, siges at følge en t-fordeling med parametre k, μ og β, hvis tætheden for X er af formen:

$$g(x) = \frac{\Gamma\{(k+1)/2\}}{\beta\sqrt{k\pi}\,\Gamma(k/2)} \left\{1 + \frac{(x-\mu)^2}{k\beta^2}\right\}^{-(k+1)/2}$$

hvor $k > 0$, $-\infty < \mu < \infty$ og $\beta > 0$.

Kort skriver vi $X \in \mathrm{t}(k, \mu, \beta)$.

Såfremt $X \in \mathrm{t}(k, \mu, \beta)$ og $Y = (X - \mu)/\beta$, vil $Y \in \mathrm{t}(k, 0, 1)$. Parametrene μ og β er således position og skalaparameter for fordelingen.



Figur 3.1. Tætheden for $\mathrm{T}(k,0,1)$-fordelingen svarende til forskellige værdier af k

Figur 3.1 viser tæthederne for $\mathrm{t}(k,0,1)$ fordelinger svarende til forskellige værdier af k.

$\mathrm{t}(1,\mu,\beta)$-fordelingen kaldes også en Cauchyfordeling.

Såfremt $X \in \mathrm{N}(\mu,\sigma^2)$ og $Y \in \mathrm{G}(f/2, 2/(f\beta^2))$ er uafhængige variable, og vi sætter

$$Z = \frac{X-\mu}{\sqrt{Y/f}},$$

da vil $Z \in \mathrm{t}(f,0,\beta\sigma)$,der for $\beta = 1/\sigma$ netop er den fra Statistik 1 kendte Student's t-fordeling med f frihedsgrader. $\mathrm{t}(k,\mu,\beta)$ fremkommer således blot ved en positions- og skalaændring af Student's t-fordeling. $\mathrm{t}(k,\mu,\beta)$ må ikke forveksles med den ikke-centrale t-fordeling.

Såfremt $X \in \mathrm{t}(k,\mu,\beta)$ og

$$Y = \frac{k}{k+(X-\mu)^2/\beta^2},$$

da vil $Y \in \mathrm{Be}(k/2, 1/2)$.

3.3.1 t-fordelingen som resultat af mikstur

Sætning 3.3.1 *t-fordelingen som resultat af mikstur*

Såfremt den betingede fordeling af X for givet $Y = y$ er en $\mathrm{N}(\mu, y)$-fordeling, og såfremt yderligere den marginale fordeling af variansen Y er en $\mathrm{RG}(f/2, \beta^2 f/2)$-fordeling, da er den marginale fordeling af X en $\mathrm{t}(f, \mu, \beta)$-fordeling.

Bevis:

Følger ved opskrivning af miksturen og integration over fordelingen af Y. □

Sætningen begrunder, hvorfor t-fordelingen har tykkere haler, end en normalfordeling med samme varians.

Såfremt $X \in \mathrm{t}(k, \mu, \beta)$ med $k > 2$ har man,

$$\mathrm{E}[X] = \mu, \quad \mathrm{V}[X] = \frac{k}{k-2}\,\beta^2$$

Middelværdien i fordelingen eksisterer for $1 < k$, og variansen eksisterer for $1 < k$.

3.3.2 Fordelingsfunktion

Vi betegner fordelingsfunktionen for $\mathrm{t}(k, \mu, \beta)$-fordelingen med

$$\mathrm{t}(c; k, \mu, \beta) = \int_{-\infty}^{c} \frac{k^{k/2}}{\beta\,\mathrm{B}(1/2, k/2)} \left\{ k + \frac{x-\mu)^2}{\beta^2} \right\}^{-(k+1)/2} dx$$

Da μ og σ er henholdsvis position og skalaparameter for familien gælder

$$\mathrm{t}(c; k, \mu, \beta) = \mathrm{t}\left(\frac{c-\mu}{\beta}; k, 0, 1 \right)$$

Vi kan altså bestemme den kumulerede fordeling for $\mathrm{t}(k, \mu, \beta)$ ved hjælp af tabellerne over Students t-fordeling med k frihedsgrader.

For $c < 0$ kan man bestemme den kumulerede fordeling ud fra den kumulerede Beta-fordeling ved

$$\mathrm{t}(c; k, 0, 1) = \mathrm{Be}(k/(k + c^2); k/2, 1/2)$$

For store værdier af k kan man benytte normalfordelingsapproksimationen

$$\mathrm{t}(c; k, \mu, \beta) \approx \Phi((c-\mu)/\beta)$$

3.3.3 Ufuldstændige momenter

Det første ufuldstændige moment for $\mathrm{t}(k,\mu,\beta)$-fordelingen er for $k>1$

$$\begin{aligned}\mu_1'(c) &= \int_{-\infty}^{c} x g(x)dx \\ &= \mu \mathrm{T}(t;k,0,1) - \beta\,\frac{k+t^2}{k-1}\,\frac{k^{k/2}}{\mathrm{B}(1/2,k/2)}\,(k+t^2)^{-(k+1)/2}\end{aligned}$$

med $t=(c-\mu)/\beta$.

Det tilsvarende centrale moment er

$$\mu_1(c)=\int_{-\infty}^{c}(x-\mu)g(x)dx=-\beta\,\frac{k+t^2}{k-1}\,\frac{k^{k/2}}{\mathrm{B}(1/2,k/2)}\,(k+t^2)^{-(k+1)/2}$$

3.4 Den flerdimensionale t-fordeling

En p-dimensional kontinuert fordelt stokastisk variabel, X, der kan antage alle værdier i $\mathbb{R}^p$, siges at følge en r-dimensional t-fordeling med parametre $k,\boldsymbol{\mu}$ og $\boldsymbol{\Sigma}$, hvis tætheden for X er af formen:

$$g(x)=\frac{\Gamma((k+p)//2)}{\Gamma(k/2)(k\pi)^{p/2}\sqrt{\det(\boldsymbol{V})}}\left\{1+\frac{1}{k}(x-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})\right\}^{-(k+p)/2}$$

hvor $k>0$, $\boldsymbol{\mu}$ er en vilkårlig p-dimensional vektor, og $\boldsymbol{\Sigma}$ er en symmetrisk, positiv definit $p\times p$ matrix.

Kort skriver vi $X\in \mathrm{t}_p(k,\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Såfremt $X\in \mathrm{N}_p(\mathbf{0},\boldsymbol{\Sigma})$, hvor $\boldsymbol{\Sigma}$ ikke er udartet, og såfremt yderligere $Z\in \mathrm{G}(k/2,2)$ er uafhængig af X, da gælder for

$$Y_i=\frac{X_i}{\sqrt{Z/k}}+\mu_i,\quad (i=1,2,\ldots,p)$$

at $Y=(Y_1,Y_2,\ldots,Y_p)^T\in \mathrm{t}_p(k,\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Såfremt $X\in \mathrm{t}_p(k,\boldsymbol{\mu},\boldsymbol{\Sigma})$ med $k>2$ gælder

$$\mathsf{E}[X]=\boldsymbol{\mu},\qquad \mathbf{D}[X]=\frac{k}{k-2}\,\boldsymbol{\Sigma}$$

Sætter vi

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \text{og} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

hvor X_1 og $\boldsymbol{\mu}_1$ er m-dimensionale, og $\boldsymbol{\Sigma}_{11}$ er $m \times m$ dimensional, da vil den marginale fordeling af X_1 være en $\mathrm{t}_m(k, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$-fordeling, og den betingede fordeling af X_1 givet $X_2 = x_2$ er en $\mathrm{t}_m(k + p - m, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_{00})$-fordeling med

$$\begin{aligned} \boldsymbol{\mu}_0 &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_2 - \boldsymbol{\mu}_2) \\ \text{og} & \\ \boldsymbol{\Sigma}_{00} &= \frac{k + (x_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}(x_2 - \boldsymbol{\mu}_2)}{k + p - m} (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}). \end{aligned}$$

Såfremt $X \in \mathrm{t}_p(k, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ og

$$Y = \frac{1}{k}(X - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}),$$

da vil $Y \in \mathrm{F}(p, k)$, hvor F betegner den sædvanlige F-fordeling.

Såfremt $X \in \mathrm{t}_p(k, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ og $Y = \mathbf{A}X$, hvor $\mathbf{A}$ er en $m \times k$ matrix sådan at $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ er ikke-singulær, da vil $Y \in \mathrm{t}_m(k, \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$

3.5 Wishartfordelingen

En kontinuert fordelt, symmetrisk, positiv definit stokastisk $k \times k$ matrix $\mathbf{X}$ siges at følge en Wishartfordeling med parametre n og $\boldsymbol{\Sigma}$, hvis tætheden for $\mathbf{X}$ er af formen

$$g(\mathbf{x}) = c \det(\boldsymbol{\Sigma})^{-n/2} \det(\mathbf{x})^{(n-k-1)/2} \exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{x})\}$$

for $\mathbf{x}$ symmetrisk, positiv definit $k \times k$.

I ovenstående udtryk angiver $\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{x})$ sporet af matricen $\boldsymbol{\Sigma}^{-1}\mathbf{x}$; parameteren $\boldsymbol{\Sigma}$ angiver en symmetrisk, positiv definit $k \times k$ matrix, og c er givet ved

$$\frac{1}{c} = 2^{nk/2} \pi^{k(k-1)/4} \prod_{i=1}^{k} \Gamma\left(\frac{n + 1 - i}{2}\right)$$

med $n > k - 1$.

Kort skriver vi $\mathbf{X} \in \mathrm{W}_k(n, \boldsymbol{\Sigma})$.

Den karakteristiske funktion for $\mathrm{W}_k(n, \boldsymbol{\Sigma})$-fordelingen er

$$\phi(\mathrm{t}) = \left\{ \frac{\det(\boldsymbol{\Sigma}^{-1})}{\det(\boldsymbol{\Sigma}^{-1} - \mathrm{it})} \right\}^{n/2}$$

med

$$\mathrm{t} = \begin{pmatrix} 2t_{11} & t_{12} & \dots & t_{1k} \\ t_{12} & 2t_{22} & \dots & t_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1k} & t_{1,k-1} & \dots & 2t_{kk} \end{pmatrix}$$

Vi har derfor

$$\mathsf{E}[\mathbf{X}] = n\boldsymbol{\Sigma}$$

Vi bemærker, at fordelingen blot er en fordeling af en $k(k+1)/2$-dimensional vektor, idet de øvrige elementer i matricen er bestemt af symmetrien.

Såfremt den stokastiske vektor $X \in \mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$ og vi sætter

$$\mathbf{Y} = XX^T,$$

da vil $\mathbf{Y} \in \mathrm{W}_k(1, \boldsymbol{\Sigma})$, hvorfor vi kan opfatte Wishartfordelingen som en flerdimensional generalisering af χ^2-fordelingen.

Sætning 3.5.1 ***Additionssætningen for Wishartfordelingen***

Såfremt $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_r$ er uafhængige $k \times k$-dimensionale matricer, sådan at $\mathbf{X}_i \in \mathrm{W}_k(n_i, \boldsymbol{\Sigma})$, og vi sætter

$$\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2, + \cdots + \mathbf{X}_r,$$

da vil

$$\mathbf{Y} \in \mathrm{W}_k(n_1 + \cdots + n_r, \boldsymbol{\Sigma}).$$

□

Sætning 3.5.2 *Fordeling af gennemsnit og empirisk dispersion for uafhængige flerdimensionale normalfordelte størrelser*

Såfremt $X_1, X_2, \ldots, X_n$ er uafhængige k-dimensionale vektorer, sådan at

$$X_i \in \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, 2, \ldots, n$$

og vi sætter

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

og

$$\mathbf{S} = \sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T,$$

da gælder, at genemsnittet, $\overline{X}$, og $\mathbf{S}$ er stokastisk uafhængige, og

$$\overline{X} \in \mathrm{N}_k(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

samt

$$\mathbf{S} \in \mathrm{W}_k(n-1, \boldsymbol{\Sigma})$$

dvs. specielt, at

$$\mathsf{E}\left[\frac{\mathbf{S}}{n-1}\right] = \boldsymbol{\Sigma}$$

□

Sætning 3.5.3 *Lineær transformation af Wishart-fordelte størrelser*

Såfremt $\mathbf{S} \in \mathrm{W}_k(n, \boldsymbol{\Sigma})$ og

$$\mathbf{Y} = \mathbf{A}\mathbf{S}\mathbf{A}^T,$$

hvor $\mathbf{A}$ er en given $m \times k$ matrix, da gælder

$$\mathbf{Y} \in \mathrm{W}_m(n, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

□

Sætning 3.5.4 *Uafhængighed i Wishart-familier*

Lad $\mathbf{S} \in \mathrm{W}_k(n, \mathbf{\Sigma})$ og antag, at $\mathbf{S}$ og $\mathbf{\Sigma}$ er opdelt

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \quad \text{og} \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

hvor $\mathbf{S}_{11}$ og $\mathbf{\Sigma}_{11}$ er $m \times m$ matricer. Lad tilsvarende præcisionsmatricen $\mathbf{\Delta} = \mathbf{\Sigma}^{-1}$ være opdelt

$$\mathbf{\Delta} = \begin{pmatrix} \mathbf{\Delta}_{11} & \mathbf{\Delta}_{12} \\ \mathbf{\Delta}_{21} & \mathbf{\Delta}_{22} \end{pmatrix}.$$

Da gælder, at $\mathbf{\Sigma}_{11}$ og $(\mathbf{\Delta}_{12}, \mathbf{\Delta}_{22})$ er likelihood-uafhængige. Endvidere er fordelingen af $\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ og $(\mathbf{S}_{12}, \mathbf{S}_{22})$ stokastisk uafhængige.
Bevis:
Se Barndorff Nielsen (1978), side 149.

□

Sætning 3.5.5 ***Marginal fordeling af delmatrix af Wishart-fordelte størrelser***

Såfremt $\mathbf{S} \in \mathrm{W}_k(n, \mathbf{\Sigma})$ og

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \quad \text{og} \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix},$$

hvor $\mathbf{S}_{11}$ og $\mathbf{\Sigma}_{11}$ er $m \times m$ matricer, da er den marginale fordeling af $\mathbf{S}_{11}$ en $\mathrm{W}_k(n, \mathbf{\Sigma}_{11})$-fordeling.
Bevis:
Følger af foranstående sætning

□

3.6 Hotellings T^2-fordeling

Lad $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ betegne n indbyrdes uafhængige og identisk fordelte p-dimensionale stokastiske vektorer med $\mathbf{X}_i \in \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Lad

$$\overline{\mathbf{X}} = \tfrac{1}{n} \sum_1^n \mathbf{X}_i \quad \text{og} \quad \mathbf{SAK} = \sum_1^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$$

da gælder

$$\overline{\mathbf{X}} \in \mathrm{N}_p(\boldsymbol{\mu}, \tfrac{1}{n}\,\mathbf{\Sigma}) \quad \text{og} \quad \mathbf{SAK} \in \mathrm{W}_p(n-1, \mathbf{\Sigma})$$

og endvidere er $\overline{\mathbf{X}}$ og $\mathbf{SAK}$ stokastisk uafhængige.

Indfører vi den centrale estimator, $\mathbf{S}$ for $\boldsymbol{\Sigma}$, bestemt ved $\mathbf{S} = \frac{1}{n-1}\,\mathbf{SAK}$, og sætter vi

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \tag{3.6.1}$$

da gælder

$$\frac{n-p}{(n-1)p}\,T^2 \in \mathrm{F}(p,\ n-p,\ (\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\ ) \tag{3.6.2}$$

hvor $\mathrm{F}(p, n-p, \delta)$ angiver den ikke-centrale F-fordeling med $(p, n-p)$ frihedsgrader og med skævhedsparameter δ.

Bemærkning 1 ***Beregning af T^2-størrelsen***

Til beregning af T^2-størrelsen kan man i stedet benytte

$$T^2 = \frac{\det(\mathbf{S} + n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)')}{\det(\mathbf{S})} - 1$$

□

3.7 Fordeling af empiriske varianser af normalfordelte observationer

Betragt et observationssæt bestående af n uafhængige observationer, $X_1, \ldots, X_n$ fra samme normalfordeling, dvs $X_i \in \mathrm{N}(\mu, \sigma^2)$.

Vi betragter

$$S^2 = \sum_{i=1}^{n}(X_i - \overline{X}.)^2 \Big/ (n-1)\,, \tag{3.7.1}$$

hvor $\overline{X}. = \sum X_i/n$.

Vi ved fra Introduktion til Statistik, Bind 1, at $S^2 \in \sigma^2\,\chi^2(f)/f$-fordeling med $f = n-1$. Imidlertid er $\chi^2(f)$-fordelingen blot en $\mathrm{G}(f/2, 2)$-fordeling (se afsnit 4.9), hvorfor man har

$$S^2 \in \mathrm{G}(f/2, \sigma^2/(f/2)) \tag{3.7.2}$$

Sætning 3.7.1 ***Fordeling af empiriske varianser fra normalfordelte observationer som eksponentiel dispersionsmodel***

Familien af fordelinger for S^2 er en reproduktiv eksponentiel dispersionsparameterfamilie med middelværdiparameter

$$\mathsf{E}[S^2] = \sigma^2 ,$$

enhedsvariansfunktionen $V_G(\sigma^2) = (\sigma^2)^2$ og med dispersionsparameter $\delta = 2/f$.

Den kanoniske link svarende til familien er

$$\eta = -1/\sigma^2$$

Ofte bruger man bare den reciproke funktion som linkfunktion, idet man alligevel vil lade funktionsværdierne indgå i en lineær (affin) relation.

Da dispersionsparameteren er på formen (1.3.23) vælger man at parametrisere ved en fast dispersionsparameter (nemlig 1), og bruge vægten $w = f/2$.

Bevis:
Resultatet følger af egenskaberne for gammafordelingen, afsnit 4.9. □

Det følger specielt, at

$$\mathsf{E}[S^2] = \sigma^2 , \quad \text{og} \quad \mathsf{V}[S^2] = \frac{2(\sigma^2)^2}{f}$$

med $f = n - 1$.

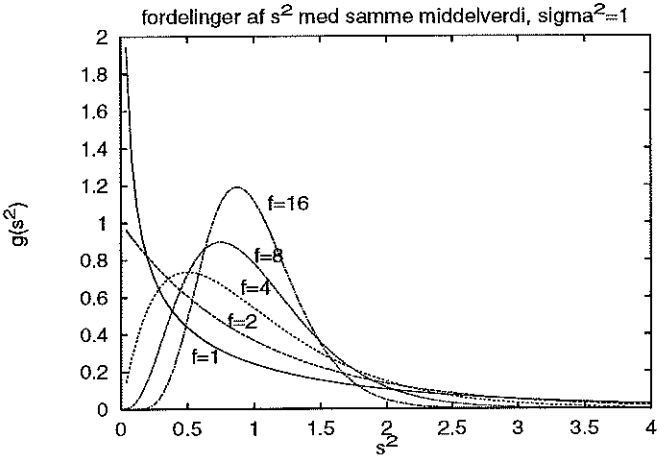
Det følger af betragtningerne i afsnit 4.9, at der er en enentydig sammenhæng mellem variationskoefficienten i fordelingen og formparameteren, $f/2$,

nemlig

$$\frac{2}{f} = \frac{\mathsf{V}[S^2]}{(\mathsf{E}[S^2])^2} = \mathsf{V}[S^2/\sigma^2] \tag{3.7.3}$$

Fordelingen af den empiriske varians for normaltfordelte observationer, S^2, er således fastlagt ved sin middelværdi (dvs variansen σ^2 i den underliggende normalfordeling af X'erne) og af variationskoefficienten, $\sqrt{2/f}$ i fordelingen af S^2.

Nedenstående figur viser tæthederne svarende til fordelingen af S^2 for $\sigma^2 = 1$ og forskellige antal frihedsgrader. Tæthederne svarende til andre værdier af σ^2 fremkommer ved en skalatransformation.



3.8 Den empiriske standardafvigelse for normalfordelte observationer

Betragt et observationssæt bestående af n uafhængige observationer, $X_1, \ldots, X_n$ fra samme normalfordeling, dvs $X_i \in \mathrm{N}(\mu, \sigma^2)$.

Traditionelt betragter man estimatoren S^2 (formel (3.7.1)) for σ^2. En ofte fremført begrundelse er, at denne estimator er central for σ^2, dvs at $\mathsf{E}[S^2] = \sigma^2$.

Denne begrundelse kan undertiden forekomme lidt søgt, da man sædvanligvis er mere interesseret i størrelsen σ, der jo måles i samme enheder, som observationerne.

Man baserer sig derfor ofte på den empiriske standardafvigelse,

$$S = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)} \; .$$

Da fordelingen af S^2 er en gammafordeling (jvf afsnit 3.7) finder man momenterne for S

$$\mathsf{E}[S^p] = \frac{\Gamma\{(n+p-1)/2\}}{\Gamma\{(n-1)/2\}} \left(\frac{2\sigma^2}{n-1}\right)^{p/2} \quad \text{for } p > 0$$

Specielt har man

$$\mathsf{E}[S] = c_4(n)\sigma; \quad \mathsf{V}[S] = \{1 - c_4^2(n)\}\sigma^2$$

hvor

$$c_4(n) = \sqrt{\frac{2}{n-1}} \times \frac{\Gamma(n/2)}{\Gamma\{(n-1)/2\}} \tag{3.8.1}$$

Man har således, at $S/c_4(n)$ er en central estimator for spredningen σ med

$$\mathsf{V}[S/c_4(n)] = \frac{1 - c_4^2(n)}{c_4^2(n)}\sigma^2$$

Størrelsen $c_4(n)$ er tabelleret i tabel 3.1 på side 122. For store værdier af n gælder $c_4(n) \approx (4n-4)/(4n-3)$.

Der gælder derfor $\mathsf{V}[S/c_4] \approx 2\sigma^4/(4n-3)$

Programsystemet SAS/QC indeholder beregningen af c_4 som en indbygget funktion benævnt `C4(n)`.

3.9 Variationsbredden

3.9.1 Vilkårlig kontinuert fordeling

Vi betragter et sæt $X_1, X_2, \ldots, X_n$ af uafhængige identisk fordelte stokastiske variable med fordelingsfunktion $F(\cdot)$ og med den kontinuerte tæthed $f(\cdot)$. De tilsvarende ordnede observationer betegnes med $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. følgende

Definition 3.9.1 ***Variationsbredden***

Variationsbredden svarende til observationssættet $X_1, X_2, \ldots, X_n$ er den stokastiske variable $R = X_{(n)} - X_{(1)}$.

Den engelske term for variationsbredde er *range*. □

Sætning 3.9.1 ***Fordeling af variationsbredde*** Lad $X_1, X_2, \ldots, X_n$ være uafhængige identisk fordelte med fordelingsfunktion $F(\cdot)$ og med den kontinuerte tæthed $f(\cdot)$.

Fordelingsfunktionen for variationsbredden, R, er bestemt ved

$$\mathsf{P}[R \leq r] = \int_{-\infty}^{\infty} n\big\{F(r+x) - F(x)\big\}^{n-1} f(x)dx \quad \text{for } r > 0$$

og 0 ellers.

Fordelingens forventningsværdi og varians fås af

$$\begin{aligned} \mathsf{E}[R] &= \int_{-\infty}^{\infty} \big[1 - \{1 - F(x)\}^n\big]dx \\ \mathsf{E}[R^2] &= \int_{-\infty}^{\infty}\int_{-\infty}^{y} \big[1 - F(y)^n - \{1 - F(x)\}^n + \{F(y) - F(x)\}^n\big]dxdy \end{aligned} \tag{3.9.1}$$

Bevis:
Beviset følger ved at bemærke, at den simultane fordeling af $Y = X_{(i)}$ og $Z = X_{(j)}$ for $i < j$ har tætheden

$$g(y,z) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \times \{F(y)\}^{i-1}\{F(z)-F(y)\}^{j-i-1}\{1-F(z)\}^{n-j} f(y)f(z)$$

for $y < z$
og $g(y,z) = 0$ ellers.

Man har således specielt, at den simultane fordeling af $X_{(1)}$ og $X_{(n)}$ har tætheden

$$f(x_{(1)}, x_{(n)}) = n(n-1)\{F(x_{(n)})-F(x_{(1)})\}^{n-2} f(x_{(n)})f(x_{(1)}) \text{ for } x_{(1)} < x_{(n)}$$

se f.eks. Dudewicz og Mishra (1988) og Tippet (1925). □

3.9.2 Variationsbredden for normalfordelte observation

Eksempel 3.9.1 ***Variationsbredden for observationer fra den standardiserede normalfordeling***

Lad $U_1, U_2, \ldots, U_n$ være uafhængige N(0,1)-fordelte variable. Vi benytter her symbolet W for variationsbredden.

$W = U_{(n)} - U_{(1)}$ har da fordelingsfunktionen

$$\mathrm{P}[W \le w] = \int_{-\infty}^{\infty} n\{\Phi(w+x) - \Phi(x)\}^{n-1}\phi(x)dx \text{ for } w > 0 \tag{3.9.2}$$

og 0 ellers

□

Definition 3.9.2 ***Den standardiserede variationsbreddefordeling*** $W(n)$

En stokastisk variabel W med fordelingsfunktionen (3.9.2) siges at følge den standardiserede variationsbreddefordeling svarende til n observationer, og vi skriver $W \in \mathrm{W}(n)$ □

Sætning 3.9.2 ***Fordeling af variationsbredden for normalfordelte observationer***

Lad $X_1, X_2, \ldots, X_n$ være uafhængige $\mathrm{N}(\mu, \sigma^2)$-fordelte variable og lad R betegne variationsbredden, $R = X_{(n)} - X_{(1)}$.

Da vil $R/\sigma \in \mathrm{W}(n)$, hvorfor specielt

$$\mathsf{E}[R/\sigma] = d_2(n) \qquad \text{og} \qquad \mathsf{V}[R/\sigma] = (d_3(n))^2 \,,$$

hvor

$$d_2(n) \overset{\text{DEF}}{=} \mathsf{E}[W(n)] = \int_{-\infty}^{\infty} \left[1 - \{\Phi(-u)\}^n - \{\Phi(-u)\}^n\right] du \tag{3.9.3}$$

(3.9.4)

$$\begin{aligned} d_3^2(n) &\overset{\text{DEF}}{=} \mathsf{V}[W(n)] \\ &= 2\int_{u=-\infty}^{\infty} \int_{v=-\infty}^{u} \\ &\quad \left[1 - \{\Phi(-u)\}^n - \{\Phi(-v)\}^n + \{\Phi(u)n - \{\Phi(v)\}^n\right] dvdu - d_2^2 \end{aligned}$$

Bevis:
Beviset følger ved indsættelse i (3.9.1). □

Parametrene d_2 og d_3 er tabelleret i tabel 3.1 på side 122.

Programsystemet SAS/QC indeholder parametrene d_2 og d_3 som indbyggede funktioner (benævnt hhv. `D2(n)` og `D3(n)`).

Bemærkning 1 ***En central estimator for*** σ

Såfremt $X_1, X_2, \ldots, X_n$ er uafhængige $\mathrm{N}(\mu, \sigma^2)$-fordelte variable, er størrelsen $\tilde{\sigma} = R/d_2(n)$ et centralt estimat for σ med

$$\mathsf{E}[R/d_2(n)] = \sigma; \qquad \mathsf{V}[R/d_2(n)] = \sigma^2(d_3(n)/d_2(n))^2$$

□

Estimatoren $\tilde{\sigma} = R/d_2(n)$ er i almindelighed ikke efficient.

Tabel 3.2 på side 123 viser efficiensen af $\tilde{\sigma} = R/d_2(n)$ i forhold til estimatoren $S/c_4(n)$.

Sætning 3.9.3 ***Asymptotisk fordeling af den standardiserede variationsbredde svarende til normaltfordelte observationer***

Lad $W(n)$ følge den standardiserede variationsbreddefordeling givet ved (3.9.2). Da gælder for $n \to \infty$

$$\mathsf{P}\left[\frac{W(n) - 2a_n}{b_n} \leq w\right] \to \int_{-\infty}^{\infty} \exp(-e^{x-w}) d[\exp(-e^{-x})] \tag{3.9.5}$$

med

$$\begin{aligned} a_n &= \sqrt{2\ln n} - \frac{1}{2}\,\frac{\ln\ln n + 4\ln(4\pi)}{\sqrt{2\ln n}} \\ b_n &= \frac{1}{\sqrt{2\ln n}} \end{aligned} \tag{3.9.6}$$

Grænsefordelingen på højre side må bestemmes ved numerisk integration.

Bevis:

Med betegnelserne fra eksempel 3.9.1 har vi, at

$$\begin{aligned} Z_n &= \frac{U(n) - a_n}{b_n} \overset{\text{as}}{\in} \mathrm{Max}_1(0,1) \\ Y_n &= \frac{U(1) + a_n}{b_n} \overset{\text{as}}{\in} \mathrm{Min}_1(0,1) \end{aligned}$$

og at Z_n og Y_n er asymptotisk uafhængige (se f.eks. Galambos (1978) p.109).

Det følger da af egenskaberne for Max_1 og Min_1-fordelingerne (afsnit 4.13), at

$$\mathsf{E}[Z_n] \rightarrow \gamma\,; \qquad \mathsf{V}[Z_n] \rightarrow \frac{\pi^2}{6}$$

$$\mathsf{E}[Y_n] \rightarrow -\gamma\,; \qquad \mathsf{V}[Y_n] \rightarrow \frac{\pi^2}{6}\,,$$

hvor γ angiver Eulers konstant, $\gamma \approx 0.57722$. jvf Introduktion til Statistik, Bind 1.

Vi har derfor

$$\begin{aligned}\mathsf{E}[W((n)] &\approx 2(a_n + b_n) = 2\sqrt{2\ln n} - \frac{\ln\ln n + 4\ln(4\pi) - 2\gamma}{\sqrt{2\ln n}} \\ \mathsf{V}[W(n)] &\approx 2b_n^2 . \frac{\pi^2}{6} = \frac{\pi^2}{6\ln n}\,,\end{aligned}$$

Vedrørende beviset for grænsefordelingen, se Galambos (1978).

□

Bemærkning 1 ***Approksimativt udtryk for variansen for*** $\widetilde{\sigma}$

Det følger af sætningen, at

$$\begin{aligned}d_2(n) &\approx 2(a_n + b_n) = 2\sqrt{2\ln n} - \frac{\ln\ln n + 4\ln(4\pi) - 2\gamma}{\sqrt{2\ln n}} \\ d_3^2(n) &\approx 2b_n^2\,\frac{\pi^2}{6} = \frac{\pi^2}{6\ln n}\end{aligned}$$

Vi har derfor det approksimative udtryk for variansen på den centrale estimator $\widetilde{\sigma} = R/d_2(n)$

$$\begin{aligned}\mathsf{V}[R/d_2(n)] &= \sigma^2\,\frac{2\pi^2}{6}\,\frac{b_n^2}{d_2^2} = \frac{\pi^2}{12}\,\frac{1}{(a_n/b_n + \gamma)^2} \\ &\approx \frac{\pi^2}{12}\,\frac{1}{4(\ln n)^2 + 2(\ln n)(\ln\ln n)\;2\ln(n)(\ln(4\pi) - 2\gamma)} \\ &\approx \frac{\pi^2}{48(\ln n)^2}\end{aligned}$$

Efficiensen af $\widetilde{\sigma} = R/d_2(n)$ i forhold til $S/c_4(n)$ bliver da

$$\frac{\mathrm{V}[S/c_2(n)]}{\mathrm{V}[R/d_2(n)]} = \frac{1-c_4^2(n)}{c_4^2(n)} \Big/ \frac{\pi^2(a_n/b_n+\gamma)^2}{12}$$
$$\approx \frac{2}{4n-3}\,\frac{48(\ln n)^2}{\pi^2} = \frac{24}{\pi^2}\,\frac{(\ln n)^2}{n-3/4}$$

Udtrykket angiver størrelsesordenen, men er ikke egnet som en numerisk approksimation. □

Bemærkning 2 ***Grænsefordelingen for midrange*** Som et kuriosum kan vi nævne, at grænsefordelingen for den såkaldte midrange, defineret ved

$$M_n = \frac{U_{(1)}+U_{(n)}}{2}$$

er den logistiske fordeling, $\mathrm{L}(0,1)$, se afsnit 4.5.

$$\mathrm{P}[2M_n/b_n \le m] \to \int_{-\infty}^{\infty} \exp(-e^{m-x})d[\exp(-e^{-x})] = \frac{1}{1+e^{-m}}$$

Bevis:
Se Galambos (1978) □

3.10 Marginal fordeling af empiriske varianser ved mikstur

Betragt en situation, hvor der foreligger k stikprøver fra $\mathrm{N}(\mu_i,\sigma^2)$-fordelinger, $i=1,2,\ldots,k$.

For hver stikprøve bestemmes den empiriske varians, S_i^2 i overensstemmelse med (3.7.1).

Skønnene, S_i^2 afhænger ikke af eventuelle forskelle mellem middelværdierne, μ_i.

Såfremt antagelsen om varianshomogenitet holder, vil fordelingen af de empiriske varianser være en $\chi^2(f)/f$-fordeling. Imidlertid oplever man undertiden, at den observerede fordeling har tykkere haler, end $\chi^2(f)/f$-fordelingen, dvs man kan ikke antage at der er varianshomogenitet.

Undertiden kan en sådan variansinhomogenitet forklares ved en tilfældig model for varianserne.

Der gælder

Sætning 3.10.1 Den marginale fordeling af s^2 ved reciprok gamma strukturfordeling af σ^2

Såfremt

$$s^2|\sigma^2 \in \sigma^2\chi^2(f)/f \qquad \text{og } 1/\sigma^2 \in \frac{1}{2\beta}\,\chi^2(\nu)\,,$$

da er den marginale fordeling af s^2 en RBet$(\nu/2,\ f/2,\ 2\beta/f)$-fordeling.

Såfremt $\nu \leq 2$ har fordelingen af s^2 ingen middelværdi. For $2 < \nu$ har fordelingen af s^2 middelværdien

$$\mathsf{E}[s^2] = \mathsf{E}[\sigma^2] = \frac{\beta}{\nu/2-1} \tag{3.10.1}$$

For $\nu \leq 4$ har fordelingen ingen varians. Såfremt $4 < \nu$, har fordelingen af s^2 variansen

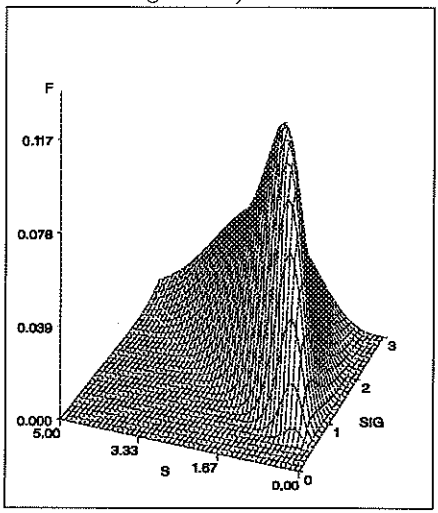
$$\mathsf{V}[s^2] = \left(\frac{\beta}{\nu/2-1}\right)^2 \frac{1}{\nu/2-2}\left[1+\frac{2(\nu/2-1)}{f}\right] \tag{3.10.2}$$

Bevis:
Se Introduktion til Statistik, Bind 3. □

Den simultane fordeling af spredningen σ og af stikprøvespredningen s_i er illustreret i figur 3.2

Figur 3.3 viser den betingede fordeling af stikprøvespredningen svarende til en givet værdi af σ, og figur 3.4 viser den marginale fordeling af stikprøvespredningen s svarende til fordelingen i figur 3.2. Det ses, at den marginale fordeling af s har tykkere haler, end den betingede fordeling i figur 3.3.

Figur 3.2. Simultan fordeling af stikprøvespredning, s, bestemt i stikprøve på $n = 5$ og sand spredning σ
(Strukturfordeling af σ som i figur 3.5.)



Bemærkning 1 ***Den marginale fordeling af s^2 udtrykt ved F-fordelingen***

Ved at udnytte relationen mellem RBet-fordelingen og F-fordelingen (se bemærkningen på side 180) finder man, at der gælder

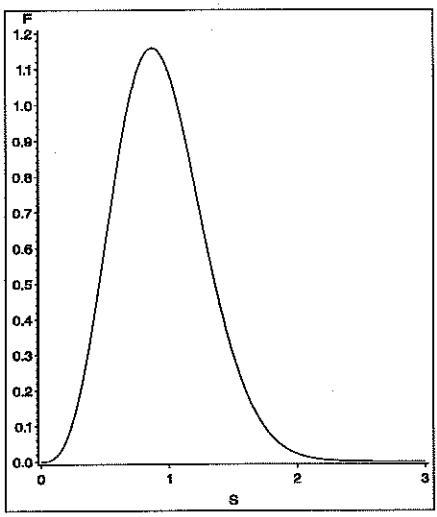
$$\frac{1}{S^2} \in \frac{\nu}{2\beta} \mathrm{F}(\nu, f) \tag{3.10.3}$$

der ved indsættelse af udtrykket (3.10.8) for β kan udtrykkes som

$$\frac{\mathrm{E}[\sigma^2]}{S^2} \in \frac{\nu}{\nu - 2} \mathrm{F}(\nu, f) \tag{3.10.4}$$

□

Figur 3.3. Betinget fordeling af af stikprøvespredning, s, bestemt i stikprøve på $n = 5$ for en sand spredning, $\sigma = 1$.



3.10.1 Strukturfordelingen af σ

Strukturfordelingen af σ^2 er en $2\beta/\chi^2(\nu)$-fordeling.

Imidlertid vil det ofte være lettere at forholde sig til fordelingen af σ, som altså er en $\sqrt{2\beta/\chi^2(\nu)}$-fordeling.

Figur 3.5 viser et eksempel på fordelingen af σ.

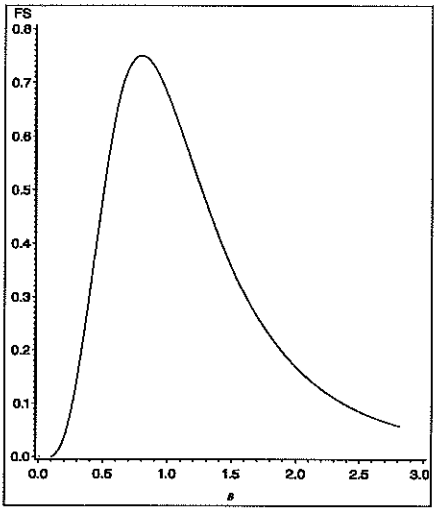
De første momenter i fordelingen af σ er

$$\mathsf{E}[\sigma_i] = \frac{\Gamma\big((\nu-1)/2\big)}{\Gamma(\nu/2)}\sqrt{\beta} \tag{3.10.5}$$

samt

$$\mathsf{V}[\sigma_i] = \beta\Big[\frac{\Gamma(\nu/2)\Gamma(\nu/2-1) - \big\{\Gamma\big((\nu-1)/2\big)\big\}^2}{\big\{\Gamma(\nu/2)\big\}^2}\Big] \tag{3.10.6}$$

Figur 3.4. Marginal fordeling af stikprøvespredning, s, bestemt i stikprøve på $n = 5$
(Strukturfordeling af σ som i figur 3.5.)



Variansen i fordelingen af σ eksisterer således kun for $\nu > 2$.

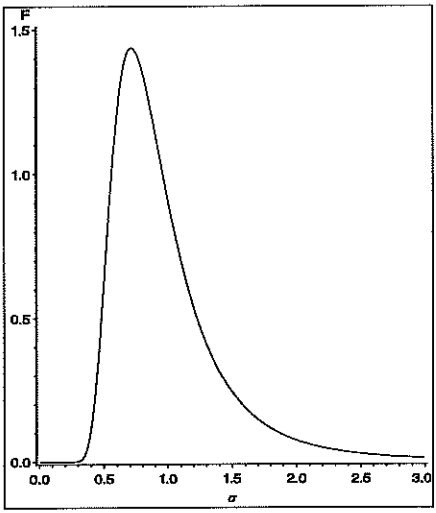
Variationskoefficienten i fordelingen af σ er

$$\frac{\sqrt{\mathsf{V}[\sigma_i]}}{\mathsf{E}[\sigma]} = \sqrt{\frac{\Gamma(\nu/2)\Gamma(\nu/2-1)}{\left\{\Gamma\big((\nu-1)/2\big)\right\}^2} - 1} \tag{3.10.7}$$

Variationskoefficienten (dvs den relative spredning af σ) afhænger ikke af β.

Figur 3.6 viser sammenhængen mellem parameteren ν og den relative spredning i fordelingen af σ. Det ses, at der er en monoton sammenhæng mellem den relative spredning og antallet af frihedsgrader, ν. Jo større antal frihedsgrader, desto mindre er den relative spredning.

Figur 3.5. Strukturfordeling af sand spredning, σ for $\nu = 4$, $\beta = 1$



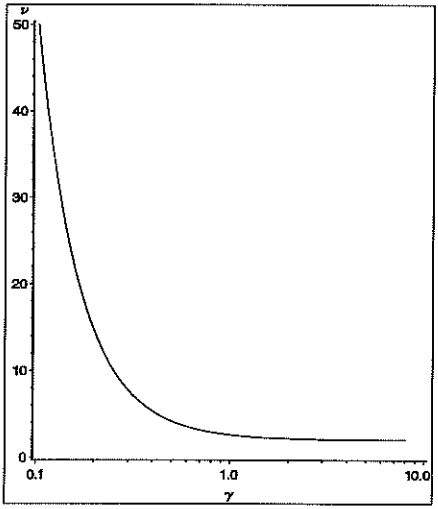
Hvis man kender den relative spredning i fordelingen af σ kan man således bestemme antallet af frihedsgrader, ν, og parameteren β kan da bestemmes ud fra middelværdien af σ eller af σ^2 fra (3.10.5) eller (3.10.6) som

$$\beta = \left[\frac{\Gamma(\nu/2)}{\Gamma\big((\nu-1)/2\big)}\right]^2 \Big(\mathrm{E}[\sigma]\Big)^2 \tag{3.10.8}$$

og

$$\beta = (\nu/2 - 1)\ \mathrm{E}[\sigma^2] \tag{3.10.9}$$

Figur 3.6. Sammenhæng mellem "frihedsgrader" og relativ spredning i strukturfordeling af σ



Tabel 3.1. Størrelserne c_4, d_2 og d_3 til brug ved estimation af spredningen i normalfordelingen

Stik-prøve-stør-relse					
n	c_4	$1/c_4$	d_2	$1/d_2$	d_3
2	0,7979	1,2533	1,128	0,8865	0,8525
3	0,8862	1,1284	1,693	0,5907	0,8884
4	0,9213	1,0854	2,059	0,4857	0,8798
5	0,9400	1,0638	2,326	0,4299	0,8641
6	0,9515	1,0510	2,534	0,3946	0,8480
7	0,9594	1,0423	2,704	0,3698	0,8332
8	0,9650	1,0363	2,847	0,3512	0,8198
9	0,9693	1,0317	2,970	0,3367	0,8078
10	0,9727	1,0281	3.078	0,3249	0,7971
11	0,9754	1,0252	3,173	0,3152	0,7873
12	0,9776	1,0229	3,258	0,3069	0,7785
13	0,9794	1,0210	3,336	0,2998	0,7704
14	0,9810	1,0194	3,407	0,2935	0,7630
15	0,9823	1,0180	3,472	0,2880	0,7562
16	0,9835	1,0168	3,532	0,2831	0,7499
17	0,9845	1,0157	3,588	0,2787	0,7441
18	0,9854	1,0148	3,640	0,2747	0,7386
19	0,9862	1,0140	3,689	0,2711	0,7335
20	0,9869	1,0133	3,735	0,2677	0,7287
21	0,9876	1,0126	3,778	0,2647	0,7242
22	0,9882	1,0119	3,819	0,2618	0,7199
23	0,9887	1,0114	3,858	0,2592	0,7159
24	0,9892	1,0109	3,895	0,2567	0,7121
25	0,9896	1,0105	3,931	0,2544	0,7084

Tabel 3.2. Forholdet $V[S/c_2(n)]/V[R/d_2(n)]$ for forskellige værdier af stikprøvestørrelsen, n.

n							
2	3	4	5	6	7	8	9
1,000	0,993	0,976	0,955	0,933	0,911	0,890	0,870

n							
10	11	12	13	14	15	16	17
0,849	0,830	0,811	0,797	0,779	0,767	0,750	0,738

n							
18	19	20	21	22	23	24	25
0,727	0,712	0,703	0,688	0,675	0,667	0,658	0,650

3.11 Referencer

Barndorff-Nielsen, O.E. (1978): *Information and Exponential Families.* Wiley, Chichester.

Dudewicz, E.J. and Mishra, S.N. (1988): *Modern Mathematical Statistics,* John Wiley & Sons, Inc., New York

Galambos, J. (1978): *The Asymptotic Theory of Extreme Order Statistics,* John Wiley & Sons, Inc., New York

Tippet, L.H.C. (1925): On the Extreme Individuals and the Range of Samples taken from a Normal Population, *Biometrika* **17** pp 364-387.

Afsnit 4

Fordelinger og miksturer af fordelinger

Fil: /tex/stat3/fordbog/fordb.tex 1998-01-13

4.1 Indledning

I dette afsnit gives en oversigt over fordelinger, der optræder i statistiske anvendelser. Afsnittet er imidlertid langt fra at være færdigt. Man henvises til encyklopædien udarbejdet af Kotz, Johnson og Read (1988) samt de omfattende oversigter udarbejdet af Johnson, Kotz og Kemp (1993) og Johnson, Kotz og Balakrishnan (1995).

4.2 Den inverse Gaussfordeling

Definition 4.2.1 ***Invers Gaussfordeling*** Vi vil kort introducere en familie af fordelinger, som ikke blev behandlet i Statistik I, nemlig familien af inverse Gaussfordelinger, en familie af fordelinger med støtte på de ikke-negative reelle tal.

En kontinuert stokastisk variabel X, der kan antage alle reelle ikke-negative værdier, siges at følge den inverse Gaussfordeling med parametre μ og λ, hvis tætheden for X (med hensyn til Lebesguemålet) er af formen

$$g(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right\} \times \frac{1}{x^{3/2}} \quad \text{for } x \in \mathbb{R}_+ \tag{4.2.1}$$

hvor $\mu \in \mathbb{R}_+$ og $\lambda \in \mathbb{R}_+$.
Kort skriver vi $X \in \mathrm{IG}(\mu,\lambda)$.

□

Det gælder, at parameteren λ kan udtrykkes som

$$\lambda = \mathsf{E}\left[\frac{1}{X}\right] - \frac{1}{\mathsf{E}[X]}$$

Tætheden er unimodal med modus i

$$x_m = -\frac{3\mu^2}{2\lambda} + \mu\left(1 + \frac{9\mu^2}{4\lambda^2}\right)^{1/2}$$

4.2.1 Estimation af parametrene μ og λ

Såfremt $X_1, X_2, \ldots, X_n$ er uafhængige og $X_i \in \mathrm{IG}(\mu,\lambda)$ er maksimum likehood estimaterne $\widehat{\mu}$ og $\widehat{\lambda}$ for μ og λ bestemt ved

$$\begin{aligned} \widehat{\mu} &= \overline{X}_. \\ \frac{1}{\widehat{\lambda}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{X_i} - \frac{1}{\overline{X}_.}\right) \end{aligned} \tag{4.2.2}$$

med $\overline{X}_. = \sum X_i/n$.

Estimatorerne $\widehat{\mu}$ og $\widehat{\lambda}$ er stokastisk uafhængige, og der gælder

$$\widehat{\mu} \in \mathrm{IG}(\mu, n\lambda) \quad \text{og } \frac{n\lambda}{\widehat{\lambda}} \in \chi^2(n-1) \tag{4.2.3}$$

se Tweedie (1957) og Barndorff-Nielsen (1978).

Vi bemærker analogien med fordelingen af $\overline{X}_\cdot$ og S^2 for normalfordelte observationer.

4.2.2 Den stabile fordeling

Sætter man i $\vartheta = 0$ (svarende til $\mu = \infty$) i udtrykket (4.2.9) for tætheden får man en familie af fordelinger med tæthed

$$g(x; 0, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{-\frac{\lambda}{2x}\right\} \quad \text{for } x \in \mathbb{R}_+ \tag{4.2.4}$$

Denne familie kaldes familien af positive stabile fordelinger (engelsk: *stable distribution*) med indeks (karakteristisk eksponent) $1/2$ og skalaparameter $1/\lambda$. Denne familie kan altså også betegnes $\mathrm{IG}(\infty, \lambda)$. Fordelingens middelværdi og varians er uendelig store.

4.2.3 Alternativ parametrisering af den inverse Gaussfordeling:

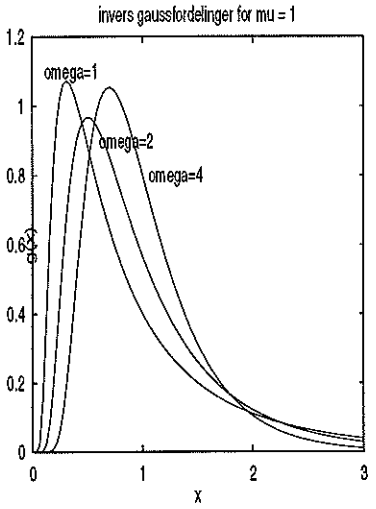
Sætter vi i (4.2.1) $\omega = \lambda/\mu$ finder vi udtrykket for tætheden

$$f(x; \mu, \omega) = \left(\frac{\mu\omega}{2\pi x^3}\right)^{1/2} \exp\left\{-\frac{\omega}{2}\left(\frac{x}{\mu} + \frac{\mu}{x}\right) + \omega\right\} \tag{4.2.5}$$

der viser, at parameteren $\omega = \lambda/\mu$ er en formparameter for fordelingen, og μ er en skalaparameter. Med denne parametrisering er modus

$$x_m = -\mu\left[\frac{3}{2\omega} + \sqrt{1 + \left(\frac{3}{2\omega}\right)^2}\right]$$

Da parametriseringen ved μ og λ er udbredt i litteraturen, har vi dog valgt at bibeholde denne parametrisering fremfor den mere bekvemme parametrisering ved form- og skalaparameter.



Sætning 4.2.1 ***Momenter for IG*** (μ, λ) ***fordeling*** Lad $X \in \mathrm{IG}(\mu, \lambda)$. Da er den karakteristiske funktion

$$\phi(t) = \exp\left[\frac{\lambda}{\mu}\left\{1 - \left(1 + \frac{2\mu^2}{\lambda}\,\mathrm{i}\,t\right)^{1/2}\right\}\right]$$

Middelværdi og varians er

$$\mathsf{E}[X] = \mu; \qquad \mathsf{V}[X] = \mu^3/\lambda \tag{4.2.6}$$

Bevis:
Beviset følger direkte

□

Da det er ofte af interesse at betragte den reciproke værdi af en IG-fordelt variabel, anfører vi de første momenter for den reciproke variabel:

Sætning 4.2.2 ***Reciproke momenter for IG(μ, λ) fordeling*** Lad $X \in \mathrm{IG}(\mu, \lambda)$. Da gælder

$$\mathsf{E}[1/X] = \frac{1}{\mu} + \frac{1}{\lambda}; \qquad \mathsf{V}[1/X] = \frac{\lambda + 2\mu}{\mu\lambda^2} \tag{4.2.7}$$

Bevis:
Beviset følger direkte

□

Udtrykt ved formparameteren $\omega = \lambda/\mu$ og skalaparameteren μ finder vi udtrykkene for momenterne

$$\mathsf{E}[X] = \mu \qquad \mathsf{E}[1/X] = \frac{\omega + 1}{\mu\omega}$$
$$\mathsf{V}[X] = \mu^2/\omega \qquad \mathsf{V}[1/X] = \frac{\omega + 2}{\mu^2\omega^2}$$

4.2.4 Fordelingsfunktion

Den kumulerede fordelingsfunktion for IG-fordelingen kan bestemmes ved hjælp af den kumulerede standardiserede normalfordeling. For $X \in \mathrm{IG}(\mu, \lambda)$ gælder

$$\mathsf{P}[X \le x] = \Phi\{\sqrt{\lambda/x}\,(x/\mu - 1)\} + \exp(2\lambda/\mu)\,\Phi\{-\sqrt{\lambda/x}\,(x/\mu + 1)\}$$

se Chhikara and Folks (1977).

4.2.5 Den inverse Gaussfordeling som eksponentiel familie

Ved omskrivning af tætheden (4.2.1) til

$$f(x;\alpha,\lambda) = \left(\frac{\lambda}{2\pi}\right)^{1/2} \times \frac{1}{x^{3/2}} \times \exp\left\{\sqrt{\alpha\,\lambda} - \alpha x/2 - \lambda/(2x)\right\} \tag{4.2.8}$$

med $\alpha = \lambda/\mu^2$ ser man, at familien af $\mathrm{IG}(\mu, \lambda)$-fordelinger med $(\mu, \lambda) \in \mathbb{R}_+ \times \mathbb{R}_+$ udgør en eksponentiel familie af orden 2 med kanonisk parameter

$$\boldsymbol{\vartheta} = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \end{pmatrix} = \begin{pmatrix} -\alpha/2 \\ -\lambda/2 \end{pmatrix},$$

kanonisk stikprøvefunktion

$$\mathbf{t} = \begin{pmatrix} t_1(x) \\ t_2(x) \end{pmatrix} = \begin{pmatrix} x \\ 1/x \end{pmatrix}$$

og kumulantfrembringer

$$\kappa(\alpha, \lambda) = \frac{1}{2} \ln(\lambda) + \sqrt{\alpha\lambda}\,.$$

Familien er ikke fuld. Den fulde familie fremkommer ved at tillade værdien $\alpha = 0$ svarende til den stabile fordeling med indeks $1/2$ og skalaparameter $1/\lambda$.

Familien er stejl, men den fulde familie er ikke regulær.

4.2.6 Den inverse Gaussfordeling som eksponentiel dispersionsmodel

Familien af $\mathrm{IG}(\mu, \lambda)$-fordelinger med $(\mu, \lambda) \in \mathbb{R}_+ \times \mathbb{R}_+$ udgør en reproduktiv eksponentiel dispersionsmodel med middelværdi μ, enhedsvariansfunktion $V_{IG}(\mu) = \mu^3$ og med dispersionsparameter $\sigma^2 = 1/\lambda$. Den kanoniske parameter er $\vartheta = -1/(2\mu^2)$ med det kanoniske parameterområde $D =]-\infty, 0]$. Middelværdirummet er $\mathcal{M} = \mathbb{R}_+$ og modellen er stejl. Udtrykt ved den kanoniske parameter $\vartheta = -1/(2\mu^2)$ og λ er tætheden

$$g(y; \vartheta, \lambda) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left\{-\frac{\lambda}{2y} + \lambda\left(\vartheta y + \sqrt{-2\vartheta}\right)\right\} \quad \text{for } y \in \mathbb{R}_+ \tag{4.2.9}$$

De vigtigste størrelser i fortolkningen af familien af $\mathrm{IG}(\mu, \lambda)$-fordelinger som en reproduktiv eksponentiel dispersionsmodel er anført i nedenstående oversigt:

$IG(\mu,\lambda)$-fordelingen som reproduktiv eksponentiel dispersionsmodel				
Kanonisk parameter ϑ	Kumulant-frembringer $\kappa(\vartheta)$	Middelværdi-afb. $\mu=\tau(\vartheta)$	Enheds-variansfunktion $V_{IG}(\mu)$	dispersions-parameter σ^2
$-1/(2\mu^2)$	$-\sqrt{-2\vartheta}$	$1/\sqrt{-2\vartheta}$	μ^3	$1/\lambda$

Den kanoniske link er kvadratet på den reciproke afbildning:

$$\vartheta = \tau^{-1}(\mu) = -\frac{1}{2\mu^2}$$

Enhedsdeviansen for IG(μ,λ)-fordelingen er

$$d(y;\mu) = \frac{(y-\mu)^2}{y\mu^2} \tag{4.2.10}$$

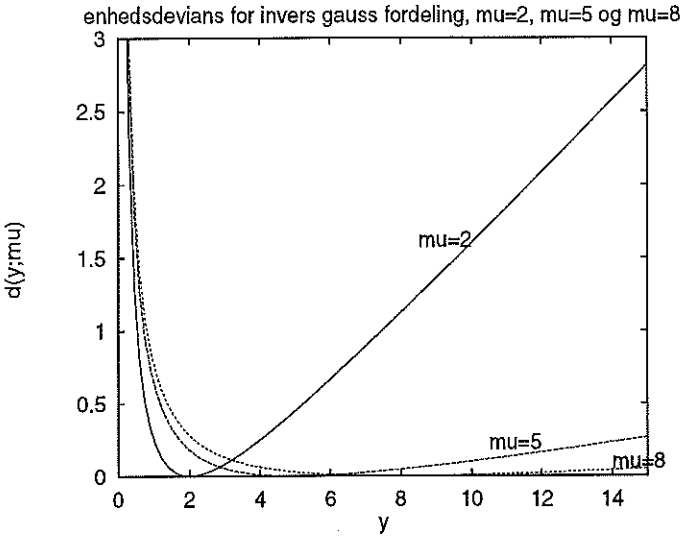
Vi kan udtrykke tætheden for en IG(μ,λ)-fordeling ved enhedsdeviansen på formen (1.3.20) som

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\frac{(y-\mu)^2}{2y\mu^2\sigma^2}\right\}, \tag{4.2.11}$$

hvor

$$a(y;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}y\sqrt{y}}$$

Nedenstående figur viser enhedsdeviansen svarende til $\mu = 2$, $\mu = 5$ og $\mu = 8$.



Bemærkning 1 ***Skalatransformation af den inverse Gaussfordeling***

Såfremt Y følger en $\mathrm{IG}(\mu, 1/\sigma^2)$-fordeling, da vil fordelingen af cY for $c > 0$ ligeledes følge en invers Gauss fordeling, men med middelværdien $c\mu$ og dispersionsparameteren σ^2/c. □

4.2.7 Genesis

Betragt en partikel, der følger en Brownsk bevægelse med positiv drift ν [længde/tidsenhed] og varians σ^2 (dvs. diffusionskonstant σ). Afstanden X, som partiklen tilbagelægger i et tidsrum af længden t, er en da en $\mathrm{N}(\nu t, \sigma^2\, t)$-fordelt stokastisk variabel med tæthed

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left[-\frac{(x-\nu t)^2}{2\sigma^2 t}\right] \tag{4.2.12}$$

Ventetiden T, indtil partiklen har flyttet sig afstanden x fra sin udgangsposition, har tætheden

$$g_T(t) = \left(\frac{x^2}{2\pi\sigma^2 t^3}\right)^{1/2} \exp\left\{-\frac{(x-\nu t)^2}{2\sigma^2 t}\right\} \tag{4.2.13}$$

Der gælder således, at $T \in \mathrm{IG}(x/\nu, (x/\sigma)^2)$.

Den inverse relation mellem den karakteristiske funktion for X og den karakteristiske funktion for T har givet anledning til betegnelsen invers Gaussfordeling (Tweedie (1957))

Denne fortolkning som ventetidsfordelingen til overskridelse af en tærskelværdi begrunder anvendelsen af den inverse Gauss fordeling til beskrivelse af levetider.

Hastigheden, hvormed partiklen når frem til tærskelværdien x, er $1/T$. Fordelingen af $Y = 1/T$ benævnes somme tider random walk fordelingen. Momenterne af $Y = 1/T$ er givet ved (4.2.7).

4.2.8 IG-fordelingen som grænsefordeling

Ventetidsfordelingen, hvis tæthed er udtrykt ved (4.2.13) optræder bl.a. som grænsefordeling for stikprøvestørrelsen i sekvenstestet.

Lad $Z_1, Z_2, \ldots$ være en følge af uafhængige identisk fordelte variable med forventningsværdi $\mathsf{E}[Z] > 0$ og med en ikke-udartet fordeling, dvs. med $\mathsf{V}[Z] > 0$. Betragt de successive summer $S_1 = Z_1, \quad S_2 = Z_1 + Z_2, \quad , \ldots, S_i = Z_1 + Z_2 + \cdots + Z_i$ og lad N betegne det første index i, for hvilket $S_i \geq A$, hvor $A > 0$. (Dvs. $S_1 < A, S_2 < A, \ldots S_{N-1} < A$ og $S_N \geq A$.

Det gælder da, at $\mathsf{E}[N] \times \mathsf{E}[Z] = A$, dvs.

$$\mathsf{E}[N] = A/\mathsf{E}[Z]$$

og

$$\lim_{\mathsf{E}[N]\to\infty} \mathsf{P}[N/\mathsf{E}[N] \leq x] = \mathsf{P}[\mathrm{IG}(1,\omega) \leq x]$$

med $\omega = A\,\mathsf{E}[Z]/\mathsf{V}[Z]$.

Grænsefordelingen for $N/\mathsf{E}[N]$ er således en invers Gauss fordeling med skalaparameter 1 og med formparameter $A\,\mathsf{E}[Z]/\mathsf{V}[Z]$.

Da egenskaberne for sekvenstestet blev beskrevet først af Wald(1947), kaldes $\mathrm{IG}(1,\omega)$-fordelingen ofte for Wald-fordelingen.

4.2.9 IG-fordelingen som levetidsfordeling

For $T \in \mathrm{IG}(\mu, \lambda)$ gælder

$$\overline{F}(t) = \Phi[\sqrt{\lambda/t}\,(1 - t/\mu)] - \exp(2\lambda/\mu)\,\Phi[-\sqrt{\lambda/tx}\,(t/\mu + 1)] \tag{4.2.14}$$

$$\lambda(t) = \frac{\sqrt{\lambda/(2\pi t^3)}\,\exp[-\lambda(t-\mu)^2/(2\mu^2 t)]}{\overline{F}(t)} \tag{4.2.15}$$

$$\mathsf{E}[T] = \mu \tag{4.2.16}$$

$$\mathsf{V}[T] = \mu^3/\lambda \tag{4.2.17}$$

Der gælder at $\lambda(\cdot)$ først er voksende, og derefter aftagende. Maksimumværdien for $\lambda(\cdot)$ fås for løsningen t til

$$\frac{\lambda}{2\mu^2} + \frac{3}{2t} - \frac{\lambda}{2t^2} = 0$$

Der gælder endvidere $\lambda(t) \to \lambda/(2\mu^2)$ for $t \to \infty$. (Se Chhikara and Folks (1977)).

4.3 Den generaliserede inverse Gaussfordeling

Den modificerede Bessel funktion af tredie orden med index $\lambda \in \mathbb{R}$ er funktionen

$$\mathrm{K}_\lambda(\omega) = \frac{1}{2}\int_0^\infty x^{\lambda-1}\exp[-\frac{1}{2}\omega(x + x^{-1})]\,dx \tag{4.3.1}$$

for $\omega > 0$. Der gælder

$$\begin{aligned}
\mathrm{K}_\lambda(\omega) &= \mathrm{K}_{-\lambda}(\omega) \\
\mathrm{K}_{\lambda+1}(\omega) &= \frac{2\lambda}{\omega}\mathrm{K}_\lambda(\omega) + \mathrm{K}_{\lambda-1}(\omega) \\
\mathrm{K}_{\lambda-1}(\omega) + \mathrm{K}_{\lambda+1}(\omega) &= -2\mathrm{K}'_\lambda(\omega) \\
\mathrm{K}_{1/2}(\omega) &= \mathrm{K}_{-1/2}(\omega) = \sqrt{\pi/(2\omega)}\,\exp(-\omega)
\end{aligned}$$

Definition 4.3.1 ***Generaliseret invers Gaussfordeling***

Den positive stokastiske variable, X siges at følge en generaliseret invers Gaussfordeling med parametrene (λ, χ, ψ) hvis tætheden for X er på formen

$$f(x;\lambda,\chi,\psi) = \frac{(\lambda/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\chi\psi})}\, x^{\lambda-1} \exp\left\{-\frac{1}{2}\left(\chi x^{-1} + \psi x\right)\right\} \tag{4.3.2}$$

for $x > 0$, hvor $\mathrm{K}_\lambda(\cdot)$ er den modificerede Bessel funktion af tredie orden med index λ.

Parameterområdet for familien er

$$\lambda \in \mathbb{R}\,, \quad (\chi,\psi) \in \Omega_\lambda\,,$$

hvor

$$\Omega_\lambda = \begin{cases} \{(\chi,\psi) : \chi \geq 0\,,\ \psi > 0\} & \text{for } \lambda > 0 \\ \{(\chi,\psi) : \chi > 0\,,\ \psi > 0\} & \text{for } \lambda = 0 \\ \{(\chi,\psi) : \chi > 0\,,\ \psi \geq 0\} & \text{for } \lambda < 0 \end{cases}$$

□

For $\chi = 0\,, \lambda > 0$ fremkommer familien af gammafordelinger (afsnit 4.9); for $\psi = 0\,, \lambda = 0$ fremkommer familien af reciproke gammafordelinger (afsnit 4.11.1), og for $\lambda = -1/2$ fremkommer de inverse gaussfordelinger, mens fordelingen af den reciproke af en invers gaussfordelt størrelse fremkommer for $\lambda = 1/2$.

Bemærkning 1 ***Familien er afsluttet over skalatransformationer og reciprokdannelse***

Såfremt X følger en generaliseret invers gaussfordeling med parametrene (λ, χ, ψ), da vil fordelingen af X^{-1} være en generaliseret invers gaussfordeling med parametrene $(-\lambda, \chi, \psi)$.

Såfremt X følger en generaliseret invers gaussfordeling med parametrene (λ, χ, ψ), da vil fordelingen af βX med $\beta > 0$ være en generaliseret invers gaussfordeling med parametrene $(\lambda, \beta\chi, \psi/\beta)$. □

Undertiden benyttes en parametrisering ved

$$\omega = \sqrt{\chi\psi}\,, \quad \eta = \sqrt{\chi/\psi}\,,$$

hvor $\omega = 0$ hvis enten $\chi = 0$, , $\lambda > 0$, eller hvis $\psi = 0$, $\lambda > 0$. For $\omega > 0$ har man tætheden

$$f(x;\lambda,\omega,\eta) = \frac{\eta^{-\lambda}}{2K_\lambda(\omega)}\, x^{\lambda-1} \exp\Big\{-\frac{\omega}{2}(\eta^{-1}x + \eta x^{-1})\Big\}\,; \quad 0 < x < \infty \tag{4.3.3}$$

Det følger af den foregående bemærkning, at parameteren ω ikke ændres, når man foretager en skalatransformation af den variable, eller når man betragter fordelingen af den reciproke variabel. Man kalder derfor ofte ω for formparameteren for familien.

Jørgensen (1982) har givet en indgående beskrivelse af de statistiske egenskaber for generaliseret inverse Gaussfordelinger.

4.4 Den logaritmisk normale fordeling

4.4.1 Den logaritmisk normale fordeling som levetidsfordeling

Den logaritmisk normale fordeling er introduceret i Introduktion til Statistik, Bind 1. Vi vil her blot supplere med enkelte betragtninger i forbindelse med anvendelsen af den logaritmisk normale fordeling som levetidsfordeling.

En mere omfattende, generel behandling af den logaritmisk normale fordeling er givet af Crow og Shimizu (1988).

Definition 4.4.1 ***Den logaritmisk normale fordeling***

En stokastisk variabel T følger den logaritmisk normale fordeling med parametre (α, β^2), hvis tætheden for T er

$$f(t) = \frac{1}{t\beta\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(t)-\alpha}{\beta}\right)^2\right] \quad \text{for} \quad t > 0$$

og $f(t) = 0$ ellers.

Vi skriver kort $T \in \mathrm{LN}(\alpha, \beta^2)$. □

Sætning 4.4.1 *Overlevelsesfunktion og hændelsesrate for LN-fordelingen*

Lad $T \in \mathrm{LN}(\alpha, \beta^2)$. Da gælder:

$$\overline{F}(t) = 1 - \Phi\Big(\frac{\ln(t) - \alpha}{\beta}\Big) \tag{4.4.1}$$

$$\lambda(t) = \frac{f(t)}{\overline{F}(t)} \tag{4.4.2}$$

$$\lambda(0) = 0; \qquad \lambda(t) \to 0 \quad \text{for} \quad t \to \infty;$$

$$\mathsf{E}[T] = \exp(\alpha + \tfrac{1}{2}\beta^2); \quad \mathsf{V}[T] = \mathsf{E}[T]^2(\exp(\beta^2) - 1)$$

Der gælder, at $\lambda(t)$ først er voksende fra nul mod et maksimum, og derefter aftager $\lambda(t)$ mod nul, når t vokser mod ∞. På trods heraf benyttes den logaritmisk normale fordeling undertiden til beskrivelse af levetidsfordelinger. □

4.5 Den logistiske fordeling

Definition 4.5.1 *Logistisk fordeling*

En stokastisk variabel X følger den logistiske fordeling[1] med parametre (0,1), hvis tætheden for X er

$$f(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2} \qquad \text{for} \quad -\infty < x < \infty$$

Vi skriver kort $X \in \mathrm{L}(0,1)$. Fordelingen med positionsparameter α og skalaparameter $\beta > 0$ benævnes $\mathrm{L}(\alpha, \beta)$.

Den logististiske fordeling bruges undertiden som levetidsfordeling til trods for at fordelingens støtte omfatter negative værdier. Det ses umiddelbart, at overlevelsesfunktionen for en $\mathrm{L}(\alpha, \beta)$-fordelt levetid er

$$\overline{F}(t) = \frac{\exp(-\frac{t-\alpha}{\beta})}{1 + \exp(-\frac{t-\alpha}{\beta})} \qquad \text{for } -\infty < t < \infty$$

[1] En række forfattere benytter betegnelsen logistisk fordeling for fordelingen af $-X$. Vi har i denne fremstilling valgt en parametrisering svarende til Introduktion til Statistik, Bind 1.

Sætning 4.5.1 ***Forventningsværdi og varians for $L(\alpha, \beta)$-fordelingen***

Lad $X \in \mathrm{L}(\alpha, \beta)$. Da er

$$\mathsf{E}[X] = \alpha; \quad \mathsf{V}[X] = \frac{\beta^2 \pi^2}{3}$$

Bevis:
Forbigås

□

Bemærkning 1 ***Brug af logistisk fordeling som substitut for normalfordeling.*** Den logistiske fordeling anvendes i en række sammenhænge i stedet for den normale fordeling. Dette begrundes hovedsageligt i den mere direkte udregning af fordelingsfunktion (og overlevelsesfunktion) for den logistiske fordeling og i at fordelingsfunktionen for $\mathrm{L}(\alpha, \beta)$-fordelingen og for $\mathrm{N}(\alpha, \pi^2\beta^2/3)$-fordelingen ikke adskiller sig væsentligt fra hinanden. □

4.6 Den log-logistiske fordeling

Definition 4.6.1 ***$LL(\alpha, \beta)$-fordelingen***

En positiv stokastisk variabel T siges at følge en log-logistisk fordeling med parametrene $\alpha > 0$ og $\beta > 0$, såfremt $Y = -\ln(T)$ følger en $\mathrm{L}(-\ln(\beta), 1/\alpha)$-fordeling.

Kort skriver vi $T \in \mathrm{LL}(\alpha, \beta)$.

Den log-logistiske fordeling bruges undertiden som levetidsfordeling.

Det ses umiddelbart, at overlevelsesfunktionen for en $\mathrm{LL}(\alpha, \beta)$-fordelt levetid er:

$$\overline{F}(t) = \frac{1}{1 + (t/\beta)^\alpha} \quad \text{for} \quad t > 0$$

og at tætheden er

$$f(t) = \frac{\alpha}{\beta} \frac{(t/\beta)^{\alpha-1}}{[1 + (t/\beta)^\alpha]^2} \quad \text{for} \quad t > 0$$

Sætning 4.6.1 ***Hændelsesrate for LL(α, β)-fordelingen***

Lad $T \in \mathrm{LL}(\alpha, \beta)$. Da er hændelsesraten:

$$\lambda(t) = \frac{\alpha}{\beta} \frac{(t/\beta)^{\alpha-1}}{[1 + (t/\beta)^{\alpha}]}$$

og den kumulerede hændelsesrate:

$$\Lambda(t) = -\ln(1 + (t/\beta)^{\alpha})$$

Hændelsesraten er aftagende med t for $\alpha \leq 1$, og for $\alpha > 1$ gælder at $\lambda(0) = 0$, og at hændelsesraten derefter vokser mod et maksimum, hvorefter hændelsesraten aftager mod $\lambda(\infty) = 0$.

Bevis:

Bevises direkte.

□

Bemærkning 1 ***Sammenligning med lognormalfordelingen.*** Det ses, at forløbet af hændelsesraten er analogt til forløbet af hændelsesraten for logaritmisk normalt fordelte levetider fejl, svarende til at den logistiske fordeling minder om normalfordelingen. □

4.7 Den hyperbolske secantfordeling

Definition 4.7.1 ***Den hyperbolske secantfordeling***

Vi indleder med at definere den simple hyperbolske secantfordeling.

Lad $Y \in \mathrm{Be}(0.5 + \vartheta/\pi, 0.5 - \vartheta/\pi)$, med $|\vartheta| < \pi/2$, og lad

$$X = \ln[Y/(1-Y)]/\pi$$

Da har X tætheden

$$f(x;\vartheta) = \frac{\exp\{\vartheta x + \ln[\cos(\vartheta)]\}}{2\cosh(\pi x/2)}\ , \quad -\infty < x < \infty \tag{4.7.1}$$

med hensyn til Lebesguemålet.

Fordelingen af X kaldes den hyperbolske secantfordeling

□

For $\vartheta = 0$ er fordelingen symmetrisk omkring nul med varians 1 og med kumuleret fordeling

$$F(x;0) = \frac{1}{\pi}\arctan[\sinh(\pi x/2)] + 1/2$$

Familien af fordelinger med tætheder (4.7.1) udgør en naturlig eksponentiel familie med kumulantfrembringeren $\kappa(\vartheta) = -\ln(\cos(\vartheta))$ for $|\vartheta| < \pi/2$. Det kanoniske parameterområde er $\Theta =]-\pi/2, \pi/2[$, og støtten og middelværdirummet er begge $\mathbb{R}$.

Den hyperbolske secantfordeling er uendelig delbar (se Feller (1971)).

4.7.1 Den generaliserede hyperbolske secantfordeling

Familien (4.7.1) frembringer en additiv eksponentiel dispersionsmodel med indeksmængde $\Lambda = \mathbb{R}_+$.

Den herved frembragte familie kaldes den generaliserede hyperbolske secantfordeling. Familien har tætheden

$$f_Z(z;\vartheta,\lambda) = a^*(z;\lambda)\,\frac{\exp\{\vartheta z + \lambda\ln[\cos(\vartheta)]\}}{2\cosh(\pi z/2)}\,,\ z \in \mathbb{R} \tag{4.7.2}$$

med

$$\begin{aligned} a^*(z;\lambda) &= \frac{2^{\lambda-2}|\Gamma(\lambda/2 + \mathrm{i}z/2)|^2}{\pi\Gamma(\lambda)} \\ &= \frac{2^{\lambda-2}\{\Gamma(\lambda/2)\}^2}{\pi\Gamma(\lambda)}\prod_{j=0}^{\infty}\left\{1 + \frac{z^2}{(\lambda+2j)^2}\right\}^{-1} \end{aligned} \tag{4.7.3}$$

Den reproduktive form for den generaliserede hyperbolske secantfordeling har tætheden

$$f_Y(y;\mu,\sigma^2) = \frac{a^*(y/\sigma^2;1/\sigma^2)}{\sigma^2}\exp\left[\frac{1}{\sigma^2}\left\{y\arctan\mu - \frac{1}{2}\ln(1+\mu^2)\right\}\right] \tag{4.7.4}$$

for $y \in \mathbb{R}$, hvor $a^*(z,\lambda)$ er givet ved (4.7.3).

Såfremt fordelingen af Y har tætheden (4.7.4) med hensyn til Lebesguemålet siger vi, at Y følger en generaliseret hyperbolsk secantfordeling med parametrene μ og σ^2. Kort skriver vi $Y \in \mathrm{GHS}(\mu, \sigma^2)$.

De vigtigste størrelser i fortolkningen af familien af $\mathrm{GHS}(\mu, \sigma^2)$-fordelinger som en reproduktiv eksponentiel dispersionsmodel er anført i nedenstående oversigt:

$\mathrm{GHS}(\mu, \sigma^2)$-fordelingen som reproduktiv eksponentiel dispersionsmodel

Kanonisk parameter ϑ	Kumulant-frembringer $\kappa(\vartheta)$	Middelværdi-afb. $\mu = \tau(\vartheta)$	Enheds-variansfunktion $V_{GHS}(\mu)$	dispersions-parameter σ^2
$\arctan(\mu)$	$-\ln\cos(\vartheta)$	$\tan(\vartheta)$	$1+\mu^2$	σ^2

Den kanoniske link er

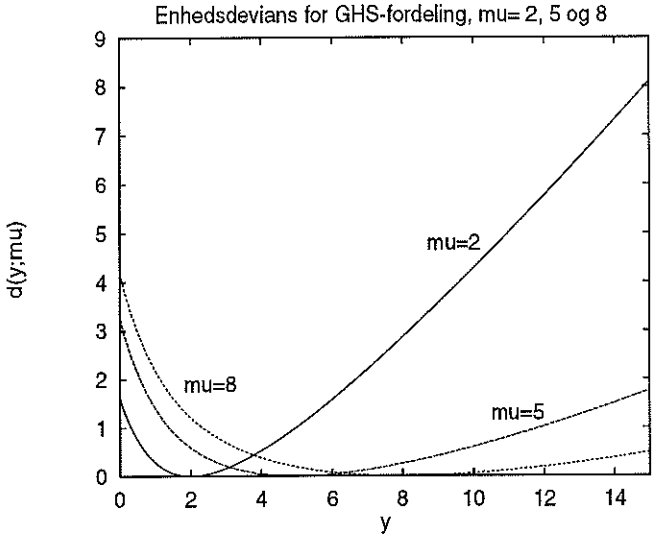
$$\vartheta = \tau^{-1}(\mu) = \arctan(\mu)$$

Enhedsdeviansen for $\mathrm{GHS}(\mu, \sigma^2)$-fordelingen er

$$d(y;\mu) = 2y\{\arctan(y) - \arctan(\mu)\} + \ln\left(\frac{1+\mu^2}{1+y^2}\right) \tag{4.7.5}$$

Tætheden (4.7.4) for den reproduktive form af den generaliserede hyperbolske secantfordeling er netop udtrykt ved enhedsdeviansen på formen (1.3.20)

Nedenstående figur viser enhedsdeviansen svarende til $\mu = 2$, $\mu = 5$ og $\mu = 8$.



Sætning 4.7.1 ***Konvergensegenskaber***

Lad fordelingen af Y være en GHS(μ, σ^2)-fordeling. Da gælder, at fordelingen af $X = Y/c$ vil konvergere mod en G(α, β)-fordeling med

$$\alpha = \frac{1}{\sigma^2} \; ; \quad \beta = \mu \times \sigma^2$$

dvs. en gammafordeling med middelværdi μ og med relativ spredning (variationskoefficient) σ.

Bevis:

Beviset følger af sætning 1.3.7 □

4.7.2 Den generaliserede hyperbolske secantfordeling med $\vartheta = 0$

Lad specielt X_1 og X_2 være uafhængige identisk fordelt med tætheden (4.7.1) og lad $Z = X_1 + X_2$, da har Z tætheden

$$f_Z(z, 0, 2) = \frac{z}{2\sinh(\pi z/2)} \; , \; -\infty < z < \infty$$

Lad $X_1, X_2, \ldots, X_{n+2}$ være uafhængige identisk fordelt med tætheden (4.7.1), og lad $Z = X_1 + X_2 + \cdots + X_{n+2}$, da har Z tætheden

$$f_Z(z; 0, n+2) = \frac{z^2 + n^2}{n(n+1)} f_Z(z; n, 0), \ -\infty < z < \infty$$

For heltallige værdier, r, af formparameteren λ gælder åbenbart at $f_Z(z; 0, r)$ er et r'te gradspolynium divideret med $\cosh(\pi z/2)$ såfremt r er ulige, og divideret med $\sinh(\pi z/2)$ såfremt r er lige. Såfremt formparameteren λ ikke er heltallig kan $f_Z(x; 0, \lambda)$ udtrykkes som et uendeligt produkt

$$f_Z(z; 0, \lambda) = \frac{2^{\lambda-2}\{\Gamma(\lambda/2)\}^2}{\pi\Gamma(\lambda)} \prod_{j=0}^{\infty} \{1 + x^2/(\lambda + 2j)^2\}^{-1}$$

4.8 Eksponentialfordelingen

Den kontinuerte analog til den geometriske fordeling er eksponentialfordelingen.

Definition 4.8.1 ***Eksponentialfordeling*** En kontinuert stokastisk variabel X, der kan antage alle reelle ikke-negative værdier, siges at følge en eksponentialfordeling med parameter β, hvis tætheden for X er af formen

$$g(x) = \frac{1}{\beta} \exp(-x/\beta) \quad \text{for} \quad 0 < x \tag{4.8.1}$$

hvor $0 < \beta$.

Kort skriver vi $X \in \mathrm{Ex}(\beta)$. □

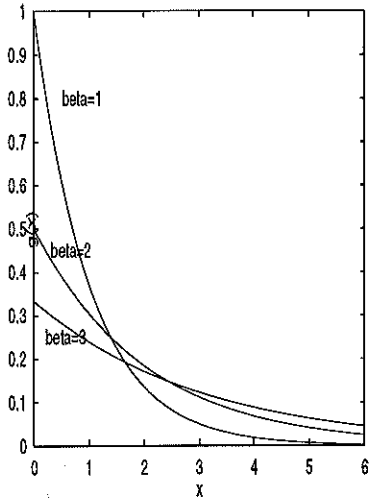
Den karakteristiske funktion er

$$\phi(t) = \frac{1}{1 - \mathrm{i}\,\beta t}$$

Såfremt $X \in \mathrm{Ex}(\beta)$, og

$$Y = \alpha X$$

med $0 < \alpha$, vil $Y \in \mathrm{Ex}(\alpha\beta)$. Parameteren β er således en skalaparameter.



Figur 4.1. Tætheden for Ex(β)-fordelingen svarende til forskellige værdier af β

Figur 4.1 viser tæthederne for eksponentialfordelinger svarende til forskellige værdier af skalaparameteren β.

Såfremt $X \in \mathrm{Ex}(\beta)$ gælder

$$\mathrm{E}[X] = \beta; \qquad \mathrm{V}[X] = \beta^2 \tag{4.8.2}$$

4.8.1 Fordelingsfunktion og ufuldstændige momenter

Fordelingsfunktionen for $\mathrm{Ex}(\beta)$-fordelingen betegnes med

$$\mathrm{Ex}(c;\beta) \overset{\mathrm{DEF}}{=} \int_0^c \frac{1}{\beta}\exp(-t/\beta)dt = 1 - \exp(-c/\beta)$$

Da β er en skalaparameter gælder

$$\mathrm{Ex}(c;\beta) = \mathrm{Ex}(c/\beta;1)$$

Det første ufuldstændige moment for $\mathrm{Ex}(\beta)$-fordelingen er

$$\mu_1'(c) = \int_0^c xg(x)dx = \beta \mathrm{G}(c/\beta;2,1)$$

hvor $\mathrm{G}(c;\alpha,\beta)$ angiver fordelingsfunktionen for $\mathrm{G}(\alpha,\beta)$-fordelingen.

Det tilsvarende centrale moment er

$$\mu_1(c) = \int_0^c (x-\beta)g(x)dx = -c\exp(-c/\beta)$$

4.8.2 Eksponentialfordelingen som eksponentiel familie

Familien af $\mathrm{Ex}(\beta)$-fordelinger for $0 < \beta < \infty$ udgør en naturlig eksponentiel familie med kanonisk parameter $\vartheta = -1/\beta$, kanonisk stikprøvefunktion $t(x) = x$, kanonisk parameterområde $D =]-\infty, 0[$ og kumulantfrembringer

$$\kappa(\vartheta) = \ln(-\vartheta)\ ,$$

hvorfor $\tau(\vartheta) = -1/\vartheta$ og $\mathsf{V}[X] = 1/\vartheta^2$. Den sædvanlige parametrisering er ved middelværdiparameteren $\mu = \mathsf{E}[X] = \beta = -1/\vartheta$.

Idet $\vartheta = \tau^{-1}(\mu) = -1/\mu$ ser vi, at den kanoniske link er den reciproke funktion.

Vi har variansen $\mathsf{V}[X] = \beta^2 = \mu^2$, dvs variansfunktionen er

$$V_G(\mu) = \mu^2$$

4.8.3 Reproduktivitetsegenskaber

Eksponentialfordelingen er elementarfordeling for Gammafordelingen (se afsnit 4.9.3), idet der gælder, at såfremt $X_1, X_2, \ldots, X_k$ er uafhængige variable, hvor $X_i \in \mathrm{Ex}(\beta)$, for $i = 1, 2, \ldots, k$, og vi sætter

$$Y = X_1 + X_2 + \cdots + X_k,$$

da vil $Y \in \mathrm{G}(k, \beta)$.

Såfremt $X_1, X_2, \ldots$ er en uendelig følge af uafhængige variable, hvor $X_i \in \mathrm{Ex}(\beta)$, for $i = 1, 2, \ldots$, og vi sætter

$$Y = \max\{n : X_1 + X_2 + \cdots + X_n \leq t\},$$

da vil $Y \in \mathrm{P}(t/\beta)$.

Dvs såfremt ventetiderne mellem successive hændelser er uafhængige og identisk fordelt som eksponentialt fordelte variable, da vil antallet af hændelser, der indtræffer i et tidsrum af længden t være Poissonfordelt med intensiteten t/β, hvor β angiver den forventede ventetid mellem to hændelser. Ved anvendelser i forbindelse, fx med reparerbare systemer, kaldes β derfor MTBF (Mean Time Between Failures) og $1/\beta$ betegnes undertiden failure-rate.

4.8.4 Eksponentialfordelingen som levetidsfordeling

Lad $T \in \mathrm{Ex}(\beta)$. Da er

$$f(t) = \frac{1}{\beta}\exp(-t/\beta), \quad \text{og} \quad \overline{F}(t) = \exp(-t/\beta)$$

Det er ofte naturligt, at tænke på T som havende dimensionen [tid]. Skalaparameteren β har da ligeledes dimensionen [tid], $f(t)$ får da dimensionen $[\text{tid}]^{-1}$, (svarende til at sandsynlighedselementet $f(t)dt$ er dimensionsløst). Den variable s i Laplacetransformen har dimensionen $[\text{tid}]^{-1}$, og Laplacetransformen og den kumulerede fordelingsfunktion bliver da dimensionsløse størrelser.

Der gælder:

$$\begin{array}{ll} \mathsf{E}[T] = \beta; & \mathsf{V}[T] = \beta \\ \lambda(t) = 1/\beta; & r(t) = \beta \end{array}$$

hvor $\lambda(\cdot)$ angiver hændelsesraten, og $r(\cdot)$ angiver den forventede restlevetid.

Vi noterer, at parameteren β er en skalaparameter, og at β netop angiver middellevetiden. Vi ser, at eksponentialfordelingen er karakteriseret ved at have konstant hændelsesrate $1/\beta$.

For den eksponentielle levetidsfordeling gælder, at fordelingen af restlevetiden ikke afhænger af den opnåede alder. Vi har nemlig

$$\overline{F}(x|t) = \overline{F}(x) \quad \text{for alle } x, t \geq 0$$

Omvendt ses, at funktionalligningen

$$\overline{F}(t+x) = \overline{F}(t)\overline{F}(x),$$

netop har løsningen $\overline{F}(x) = \exp(-\lambda x)$.

Eksponentialfordelingen er således netop karakteriseret ved at fordelingen af restlevetiden ikke afhænger af den opnåede alder.

Genesis: afstand ml. upcrossings (Cram. Leadbetter, Lindgren, Rootzen.)

Sætning 4.8.1 ***Minimumsfordelinger og eksponentialfordelingen***

$$T \in \mathrm{Ex}(\beta) \quad \text{og} \quad Y = \ln(T) \quad \Rightarrow \quad Y \in \mathrm{Min}_1(\ln(\beta), 1)$$

$$(T_1, T_2) \in \mathrm{Ex}(\beta), T_1, T_2 \text{ uafhængige} \quad \Rightarrow \quad \min(T_1, T_2) \in \mathrm{Ex}(\beta/2)$$

Bevis:
Sætningen bevises direkte ved anvendelse af definitionerne. □

Bemærkning:
Transformationen $Y = \ln(T)$ fører således skalaparameteren β for T over i positionsparameteren $\ln(\beta)$ for Y.

Sætning 4.8.2 ***Fordeling af afstand mellem ordnede observationer***

Lad $T_1, T_2, \ldots, T_n$ angive en række uafhængige, $\mathrm{Ex}(\beta)$-fordelte stokastiske variable, og lad $T_{n,(1)}, T_{n,(2)}, \ldots, T_{n,(n)}$ angive de tilsvarende ordnede observationer, og lad $D_{n,1}, D_{n,2}, \ldots, D_{n,n}$ angive afstandene mellem de ordnede observationer, dvs

$$D_{n,1} = T_{n,(1)}; \quad D_{n,2} = T_{n,(2)} - T_{n,(1)}; \quad \ldots; D_{n,n} = T_{n,(n)} - T_{n,(n-1)}$$

Da er $D_{n,1}, D_{n,2}, \ldots, D_{n,n}$ stokastisk uafhængige, og $D_{n,k} \in \mathrm{Ex}(\beta/(n-k+1))$.

Der gælder derfor:

$$\mathsf{E}[D_{n,k}] = \beta/(n-k+1); \quad \mathsf{V}[D_{n,k}] = \beta/(n-k+1)$$

samt

$$\mathsf{E}[T_{n,(k)}] = \beta\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-k+1}\right)$$

og

$$\mathsf{V}[T_{n,(k)}] = \beta\left(\frac{1}{n^2} + \frac{1}{(n-1)^2} + \cdots + \frac{1}{(n-k+1)^2}\right)$$

Bevis:
Sætningen bevises direkte ved opskrivning af den simultane tæthed for $T_1, T_2, \ldots, T_n$.

□

Sætning 4.8.3 ***Fordeling af total testtid***

Lad $T_1, T_2, \ldots, T_n$ angive en række uafhængige, $\mathrm{Ex}(\beta)$-fordelte stokastiske variable, og lad

$$Z = T_{n,(1)} + T_{n,(2)} + \cdots + (n-r+1)T_{n,(r)},$$

hvor $T_{n,(1)}, T_{n,(2)}, \ldots, T_{n,(n)}$ angiver de tilsvarende ordnede observationer.

Der gælder da, at $Z \in \mathrm{G}(r, \beta)$.

Bevis:

Resultatet fås af sætning 4.8.2 ved at bemærke, at

$$Z = n\, D_{n,1} + (n-1)D_{n,2} + \cdots + (n-r)D_{n,r-1} + (n-r+1)D_{n,r}$$

□

Bemærkning

Hvis T_i angiver levetiden af en komponent, da angiver Z totaltesttiden for en prøvning af n komponenter, der afsluttes, når den r'te komponent er fejlet (se bemærkning 1 på side 78).

4.9 Gammafordelingen

Definition 4.9.1 ***Gammafordeling***

En kontinuert stokastisk variabel X, der kan antage alle reelle ikke-negative værdier, siges at følge en gammafordeling med parametre α og β, hvis tætheden for X er af formen

$$g(x) = \frac{1}{\beta\Gamma(\alpha)} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp(-x/\beta) \quad \text{for} \quad x \in \mathbb{R}_+ \tag{4.9.1}$$

hvor $\alpha \in \mathbb{R}_+$ og $\beta \in \mathbb{R}_+$.

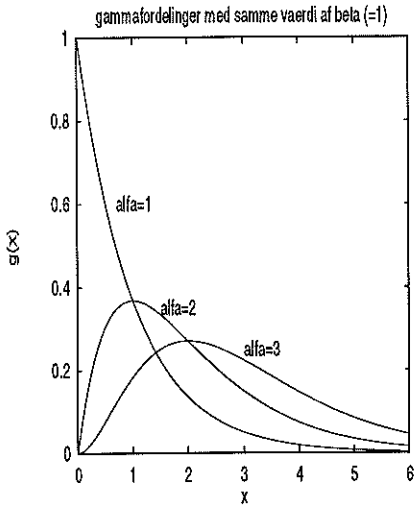
Kort skriver vi $X \in \mathrm{G}(\alpha, \beta)$. □

Den karakteristiske funktion for $\mathrm{G}(\alpha, \beta)$-fordelingen er

$$\phi(t) = \frac{1}{(1 - \mathrm{i}t\beta)^{\alpha}}$$

Såfremt $X \in \mathrm{G}(\alpha, \beta)$ gælder

$$\mathsf{E}[X] = \alpha\beta\,; \quad \mathsf{V}[X] = \alpha\beta^2$$



Figur 4.2. Tætheden for $G(\alpha, 1)$-fordelingen svarende til forskellige værdier af formparameteren α

Såfremt $X \in G(\alpha, \beta)$, og vi sætter $Y = \beta_1 X$, vil $Y \in G(\alpha, \beta_1\beta)$. Parameteren β er altså en skalaparameter for fordelingen. Parameteren α kaldes formparameteren.

$G(1, \beta)$-fordelingen er netop $Ex(\beta)$ fordelingen. $G(f/2, 2)$-fordelingen er den fra statistikken kendte χ^2-fordeling med f frihedsgrader.

Figur 4.2 viser tæthederne for gammafordelinger med samme skalaparameter, $\beta = 1$, men forskellige værdier af formparameteren, α.

Gammafordelingen er uendeligt delbar.

4.9.1 Fordelingsfunktion og ufuldstændige momenter

Fordelingsfunktionen for $G(k,\beta)$-fordelingen betegnes med

$$G(c;k,\beta) \stackrel{DEF}{=} \int_0^c \frac{1}{\beta\Gamma(k)} \left(\frac{x}{\beta}\right)^{k-1} \exp(-x/\beta)dx$$

Da β er en skalaparameter gælder

$$G(c;k,\beta) = G(c/k;1) = \int_0^{c/\beta} \frac{1}{\Gamma(k)}\, x^{k-1}\exp(-x)dx = \frac{\Gamma_{c/\beta}(k)}{\Gamma(k)}$$

hvor den ufuldstændige Gammafunktion $\Gamma_x(k)$ er defineret ved

$$\Gamma_x(k) \stackrel{DEF}{=} \int_0^x \frac{1}{\Gamma(k)}\, u^{k-1}\exp(-u)du$$

For heltallige værdier af k finder man ved delt integration

$$G(c;k,\beta) = 1 - P(k-1;c/\beta)$$

For $k \to \infty$ kan man approksimere gammafordelingen med en normalfordeling med samme middelværdi og varians:

$$G(c;k,\beta) = \Phi(u)$$

med

$$u = \frac{c/\beta - k}{\sqrt{k}}$$

Approksimationen kan bruges for store værdier af k og moderate værdier af β.

Det første ufuldstændige moment for $G(k,\beta)$-fordelingen er

$$\mu_1'(c) = \int_0^c xg(x)dx = k\beta G(c/\beta;k+1,1)$$

Det tilsvarende centrale moment er

$$\mu_1(c) = \int_0^c (x-k\beta)g(x)dx = k\beta[G(c/\beta;k+1,1) - G(c/\beta;k,1)]$$

der for heltallige værdier af k reduceres til

$$\mu_1(c) = -k\beta\frac{(c/\beta)^k}{k!}\exp(-c/\beta).$$

4.9.2 Gammafordelingen som eksponentiel familie

Det fremgår af (4.9.1), at $\mathrm{G}(\alpha,\beta)$-fordelingen har tætheden

$$g(x) = \frac{1}{x} \times \exp\{\alpha \ln(x) - x/\beta - \alpha \ln(\beta) - \ln(\Gamma(\alpha))\} \quad \text{for } x \in \mathbb{R}_+$$

med hensyn til Lebesguemålet.

Familien af gammafordelinger er derfor en eksponentiel familie af orden 2 med de kanoniske parametre $(\vartheta_1,\vartheta_2) = (\alpha,-1/\beta)$, de kanoniske stikprøvefunktioner $\big(t_1(x),t_2(x)\big) = \big(\ln(x),x\big)$, kanonisk parameterområde $\Omega = \mathbb{R}_+\times]-\infty,0[$, og med $\kappa(\vartheta) = \ln\Gamma(\vartheta_1) - \vartheta_1\ln(-\vartheta_2)$, hvorfor

$$\tau(\vartheta) = \mathsf{E}\begin{pmatrix} \ln(X) \\ X \end{pmatrix} = \begin{pmatrix} \Psi(\vartheta_1) - \ln(\vartheta_2) \\ -\vartheta_1/\vartheta_2 \end{pmatrix}$$

og

$$\mathbf{V}(\vartheta) = \mathsf{D}\begin{pmatrix} \ln(X) \\ X \end{pmatrix} = \begin{pmatrix} \Psi'(\vartheta_1) & -1/\vartheta_2 \\ -1/\vartheta_2 & \vartheta_1/\vartheta_2^2 \end{pmatrix},$$

hvor $\Psi(\cdot)$ angiver digammafunktionen, $\Psi(x) = \Gamma'(x)/\Gamma(x)$.

4.9.3 Gammafordelingen som eksponentiel dispersionsmodel

Vi minder om, at familien af $\mathrm{Ex}(\beta)$-fordeling udgør en endimensional naturlig eksponentiel familie.

Det følger af additionsegenskaberne for eksponentialfordelingen (afsnit 4.8.3), at den additive eksponentielle dispersionsmodel frembragt af eksponentialfordelingen netop er familien af gammafordelinger.

Indeksmængden er $\Lambda = \mathbb{R}_+$, den kanoniske parameter er $\vartheta = -1/\beta$, enhedskumulantfrembringeren

$$\kappa(\vartheta) = \ln(-\vartheta)$$

og enhedsvariansfunktionen er

$$V_G(\mu) = \mu^2$$

Såfremt $Z \in \mathrm{G}(\alpha, \beta)$ opfattes som en additiv dispersionsmodel finder vi tætheden for Z på formen

$$f^*(z; \vartheta, \alpha) = \frac{1}{\Gamma(\alpha)}\, z^{\alpha-1} \exp\{\vartheta z + \alpha \ln(-\vartheta)\} \tag{4.9.2}$$

med $\vartheta < 0$. Udtrykket er netop på formen (1.3.8), idet vi har benyttet symbolet α for indeksparameteren.

Ved den sædvanlige transformation $Y = Z/\alpha$ føres den additive model over i en reproduktiv eksponentiel dispersionsmodel for Y. Dette er jo blot en skalatransformation af fordelingen. Fordelingen af Y bliver da en $\mathrm{G}(\alpha, \beta/\alpha)$-fordeling.

Såfremt $Y \in \mathrm{G}(\alpha, \beta/\alpha)$, kan vi udtrykke tætheden for Y som

$$f(y; \mu, \sigma^2) = h(y; \sigma^2)\ \exp\Big\{ -\ \frac{1}{\sigma^2}\Big(\frac{y}{\mu} + \ln(\mu)\Big)\Big\} \tag{4.9.3}$$

med

$$h(y; \sigma^2) = \frac{1}{\Gamma(1/\sigma^2)} \Big(\frac{y}{\sigma^2}\Big)^{1/\sigma^2} \frac{1}{y}$$

og hvor

$$\mu = \beta\,, \quad \sigma^2 = 1/\alpha$$

Dette er netop på formen (1.3.22)

Familien af gammafordelinger $\mathrm{G}(\alpha, \beta/\alpha)$ med $\alpha \in \mathbb{R}_+$ og $\beta \in \mathbb{R}_+$ er en reproduktiv eksponentiel dispersionsmodel med kanonisk parameter $\vartheta = -1/\beta$ og dispersionsparameter $\sigma^2 = 1/\alpha$.

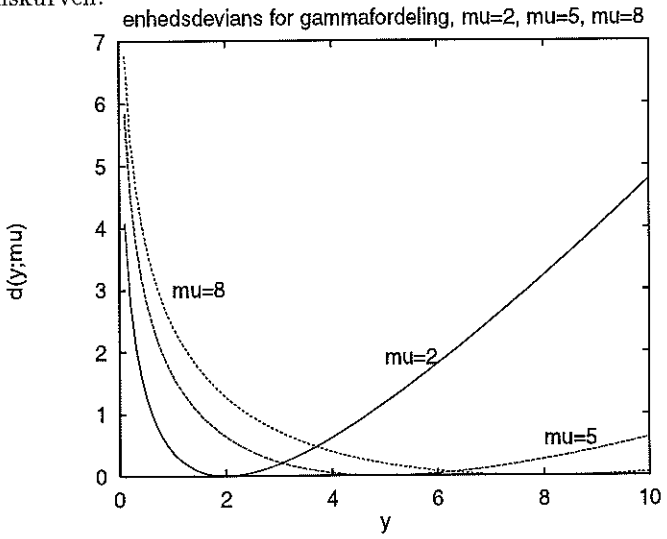
Vi har enhedskumulantfrembringeren $\kappa(\vartheta) = -\ln(-\vartheta)$, middelværdien $\mu = \beta =$ og variansen,

$$\mathsf{V}[Y] = \sigma^2 V_G(\mu) = \sigma^2 \mu^2 = \frac{\beta^2}{\alpha}$$

Enhedsdeviansen for $\mathrm{G}(\alpha, \beta/\alpha)$-fordelingen er

$$d(y;\mu) = 2\Big(\frac{y}{\mu} - \ln\Big(\frac{y}{\mu}\Big) - 1\Big) \tag{4.9.4}$$

Nedenstående figur viser enhedsdeviansen svarende til forskellige værdier af middelværdien μ. Man ser, at jo større værdi af μ, desto fladere er grafen af devianskurven.



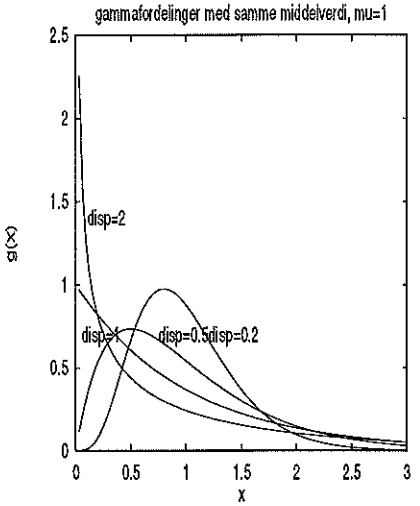
Vi kan udtrykke tætheden for en $\mathrm{G}(\alpha, \beta/\alpha)$ fordelt variabel på formen (1.3.20) som

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\ \exp\Big\{-\frac{1}{2\sigma^2}\Big(\frac{y}{\mu} - \ln\Big(\frac{y}{\mu}\Big) - 1\Big)\Big\}$$

med

$$a(y;\sigma^2) = \frac{\exp(-1/\sigma^2)}{(\sigma^2)^{1/\sigma^2}\Gamma(1/\sigma^2)}\ \frac{1}{y}$$

De vigtigste størrelser i fortolkningen af familien af $\mathrm{G}(\alpha, \beta/\alpha)$-fordelinger som en eksponentiel dispersionsmodel er anført i nedenstående oversigt:



Figur 4.3. Tætheden for gammafordelinger svarende til $\mu = 1$ for forskellige værdier af dispersionsparameteren σ^2

$G(\alpha, \beta/\alpha)$-fordelingen som reproduktiv eksponentiel dispersionsmodel				
Kanonisk parameter ϑ	Kumulant-frembringer $\kappa(\vartheta)$	Middelværdi-afb. $\mu = \tau(\vartheta)$	Enheds-variansfunktion $V_G(\mu)$	Dispersions-parameter σ^2
$-1/\beta$	$-\ln(-\vartheta)$	$-1/\vartheta$	μ^2	$1/\alpha$

Kvadratroden σ af dispersionsparameteren er variationskoefficienten

$$\sigma = \frac{\sqrt{\mathrm{V}[Y]}}{\mathrm{E}[Y]} = \frac{1}{\sqrt{\alpha}}$$

Formparameteren α i gamma-fordelingen udtrykker netop præcisionen $1/\sigma^2$ ved en fortolkning af familien af gammafordelinger som en eksponentiel dispersionsparameterfamilie.

Figur 4.3 viser tæthederne for gammafordelinger med samme middelværdi μ og forskellige værdier af dispersionsparameteren σ^2.

Bemærkning 1 ***Oversættelse mellem sædvanlig parametrisering og parametrisering som reproduktiv dispersionsmodel***

Da en række resultater vedrørende gammafordelingerne har en simpel fortolkning ved en repræsentation som en reproduktiv dispersionsmodel, kan det være nyttigt at kunne skifte mellem de to repræsentationer.

Vi har, at hvis Y følger en reproduktiv dispersionsmodel med middelværdi μ, variansfunktion $V_G(\mu) = \mu^2$ og dispersionsparameter $1/\sigma^2$, da er $Y \in \mathrm{G}(1/\sigma^2, \mu\sigma^2)$.

Omvendt, hvis $Y \in \mathrm{G}(\alpha, \beta/\alpha)$, da følger Y en reproduktiv dispersionsmodel med middelværdi $\mu = \mathsf{E}[Y] = \beta$, enhedsvariansfunktion $V(\mu) = \mu^2$ og dispersionsparameter $\sigma^2 = 1/\alpha$.

Hvis $Z \in \mathrm{G}(\alpha, \beta)$, da følger fordelingen af Z en additiv dispersionsmodel med indeksparameter α, middelværdi $\mathsf{E}[Z] = \alpha\mu$, enhedsvariansfunktion $V(\mu) = \mu^2$ og $\mathsf{V}[Z] = \alpha V(\mu)$. □

Bemærkning 2 ***Dispersionen ændres ikke ved en skalatransformation***

Hvis Y følger en gammafordeling med middelværdi μ og dispersionsparameter σ^2, da vil fordelingen af cY følge en gammafordeling med middelværdi $c\mu$ og med en uforandret dispersionsparameter, σ^2.

Dette følger af den sædvanlige skalatransformationsegenskab for gammafordelingen ved at bemærke, at $Y \in \mathrm{G}(1/\sigma^2, \mu\sigma^2)$. Resultatet udtrykker blandt andet, at formparameteren for gammafordelingen ikke ændres ved en skalatransformation. □

4.9.4 Reproduktivitetsegenskaber for Gammafordelingen

Vi minder om additionssætningen

Sætning 4.9.1 ***Additionssætningen for Gammafordelingen***

Såfremt $Z_1, Z_2, \ldots, Z_k$ er uafhængige variable, hvor $Z_i \in \mathrm{G}(\alpha_i, \beta)$ for $i = 1, 2, \ldots, k$, og vi sætter

$$Z_+ = X_1 + X_2 + \cdots + X_k,$$

da vil $Z_+ \in \mathrm{G}(\alpha_+, \beta)$, hvor $\alpha_+ = \alpha_1 + \alpha_2 + \cdots + \alpha_k$.
Såfremt addenderne følger gammafordelinger med samme skalaparameter, vil summen altså atter følge en gammafordeling, der fremkommer ved at addere formparametrene.

□

Specielt får man altså additionssætningen for $\mathrm{Ex}(\beta)$-fordelingen ved at sætte $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 1$. $\mathrm{G}(\alpha, \beta)$-fordelingen for α heltallig kan altsaa opfattes som fordelingen for en sum af α uafhængige $\mathrm{Ex}(\beta)$-fordelte størrelser. Man siger, at $\mathrm{Ex}(\beta)$-fordelingen er elementarfordeling for gammafordelingen.

Udtrykkes ovenstående sætning ved den reproduktive eksponentielle dispersionsmodel får man den generelle reproduktivitetsegenskab svarende til sætning 1.3.4 :

Bemærkning 1 ***Foldning (addition) af fordelinger med samme middelværdi***

Lad $Y_1, \ldots, Y_k$ være uafhængige, og lad Y_i følge en $\mathrm{G}(w_i/\sigma^2, \mu\sigma^2/w_i)$-fordeling og betragt

$$\overline{Y}_w = \frac{1}{w_{tot}} \sum_{i=1}^{k} w_i Y_i$$

med

$$w_{tot} = \sum_{i=1}^{k} w_i$$

Da vil fordelingen af $\overline{Y}_w$ være en $\mathrm{G}(w_{tot}/\sigma^2, \mu\sigma^2/w_{tot})$-fordeling. □

Det er netop dette resultat, der bruges, når man bestemmer et fælles estimat, s_+^2 for variansen i en normalfordeling på basis af en række uafhængige estimater, s_i^2.

4.9.5 Estimation i gammafordelingen

Sætning 4.9.2 ***Maksimum-likelihood estimatorer for uafhængige, identisk fordelte observationer***

Lad $Y_1, Y_2, \ldots, Y_k$ være uafhængige, og lad Y_i følge en $\mathrm{G}(1/\sigma^2, \mu\sigma^2)$-fordeling.

Da er maksimum-likelihood estimatoren μ

$$\widehat{\mu} = \overline{Y}$$

og maksimum-likelihood estimatoren for σ^2 fås som løsning til

$$\Psi(\frac{1}{\sigma^2}) - \ln\left(\frac{1}{\sigma^2}\right) = \ln\left(\overline{Y}/\overline{Y}_G\right) ,$$

hvor $\Psi(\cdot)$ angiver digammafunktionen (se side 152), $\overline{Y}$ angiver gennemsnittet af Y_i, $i = 1, \ldots, k$ og $\overline{Y}_G$ angiver det geometriske gennemsnit af Y'erne,

$$\overline{Y}_G = \Big(\prod_{i=1}^{k} Y_i\Big)^{1/k}$$

Bevis:
Se Cox og Lewis (1966). □

4.9.6 Gammafordelingen som miksturfordeling

Når Gammafordelingen optræder som en miksturfordeling i forbindelse med Poissonfordelingen er det ofte af interesse at benytte en parametrisering ved middelværdien, m og forholdet mellem variansen og middelværdien.

For $\mu \in \mathrm{G}(\alpha, m/\alpha)$ vil man betragte "signal/støj" forholdet ved Poissonfordelt støj,

$$\gamma(m, \alpha) \stackrel{\text{DEF}}{=} \frac{\mathsf{V}[\mu]}{\mathsf{E}[V_P(\mu)]} = \frac{m}{\alpha}\beta .$$

4.9.7 Gammafordelingen som levetidsfordeling

Sætning 4.9.3 ***Overlevelsesfunktion og hændelsesrate for $G(k, \beta)$-fordelingen***

For $T \in \mathrm{G}(k, \beta)$ gælder:

$$\overline{F}(t) = \frac{1}{\Gamma(k)} \int_{t/\beta}^{\infty} u^{k-1} \exp(-u) du \tag{4.9.5}$$

$$\lambda(t) = \frac{f(t)}{\overline{F}(t)} \tag{4.9.6}$$

(4.9.7)

$$\lambda(t) \to \frac{1}{\beta} \quad \text{for} \quad t \to \infty; \quad \lambda(t) \to \begin{cases} \infty & \text{for } 0 < k < 1 \\ 0 & \text{for } 1 < k \end{cases} \quad \text{for} \quad t \to 0;$$
$$\mathsf{E}[T] = k\beta; \qquad \mathsf{V}[T] = k\beta^2$$

For $0 < k < 1$ er $\lambda(t)$ monotont aftagende, og for $1 < k$ er $\lambda(t)$ monotont voksende i t.

Den Laplacetransformerede er $\psi_T(s) = (1 + s\beta)^{-k}$. □

Sætning 4.9.4 ***Mikstur af eksponentialfordelte hændelsesrater***

Lad den betingede fordeling af T givet $Y = y$ være en $\mathrm{Ex}(1/y)$-fordeling, og lad den stokastiske middellevetid Y have fordelingen $Y \in \mathrm{G}(\alpha, \beta)$. Da vil den marginale fordeling af T have hændelsesraten

$$\lambda(t) = \frac{\alpha\beta}{1 + \beta t}$$

Bevis:
Følger ved at opskrive den marginale tæthed,

$$f(t) = \frac{\alpha\beta}{(1 + \beta t)^{\alpha+1}} \qquad \text{for } 0 < t$$

□

Bemærkning:
den marginale fordeling af T er et specialtilfælde af den generaliserede Pareto-fordeling (se Davis og Feldstein 1979)

4.10 Log-gamma fordelingen

Definition 4.10.1 ***log-gamma fordelingen***

En stokastisk variabel W siges at følge en log-gamma fordeling med parameteren k, såfremt $T = \exp(W)$ er $\mathrm{G}(k, 1)$-fordelt. Symbolsk skriver vi:

$$W \in \mathrm{LG}(k) \Leftrightarrow \exp(W) \in \mathrm{G}(k, 1)$$

Såfremt $W \in \mathrm{LG}(k)$, er tætheden for W:

$$f(w) = \frac{1}{\Gamma(k)} \exp(-\mathrm{e}^w) \quad \text{for} -\infty < w < \infty$$

og den kumulerede fordelingsfunktion er

$$F(w) = \frac{1}{\Gamma(k)} \int_{-\infty}^{\exp(w)} u^{k-1} \exp(-u) du$$

□

Der gælder

$$\mathsf{E}[W] = \Psi(k), \quad \text{og} \quad \mathsf{V}[W] = \Psi'(k)$$

hvor $\Psi(\cdot)$ og $\Psi'(\cdot)$ angiver digammafunktionen og trigammafunktionen,

$$\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}; \quad \Psi'(x) = \frac{\Gamma''(x)\Gamma(x) - \{\Gamma'(x)\}^2}{\Gamma(x)^2}$$

Der gælder de asymptotiske relationer:

$$\Psi(x) = \ln(x) - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \cdots$$

$$\Psi'(x) = \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{62x^3} + \cdots$$

Bemærkning:
Betegnelsen log-gamma fordeling skyldes Bartlett og Kendall (1946). Betegnelsen hentyder til, at fordelingen beskriver fordelingen af logaritmen til en gammafordelt variabel. Betegnelsen er således ikke analog til log-normalfordelingen, der angiver fordelingen af en variabel, hvis logaritme er normalt fordelt. □

Sætning 4.10.1 ***log-normalfordelingen som grænsefordeling for log-gammafordelingen***

Lad $W \in \mathrm{LG}(k)$. Da gælder:

$$\lim_{k \to \infty} P[(W - \ln k)\sqrt{k} \leq u] = \Phi(u).$$

Bevis:
Se fx Lawless (1982).

□

Bemærkning:
Sætningen siger, at log-gammafordelingen asymptotisk svarer til en $\mathrm{LN}(\ln k, (1/\sqrt{k})^2)$-fordeling.

□

4.11 Den reciproke gammafordeling

Definition 4.11.1 ***Reciprok Gammafordeling***

En kontinuert stokastisk variabel X, der kan antage alle reelle ikke-negative værdier, siges at følge den reciproke gammafordeling med parametre α og β, hvis tætheden for X er af formen

$$g(x) = \frac{1}{\beta\Gamma(\alpha)} \left(\frac{\beta}{x}\right)^{\alpha+1} \exp(-\beta/x) \quad \text{for } 0 < x \tag{4.11.1}$$

hvor $0 < \alpha$ og $0 < \beta$.

Kort skriver vi $X \in \mathrm{RG}(\alpha, \beta)$

□

Bemærkning 1 ***Relation til gammafordelingen***

Såfremt $X \in \mathrm{G}(\alpha, 1/\beta)$, og vi sætter $Y = 1/X$, da vil $Y \in \mathrm{RG}(\alpha, \beta)$.

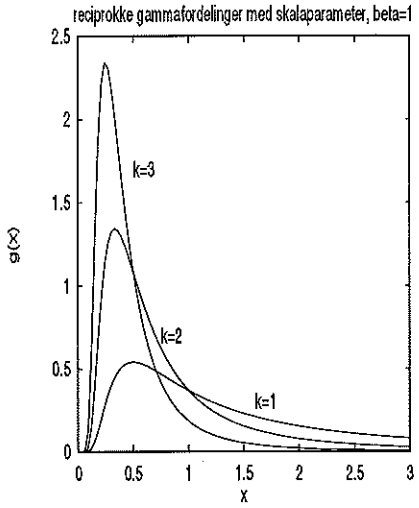
Den reciproke gammafordeling beskriver således fordelingen af den reciproke af en $\mathrm{G}(\alpha, 1/\beta)$-fordelt variabel.

□

Bemærkning 2 ***Skalatransformation i den reciproke gammafordeling***

Såfremt $X \in \mathrm{RG}(\alpha, \beta)$, og vi sætter $Y = \beta_1 X$ med $\beta_1 > 0$, da vil $Y \in \mathrm{RG}(\alpha, \beta_1\beta)$. Parameteren β er således en skalaparameter for fordelingen.

□



Figur 4.4. Tætheden for RG$(k,1)$-fordelingen svarende til forskellige værdier af formparameteren k

Såfremt $X \in \mathrm{RG}(\alpha,\beta)$, har vi

$$\mathsf{E}[X] = \frac{\beta}{\alpha-1}, \quad \mathsf{V}[X] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} = (\mathsf{E}[X])^2\,\frac{1}{\alpha-2},$$

hvor forventningsværdien existerer for $1 < \alpha$, og variansen existerer såfremt $2 < \alpha$.

Figur 4.4 viser tæthederne for reciprokke gammafordelinger med samme skalaparameter, $\beta = 1$, og forskellige værdier af formparameteren, k.

Parametriserer vi familien af RG(α,β)-fordelinger ved middelværdien finder vi, idet

$$\mu = \frac{\beta}{\alpha-1},$$

at variansen er

$$\mathsf{V}[X] = \mu^2 \frac{1}{\alpha - 2},$$

Når den variable $X \in \mathrm{RG}(\alpha, \beta)$ optræder som miksturfordeling er det ofte af interesse at benytte en parametrisering ved middelværdien m og "signal/støj forholdet" ved gammafordelt støj,

$$\gamma(\alpha) \stackrel{\text{DEF}}{=} \frac{\mathsf{V}[X]}{\mathsf{E}[X^2]} = \frac{1}{\alpha - 1},$$

Fordelingsfunktionen for $\mathrm{RG}(\alpha, \beta)$-fordelingen kan bestemmes ud fra fordelingsfunktionen for gammafordelingen ved

$$\mathsf{P}[\mathrm{RG}(\alpha, \beta) \leq x] = 1 - \mathsf{P}[\mathrm{G}(\alpha, 1/\beta) \leq 1/x] = 1 - \mathsf{P}[\mathrm{G}(\alpha, 1) \leq \beta/x]$$

Det første ufuldstændige moment for $\mathrm{RG}(\alpha, \beta)$-fordelingen er

$$\mu_1'(c) = \int_0^c x\, g(x)\, dx = \frac{\beta}{\alpha - 1} \mathsf{P}[\mathrm{RG}(\alpha - 1, \beta) \leq c]$$

og det tilsvarende centrale moment er

$$\begin{aligned} \mu_1(c) &= \int_0^c \{x - \frac{\beta}{\alpha - 1}\}\, g(x)\, dx \\ &= \frac{\beta}{\alpha - 1} \{\mathsf{P}[\mathrm{RG}(\alpha - 1, \beta) \leq c] - \mathsf{P}[\mathrm{RG}(\alpha, \beta) \leq c]\} \\ &= \frac{\beta}{\alpha - 1} \{\mathsf{P}[\mathrm{G}(\alpha, 1) \leq \beta/c] - \mathsf{P}[\mathrm{G}(\alpha - 1, 1) \leq \beta/c]\} \end{aligned}$$

der for heltallige værdier af α reduceres til

$$\mu_1(c) = -\frac{\beta}{\alpha - 1} \frac{(\beta/c)^{\alpha - 1}}{(\alpha - 1)!} \exp(-\beta/c)$$

4.12 Den generaliserede gammafordeling

Definition 4.12.1 *Den generaliserede gammafordeling*

En positiv stokastisk variabel T siges at følge en generaliseret gammafordeling med parametrene k, β og ν, hvis fordelingen af T har tætheden

$$f(t) = \frac{\nu}{\beta\Gamma(k)}(t/\beta)^{k\nu-1}\exp\{-(t/\beta)^{\nu}\} \quad \text{for} \quad t > 0$$

hvor $\beta > 0, k > 0, \nu > 0$.

Kort skrives $T \in \mathrm{GG}(\nu, k, \beta)$ □

Bemærkning 1 ***Den generaliserede gammafordeling som levetidsfordeling***

Generaliseringen er foreslået af Stacey (1962) med henblik på at formulere en klasse af fordelinger, der omfatter de mest almindelige levetidsfordelinger.

Fastsættes formparameteren $\nu = 1$ fremkommer således klassen af $\mathrm{G}(k, \beta)$-fordelinger, dvs specielt fås for $\nu = k = 1$, at $T \in \mathrm{Ex}(\beta)$. For $k = 1$ fås klassen af $\mathrm{We}(\nu, \beta)$-fordelinger. □

Der gælder endvidere

Sætning 4.12.1 ***Transformation af GG-fordelingen***

Lad $T \in \mathrm{GG}(\nu, k, \beta)$. Da vil $Y = (T/\beta)^{\nu}$ følge en $\mathrm{G}(k, 1)$-fordeling, dvs at

$$\mathsf{P}[T \geq t] = \overline{F}_Y((t/\beta)^{\nu}) = \mathsf{P}[\mathrm{G}(k,1) \geq (t/\beta)^{\nu}] = \frac{1}{\Gamma(k)}\int_{(t/\beta)^{\nu}}^{\infty} u^{k-1}\exp(-u)du$$

og $W = \nu\ln(T/\beta)$ følger en $\mathrm{LG}(k)$-fordeling.

□

Specielt gælder derfor for $k \to \infty$

$$\lim_{k\to\infty} P[(W - \ln k)\sqrt{k} \leq u] = \Phi(u)$$

dvs at $W = \nu\ln(T/\beta)$ asymptotisk er en normalfordeling, hvorfor den asymptotiske fordeling af T er en

$$\mathrm{LN}\Big(\frac{\nu\ln\beta + \ln k}{\nu}, \frac{1}{(\nu\sqrt{k})^2}\Big)\text{fordeling}\,.$$

Såfremt nu $k \to \infty, \nu \to 0$, og $\beta \to 0$ på en sådan måde, at $\nu\sqrt{k} \to 1/\gamma$ og $\beta k^{1/\nu} \to \alpha$, da vil $\mathrm{GG}(\nu, k, \beta)$-fordelingen nærme sig $\mathrm{LN}(\ln\alpha, \gamma^2)$-fordelingen.

Sætter vi specielt $k = 1/\lambda^2, \nu = \lambda$ og $\beta = \lambda^{2/\lambda}$ finder vi således, at $\mathrm{GG}(\lambda, 1/\lambda^2, \lambda^{2/\lambda})$-fordelingen svarer til $\mathrm{We}(1,1)$-fordelingen, såfremt $\lambda = 1$, og til $\mathrm{LN}(0,1)$-fordelingen, såfremt $\lambda = 0$.

4.13 Min_1 fordeling

Fordelingen kaldes ofte extremværdi eller dobbelteksponentiel fordeling

$$T \in \mathrm{Min}_1(\alpha,\beta) \Leftrightarrow F(t) = 1 - \Lambda_1\Big(-\frac{t-\alpha}{\beta}\Big)$$

med $\Lambda_1(u) = \exp(\exp(-u))$
For $T \in \mathrm{Min}_1(\alpha,\beta)$ gælder:

$$\overline{F}(t) = \Lambda_1(-\frac{t-\alpha}{\beta}) \tag{4.13.1}$$

$$f(t) = \frac{1}{\beta}\exp\left(\frac{t-\alpha}{\beta} - \exp\left(\frac{t-\alpha}{\beta}\right)\right) \tag{4.13.2}$$

$$\lambda(t) = \frac{1}{\beta}\exp\left(\frac{t-\alpha}{\beta}\right);\ \Lambda(t) = \exp\left(\frac{t-\alpha}{\beta}\right) \tag{4.13.3}$$

$$\mathsf{E}[T] = \alpha - \beta\gamma;\quad (\gamma = 0.522) \tag{4.13.4}$$

$$\mathsf{V}[T] = \frac{\pi^2}{6}\beta^2 \tag{4.13.5}$$

Bemærkning 1 ***Parameteren α er positionsparameter, og β er skalaparameter i $\mathrm{Min}_1(\alpha,\beta)$-fordelingen.*** □

Bemærkning 2 ***Gompertz-fordeling***

En positiv stokastisk variabel T siges at følge en Gompertz-fordeling med parametrene α_0 og β_0, såfremt

$$T \in \mathrm{Min}_1\Big(\frac{\ln(\beta_0) - \alpha_0}{\beta_0}, \frac{1}{\beta_0}\Big)\,.$$

Det ses, at hændelsesraten for Gompertz-fordelingen er $\lambda(t) = \exp(\alpha_0 + \beta_0 t)$. Gompertz-fordelingen er således en Min_1-fordeling, der er parametriseret sådan at logaritmen til hændelsesraten får en simpel form. □

4.13.1 Genesis:

Vi erindrer om sætningen

Sætning 4.13.1 ***Min_1-fordelingen som grænsefordeling***

Lad $T_1, T_2, \ldots, T_n$ være en følge af uafhængige identisk fordelte stokastiske variable med fordelingsfunktion $F(\cdot)$, og lad $T = \min(T_1, T_2, \ldots, T_n)$. Hvis F er af eksponentiel type, da vil for $n \to \infty$

$$\mathsf{P}\Big[\frac{T_n - \alpha_n}{\beta_n} \leq t\Big] \to \mathsf{P}[\mathrm{Min}_1(0,1) \leq t],$$

hvor α_n og β_n bestemmes ved

$$F(\alpha_n) = \frac{1}{n} \quad \beta_n = \frac{n}{nF'(\alpha_n)}$$

Bevis:
Se fx David (1970). □

Sætning 4.13.2 ***Reproduktivitetsegenskaber for Min_1-fordelingen***

Lad T_1 og T_2 være stokastisk uafhængige $\mathrm{Min}_1(\alpha, \beta)$ -fordelte variable, og lad $T = \min(T_1, T_2)$. Da gælder $T \in \mathrm{Min}_1(\alpha - \beta \ln 2, \beta)$.
Bevis:
Sætningen bevises direkte.

□

4.13.2 Max_1-fordelingen

Max_1-fordelingen indføres ved

$$T \in \mathrm{Max}_1(\alpha, \beta) \Leftrightarrow F(t) = \Lambda_1\Lambda_1\Big(\frac{t-\alpha}{\beta}\Big),$$

hvor $\Lambda_1(u) = \exp(\exp(-u))$.

Såfremt $T \in \mathrm{Max}_1(\alpha, \beta)$ vil $-T \in \mathrm{Min}_1(-\alpha, \beta)$. Egenskaberne ved Max_1-fordelingen følger da af egenskaberne for Min_1-fordelingen (afsnit 4.13).

4.14 Weibull-fordeling

Vi minder om definitionen på Weibull-fordelingen

$$T \in \mathrm{We}(k,\beta) \Leftrightarrow F(t) = 1 - \exp\big(-(t/\beta)^k\big)$$

For $T \in \mathrm{Min}_1(\alpha,\beta)$ gælder:

$$\overline{F}(t) = 1 - \exp(-(x/\beta)^k) \tag{4.14.1}$$

$$f(t) = \frac{k}{\beta}\left(\frac{t}{\beta}\right)^{k-1} \exp\left(-\left(\frac{t}{\beta}\right)^k\right) \quad \text{for} \quad t > 0 \tag{4.14.2}$$

$$\lambda(t) = (t/\beta)^{k-1} \quad \Lambda(t) = \frac{k}{\beta}(t/\beta)^k \tag{4.14.3}$$

$$\mathrm{E}[T] = \beta\,\Gamma(1+\frac{1}{k}) \tag{4.14.4}$$

$$\mathrm{V}[T] = \beta^2\Big[\Gamma\Big(1+\frac{2}{k}\Big) - \Gamma^2\Big(1+\frac{1}{k}\Big)\Big] \tag{4.14.5}$$

Bemærkning:
Vi noterer, at parameteren β er skalaparameter i $\mathrm{We}(k,\beta)$-fordelingen.

Sætning 4.14.1 ***Relation mellem We-fordeling og Min_1-fordeling***

Lad $T \in \mathrm{We}(k,\beta)$ og sæt $Y = \ln(T)$, da gælder $Y \in \mathrm{Min}_1(\ln\beta, 1/k)$.

Bevis:
Bevises direkte.

□

Bemærkning:
Ved logaritmetransformationen føres skalaparameteren β i fordelingen af T over i positionsparameteren $\ln\beta$ i fordelingen af $\ln(T)$, og formparameteren k bliver skalaparameter i fordelingen af $\ln(T)$.

4.14.1 Genesis:

Vi erindrer om sætningen

Sætning 4.14.2 ***We-fordelingen som grænsefordeling***

Lad $T_1, T_2, \ldots, T_n$ være en følge af uafhængige identisk fordelte stokastiske variable med fordelingsfunktion $F(\cdot)$ og lad $T_n = \min(T_1, T_2, \ldots, T_n)$. Hvis der eksisterer en positiv konstant k, således, at der for alle $a > 0$ gælder

$$\frac{F(ax)}{F(x)} \to a^k \quad \text{for} \quad x \to 0$$

da vil for $n \to \infty$,

$$P[T_n/\beta_n \leq t] \to P[\text{We}(k,1) \leq t]$$

hvor β_n bestemmes ved $F(\beta_n) = 1/n$.

Bevis:
Se fx Gnedenko (1943).

□

Der gælder følgende resultater vedrørende transformation af Weibull-fordelte størrelser:

Sætning 4.14.3 ***Transformation af We-fordelte størrelser***

Lad $T \in \text{We}(k, \beta)$. For $Y = \beta_1 T$, gælder $Y \in \text{We}(k, \beta\beta_1)$.

For $Z = T^{k_2}$ gælder $Z \in \text{We}(k/k_2, \beta^{k_2})$.

Specielt gælder for $V = (T/\beta)^k$, at $V \in \text{We}(1,1) = \text{Ex}(1)$

Bevis:
Sætningen bevises direkte. □

Bemærkning:
En Weibull-fordelt levetid kan således føres over i en eksponentialfordelt levetid ved en potenstransformation af tidsaksen. Ved denne transformation føres den aldersafhængige hændelsesrate over i en aldersuafhængig hændelsesrate.

4.15 Polyafordelingen

En diskret fordelt stokastisk variabel X, der kan antage værdierne $0, 1, 2, \ldots, n$, siges at følge en Polyafordeling, hvis X har frekvensfunktionen

$$g(x) = \binom{n}{x} \frac{\Gamma(\alpha + x)}{\Gamma(\alpha)} \frac{\Gamma(\beta - \alpha + n - x)}{\Gamma(\beta - \alpha)} \frac{\Gamma(\beta)}{\Gamma(\beta + n)} \quad \text{for} \quad x = 0, 1, 2, \ldots, n \tag{4.15.1}$$

hvor $\alpha > 0$ og $\beta > 0$ er reelle tal og hvor n er et positivt heltal.

Kort skriver vi $X \in \mathrm{PL}(n, \alpha, \beta)$

Polyafordelingen med heltallige værdier af α og β kan opfattes som en ventetidsfordeling ved stikprøveudtagning uden tilbagelægning. Antag nemlig at en population indeholder $\beta + n - 1$ elementer, hvoraf de $\beta - 1$ elementer er mærkede. Såfremt der udtages ét elemenent ad gangen fra populationen ved simpel tilfældig stikprøveudtagning uden tilbagelægning, da vil antallet X af umærkede elementer, der udtages, inden det α'te mærkede element udtages, kunne beskrives ved en $\mathrm{PL}(n, \alpha, \beta)$-fordelt stokastisk variabel. Vi har nemlig

$$\mathrm{P}[X = x] = \binom{\alpha + x - 1}{\alpha - 1} \binom{\beta + n - 1 - \alpha - x}{\beta - \alpha - 1} \Big/ \binom{\beta + n - 1x}{\beta - 1}$$

der omformes til frekvensfunktionen for en $\mathrm{PL}(n, \alpha, \beta)$ fordelt stokastisk variabel. For heltallige værdier af α og β er $\mathrm{PL}(n, \alpha, \beta)$-fordelingen således identisk med $\mathrm{NHyp}(\alpha, \beta + n - 1, \beta - 1)$-fordelingen.

Den anvendelse, vi hyppigst vil gøre af Polyafordelingen, er som en compound fordeling.

Der gælder

Sætning 4.15.1 ***Polya-fordeling som compound fordeling***

Lad X og p være stokastiske variable således at $p \in \mathrm{Be}(\alpha, \beta - \alpha)$ og den betingede fordeling $X|p \in \mathrm{B}(n, p)$. Da vil den marginale fordeling af X være $X \in \mathrm{PL}(n, \alpha, \beta)$

Bevis
Resultatet findes ved at betragte den marginale frekvensfunktion:

$$k(x) = \int_0^1 \binom{n}{x} \frac{\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta - \alpha)} p^{x+\alpha-1}(1-p)^{\beta-\alpha+n-x-1}\, dp$$

Polyafordelingen giver således af det samlede resultat ved n afhængige gentagelser, idet den betingede fordeling af X for givet p netop refererer til n uafhængige gentagelser af et Bernoulliekseriment. I den marginale fordeling af X bliver disse gentagelser bundet sammen af den omstændighed, at de alle er foretaget med den samme værdi af p, valgt blandt de mange mulige p-værdier.

Den model, der har givet navn til fordelingen er Polya's urnemodel.

Antag, at en urne indeholder β kugler, hvoraf α er røde og de øvrige $(\beta-\alpha)$ er hvide. En kugle trækkes tilfældigt fra urnen, hvorefter kuglen lægges tilbage, og yderligere c kugler i den pågældende farve lægges ned i urnen. Herefter trækkes en ny kugle fra urnen; kuglen og yderligere c kugler i den udtrukne farve lægges tilbage i urnen, og således fortsættes indtil man i alt har trukket n kugler fra urnen. Sandsynligheden for at der i disse n trækninger netop udtrækkes x røde kugler er da:

$$g(x) = \binom{n}{x} \times \frac{\alpha(\alpha+c)(\alpha+2c)\ldots(\alpha+(x-1)c)(\beta-\alpha+c)\ldots(\beta-\alpha+(n-x-1)c)}{\beta(\beta+c)(\beta+2c)\ldots(\beta+(n-1)c)},$$

der netop er tætheden for en $\mathrm{PL}(n,\alpha/c,\beta/c)$-fordelt stokastisk variabel. Lader vi $c \to 0$, svarende til uafhængige gentagelser, får vi tætheden for en $\mathrm{B}(n,\alpha/\beta)$-fordelt variabel, og sætter vi $c=-1$, svarende til udtrækning uden tilbagelægning, får vi tætheden for en $\mathrm{H}(n,\beta,\alpha)$-fordelt variabel.

Såfremt $X \in \mathrm{PL}(n,\alpha,\beta)$ og $Y = n - X$, vil $Y \in \mathrm{PL}(n,\beta-\alpha,\beta)$

Ved at betragte Polyafordelingen som en compound binomialfordeling finder vi ved brug af sætning 0.1.1, at

$$\mathrm{E}[X] = n\,\frac{\alpha}{\beta}, \quad \mathrm{V}[X] = n\,(n+\beta)\,\frac{\alpha(\beta-\alpha)}{\beta^2(\beta+1)} = (\mathrm{E}[X])^2\,\frac{1}{\alpha-2}$$

Vi vil betegne fordelingsfunktionen for $\mathrm{PL}(n,\alpha,\beta)$-fordelingen ved

$$\mathrm{PL}(c;n,\alpha,\beta) \overset{\mathrm{DEF}}{=} \sum_{x=0}^{c} \binom{n}{x} \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)} \frac{\Gamma(\beta-\alpha+n-x)}{\Gamma(\beta-\alpha)} \frac{\Gamma(\beta)}{\Gamma(\beta+n)} \tag{4.15.2}$$

Såfremt α og β er heltallige har man

$$\mathrm{PL}(c;n,\alpha,\beta) = \mathrm{H}(c;c+\alpha,\beta+n-1,n)$$

For vilkårlige værdier af α og β gælder

$$\mathrm{PL}(c;n,\alpha,\beta) = 1 - \mathrm{PL}(n-c-1;n,\beta-\alpha,\beta)$$

Til bestemmelse af fordelingsfunktionen ved summation af tæthederne kan man også benytte

$$g(0) = \frac{\Gamma(\beta - \alpha + n)}{\Gamma(\beta - \alpha)} \frac{\Gamma(\beta)}{\Gamma(\beta + n)}$$

og

$$g(i) = \frac{(n - i + 1)}{i} \frac{(\alpha + i - 1)}{(\beta - \alpha + n - i)} g(i - 1) \quad \text{for } i = 1, 2, \ldots, n$$

For $n \to \infty$ og $\beta \to \infty$ sådan at $\beta/(\beta + n) \to p$, vil

$$\mathrm{PL}(c; n, \alpha, \beta) \simeq \mathrm{NB}(c; n, \beta/(\beta + n)).$$

Approksimationen kan benyttes for store værdier af n og β såfremt $\beta/(\beta + n)$ ikke er for nær ved 0 eller 1.

For $n \to \infty$ og $\beta \to \infty$ sådan at $n/(\beta + n) \to 1$ og $n\alpha/\beta$ er moderat, kan man benytte en Poissonfordelingsapproksimation

$$\mathrm{PL}(c; n, \alpha, \beta) \simeq \mathrm{P}(c; n\alpha/\beta)$$

Det første ufuldstændige moment i $\mathrm{PL}(n, \alpha, \beta)$-fordelingen er

$$\mu_1'(c) = \sum_{x=0}^{c} x g(x) = n\frac{\alpha}{\beta} \mathrm{PL}(c - 1; n - 1, \alpha + 1, \beta + 1)$$

Relationen vises ved at bemærke, at

$$x(gx) = \frac{n\alpha}{\beta} \binom{n - 1}{x - 1} \frac{\Gamma(\alpha + x)}{\Gamma(\alpha + 1)} \frac{\Gamma(\beta - \alpha + n - x)}{\Gamma(\beta - \alpha)} \frac{\Gamma(\beta + 1)}{\Gamma(\beta + n)}$$

hvor højre side på nær faktoren $n\alpha/\beta$ er tætheden for en $\mathrm{PL}(n - 1, \alpha + 1, \beta + 1)$-fordelt variabel i punktet $x - 1$.

Det tilsvarende centrale moment er

$$\begin{aligned} \mu_1(c) &= \sum_{x=0}^{c} \left(x - n\frac{\alpha}{\beta} \right) g(x) \\ &= -\frac{(n - c)(\alpha + c)}{\beta} \binom{n}{c} \frac{\Gamma(\alpha + c)}{\Gamma(\alpha)} \frac{\Gamma(\beta - \alpha + n - c)}{\Gamma(\beta - \alpha)} \frac{\Gamma(\beta)}{\Gamma(\beta + n)} \end{aligned}$$

For at vise relationen bemærker vi først, at

$$x(\beta-\alpha+n-x)g(x)=(n-x+1)(\alpha+x-1)g(x-1)$$

Ved en simpel omskrivning heraf får man

$$(\beta-\alpha)\left(x-n\frac{\alpha}{\beta}\right)g(x)=-(\beta-\alpha)n\frac{\alpha}{\beta}\{g(x-1)-g(x)\}-x(n-x)g(x)$$
$$+(x-1)(n-x+1)g(x-1)-\alpha\left(x-1-n\frac{\alpha}{\beta}\right)g(x-1)$$

hvoraf vi finder ved summation fra 0 til c, at

$$\beta\mu_1(c)=-(n-c)(\alpha+c)g(c)$$

4.16 Den negative Polyafordeling

En diskret fordelt stokastisk variabel X, der kan antage værdierne $0,1,2,\ldots,n$, siges at følge en negativ Polyafordeling, hvis X har frekvensfunktionen

$$g(x) = \binom{r+x-1}{x}\frac{\Gamma(\alpha+x)}{\Gamma(\alpha)}\times \frac{\Gamma(\beta-\alpha+r)}{\Gamma(\beta-\alpha)}\frac{\Gamma(\beta)}{\Gamma(\beta+r+x)} \quad \text{for} \quad x=0,1,2,\ldots,n \quad (4.16.1)$$

hvor $\alpha>0$ og $\beta>0$ er reelle tal og r er et positivt heltal.

Kort skriver vi $X\in \mathrm{NPL}(n,\alpha,\beta)$

Såfremt p og X er stokastiske variable, således at $p\in \mathrm{Be}(\alpha,\beta-\alpha)$ og den betingede fordeling $X|p\in \mathrm{NB}(r,p)$, da vil den marginale fordeling af X være en negativ Polyafordeling, $X\in \mathrm{NPL}(n,\alpha,\beta)$

Ved at betragte den negative Polyafordeling som en compound negativ binomialfordeling finder vi ved anvendelse af sætning 0.1.1, at for $X\in \mathrm{NPL}(n,\alpha,\beta)$, vil

$$\mathrm{E}[X]=r\frac{\alpha}{\beta-\alpha-1},\quad \mathrm{V}[X]=r\frac{\alpha(\beta-1)(r+\beta-\alpha-1)}{(\beta-\alpha-1)^2(\beta-\alpha-2)}$$

Vi vil betegne fordelingsfunktionen for NPL(n, α, β)-fordelingen ved

$$\mathrm{NPL}(c; r, \alpha, \beta) \overset{\mathrm{DEF}}{=} \sum_{x=0}^{c} \binom{r+x-1}{x} \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)} \frac{\Gamma(\beta-\alpha+r)}{\Gamma(\beta-\alpha)} \frac{\Gamma(\beta)}{\Gamma(\beta+r+x)} \tag{4.16.2}$$

Der gælder

$$\mathrm{NPL}(c; r, \alpha, \beta) = \mathrm{PL}(c; c+r, \alpha, \beta) = 1 - \mathrm{PL}(r-1; c+r, \beta-\alpha, \beta)$$

Relationen vises ved at multiplicere relationen

$$\mathrm{NB}(c; r, p) = \mathrm{B}(c; c+r, 1-p) = 1 - \mathrm{B}(r-1; c+r, p)$$

med

$$\frac{1}{\mathrm{B}(\beta-\alpha, \alpha)} p^{\beta-\alpha}(1-p)^{\alpha}$$

og derefter integrere med hensyn til p, $0 \leq p \leq 1$.

Det første ufuldstændige moment i NPL(r, α, β)-fordelingen er

$$\mu_1'(c) = \sum_{x=0}^{c} x g(x) = r \frac{\alpha}{\beta-\alpha-1} NPL(c-1; r+1, \alpha+1, \beta)$$

Relationen vises ved at bemærke, at

$$x(gx) = \frac{r\alpha}{\beta-\alpha-1} \binom{r+x-1}{x-1} \frac{\Gamma(\alpha+x)}{\Gamma(\alpha+1)} \frac{\Gamma(\beta-\alpha+r)}{\Gamma(\beta-\alpha-1)} \frac{\Gamma(\beta)}{\Gamma(\beta+r+x)}$$

hvor højre side på nær faktoren $r\alpha/(\beta-\alpha-1)$ er tætheden for en NPL$(r+1, \alpha+1, \beta)$-fordelt variabel i punktet $x-1$.

Det tilsvarende centrale moment er

$$\begin{aligned} \mu_1(c) &= \sum_{x=0}^{c} \left(x - r\frac{\alpha}{\beta-\alpha-1} \right) g(x) \\ &= -\frac{(r+c)(\alpha+c)}{\beta-\alpha-1} \binom{r+c-1}{c} \frac{\Gamma(\alpha+c)}{\Gamma(\alpha)} \frac{\Gamma(\beta-\alpha+r)}{\Gamma(\beta-\alpha)} \frac{\Gamma(\beta)}{\Gamma(\beta+r+c)} \end{aligned}$$

Relationen vises ved at bemærke, at

$$x(\beta+r+x-1)g(x) = (r+x-1)(\alpha+x-1)g(x-1)$$

der omskrives til

$$(\beta-1)\left(x-r\frac{\alpha}{\beta-\alpha-1}\right)g(x) = -\alpha\left(x-1-r\frac{\alpha}{\beta-\alpha-1}\right)$$
$$= r\frac{\alpha(\beta-1)}{\beta-\alpha-1}\{g(x-1)-g(x)\}$$
$$+ (x-1)(r+x-1)g(x-1) - x(r+x)g(x)$$

Ved summation af den sidste ligning får vi

$$(\beta-\alpha-1)\mu_1(c) = -(r+c)(\alpha+c)g(c)$$

4.17 Betafordelingen

En kontinuert fordelt stokastisk variabel X, der kan antage alle reele værdier i intervallet $]0,1[$, siges at være Betafordelt med parametre (α,β), såfremt tætheden for X kan udtrykkes på formen

$$g(x) = \frac{1}{\mathrm{B}(\alpha,\beta)}\, x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0<x<1, \tag{4.17.1}$$

hvor $\alpha>0$ og $\beta>0$ og hvor $\mathrm{B}(\alpha,\beta)$ angiver betafunktionen,

$$\mathrm{B}(\alpha,\beta) \overset{\text{DEF}}{=} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,. \tag{4.17.2}$$

Kort skriver vi $X \in \mathrm{Be}(\alpha,\beta)$.

Figur 4.5 viser tæthederne for $\mathrm{Be}(\alpha,\beta)$ fordelinger svarende til forskellige værdier af α og β med $\alpha+\beta=10$.

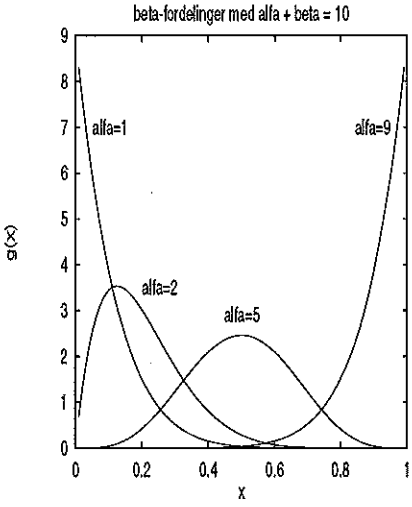
Vi bemærker, at klassen $\mathrm{Be}(\cdot,\cdot)$ udgør en eksponential familie af orden 2.

Såfremt $X \in \mathrm{Be}(\alpha,\beta)$, og $Y = 1-X$, da vil $Y \in \mathrm{Be}(\beta,\alpha)$.

Såfremt $X_1 \in \mathrm{G}(\alpha_1,\beta)$ og $X_2 \in \mathrm{G}(\alpha_2,\beta)$ er uafhængige, da gælder for

$$Y = \frac{X_1}{X_1+X_2}$$

at $Y \in \mathrm{Be}(\alpha_1,\alpha_2)$



Figur 4.5. Tætheden for Be(α, β)-fordelingen svarende til forskellige værdier af α og β

For $X \in \mathrm{Be}(\alpha, \beta)$ gælder

$$\mathsf{E}[X] = \frac{\alpha}{\alpha+\beta}; \qquad \mathsf{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4.17.3}$$

Vi bemærker i øvrigt, at

$$\mathsf{E}[X(1-X)] = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}$$

Vi vil ofte benytte en anden parametrisering af familien af beta-fordelinger.

For $X \in \mathrm{Be}(\alpha, \beta)$ sætter vi

$$\pi \overset{\text{DEF}}{=} \mathsf{E}[X] = \frac{\alpha}{\alpha+\beta} \tag{4.17.4}$$

$$\gamma \overset{\text{DEF}}{=} \frac{\mathsf{V}[X]}{\mathsf{E}[X(1-X)]} = \frac{1}{\alpha+\beta} \tag{4.17.5}$$

dvs

$$\alpha = \frac{\pi}{\gamma} \tag{4.17.6}$$

(4.17.7)

$$\beta = \frac{1-\pi}{\gamma} \tag{4.17.8}$$

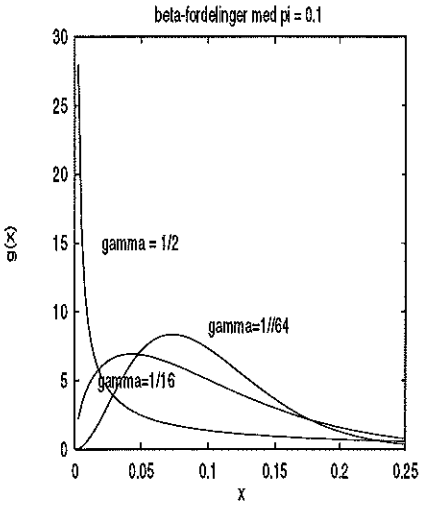
Udtrykt ved parametrene π og γ har vi:

$$\mathsf{E}[X] = \pi; \qquad \mathsf{V}[X] = \pi(1-\pi)\,\frac{\gamma}{\gamma+1}$$

Figur 4.6 viser tæthederne for $\mathrm{Be}(\alpha, \beta)$ fordelinger med samme middelværdi, $\pi = 0.1$ og forskellige værdier af γ.

Såfremt $X \in \mathrm{Be}(\alpha, \beta)$, og

$$Y = \frac{\beta X}{\alpha(1-X)},$$



Figur 4.6. Tætheden for Be(α, β)-fordelingen svarende til $\pi = 0.1$ og forskellige værdier af γ

da vil $Y \in \mathrm{F}(2\alpha, 2\beta)$, hvor $\mathrm{F}(f_1, f_2)$-fordelingen angiver den fra statistik 1 kendte F-fordeling med (f_1, f_2) frihedsgrader.

Vi vil betegne fordelingsfunktionen for $\mathrm{Be}(\alpha, \beta)$-fordelingen med

$$\mathrm{Be}(c; \alpha, \beta) = \int_0^c \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Der gælder

$$\mathrm{Be}(c; \alpha, \beta) = 1 - \mathrm{Be}(1-c; \beta, \alpha)$$

For store værdier af $\alpha + \beta$ kan man bruge en normalfordelingsapproksimation:

Det første ufuldstændige monemt for $\mathrm{Be}(\alpha, \beta)$-fordelingen er bestemt ved

$$\mu_1'(c) = \int_0^c x g(x) dx = \frac{\alpha}{\alpha+\beta} \mathrm{Be}(c; \alpha+1, \beta)$$

Det tilsvarende centrale moment er

$$\mu_1(c) = \int_0^c \left(x - \frac{\alpha}{\alpha+\beta}\right) g(x) dx = \frac{\alpha}{\alpha+\beta} \left[\mathrm{Be}(c; \alpha+1, \beta) - \mathrm{Be}(c; \alpha, \beta)\right]$$

der for heltallige værdier af α og β reduceres til

$$\mu_1(c) = -\frac{\alpha}{\alpha+\beta} \binom{\alpha+\beta-1}{\beta-1} c^\alpha (1-c)^\beta$$

4.18 Den flerdimensionale betafordeling

En k-dimensional stokastisk variabel X, der kan antage værdier i $]0,1[^k$ siges at følge en k-dimensional betafordeling, hvis X har tætheden:

$$g(x_1, x_2, \ldots, x_k) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \cdots x_k^{\alpha_k-1} \tag{4.18.1}$$

for

$$0 < x_1 < 1,\ 0 < x_2 < 1,\ \cdots, 0 < x_k < 1, \quad x_1 + x_2 + \cdots + x_k = 1$$

og $\alpha_1 > 0,\ \alpha_2 > 0,\ \cdots,\alpha_k > 0$.

Kort skriver vi $X \in \mathrm{Be}_k(\boldsymbol{\alpha})$.

Den flerdimensionale betafordeling kaldes også Dirichletfordelingen.

Såfremt $X \in \mathrm{Be}_k(\boldsymbol{\alpha})$ gælder

$$\mathsf{E}[X_i] = \frac{\alpha_i}{\alpha};\quad \mathsf{V}[X_i] = \frac{\alpha_i(\alpha-\alpha_i)}{\alpha^2(\alpha+1)};\quad \text{og}\quad \mathsf{COV}[X_i,X_j] = -\frac{\alpha_i\alpha_j}{\alpha^2(\alpha+1)} \tag{4.18.2}$$

hvor vi har sat $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_k$.

Den marginale fordeling af X_i er en $\mathrm{Be}(\alpha_i, \alpha-\alpha_i)$-fordeling.

Lad $X_1, X_2, \ldots, X_k$ være uafhængige stokastiske variable sådan at $X_i \in \mathrm{G}(\alpha_i,\beta)$ for $i = 1,2,\cdots,k$, og lad

$$Y_i = \frac{X_i}{X_1 + X_2 + \cdots + X_k};\quad i = 1,2,\cdots,k$$

Da er $Y = (Y_1, Y_2, \ldots, Y_k)^T \in \mathrm{Be}_k(\boldsymbol{\alpha})$, og Y og $(X_1 + X_2 + \cdots + X_k)$ er stokastisk uafhængige.

Lad $X = (X_1, X_2, \ldots, X_k)^T \in \mathrm{Be}_k(\boldsymbol{\alpha})$, og lad

$$Y_i = \frac{X_i}{X_1 + X_2 + \cdots + X_r};\quad i = 1,2,\cdots,r$$

med $2 \leq r \leq k$. Da er $Y \in \mathrm{Be}_r(\boldsymbol{\alpha}^*)$ med $\boldsymbol{\alpha}^* = (\alpha_1, \alpha_2, \ldots, \alpha_r)^T$, og Y og $(X_1 + X_2 + \cdots + X_r)$ er stokastisk uafhængige.

4.19 Den reciproke betafordeling

Definition 4.19.1 ***Reciprok betafordeling***

En kontinuert stokastisk variabel X, der kan antage alle reelle ikke-negative værdier, siges at følge den reciproke betafordeling med parametre μ, ν og β, hvis tætheden for X er af formen

$$g(x) = \frac{1}{\mathrm{B}(\mu,\nu)}\,\frac{\beta^\mu x^{\nu-1}}{(\beta+x)^{\mu+\nu}} \quad \text{for } 0 < x \tag{4.19.1}$$

hvor $0 < \mu$, $0 < \nu$ og $0 < \beta$.

Kort skriver vi $X \in \mathrm{RBe}(\mu, \nu, \beta)$.

□

Såfremt $X \in \mathrm{RBe}(\mu, \nu, \beta)$, og vi sætter $Y = \alpha X$ med $0 < \alpha$, gælder, at $Y \in \mathrm{RBe}(\mu, \nu, \beta)$. Parameteren β er altså en skalaparameter for fordelingsklassen.

Sætning 4.19.1 ***RBe*(μ, ν, β) *fordelingen som resultat af odds-transformation***

Såfremt $X \in \mathrm{Be}(\mu, \nu)$, og vi sætter

$$Y = \beta \frac{1 - X}{X}$$

da vil $Y \in \mathrm{RBe}(\mu, \nu, \beta)$.

□

Såfremt $X \in \mathrm{RBe}(\mu, \nu, \beta)$ og man sætter $Y = 1/X$, da vil $Y \in \mathrm{RBe}(\nu, \mu, 1/\beta)$.

Bemærkning 1 ***Sammenhæng med F-fordelingen***

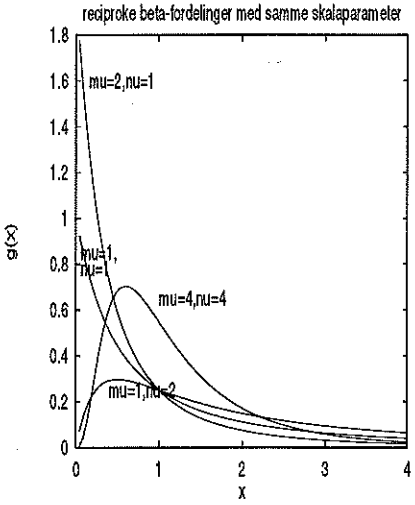
$\mathrm{RBe}(\mu, \nu, \nu/\mu)$-fordelingen er den fra Introduktion til Statistik, Bind 1kendte F-fordeling med $(2\mu, 2\nu)$ frihedsgrader. □

Sætning 4.19.2 ***Momenter for RB*(μ, ν, β) *fordeling***

Såfremt $X \in \mathrm{RBe}(\mu, \nu, \beta)$, gælder for $2 < \mu$:

$$\mathsf{E}[X] = \frac{\nu}{\mu - 1}\,\beta; \qquad \mathsf{V}[X] = \frac{\nu(\nu + \mu - 1)}{(\mu - 1)^2(\mu - 2)}\,\beta^2 \tag{4.19.2}$$

Figur 4.7 viser tæthederne for $\mathrm{RBe}(\mu, \nu, \beta)$ fordelinger med samme skalaparameter, $\beta = 1$.



Figur 4.7. Tætheden for RBe$(\mu, \nu, 1)$-fordelingen svarende til forskellige værdier af μ og ν

4.19.1 Genesis

Sætning 4.19.3 *Den reciproke betafordeling som resultat af en mikstur*

Såfremt det for to stokastiske variable T og X gælder, at $T|X = x \in \mathrm{G}(k, x)$ og at den ubetingede fordeling af X er $X \in \mathrm{RG}(\mu, \beta)$, da vil den marginale fordeling af T være en $\mathrm{RBe}(\mu, k, \beta)$-fordeling.

Bevis:

Følger umiddelbart ved opskrivning af den marginale tæthed. □

4.19.2 Approksimationer

Såfremt $\mu \to \infty$ og $\beta \to \infty$ på en sådan måde, at

$$\frac{\beta}{\mu - 1} \to m > 0$$

vil $\mathrm{RBe}(\mu, k, \beta)$-fordelingen konvergere mod en $\mathrm{G}(k, m)$-fordeling.

Ved en fortolkning af $\mathrm{RBe}(\mu, k, \beta)$-fordelingen som en miksturfordeling af $T|X = x \in \mathrm{G}(k, x)$ og $X \in \mathrm{RG}(\mu, \beta)$, svarer dette resultat til, at fordelingen af X konvergerer mod étpunktsfordelingen i $m = \beta/(\mu - 1)$, hvorfor den resulterende fordeling er en $\mathrm{G}(k, m)$-fordeling.

Figur 4.8 viser tæthederne for $\mathrm{RBe}(\mu, k, \beta)$-fordelinger med fastholdt $k = 2$ og forskellige værdier af μ og β sådan at $\beta/(\mu - 1) = 1$. Figuren viser yderligere den resulterende grænsefordeling, $\mathrm{G}(k, 1)$-fordelingen, som fremkommer for $\mu \to \infty$.

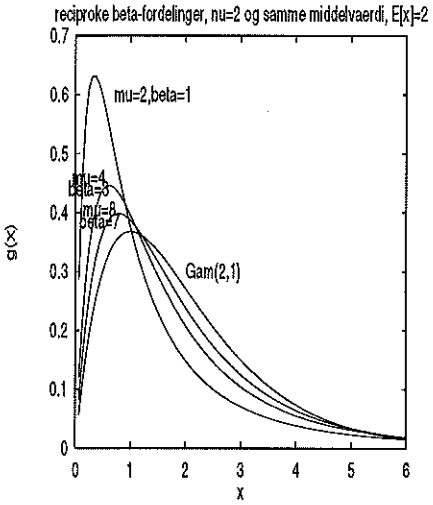
Såfremt $k \to \infty$ og $\beta \to 0$ på en sådan måde, at

$$\beta k \to b > 0$$

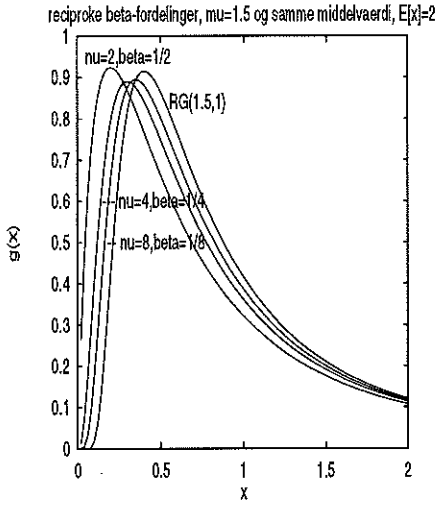
vil $\mathrm{RBe}(\mu, k, \beta)$-fordelingen konvergere mod en $\mathrm{RG}(\mu, b)$-fordeling.

Ved en fortolkning af $\mathrm{RBe}(\mu, k, \beta)$-fordelingen som en miksturfordeling af $T|X = x \in \mathrm{G}(k, x)$ og $X \in \mathrm{RG}(\mu, \beta)$, svarer dette resultat til, at fordelingen af T/k (en $\mathrm{RBe}(\mu, k, \beta/k)$-fordeling) konvergerer mod fordelingen af X, hvorfor den resulterende fordeling af T/k er en $\mathrm{RG}(\mu, b)$-fordeling.

Figur 4.9 viser tæthederne for $\mathrm{RBe}(\mu, k, \beta)$-fordelinger med fastholdt $\mu = 2$ og forskellige værdier af k og β sådan at $k\beta = 1$. Figuren viser yderligere



Figur 4.8. Tætheden for RBe(μ, ν, β)-fordelingen for $\nu = 2$ og forskellige værdier af μ og β sådan at $\beta/(\mu - 1) = 1$.



Figur 4.9. Tætheden for RBe(μ, ν, β)-fordelingen for $\mu = 2$ og forskellige værdier af ν og β sådan at $\beta * \nu = 1$.

den resulterende grænsefordeling, RG$(\mu, 1)$-fordelingen, som fremkommer for $k \to \infty$.

Fordelingsfunktionen for RBe(μ, ν, β)-fordelingen betegnes med

$$\mathrm{RBe}(c; \mu, \nu, \beta) \stackrel{\mathrm{DEF}}{=} \int_0^c \frac{1}{\mathrm{B}(\mu, \nu)} \frac{\beta^\mu x^{\nu-1}}{(\beta + x)^{\mu+\nu}}\, dx$$

Da β er en skalaparameter for fordelingen, har man

$$\mathrm{RBe}(c; \mu, \nu, \beta) = \mathrm{RBe}(c/\beta; \mu, \nu, 1)$$

og endvidere gælder som følge af sætning 4.19.1, at

$$\mathrm{RBe}(c; \mu, \nu, \beta) = 1 - Be(\beta/(c + \beta); \mu, \nu) = \mathrm{Be}(c/(c + \beta); \nu, \mu)$$

For store værdier af μ, kan man benytte normalfordelingsapproksimationen

$$\mathrm{RBe}(c; \mu, \nu, \beta) \approx \Phi(u),$$

med

$$u = \frac{c - \mathrm{E}[X]}{\sqrt{\mathrm{V}[X]}} = \left[\frac{c}{\beta}(\mu - 1) - \nu\right] \sqrt{\frac{\mu - 2}{\nu(\nu + \mu - 1)}}$$

Det første ufuldstændige moment er

$$\mu_1'(c) = \int_0^c x g(x) dx = \frac{\nu\beta}{\mu - 1} \mathrm{RBe}(c; \mu - 1, \nu + 1, \beta)$$

og det tilsvarende centrale moment er

$$\begin{aligned} \mu_1(c) &= \int_0^c \left(x - \frac{\nu\beta}{\mu - 1}\right) g(x) dx \\ &= \frac{\nu\beta}{\mu - 1}[\,\mathrm{RBe}(c; \mu - 1, \nu + 1, \beta) - \mathrm{RBe}(c; \mu, \nu, \beta)], \end{aligned}$$

der for heltallige værdier af μ og ν reduceres til

$$\mu_1(c) = -\frac{\nu\beta}{\mu - 1} \binom{\mu + \nu - 1}{\nu} \frac{\beta^{\mu-1} c^\nu}{(c + \beta)^{\mu+\nu-1}}$$

4.20 Binomialfordelingen

En diskret fordelt stokastisk variabel X, der kan antage værdierne $0, 1, 2, \ldots, n$, siges at følge en binomialfordeling med parametrene (n, p), hvis X har frekvensfunktionen

$$g(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{4.20.1}$$

Kort skriver vi $X \in \mathrm{B}(n, p)$.

Binomialfordelingen med $n = 1$ er en *Bernoullifordeling* med frekvensfunktionen

$$g(x) = p^x (1-p)^{1-x} = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$$

Hvis $X \in \mathrm{B}(n, p)$ og $Y = n - X$, gælder $Y \in \mathrm{B}(n, 1-p)$.

Den karakteristiske funktion for $\mathrm{B}(n, p)$-fordelingen er

$$\phi(t) = [1 - p + p\exp(it)]^n$$

Såfremt $X \in \mathrm{B}(n, p)$ gælder

$$\mathrm{E}[X] = np\,; \quad \mathrm{V}[X] = np(1-p)$$

4.20.1 Fordelingsfunktion og ufuldstændige momenter

Fordelingsfunktionen for $\mathrm{B}(n, p)$-fordelingen betegnes med

$$\mathrm{B}(c; n, p) \overset{\mathrm{DEF}}{=} \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

Ved delvis integration finder man

$$\mathrm{B}(c; n, p) = 1 - \mathrm{Be}(p; c+1, n-c) = \mathrm{Be}(1-p; n-c, c+1)$$

hvor $\mathrm{Be}(\cdot;\alpha,\beta)$ angiver fordelingsfunktionen for Betafordelingen, jvf afsnit 4.17.

Endvidere finder vi

$$\mathrm{B}(c;n,p) = 1 - \mathrm{B}(n-c-1;n,1-p)$$

Det første ufuldstændige moment for binomialfordelingen fås som

$$\mu_1'(c) = \sum_{x=0}^{c} xg(x) = np\mathrm{B}(c-1;n-1,p) \ ,$$

og det tilsvarende centrale moment er

$$\mu_1(c) = \sum_{x=0}^{c}(x-np)g(x) = -\ \frac{c+1}{n+1}\ \binom{n+1}{c+1}\ p^{c+1}(1-p)^{n-c} \ .$$

Den sidste relation fås ved at bemærke, at

$$x(1-p)g(x) = (n-x+1)pg(x) \ .$$

Summerer man denne relation for $x = 0,1,\ldots,c$, finder man, idet man benytter, at $x(1-p) = x - np + (n-x)p$

$$\begin{aligned}\sum_{x=0}^{c}(x-np)g(x) &= \sum_{x=0}^{c}[n-(x-1)]pg(x-1) - \sum_{x=0}^{c}(n-x)pg(x) \\ &= -(n-c)pg(c) \ .\end{aligned}$$

4.20.2 Binomialfordelingen som eksponentiel dispersionsmodel

For ethvert fast $n \in \mathbb{N}$ er familien af Binomialfordelinger, $\{\mathrm{B}(n,p)\}_{p\in]0,1[}$ en naturlig eksponentiel familie med den kanoniske parameter $\vartheta = \ln\{p/(1-p)\}$, kanonisk parameterområde $D = \mathbb{R}$ og med kumulantfrembringer $\kappa(\vartheta) = n\ln(1+\exp(\vartheta))$.

Lad Z være Binomialfordelt, $Z \in \mathrm{B}(n,p)$. Vi kan da udtrykke frekvensfunktionen (4.20.1) som

$$g(z;p,n) = \binom{n}{z}\ (1-p)^n[p/(1-p)]^z \quad \text{for} \quad z \in \{0,1,\ldots,n\} \tag{4.20.2}$$

Sætter vi

$$\vartheta = \ln\{p/(1-p)\} \tag{4.20.3}$$

og

$$\kappa(\vartheta) = \ln(1+\exp(\vartheta))$$

ser vi, at tætheden (4.20.2) netop er på formen (1.3.8) med ϑ og $\kappa(\cdot)$ som ovenfor, og med indeksparameteren λ lig med antalsparameteren n, dvs. $\Lambda = \mathbb{N}$.

Familien af $\mathrm{B}(n,p)$-fordelinger, $0 < p < 1$ og $n \in \mathbb{N}$ er således en additiv eksponentiel dispersionsmodel med kanonisk parameter $\vartheta = \ln(p/(1-p))$ enhedskumulantfrembringer $\kappa(\vartheta) = \ln(1+\exp(\vartheta))$.

Familien er frembragt, fx af en $\mathrm{B}(1,p)$-fordeling.

Vi har middelværdiafbildningen

$$\mu = \tau(\vartheta) = \frac{\exp(\vartheta)}{1+\exp(\vartheta)} = p$$

Parameteren p er altså netop middelværdien af en enhedsobservation.

Den kanoniske link er logitfunktionen

$$\vartheta = \tau^{-1}(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = \ln\left(\frac{p}{1-p}\right)$$

Enhedsvariansfunktionen er

$$V_{Bin}(\mu) = \tau'(\tau^{-1}(\mu)) = \mu(1-\mu)\ ,$$

der jo netop er variansen for en $\mathrm{B}(1,\mu)$-fordeling.

De vigtigste størrelser i fortolkningen af familien af $\mathrm{B}(n,p)$-fordelinger som en eksponentiel dispersionsmodel er anført i nedenstående oversigt:

$\mathrm{B}(n,p)$-fordelingen som additiv eksponentiel dispersionsmodel				
Kanonisk parameter ϑ	Kumulant-frembringer $\kappa(\vartheta)$	Middelværdi-afb. $\mu = \tau(\vartheta)$	Enheds-variansfunktion $V_{Bin}(\mu)$	Indeks-parameter λ
$\ln\{p/(1-p)\}$	$\ln(1+\exp(\vartheta))$	$\exp(\vartheta)/[1+\exp(\vartheta)]$	$\mu(1-\mu)$	n

Vi bemærker, at selv om Binomialfordelingen er veldefineret og har endelige momenter for $p = 0$ og $p = 1$, gælder fremstillingen som en eksponentiel dispersionsmodel kun for $0 < p < 1$, svarende til $\vartheta \in \mathbb{R}$. De udartede fordelinger svarende til værdierne $p = 0$ og $p = 1$ fremkommer ved en kompaktificering (afslutning) af $\mathbb{R}$ med punkterne $\vartheta = -\infty$ og $\vartheta = \infty$.

Enhedsdeviansen svarende til binomialfordelingen er

$$d(y;\mu) = 2\Big\{ y \ln\Big(\frac{y}{\mu}\Big) + (1-y)\ln\Big(\frac{1-y}{1-\mu}\Big)\Big\}\,, \tag{4.20.4}$$

hvor $y = z/n$.

Vi kan udtrykke tætheden for $\mathrm{B}(n,p)$-fordelingen ved enhedsdeviansen i overensstemmelse med (1.3.21) ved

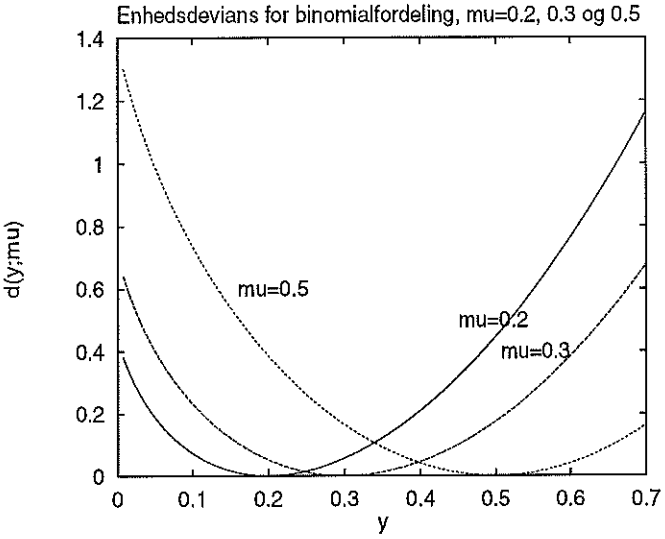
$$f(z;\xi,n) = a^*(z;n)\exp\Big\{ -\, n\Big\{ y\ln\Big(\frac{y}{\mu}\Big) + (1-y)\ln\Big(\frac{1-y}{1-\mu}\Big)\Big\}\Big\}, \tag{4.20.5}$$

hvor

$$a^*(z;n) = \binom{n}{z}$$

for $z = 0, 1, \ldots, n$ og hvor $y = z/n$ og $\xi = n\mu$

Nedenstående figur viser enhedsdeviansen svarende til $\mu = 0.2$, $\mu = 0.3$ og $\mu = 0.5$.



4.20.3 Reproduktivitetsegenskaber

Additionssætningen (sætning 1.3.5) udtrykkes her som:

Sætning 4.20.1 ***Additionssætningen for binomialfordelingen***

Lad $Z_1, Z_2, \ldots, Z_k$ væe indbyrdes uafhængige med $Z_i \in \mathrm{B}(n_i, p)$, og sæt

$$Z_+ = Z_1 + XZ2 + \cdots + Z_k \ .$$

Da gælder, at $Z_+ \in \mathrm{B}(n_1 + \cdots + n_k, p)$.

Fordelingen af en sum af binomialfordelte variable med samme sandsynlighedsparameter, p, er altså atter en binomialfordeling, der fremkommer ved at addere antalsparametrene, n_i. □

Definition 4.20.1 ***Binær addition og multiplikation for Bernoullivariable***

Lad $X_1 \in \mathrm{B}(1, p_1)$ og $X_2 \in \mathrm{B}(1, p_2)$ være uafhængige Bernoulli-fordelte variable, og betragt

$$Y = X_1 \oplus X_2$$

hvor symbolet $\oplus$ angiver den binære addition defineret ved

$Y = X_1 \oplus X_2$

	X_2	
X_1	0	1
0	0	1
1	1	1

da vil $Y \in \mathrm{B}(1, 1-(1-p_1)(1-p_2))$.

Sætter vi

$$Z = X_1 \otimes X_2$$

hvor symbolet $\otimes$ angiver den binære multiplikation defineret ved

$Y = X_1 \otimes X_2$

	X_2	
X_1	0	1
0	0	0
1	0	1

da vil $Z \in \mathrm{B}(1, p_1 p_2)$.

Vi bemærker, at der gælder

$$Z = 1 - (1 - X_1) \oplus (1 - X_2)$$

□

4.20.4 Approksimationer

For $n \to \infty$ og $p \to 0$, sådan at np er begrænset væk fra 0 og ∞ gælder approksimationen

$$\mathrm{B}(c; n, p) \approx \mathrm{P}(c; np),$$

hvor $\mathrm{P}(\cdot; \lambda)$ angiver Poissonfordelingens fordelingsfunktion. Approksimationen kan benyttes for store værdier af n og moderate værdier af np.

Approksimationen respekterer den skævhed, som er i binomialfordelingen for små værdier af p.

En intuitiv baggrund for approksimationen fås ved fx at betragte en væske med en bakterietæthed på p bakterier/ml. I stedet for at optælle det totale antal bakterier i n ml (Poissonfordelt), opdeles væsken i n delvolumener á 1 ml, og man optæller antallet af delvolumener med mindst én bakterie (binomialfordelt). Hvis delvolumenerne er så små, at der ikke er stor sandsynlighed for at finde mere en én bakterie i hvert delvolumen, vil de to opgørelsesformer stort set være ækvivalente. Hvis bakterietætheden er stor i forhold til størrelsen af delvolumenet, sådan at der ofte forekommer delvolumener med mere end én bakterie, vil de to opgørelsesformer give anledning til væsensforskellige resultater.

En nøjagtigere approksimation er $\mathrm{B}(c;n,p) \approx \mathrm{P}(c;\lambda)$ med

$$\lambda = \gamma - \frac{c(c+2+\gamma)}{6(2n-c)^2}$$

med $\gamma = -(n-c/2)\ln(1-p)$ (Molenaar, 1969).

Den centrale grænseværdisætning giver for $n \to \infty$ og p fast

$$\mathrm{B}(c;n,p) \approx \Phi\left(\frac{c-np}{\sqrt{np(1-p)}}\right) .$$

Approksimationen kan benyttes for store værdier af n og moderate værdier af $(c-np)/\sqrt{np(1-p)}$. Ofte benyttes argumentet $(c+1/2-np)/\sqrt{np(1-p)}$ for at kompensere for binomialfordelingens diskrete karakter. Approksimationen består i at erstatte binomialfordelingen med en normalfordeling med samme middelværdi og varians. Approksimationen er derfor mest velegnet for sådanne værdier af n og p, at binomialfordelingen med rimelighed kan anses for at være symmetrisk.

4.21 Multinomialfordelingen

En k-dimensional diskret fordelt variabel $X = (X_1, X_2, \ldots, X_k)^T$ siges at følge en multinomialfordeling med parametre $n, p_1, p_2, \ldots, p_k$, hvis X har frekvensfunktionen

$$g(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} \tag{4.21.1}$$

for

$$x_1 = 0, 1, \ldots, n;\ x_2 = 0, 1, \ldots, n; \ldots; x_k = 0, 1, \ldots, n; x_1 + x_2 + \cdots + x_k = n$$

hvor $0 \leq p_i \leq 1$ for $i = 1, 2, \ldots, k$ og $p_1 + p_2 + \cdots + p_k = 1$.

Kort skriver vi $X \in \text{Mult}_k(n, \mathbf{p})$.

Antag at et eksperiment kan resultere i k forskellige udfald $A_1, A_2, \ldots, A_k$, der alle gensidigt udelukker hinanden og tilsammen udtømmer alle muligheder. Antag, at vi foretager n uafhængige gentagelser af eksperimentet, sådan at $\mathsf{P}[A_j] = p_j$ for $j = 1, 2, \ldots, k$ i alle de n gentagelser. Lader vi nu X_j angive antallet af gange, udfaldet A_j forekommer i de n gentagelser, gælder for $X = (X_1, X_2, \ldots, X_k)^T$, at $X \in \text{Mult}_k(n, \mathbf{p})$.

Sætning 4.21.1 ***Additionssætningen for multinomialfordelingen***

Såfremt $X_1, X_2, \ldots, X_r$ er uafhængige, og $X_i \in \text{Mult}(n_i, \mathbf{p})$ og vi sætter

$$Y = X_1 + X_2 + \cdots + X_r\ ,$$

da vil $Y \in \text{Mult}_k(n, \mathbf{p})$ med $n = n_1 + n_2 + \cdots + n_r$ □

Sætning 4.21.2 ***Marginale fordelinger og momenter i multinomialfordelingen***

Såfremt $X = (X_1, X_2, \ldots, X_k)^T \in \text{Mult}_k(n, \mathbf{p})$, gælder

$$\mathsf{E}[X_i] = np_i\ ;\quad \mathsf{V}[X_i] = np_i(1 - p_i)\ ; \tag{4.21.2}$$

$$\mathsf{COV}[X_i X_j] = -np_i p_j \quad \text{for } i \neq j \tag{4.21.3}$$

Endvidere gælder, at den marginale fordeling af X_i er en $\text{B}(n, p_i)$-fordeling. □

Vi bemærker, at fordelingen af den k-dimensionale vektor X placerer hele massen i et k-dimensionalt underrum bestemt ved båndet $x_1 + x_2 + \cdots + x_k = n$. Vi kan derfor nøjes med at beskrive fordelingen af, fex de første $k - 1$

komponenter $X_1, X_2, \ldots, X_{k-1}$, ide X_k er bestemt ved $X_k = n - (X_1 + X_2 + \cdots + X_{k-1})$.

Familien af multinomialfordelinger, $\{\mathrm{Mult}_k(n, \mathbf{p})\}_{\mathbf{p} \in]0,1[^k}$ udgør en eksponentiel familie.

Vi vil betegne fordelingsfunktionen for $\mathrm{Mult}_k(n, \mathbf{p})$-fordelingen med

$$\mathrm{Mult}_k(\mathbf{c}; n, \mathbf{p}) \stackrel{\mathrm{DEF}}{=} \sum_{x_1=0}^{c_1} \cdots \sum_{x_k=0}^{c_k} g(x_1, x_2, \ldots, x_k)$$

for $c_k = n - (c_1 + c_2 + \cdots + c_{k-1})$. For $n \to \infty$ gælder

$$\mathrm{Mult}_k(\mathbf{c}; n, \mathbf{p}) \approx \mathrm{N}_k(\mathbf{c}; n\mathbf{p}, n\mathbf{V}) \,,$$

hvor $\mathbf{V}$ er den singulære matrix

$$\mathbf{V} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \ddots & -p_2p_k \\ \vdots & \vdots & \vdots & \vdots \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{pmatrix} .$$

4.21.1 Multinomialfordelingen som eksponentiel familie

Multinomialfordelingen er et eksempel på en flerdimensional eksponentiel familie, hvor målet ν er koncentreret på et affint underrum af $\mathbb{R}^m$.

Lad E være en endelig mængde med m elementer nummereret $1, 2, \ldots, m$, og lad μ være tællemålet på E, (dvs. $\mu\{A\}$ er antallet af elementer i A for $A \subset E$). For et sandsynlighedsmål P på E er sandsynlighedsfunktionen p, hvor

$$p(x) = P[\{x\}], \quad \text{for} \quad x \in \{1, 2, \ldots, m\}$$

netop tætheden af P med hensyn til μ.

Betragt familien

$$\{P \mid p(x) > 0 \quad \text{for alle} \quad x \in \{1, 2, \ldots, m\} \} \tag{4.21.4}$$

af sandsynlighedsmål, P, på E, der tillægger strengt positive sandsynligheder til ethvert $x \in \{1, 2, \ldots, m\}$.

Sætter vi nu

$$n_i(x) = \begin{cases} 1, & \text{for } x = i \\ 0, & \text{for } x \neq i \end{cases}$$

har vi

$$p(x) = \prod_{i=1}^{m} p(i)^{n_i(x)} = \exp\left\{ \sum_{i=1}^{m} (\ln p(i)) n_i(x) \right\}$$

Familien (4.21.4) er således en fuld eksponentiel familie med kanonisk stikprøvefunktion $t : E \to \mathbb{R}^m$

$$t(x) = \begin{pmatrix} n_1(x) \\ \vdots \\ n_m(x) \end{pmatrix} \in \mathbb{R}^m$$

og kanonisk parameter

$$\vartheta = \begin{pmatrix} \ln\, p(1) \\ \vdots \\ \ln\, p(m) \end{pmatrix} \in \mathbb{R}^m$$

For $\vartheta = (\vartheta_1, \vartheta_2, \ldots, \vartheta_m)^T \in \mathbb{R}^m$ har vi

$$\psi(\vartheta) = \sum_{j=1}^{m} \exp\left\{ \sum_{i=1}^{m} \vartheta_i n_i(j) \right\} = \sum_{j=1}^{m} \exp(\vartheta_j) < \infty$$

og

$$p(x) = \frac{\exp(\vartheta_x)}{\sum_{j=1}^{m} \exp(\vartheta_j)} \qquad \text{for } x \in \{1, 2, \ldots, m\} \tag{4.21.5}$$

Der gælder altså $D = \mathbb{R}^m$.

Støtten for $\nu = t(\mu)$ er

$$t(E) = \left\{ (t_1, t_2, \ldots, t_m)^T \in \mathbb{R}^m \mid \sum_{i=1}^{m} t_i = 1, t_i \in \{0, 1 \;\; i = 1, 2, \ldots, m \right\}$$

og den konvekse støtte er da

$$\text{konv st}(t(\mu)) = \left\{(t_1, t_2, \ldots, t_m)^T \in \mathbb{R}^m \mid \sum_{i=1}^{m} t_i = 1, t_i \geq 0 \; i = 1, 2, \ldots, m\right\}$$

med det relative indre af den konvekse støtte bestemt ved

$$\text{ri}(\text{ konv st}(t(\mu))) = \left\{(t_1, t_2, \ldots, t_m)^T \in \mathbb{R}^m \mid \sum_{i=1}^{m} t_i = 1, t_i > 0 \; i = 1, 2, \ldots, m\right\}$$

Den affine støtte er

$$A = \text{aff st}(t(\mu)) = \left\{(t_1, t_2, \ldots, t_m)^T \in \mathbb{R}^m \mid \sum_{i=1}^{m} t_i = 1\right\}$$

Den affine støtte er således begrænset til en hyperplan i $\mathbb{R}^m$. Parametrene ϑ_i er ikke identificerbare. Der gælder nemlig, at for ethvert $c \in \mathbb{R}$, vil parametrene

$$\vartheta = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \\ \vdots \\ \vartheta_m \end{pmatrix} \quad \text{og} \quad \vartheta + c = \begin{pmatrix} \vartheta_1 + c \\ \vartheta_2 + c \\ \vdots \\ \vartheta_m + c \end{pmatrix}$$

bestemme samme sandsynlighedsmål (4.21.5). Såfremt vi indfører et lineært bånd, som fx $\sum_{i=1}^{m} \vartheta_i = 0$, eller $\vartheta_1 = 0$ kan parametrene identificeres, og estimeres.

Middelværdiparameteren er

$$\tau(\vartheta) = \begin{pmatrix} \mathsf{E}_\vartheta[n_1(X)] \\ \vdots \\ \mathsf{E}_\vartheta[n_m(X)] \end{pmatrix} = \begin{pmatrix} p(1) \\ \vdots \\ p(m) \end{pmatrix} = \begin{pmatrix} \exp(\vartheta_1)/\sum_{j=1}^{m} \exp(\vartheta_j) \\ \vdots \\ \exp(\vartheta_m)/\sum_{j=1}^{m} \exp(\vartheta_j) \end{pmatrix} \tag{4.21.6}$$

og kovariansmatricen er givet ved at det (i, j)'te element er

$$V(\vartheta)_{ij} = \mathsf{COV}_\vartheta[n_i(X), n_j(X)] = \begin{cases} p(i)(1 - p(i)) & \text{for } i = j \\ -p(i)p(j) & \text{for } i \neq j \end{cases} \tag{4.21.7}$$

Kovariansmatricen er singulær på grund af overparametriseringen. Der gælder, at matricen $\mathbf{A}(\vartheta)$ givet ved

$$\mathbf{A}(\vartheta)_{ij} = \begin{cases} 1/p(i) & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \tag{4.21.8}$$

er en generaliseret invers til kovariansmatricen $V(\vartheta)$.

4.22 Den geometriske fordeling

En diskret fordelt stokastisk variabel X, der kan antage værdierne $0, 1, 2, \ldots$, siges at følge en geometrisk fordeling med parameteren p, hvis fordelingen af X har frekvensfunktionen

$$g(x) = p(1-p)^x \qquad \text{for } x = 0, 1, 2, \ldots \tag{4.22.1}$$

hvor $0 \leq p \leq 1$.

Kort skriver vi[1] $X \in \text{Geo}(p)$.

Såfremt $X \in \text{Geo}(p)$, er den karakteristiske funktion for fordelingen af X

$$\phi(t) = \frac{p}{1-(1-p)\exp(it)} .$$

Vi har

$$\mathsf{E}[X] = \frac{1-p}{p} ; \quad \mathsf{V}[X] = \frac{1-p}{p^2}$$

[1] Denne parametrisering af den geometriske fordeling svarer til den parametrisering, der er benyttet i Introduktion til Statistik, Bind 1.

4.22.1 Genesis

Såfremt $X_1, X_2, \ldots$ er en uendelig følge af uafhængige stokastiske variable sådan at $X_i \in \mathrm{B}(1,p)$ for $i = 1,2,\ldots$, og vi definerer den stokastiske variable T ved

$$T = \min\{i : X_i = 1\} ,$$

da vil $T \in \mathrm{Geo}(p)$.

Den geometriske fordeling angiver således fordelingen af antallet af gange, en hændelse $\neg A$ indtræffer før hændelsen A indtræffer for første gang i en ubegrænset serie af uafhængige Bernoulliforsøg, hvor hændelsen A indtræffer med sandsynligheden p i hvert forsøg. Den geometriske fordeling kan derfor benyttes til at beskrive ventetider, hvor parameteren p angiver sandsynligheden for den hændelse, man venter på.

Man finder den betingede sandsynlighed for at hændelsen indtræffer i det $c+1$'te forsøg, $[X = c]$, givet at den endnu ikke var indtruffet ved det c'te forsøg, $[X \geq c]$

$$\mathsf{P}[X = c | X \geq c] = p .$$

Resultatet udtrykker, at ventetidsprocessen ikke har nogen hukommelse.

4.22.2 Fordelingsfunktion, ufuldstændige momenter

Vi vil betegne fordelingsfunktionen for $\mathrm{Geo}(p)$-fordelingen med

$$\mathrm{Geo}(c;p) \stackrel{\mathrm{DEF}}{=} \sum_{x=0}^{c} p(1-p)^x$$

Der gælder

$$\mathrm{Geo}(c;p) = 1 - (1-p)^{c+1} .$$

4.22.3 Ufuldstændige momenter

Det første ufuldstændige moment i Geo(p)-fordelingen er bestemt ved

$$\mu_1'(c) = \sum_{x=0}^{c} xg(x) = \frac{1-p}{p}\left[1-(cp+1)(1-p)^c\right] .$$

Udtrykket verifceres let ved at bemærke, at $\mu_1'(c) = -p(1-p)f'(p)$,hvor

$$f(p) = \sum_{x=0}^{c}(1-p)^x = \frac{1}{p}[1-(1-p)^{c+1}] .$$

Det tilsvarende centrale moment er

$$\mu_1(c) = \sum_{x=0}^{c}\left(x - \frac{1-p}{p}\right) g(x) = -(c+1)p(1-p)^{c+1} .$$

4.22.4 Familien af geometriske fordelinger som en naturlig eksponentiel familie

Familien af geometriske fordelinger udgør en naturlig eksponentiel familie.

I forbindelse med behandling af de geometriske fordelinger som eksponentiel familie benytter man ofte en lidt anden parametrisering end den, der er benyttet i Introduktion til Statistik, Bind 1.

I disse sammenhænge siger man, at en diskret fordelt stokastisk variabel X, der kan antage værdierne $0, 1, 2, \ldots$, følger en geometrisk fordeling med parameteren p, hvis fordelingen af X har frekvensfunktionen

$$g^*(x) = p^x(1-p) \qquad \text{for } x = 0, 1, 2, \ldots \tag{4.22.2}$$

hvor $0 \leq p \leq 1$.

Vi bemærker, at i denne fremstilling er det størrelsen $1-p$, der angiver sandsynligheden for den hændelse, man venter på, og p angiver sandsynligheden for den hændelse, hvis forekomst, der tælles.

Vi vil benytte betegnelsen $X \in \text{Geo}^*(p)$ når frekvensfunktionen for fordelingen af X har formen (4.22.2). Det er klart, at de to repræsentationer fremstiller samme familie. Der gælder åbenbart, at

$$X \in \text{Geo}(1-p) \Longleftrightarrow X \in \text{Geo}^*(p) .$$

Familien af Geo*(p)-fordelinger udgør en naturlig eksponentiel familie, idet tætheden (4.22.2) kan udtrykkes på formen

$$g^*(x) = \exp\{\vartheta x + \ln\big(1 - \exp(\vartheta)\big) \qquad \text{for } x = 0, 1, 2, \ldots ,$$

hvor den kanoniske parameter $\vartheta = \ln(p)$, kumulantfrembringeren er $\kappa(\vartheta) = -\ln\Big(1 - \exp(\vartheta)\Big)$ og middelværdiafbildningen er

$$\tau(\vartheta) = \frac{\exp(\vartheta)}{1 - \exp(\vartheta)}$$

med middelværdirummet $\mathcal{M} = \mathbb{R}_+$. Det kanoniske parameterområde er $\Theta =]-\infty, 0[$. Den naturlige eksponentielle familie omfatter således ikke randvædierne $p = 0$ og $p = 1$.

Familien er stejl med den konvekse støtte $[0, \infty)$.

Ved differentiation finder man

$$\tau'(\vartheta) = \frac{\exp(\vartheta)}{\Big(1 - \exp(\vartheta)\Big)^2} ,$$

hvorfor man har, at variansfunktionen er

$$V(\mu) = \mu(1 + \mu)$$

Udtrykt ved sandsynligheden p for den hændelse, der tælles, har man

$$\mu = \mathsf{E}[X] = \frac{p}{1 - p} ; \quad \mathsf{V}[X] = V(\mu) = \frac{p}{(1 - p)^2}$$

4.22.5 Approksimationer

For $p \to 0$ gælder

$$\mathrm{Geo}(c; p) \approx 1 - \exp[-(c + 1)p] ,$$

der knytter forbindelsen mellem den diskrete og den kontinuerte ventetidsfordeling (eksponentialfordelingen).

4.23 Den negative binomialfordeling

En diskret fordelt stokastisk variabel X, der kan antage værdierne $0, 1, 2, \ldots$, siges at følge en negativ binomialfordeling med parametrene (r, p), hvis X har frekvensfunktionen

$$g(x) = \binom{r+x-1}{x} p^r (1-p)^x \qquad \text{for } x = 0, 1, 2, \ldots \tag{4.23.1}$$

hvor r er heltallig positiv, og $0 \leq p \leq 1$.

Kort skriver vi[1] $X \in \mathrm{NB}(r, p)$. For $r = 1$ fås den geometriske fordeling.

Den karakteristiske funktion for $\mathrm{NB}(r, p)$-fordelingen er

$$\phi(t) = \left(\frac{p}{1-(1-p)\exp(it)} \right)^r .$$

For $X \in \mathrm{NB}(r, p)$ gælder

$$\mathsf{E}[X] = r\,\frac{1-p}{p}\,; \quad \mathsf{V}[X] = r\,\frac{1-p}{p^2}$$

Såfremt $X \in \mathrm{NB}(r, p)$ og $Y = X + r$, siges Y at være Pascalfordelt med parametrene r og $1 - p$.

Ved konventionen

$$\binom{y}{x} \stackrel{\text{DEF}}{=} \frac{\Gamma(y+1)}{\Gamma(y+1-x)\,x\,!} = \frac{y(y-1)\cdots(y-x+1)}{x!}$$

for y reel, og x heltallig, ser vi, at vi kan skrive

$$\binom{r+x-1}{x} = \frac{\Gamma(r+x)}{\Gamma(r)x!} ,$$

hvilket antyder, hvorledes vi kan udvide definitionen af den negative binomialfordeling til reelle, positive værdier af r, se afsnit 4.23.2.

[1] Denne parametrisering af den negative binomialfordeling svarer til den parametrisering, der er benyttet i Introduktion til Statistik, Bind 1.

4.23.1 Fordelingsfunktion og ufuldstændige momenter

Vi betegner fordelingsfunktionen for NB(r,p)-fordelingen ved

$$\mathrm{NB}(c;r,p) \stackrel{\mathrm{DEF}}{=} \sum_{x=0}^{c} \binom{r+x-1}{x} p^r(1-p)^x \ .$$

Der gælder

$$\mathrm{NB}(c;r,p) = 1 - \mathrm{B}(r-1;c+r,p) = \mathrm{B}(c;c+r,1-p) \ ,$$

hvorfor vi også har

$$\mathrm{NB}(c;r,p) = \mathrm{Be}(p;r,c+1) \ .$$

Det første ufuldstændige moment for NB(r,p)-fordelingen fås af

$$\mu_1'(c) = \sum_{x=0}^{c} xg(x) = r\,\frac{1-p}{p}\,\mathrm{NB}(c-1;r+1,p) \ .$$

Relationen vises ved at bemærke, at

$$xg_r(x) = r\,\frac{1-p}{p}\binom{r+1+x-1-1}{x-1} p^{r+1}(1-p)^{x-1} \ ,$$

hvor højre side på nær faktoren $r(1-p)/p$ er tætheden for en NB$(r+1,p)$-fordelt variabel i punktet $x-1$.

Endvidere finder vi det ufuldstændige centrale moment

$$\mu_1(c) = \sum_{x=0}^{c}\left(x - r\,\frac{1-p}{p}\right) g(x) = -(c+1)\binom{r+c}{c+1} p^r(1-p)^{c+1} \ .$$

Relationen vises ved at bemærke, at

$$xg(x) = (1-p)(r+x-1)g(x-1)$$

hvorfor

$$\begin{aligned}\left(x - r\,\frac{1-p}{p}\right) g(x) \quad &= \quad -r\,\frac{1-p}{p}\,\{g(x) - g(x-1)\} \\ &\quad +(1-p)\left(x-1-r\,\frac{1-p}{p}\right) g(x-1) \ .\end{aligned}$$

Ved summation af denne relation finder man

$$p\mu_1(c) = -(r+c)(1-p)g(c) \ ,$$

hvoraf $\mu_1(c)$ bestemmes.

4.23.2 Den negative binomialfordeling som eksponentiel dispersionsmodel

I forbindelse med behandlingen af negative binomialfordelinger som en eksponentiel dispersionsmodel vil vi i analogi med afsnit 4.22.4 benytte en lidt anden parametrisering end den, der er benyttet i Introduktion til Statistik, Bind 1. For at sondre mellem de to parametriseringer benyttes betegnelsen NB^* til at betegne den additive dispersionsmodel frembragt af $\mathrm{Geo}^*(p)$-fordelingen.

En diskret fordelt stokastisk variabel, Z, der kan antage værdierne $0, 1, \ldots$, siges at følge en negativ binomialfordeling med parametrene α og p, hvis Z har frekvensfunktionen

$$g(z;\alpha,p) = \frac{\Gamma(\alpha+z)}{\Gamma(\alpha)\, z\,!}\,(1-p)^{\alpha}p^{z} \quad \text{for} \quad z \in \{0,1,\ldots,\}\,, \tag{4.23.2}$$

hvor $\alpha \in \mathbb{R}_+$ og $0 \leq p \leq 1$.

Vi bruger betegnelsen $Z \in \mathrm{NB}^*(\alpha,p)$, når frekvensfunktionen for fordelingen af Z har formen (4.23.2).

For heltallige vædier af α har vi

$$Z \in \mathrm{NB}(\alpha, 1-p) \Longleftrightarrow Z \in \mathrm{NB}^*(\alpha,p)$$

Bemærk, at i lighed med $\mathrm{Geo}^*(p)$-fordelingen er $\mathrm{NB}^*(\alpha,p)$-fordelingen parametriseret ved sandsynligheden p for den hændelse, hvis antal forekomster, man tæller.

Familien af $\mathrm{NB}^*(\alpha,p)$-fordelinger for $\alpha \in \mathbb{R}_+$ og $0 < p < 1$ er netop den additive eksponentielle dispersionsmodel frembragt af en $\mathrm{Geo}^*(p)$-fordeling. Den kanoniske parameter er $\vartheta = \ln(p)$, indeksmængde $\Lambda = \mathbb{R}_+$ og kanonisk parameterområde $\Theta =]-\infty, 0[$. Vi bemærker, at den eksponentielle dispersionsmodel ikke omfatter fordelingerne svarende til $p = 0$ og $p = 1$.

En $\mathrm{NB}^*(\alpha,p)$-fordeling er uendeligt delbar..

De vigtigste størrelser i fortolkningen af familien af $\mathrm{NB}^*(\alpha,p)$-fordelinger som en additiv eksponentiel dispersionsmodel er anført i nedenstående oversigt:

NB*(α, p)-fordelingen som additiv eksponentiel dispersionsmodel				
Kanonisk parameter ϑ	Kumulantfrembringer $\kappa(\vartheta)$	Middelværdiafb. $\mu = \tau(\vartheta)$	Enhedsvariansfunktion $V_{NB}(\mu)$	Indeksparameter λ
$\ln(p)$	$-\ln\big(1-\exp(\vartheta)\big)$	$\exp(\vartheta)/[1-\exp(\vartheta)]$	$\mu(1+\mu)$	α

Specielt finder man middelværdi og varians for NB*(α, p)-fordelingen

$$\mathsf{E}[Z] = \alpha\mu \; ; \quad \mathsf{V}[Z] = \alpha V_{NB}(\mu) = \alpha\mu(1+\mu) \; ,$$

hvor $\mu = p/(1-p)$.

Den kanoniske link for den negative binomialfordeling er

$$\vartheta = \tau^{-1}(\mu) = \ln\Big(\frac{\mu}{1+\mu}\Big)$$

Enhedsdeviansen svarende til NB*(α, p)-fordelingen er

$$d(y;\mu) = 2\Big\{y \ln\Big(\frac{y(1+\mu)}{(1+y)\mu}\Big) + \ln\Big(\frac{1+\mu}{1+y}\Big)\Big\} \; , \tag{4.23.3}$$

hvor $y = z/\alpha$.

Vi kan udtrykke tætheden for NB*(α, p)-fordelingen ved enhedsdeviansen i overensstemmelse med (1.3.21) ved

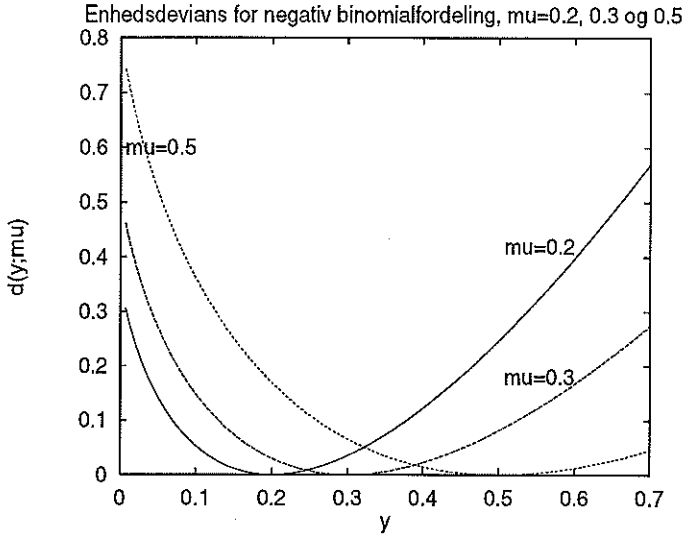
$$f(z;\xi,\alpha) = a^*(z;\alpha) \exp\Big[-\alpha\Big\{y \ln\Big(\frac{y(1+\mu)}{(1+y)\mu}\Big) + \ln\Big(\frac{1+\mu}{1+y}\Big)\Big\}\Big] \; , \tag{4.23.4}$$

hvor

$$a^*(z;\alpha) = \frac{\Gamma(\alpha+z)}{\Gamma(\alpha)z\,!}$$

for $z = 0, 1, \ldots, n$ og hvor $y = z/\alpha$ og $\xi = \alpha\mu$

Nedenstående figur viser enhedsdeviansen svarende til $\mu = 0.2$, $\mu = 0.3$ og $\mu = 0.5$.



Vi betegner fordelingsfunktionen for NB*(α, p)-fordelingen ved

$$\mathrm{NB}^*(c;\alpha,p) \stackrel{\mathrm{DEF}}{=} \sum_{z=0}^{c} \frac{\Gamma(\alpha+z)}{\Gamma(\alpha)\, z\,!}(1-p)^{\alpha}p^z \ .$$

Der gælder

$$\mathrm{NB}^*(c;\alpha,p) = 1 - \mathrm{Be}(p;\alpha,c+1) \ .$$

For heltallige værdier af α gælder desuden

$$\mathrm{NB}^*(c;\alpha,p) = \mathrm{B}(c;c+\alpha,p) \ .$$

4.23.3 Reproduktivitetsegenskaber

Additionssætningen (sætning 1.3.5) udtrykkes her som:

Sætning 4.23.1 ***Addition af negativ binomialt fordelte variable***

Hvis $Z_1, Z_2, \ldots, Z_k$ er uafhængige variable med $Z_i \in \mathrm{NB}^*(\alpha_i, p)$ og vi sætter

$$Z_+ = Z_1 + Z_2 + \cdots + Z_k \ ,$$

da er $Z_+ \in \mathrm{NB}^*((\alpha_1 + \alpha_2 + \cdots + \alpha_k, p)$. □

4.23.4 Approksimationer

For $p \to 0$ og $\alpha \to \infty$, sådan at $\alpha p \to \lambda$, hvor λ er begrænset væk fra 0 og ∞, gælder

$$\mathrm{NB}^*(c; \alpha, p) \approx \mathrm{P}(c; \alpha\mu) \ ,$$

hvor $\mu = p/(1-p)$. Denne approksimation bevarer skævheden i fordelingen.

For $\alpha \to \infty$ får vi ved den centrale grænseværdisætning:

$$\mathrm{NB}*(c; r, p) \approx \Phi(u)$$

med

$$u = \frac{(c + 1/2)(1-p) - \alpha p)}{\sqrt{\alpha p}} \ .$$

Approksimationen kan benyttes for moderate værdier af u.

4.24 Poissonfordelingen

En diskret fordelt stokastisk variabel X, der kan antage værdierne $0, 1, 2, \ldots,$, siges at følge en Poissonfordeling med parameteren λ, hvis X har frekvensfunktionen

$$g(x) = \frac{\lambda^x}{x!} \mathrm{e}^{-\lambda} \tag{4.24.1}$$

for $x = 0, 1, 2, \ldots$, hvor $0 < \lambda$.

Kort skriver vi $X \in \mathrm{P}(\lambda)$.

Den karakteristiske funktion for $\mathrm{P}(\lambda)$-fordelingen er

$$\phi(t) = \exp\left(\lambda \mathrm{e}^{\mathrm{i}t} - 1\right) \ .$$

Såfremt $X \in \mathrm{P}(\lambda)$ har vi

$$\mathsf{E}[X] = \lambda \; ; \quad \mathsf{V}[X] = \lambda \; .$$

Såfremt $Y_1, Y_2, \ldots$ er en uendelig følge af uafhængige stokastiske variable, hvor $Y_i \in \mathrm{Ex}(\beta)$, og vi sætter

$$X(t) = \max\{n : Y_1 + Y_2 \cdots + Y_n \leq t\} \; ,$$

da vil $X(t) \in \mathrm{P}(t/\beta)$.

Poissonfordelingen kan således beskrive fordelingen af antallet af hændelser, der indtræffer i et givet tidsrum, når de enkelte hændelser er "tilfældigt fordelt" i tid.

4.24.1 Fordelingsfunktion og ufuldstændige momenter

Fordelingsfunktionen for $\mathrm{P}(\lambda)$-fordelingen betegnes

$$\mathrm{P}(c;\lambda) \stackrel{\mathrm{DEF}}{=} \sum_{x=0}^{c} \frac{\lambda^x}{x!} \exp(-\lambda)$$

Ved ledvis integration af gammafordelingens fordelingsfunktion finder man

$$\mathrm{P}(c;\lambda) = 1 - \mathrm{G}(\lambda; c+1, 1) \; .$$

Det første ufuldstændige moment i $\mathrm{P}(\lambda)$-fordelingen er bestemt ved

$$\mu_1'(c) = \sum_{x=0}^{c} x g(x) = \lambda \, \mathrm{P}(c-1;\lambda)$$

for $c > 0$.

Det tilsvarende centrale moment er

$$\mu_1(c) = \sum_{x=0}^{c} (x-\lambda) g(x) = -\lambda \frac{\lambda^c}{c!} \mathrm{e}^{-\lambda}$$

4.24.2 Poissonfordelinger som eksponentiel dispersionsmodel

Lad Z være Poissonfordelt, $Z \in \mathrm{P}(\mu)$.

Familien af $\mathrm{P}(\mu)$-fordelinger for $\mu \in \mathbb{R}_+$ er en naturlig eksponentiel familie med kanonisk parameter $\vartheta = \ln(\mu)$, kanonisk parameterområde $\Theta = \mathbb{R}$, kumulantfrembringer $\kappa(\vartheta) = \exp(\vartheta)$ og middelværdiafbildning $\tau(\vartheta) = \exp(\vartheta)$.

Den additive eksponentielle dispersionsmodel frembragt af en $\mathrm{P}(\mu)$-fordeling er netop denne naturlige eksponentielle familie.

Den sædvanlige parametrisering er netop ved middelværdiparameteren.

En $\mathrm{P}(\mu)$-fordeling er uendeligt delbar.

De vigtigste størrelser i fortolkningen af familien af $\mathrm{P}(\mu)$-fordelinger som en eksponentiel dispersionsmodel er anført i nedenstående oversigt:

$\mathrm{P}(\mu)$-fordelingen som additiv eksponentiel dispersionsmodel				
Kanonisk parameter ϑ	Kumulant-frembringer $\kappa(\vartheta)$	Middel-værdi-afb. $\mu = \tau(\vartheta)$	Enheds-varians-funktion $V_P(\mu)$	Indeks-parameter λ
$\ln(\mu)$	$\exp(\vartheta)$	$\exp(\vartheta)$	μ	*

* Indeksparameteren og middelværdiparameteren kan ikke umiddelbart adskilles

Den kanoniske link for Poissonfordelingen er logaritmefunktionen,

$$\vartheta = \tau^{-1}(\mu) = \ln(\mu)$$

Enhedsdeviansen svarende til Poissonfordelingen er

$$d(y;\mu) = 2\left\{y \ln\left(\frac{y}{\mu}\right) - (y-\mu)\right\}, \tag{4.24.2}$$

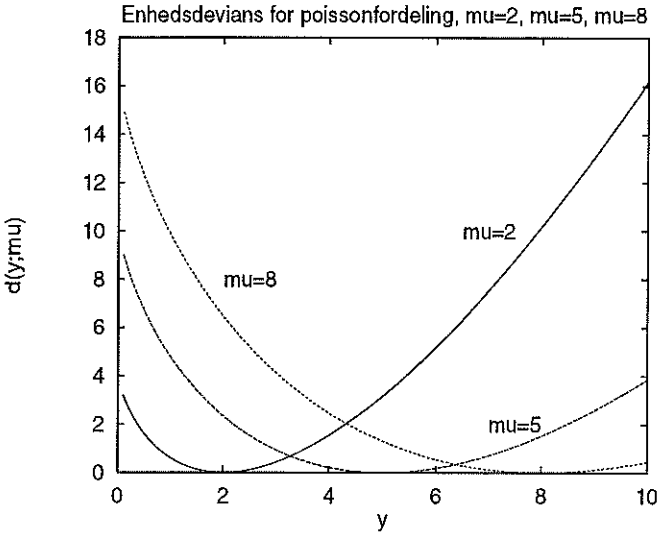
Vi kan udtrykke tætheden for $\mathrm{P}(\mu)$-fordelingen ved enhedsdeviansen i overensstemmelse med (1.3.21) ved

$$f^*(z;\mu) = a^*(z)\exp\left[\left\{y\ln\left(\frac{y}{\mu}\right) - (y-\mu)\right\}\right], \tag{4.24.3}$$

hvor

$$a^*(z) = \frac{1}{z\,!}$$

Nedenstående figur viser enhedsdeviansen svarende til $\mu = 2$, $\mu = 5$ og $\mu = 8$.



Vi bemærker, at der ikke indgår en indeksparameter i familien. Familien af Poissonfordelinger er den eneste eksponentielle dispersionsmodel med denne egenskab. En forklaring på dette fænomen kan findes i at variansfunktionen $V(\mu)$ for Poissonfordelingen er identiteten.

Lad nemlig $Z_1, Z_2, \ldots, Z_k$ være uafhængige med $Z_i \in \mathrm{P}(\mu)$. Da vil summen $Z_+ = Z_1 + \cdots + Z_k$ følge en $\mathrm{P}(k\mu)$-fordeling med forventningsværdi, $\xi = k\mu$ og varians, $\mathrm{V}[Z_+] = \xi$. Funktionalligningen

$$\mathrm{V}[Z] = kV(\xi/k)$$

svarende til (1.3.12) bliver blot

$$\xi = k\xi/k \ ,$$

der trivielt er opfyldt for enhver faktorisering, $\xi = k\mu$.

Bemærkning 1 ***Parametrisering af Poissonfordelinger med middelværdier proportionale med kendte størrelser, offset af kanonisk parameter***

Undertiden (fx ved analyse af generaliserede lineære modeller) har man behov for at parametrisere Poissonfordelingen sådan at middelværdien er proportional med en kendt proportionalitetsfaktor, w. Da indeksparameteren er ubestemt, kan man ikke opnå dette ved en sædvanlig vægtning. Det kan imidlertid opnå ved en nulpunktsforskydning, (offset) af den kanoniske parameter.

Lad $Z \in \mathrm{P}(w\mu)$. Vi kan da udtrykke frekvensfunktionen for fordelingen af Z som

$$f^*(z;\vartheta,w) = \frac{w^z}{z\,!}\exp[\vartheta z - w\exp(\vartheta)] \tag{4.24.4}$$

Ved skift af nulpunkt for ϑ til $\vartheta' = \vartheta + \ln(w)$ bliver udtrykket for tætheden

$$f^*(z;\vartheta',w) = \frac{1}{z\,!}\exp[\vartheta' z - \exp(\vartheta')] \ .$$

Dette nulpunktsskift for den kanoniske parameter kaldes også en offset af den kanoniske parameter. □

4.24.3 Reproduktivitetsegenskaber

Additionssætningen (sætning 1.3.5) udtrykkes her som:

Såfremt $Z_1, Z_2, \ldots, Z_k$ er uafhængige sådan at $Z_i \in \mathrm{P}(\mu_i)$, og vi sætter

$$Z_+ = Z_1 + Z_2 + \cdots + Z_k \ ,$$

da vil $Z_+ \in \mathrm{P}(\mu_1 + \mu_2 + \cdots + \mu_k)$.

4.24.4 Approksimationer

Den direkte normalfordelingsapproksimation til Poissonfordelingen er

$$P(c;\lambda) \approx \Phi\left(\frac{c+1/2-\lambda}{\sqrt{\lambda}}\right),$$

der fås ved den centrale grænseværdisætning. En bedre approksimation er

$$P(c;\lambda) \approx \Phi(2\sqrt{c+1}-2\sqrt{\lambda}).$$

4.25 Paretofordelingen

En kontinuert fordelt stokastisk variabel X siges at følge en Paretofordeling med parametre (α,β), hvis tætheden for X er af formen

$$g(x) = \begin{cases} \dfrac{\alpha\beta^\alpha}{x^{\alpha+1}} & \text{for } \beta \le x \\ 0 & \text{ellers} \end{cases} \tag{4.25.1}$$

hvor $0 < \beta$, $0 < \alpha$.

Kort skriver vi[1] $X \in \mathrm{Par}(\alpha,\beta)$.

Såfremt $X \in \mathrm{Par}(\alpha,\beta)$, og $Y = \beta_1 X$ med $\beta_1 > 0$, da vil $Y \in \mathrm{Par}(\alpha,\beta_1\beta)$. Parameteren β er således en skalaparameter for fordelingen.

Såfremt $X - \ln(\beta) \in \mathrm{Ex}(1/\alpha)$, og vi sætter $Y = \exp(X)$, da vil $Y \in \mathrm{Par}(\alpha,\beta)$.

Såfremt $X \in \mathrm{Par}(\alpha,\beta)$ og vi sætter $Y = 1/X^\alpha$, da vil $Y \in \mathrm{U}(0,1/\beta^\alpha)$, hvor $\mathrm{U}(a,b)$ angiver ligefordelingen på intervallet $[a,b]$.

Specielt har vi altså for $X \in \mathrm{Par}(\alpha,\beta)$ og $Z = (\beta/X)^\alpha$, at $Z \in \mathrm{U}(0,1)$.

Såfremt $X \in \mathrm{Par}(\mu,\beta)$ og $Y \in \mathrm{Par}(\nu,\beta)$ er uafhængige, da gælder for $Z = \min\{X,Y\}$, at $Z \in \mathrm{Par}(\mu+\nu,\beta)$.

Såfremt $X_1, X_2, \ldots, X_n$ er uafhængige variable, og vi sætter

$$Y = \prod_{i=1}^{n}(X_i/\beta),$$

[1] Parametriseringen adskiller sig fra parametriseringen i noten om Statistisk beslutningsteori fra 1991. Vi har byttet om på positionerne af α og β i forhold til den gamle note

da vil $\ln Y \in \mathrm{G}(n, 1/\alpha)$.

For $X \in \mathrm{Par}(\alpha, \beta)$ med $2 < \alpha$ gælder

$$\mathsf{E}[X] = \frac{\alpha\beta}{\alpha - 1}\,; \quad \mathsf{V}[X] = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$$

Fordelingsfunktionen for $\mathrm{Par}(\alpha, \beta)$-fordelingen betegnes med

$$\mathrm{Par}(c; \alpha, \beta) \stackrel{\mathrm{DEF}}{=} \int_\beta^c \frac{\alpha\beta^\alpha}{x^{\alpha+1}} dx = 1 - (\beta/c)^\alpha$$

for $\beta < c$.

Da β er en skalaparameter gælder

$$\mathrm{Par}(c; \alpha, \beta) = \mathrm{Par}(c/\beta; \alpha, 1)$$

Det første ufuldstændige moment for $\mathrm{Par}(\alpha, \beta)$-fordelingen er

$$\mu_1'(c) = \int_\beta^c x g(x) dx = \frac{\alpha\beta}{\alpha - 1} \mathrm{Par}(c; \alpha - 1, \beta)$$

og det tilsvarende centrale moment er

$$\mu_1(c) = \int_\beta^c \left(x - \frac{\alpha\beta}{\alpha-1}\right) g(x) dx = \frac{\alpha\beta}{\alpha-1} \left(\frac{\beta}{c} - 1\right) \left(\frac{\beta}{c}\right)^{\alpha-1}.$$

4.26 Referencer

Barndorff-Nielsen, O.E. (1978): *Information and Exponential Families.* Wiley, Chichester.

Cox, D.R. and Lewis, P.A.W. (1966): *The Statistical Analysis of Series of Events*, Methuen, London.

Johnson, N.L., Balakrishnan, N. and Kotz, S. (1995): *Continuous Univariate Distributions*, Vol 1-2, Wiley, New York

Johnson, N.L., Kotz, S. and Kemp, A.W. (1993): *Univariate Discrete Distributions*, Vol 1-3, 2.ed., Wiley, New York

Jørgensen, B. (1982): *Statistical Properties of the Generalized Inverse Gaussian Distribution*, Lecture notes in statistics, Vol 9, Springer Verlag, New York.

Kotz, S., Johnson, N.L. and Read, C.B., eds. (1988): *Encyclopedia of Statistical Sciences*, 9 vols., Wiley, New York

Tweedie, M.C.K. (1957): Statistical properties of inverse Gaussian distributions, *Annals of Mathematical Statistics*, **28**, pp. 362-377

Chhikara, R.S. and Folks, J.L. (1977): The Inverse Gaussian Distribution as a Lifetime Model, *Technometrics* **18**, pp. 189 -193

Wald, A. (1947): *Sequential Analysis*, Wiley & Sons, Inc. New York

Crow, E.L and Shimizu, K, eds. (1988): *Lognormal Distributions*, Marcel Dekker, New York

Afsnit 5

Momentfordelinger og Lorenzkurver

Fil: lorenz.tex 1998-01-13

Ved betragtning af endelige populationer (eller eventuelt bare af sædvanlige stikprøver) kan det være af interesse at vurdere, hvorledes populationens totalværdi er fordelt på de enkelte observationsenheder.

De såkaldte Lorenzkurver kan ofte være et nyttigt hjælpemiddel til vurdering af en sådan fordeling.

Til brug for behandlingen af Lorenzkurverne vil vi indledningsvist resumere teorien for momentfordelinger.

5.1 Momentfordelinger

Vi erindrer om definitionen af momentfordelingen for en positiv stokastisk variabel:

Lad den stokastiske variabel X have fordelingsfunktionen $F(\cdot)$ og tæthed (eller frekvensfunktion) $f(\cdot)$. Ved den ν 'te momentfordeling for X forstås

fordelingen med fordelingsfunktion

$$F_\nu(x) = \int_0^x t^\nu dF(t)/\mathsf{E}[X^\nu] \tag{5.1.1}$$

Det fremgår således, at $F_\nu(\cdot)$ er fordelingsfunktionen for fordelingen med tæthed (eller frekvensfunktion)

$$f_\nu(x) = x^\nu f(x)/\mathsf{E}[X^\nu] \tag{5.1.2}$$

Vi bemærker, at den ν 'te momentfordeling adskiller sig fra fordelingen af den stokastiske variabel $Y = X^\nu$. Vi har således

$$\mathsf{P}[X^\nu \leq y] = F_X(y^{1/\nu})$$

der i almindelighed er forskelligt fra $F_\nu(y^{1/\nu})$

For en endelig population med elementerne $x_1, x_2, \ldots, x_N$ kan relationen mellem (antals) fordelingen af X og den ν 'te momentfordeling for X fortolkes ved hjælp af konkrete populationsandele.

Fordelingen $F(\cdot)$ af populationsværdier er givet ved

$$F(x) = \Sigma_{i=1}^N \mathrm{I}_{[x_i \leq x]}/N$$

hvor $\mathrm{I}_{[\ ]}$ som sædvanligt angiver indikatorfunktionen for hændelsen i den kantede parentes.

$F(x)$ angiver således andelen af analyseenheder i populationen med værdier $\leq x$.

Den ν'te momentfordeling er givet ved

$$F_\nu(x) = \Sigma_{i=1}^N x_i^\nu \ \mathrm{I}_{[x_i \leq x]}/\Sigma_{i=1}^N x_i^\nu$$

$F_\nu(x)$ angiver andelen af værdier af x^ν, der er mindre end x , hvor andelen beregnes i forhold til den samlede masse af værdier af x^ν.

Eksempel 5.1.1 ***Det samlede landbrugsareal i Danmark fordelt på brugsstørrelser***

Nedenstående tabel angiver fordelingen på brug af det dyrket landbrugsareal i Danmark i 1989 fordelt efter brugsstørrelse. (Kilde: Danmarks Statistik, Statistisk Årbog 1992).

Brugs-størrelse (dyrket areal)	Antal Brug	Andel Brug %	Kumuleret andel %
< 5,0 ha	2 232	2.75	2.75
5,0 -9,9 ha	12 517	15.40	18.15
10,0 - 19,9 ha	19 605	24.12	42.27
20,0 - 29,9 ha	14 195	17.47	59.74
30,0 - 49,9 ha	17 153	21.11	80.85
50,0 - 99,9 ha	12 162	14.97	95.81
> 100,0 ha	3 403	4.19	100.00
ialt	81 267	100.00	

Tabellen er et eksempel på en sædvanlig antalsfordeling. Observationsenheden i er et landbrug og interessevariablen x_i er brugsstørrelsen. Figur 5.1 viser den kumulerede fordeling (efter antal) af brugsstørrelserne.

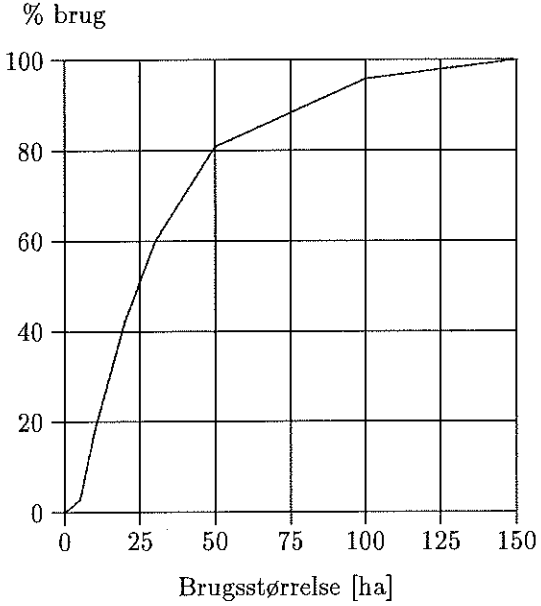
Idet det oplyses, at det samlede dyrkede areal udgør 2 774 127 [ha], kan man bestemme det gennemsnitlige dyrket areal:

$$\xi_x = \frac{2\ 774\ 127}{81\ 267} = 34,136\ [\text{ha}]$$

Såfremt det samlede dyrkede areal ikke havde været oplyst, kunne vi have approksimeret det gennemsnitlige dyrkede areal ved at tillægge alle brug i en given størrelsesgruppe den størrelse, der svarer til klassemidtpunktet. (For gruppen af brug større end 100 [ha] måtte vi skønne et passende klassemidtpunkt).

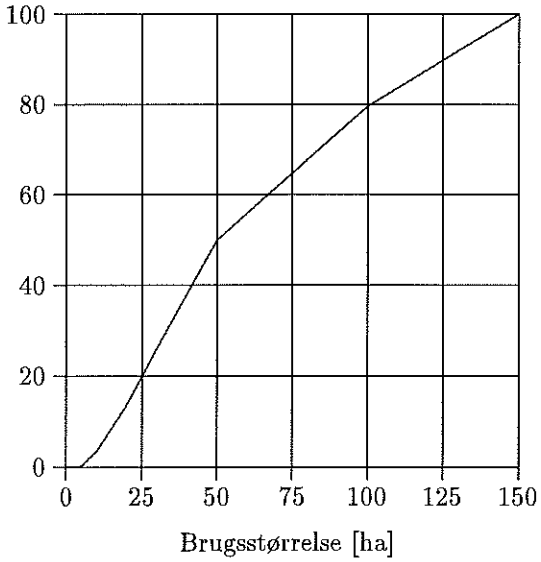
For hver af de anførte størrelsesklasser foreligger der endvidere oplysning om det samlede dyrkede areal for brugene i den pågældende klasse. Den herved fastlagte fordeling af det samlede dyrkede areal fordelt på brugsstørrelser er anført i nedenstående tabel

Figur 5.1. **Kumuleret antalsfordeling af brugsstørrelser**



Brugs-størrelse (dyrket areal)	Antal Brug	Dyrket areal (samlet) ha	Andel af samlet areal %	Kum. areal andel %
< 5,0 ha	2 232	3 748	0.14	0.14
5,0 - 9,9 ha	12 517	91 714	3.31	3.44
10,0 - 19,9 ha	19 605	284 791	10.27	13.71
20,0 - 29,9 ha	14 195	348 252	12.55	26.26
30,0 - 49,9 ha	17 153	659 604	23.78	50.04
50,0 - 99,9 ha	12 162	818 355	29.50	79.54
> 100,0 ha	3 403	567 662	20.46	100.00
ialt	81 267	2 774 127	100.0	

Figur 5.2. **Kumuleret arealfordeling efter brugsstørrelser**



Den sidste kolonne i tabellen angiver den kumulerede fordeling af arealenheder, fordelt efter brugsstørrelser, altså netop den første momentfordeling svarende til fordelingen af brugsstørrelser. Den kumulerede fordeling af arealenheder er vist i figur 5.2.

Danmarks Statistik har endvidere oplyst, at den gennemsnitlige brugsstørrelse pr. arealenhed er 83,46 [ha].

Såfremt denne værdi ikke havde været oplyst, havde man kunnet danne en approksimativ værdi ved at bestemme den gennemsnitlige brugsstørrelse for hver af de anførte klasser, og derefter bestemme det vægtede gennemsnit af disse ved at vægte med den andel af det samlede areal, der er knyttet til pågældende klasse. □

Eksempel 5.1.2 ***Første momentfordeling af fiberlængder***

Betragt en population af fibre med længder $x(i), i = 1, 2, \ldots, N$. Hyppighedsfordelingen, $F(x)$, af fiberlængden angiver den andel (målt som antalsandel) af fibrene, der har en længde $\leq x$. Den første momentfordeling, $F_1(x)$, angiver den andel (målt som længdeandel) af den samlede fiberlængde, der hidrører fra fibre med længder $\leq x$. □

Eksempel 5.1.3 ***Tredie momentfordeling af kugleformede partikler***

Betragt en population af kugleformede partikler med diametre $x(i), i = 1, 2, \ldots, N$

Hyppighedsfordelingen $F(x)$ af partikeldiametrene angiver den andel (målt som antalsandel) af partiklerne, der har en diameter $\leq x$.

Antages at partiklerne har samme massefylde, vil værdierne svarende til den tredje momentfordeling, $F_3(x)$, angive den andel (målt som vægtandel) af populationens samlede vægt, der hidrører fra kugler med diametre $\leq x$. □

Eksempel 5.1.4 ***Første momentfordeling af udgiftsposter***

Betragt en population af udgiftsposter med beløb $x(i), i = 1, 2, \ldots, N$

Hyppighedsfordelingen $F(x)$ af beløbene angiver den andel (målt som antalsandel) af udgiftsposterne, der har et beløb $\leq x$.

Momentfordelingen, $F_1(x)$, angiver den andel (målt som beløbsandel) af de samlede udgifter, der hidrører fra udgiftsposter med beløb $\leq x$. □

Idet vi sætter

$$\mathsf{E}_{\mathrm{M}_\nu}[X] = \int_0^\infty x f_\nu(x)dx = \int_0^\infty x^{\nu+1} f(x)dx / \mathsf{E}[X^\nu]$$

finder vi, at forventningsværdien i den ν'te momentfordeling er forholdet mellem det $\nu+1$'te og det ν'te moment i fordelingen af X, dvs

$$\mathsf{E}_{\mathrm{M}_\nu}[X] = \mathsf{E}[X^{\nu+1}]/\mathsf{E}[X^\nu] \tag{5.1.3}$$

hvor vi har benyttet fodtegnet M_ν for at indikere, at forventningsværdien beregnes i den ν'te momentfordeling.

Specielt gælder, at forventningsværdien i den første momentfordeling er:

$$\mathsf{E}_{\mathrm{M}_1}[X] = \mathsf{E}[X]\{1 + \mathsf{V}[X]/(\mathsf{E}[X])^2\} = \xi_X(1+\gamma_X^2) \tag{5.1.4}$$

hvor ξ_X og γ_X som sædvanligt betegner forventningsværdi og relativ varians i fordelingen af X (fordelt efter antal).

Eksempel 5.1.5 ***Momentfordeling for RG-fordeling***

Antag, at fiberlængderne, X, i en population af fibre kan beskrives ved en $\mathrm{RG}(\alpha,\beta)$-fordeling. Man finder da, at den første momentfordeling, dvs. fordelingen af fibre efter længdeandele er en $\mathrm{RG}(\alpha-1,\beta)$-fordeling Der gælder da at den gennemsnitlige fiberlængde er

$$\mathsf{E}[X] = \mathsf{E}[\,\mathrm{RG}(\alpha,\beta)] = \frac{\beta}{\alpha-1}$$

og variansen i fordelingen af fiberlængder er

$$\mathsf{V}[X] = \left(\frac{\beta}{\alpha-1}\right)^2 \frac{\beta}{\alpha-2}$$

Forventningsværdien af fiberlængden X i momentfordelingen af X udtrykker den gennemsnitlige fiberlængde, der en tilknyttet en længdeenhed. Man finder, at forventningsværdien af X i den første momentfordeling er forventningsværdien i $\mathrm{RG}(\alpha,\beta)$-fordelingen, dvs

$$\mathsf{E}_{\mathrm{M}_1}[X] = \mathsf{E}[\,\mathrm{RG}(\alpha-1,\beta)] = \frac{\beta}{\alpha-2}$$

Tilsvarende udtrykker variansen af fiberlængden X i momentfordelingen af X, hvor stor variation, der er i de fiberlængder, der er tilknyttet en længdeenhed.

Man finder

$$\mathsf{V}_{\mathrm{M}_1}[X] = \mathsf{V}[\,\mathrm{RG}(\alpha-1,\beta)] = \left(\frac{\beta}{\alpha-2}\right)^2 \quad \frac{\beta}{\alpha-3}$$

□

Betragter vi den variable $Y = 1/X$ finder vi, at forventningsværdi og varians af Y i den første momentfordeling af X er

$$\mathsf{E}_{\mathrm{M}_1}[1/X] = 1/\mathsf{E}[X] \tag{5.1.5}$$

$$\mathsf{V}_{\mathrm{M}_1}[1/X] = \frac{\mathsf{E}[X]\,\mathsf{E}[1/X]-1}{\mathsf{E}[X]^2} \tag{5.1.6}$$

Bevis:
Følger ved at bemærke, at

$$\mathsf{E}_{\mathrm{M}_1}[1/X^2] = \frac{\mathsf{E}[1/X]}{\mathsf{E}[X]}$$

Eksempel 5.1.6 ***Fordeling af reciprok fiberlængde i momentfordeling for RG -fordeling***

Antag som i det foregående eksempel, at fiberlængderne, X, i en population af fibre kan beskrives ved en $\mathrm{RG}(\alpha,\beta)$-fordeling.

På grund af relationen mellem den reciproke gammafordeling og gammafordelingen $1/\mathrm{RG}(\alpha,\beta) = \mathrm{G}(\alpha,1/\beta)$, finder vi, at forventningsværdien af den reciprokke fiberlængde, $1/X$, i momentfordelingen bliver $\mathsf{E}[\mathrm{G}(\alpha-1,1/\beta)]$, hvorfor vi har

$$\mathsf{E}_{\mathrm{M}_1}[1/X] = \frac{\alpha-1}{\beta} \quad \text{og} \quad \mathsf{V}_{\mathrm{M}_1}[1/X] = \mathsf{V}[\mathrm{G}(\alpha-1,1/\beta)] = \frac{\alpha-1}{\beta^2}$$

□

Eksempel 5.1.7 ***Momentfordeling for* LN*-fordeling***

Betragt en population af partikler og antag, at fordelingen af partikelvægtene X kan beskrives ved en $\mathrm{LN}(\alpha, \beta^2)$-fordeling med forventningsværdien $\mathsf{E}[X] = \exp(\alpha + \frac{1}{2}\beta^2)$ og $\mathsf{V}[X] = (\mathsf{E}[X])^2 \{\exp(\beta^2) - 1\}$

Den første momentfordeling af X er da en $\mathrm{LN}(\alpha + \beta^2, \beta^2)$-fordeling, og forventningsværdi og varians i den første momentfordeling bliver

$$\begin{aligned} \mathsf{E}_{\mathrm{M}_1}[X] &= \exp(\alpha + 3\beta^2/2) = \mathsf{E}[X]\exp(\beta^2) \\ \mathsf{V}_{\mathrm{M}_1}[X] &= \exp(2\alpha + 3\beta^2/2)\{\exp(\beta^2) - 1\} \end{aligned}$$

Tilsvarende finder man

$$\mathsf{E}_{\mathrm{M}_1}[1/X] = \exp(-\alpha - \frac{1}{2}\beta^2) \quad \text{og} \quad \mathsf{V}_{\mathrm{M}_1}[1/X] = \exp(-2\alpha - 3\beta^2)\{\exp(\beta^2) - 1\}$$

□

5.2 Lorenzkurver og Gini-index

Relationen mellem den første momentfordeling af X og antalsfordelingen af X illustreres ofte ved at foretage en grafisk afbildning af $F_1(x)$ mod $F(x)$. Et sådant diagram kaldes et *Lorenz-diagram* efter økonomen M. O. Lorenz, der foreslog denne afbildning til beskrivelse af fordelingen af formuemassen eller indtægtsmassen i en befolkning med henblik på vurdering af befolkningens skatteevne.

Definition 5.2.1 ***Lorenzkurven***

For en fordeling $F(\cdot)$ med en endelig forventningsværdi defineres Lorenzkurven $L_F(\cdot)$ ved

$$L_F(p) \stackrel{\mathrm{DEF}}{=} \int_0^p F^{-1}(v)dv/\mu \tag{5.2.1}$$

hvor $\mu = \int_0^\infty x dF(x)$ angiver forventningsværdien i fordelingen.

(Vi bemærker at definitionen er gyldig, også selv om $F(\cdot)$ er konstant på enkelte intervaller, idet de tilsvarende værdier af $v = F(x)$ ikke bidrager til integralet $L_F(p)$).

Såfremt fordelingsfunktionen $F(\cdot)$ er kontinuert med den kontinuerte tæthed $f(\cdot)$, finder man ved at indføre transformationen $t = F^{-1}(v)$, $v = F(t)$, $dv = f(t)dt$, at

$$L_F(p) = \int_0^p F^{-1}(v)dv/\mu = \int_{-\infty}^{F^{-1}(p)} tf(t)dt/\mu \qquad (5.2.2)$$

Man kan således konstruere Lorenzkurven ved at indtegne sammenhørende værdier af $p(x) = F(x)$ og

$$L(x) = L_F(F(x)) = \int_{t=0}^{x} tf(t)dt/\mu \qquad (5.2.3)$$

□

Eksempel 5.2.1 ***Det samlede landbrugsareal i Danmark fordelt på brugsstørrelser (fortsat)***

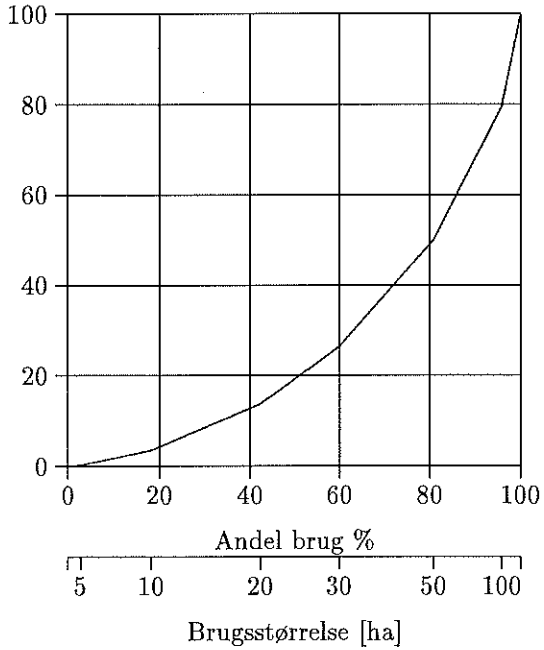
Vi betragter data fra eksempel 5.1.1

Ved hjælp af tabellerne i eksemplet kan vi danne nedenstående tabel over samhørende værdier af den kumulerede antalsandel og den kumulerede arealandel for hver størrelseskategori.

Brugs-størrelse (dyrket areal)	Antal Brug	Andel Brug %	Kum. andel brug %	Kum. andel areal %
< 5,0 ha	2 232	2.75	2.75	0.14
5,0 -9,9 ha	12 517	15.40	18.15	3.44
10,0 - 19,9 ha	19 605	24.12	42.27	13.71
20,0 - 29,9 ha	14 195	17.47	59.74	26.26
30,0 - 49,9 ha	17 153	21.11	80.85	50.04
50,0 - 99,9 ha	12 162	14.97	95.81	79.54
> 100,0 ha	3 403	4.19	100.00	100.00
ialt	81 267	100.00		

Figur 5.3 viser den resulterende Lorenzkurve for fordelingen af størrelsen af landbrug. Det ses, at de 20 % største brug omfatter ca. 50 % af det samlede landbrugsareal. □

Figur 5.3. **Lorenzkurve for fordelingen af brugsstørrelser**



Sætning 5.2.1 *Lorenzkurven er en konveks funktion af p*

Såfremt fordelingsfunktionen $F(\cdot)$ er kontinuert med den kontinuerte tæthed $f(\cdot)$, gælder

$$L'_F(p) = F^{-1}(p)/\mu \tag{5.2.4}$$

$$L''_F(p) = \frac{1}{f(F^{-1}(p))\mu} \tag{5.2.5}$$

Bevis:
Ved differentiation af (5.2.3) med hensyn til x finder vi

$$L'(x) = L'_F(F(x))F'(x) = xf(x)/\mu$$

hvorfor der gælder for $f(x) > 0$

$$L'_F(F(x)) = x/\mu \tag{5.2.6}$$

Idet $x = F^{-1}(p)$ fås den første del af resultatet. Differentierer vi nu (5.2.4) med hensyn til p og benytter, at

$$\frac{dx}{dp} = \frac{1}{dp/dx} = \frac{1}{F'(x)} = \frac{1}{f(F^{-1}(p))}$$

finder vi (5.2.5) □

Det gælder således, at Lorenzkurven svarende til en kontinuert fordeling er strengt voksende og konveks som funktion af p.

Bemærkning 1 *Alternative definitioner af Lorenzkurven*

Ovenstående definition af Lorenzkurven sammenholder andelen p af de mindste værdier af X med den tilsvarende andel $L_F(p)$ i momentfordelingen.

I nogle sammenhænge er interessen snarere rettet mod de største værdier af X, og man betragter derfor den komplementære Lorenzkurve $L^*_F(\cdot)$, der afbilder $1 - F_1(x)$ mod $1 - F(x)$:

$$L^*_F(p^*) \stackrel{\text{DEF}}{=} \int_{1-p^*}^{1} F^{-1}(v)dv/\mu = \int_{F^{-1}(1-p^*)}^{\infty} tf(t)dt/\mu = 1 - L_F(1-p^*)$$

□

Bemærkning 2 *Vurdering af Lorenzkurvens asymmetri*

For bedre at kunne vurdere asymmetrien i en Lorenzkurve, kan man vælge at betragte Lorenzkurven i et koordinatsystem med første akse ud af vinkelhalveringslinien (svarende til den egale fordeling, etpunktsfordelingen) og anden aksen vinkelret derpå, dvs

$$\begin{Bmatrix} p^\diamond \\ L^\diamond \end{Bmatrix} = \begin{Bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{Bmatrix} \begin{Bmatrix} p \\ L \end{Bmatrix}$$

□

Gini's indeks

Som et mål for uegaliteten i en indkomsten foreslog den italienske økonom Gini at betragte arealet mellem Lorenzkurven svarende til étpunktsfordelingen ($L(p) = p$) og Lorenzkurven svarende til den aktuelle fordeling, $L_F(\cdot)$

Definition 5.2.2 *Gini indeks*

Lad X have fordelingsfunktionen $F_X(\cdot)$ og den tilsvarende Lorenzkurve $L_F(\cdot)$. Gini-indekset G_F for X er da

$$G_F \stackrel{\text{DEF}}{=} \int_0^1 [p - L_F(p)]dp \tag{5.2.7}$$

□

Der gælder

Sætning 5.2.2 *Alternative formuleringer af Gini-indekset*

Lad X have fordelingsfunktionen $F_X(\cdot)$ og Gini-indeks G_F. Da gælder

$$G_F = 2 \int F_X(t)[1 - F_X(t)]dt = 4\mathsf{COV}[X, F_X(X)] \tag{5.2.8}$$

Såfremt X_1 og X_2 er uafhængige identisk fordelte med fordelingsfunktion $F_X(\cdot)$ gælder desuden

$$G_F = \mathsf{E}[\,|X_1 - X_2|\,] \tag{5.2.9}$$

Bevis:
Følger umiddelbart □

Vi bemærker, at variansen $\mathsf{V}[X]$ kan udtrykkes i analogi med (5.2.9) som $\mathsf{V}[X] = (1/2)\mathsf{E}[(X_1 - X_2)^2]$. Gini-indekset kan således tages som udtryk for spredningen i fordelingen, udtrykt ved numeriske afvigelser.

Relationerne (5.2.8) udtrykker, at Gini-indekset alternativt kan formuleres som en vægtning af observationen X med dens rang $F(\cdot)$, eller som en vægtning af rangen $F(\cdot)$ med differenser af de ordnede observationer $1 - F(\cdot)$. Disse alternative formuleringer giver tilsvarende anledning til alternative formuleringer af estimatorer. (Se f.eks. Lerman og Yitzhaki (1984)).

Definition 5.2.3 ***Gini stikprøvefunktionen***

Lad $X_1, X_2, \ldots, X_n$ angive et sæt observationer og lad $X_{(1)}.X_{(2)}, \ldots, X_{(n)}$ angive de tilsvarende ordnede observationer. Gini-stikprøvefunktionen G_n er

$$G_n(X_1, X_2, \ldots, X_n) \stackrel{\text{DEF}}{=} \frac{\sum_{i=1}^{n-1} i(n-i)(X_{(i+1)} - X_{(i)})}{(n-1)\sum_{i=1}^{n} X_i} \tag{5.2.10}$$

□

Eksempel 5.2.2 ***Lorenzkurven for en étpunktsfordeling***

Såfremt $F(\cdot)$ er en étpunktsfordeling, der koncentrerer hele massen i μ, bliver Lorenzkurven den identiske afbildning

$$L_F(p) = p \quad \text{for} \quad 0 \leq p \leq 1$$

Enhver andel, p, af populationen bidrager altså netop med andelen p af den totale sum af x-værdier. □

Eksempel 5.2.3 ***Lorenzkurven for en ligefordeling***

Såfremt $F(\cdot)$ er en ligefordeling over intervallet $0 \leq x \leq 2\mu$ med forventningsværdien $\mathsf{E}[X] = \mu$, finder vi, at den første momentfordeling er fordelingen med tæthed

$$f(x) = \begin{cases} x/(2\mu) & \text{for} \quad 0 \leq x \leq 2\mu \\ 0 & \text{ellers} \end{cases}$$

Forventningsværdien i den første momentfordeling er derfor
$\mathsf{E}[X^2]/\mathsf{E}[X] = \frac{4}{3}\mu$.

Lorenzkurven er bestemt som

$$L_F(p) = p^2 \quad \text{for} \quad 0 \leq p \leq 1$$

□

Eksempel 5.2.4 ***Lorenzkurven for en Paretofordeling***

Såfremt $F(\cdot)$ er en Pareto-fordeling, $\mathrm{Par}(\beta, \alpha)$

$$F(x) = \begin{cases} 0 & \text{for} \quad x < \beta \\ 1 - (\beta/x)^{2\alpha} & \text{for} \quad \beta \leq x \end{cases}$$

med forventningsværdien

$$\mathsf{E}[X] = \frac{\alpha}{\alpha - 1}\,\beta\,,$$

finder vi for $\alpha > 1$, at første momentfordeling er en $\mathrm{Par}(\beta, \alpha - 1)$- fordeling. For $\alpha > 2$ eksisterer forventningsværdien $\dfrac{\alpha - 1}{\alpha - 2}\,\beta$ i den første momentfordeling. Jo mindre værdi af α, desto mere vægt vil der ligge på de store x-værdier i momentfordelingen.

For $\alpha > 1$ eksisterer Lorenzkurven og kurven er bestemt ved

$$L_{\mathrm{Par}}(p) = 1 - (1 - p)^{(\alpha-1)/\alpha} \quad \text{for} \quad 0 \leq p \leq 1$$

Funktionen er konveks, da $(\alpha - 1)/\alpha < 1$. Lorenzkurven svarer i øvrigt netop til den kumulerede fordelingsfunktion for en $\mathrm{Be}(1, (\alpha - 1)/\alpha)$-fordelt variabel.

Såfremt fordelingen af de personlige formuer i befolkningen kan beskrives ved en $\mathrm{Par}(\beta, \alpha)$-fordeling med $\alpha = 1.16$ finder man ved indsættelse af

$p = 0.8$, at $L_{\mathrm{Par}}(p) = 0.2$, altså at den samlede formue, der besiddes af de 80 % mindst formuende personer i befolkningen kun udgør 20 % af befolkningens samlede formue. Eller - omvendt, da $L^*_{\mathrm{Par}}(0.2) = 0.8$, at de største 80 % af den samlede formue er koncentreret på 20 % af befolkningen. Denne relation, kaldet 80/20-reglen, der blev fundet af Lorenz, kaldes ofte "Paretos lov". □

Eksempel 5.2.5 ***Lorenzkurven for en LN -fordeling***

Såfremt $F(\cdot)$ er en $\mathrm{LN}(\alpha, \beta^2)$-fordeling, bliver første momentfordeling en $\mathrm{LN}(\alpha+\beta, \beta^2)$-fordeling. Der gælder $\mathsf{E}[X] = \exp(\alpha + \frac{1}{2}\beta^2)$ og forventningsværdien i den første momentfordeling for X er $\exp(\alpha+\frac{3}{2}\beta^2)$. Tyngdepunktet i den første momentfordeling er således øget med faktoren $\exp(\beta^2)$ i forhold til fordelingen af X.

Man har Lorenzkurven

$$L_{\mathrm{LN}}(p) = \Phi(z_p - \beta) \quad \text{for} \quad 0 \leq p \leq 1$$

hvor z_p angiver p-fraktilen i den normerede normale fordeling.

Vi bemærker, at $\mathsf{P}[X \leq \mathsf{E}[X]] = \Phi(\beta/2)$. Den værdi af Lorenzkurven, der svarer til $x = \mathsf{E}[X]$, er således $L_{\mathrm{LN}}(\Phi(\beta/2)) = \Phi(-\beta/2) = 1 - \Phi(\beta/2)$.

For logaritmisk normalfordelte populationsværdier gælder åbenbart, at andelen af analyseenheder, hvis værdier er mindre end populationsmiddelværdien er een minus den andel af populationsværdier, der besiddes af disse enheder. □

Eksempel 5.2.6 ***Lorenzkurven for en Gamma-fordeling***

Såfremt fordelingen $F(\cdot)$ er en gammafordeling $\mathrm{G}(\alpha, \beta)$, bliver første momentfordeling en $\mathrm{G}(\alpha+1, \beta)$-fordeling. Der gælder $\mathsf{E}[X] = \alpha\beta$, og forventningsværdien i den første momentfordeling er $(\alpha+1)\beta$. Tyngdepunktet i den første momentfordeling er således øget med faktoren β/α i forhold til fordelingen af X.

Lorenzkurven bestemmes som sammenhørende værdier af den kumulerede fordelingsfunktion for $\mathrm{G}(\alpha, \beta)$-fordelingen og $\mathrm{G}(\alpha+1, \beta)$-fordelingen.

Specielt finder vi for $\mathrm{Ex}(\beta)$-fordelingen, at Lorenzkurven er givet ved sammenhørende værdier af

$$p(x) = \mathsf{P}[\mathrm{Ex}(\beta) \leq x] = 1 - \exp(-x/\beta)$$

og

$$L(x) = \mathsf{P}[\mathrm{G}(2,\beta) \leq x] = 1 - \mathsf{P}[\mathrm{P}(x/\beta) \leq 1] = 1 - \exp(-x/\beta)\{1 + x/\beta\}$$

hvoraf vi får

$$L_{\mathrm{Ex}}(p) = p + (1-p)\ln(1-p) \tag{5.2.11}$$

□

En afbildning af en fordeling ved en Lorenzkurve kan ofte være et nyttigt værktøj ved planlægning af en stikprøveundersøgelse.

Definition 5.2.4 ***Lorenz koncentrationskurver***

Lad X være en positiv stokastisk variabel og lad $g(\cdot)$ være positiv funktion på $\mathbb{R}_+$ sådan at $\mathsf{E}[g(X)]$ er endelig. Koncentrationskurven $L_g(\cdot)$ for $g(\cdot)$ angiver den kumulative andel af $g(x)$ ordnet efter voksende værdier af x. Såfremt fordelingen af X er kontinuert med tætheden $f(\cdot)$, defineres koncentrationskurven ved

$$L_g(p) \stackrel{\mathrm{DEF}}{=} \int_{-\infty}^{F^{-1}(p)} g(t)f(t)dt / \mathsf{E}[g(X)] \tag{5.2.12}$$

□

Vi bemærker, at den sædvanlige Lorenzkurve fås for $g(x) = x$.

Sætning 5.2.3 ***Konveksitet af koncentrationskurven***

Såfremt $g(\cdot)$ er en strengt voksende funktion med $g(0) = 0$ gælder at koncentrationskurven er konveks, og $L_g(p) \leq p$.
Såfremt $g(\cdot)$ er en strengt aftagende funktion med $g(0) = 0$ gælder at koncentrationskurven er konkav, og $L_g(p) \geq p$.
Lad $g(\cdot)$ og $h(\cdot)$ være positive funktioner af X sådan at $\mathsf{E}[g(X)]$ og $\mathsf{E}[h(X)]$ eksisterer. Da gælder

$$L_g(p) \geq L_h(p) \quad \text{for} \quad 0 \leq p \leq 1 \Leftrightarrow \frac{g'(x)}{g(x)} < \frac{h'(x)}{h(x)} \quad 0 < x$$

og omvendt

$$L_g(p) \leq L_h(p) \quad \text{for} \quad 0 \leq p \leq 1 \Leftrightarrow \frac{g'(x)}{g(x)} > \frac{h'(x)}{h(x)} \quad 0 < x$$

Bevis:
se Kakwani (1980). □

5.3 Referencer

Kakwani, N. C. (1980). *Income Inequality and Poverty.* Oxford University Press, Oxford.

Lerman, R.I. and Yitzhaki, S (1984): A note on the calculation and interpretation of the Gini index, *Economics Letters*, **15**, pp. 363-368.

Appendiks A

Sandsynlighedsmål, stokastiske variable og fordelingsfunktioner

Fil: /tex/stat3/fordbog/fordapp.tex 1998-01-04

A.1 Sandsynlighedsmål

Definition A.1.1 *σ-algebra*

Lad E være en vilkårlig mængde og lad $\mathcal{A}$ være et system af delmængder af E. Mængdesystemet $\mathcal{A}$ kaldes en σ-algebra over E, såfremt der gælder:

1. $E \in \mathcal{A}$ og $\emptyset \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$
3. $A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$

□

Definition A.1.2 *Måleligt rum* Lad E være en vilkårlig mængde og lad $\mathcal{A}$ være en σ-algebra over E, da kaldes parret $(E, \mathcal{A})$ for et måleligt rum og mængderne i $\mathcal{A}$ kaldes målelige mængder. □

Grunden til denne betegnelse er, at med mindre der er defineret et system af delmængder af E (en σ-algebra), kan man ikke definere et meningsfuldt mål på E

Definition A.1.3 *Målelig afbildning*

Lad $(E, \mathcal{A})$ og $(F, \mathcal{B})$ være to målelige rum og lad f være en afbildning fra E til F. Såfremt f tilfredsstiller

$$f^{-1}(B) \in \mathcal{A} \quad \text{for ethvert } B \in \mathcal{B} \tag{A.1.1}$$

siges afbildningen $f : E \to F$ at være målelig.

□

Definition A.1.4 *Mål*

Et mål på et måleligt rum $(E, \mathcal{A})$ er en afbildning $\mu : \mathcal{A} \to [0, \infty]$, der er σ-additiv, og som har egenskaben $\mu\{\emptyset\} = 0$.

□

Definition A.1.5 *σ-endeligt mål*

Lad μ være et mål på det målelige rum $(E, \mathcal{A})$. Målet μ kaldes σ-endeligt, hvis der findes en tællelig samling $(A_i)_{i \in I}$ af målelige delmængder af E, sådan at

$$\bigcup_{i \in I} A_i = E$$

og $\mu(A_i) < \infty$ for $i \in I$.

I det følgende vil vi kun betragte σ-endelige mål. □

Eksempel A.1.1 *Lebesgue målet*

Lad $E = \mathbb{R}$ og $\mathcal{B}$ være Borelsigmaalgebraen på $\mathbb{R}$ (dvs σ-algebraen over alle åbne mængder). Der findes da netop ét mål på $(\mathbb{R}, \mathcal{B})$, hvis værdi på et interval $]a, b]$ med $a < b$ netop er intervallængden $b - a$. Dette mål kaldes Lebesguemålet på $\mathbb{R}$ og betegnes ofte med λ. □

Eksempel A.1.2 *Tællemålet*

Lad M være en mængde med et (højst) tælleligt antal elementer, og lad $\mathcal{D}(M)$ være σ-algebraen af alle delmængder af M. Funktionen μ bestemt ved

$$\mu(A) = \text{antal elementer i } A$$

er da et mål på $(M, \mathcal{D}(M))$. □

Definition A.1.6 *Integral mht mål*

Lad $(E, \mathcal{A})$ være et måleligt rum og lad $\mu(\cdot)$ være et mål på $(E,)$ Lad f være en ikke-negativ reel funktion $f(\cdot)$ på $(E, \mathcal{A})$, som er målelig med hensyn til Borel-sigmaalgebraen (σ-algebraen af åbne mængder i $\mathbb{R}$).

Man kan da entydigt definere en størrelse,

$$I(f) = \lim_{n\to\infty} \sum_{q=0}^{n2^n-1} \frac{q}{2^n} \mu\Big(\Big\{x \Big| \frac{q}{2^n} \le f(x) < \frac{q+1}{2^n}\Big\}\Big) + n\mu(\{x|f(x) \ge n\})$$

Størrelsen $I(f)$ er et ikke-negativ tal, eventuelt ∞.

For en vilkårlig målelig funktion f på $(E, \mathcal{A})$ (dvs ikke nødvendigvis ikke-negativ) er der følgende tre muligheder (hvor $f^+(x) = \max\{f(x), 0\}$ og $f^- = \max\{-f(x), 0\}$ betegner henholdsvis den positive og den negative del af f):

1. $I(f^+) < \infty$ og $I(f^-) < \infty$
   I dette tilfælde siges f at være integrabel, og integralet af f sættes til
   $$I(f) = I(f^+) - I(f^-)$$

2. Enten er $I(f^+) < \infty$ eller $I(f^-) < \infty$, men ikke begge.
   I dette tilfælde siges f ligeledes at være integrabel, og integralet af f sættes til $I(f) = I(f^+) - I(f^-)$. (Denne størrelse har mening (evt $\pm\infty$), da kun ét af leddene kan være uendeligt.

3. Såvel $I(f^+)$ som $I(f^-)$ er uendelige. I dette tilfælde er $I(f)$ ikke defineret. Man siger, at f ikke er integrabel.

Såfremt f er integrabel jvf ovenstående, kalder man størrelsen $I(f)$ for integralet af f med hensyn til målet μ, og man bruger symbolet $\int_E f d\mu$, $\int f d\mu$, $\int f(x)\mu\{dx\}$ for integralet af f. □

Bemærkning 1 *Sammenligning med Riemann-integralet*

Forskellen på det herved introducerede integralbegret og det sædvanlige (Riemann) integral er, at Riemann-integralet tager udgangspunkt i en intervalinddeling af x-aksen og en tilnærmelse af funktionen f ved en trappefunktion, der er konstant på disse intervaller, hvorimod det mere generelle integralbegreb tager udgangspunkt i en intervalinddeling af y-aksen, og en tilnærmelse af f ved en funktion, der er konstant på de tilsvarende x-mængder (der skal være målelige).

Det er således ikke nødvendigt at forudsætte kontinuitet af f for at opnå integrabilitet ved det generelle integralbegreb. Det er nok at f er målelig. Hvis feks. E blot har et tælleligt antal elementer (og μ er tællemålet), kan vi altså stadig tale om et integral. (I dette tilfælde udarter integralet til en sum). □

Definition A.1.7 *Restriktion af mål*

Lad $(E, \mathcal{A})$ være et måleligt rum, og lad A_0 være en målelig delmængde af E. Da er μ's restriktion til $A_0 \cap \mathcal{A})$ et mål på det målelige rum $(A_0, A_0 \cap \mathcal{A})$. Dette mål kaldes restriktionen af μ til A_0 og betegnes μ_{A_0}. □

Sætning A.1.1 *Integration med hensyn til restriktion af mål*

Lad f være en funktion, der er målelig m.h.t. $(A_0, A_0 \cap \mathcal{A})$. Da gælder, at f er integrabel, hvis og kun hvis funktionen $f^{\square}$ givet ved

$$f^{\square}(x) = \begin{cases} f(x) & \text{for } x \in A_0 \\ 0 & \text{for } x \notin A_0 \end{cases}$$

er integrabel m.h.t. μ. Såfremt dette er tilfældet, gælder

$$\int f d\mu_{A_0} = \int f^{\square} d\mu$$

□

Definition A.1.8 *Transformation af mål*

Lad $(E, \mathcal{A})$ være et måleligt rum med målet μ, og lad $t : (E, \mathcal{A}) \to (F, \mathcal{B})$ være en målelig afbildning. Da er funktionen

$$\nu(B) = \mu\{t^{-1}(B))\}, \quad \text{for } B \in \mathcal{B}$$

et mål på $(F, \mathcal{B})$. Dette mål kaldes det ved t transformerede mål af μ og betegnes $t\mu$.

Lad yderligere $(C, \mathcal{C})$ være et måleligt rum og lad $s : (F, \mathcal{B}) \to (C, \mathcal{C})$ være en målelig afbildning. Der gælder da, at

$$s(t(\mu)) = s \circ t(\mu)$$

□

Sætning A.1.2 ***Integration med hensyn til transformeret mål***

Der gælder, at en funktion g, der er målelig m.h.t. $(F, \mathcal{B})$ er integrabel m.h.t. $t(\mu)$ hvis og kun hvis funktionen $g \circ t$ er integrabel m.h.t. μ. Hvis dette er tilfældet, gælder at

$$\int g \, dt(\mu) = \int g \circ t \, d\mu$$

□

Definition A.1.9 ***Tæthed af mål*** Lad μ være et mål $(E, \mathcal{A})$ og lad f være en ikke-negativ funktion på E, der er målelig med hensyn til $(E, \mathcal{A})$. Relationen

$$\nu(A) = \int (1_A f) \, d\mu, \quad \text{for } A \in \mathcal{A},$$

hvor 1_A betegner indikatorfunktionen for mængden A, definerer et mål ν på $(E, \mathcal{A})$. Målet betegnes $f\mu$. Målet siges at have tætheden f med hensyn til μ. □

Såfremt h er en ikke-negativ funktion på E, der er målelig med hensyn til $(E, \mathcal{A})$. gælder, at

$$h(f\mu) = (hf)\mu (= f(h\mu))$$

Sætning A.1.3 ***Integration med hensyn til mål med tæthed***

Der gælder, at en funktion g, der er målelig m.h.t. $(E, \mathcal{A})$ er integrabel m.h.t. $f\mu$ hvis og kun hvis funktionen gf er integrabel m.h.t. μ. Hvis dette er tilfældet, gælder at

$$\int g \, df\mu = \int gf \, d\mu$$

□

Definition A.1.10 ***Produktmål***

Lad $(E, \mathcal{A})$ være en målelig mængde med målet μ og $(F, \mathcal{B})$ tilsvarende være en målelig mængde med målet ν, og lad $\mathcal{A} \otimes \mathcal{B}$ angive produktsigmalegemet af $\mathcal{A}$ og $\mathcal{B}$ (dvs. det mindste sigmalegeme udspændt af mængder $A \times B$ med $A \in \mathcal{A}$ og $B \in \mathcal{B}$.

Målet η på $(E \times F, \mathcal{A} \otimes \mathcal{B})$, defineret ved

$$\eta(A \times B) = \mu(A)\nu(B) \quad \text{for } A \in \mathcal{A},\ B \in \mathcal{B} \tag{A.1.2}$$

kaldes produktmålet af μ og ν. Målet betegnes med $\mu \otimes \nu$. □

Sætning A.1.4 ***Entydighed af produktmål***

Der findes ét og kun ét mål η på $(E \times F, \mathcal{A} \otimes \mathcal{B})$, der opfylder (A.1.2). □

Såfremt ω er et mål på det målelige rum $(G, \mathcal{G})$ gælder, at

$$\mu \otimes (\nu \otimes \omega) = (\mu \otimes \nu) \otimes \omega$$

Man behøver derfor ikke sætte parenteser, men kan uden misforståelse skrive $\mu \otimes \nu \otimes \omega$.

Sætning A.1.5 ***Integration m.h.t. produktmål***

Lad μ og ν være mål på hhv $(E, \mathcal{A})$ og $(F, \mathcal{B})$, og lad f være en ikke-negativ funktion på $E \times F$, der er målelig med hensyn til $(E \times F, \mathcal{A} \otimes \mathcal{B})$. Da gælder, at funktionen $h : E \to \mathbb{R}_+ \cup \infty$ defineret ved

$$h(x) = \int f(x, y)\ d\nu(y)$$

være ikke-negativ og målelig med hensyn til $(E, \mathcal{A})$. Der gælder, at

$$\int h(x)\ d\mu(x) = \int f(x, y)\ d(\mu \otimes \nu)\ (x, y)$$

Indfører vi udtrykket for h, udtrykker sætningen, at

$$\int (\int f(x, y)\ d\nu(y))\ d\mu(x) = \int f(x, y)\ d(\mu \otimes \nu)\ (x, y)$$

dvs. at det er tilladt at udføre integrationen ved succesiv integration. □

Sætningen kaldes Tonellis sætning.

Sætning A.1.6 ***Fubinis sætning***

Lad μ og ν være mål på hhv $(E, \mathcal{A})$ og $(F, \mathcal{B})$, og lad f være en funktion på $E \times F$, der er integrabel med hensyn til $(E \times F, \mathcal{A} \otimes \mathcal{B})$. Da gælder,

a) funktionen $g : F \to \mathbb{R}$ defineret ved

$$g(y|x) = f(x, y) \quad \text{for } y \in F,\ x \in E$$

være integrabel med hensyn til ν for alle $x \in E$, pånær evt. en mængde med μ-målet nul.

b) funktionen $h : E \to \mathbb{R}$ defineret ved

$$h(x) = \int f(x, y)\ d\nu(y) \quad \text{for } x \in E$$

er integrabel med hensyn til μ og

$$\int \Big(\int f(x, y)\ d\nu(y) \Big)\ d\mu(x) = \int f(x, y)\ d(\mu \otimes \nu)\ (x, y)$$

□

Sætningen udtrykker, at for integrable funktioner (dvs funktioner for hvilke integralerne eksisterer), da kan integration med hensyn til et produktmål udføres ved successive integrationer med hensyn til de enkelte komponenter, og at rækkefølgen af disse integrationer er ligegyldig.

Definition A.1.11 ***Sandsynlighedsfelt***

Lad $(\Omega, \mathcal{A})$ være et måleligt rum, og lad P være et mål på $(\Omega, \mathcal{A})$ for hvilket det gælder at

$$\mathsf{P}\{\Omega\} = 1$$

Sættet $(\Omega, \mathcal{A}, \mathsf{P})$ kaldes da et sandsynlighedsfelt (jvf. Jørsboe p.8). □

Vi bemærker, at ethvert mål, μ, der tilfredsstiller $\mu\{\Omega\} < \infty$, kan omformes til et sandsynlighedsmål P ved

$$\mathsf{P}\{A\} = \frac{1}{\mu\{\Omega\}} \mu\{A\} \quad \text{for } A \in \mathcal{A}$$

Et sandsynlighedsmål på $(\Omega, \mathcal{A})$ kaldes også en sandsynlighedsfordeling, eller blot en fordeling på $(\Omega, \mathcal{A})$.

Definition A.1.12 *Stokastisk variabel*

Lad $(\Omega, \mathcal{A}, \mathsf{P})$ være et sandsynlighedsfelt og lad $(E, \mathcal{B})$ være et måleligt rum.

En stokastisk variabel X er en målelig afbildning $X : (\Omega, \mathcal{A}) \to (E, \mathcal{B})$.

Det transformerede mål, $P = X(\mathsf{P})$ kaldes fordelingen for den stokastiske variable.

Vi bemærker, at såfremt specielt $E = \mathbb{R}$ og $\mathcal{B}$ er Borelsigmaalgebraen på $\mathbb{R}$, da svarer definitionen til Jørsboe (1988) p. 19. □

Ovenstående mere generelle definition gør det muligt at formulere alle udsagn om sandsynlighedsmål som udsagn om stokastiske variable.

Traditionelt har man brugt betegnelsen stokastisk variabel for at sondre mellem sandsynlighedsfeltet $(\Omega, \mathcal{A}, \mathsf{P})$, hvor den målelige afbildning er defineret, og afbildningen ind i de reelle tal, hvor man kan foretage sædvanlig integration eller summation. Ved at benytte den mere generelle mål- og integralteori, kan vi tillade os lade værdimængden at være et vilkårligt måleligt rum, og det er derfor ikke så væsentligt at skelne mellem det underliggende sandsynlighedsfelt $(\Omega, \mathcal{A}, \mathsf{P})$ og billedet $(E, \mathcal{B}, (X\mathsf{P})$.

Specielt kan man opfatte et sandsynlighedsmål P som fordelingen af en stokastisk variabel. Hvis nemlig $(\Omega, \mathcal{A}, \mathsf{P})$ er et sandsynlighedsfelt, da er den identiske afbildning af Ω på sig selv en målelig afbildning med fordeling P.

Formuleringen "Lad X være en stokastisk variabel med fordeling P " skal således opfattes som betydende: Vi betragter et sandsynlighedsfelt $(E, \mathcal{A}, P)$, hvor P opfattes som fordelingen af en stokastisk variabel.

Ovenstående sætninger om mål kan derfor formuleres som udsagn om fordelinger af stokastiske variable.

A.2 Referencer

Jørsboe, O.G. (1984): *Sandsynlighedsregning*, Matematisk Institut, DTH

Indeks