SANNOLIKHETSTEORI

TILLÄGGSMATERIAL TILL BLOM, BOK A

Våren 2001

LUNDS UNIVERSITET

Matematikcentrum
Matematisk statistik

CENTRUM SCIENTIARUM MATHEMATICARUM

SANNOLIKHETSTEORI

TILLÄGGSMATERIAL TILL BLOM, BOK A

Innehåll

1 Dubbelintegraler, partiella derivator, m.m. 1
1.1 Riemannintegralen . . . 1
1.2 Itererade enkelintegraler. . . . 3
1.3 Partiella derivator. . . . 5
1.4 Variabelsubstitution(*). . . . 6

2 Några fördelningar 9
2.1 Negativ binomialfördelning . . . 9

3 Simulering och slumptalsgenerering 11
3.1 Vad är simulering? . . . 11
3.2 Allmänt om slumptalsgenerering . . . 12
3.3 Likformigt fördelade slumptal . . . 13
3.4 Slumptal från andra fördelningar . . . 13
3.4.1 Inversa transformationsmetoden. . . . 13
3.4.2 Funktioner av stokastiska variabler. . . . 15
3.4.3 Rejektionsmetoden. . . . 18

4 Mera om Markovkedjor 20
4.1 Beständiga och icke-beständiga tillstånd . . . 20
4.2 Partitionering av tillståndsrummet(*) . . . 22
4.3 Stationär(a) fördelning(ar) . . . 23
4.4 Asymptotisk fördelning . . . 25
4.5 Övergångstider . . . 27
4.6 Övningsuppgifter . . . 32

5 Diskreta Markovprocesser i kontinuerlig tid 35
5.1 Definition. Övergångs- och intensitetsmatriser. . . . 35
5.2 Tidsberoende hos övergångs- och absoluta sannolikheter.(*) . . . 38
5.3 Stationära sannolikheter. . . . 40
5.4 Irreducibilitet och asymptotiska fördelningar. . . . 41
5.5 Inbäddad Markovkedja, tid mellan hopp. . . . 42
5.6 Övergångstider . . . 44
5.7 Övningsuppgifter . . . 45

6 Tillförlitlighet 48
6.1 Inledning . . . 48
6.2 Icke-underhållna system . . . 48
6.2.1 Funktionssannolikhet, intensitetsfunktion och förväntad livslängd . . . 48

6.2.2 Redundans 50
6.3 Underhållna system 52
6.4 Övningsuppgifter 54

7 Svar till övningsuppgifter **56**

Kapitel 1

Dubbelintegraler, partiella derivator, m.m.

1.1 Riemannintegralen

Låt $f : \mathbb{R}^2 \to \mathbb{R}$ vara en icke-negativ funktion av två variabler. Vi vill beräkna den s.k. dubbelintegralen

$$V = \iint\limits_{(x,y)\in A} f(x,y)\,dx\,dy,$$

över ett område $A \subset \mathbb{R}^2$. Geometriskt tolkar vi $0 \leq V \leq \infty$ som volymen av området

$$B = \{(x,y,z);\ (x,y) \in A, 0 \leq z \leq f(x,y)\}$$

i $\mathbb{R}^3$, där vi tänker oss grafen av f inritad som en yta i ett tredimensionellt koordinatsystem.

För att ge en formell definition av dubbelintegralen, ska vi betrakta den s.k. Riemannintegralen. Vi börjar med att anta att $A = [a,b] \times [c,d]$ är en rektangel ($a < b, c < d$). Precis som för enkelintegraler, kan vi definiera V som gränsvärdet av Riemannsummor. Givet ett positivt heltal n, delas A först in i n^2 lika stora rektanglar

$$A_{ij}^n = \left[a + \frac{i-1}{n}(b-a),\ a + \frac{i}{n}(b-a)\right] \times \left[c + \frac{j-1}{n}(d-c),\ c + \frac{j}{n}(d-c)\right], \quad 1 \leq i,j \leq n.$$

Välj därefter, för varje n, en följd

$$\Delta_n = \{(x_{(i,j)}^n, y_{(i,j)}^n);\ (x_{(i,j)}^n, y_{(i,j)}^n) \in A_{ij}^n, 1 \leq i,j \leq n\}$$

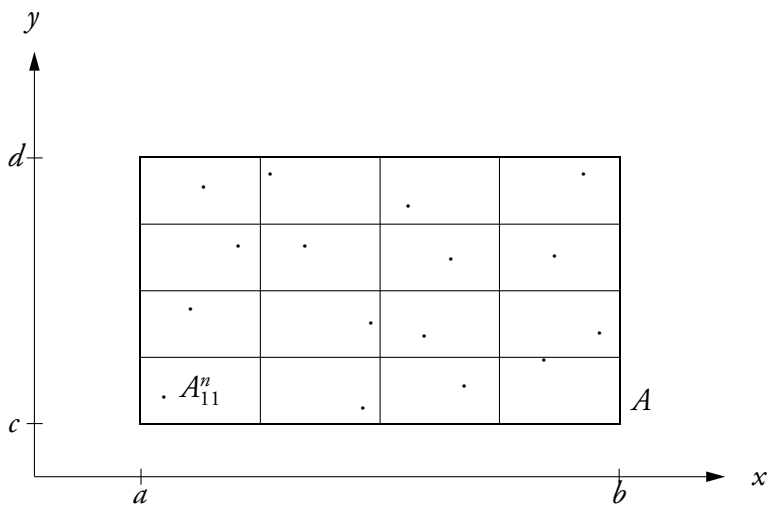
av n^2 punkter från de olika rektanglarna (se Figur 1.1) samt definiera Riemannsumman

$$V(\Delta_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} f(x_{(i,j)}^n,\, y_{(i,j)}^n)\frac{(b-a)(d-c)}{n^2}. \tag{1.1}$$

Tydligen är $V(\Delta_n)$ summan av volymerna för ett antal rätblock med basyta A_{ij}^n och höjd $f(x_{(i,j)}^n,\, y_{(i,j)}^n)$. Vi säger att Riemannintegralen av f över området A existerar om

$$\lim_{n\to\infty} V(\Delta_n) = V$$

oavsett valet av punkter $(x_{(i,j)}^n,\, y_{(i,j)}^n)$. (Alternativt kan man använda så kallade över- och undersummor, men det leder till precis samma definition.)



Figur 1.1: *Rektangeln* $A = [a,b] \times [c,d]$ *uppdelad i* $n^2 = 4^2$ *rektanglar* A_{ij}^n *med en tänkbar punktmängd* Δ_n *i vilken* $f(x,y)$ *skall evalueras.*

För en kompakt mängd $A_{\text{komp}} \subset \mathbb{R}^2$ kan vi först hitta en rektangel $A \supseteq A_{\text{komp}}$, och sedan definiera

$$\iint\limits_{(x,y)\in A_{\text{komp}}} f(x,y)\,dx\,dy = \iint\limits_{(x,y)\in A} g(x,y)\,dx\,dy,$$

med

$$g(x,y) = \begin{cases} f(x,y), & (x,y) \in A_{\text{komp}} \\ 0, & (x,y) \in A \setminus A_{\text{komp}}. \end{cases}$$

Slutligen, för en godtycklig mängd $A \subseteq \mathbb{R}^2$, kan vi definiera en växande följd av följd $A_{\text{komp}}^1 \subseteq A_{\text{komp}}^2 \subseteq \ldots \subseteq A_{\text{komp}}^n \ldots$ av kompakta mängder så att $\cup_{n\geq 1} A_{\text{komp}}^n = A$. (Ta exempelvis $A_{\text{komp}}^n = A \cap [-n,n] \times [-n,n]$.) Låt sedan

$$V = \iint\limits_{(x,y)\in A} f(x,y)\,dx\,dy = \lim_{n\to\infty} \iint\limits_{(x,y)\in A_{\text{komp}}^n} f(x,y)\,dx\,dy.$$

Om f är Riemannintegrerbar över vart och ett av de kompakta områdena A_{komp}^n så är högerledet gränsvärdet av en icke-avtagande talföljd.

Riemannintegralen existerar ($0 \leq V \leq \infty$) exempelvis för alla kontinuerliga funktioner, samt för funktioner som är kontinuerliga så när som på ett ändligt antal linjer och cirkelbågar.

Slutligen nämner vi att Riemannintegralen definieras på samma sätt även då f tillåts anta negativa värden (då kan emellertid f inte vara en tvådimensionell täthetsfunktion). Man ska bara komma ihåg att i (1.1) blir $f(x_{(i,j)}^n, y_{(i,j)}^n) \cdot (b-a)(d-c)/n^2$ volymen av ett rätblock med omvänt tecken då $f(x_{(i,j)}^n, y_{(i,j)}^n) < 0$, så i definitionen av $V(\Delta_n)$ kommer både positiva och negativa termer att ingå. Tar vi gränsvärdet $n \to \infty$ kan området A delas upp i två delområden

$$A_+ = \{(x,y);\ f(x,y) \geq 0\} \quad \text{och} \quad A_- = \{(x,y);\ f(x,y) < 0\}$$

beroende på tecknet hos f. Det visar sig att

$$\iint\limits_{(x,y)\in A} f(x,y)\,dx\,dy = \iint\limits_{(x,y)\in A_+} f(x,y)\,dx\,dy - \iint\limits_{(x,y)\in A_-} (-f(x,y))\,dx\,dy,$$

där de två termerna i högerledet svarar mot volymen av den del av området mellan f och xy-planet där f är positiv respektive negativ. Eftersom dessa båda volymer subtraheras med varandra kan $\iint_A f(x,y)\,dx\,dy$ bli noll även om f ej är identiskt lika med noll.

1.2 Itererade enkelintegraler.

I praktiken är definitionen av V alldeles för bökig att använda för att räkna ut dubbelintegraler. Med hjälp av Fubinis sats kan V beräknas med itererade enkelintegraler. Definiera, för fixa x_0 och y_0, snitten

$$I_{x_0} = \{y;\ (x_0, y) \in A\} \quad \text{och} \quad I_{y_0} = \{x;\ (x, y_0) \in A\}.$$

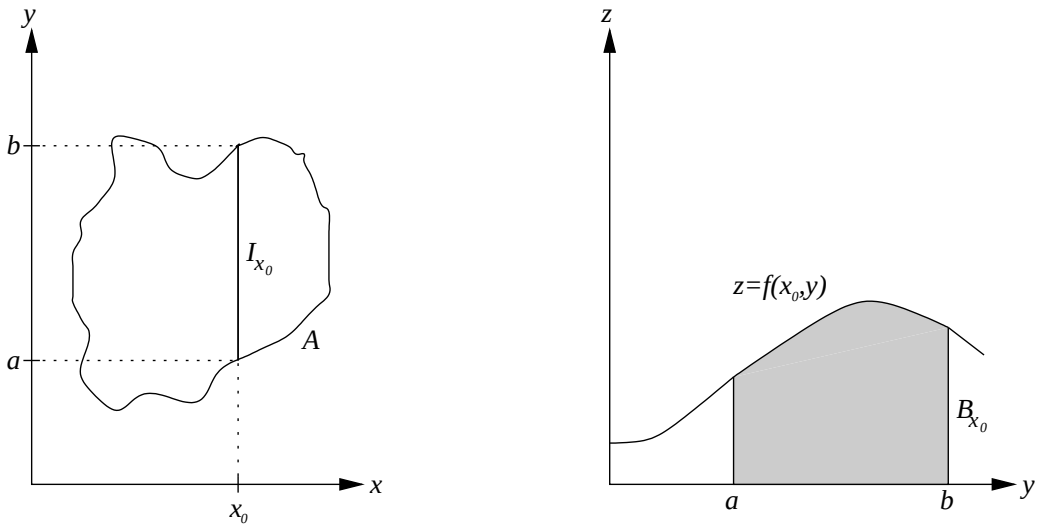
Vi kan tolka

$$\int_{y \in I_{x_0}} f(x,y)\,dy \quad \text{och} \quad \int_{x \in I_{y_0}} f(x,y)\,dx$$

som arean av områdena

$$B_{x_0} = B \cap \{x = x_0\} \quad \text{respektive} \quad B_{y_0} = B \cap \{y = y_0\}.$$

Vi ser att B_{x_0} är ett snitt av B taget i punkten x_0, ungefär som en oändligt smal brödskiva utskuren ur en limpa längs yz-planet (se Figur 1.2). Området B_{y_0} svarar mot att man skär åt andra hållet, dvs längs xz-planet.



Figur 1.2: *Integrationsområdet A med skivan $I_{x_0} = [a,b]$ (vänster) och skivans yta $B_{x_0} = \{(x,y,z);\ x = x_0,\ a \le y \le b,\ 0 \le z \le f(x_0,y)\}$ (höger).*

Fubinis sats säger nu att vi kan räkna ut volymen V genom att integrera arean av B_x med avseende på x eller arean av B_y med avseende på y. För att återgå till liknelsen med limpan, så är dess volym lika med summan av volymerna hos ett stort antal tunna utskurna skivor.

Sats 1.1 (Fubinis sats för positiva funktioner.) *Om $f \geq 0$ är Riemannintegrerbar över A, så gäller*

$$\begin{aligned} V &= \int \text{area}(B_x)\, dx = \int \left(\int_{y \in I_x} f(x,y)\, dy \right) dx \\ &= \int \text{area}(B_y)\, dy = \int \left(\int_{x \in I_y} f(x,y)\, dx \right) dy. \end{aligned}$$

□

Exempel 1.1 (Oberoende stokastiska variabler.) Låt $f = f_{X,Y} = f_X f_Y$ vara den simultana täthetsfunktionen för två oberoende kontinuerliga stokastiska variabler. Om $A = I_1 \times I_2$ fås

$$\begin{aligned} \mathsf{P}\left((X,Y) \in A\right) &= \iint_A f_{X,Y}(x,y)\, dx\, dy = \int_{x \in I_1} \left(\int_{y \in I_2} f_X(x) f_Y(y)\, dy \right) dx \\ &= \int_{x \in I_1} f_X(x)\, dx \int_{y \in I_2} f_Y(y)\, dy = \mathsf{P}\left(X \in I_1\right) \mathsf{P}\left(Y \in I_2\right). \end{aligned}$$

□

I fortsättningen kommer vi att utelämna parenteser och underförstå att vi avser en itererad integral. I exemplet ovan hade vi $I_x \equiv I_1$ och $I_y \equiv I_2$. Ofta är det är det enklare att bestämma integrationsområdena åt ena hållet:

Exempel 1.2 Låt X och Y ha simultan täthetsfunktion

$$f_{X,Y}(x,y) = \begin{cases} 2\mathrm{e}^{-x-y}, & 0 \leq x \leq y \\ 0, & \text{annars.} \end{cases}$$

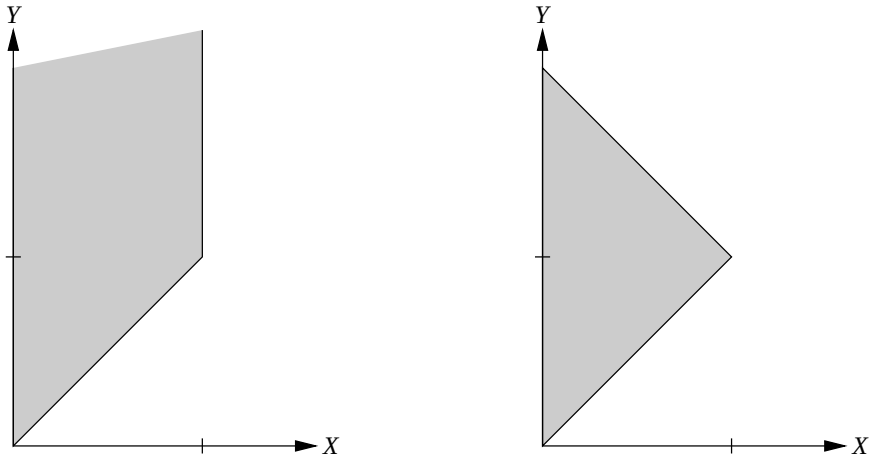
Låt oss beräkna $\mathsf{P}(X \leq 1)$. Enligt förutsättningarna skall vi integrera funktionen $2\mathrm{e}^{-x-y}$ över området $\{x \leq y\} \cap \{0 \leq x \leq 1\}$ (jfr. Figur 1.3). Vi får

$$\mathsf{P}(X \leq 1) = \int_0^1 \int_x^\infty 2\mathrm{e}^{-x-y}\, dy\, dx = \int_0^1 2\mathrm{e}^{-2x}\, dx = 1 - \mathrm{e}^{-2}.$$

Om vi i stället väljer att integrera med avseende på x i den inre integralen får vi dela upp i två fall beroende på om y är större eller mindre än 1:

$$\begin{aligned} \mathsf{P}(X \leq 1) &= \int_0^1 \int_0^y 2\mathrm{e}^{-x-y}\, dx\, dy + \int_1^\infty \int_0^1 2\mathrm{e}^{-x-y}\, dx\, dy \\ &= \int_0^1 2\mathrm{e}^{-y}(1 - \mathrm{e}^{-y})\, dy + \int_1^\infty 2(1 - \mathrm{e}^{-1})\mathrm{e}^{-y}\, dy \\ &= 2(1 - \mathrm{e}^{-1}) - (1 - \mathrm{e}^{-2}) + 2(1 - \mathrm{e}^{-1})\mathrm{e}^{-1} = 1 - \mathrm{e}^{-2}. \end{aligned}$$

□



Figur 1.3: *Integrationsområdena* $\{(x,y);\ 0 \le x \le 1,\ x \le y\}$ *och* $\{(x,y);\ 0 \le x \le 1,\ x \le y \le 2-x\}$.

Om vi vill beräkna $\mathsf{P}(Y \le 2 - X)$ $(= \mathsf{P}(X \le Y \le 2 - X))$ visar det sig också lämpligast att börja integrera över $I_x = [x, 2-x]$ (se Figur 1.3). Vi har

$$\mathsf{P}(Y \le 2 - X) = \int_0^1 \int_x^{2-x} 2\mathrm{e}^{-x-y}\, dy\, dx = \int_0^1 2(\mathrm{e}^{-2x} - \mathrm{e}^{-2})\, dx = 1 - 3\mathrm{e}^{-2}.$$

Exempel 1.3 (Beräkning av väntevärden genom betingning.(*)) I Blom A sid 138–139 bevisas formeln

$$\mathsf{E}(X) = \mathsf{E}\big[\mathsf{E}(X \mid Y)\big] \tag{1.2}$$

för diskreta stokastiska variabler. Man kan alltså beräkna $\mathsf{E}(X)$ genom att först beräkna det betingade väntevärdet $\mathsf{E}(X \mid Y = y)$ för alla y. Eftersom $\mathsf{E}\big[\mathsf{E}(X \mid Y)\big]$ blir en funktion av Y är den i sig en stokastisk variabel, som enligt (1.2) har samma väntevärde som X. För kontinuerliga stokastiska variabler kan vi bevisa (1.2) på följande sätt:

$$\begin{aligned}
\mathsf{E}(X) &= \int x f_X(x)\, dx = \iint x f_{X,Y}(x,y)\, dy\, dx = \iint x f_{X,Y}(x,y)\, dx\, dy \\
&= \iint x \frac{f_{X,Y}(x,y)}{f_Y(y)}\, dx\, f_Y(y)\, dy = \iint x f_{X|Y}(x \mid y)\, dx\, f_Y(y)\, dy \\
&= \int \mathsf{E}(X \mid Y = y) f_Y(y)\, dy = \mathsf{E}\big[\mathsf{E}(X \mid Y)\big],
\end{aligned}$$

där vi i tredje ledet utnyttjade Fubinis sats och i det femte ledet formeln $f_{X|Y}(x \mid y) = f_{X,Y}(x,y)/f_Y(y)$ för betingade täthetsfunktioner. □

1.3 Partiella derivator.

För en funktion $F : \mathbb{R}^2 \to \mathbb{R}$ definierar vi den *partiella derivatan* av F med avseende på x som

$$\frac{\partial F(x,y)}{\partial x} = \lim_{h \to 0} \frac{F(x+h,y) - F(x,y)}{h},$$

dvs den vanliga derivatan om vi fryser y och varierar x. Analogt fås den partiella derivatan med avseende på y som

$$\frac{\partial F(x,y)}{\partial y} = \lim_{h \to 0} \frac{F(x,y+h) - F(x,y)}{h}.$$

Precis som endimensionella funktioner kan F deriveras flera gånger. Andra ordningens partiella derivata med avseende på x och y definieras exempelvis som

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = \lim_{h\to 0} \frac{\partial F(x,y+h)/\partial x - \partial F(x,y)/\partial x}{h}.$$

Vidare kan vi definiera $\partial^2 F/\partial x \partial x = \partial^2 F/\partial x^2$ genom att derivera med avseende på x två gånger, och analogt $\partial^2 F/\partial y \partial y = \partial^2 F/\partial y^2$ och $\partial^2 F/\partial y \partial x$. För de mixade partiella derivatorna kan man visa att det inte spelar någon roll i vilken ordningsföljd deriveringen sker, dvs

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = \frac{\partial^2 F(x,y)}{\partial y \partial x}.$$

Exempel 1.4 (Simultan täthetsfunktion.) Låt $F = F_{X,Y}$ vara den simultana fördelningsfunktionen för två kontinuerliga stokastiska variabler, dvs

$$F_{X,Y}(x,y) = \iint\limits_{u\le x,\, v\le y} f_{X,Y}(u,v)\, du\, dv = \int\limits_{u=-\infty}^{x} \int\limits_{v=-\infty}^{y} f_{X,Y}(u,v)\, dv\, du.$$

Med hjälp av den sista omskrivningen (Fubinis sats) kan vi lättare derivera partiellt med avseende på x,

$$\frac{\partial F(x,y)}{\partial x} = \int\limits_{v=-\infty}^{y} f_{X,Y}(x,v)\, dv,$$

och ytterligare en partiell derivering med avseende på y ger

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = f_{X,Y}(x,y).$$

Givet $F_{X,Y}$ får vi alltså $f_{X,Y}$ genom att derivera partiellt två gånger. □

1.4 Variabelsubstitution(*).

Många gånger är en dubbelintegral enklare att beräkna om vi först byter variabler, från (x,y) till (u,v). Vi antar att funktionen $(u,v) \to (x(u,v),\ y(u,v))$ är inverterbar, samt att dess komponenter är partiellt deriverbara. För att kunna byta integrationsvariabler behöver vi införa den så kallade *Jacobianen*

$$J(u,v) = \begin{vmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{vmatrix}.$$

Formeln för variabelsubstitution blir nu

$$\iint\limits_A f(x,y)\, dx\, dy = \iint\limits_{A'} f(x(u,v),\ y(u,v))\, |J(u,v)|\, du\, dv,$$

med

$$A' = \{(u,v);\ \ (x(u,v),\ y(u,v)) \in A\}.$$

Anledningen till att Jacobianen dyker upp i formeln är att (formellt skrivet)

$$|J(u,v)|\, du\, dv = dx\, dy.$$

Ett litet volymselement $dx\, dy$ är alltså faktorn $|J(u,v)|$ gånger större än $du\, dv$. Vi tänker oss att $dx\, dy$ och $du\, dv$ är positiva, därav beloppstecknet kring $J(u,v)$. Jämför med det endimensionella fallet, då $dx = |x'(u)|\, du$.

Exempel 1.5 (Rayleighfördelning.) Anta att $X, Y \in N(0,\,1)$ är två oberoende s.v. Vi vill beräkna fördelnings- och täthetsfunktion för $R = \sqrt{X^2 + Y^2}$ och börjar med

$$F_R(z) = \mathsf{P}(X^2 + Y^2 \le z^2) = \iint\limits_A f_X(x) f_Y(y)\,dx\,dy = \frac{1}{2\pi} \iint\limits_A \mathrm{e}^{-(x^2+y^2)/2}\,dx\,dy, \quad z \ge 0,$$

där integrationsområdet ges av cirkelskivan

$$A = \{(x, y);\, x^2 + y^2 \le z^2\}.$$

Här är det fördelaktigt att gå över till polära koordinater, $x = v\cos u$ och $y = v\sin u$, med

$$J(v, u) = \begin{vmatrix} \cos\vartheta & -v\sin u \\ \sin u & v\cos u \end{vmatrix} = v,$$

och

$$A' = [0, z^2] \times [0, 2\pi].$$

Formeln för variabelsubstitution och Fubinis sats ger nu

$$\begin{aligned} F_R(z) &= \iint\limits_{A'} v f_X(v\cos u) f_Y(v\sin u)\,dv\,du \\ &= \frac{1}{2\pi}\int_0^z v\mathrm{e}^{-v^2/2}\int_0^{2\pi} du\,dv = \int_0^z v\mathrm{e}^{-v^2/2}\,dv = 1 - \mathrm{e}^{-z^2/2}. \end{aligned}$$

Derivering med avseende på z ger slutligen

$$f_R(z) = z\mathrm{e}^{-z^2/2}. \tag{1.3}$$

Detta är en så kallad Rayleighfördelning, och den tillhör familjen av Weibullfördelningar. □

Det går också att räkna ut den *simultana täthetsfunktionen* $f_{U,V}(u, v)$, genom att utnyttja formeln för variabelsubstitution. Låt

$$A = \{(x, y);\, (x, y) = (x(u, v),\, y(u, v)) \text{ för något } u \le u_0,\, v \le v_0\}.$$

Vi får

$$\begin{aligned} F_{U,V}(u_0, v_0) &= \mathsf{P}(U \le u_0,\, V \le v_0) = \mathsf{P}\big((X, Y) \in A\big) = \iint\limits_A f_{X,Y}(x, y)\,dx\,dy \\ &= \iint\limits_{u\le u_0,\, v\le v_0} |J(u, v)| f_{X,Y}(x(u, v),\, y(u, v))\,du\,dv = \iint\limits_{u\le u_0,\, v\le v_0} f_{U,V}(u, v)\,du\,dv. \end{aligned}$$

Vi har alltså erhållit

$$f_{U,V}(u, v) = |J(u, v)| f_{X,Y}(x(u, v),\, y(u, v)). \tag{1.4}$$

Exempel 1.6 (Faltningsformeln.) Låt X och Y vara två kontinuerliga s.v. med given simultan täthetsfunktion $f_{X,Y}$. Vi vill härleda faltningsformeln, dvs bestämma täthetsfunktionen f_V för summan $V = X + Y$. Byt variabler till

$$u = x \quad \text{och} \quad v = x + y.$$

Invertering av denna variabeltransformation ger

$$x = u \quad \text{och} \quad y = v - u \quad \text{med} \quad J(u,v) = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

Formeln (1.4) ger nu

$$f_{U,V}(u,v) = f_{X,Y}(u, v-u).$$

Slutligen erhåller vi den marginella tätheten f_V genom integrering av $f_{U,V}$:

$$f_V(v) = \int_{-\infty}^{\infty} f_{U,V}(u,v)\,du = \int_{-\infty}^{\infty} f_{X,Y}(u, v-u)\,du.$$

□

Exempel 1.7 (Täthet för kvot av s.v.) Låt X och Y vara två oberoende kontinuerliga s.v. med givna täthetsfunktioner f_X och f_Y, samt $Y > 0$. Låt oss nu, med samma metod som i föregående exempel, bestämma f_V för $V = X/Y$. Vi inför

$$u = y \quad \text{och} \quad v = x/y.$$

Invertering ger

$$x = uv \quad \text{och} \quad y = u \quad \text{med} \quad J(u,v) = \begin{vmatrix} v & u \\ 1 & 0 \end{vmatrix} = u.$$

Utnyttjar vi nu (1.4) erhålls

$$f_{U,V}(u,v) = u f_X(uv) f_Y(u),$$

och efter integrering,

$$f_V(v) = \int_0^{\infty} f_{U,V}(u,v)\,du = \int_0^{\infty} u f_X(uv) f_Y(u)\,du.$$

□

Kapitel 2

Några fördelningar

2.1 Negativ binomialfördelning

Vid ett experiment har händelsen A sannolikheten p att inträffa. Låt X beteckna antalet oberoende misslyckade försök som utförs *innan* A inträffar för r:te gången. Vi kan då skriva

$$X = Y_1 + Y_2 + \ldots + Y_r,$$

där Y_i är antalet misslyckade försök som utförs mellan det att A inträffar för $(i-1)$:ta och i:te gången. Eftersom alla försök är oberoende, blir också $Y_1, \ldots, Y_r$ oberoende stokastiska variabler med en geometrisk fördelning, dvs om $q = 1 - p$,

$$p_{Y_i}(k) = q^k p, \quad k = 0, 1, 2, \ldots$$

Genom att använda faltningsformeln upprepade gånger och ett induktionsresonemang kan man visa att

$$p_X(k) = \binom{k+r-1}{k} p^r q^k = \binom{k+r-1}{r-1} p^r q^k, \qquad k = 0, 1, 2, \ldots, \tag{2.1}$$

vilket ger den så kallade *negativa binomialfördelningen*.

Det finns ett direkt kombinatoriskt resonemang som leder till (2.1). Låt varje försöksserie representeras av en binär följd, där en etta innebär att A inträffar. Då blir X antalet nollor innan den r:te ettan inträffar, exempelvis $X = 8$ om $r = 3$ för sekvensen

0 1 0 0 1 0 0 0 0 0 1

Om $X = k$ så kommer den r:te ettan att ha position $k + r$ (den föregås av $r - 1$ ettor och k nollor). Det finns tydligen $\binom{k+r-1}{k}$ sätt att placera ut k nollor och $r - 1$ ettor i de $k + r - 1$ första positionerna. Sannolikheten för varje sådan sekvens med längd $k + r$ blir $p^r q^k$ eftersom de olika försöken är oberoende. Slutligen är händelserna svarande mot olika sådana sekvenser disjunkta, varför $\mathsf{P}(X = k)$ ges av (2.1).

Om vi istället intresserar oss för totala antalet försök Z som utförs inklusive det r:te lyckade, dvs

$$Z = X + r,$$

får vi

$$p_Z(k) = p_X(k-r) = \binom{k-1}{r-1} p^r q^{k-r}, \quad k = r, r+1, r+2, \ldots.$$

Även denna fördelning kallas negativ binomialfördelning, vilket ibland orsakar viss förvirring. För att inte sammanblanda de två fördelningarna säger vi att Z har en *för r:te gången fördelning* (frg), vilket svarar mot distinktionen mellan geometriska och ffg-fördelningar i Blom A. Notera att

$$Z = (Y_1 + 1) + \ldots + (Y_r + 1),$$

och eftersom $Y_i + 1 \in \mathit{ffg}(p)$ så är Z summan av r oberoende och $\mathit{ffg}(p)$-fördelade stokastiska variabler.

Kapitel 3

Simulering och slumptalsgenerering

3.1 Vad är simulering?

Simulering är en mycket vanlig teknik att lösa problem som är så komplicerade att en analytisk metod är svårgenomförbar. Vi inleder med ett exempel för att beskriva förfarandet:

Exempel 3.1 (Simulering av kömodell.) Vi vill bestämma antalet kassor m som bör vara bemannade i en affär. Vi ställer upp en slumpmodell som innebär att kunder anländer med tidsavstånd $X_1, X_2, \ldots$, som är oberoende s.v. med en viss fördelning F_X samt att betjäningstiderna $Y_1, Y_2, \ldots$ för kund nummer $1, 2, \ldots$ är oberoende s.v. med en annan fördelning F_Y. För varje m ställer vi upp kostnadsfunktionen

$$g(m) = am + b\mathsf{E}(W), \quad m = 1, 2, 3, \ldots$$

där W är väntetiden (kötid + betjäning) för en slumpmässigt vald kund, och a och b är positiva konstanter. (Givetvis kommer fördelningen för W och därmed även $\mathsf{E}(W)$ att bero av m.) Vi vill hitta det m som minimerar $g(m)$. Den första termen i $g(m)$ svarar mot kostnaden att ha många kassor i funktion, och den andra mot kostnaden av att kunder får vänta för länge (så att de kanske inte återvänder till samma affär). I vissa specialfall kan detta problem lösas analytiskt (t.ex. då F_X och F_Y är exponentialfördelade), men i allmänhet är detta mycket svårt. Vi kan då simulera ankomst och betjäning för n stycken kunder, vilket innebär att slumpvariabler $\{X_i\}_{i=1}^n$ och $\{Y_i\}_{i=1}^n$ genereras, samt en mekanism som anger hur kunder väljer kassa (exempelvis kassan med kortast kö, och om flera kassor har samma kölängd väljs slumpmässigt en av dessa). Sedan uppskattas $g(m)$ för olika m med

$$g^*(m) = am + \frac{b}{n}\sum_{i=1}^{n} W_i,$$

dvs väntevärdet $\mathsf{E}(W)$ har ersatts med den genomsnittliga väntetiden för de kunder som ingår i simuleringsundersökningen. Slutligen väljer vi det m som minimerar g^*. □

Vi kan nu formalisera förfarandet i exemplet med följande *schema för en simuleringsundersökning*:

1. Formulera det praktiska problemet.
2. Bygg en slumpmodell.
3. Formulera en eller flera funktioner g av intresse.
4. Dimensionera experimentet n.

5. Generera n uppsättningar av slumptal från de fördelningar som ingår i modellen.
6. Beräkna g för varje omgång.
7. Sammanställ de beräknade g-värdena och analysera dem.
8. Drag praktiska slutsatser.

Det praktiska problemet är ofta formulerat som en fråga: ”Hur många kassor ska vara bemannade?”. Många gånger är det fråga om att testa hypoteser, t.ex. ”vilket av två typer av vinterdäck ger säkrast väghållning?”, eller, om uppskattning av effekter, ”vilken dos bör ges av ett läkemedel för bäst behandling?”. Problemformuleringen är mycket viktig, eftersom den påverkar valet av slumpmodell och sedan analys och slutsatser.

Det karakteristiska för en simuleringsundersökning är punkt 5 — generering av slumptal. I ett traditionellt experiment har man i stället insamlade data att tillgå. Vi kommer i de senare avsnitten att uppehålla oss vid slumptalsgenerering.

En viktig aspekt hos simuleringsexperiment är repeterbarheten hos experimentet, dvs repeterbarhet hos slumptalsgenereringen. Möjligheten att upprepa en slumpmässig sekvens är viktig av minst två orsaker:

1. Det är lättare att finna fel i simuleringsprogram om man exakt kan repetera en tidigare körning.
2. Jämförelser mellan två eller flera alternativa metoder kan göras mer precist om varje metod kan simuleras under samma slumpmässiga händelser.

3.2 Allmänt om slumptalsgenerering

Det finns olika sätt att generera synbarligen slumpmässiga tal.

1. Ett sätt att generera slumptal är med fysikaliska metoder. Utgången från en synbarligen slumpmässig mekanism, t.ex. utfallet av ett myntkast eller tärning, utgången hos en räknare av kosmisk strålning eller den minst signifikanta siffran (eller bits) hos den digitala klockan i en dator. Andra exempel på sådana är om man observerar brus och låter nivån på bruset (lämpligt normerat) utgöra slumpvariablens värde. Nackdelarna med dessa är att de i regel är ganska långsamma och de är svåra att repetera. Dessutom kan det vara svårt att bedöma huruvida de är slumpmässiga eller ej.

2. Ett annat sätt att generera likformigt fördelade (eller synonymt: rektangelfördelade) slumptal är att använda tabellverk över tal vars slumpmässighet och likformighet i intervallet $(0, 1)$ har validerats. Exempel på sådana tabellverk med slumptal är exempelvis Rand Corporation (1955). *A Million Random Digits with 100 000 Normal Deviates*. Free Press: Glencoe, Ill. En klar nackdel med detta förfaringssätt är att proceduren blir rätt långsam.

3. Ett tredje sätt är att betrakta så kallade *pseudoslumptal*, dvs tal som genereras med en deterministisk rekursiv algoritm och är approximativt likformigt fördelade. Metoden är snabb, och har dessutom fördelen av repeterbarhet. Vidare kan slumpmässigheten hos metoderna kan prövas. I nästa avsnitt beskrivs pseudoslumptalsgeneratorer mer ingående.

I de flesta programmeringsspråk finns tillgång till slumptalsgeneratorer. I vissa fall har man tillgång till en hel uppsättning generatorer som genererar slumptal från olika fördelningar. I andra fall har man åtminstone tillgång till en slumpgenerator som genererar likformigt fördelade slumptal i intervallet $(0, 1)$. I ett senare avsnitt skall vi se hur man med utgångspunkt från sådana slumptal kan generera slumptal från andra fördelningar.

Det är också viktigt att kontrollera eller validera de valda slumptalen, t.ex. om de har rätt fördelning och om de är oberoende. Sådana testmetoder behandlas inom ramen för inferensteorin, så kallade *goodness-of-fit*-tester respektive oberoendetester.

3.3 Likformigt fördelade slumptal

Som nämnts genereras pseudoslumptal genom rekursiva deterministiska algoritmer, men på ett sådant sätt att de får slumpliknande egenskaper. En sådan algoritm är *linjära kongruensmetoden*: För ett givet startvärde x_0 och givna icke-negativa konstanter a, c och m definieras

$$x_{i+1} = (ax_i + c) \mod m$$

dvs om k_i är heltalsdelen av $(ax_i + c)/m$ så ges x_{i+1} av

$$x_{i+1} = ax_i + c - mk_i.$$

Genom att normera den på detta sätt uppkomna sekvensen $\{x_i\}$ med m får vi en sekvens $\{u_i\}$, där

$$u_i = x_i/m$$

som alla ligger mellan 0 och 1 och med egenskaper liknande oberoende och rektangelfördelade slumptal mellan 0 och 1.

Den rekursiva algoritmen genererar en sekvens av tal som efter en viss längd upprepas. Längden hos en sådan cykel (perioden) beror på hur a, c och m väljs.

Man kan visa[1] att korrelationskoefficienten mellan x_i och x_{i+1} är begränsad till intervallet

$$(\frac{1}{a} - \frac{6c}{am}(1 - \frac{c}{m}) - \frac{a}{m}, \frac{1}{a} - \frac{6c}{am}(1 - \frac{c}{m}) + \frac{a}{m})$$

För stora värden på a och m ($m \gg a$) kommer således den seriella korrelationen att vara nära noll.

Goda egenskaper hos genererade slumptal har erhållits med $m = 2^{35}$, $a = 2^7 + 1$ och $c = 1$ för binära datorer och också med $m = 2^b$, $a = 101$ och $c = 1$ för decimala datorer med ordlängd b.

Specialfallet $c = 0$ ger den s.k. *multiplikativa kongruensgeneratorn*

$$x_{i+1} = ax_i \mod m$$

$$u_i = x_i/m$$

Goda resultat har här erhållits för $a = 16807$ och $m = 2^{31} - 1$.

3.4 Slumptal från andra fördelningar

3.4.1 Inversa transformationsmetoden.

Låt U vara en s.v. sådan att $U \in R(0, 1)$. Då gäller att $\mathsf{P}(U \leq x) = x$, $0 \leq x \leq 1$. Låt F vara en godtycklig fördelningsfunktion med inversen F^{-1}. (Om F inte är kontinuerlig och strängt växande så är inte inversen väldefinierad/entydigt bestämd. Om vi skriver

$$F^{-1}(u) = \sup\{x;\, F(x) \leq u\} \tag{3.1}$$

[1] Greenberger, M. (1961), An a priori determination of serial correlation in computer generated random numbers, Mathematics of Computation, 15, 383–389

så existerar alltid F^{-1}, och den sammanfaller med den vanliga inversen när denna existerar entydigt.) Bilda nu

$$X = F^{-1}(U).$$

Då är[2] X en s.v. med fördelningsfunktion F.

Resultatet kan användas för att generera slumptal från en godtycklig fördelning. Metoden som använder detta samband kallas *inversa transformationsmetoden*. Den är mest effektiv om inversen till fördelningsfunktionen F kan uttryckas explicit.

Exempel 3.2 (Exponentialfördelade slumptal.) Låt $X \in Exp(m)$ dvs fördelningsfunktionen ges av

$$F_X(x) = \begin{cases} 1 - \mathrm{e}^{-x/m} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Den inversa funktionen ges av

$$x = F_X^{-1}(u) = -m \ln(1-u)$$

Om vi således har rektangelfördelade slumptal $u_1, u_2, \ldots, u_n$ i intervallet $(0, 1)$ kan vi åstadkomma slumptal $x_1, x_2, \ldots, x_n$ genom

$$x_i = -m \ln(1-u_i)$$

som är fördelade enligt F_X. Eftersom $1 - u_i$ också är rektangelfördelad i $(0, 1)$ kan man ännu enklare bilda exponentialfördelade slumptal genom

$$x_i = -m \ln u_i$$

□

Exempel 3.3 (Rayleighfördelade slumptal.) En Rayleighfördelning har fördelningsfunktion $F(z) = 1 - \mathrm{e}^{-z^2/2}$. Den inversa funktionen blir

$$x = F^{-1}(u) = \sqrt{-2\ln(1-u)}.$$

Precis som i föregående exempel kan vi ersätta u med $1 - u$. Om $U \in R(0,\ 1)$ så har $\sqrt{-2 \ln U}$ en Rayleighfördelning. □

Vi ska nu använda inversa transformationsmetoden för diskreta stokastiska variabler. Fördelningsfunktionen F_X för en diskret s.v. är trappstegsformad, dvs

$$F_X(x) = \sum_{j=1}^{k} p_X(x_j), \quad x_k \leq x < x_{k+1},$$

om $p_X(x_j) = \mathsf{P}(X = x_j)$ och $x_1, \ldots, x_p, \ldots$ är de möjliga värdena på X. Det betyder att (den generaliserade) inversen (3.1) ges av

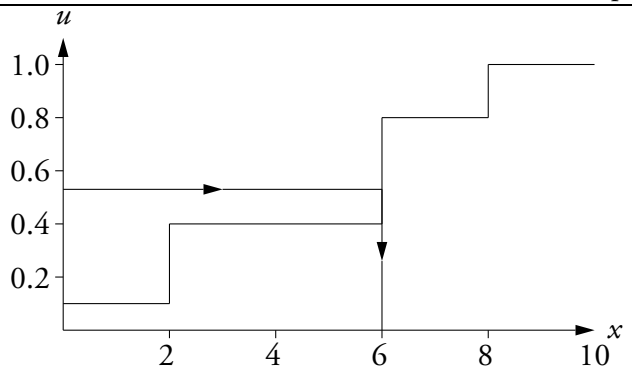
$$x = F_X^{-1}(u) = x_k$$

om

$$p_X(x_1) + p_X(x_2) + \cdots + p_X(x_{k-1}) \leq u < p_X(x_1) + p_X(x_2) + \cdots + p_X(x_k), \tag{3.2}$$

se Figur 3.1.

[2]Om F är strikt växande gäller att $\mathsf{P}(X \leq y) = \mathsf{P}(F^{-1}(U) \leq y) = \mathsf{P}(U \leq F(y)) = F(y)$ där sista ledet följer av att $U \in R(0,\ 1)$ och således $\mathsf{P}(U \leq z) = z$. Således gäller att $F_X(y) = \mathsf{P}(X \leq y) = F(y)$. För allmänt F (icke strängt växande) är beviset något längre.



Figur 3.1: *Illustrering av inversa transformationsmetoden för en fyrpunktsfördelning med* $p_X(0) = 0.1$, $p_X(2) = 0.3$, $p_X(6) = 0.4$ *och* $p_X(8) = 0.2$. *Slumptalet* $u = 0.53$ *ger* $x = 6$.

Exempel 3.4 (Poissonfördelade slumptal.) Låt X vara en Poissonfördelad s.v. med parameter (väntevärde) m. Enligt (3.2) ges ett poissonfördelat slumptal av k där

$$\sum_{j=0}^{k-1} \mathrm{e}^{-m} \frac{m^j}{j!} \leq u < \sum_{j=0}^{k} \mathrm{e}^{-m} \frac{m^j}{j!}$$

och u är ett rektangelfördelat slumptal från $(0, 1)$. □

Exempel 3.5 (Binomialfördelade slumptal.) Låt X vara binomialfördelad s.v. med parametrar n och p. Enligt (3.2) ges ett binomialfördelat slumptal av k där

$$\sum_{j=0}^{k-1} \binom{n}{j} p^j (1-p)^{n-j} \leq u < \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}$$

och u är ett rektangelfördelat slumptal från $(0, 1)$[3]. □

Exempel 3.6 (Geometriskt fördelade slumptal.) Om $X \in Ge(p)$ så har fördelningsfunktionen ett explicit uttryck; $F_X(k) = 1 - (1-p)^{k+1}$, $k = 0, 1, 2, \ldots$ Vi väljer slumptalet k om $1 - (1-p)^k \leq u < 1 - (1-p)^{k+1}$, vilket kan skrivas

$$k = \left[\frac{\ln(1-u)}{\ln(1-p)} \right],$$

där $[x]$ anger heltalsdelen av x. Eftersom $1 - u$ också är ett $R(0,\ 1)$-slumptal får vi ännu enklare

$$k = \left[\frac{\ln(u)}{\ln(1-p)} \right]$$

som ett geometriskt fördelat slumptal. Jämförelse med Exempel 3.2 ger för övrigt $k = [y]$, där y är ett slumptal från $Exp(1/\ln((1-p)^{-1}))$. □

3.4.2 Funktioner av stokastiska variabler.

Om en s.v. X kan uttryckas som en funktion av andra stokastiska variabler $Y_1, Y_2, \ldots, Y_p$ dvs

$$X = g(Y_1, Y_2, \ldots, Y_p)$$

[3]Ett alternativt och egentligen enklare sätt att generera $X \in Bin(n, p)$ är att simulera n stycken slantsinglingar genom användning av n stycken rektangelfördelade slumptal $u_1, \ldots, u_n$.

och vi har möjlighet att generera slumptal med samma fördelning som $Y_1, Y_2, \ldots, Y_p$ så kan vi erhålla slumptal med samma fördelning som X genom

$$x = g(y_1, y_2, \ldots, y_p)$$

där $y_1, y_2, \ldots, y_p$ är slumptal genererade med fördelning enligt $Y_1, Y_2, \ldots, Y_p$.

Exempel 3.7 (Normalfördelade slumptal enligt Box-Müller.) Om U_1 och U_2 är rektangelfördelade i $(0, 1)$ och vi definierar

$$X_1 = \sqrt{-2 \ln U_2}\, \cos(2\pi U_1) \tag{3.3}$$

$$X_2 = \sqrt{-2 \ln U_2}\, \sin(2\pi U_1) \tag{3.4}$$

så kan man visa att X_1 och X_2 oberoende standardiserat normalfördelade slumptal[4]. Vi kan erhålla normalfördelade slumptal från $N(m,\ \sigma)$ genom att bilda

$$X_1 = m + \sigma\sqrt{-2 \ln U_2}\, \cos(2\pi U_1)$$

$$X_2 = m + \sigma\sqrt{-2 \ln U_2}\, \sin(2\pi U_1)$$

□

På grund av att de rektangelfördelade slumptalen genereras som pseudoslumptal med någon deterministisk rekursiv algoritm har det visat sig sämre att generera enligt

$$X_1 = \sqrt{-2 \ln U_1}\, \cos(2\pi U_2)$$

$$X_2 = \sqrt{-2 \ln U_1}\, \sin(2\pi U_2)$$

än enligt (3.3)–(3.4), där U_2 genereras direkt efter U_1.

Skall ett större antal slumptal genereras kan det vara lämpligt att modifiera Box-Müllers metod så att man slipper de tidskrävande trigonometriska funktionsberäkningarna. En snabbare variant är Marsaglias metod.

Exempel 3.8 (Normalfördelade slumptal enligt Marsaglia.) Låt Z_1 och Z_2 vara oberoende rektangelfördelade slumptal på intervallet $(-1, 1)$. Om $Z_1^2 + Z_2^2 \leq 1$ får vi två oberoende slumptal X_1 och X_2 från $N(0,\ 1)$ genom att sätta

$$X_1 = Z_1\sqrt{\frac{-2\ln(Z_1^2 + Z_2^2)}{Z_1^2 + Z_2^2}}$$

$$X_2 = Z_2\sqrt{\frac{-2\ln(Z_1^2 + Z_2^2)}{Z_1^2 + Z_2^2}}$$

men då $Z_1^2 + Z_2^2 > 1$ genererar man nya Z_1, Z_2 tills villkoret $Z_1^1 + Z_2^2 \leq 1$ är uppfyllt[5]. □

[4] Observera att $\sqrt{X_1^2 + X_2^2} = \sqrt{-2 \ln U_2}$. Jämförelse med Exempel 3.3 ger att $\sqrt{X_1^2 + X_2^2}$ är Rayleighfördelad.

[5] Genom att förkasta de slumptalspar som hamnar utanför enhetscirkeln har vi på ett enkelt sätt erhållit likformigt fördelade slumptal Z_1, Z_2 över cirkelytan. När vi har ett slumptalspar Z_1, Z_2 innanför cirkelytan får vi i polära koordinater att $(Z_1, Z_2) = (R\cos\Theta,\ R\sin\Theta)$ där $R^2 = Z_1^2 + Z_2^2$. Man kan visa att fördelningen för vinkeln Θ blir $R(0,\ 2\pi)$ och att den kvadrerade radien R^2 blir $R(0,\ 1)$. De blir också oberoende. Då Θ och R^2 är fördelade på detta sätt blir $\sqrt{-2\ln(R^2)}\, \cos\Theta$ och $\sqrt{-2\ln(R^2)}\, \sin\Theta$ normalfördelade enligt Exempel 3.7. Eftersom vi kan skriva

$$X_1 = \sqrt{-2\ln(R^2)}\, \cos\Theta = \sqrt{\frac{-2\ln(R^2)}{R^2}}\, R\cos\Theta = Z_1\sqrt{\frac{-2\ln(Z_1^2 + Z_2^2)}{Z_1^2 + Z_2^2}}$$

och motsvarade för X_2 har vi därmed visat resultatet.

Tidsvinsten man gör genom att slippa beräkna cosinus resp sinus kompenserar mer än väl den förlust vi gör genom att i genomsnitt kasta bort andelen $1-\pi/4$ av de genererade rektangelfördelade slumptalen. I simuleringsundersökningar är dock oftast inte själva slumptalsgenereringen den mest tidskrävande beräkningen.

Exempel 3.9 (Approximativt normalfördelade slumptal med CGS) Om $X_1, X_2, \ldots, X_n$ är oberoende och alla fördelade som $R(0,\ 1)$ så är

$$\frac{X_1 + X_2 + \cdots + X_n - n/2}{\sqrt{n/12}}$$

approximativt fördelad som $N(0,\ 1)$ enligt centrala gränsvärdessatsen. Summor av rektangelfördelade variabler närmar sig tämligen snabbt normalfördelningen och om man sätter $n = 12$ får man att

$$y = x_1 + x_2 + \cdots + x_{12} - 6$$

är ett slumptal som är approximativt $N(0,\ 1)$-fördelat. □

Exempel 3.10 (Lognormalfördelade slumptal.) Om Y är normalfördelad enligt $N(m,\ \sigma)$ så gäller att

$$X = \mathrm{e}^Y$$

är lognormalfördelad enligt $\Lambda(m,\ \sigma)$, vilket innebär att

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma x}\,\mathrm{e}^{-(\ln x - m)^2/2\sigma^2}, \quad x > 0.$$

□

Med hjälp av exponentialfördelade slumptal kan vi bilda gammafördelade slumptal:

Exempel 3.11 (Gammafördelade slumptal.) Låt $X_i \in \mathit{Exp}(a)$, $i = 1, 2, \ldots, n$. Då gäller att $\sum_{i=1}^{n} X_i \in \Gamma(n,\ a)$. Gammafördelade slumptal $\Gamma(n, a)$, (för heltal n) kan således bildas som

$$y_i = -\sum_{i=1}^{n} a \ln(u_i)$$

där u_i är rektangelfördelad i $(0, 1)$, och således $-\ln(u_i)$ är $\mathit{Exp}(1)$ (se Exempel 3.2). □

Exempel 3.12 (Weibullfördelade slumptal.) Om $Y \in \mathit{Weibull}(a,\ c)$ så gäller att $(Y/a)^c \in \mathit{Exp}(1)$. Weibullfördelade slumptal kan således genereras genom

$$y = a(-\ln(1-u))^{1/c}$$

där u är rektangelfördelad i $(0, 1)$, och således $-\ln(1-u)$ är $\mathit{Exp}(1)$ (se Exempel 3.2). Även $a(-\ln u)^{1/c}$ är Weibullfördelad $\mathit{Weibull}(a,\ c)$. (Pröva även att härleda detta resultat med inversa transformationsmetoden med hjälp av $F_Y(x) = 1 - \mathrm{e}^{-(x/a)^c}$, $x \geq 0$.) □

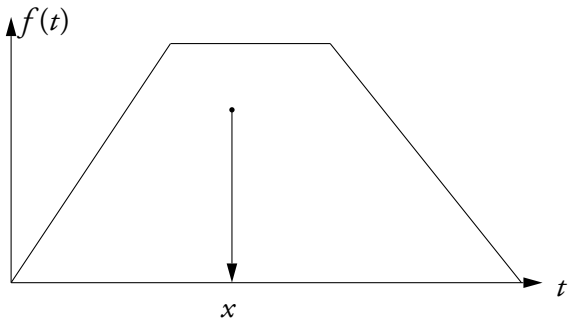
3.4.3 Rejektionsmetoden.

Ibland är rejektionsmetoden ett bättre alternativ än metoderna som beskrivits ovan då man vill generera slumptal från en fördelning med täthetsfunktionen $f(t)$. Metoden är enkel att beskriva: Välj en punkt slumpvis i området som begränsas av t-axeln och täthetsfunktionen $f(t)$ enligt Figur 3.2. Slumptalet x väljs som första koordinaten hos den valda punkten.

Att metoden fungerar följer av att slumptalet X har fördelningsfunktion

$$F_X(a) = \mathsf{P}(X \leq a) = \frac{\text{Ytan till vänster om } a}{\text{Totala ytan}} = \frac{\int_{-\infty}^{a} f(x)\,dx}{1} = F(a).$$

Svårigheten är att hitta en metod som väljer en punkt slumpvis under kurvan. Man utgår från en fördelning från vilken det är lätt att generera slumptal och som har täthetsfunktion $g(x)$ som uppfyller $f(x) \leq Mg(x)$ för alla x och någon konstant M. En slumpmässigt vald punkt under kurvan $Mg(x)$ erhålls genom att ta ett slumptal x från fördelningen som ges av $g(x)$ och ett rektangelfördelat slumptal y i intervallet $(0, M)$ för att få punkten $(\mathsf{x}, \mathsf{y}g(\mathsf{x}))$. Om denna punkt också ligger under kurvan $f(x)$ dvs om $f(\mathsf{x}) \geq \mathsf{y}g(\mathsf{x})$, väljs x som slumptal, i annat fall upprepas proceduren tills detta inträffar. Av beskrivningen förstår man att täthetsfunktionen $g(x)$ skall väljas så att konstanten M blir så liten som möjligt för att minimera antalet slumptal som behöver användas.



Figur 3.2: *Välj en punkt slumpmässigt (likformigt) från området under täthetsfunktionen f. Projektionen x av den valda punkten blir ett slumptal från en fördelning med täthetsfunktion f.*

Exempel 3.13 Vi önskar generera ett slumptal från en s.v. X med täthetsfunktion

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & -1 \leq x \leq 1 \\ 0 & \text{för övrigt} \end{cases}$$

Rejektionsmetoden blir speciellt lätt då den s.v. endast antar värden i ett begränsat intervall (a, b), eftersom $g(x)$ då kan väljas som

$$g(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{för övrigt} \end{cases}$$

I detta fall är $a = -1$ och $b = 1$ varför vi väljer $g(x) = 1/2$, $-1 < x < 1$ och $M = 4/\pi$. Inversmetoden ovan ger $\mathsf{x} = (b-a)u + a$ där u är ett slumptal från $R(0,\ 1)$. Slumptalen $u_1 = 0.08959$ och $u_2 = 0.49051$ ger därför värdena $\mathsf{x} = 2u_1 - 1 = 0.82082$ respektive $\mathsf{y} = 4u_2/\pi = 0.62454$, och punkten $(-0.82082,\ 0.31227)$. Eftersom $f_X(-0.82082) = 0.36363 \geq 0.31227$ ligger punkten under kurvan och slumptalet väljs som -0.82082. □

Exempel 3.14 Vi önskar ett slumptal från den s.v. $Y = |X|$, där X är en standardiserad normalfördelad variabel. Täthetsfunktionen för Y erhålls som

$$f_Y(x) = \begin{cases} \sqrt{\frac{2}{\pi}}\, e^{-x^2/2} & x \geq 0 \\ 0 & \text{för övrigt} \end{cases}$$

Eftersom $x^2 - 2x + 1 = (x-1)^2 \geq 0$ följer olikheten

$$\sqrt{\frac{2}{\pi}}\, e^{-x^2/2} \leq \sqrt{\frac{2}{\pi}}\, e^{(1-2x)/2}$$

Vi kan alltså välja $g(x) = e^{-x}$, $x \geq 0$ och $M = \sqrt{\frac{2}{\pi}}\, e^{1/2}$. Exponentialfördelade slumptal framställdes i Exempel 3.2 och därur får vi att ur två slumptal u_1 och u_2 från $R(0,\ 1)$ erhålls den slumpvis valda punkten $(-\ln(1-u_1),\ u_2\sqrt{\frac{2}{\pi}}\, e^{1/2}(1-u_1))$. Slumptalen $u_1 = 0.08959$ och $u_2 = 0.49051$ ger punkten $(0.09386,\ 0.58745)$ och eftersom $f_Y(0.09386) = 0.79438 \geq 0.58745$ har vi även här haft turen att få en punkt under kurvan, och vi kan välja 0.09386 som slumptal från fördelningen f_Y ovan. □

Kapitel 4

Mera om Markovkedjor

4.1 Beständiga och icke-beständiga tillstånd

Låt $f_{ii}^{(n)}$ vara sannolikheten för att processen vid tidpunkt n för första gången efter starten i E_i åter befinner sig i E_i,

$$f_{ii}^{(n)} = \mathsf{P}(X(k+1) \neq E_i,\ \ldots,\ X(k+n-1) \neq E_i,\ X(k+n) = E_i \,|\, X(k) = E_i).$$

Då är

$$f_{ii} = \sum_{n=1}^{\infty} f_{ii}^{(n)}$$

sannolikheten för att processen vid start i E_i förr eller senare åter befinner sig i E_i. Beständiga tillstånd karakteriseras ju av att $f_{ii} = 1$, dvs man återvänder säkert till E_i. Tiden för återvändande betecknas Y_{ii} och kallas *återkomsttiden*. Man kan dela upp beständiga tillstånd i två kategorier, beroende på om Y_{ii} har ändligt väntevärde eller ej:

Definition 4.1 *Ett beständigt tillstånd kallas positivt beständigt om*

$$\mu_i = \mathsf{E}(Y_{ii}) = \sum_{n=1}^{\infty} n f_{ii}^{(n)} < \infty.$$

Om $\mathsf{E}(Y_{ii}) = \infty$ *sägs tillståndet vara nollbeständigt.* □

Vi har alltså delat upp tillstånden i tre kategorier: Obeständiga, nollbeständiga och positivt beständiga.

Exempel 4.1 (Typer av tillstånd.) Låt $\{X(n),\ n \geq 0\}$ vara en Markovkedja med tillstånd E_1, E_2, E_3, E_4 och övergångsmatris

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1/3 & 2/3 & 0 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix}.$$

I detta fall är $f_{44}^{(n)} = 0$ för alla n så tillstånd E_4 är ett obeständigt tillstånd. För tillstånd E_3 gäller att $f_{33}^{(1)} = \frac{2}{3}$ och $f_{33}^{(n)} = 0$ för $n \geq 2$ så tillstånd E_3 är obeständigt. För tillstånd E_1 och E_2 gäller att

$$f_{11} = f_{11}^{(1)} + f_{11}^{(2)} = \frac{1}{2} + \frac{1}{2} = 1$$

$$f_{22} = f_{22}^{(1)} + f_{22}^{(2)} + \cdots + f_{22}^{(k)} + \cdots = 0 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = 1.$$

och är således beständiga tillstånd. Den förväntade återkomsttiden för dessa tillstånd är

$$\mu_1 = \sum_{n=1}^{\infty} n f_{11}^{(n)} = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2}$$

och

$$\mu_2 = \sum_{n=1}^{\infty} n f_{22}^{(n)} = \sum_{n=1}^{\infty} n \frac{1}{2^{n-1}} = 3.$$

Således är såväl tillstånd E_1 som E_2 positivt beständiga. □

Exempel 4.2 (Slumpvandring.) En partikel startar i origo och rör sig sedan vid varje tidpunkt $n = 1, 2, 3, \ldots$ antingen uppåt eller nedåt, med sannolikhet p respektive $q = 1 - p$, $0 \leq p \leq 1$. De olika förflyttningarna är oberoende av varandra. Positionen vid tidpunkt n ges av

$$X(n) = \sum_{i=1}^{n} Y_i,$$

där Y_i är oberoende och likafördelade s.v. med $p_{Y_i}(-1) = q$ och $p_{Y_i}(1) = p$. Detta är en Markovkedja med tillstånd $\ldots, E_{-1}, E_0, E_1, \ldots$. Av symmetriskäl inses att $f_{ii}^{(n)}$ inte beror av i, eftersom slumpvandringen hela tiden startar om från där den befinner sig. Man kan visa (exempelvis med sannolikhetsgenererande funktioner) att

$$f_{ii} = \begin{cases} 1, & p = 1/2 \\ 1 - \sqrt{1 - 4pq}, & p \neq 1/2 \end{cases}$$

och

$$\mathsf{E}(Y_{ii}) = \sum_{n=1}^{\infty} n f_{ii}^{(n)} = \infty, \quad p = 1/2.$$

I en symmetrisk slumpvandring ($p = 1/2$) är således alla tillstånd nollbeständiga, dvs vi kommer säkert tillbaka från en given startpunkt, även om det tar oändligt lång tid i genomsnitt. För icke-symmetriska slumpvandringar ($p \neq 1/2$) är alla tillstånd obeständiga, dvs det finns en positiv sannolikhet att vi aldrig kommer tillbaka.

Med stora talens lag kan vi ge en motivering: För stora n gäller $X(n)/n \approx \mathsf{E}(Y_1) = p - q$. Om $p > 1/2$ så är $p - q$ positiv och $X(n)$ driver iväg mot ∞ med hastighet $p - q$ i genomsnitt. På samma sätt driver $X(n)$ iväg mot $-\infty$ om $p < 1/2$. Slutligen kommer, om $p = 1/2$, fördelningen för $X(n)$ vara centrerad kring 0, men oscillationerna är så pass stora att det tar oändligt lång tid i genomsnitt att återvända till origo.

Man kan visa att för en symmetrisk slumpvandring i d-dimensioner (dvs varje koordinat beskriver en slumpvandring enligt ovan med $p = 1/2$) så kommer alla tillstånd att vara nollbeständiga om $d = 1$ eller 2, och obeständiga om $d \geq 3$. En slumpvandring i planet återvänder alltså alltid till en given punkt, medan en slumpvandring i rummet med positiv sannolikhet aldrig återvänder. Intuitivt beror detta på att högre dimensioner innehåller mer rymd än lägre. □

4.2 Partitionering av tillståndsrummet(*)

Låt

$$\mathcal{G} = \{E_1, E_2, \ldots\}$$

beteckna hela tillståndsrummet (ändligt eller uppräkneligt oändligt). Det visar sig att för en allmän Markovkedja kan $\mathcal{G}$ delas upp i komponenter på ett mycket intuitivt sätt. Vi behöver först följande definition:

Definition 4.2 *En mängd av tillstånd $\mathcal{G}_1 \subseteq \mathcal{G}$ kallas* sluten *om $p_{ij} = 0$ så snart $E_i \in \mathcal{G}_1$ och $E_j \notin \mathcal{G}_1$ och irreducibel om alla tillstånd i $\mathcal{G}_1$ kommunicerar tvåsidigt med varandra.* □

Det går alltså inte att ta sig ut ur en sluten mängd av tillstånd när man väl har hamnat där. Vidare kan en irreducibel mängd av tillstånd inte delas upp i mindre, slutna komponenter. Det visar sig att irreducibla och slutna komponenter har stor betydelse för att förstå teorin för Markovkedjor. Vi inleder med följande resultat:

Sats 4.1 (Irreducibla och slutna komponenter.) *Alla tillstånd i en irreducibel och sluten komponent är av samma slag, dvs antingen obeständiga, nollbeständiga eller positivt beständiga.* □

Följande sats ger en uppdelning av $\mathcal{G}$ i slutna och irreducibla komponenter med beständiga tillstånd samt en (ej nödvändigtvis irreducibel eller sluten) komponent med obeständiga tillstånd:

Sats 4.2 (Dekompositionssatsen) *Tillståndsrummet $\mathcal{G}$ kan delas upp i disjunkta delmängder enligt*

$$\mathcal{G} = (\mathcal{A}_1 \cup \mathcal{A}_2 \cup \ldots) \cup (\mathcal{B}_1 \cup \mathcal{B}_2 \cup \ldots) \cup \mathcal{C},$$

där $\mathcal{C}$ består av obeständiga tillstånd, $\mathcal{A}_j$ är en irreducibel, sluten mängd av positivt beständiga tillstånd, och $\mathcal{B}_j$ är en irreducibel, sluten mängd av nollbeständiga tillstånd. □

Bevisskiss. Dela upp tillståndsrummet enligt $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{C}$, där $\mathcal{G}_1$ innehåller alla beständiga tillstånd och $\mathcal{C}$ alla obeständiga tillstånd. Vi inför så relationen

$$E_i \leftrightarrow E_j \text{ om } E_i \text{ och } E_j \text{ kommunicerar tvåsidigt }, \forall E_i, E_j \in \mathcal{G}_1.$$

Det visar sig att $\leftrightarrow$ blir en så kallad ekvivalensrelation, och enligt ett fundamentalt resultat för ekvivalensrelationer kan $\mathcal{G}_1$ partitioneras i komponenter inom vilka samtliga element kommunicerar tvåsidigt men inga tillstånd från olika komponenter kommunicerar tvåsidigt. Sats 4.1 ger slutligen att varje komponent består av antingen enbart positivt beständiga eller också enbart nollbeständiga tillstånd. □

Om vi startar Markovkedjan i $\mathcal{C}$ kommer vi antingen att stanna i $\mathcal{C}$ (då måste $\mathcal{C}$ innehålla oändligt många tillstånd), eller också hamnar vi så småningom i någon av de slutna, irreducibla mängderna $\mathcal{A}_j$ eller $\mathcal{B}_j$ och stannar sedan där. En schematisk illustration av dekompositionssatsen ges i följande exempel:

Exempel 4.3 (Blockuppdelning av övergångsmatrisen.) Anta för enkelhets skull att det finns vardera två komponenter av positivt betständiga och nollbeständiga tillstånd, dvs $\mathcal{G} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup B_1 \cup \mathcal{B}_2 \cup \mathcal{C}$. Då kan övergångsmatrisen delas upp i block enligt

$$P = \begin{pmatrix} P_{\mathcal{A}_1\mathcal{A}_1} & 0 & 0 & 0 & 0 \\ 0 & P_{\mathcal{A}_2\mathcal{A}_2} & 0 & 0 & 0 \\ 0 & 0 & P_{\mathcal{B}_1\mathcal{B}_1} & 0 & 0 \\ 0 & 0 & 0 & P_{\mathcal{B}_2\mathcal{B}_2} & 0 \\ P_{\mathcal{C}\mathcal{A}_1} & P_{\mathcal{C}\mathcal{A}_2} & P_{\mathcal{C}\mathcal{B}_1} & P_{\mathcal{C}\mathcal{B}_2} & P_{\mathcal{C}\mathcal{C}} \end{pmatrix},$$

där $P_{\mathcal{A}_1\mathcal{A}_1}$ är den delmatris av P som innehåller alla övergångssannolikheter mellan tillstånd i $\mathcal{A}_1$ osv. Nollan för blocket i position $(1,2)$ innebär exempelvis att $P_{\mathcal{A}_1\mathcal{A}_2} = 0$, dvs för varje $E_i \in \mathcal{A}_1$ och $E_j \in \mathcal{A}_2$ gäller $p_{ij} = 0$. □

En ändlig irreducibel Markovkedja har $\mathcal{C} = \emptyset$ och endast en sluten komponent. Vidare kan man visa att tillstånden i denna komponent måste vara positivt beständiga. Mer allmänt gäller följande:

Sats 4.3 *Följande egenskaper gäller för en Markovkedja:*

(i) *Om en irreducibel och sluten komponent $\mathcal{B}_i$ med obeständiga tillstånd existerar har den oändligt många element.*

(ii) *För varje ändlig Markovkedja kan tillståndsrummet delas upp enligt*

$$\mathcal{G} = (\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_m) \cup \mathcal{C},$$

med samma beteckningar som i Sats 4.2 och $m \geq 1$, dvs alla tillstånd kan inte vara obeständiga.

(iii) *För en irreducibel och ändlig Markovkedja är samtliga tillstånd positivt beständiga, dvs $\mathcal{G} = \mathcal{A}_1$.*

□

Den svåra delen av Sats 4.3 är (i), sedan följer (ii) och (iii) direkt av (i) och Dekompositionssatsen. Notera att (iii) skärper något det som står på sid 235 i Blom A — att samtliga tillstånd i en irreducibel, ändlig Markovkedja är beständiga.

Exempel 4.4 (Forts. av Exempel 4.1.) Vi har uppdelningen

$$\mathcal{G} = \{E_1, E_2\} \cup \{E_3, E_4\} = \mathcal{A}_1 \cup \mathcal{C}.$$

Eftersom $\mathcal{C} \neq \emptyset$ är Markovkedjan inte irreducibel, men $\mathcal{A}_1$ är en irreducibel komponent. Vidare är kedjan ändlig, och då kan den enligt Sats 4.3 inte innehålla någon irreducibel och sluten komponent $\mathcal{B}_j$ med nollbeständiga tillstånd. Detta stämmer med vad vi fann i Exempel 4.1; μ_1 och μ_2 var båda ändliga. □

En irreducibel komponent av nollbeständiga tillstånd, $\mathcal{B}_i$, måste enligt Sats 4.2 och 4.3 vara oändlig. Ett exempel på detta är symmetriska slumpvandringar:

Exempel 4.5 (Slumpvandring, forts.) För en symmetrisk slumpvandring ($p = 1/2$) i en dimension gäller

$$\mathcal{G} = \{\ldots, E_{-1}, E_0, E_1, \ldots\} = \mathcal{B}_1,$$

dvs tillståndsrummet är irreducibelt och består av nollbeständiga tillstånd. Om i stället $p \neq 1/2$ får vi $\mathcal{G} = \mathcal{C}$, eftersom alla tillstånd är obeständiga. □

4.3 Stationär(a) fördelning(ar)

För en Markovkedja med flera slutna och irreducibla komponenter existerar ofta flera stationära fördelningar.

Exempel 4.6 Betrakta en Markovkedja med tre tillstånd och övergångsmatris

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

I detta fall har vi två slutna och irreducibla mängder med positivt beständiga tillstånd

$$\mathcal{G} = \{E_1, E_2\} \cup E_3 = \mathcal{A}_1 \cup \mathcal{A}_2.$$

(Att tillstånden måste vara positivt beständiga följer av Sats 4.3 eller med liknande resonemang som i Exempel 4.1.) Man ser att $\pi_{\mathcal{A}_1} = (0.5, 0.5, 0)$ och $\pi_{\mathcal{A}_2} = (0, 0, 1)$ är två olika stationära fördelningar svarande mot att vi säkert befinner oss i $\mathcal{A}_1$ respektive $\mathcal{A}_2$. Vidare kommer alla konvexa linjärkombinationer

$$(1-t)\pi_{\mathcal{A}_1} + t\pi_{\mathcal{A}_2}, \quad 0 < t < 1$$

att ge stationära fördelningar, eftersom

$$\big((1-t)\pi_{\mathcal{A}_1} + t\pi_{\mathcal{A}_2}\big)\, P = (1-t)\pi_{\mathcal{A}_1}P + t\pi_{\mathcal{A}_2}P = (1-t)\pi_{\mathcal{A}_1} + t\pi_{\mathcal{A}_2}.$$

Markovkedjan har alltså oändligt många stationära fördelningar! □

Mer allmänt gäller följande för existens av stationära fördelningar:

Sats 4.4 (Beskrivning av samtliga stationära fördelningar.(*)) *Betrakta en Markovkedja med uppdelning av tillståndsrummet enligt Sats 4.2. Då finns till varje irreducibel och sluten komponent av positivt beständiga tillstånd $\mathcal{A}_i$ precis en stationär fördelning $\pi_{\mathcal{A}_i} = \{\pi_{\mathcal{A}_i,j};\ E_j \in \mathcal{G}\}$ med all sannolikhetsmassa i $\mathcal{A}_i$, och den ges av*

$$\pi_{\mathcal{A}_i,j} = \begin{cases} \mu_j^{-1}, & E_j \in \mathcal{A}_i \\ 0, & E_j \notin \mathcal{A}_i, \end{cases} \tag{4.1}$$

där μ_j är den förväntade återkomsttiden (se Definition 4.1) till E_j. Mängden av alla stationära fördelningar till Markovkedjan är

$$\{\pi = \sum_i t_i \pi_{\mathcal{A}_i};\ t_i \geq 0, \sum_i t_i = 1\},$$

där $\mathcal{A}_i$ genomlöper alla irreducibla och slutna komponenter med positivt beständiga tillstånd i summan ovan.

Sats 4.4 har en mängd viktiga konsekvenser, vilka vi formulerar separat i en följdsats:

Följdsats 4.5 *För en Markovkedja gäller följande:*

(i) *Det finns noll, en respektive oändligt många stationära fördelningar beroende på om antalet irreducibla och slutna komponenter $\mathcal{A}_i$ med positivt beständiga tillstånd är noll, ett respektive större än ett.(*)*

(ii) *En irreducibel Markovkedja med positivt beständiga tillstånd har en entydig stationär fördelning $\pi = (\pi_1, \pi_2, \ldots)$ med komponenter $\pi_i = 1/\mu_i$ för alla $E_i \in \mathcal{G}$, där μ_i är den förväntade återkomsttiden för E_i.*

(iii) *En irreducibel Markovkedja med antingen nollbeständiga eller obeständiga tillstånd saknar stationär fördelning.*

□

Egenskapen (ii) i Följdsats 4.5 är intuitivt rimlig: Ju längre tid det tar att återvända till E_i, desto mindre är sannolikheten π_i att befinna sig i E_i vid jämvikt.

Exempel 4.7 (Forts av Exempel 4.1.) Vi bestämmer den stationära fördelningen. Ekvationssystemet

$$\begin{cases} \pi = \pi P \\ \sum_i \pi_i = 1 \end{cases} \quad \Leftrightarrow \quad \begin{cases} \pi_1 = \frac{1}{2}\,\pi_1 + \pi_2 + \frac{1}{2}\,\pi_4 \\ \pi_2 = \frac{1}{2}\,\pi_1 + \frac{1}{3}\,\pi_3 \\ \pi_3 = \frac{2}{3}\,\pi_3 + \frac{1}{2}\,\pi_4 \\ \pi_4 = 0 \\ \pi_1 + \pi_2 + \pi_3 + \pi_4 = 0 \end{cases}$$

har lösningen

$$\begin{cases} \pi_1 = 2/3 \\ \pi_2 = 1/3 \\ \pi_3 = 0 \\ \pi_4 = 0 \end{cases}$$

Eftersom det finns precis en komponent $\mathcal{A}_1 = \{E_1, E_2\}$ med positivt beständiga tillstånd (se Exempel 4.4), stämmer det med Följdsats 4.5(i) att den stationära fördelningen är entydig. Vidare är

$$\mu_1 = 1/\pi_1 = \frac{3}{2}, \qquad \mu_2 = 1/\pi_2 = 3$$

i överensstämmelse med (4.1) och de värden på μ_1 och μ_2 som vi explicit räknade ut i Exempel 4.1. □

Exempel 4.8 (Slumpvandring, forts.) En endimensionell slumpvandring är alltid irreducibel. Eftersom tillstånden antingen är nollbeständiga ($p = 1/2$) eller obeständiga ($p \neq 1/2$) så saknas alltså, enligt Följdsats 4.5(iii), stationär fördelning. Det beror på att sannolikhetsmassan rinner ut i oändligheten. I det asymmetriska fallet är detta inte så förvånande — Markovkedjan driver ju iväg mot antingen ∞ ($p > 1/2$) eller $-\infty$ ($p < 1/2$) enligt vad vi fann i Exempel 4.2. I det symmetriska fallet är visserligen $X(n)$:s fördelning centrerad kring origo, men den blir alltmer utbredd kring 0 då n växer. Något oegentligt kan vi säga att $\pi_i = 0$ (och då är alltså inte vektorn π en sannolikhetsfördelning). Låt oss motivera detta då $i = 0$ genom att titta på gränsvärdet av $p_0^{(n)}$ då n växer. Eftersom slumpvandringen startar i origo gäller $(X(n) + n)/2 \in Bin(n, p)$ (visa gärna detta), och därför $p_0^{(n)} = 0$ då n är udda och

$$p_0^{(n)} = \binom{n}{n/2} p^{n/2} q^{n/2} = \frac{n!}{((n/2)!)^2} p^{n/2} q^{n/2} \sim \frac{(4pq)^{n/2}}{2\sqrt{2\pi n}} \leq \frac{1}{2\sqrt{2\pi n}} \to 0$$

då $n \to \infty$ och n är jämn. Här betyder $\sim$ att kvoten mellan höger- och vänsterleden går mot 1. Vidare utnyttjade vi Stirlings formel $n! \sim \sqrt{2\pi}\, n^{n+1/2} \mathrm{e}^{-n}$ för att uppskatta binomialkoefficienten, samt $pq \leq 1/4$ i den sista olikheten. För övriga tillstånd E_i visas analogt att $p_i^{(n)} \to 0$ då $n \to \infty$. □

4.4 Asymptotisk fördelning

Det bevisades i Sats 5 i Bloms A att varje asymptotisk fördelning måste vara stationär. Omvändningen gäller inte alltid. I Exempel 4.6 hittade vi många stationära fördelningar, och givetvis existerar då ingen asymptotisk fördelning π, eftersom definitionen kräver konvergens mot π oberoende av startvektorn.

Ett nödvändigt krav för existensen av en asymptotisk fördelning är alltså att det finns precis en stationär fördelning. Detta krav (som enligt Följdsats 4.5 innebär att det finns precis en komponent $\mathcal{A}_1$ med positivt beständiga tillstånd) är emellertid inte tillräckligt. Markovkedjan måste dessutom vara aperiodisk, det får inte finnas några nollbeständiga tillstånd och de obeständiga tillstånden måste lämnas med sannolikhet 1:

Sats 4.6 (Existens av asymptotisk fördelning.(*)) *En Markovkedja har en asymptotisk fördelning om och endast om den är aperiodisk, tillståndsrummet har dekompositionen*

$$\mathcal{G} = \mathcal{A}_1 \cup \mathcal{C}$$

och Markovkedjan med sannolikhet ett lämnar $\mathcal{C}$ om den startar där (se Sats 4.2 för beteckningarna $\mathcal{A}_1$ och $\mathcal{C}$). Den asymptotiska fördelningen ges av $\pi = (\pi_1, \pi_2, \ldots)$, med

$$\pi_i = \begin{cases} \mu_i^{-1}, & E_i \in \mathcal{A}_1, \\ 0, & E_i \in \mathcal{C}, \end{cases}$$

där μ_i^{-1} är den förväntade återkomsttiden för tillstånd E_i. □

Exempel 4.9 (Förgreningsprocesser.(*)) Läs igenom Avsnitt 11.5 i Blom A innan du studerar detta exempel. Där visas att antalet individer i olika generationer $\{X(n),\ n = 0, 1, 2, \ldots\}$ är en Markovkedja med tillstånd

$$E_i = \text{"}i \text{ individer i en generation"}, \quad i = 0, 1, 2, \ldots$$

Då $n \to \infty$ kommer processen antingen att växa över alla gränser eller att dö ut. Det betyder att E_0 är ett absorberande tillstånd och därmed positivt beständigt, medan $E_1, E_2, \ldots$ alla är obeständiga. Vi får alltså dekompositionen

$$\mathcal{G} = \mathcal{A}_1 \cup \mathcal{C} = \{E_0\} \cup \{E_1, E_2, \ldots\}.$$

Man kan visa att en förgreningsprocess är aperiodisk. Om processen lämnar $\mathcal{C}$ med sannolikhet ett så dör den säkert ut, och detta sker då det förväntade antalet barn $m = \mathsf{E}(Y)$ till en individ är mindre eller lika med ett. Enligt Sats 4.6 existerar då en asymptotisk fördelning, nämligen $\pi = (1, 0, 0, \ldots)$. Om å andra sidan processen ej lämnar $\mathcal{C}$ med sannolikhet ett (fallet $m > 1$) existerar ingen asymptotisk fördelning beroende på att processen med positiv sannolikhet växer över alla gränser. □

För ändliga Markovkedjor får vi följande mycket viktiga följdsats till Sats 4.6.

Följdsats 4.7 (Asymptotiska fördelningar för ändliga Markovkedjor.) *En ändlig Markovkedja har en asymptotisk fördelning om och endast om den är aperiodisk och har precis en irreducibel komponent. Speciellt har en aperiodisk, ändlig och irreducibel Markovkedja alltid en asymptotisk fördelning.* □

Bevis.(*) Enligt Sats 4.3 har varje ändlig Markovkedja dekompositionen $\mathcal{G} = \mathcal{A}_1 \cup \ldots \cup \mathcal{A}_m \cup \mathcal{C}$ med $m \geq 1$ och $\mathcal{A}_i$ en irreducibel och sluten komponent av positivt beständiga tillstånd. Eftersom $\mathcal{C}$ är ändlig grupp av obeständiga tillstånd så lämnar Markovkedjan $\mathcal{C}$ med sannolikhet 1 om den startar i $\mathcal{C}$. Men då ger Sats 4.6 att en asymptotisk fördelning existerar om och endast om Markovkedjan är aperiodisk och har en irreducibel komponent ($m = 1$). □

Exempel 4.10 (En periodisk Markovkedja.) Vi ska nu beskriva ett fall där aperiodicitetsvillkoret i Följdsats 4.7 fallerar. Betrakta den irreducibla och periodiska Markovkedjan med övergångsmatris

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

Man ser att det existerar precis en stationär fördelning $\pi = (1/3, 1/3, 1/3)$. Den kan dock inte svara mot en asymptotisk fördelning. Om t.ex. $p^{(0)} = (1, 0, 0)$ så gäller

$$p^{(n)} = \begin{cases} (1,0,0), & n = 3k, \\ (0,1,0), & n = 3k+1, \\ (0,0,1), & n = 3k+2, \end{cases}$$

där $k = [n/3]$ är ett icke-negativt heltal. Vi ser att den absoluta sannolikhetsvektorn $p^{(n)}$ inte konvergerar, vilket beror på att Markovkedjan hoppar runt i en cykel $E_1 \to E_2 \to E_3 \to E_1 \to \ldots$ helt deterministiskt. □

Som ett alternativt kriterium för existens av asymptotiska fördelningar för ändliga Markovkedjor kan man använda följande sats:

Sats 4.8 (Asymptotiska fördelningar för ändliga Markovkedjor.) *För en ändlig Markovkedja med N tillstånd existerar en asymptotisk fördelning om och endast om $P^{(N-1)^2}$ har minst en kolonn med positiva* (> 0) *element.* □

Med denna sats i bagaget kan man lägga mindre vikt vid Sats 4 och Sats 6 i Blom A, Kapitel 11 (vilka också gäller för ändliga Markovkedjor).

Observera att vi nu har två karakteriseringar av ändliga Markovkedjor med asymptotiska fördelningar; Följdsats 4.7 och Sats 4.8. Den första satsen ger teoretisk förståelse, medan den sistnämnda är praktiskt mer användbar[1]. Man skriver nämligen lätt ett datorprogram som beräknar $P^{(N-1)^2}$ och sedan kontrollerar att någon kolumn har positiva element. (Se vidare övningsboken för exempel och ytterligare diskussion.)

4.5 Övergångstider

Vi delar upp tillståndsrummet $\mathcal{G}$ i två grupper $\mathcal{G}_1$ och $\mathcal{G}_2$. För enkelhets skull numreras tillstånden så att de i $\mathcal{G}_1$ betecknas $E_1, E_2, \ldots, E_N$ medan de i $\mathcal{G}_2$ har index $> N$. Utgå från tillståndet $E_i \in \mathcal{G}_1$, och låt

$$Y_i = \text{"tiden till första besöket i } \mathcal{G}_2 \text{ givet start i } E_i\text{"}.$$

Vi vill beräkna $\mathsf{E}(Y_i)$, den så kallade medelövergångstiden från E_i till $\mathcal{G}_2$. Vi inför hjälpvariablerna

$$\tau_{ij} = \text{"antal besök i } E_j \in \mathcal{G}_1 \text{ innan } \mathcal{G}_2 \text{ nås"}.$$

För τ_{ii} inkluderas besöket vid starten i räkningen. Uppenbarligen gäller $Y_i = \tau_{i1} + \cdots + \tau_{iN}$, eftersom tiden det tar att nå $\mathcal{G}_2$ är summan av antal besök i de olika tillstånden i $\mathcal{G}_1$ som föregår första hoppet till $\mathcal{G}_2$. Bilda sedan väntevärdet i båda leden:

$$\mathsf{E}(Y_i) = \mathsf{E}(\tau_{i1}) + \cdots + \mathsf{E}(\tau_{iN}). \tag{4.2}$$

Vi ska nu härleda hur man beräknar $\mathsf{E}(\tau_{ij})$. Övergångsmatrisen (efter omnumreringen) delas upp i block enligt

$$P = \begin{pmatrix} Q & P_{\mathcal{G}_1\mathcal{G}_2} \\ P_{\mathcal{G}_2\mathcal{G}_1} & P_{\mathcal{G}_2\mathcal{G}_2} \end{pmatrix}$$

där Q består av övergångssannolikheter mellan tillstånden i $\mathcal{G}_1$, $P_{\mathcal{G}_1\mathcal{G}_2}$ övergångssannolikheter från tillstånd i $\mathcal{G}_1$ till tillstånd i $\mathcal{G}_2$ osv. Vidare inför vi matrisen $\tau = (\tau_{ij})_{i,j=1}^{N}$ och I, enhetsmatrisen av ordning N.

[1] I själva verket kan man visa Sats 4.8 med hjälp av Följdsats 4.7, se Uppgifterna 4.8–4.9 för en bevisskiss av en enklare version av Sats 4.8, nämligen Sats 4 i Blom A, Kapitel 11.

Sats 4.9 *Anta att man alltid lämnar $\mathcal{G}_1$ någon gång, oavsett vilket tillstånd man startar ifrån. Då är $I-Q$ icke-singulär, och*

$$(I-Q)^{-1} = \mathsf{E}(\tau) = \begin{pmatrix} \mathsf{E}(\tau_{11}) & \dots & \mathsf{E}(\tau_{1N}) \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \mathsf{E}(\tau_{N1}) & \dots & \mathsf{E}(\tau_{NN}) \end{pmatrix}.$$

Speciellt ges $\mathsf{E}(Y_i)$ av i:te radsumman i $(I-Q)^{-1}$.

Bevis:(*) Antag att $I-Q$ är singulär. Då finns en nollskild radvektor $\pi = (\pi_1, \dots, \pi_N)$ (ej nödvändigtvis med positiva koordinater) som uppfyller $\pi = \pi Q$. Eftersom π kan ersättas med $-\pi$ kan vi utan inskränkning anta att minst en koordinat π_i är positiv. Låt I vara den delmängd av $\{1, \dots, N\}$ som innehåller de positiva koordinaterna hos π. Då fås

$$\sum_{i\in I} \pi_i \;=\; \sum_{j\in I} (\pi Q)_j = \sum_{j\in I} \sum_{i=1}^{N} \pi_i p_{ij} \le \sum_{i\in I} \pi_i \sum_{j\in I} p_{ij} \le \sum_{i\in I} \pi_i,$$

där vi i första olikheten utnyttjade att $\pi_i \le 0$ då $i \notin I$, och i sista olikheten användes $\sum_{j\in I} p_{ij} \le \sum_{j=1}^{N} p_{ij} \le 1$. Tydligen måste $\sum_{j\in I} p_{ij} = 1$ gälla för alla $i \in I$, men det innebär att gruppen av tillstånd med indexmängd I är sluten, vilket motsäger att vi kan nå $\mathcal{G}_2$ från alla tillstånd i $\mathcal{G}_1$. Allstå måste $I-Q$ vara icke-singulär.

Anta nu $i \neq j$. Då fås

$$\begin{aligned} \mathsf{E}(\tau_{ij}) &= \sum_k \mathsf{E}(\tau_{ij} \mid \text{första övergång till } E_k) p_{ik} \\ &= \sum_{k=1}^{N} \mathsf{E}(\tau_{kj}) p_{ik} + \sum_{k\ge N+1} 0 \cdot p_{ik} \end{aligned}$$

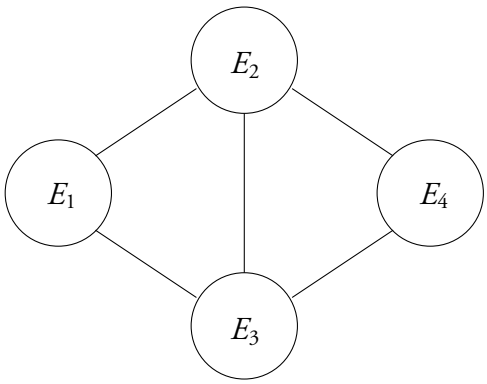
Då $i = j$ erhålls

$$\begin{aligned} \mathsf{E}(\tau_{ii}) &= \sum_k \mathsf{E}(\tau_{ii} \mid \text{första övergång till } E_k) p_{ik} \\ &= \sum_{k=1}^{N} (1 + \mathsf{E}(\tau_{ki})) p_{ik} + \sum_{k\ge N+1} 1 \cdot p_{ik} \\ &= \sum_{k=1}^{N} p_{ik} + \sum_{k=1}^{N} \mathsf{E}(\tau_{ki}) p_{ik} + 1 - \sum_{k=1}^{N} p_{ik} \\ &= \sum_{k=1}^{N} \mathsf{E}(\tau_{ki}) p_{ik} + 1. \end{aligned}$$

På matrisform kan detta sättas samman till

$$\mathsf{E}(\tau) = Q \cdot \mathsf{E}(\tau) + I \Leftrightarrow (I-Q)\mathsf{E}(\tau) = I \Leftrightarrow \mathsf{E}(\tau) = (I-Q)^{-1}.$$

□



Figur 4.1: *Tillståndsdiagram för Exempel 4.11.*

Exempel 4.11 I Figur 4.1 visas tillståndsdiagrammet för en Markovkedja där man från ett tillstånd går till de närmaste tillstånden med lika sannolikhet. Övergångsmatrisen blir då

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

Tag $\mathcal{G}_1 = \{E_3, E_4\}$ och $\mathcal{G}_2 = \{E_1, E_2\}$. Eftersom $\mathcal{G}_1$ i detta exempel ges av de två sista tillstånden fås Q som den 2×2 undre delmatrisen av P, dvs

$$Q = \begin{pmatrix} 0 & 1/3 \\ 1/2 & 0 \end{pmatrix}, \qquad I-Q = \begin{pmatrix} 1 & -1/3 \\ -1/2 & 1 \end{pmatrix} \qquad \text{och} \qquad (I-Q)^{-1} = \frac{6}{5}\begin{pmatrix} 1 & 1/3 \\ 1/2 & 1 \end{pmatrix}.$$

Medelövergångstiden från E_3 till den första av E_1 och E_2 är $^6/_5 + {}^2/_5 = {}^8/_5$ och medelövergångstiden från E_4 till den första av E_1 och E_2 är $^3/_5 + {}^6/_5 = {}^9/_5$. □

En viktig tillämpning av Sats 4.9 är s.k. *absorptionstider*, vilka vi nu helt kort behandlar:

Definition 4.3 *Om $\mathcal{G}_2$ är sluten (se Definition 4.2) säger man att $\mathcal{G}_2$ utgör en absorberande grupp av tillstånd, och absorptionstiden definieras som*

$$Y = \min\{n \geq 0;\, X(n) \in \mathcal{G}_2\},$$

givet $X(0) \notin \mathcal{G}_2$. □

Som en enkel konsekvens av Sats 4.9 får vi följande uttryck för beräkning av den genomsnittliga absorptionstiden:

Följdsats 4.10 *Den genomsnittliga absorptionstiden (se Definition 4.3) ges av*

$$\mathsf{E}(Y) = \sum_{i=1}^{N} \mathsf{P}(X(0) = E_i) \cdot \mathsf{E}(Y \,|\, X(0) = E_i) = \sum_{i=1}^{N} p_i^{(0)} E(Y_i),$$

där $\mathsf{E}(Y_i)$ bestäms ur (4.2) och Sats 4.9. □

Exempel 4.12 Låt

$$P = \begin{pmatrix} q & p \\ 0 & 1 \end{pmatrix},$$

med $q = 1 - p$, $0 < p < 1$. Här är $\mathcal{G}_2 = \{E_2\}$ ett absorberande tillstånd och $\mathcal{G}_1 = \{E_1\}$. I detta enkla fall kan vi räkna ut fördelningen för Y exakt, nämligen

$$Y \in \mathit{ffg}(p)$$

eftersom vi upprepar oberoende försök att hamna i $\mathcal{G}_2$ tills de lyckas, dvs tills vi lämnar E_1 och hamnar i E_2. Eftersom $Q = (q)$ ger Sats 4.9

$$\mathsf{E}(Y) = (1-q)^{-1} = p^{-1},$$

vilket stämmer med vad vi vet om väntevärdet för en ffg-fördelning. □

Exempel 4.13 (Kupongsamlarproblemet.) Anta att varje cornflakespaket innehåller en av N möjliga popstjärnor med samma sannolikhet. Hur många paket måste man köpa i genomsnitt tills man fått en fullständig samling? Inför

$$E_i = \text{”samlingen innehåller } i-1 \text{ olika popstjärnor”}, \quad i = 1, \ldots, N+1.$$

Tydligen blir

$$X(n) = \text{”antal popstjärnor i samlingen efter } n \text{ köpta paket”}$$

en Markovkedja som startar i E_1 ($X(0) = E_1$) och har övergångsmatrisen

$$P = \begin{pmatrix} q_1 & p_1 & 0 & \ldots & 0 & 0 \\ 0 & q_2 & p_2 & \ldots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & \ldots & q_N & p_N \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix},$$

med $p_i = (N-i+1)/N$, $i = 1, \ldots, N$ och $q_i = 1 - p_i$. Vid starten i E_1 kommer vi direkt att hoppa till E_2, eftersom första inköpet alltid utökar samlingen. När E_2 nåtts kommer vi att tillbringa en $\mathit{ffg}(p_2)$-fördelad tid där tills vi hoppar över till E_3, eftersom vi inte nödvändigtvis utökar vår samling direkt när den redan innehåller en popstjärna. Vi kan ju ha otur och köpa ytterligare ett eller flera paket med samma person. Analogt tillbringar vi sedan en $\mathit{ffg}(p_i)$-fördelad tid τ_{1i} i E_i, $i = 3, \ldots, N$. Tiden tills vi når det absorberande tillståndet $\mathcal{G}_2 = E_{N+1}$ (full samling) är därför en summa av oberoende och ffg-fördelade stokastiska variabler. Utnyttjar vi så väntervärdet för en ffg-fördelning får vi $\mathsf{E}(\tau_{1i}) = p_i^{-1}$. Medeltiden till absorption blir således

$$\mathsf{E}(Y_1) = \sum_{i=1}^{N} \mathsf{E}(\tau_{1i}) = \sum_{i=1}^{N} p_i^{-1} = \sum_{i=1}^{N} \frac{N}{N-i+1} = \sum_{i=1}^{N} \frac{N}{i} \approx N \ln(N). \tag{4.3}$$

Notera att $\mathsf{E}(Y_1)$ är (approximativt) en faktor $\ln(N)$ större än om vi var garanterade utökad samling vid varje inköp ($Y_1 = N$).

Vi ska nu använda Sats 4.9 för att kontrollera uttrycket för $\mathsf{E}(Y_1)$. Med $\mathcal{G}_1 = \{E_1, \ldots, E_N\}$ får vi

$$Q = \begin{pmatrix} q_1 & p_1 & 0 & \ldots & 0 \\ 0 & q_2 & p_2 & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \ldots & q_N \end{pmatrix}.$$

Vi startar ju i E_1, så medeltiden till absorption ges av första radsumman i $\mathsf{E}(\tau) = (I-Q)^{-1}$. Vi behöver alltså inte bestämma hela inversen, utan det räcker att lösa ut första raden:

$$\begin{aligned} &(\mathsf{E}(\tau_{11}), \ldots, \mathsf{E}(\tau_{1N}))\,(I-Q) = (1,0,\ldots,0) \quad \Longleftrightarrow \\ &\begin{cases} \mathsf{E}(\tau_{1N})p_1 &= 1, \\ -\mathsf{E}(\tau_{1,i-1})p_{i-1} + \mathsf{E}(\tau_{1i})p_i &= 0, \quad i = 2,\ldots,N \end{cases} \quad \Longleftrightarrow \\ &\mathsf{E}(\tau_{1i}) = p_i^{-1}, \quad i = 1,\ldots,N. \end{aligned}$$

Summering av elementen i första raden ger sedan (4.3). □

I de två senaste exemplen var direkta metoder (baserade på ffg-fördelningar) egentligen enklare att använda än Sats 4.9. Vi ska nu ge ett exempel där så inte är fallet[2]:

Exempel 4.14 (Antal par skor vid dörren.) En man har fyra par skor och hans hus har två ytterdörrar. Varje gång han tar en promenad väljer han slumpmässigt (sannolikhet 0.5) vilken dörr han ska gå ut och in genom. Om han börjar med två par skor vid respektive dörr, hur många promenader tar det i genomsnitt innan han finner att den dörr som han tänkt gå ut igenom saknar skor?

Låt oss införa tillstånden

E_1 : ”två par skor vid vardera dörr efter förra promenaden”
E_2 : ”tre par skor vid ena dörren och ett par vid den andra efter förra promenaden”
E_3 : ”fyra par skor vid ena dörren och inga vid den andra efter förra promenaden”
E_4 : ”vid någon tidigare promenad har man måst byta dörr före utgång”

Vi startar i E_1, och så länge vi inte behöver byta dörr före en promenad befinner vi oss i $\mathcal{G}_1 = \{E_1, E_2, E_3\}$. Så snart vi måste byta dörr (pga att den dörr vi först gick till saknade skor) hamnar vi i $\mathcal{G}_2 = \{E_4\}$, som är ett absorberande tillstånd. Eftersom vi intresserar oss för antalet promenader Z *innan* vi finner en dörr tom gäller

$$Z = Y_1 - 1$$

eftersom absorptionstiden Y_1 givet start i E_1 även räknar med den sista promenaden efter dörrbyte. Markovkedjan får övergångsmatrisen

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & 0 \\ 0 & 1/4 & 1/4 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

(visa gärna detta). Då ges Q av den övre 3×3 delmatrisen, dvs

$$Q = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1/4 & 1/4 \end{pmatrix} \qquad I - Q = \begin{pmatrix} 1/2 & -1/2 & 0 \\ -1/4 & 1/2 & -1/4 \\ 0 & -1/4 & 3/4 \end{pmatrix}.$$

[2]Samma problem förekommer som en simuleringsuppgift i Blom A, sid 223.

Invertering ger

$$(I-Q)^{-1} = \begin{pmatrix} 5 & 6 & 2 \\ 3 & 6 & 2 \\ 1 & 2 & 2 \end{pmatrix}.$$

Slutligen ger Sats 4.9

$$\mathsf{E}(Z) = \mathsf{E}(Y_1) - 1 = 5 + 6 + 2 - 1 = 12.$$

□

4.6 Övningsuppgifter

4.1. I en frågesport är sannolikheten p att svara rätt på frågorna. Den som vinner belönas med en skiva plus samtliga skivor i potten. Svarar man fel läggs en skiva till potten. Inför en Markovkedja med tillstånd

$$E_i = \text{"}i \text{ skivor i potten"}, \quad i = 0, 1, 2, \ldots$$

Beräkna övergångsmatrisen P samt $f_{00}^{(n)}$, $n = 1, 2, \ldots$ Givet att potten är tom, hur lång tid tar det i genomsnitt (μ_0) innan den töms nästa gång? Kontrollera resultatet genom att beräkna den stationära fördelningen.

4.2. Givet en Markovkedja med

$$P = \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.9 & 0.1 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix}.$$

Beräkna $f_{22}^{(5)}$, sannolikheten att man för första gången återvänder till E_2 efter 5 steg.

4.3. Klassificera alla tillstånd i Markovkedjan med

$$P = \begin{pmatrix} 1/3 & 2/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 \end{pmatrix}.$$

4.4. Beräkna den stationära fördelningen till

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 0 & 1/2 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

Hur lång tid tar det i genomsnitt att återvända till E_1?

4.5. Vilka av följande Markovkedjor har en asymptotisk fördelning?

(a)

$$P = \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.8 & 0.2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

(b)

$$P = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.5 & 0 & 0.5 \\ 0.1 & 0.7 & 0.2 \end{pmatrix},$$

(c)

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

(d)

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

4.6. Vilka av kedjorna i föregående uppgift har en entydig stationär fördelning?

4.7. Bestäm (i den mån de existerar) den stationära och asymptotiska fördelningen till Markovkedjan med

$$P = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

4.8. (*) Anta att i:te kolumnen i P^r endast har positiva element. Visa att tillståndsrummet kan delas upp enligt $\mathcal{G} = \mathcal{A}_1 \cup \mathcal{C}$, med

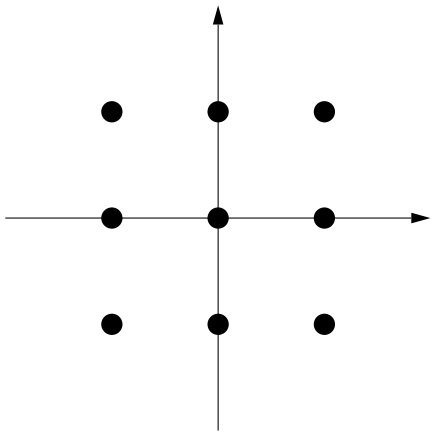
$$\begin{aligned} \mathcal{A}_1 &= \{E_j;\, E_i \to E_j\} \\ \mathcal{C} &= \{E_j;\, E_i \,|\, E_j\}, \end{aligned}$$

och $\mathcal{A}_1$ är sluten och irreducibel, medan $\mathcal{C}$ innehåller obeständiga tillstånd. Vi använder här beteckningarna

$$\begin{aligned} E_i \to E_j &: \quad p_{ij}^{(n)} > 0 \text{ för något } n > 0 \\ E_i \,|\, E_j &: \quad p_{ij}^{(n)} = 0 \text{ för alla } n > 0 \\ E_i \leftrightarrow E_j &: \quad E_i \to E_j \text{ och } E_j \to E_i \text{ (tvåsidig kommunikation)} \\ E_i \overset{n}{\to} E_j &: \quad p_{ij}^{(n)} > 0. \end{aligned}$$

4.9. (*) Med samma förutsättningar som i föregående uppgift, visa att tillstånden i $\mathcal{A}_1$ är aperiodiska. (Du får anta att alla tillstånd i en irreducibel komponent har samma period.)

4.10. Man hoppar omkring på rutmönstret nedan, varje gång så att alla angränsande tillstånd till höger, vänster, nedåt eller uppåt har samma sannolikhet att bli valda. Om man startar i mittpunkten, hur många gånger besöker man i genomsnitt en kantpunkt innan någon av hörnpunkterna nås. (Ledning: Av symmetriskäl räcker det att införa tre tillstånd.)



4.11. Betrakta Markovkedjan med

$$P = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.4 & 0.2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hur lång tid tar det i genomsnitt att nå E_3, givet att man startar i E_1.

4.12. (*) Två personer spelar flera omgångar av ett spel, tills någon vunnit N fler gånger än den andra. Vinstchansen är 0.5 för var och en i varje omgång. Hur många spelomgångar krävs i genomsnitt?

Kapitel 5

Diskreta Markovprocesser i kontinuerlig tid

5.1 Definition. Övergångs- och intensitetsmatriser.

Betrakta en diskret stokastisk process $\{X(t),\ t \geq 0\}$ i kontinuerlig tid med tillståndsrum

$$\mathcal{G} = \{E_0, E_2, \ldots, E_{N-1}\},$$

där antalet tillstånd N kan vara ändligt eller oändligt. Med $\{X(t) = E_i\}$ menas händelsen att processen vid tidpunkten t befinner sig i tillstånd E_i. För födelsedödsprocesser ändrar sig $X(\cdot)$ uppåt eller nedåt med högst en enhet i taget. Vi ska nu behandla det allmänna fallet, då $X(\cdot)$ kan göra större hopp och vandra mellan godtyckliga tillstånd E_i och E_j. Vi börjar med några definitioner:

Definition 5.1 (Markovvillkoret i kontinuerlig tid.) *Vi säger att* $\{X(t),\ t \geq 0\}$ *uppfyller Markovvillkoret om för varje följd av tidpunkter* $0 < t_1 < \ldots < t_n$ *(*$n \geq 3$*)*

$$\mathsf{P}(X(t_n) = x_n \,|\, X(t_1) = x_1, \ldots, X(t_{n-1}) = x_{n-1}) = \mathsf{P}(X(t_n) = x_n \,|\, X(t_{n-1}) = x_{n-1}).$$

□

Det innebär att om vi känner $X(t_{n-1})$ får vi ingen ytterligare information om framtida värden på processen genom att känna till dess beteende före tiden t_{n-1}. Om Markovvillkoret är uppfyllt säger vi att $X(\cdot)$ är en diskret Markovprocess i kontinuerlig tid.

När vi definierar övergångssannolikheter för Markovprocesser behöver vi endast betinga på ett tidsvärde (de tidigare tillför ju ingen information). Låt $0 \leq s < t$ och inför

$$p_{ij}(s,t) = \mathsf{P}(X(t) = E_j \,|\, X(s) = E_i)$$

som övergångssannolikheten mellan E_i och E_j från tiden s till t. Vi kommer fortsättningsvis endast att studera Markovprocesser vars egenskaper inte förändras med tiden:

Definition 5.2 (Tidshomogenitet.) *En diskret Markovprocess i kontinuerlig tid är tidshomogen om*

$$p_{ij}(s,t) = p_{ij}(0, t-s) \textit{ för alla } 0 \leq s < t.$$

□

Sannolikheten att hoppa från E_i och E_j under ett visst tidsintervall beror alltså inte på när vi startar intervallet utan endast på dess längd. För tidshomogena Markovprocesser är det onödigt att låta p_{ij} ha två argument. Vi sätter i stället (med viss symbolkollision)

$$p_{ij}(t) = p_{ij}(0,t).$$

Definiera härnäst *övergångsmatrisen* av ordning t som

$$P(t) = \big(p_{ij}(t)\big)_{i,j=0}^{N-1}$$

de absoluta sannolikheterna

$$p_i(t) = \mathsf{P}(X(t) = E_i),$$

samt motsvarande vektor $p(t) = (p_0(t), p_1(t), \ldots, p_{N-1}(t))$. Här kan N vara ändlig eller oändlig. Analogt med i diskret tid, kan vi med hjälp av övergångsmatrisen $P(t)$ och $p(0)$ bestämma $p(t)$ vid en godtycklig tidpunkt:

Sats 5.1 (Absoluta sannolikheter.) *Den absoluta sannolikhetsvektorn vid tiden $t > 0$ ges av*

$$p(t) = p(0)P(t). \tag{5.1}$$

□

Bevis av Sats 5.1. Enligt lagen om total sannolikhet gäller

$$p_i(t) = \mathsf{P}(X(t) = E_i) = \sum_{k=0}^{N-1} \mathsf{P}(X(0) = E_k) \cdot \mathsf{P}(X(t) = E_i \,|\, X(0) = E_k) = \sum_{k=0}^{N-1} p_k(0) p_{ki}(t).$$

Matrisvarianten av denna relation är precis (5.1). □

Exempel 5.1 (Livslängdsprocess.) Betrakta en livsängdsprocess med tillstånd $\{E_0, E_1\}$ och konstant felintensitet λ. Givet att vi befinner oss i E_0 så gäller $T \in Exp(1/\lambda)$ för den återstående tiden T i E_0 (= livslängden). Då vi aldrig kan lämna E_1 när vi en gång har kommit dit blir övergångsmatrisen

$$P(t) = \begin{pmatrix} \mathrm{e}^{-\lambda t} & 1 - \mathrm{e}^{-\lambda t} \\ 0 & 1 \end{pmatrix}.$$

Eftersom $p(0) = (1,\ 0)$, ger Sats 5.1 att $p(t) = (\mathrm{e}^{-\lambda t},\ 1 - \mathrm{e}^{-\lambda t})$. □

I det diskreta fallet kunde övergångsmatriser av alla ordningar skrivas som potenser av den minsta byggstenen $P(1)$. I kontinuerlig tid har vi inget minsta $t > 0$ att tillgå, men genom att göra en gränsövergång $t \to 0$ kan vi erhålla ett motsvarande resultat.

Vi behöver anta att högerderivatan av $p_{ij}(t)$ existerar för $t = 0$. Sätt $a_{ij} = p'_{ij}(0)$. Observera att $P(0) = I$, dvs $p_{ij}(0) = 0$ om $i \neq j$ och $p_{ii}(0) = 1$. Taylorutveckling av $p_{ij}(\cdot)$ kring 0 ger därför

$$p_{ij}(h) = \mathsf{P}(X(h) = E_j \,|\, X(0) = E_i) = \begin{cases} a_{ij}h + o(h), & i \neq j \\ 1 + a_{ii}h + o(h), & i = j. \end{cases}$$

Då $i \neq j$ kan vi tydligen tolka a_{ij} som övergångsintensiteten från E_i till E_j. Hela matrisen

$$A = (a_{ij})_{i,j=0}^{N-1}$$

benämns *intensitetsmatrisen*. Observera att $A = P'(0) = (p'_{ij}(0))_{i,j=0}^{N-1}$.

Exempel 5.2 (Poissonprocessen.) För en Poissonprocess med intensitet λ erhålls $a_{i,i+1} = -a_{ii} = \lambda$ samt $a_{ij} = 0$ annars. Det ger intensitetsmatrisen

$$A = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & \dots \\ 0 & 0 & 0 & -\lambda & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

□

Exempel 5.3 (Födelsedödsprocesser.) För en allmän födelsedödsprocess får intensitetsmatrisen utseendet

$$A = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \dots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & 0 & \dots \\ 0 & 0 & -\mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \dots \\ 0 & 0 & 0 & -\mu_3 & -(\lambda_3+\mu_3) & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.$$

Den enda nollskilda elementen hos A ligger till vänster om, på eller till höger om diagonalen. Orsaken är naturligtvis att vi inte tillåter andra förändringar än $E_i \to E_{i-1}$ och $E_i \to E_{i+1}$. □

Vi ser att radsummorna hos A är noll i de två exemplen ovan. Detta är ingen tillfällighet, vilket framgår av följande sats:

Sats 5.2 (Radsummor hos intensitetsmatarisen.) *För en diskret Markovprocess i kontinuerlig tid med ändligt många tillstånd och intensitetsmatris A gäller att*

$$\sum_{j=0}^{N-1} a_{ij} = 0,$$

dvs alla radsummor är 0. □

Bevis av Sats 5.2. Observera att

$$\begin{aligned} \sum_{j=0}^{N-1} a_{ij} &= \sum_{j=0}^{N-1} p'_{ij}(0) = \sum_{j=0}^{N-1} \lim_{h \searrow 0} \frac{p_{ij}(h) - p_{ij}(0)}{h} \\ &= \lim_{h \searrow 0} \frac{\sum_{j=0}^{N-1} p_{ij}(h) - \sum_{j=0}^{N-1} p_{ij}(0)}{h} = \lim_{h \searrow 0} \frac{1-1}{h} = 0, \end{aligned}$$

eftersom radsummorna i $P(h)$ och $P(0) = I$ är 1. □

Satsen gäller även för oändliga tillståndsrum under vissa tekniska villkor på övergångsmatriserna $P(h)$ då $h \searrow 0$ som garanterar att vi får byta plats på summation och gränsvärdesbildning i beviset ovan. Intuitivt representerar

$$\sum_{j;j\neq i} a_{ij}$$

den totala intensitet med vilken man övergår från E_i till något annat tillstånd E_j, dvs den takt med vilken $\mathsf{P}(X(t+h) \neq E_i \,|\, X(t) = E_i)$ ökar för små värden på h. Ökningen kompenseras av att $\mathsf{P}(X(t+h) = E_i \,|\, X(t) = E_i)$ ändras med hastigheten

$$a_{ii} = -\sum_{j;j\neq i} a_{ij}.$$

5.2 Tidsberoende hos övergångs- och absoluta sannolikheter.(*)

För Markovkedjor kunde vi med hjälp av rekursionsformeln $P(n+1) = P(1)P(n)$ härleda $P(n) = P(1)^n$, vilket innebär att övergångsmatriser av alla ordningar kan beskrivas med hjälp av $P(1)$. Nu ska vi i stället härleda ett system av differentialekvationer för elementen i $P(\cdot)$ som gör att övergångsmatriserna i kontinuerlig tid kan beskrivas med hjälp av A.

Sats 5.3 (Tidsutveckling hos övergångssannolikheter.) *För en diskret Markovprocess i kontinuerlig tid med ändligt tillståndsrum satisfierar övergångsmatriserna framåtekvationerna*

$$P'(t) = P(t)A, \quad t \geq 0,$$

och bakåtekvationerna

$$P'(t) = AP(t), \quad t \geq 0.$$

□

Bevis av Sats 5.3. Om vi ska göra en övergång från E_i till E_j från tiden 0 till $t+h$, kan vi betinga på var vi befinner oss vid tiden t. Enligt lagen om total sannolikhet fås

$$p_{ij}(t+h) = \sum_{k=0}^{N-1} p_{ik}(t)p_{kj}(h) = p_{ij}(t)(1 + a_{jj}h + o(h)) + \sum_{k;k\neq j} p_{ik}(t)(a_{kj}h + o(h)).$$

Flytta nu över $p_{ij}(t)$ till vänsterledet, dividera med h och låt $h \searrow 0$,

$$p'_{ij}(t) = \lim_{h\searrow 0} \frac{p_{ij}(t+h) - p_{ij}(t)}{h} = \lim_{h\searrow 0} \sum_{k=0}^{N-1} \left(p_{ik}(t)a_{kj} + p_{ik}(t)\frac{o(h)}{h} \right) = \sum_{k=0}^{N-1} p_{ik}(t)a_{kj},$$

där vi utnyttjat att $o(h)/h \to 0$ och att summan av ändligt många sådana termer också måste gå mot noll. Men detta är precis framåtekvationerna, formulerade elementvis. Genom att i stället betinga på var processen befinner sig vid tiden h bevisar man bakåtekvationerna på samma sätt. □

Strikt resonerat har vi egentligen endast behandlat högerderivator av $P(\cdot)$, men man kan modifiera resonemanget något, så att resultatet även omfattar vänsterderivatorna. Precis som Sats 5.2 gäller Sats 5.3 under vissa bivillkor även för oändliga tillståndrum.

Man kan nu lösa differentialekvationerna i Sats 5.3 med begynnelsevillkoret $P(0) = I$. Det visar sig bekvämt att generalisera exponentialfunktionen till att omfatta även matriser. Genom att utgå från Taylors formel, definierar vi

$$\mathrm{e}^G = \exp(G) = \sum_{k=0}^{\infty} \frac{G^k}{k!},$$

för en godtycklig kvadratisk matris (ändlig eller oändlig). Lösningen till framåt- resp bakåtekvationerna kan då skrivas

$$P(t) = \exp(tA) \tag{5.2}$$

för ändliga tillståndsrum, och under vissa bivillkor även för oändliga tillståndsrum. Man kontrollerar (5.2) genom att sätta in i framåt- eller bakåtekvationerna och derivera termvis. Lösningen ska jämföras med den skalära differentialekvationen

$$\begin{cases} P'(t) = aP(t), & t > 0 \\ P(0) = 1, \end{cases}$$

som har lösningen $P(t) = \exp(at)$.

Med hjälp av (5.2) kan vi nu härleda ett uttryck för hur de absoluta sannolikheterna $p(t)$ utvecklas i tiden:

Sats 5.4 (Tidsutveckling av absoluta sannolikheter.) *För ändliga tillståndsrum ges de absoluta sannolikheterna vid tiden $t > 0$ av*

$$p(t) = p(0)\exp(tA).$$

Vidare uppfyller $p(\cdot)$ systemet av differentialekvationer

$$p'(t) = p(t)A. \tag{5.3}$$

□

Den i:te koordinaten i (5.3) kan skrivas

$$p_i'(t) = \sum_{k;k\neq i} a_{ki}p_k(t) - (-a_{ii}p_i(t)). \tag{5.4}$$

Nu kan $a_{ki}p_k(t)$ tolkas som flödet av sannolikhetsmassa från E_k till E_i om $k \neq i$, dvs $\lim_{h\to 0} \mathsf{P}(X(t) = E_k,\, X(t+h) = E_i)/h$. Summan i högerledet av (5.4) svarar därför mot det totala *inflödet av sannolikhetsmassa* till E_i, medan $-a_{ii}p_i(t)$ ger *utflödet av sannolikhetsmassa* från E_i. Subtraheras in- och utflöde erhålls $p_i'(t)$, nettoförändringen av $p_i(t)$.

Bevis av Sats 5.4. Första halvan av satsen fås genom att kombinera Sats 5.1 med (5.2). Den andra identiteten (5.3) visas med hjälp av framåtekvationerna i Sats 5.3:

$$p'(t) = \frac{d}{dt}(p(0)P(t)) = p(0)P'(t) = p(0)P(t)A = p(t)A. \tag{5.5}$$

□

Exempel 5.4 (Två tillstånd.) En maskins funktionsstatus beskrivs av en Markovprocess med två tillstånd E_0 = ”maskinen är hel”, och E_1 = ”maskinen repareras”. Intensitetsmatrisen ges av

$$A = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

Här är tydligen λ den intensitet med vilken maskinen går sönder och μ den intensitet med vilken den repareras. Anta att maskinen fungerar vid $t = 0$. Enligt Sats 5.4 gäller nu

$$\begin{cases} p_0'(t) = -\lambda p_0(t) + \mu p_1(t) \\ p_1'(t) = \lambda p_0(t) - \mu p_1(t). \end{cases}$$

Eftersom $p_0(t) + p_1(t) = 1$ får vi

$$p_0'(t) = \mu - (\lambda + \mu)p_0(t),$$

en linjär differentialekvation med begynnelsevillkoret $p_0(0) = 1$ och lösningen

$$p_0(t) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}\,\mathrm{e}^{-(\lambda+\mu)t}.$$

Observera att $(p_0(t), p_1(t))$ ger första raden i övergångsmatrisen. Motsvarande förfarande med begynnelsevillkoret $p_0(0) = 0$ ger andra raden i övergångsmatrisen, dvs vi får

$$P(t) = \begin{pmatrix} \dfrac{\mu}{\lambda+\mu} + \dfrac{\lambda}{\lambda+\mu}\,\mathrm{e}^{-(\lambda+\mu)t} & \dfrac{\lambda}{\lambda+\mu} - \dfrac{\lambda}{\lambda+\mu}\,\mathrm{e}^{-(\lambda+\mu)t} \\ \dfrac{\mu}{\lambda+\mu} - \dfrac{\mu}{\lambda+\mu}\,\mathrm{e}^{-(\lambda+\mu)t} & \dfrac{\lambda}{\lambda+\mu} + \dfrac{\mu}{\lambda+\mu}\,\mathrm{e}^{-(\lambda+\mu)t} \end{pmatrix}. \tag{5.6}$$

Livslängdsprocessen i Exempel 5.1 svarar mot specialfallet $\mu = 0$ (ingen reparation sker). □

5.3 Stationära sannolikheter.

Vi övergår nu till att behandla stationära fördelningar:

Definition 5.3 (Stationär fördelning) *En* stationär fördelning $\pi = (\pi_0, \pi_1, \ldots)$ *definieras av*

$$p(0) = \pi \Rightarrow p(t) = \pi \textit{ för alla } t > 0.$$

□

Om $X(\cdot)$ vid någon tidpunkt har fördelningen π, så kommer tydligen dess fördelning att förbli densamma vid alla senare tidpunkter. Med hjälp av intensitetsmatrisen kan ett praktiskt mycket användbart villkor för den stationära fördelningen formuleras.

Sats 5.5 *För en ändlig diskret Markovkedja i kontinuerlig tid så är π en stationär fördelning om och endast om*

$$\pi A = 0. \tag{5.7}$$

□

Bevis av Sats 5.5. Anta att (5.7) är uppfyllt samt att $p(0) = \pi$. Då gäller även $\pi A^k = 0$ för alla positiva heltal k. Utnyttja så (5.2) och definitionen av $\exp(\cdot)$,

$$p(t) = \pi \exp(tA) = \pi \sum_{k=0}^{\infty} \frac{(tA)^k}{k!} = \pi + \sum_{k=1}^{\infty} \frac{t^k \pi A^k}{k!} = \pi.$$

Anta omvänt att π är en stationär fördelning. Om $p(0) = \pi$ följer då $p(t) \equiv \pi$. Ekvation (5.5) ger då

$$0 = p'(t) = p(t)A = \pi A,$$

dvs vi har visat (5.7). □

Sats 5.4 och 5.5 gäller under vissa förutsättningar (som vi utelämnar) även för oändliga tillståndsrum. Det kommer vi att utnyttja i en del exempel nedan.

Vi ser att i:te raden i ekvationssystemet (5.7) har formen

$$0 = \sum_k \pi_k a_{ki} \Leftrightarrow \pi_i(-a_{ii}) = \sum_{k;k\neq i} \pi_k a_{ki}.$$

Vi erinrar oss diskussionen efter (5.4), och ser att vid stationäritet har jämvikt uppnåtts, så att utflödet av sannolikhetsmassa från E_i (vänsterledet) lika stort som inflödet av sannolikhetsmassa till E_i (högerledet) för alla tillstånd.

Exempel 5.5 (Födelsedödsprocesser, forts.) Bestäm den stationära fördelningen till den allmänna födelsedödsprocessen i Exempel 5.3. Vi får ekvationssystemet

$$\begin{cases} \pi A = 0 \\ \sum_k \pi_k = 1 \end{cases} \Leftrightarrow \begin{cases} -\lambda_0\pi_0 + \mu_1\pi_1 = 0, \\ \lambda\pi_{k-1} - (\mu_k + \lambda_k)\pi_k + \mu_{k+1}\pi_{k+1} = 0, \quad k = 1, 2, \ldots \\ \sum_k \pi_k = 1. \end{cases}$$

Det sista ekvationssystemet (övre delen) kan lösas rekursivt. Man ser att $\pi_k = \lambda_{k-1}\pi_{k-1}/\mu_k$, vilket i sin tur ger

$$\pi_k = \frac{\lambda_0 \cdot \ldots \cdot \lambda_{k-1}}{\mu_1 \cdot \ldots \cdot \mu_k}\pi_0.$$

Slutligen kan π_0 bestäms ur villoret $\sum_k \pi_k = 1$. Det ger

$$\pi_k = \frac{\frac{\lambda_0 \cdot \ldots \cdot \lambda_{k-1}}{\mu_1 \cdot \ldots \cdot \mu_k}}{1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \cdot \ldots \cdot \lambda_{j-1}}{\mu_1 \cdot \ldots \cdot \mu_j}}$$

förutsatt att serien i nämnaren konvergerar. (Annars kommer $X(t)$ växa över alla gränser då $t \to \infty$.)
□

5.4 Irreducibilitet och asymptotiska fördelningar.

Även i kontinuerlig tid kan $\mathcal{G}$ delas upp i obeständiga, nollbeständiga och positivt beständiga tillstånd. Vi går inte in närmare på detta, utan nöjer oss med att behandla irreducibilitet och asymptotiska fördelningar.

Definition 5.4 (Irreducibilitet.) *Markovprocessen* $\{X(t),\ t \geq 0\}$ *är irreducibel om för varje par av tillstånd* E_i *och* E_j *det finns ett* $t > 0$ *sådant att* $p_{ij}(t) > 0$. □

Man kan visa att om det finns ett t-värde med $p_{ij}(t) > 0$, så gäller detta *alla* $t > 0$. Denna egenskap hos övergångssannolikheterna har ingen motsvarighet för Markovkedjor.

Asymptotisk fördelning definieras helt analogt med i det diskreta fallet, vilket framgår av följande:

Definition 5.5 (Asymptotisk fördelning.) *Markovprocessen* $\{X(t),\ t \geq 0\}$ *har en* asymptotisk fördelning π *om* $p(t) \to \pi$ *då* $t \to \infty$, *oberoende av startvektorn* $p(0)$. □

Kriterier för existens av asymptotiska fördelningar för irreducibla Markovprocesser blir faktiskt enklare i kontinuerlig tid, eftersom begreppet periodicitet inte existerar. (Som vi kommer att se i nästa avsnitt är tidskillnaden mellan två hopp en kontinuerlig stokastisk variabel, närmare bestämt exponentialfördelad.)

Sats 5.6 (Existens av asymptotisk fördelning.) *Låt* $\{X(t),\ t \geq 0\}$ *vara en irreducibel Markovprocess i kontinuerlig tid.*

(i) *Om en stationär fördelning* π *existerar är den entydig och tillika en asymptotisk fördelning.*

(ii) *Om ingen stationär fördelning existerar så gäller* $p_{ij}(t) \to 0$ *då* $t \to \infty$ *för alla par av tillstånd* E_i *och* E_j.

(iii) *Om tillståndsrummet är ändligt existerar alltid en stationär fördelning.*

□

Villkoret (ii) svarar mot att $X(t)$ divergerar mot oändligheten eller gör allt större oscillationer iväg från origo, och det kan endast inträffa för oändliga tillståndsrum.

Exempel 5.6 (Födelsedödsprocesser, forts.) Betrakta en allmän födelsedödsprocess med $\lambda_i > 0$ och $\mu_j > 0$ för alla $i \geq 0$ och $j \geq 1$. Man ser lätt att denna kedja är irreducibel. Kriterium (i) i Sats 5.6 är uppfyllt om

$$\sum_{j=1}^{\infty} \frac{\lambda_0 \cdot \ldots \cdot \lambda_{j-1}}{\mu_1 \cdot \ldots \cdot \mu_j} < \infty,$$

dvs då är den stationära fördelningen i Exempel 5.5 även en asymptotisk fördelning. Om serien ovan divergerar saknas såväl stationär som asymptotisk fördelning (kriterium (ii)). □

Exempel 5.7 (Två tillstånd, forts.) Vi utnyttjar först Sats 5.5 för att bestämma den stationära fördelningen π i Exempel 5.4;

$$\begin{cases} \pi A = 0, \\ \pi_1 + \pi_2 = 1 \end{cases} \Longrightarrow \pi = \left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu} \right).$$

Enligt Sats 5.6 är detta även en asymptotisk fördelning. Alternativt ser man detta direkt ur (5.6), ty

$$\lim_{t\to\infty} p(t) = p(0) \lim_{t\to\infty} P(t) = p(0) \begin{pmatrix} \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \\ \frac{\mu}{\lambda+\mu} & \frac{\lambda}{\lambda+\mu} \end{pmatrix} = \pi$$

oberoende av startvektorn. □

5.5 Inbäddad Markovkedja, tid mellan hopp.

En diskret Markovprocess $\{X(t),\ t \geq 0\}$ i kontinuerlig tid gör alltid språngvisa förändringar. Vi antar att dessa sker vid tidpunkterna $0 < T_1 < T_2 < \ldots$. För att underlätta en del beteckningar sätter vi också $T_0 = 0$. Om vi bara bryr oss om vilka tillstånd $X(\cdot)$ hoppar mellan kan vi definiera en stokastisk process i diskret tid $\{\tilde{X}(n),\ n \geq 0\}$ genom

$$\tilde{X}(n) = X(T_n), \quad n \geq 0. \tag{5.8}$$

Vi antar att $X(\cdot)$ har högerkontinuerliga realiseringar, dvs $X(T_n) = E_j$ om hoppet $E_i \to E_j$ äger rum vid tiden T_n. Det visar sig att $\tilde{X}$ blir en Markovkedja, den så kallade *inbäddade Markovkedjan* till $X(\cdot)$, vilket framgår av följande sats:

Sats 5.7 (Inbäddad Markovkedja och tidsavstånd.) *Den stokastiska processen $\{\tilde{X}(n),\ n \geq 0\}$ definierad i (5.8) är en Markovkedja med övergångsmatris $\tilde{P} = (\tilde{p}_{ij})$, där*

$$\tilde{p}_{ij} = \begin{cases} -a_{ij}/a_{ii}, & j \neq i, \\ 0, & j = i. \end{cases}$$

Vidare gäller

$$T_n - T_{n-1} \in \mathrm{Exp}(\frac{-1}{a_{ii}}) \text{ om } \tilde{X}(n-1) = E_i$$

och $T_n - T_{n-1}$ är oberoende av $\tilde{X}(n)$. □

För en allmän diskret Markovprocess i kontinuerlig tid är alltså avstånden mellan hopptidpunkterna exponentialfördelade med intensitet $-a_{ii}$. (Tidigare har vi sett att Poissonprocessen har samma egenskap, med $-a_{ii} = \lambda$.) Att exponentialfördelningens väntevärde blir just $1/(-a_{ii})$ förefaller rimligt med tanke på att $-a_{ii}$ beskriver intensiteten att övergå från E_i till något annat tillstånd.

Sannolikheten $\tilde{p}_{ij}$ att hoppet från E_i hamnar i E_j är proportionell mot intensiteten a_{ij}. Det är alltså troligare att hoppa till E_j om a_{ij} är stor, vilket är intuitivt rimligt. Notera att radsummorna i $\tilde{P}$ blir 1, eftersom

$$\sum_j \tilde{p}_{ij} = \frac{\sum_{j;j\neq i} a_{ij}}{-a_{ii}} = 1,$$

där sista likheten följer av att radsummorna i A är 0.

Bevisskiss av Sats 5.7.(*) Sätt $Y = T_n - T_{n-1}$ och anta att $X(T_{n-1}) = i$. Vi ska nu utnyttja (utan bevis) den så kallade *starka Markovegenskapen*, som innebär att $t_1, \ldots, t_n$ i Definition 5.1 kan ersättas med vissa typer av stokastiska tidpunkter, exempelvis $T_1, \ldots, T_n$. Den starka Markovegenskapen hos $X(\cdot)$ medför att

$$\begin{aligned}
&\mathsf{P}(Y > x+h \,|\, Y > x) = \\
&\quad = \mathsf{P}(X(t) = E_i \text{ för } t \in [T_{n-1},\, T_{n-1}+x+h] \,|\, X(t) = E_i \text{ för } t \in [T_{n-1},\, T_{n-1}+x]) \\
&\quad = \mathsf{P}(X(t) = E_i \text{ för } t \in [T_{n-1}+x,\, T_{n-1}+x+h] \,|\, X(T_{n-1}+x) = E_i) \\
&\quad = \mathsf{P}(X(t) = E_i \text{ för } t \in [T_{n-1},\, T_{n-1}+h] \,|\, X(T_{n-1}) = E_i) = \mathsf{P}(Y > h).
\end{aligned}$$

Här utnyttjade vi tidshomogeniteten hos X i näst sista ledet. Dessutom betingade vi, till skillnad från i Definition 5.1, på ett helt intervall av t-värden. Man kan visa att detta är tillåtet. Med $R_Y(x) = \mathsf{P}(Y > x)$ fås alltså

$$R_Y(x+h) = R_Y(x)R_Y(h).$$

Gränsövergång $h \searrow 0$ ger

$$R_Y'(x) = \lim_{h \searrow 0} \frac{R_Y(x+h) - R_Y(x)}{h} = R_Y(x) \lim_{h \searrow 0} \frac{R_Y(h) - 1}{h} = R_Y(x)R_Y'(0) = a_{ii}R_Y(x).$$

Löser vi denna differentialekvation med begynnelsevillkor $R_Y(0) = 1$ får vi $R_Y(x) = \exp(a_{ii}x)$, dvs $Y \in Exp(-1/a_{ii})$.

För att motivera andra delen av satsen, notera att tidshomogeniteten och den starka Markovegenskapen medför (med liknande resonemang som ovan) att

$$\mathsf{P}(E_i \to E_j \,|\, x < Y < x+h) = \mathsf{P}(E_i \to E_j \,|\, 0 < Y < h) \approx \frac{p_{ij}(h)}{1 - p_{ii}(h)} \approx \frac{a_{ij}h}{-a_{ii}h} = \tilde{p}_{ij}$$

för små h och $i \neq j$. Approximationerna övergår i likheter då $h \searrow 0$. Givet att hoppet från E_i sker efter tiden x (räknat från T_{n-1}) är alltså sannolikheten $\tilde{p}_{ij}$ att hamna i E_j oberoende av x. Vi har alltså motiverat

$$\mathsf{P}(X(T_n) = E_j \,|\, X(T_{n-1}) = E_i,\ Y = x) = \tilde{p}_{ij}.$$

Eftersom högerledet inte beror av x måste dels $X(T_n)$ vara oberoende av Y och dels

$$\mathsf{P}(X(T_n) = E_j \,|\, X(T_{n-1}) = E_i) = \tilde{p}_{ij}.$$

□

Exempel 5.8 (Inbäddad födelsedödsprocess.) Betrakta en födelsedödsprocess med konstanta födelse- och dödsintensiteter $\lambda_i \equiv \lambda$ respektive $\mu_i \equiv \mu$. Man ser att $\tilde{X}$ blir en slumpvandring som reflekteras i origo, dvs $\tilde{p}_{01} = 1$ och

$$\tilde{p}_{ij} = \begin{cases} \lambda/(\mu+\lambda), & j = i+1, \\ \mu/(\mu+\lambda), & j = i-1, \\ 0, & \text{annars}, \end{cases}$$

om $i \geq 1$. □

Sats 5.7 har stor betydelse vid simulering av Markovprocesser. Vi kan först simulera Markovkedjan $\{\tilde{X}(n),\ n \geq 0\}$, och sedan generera tidsavstånden $\{T_n - T_{n-1}\}$ som exponentialfördelade stokastiska variabler. Det n:te tidsavståndet ska då ha medelvärde $1/(-a_{ii})$ om $\tilde{X}(n-1) = E_i$. I nästa avsnitt kommer vi att utnyttja Sats 5.7 för att bestämma väntevärdet av tiden att vandra mellan två grupper av tillstånd.

5.6 Övergångstider

Dela nu (som i Avsnitt 4.5 upp tillståndsrummet i två disjunkta delmängder,

$$\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2.$$

Vi ska bestämma väntevärdet av tiden att från ett tillstånd $E_i \in \mathcal{G}_1$ övergå till $\mathcal{G}_2$. Definiera först

$$Y_i = \text{"tidpunkten för första hoppet till } \mathcal{G}_2 \text{ givet } X(0) = E_i\text{"}.$$

Vi kan tänka oss ett system som kan befinna sig i olika tillstånd och som fungerar vid tiden t precis då $X(t) \in \mathcal{G}_1$. Alltså blir Y_i livslängden hos systemet (icke underhållna system), eller tiden till första reparationen (underhållna system). Denna tillämpning behandlas utförligare i Kapitel 6.

För att bestämma $\mathsf{E}(Y_i)$ visar det sig användbart att införa hjälpvariablerna

$$\tau_{ij} = \text{"totala tiden } E_j \text{ besöks innan } \mathcal{G}_2 \text{ nås givet } X(0) = E_i\text{"}.$$

Numrera nu tillstånden så att $E_1, \ldots, E_N \in \mathcal{G}_1$ och $E_j \in \mathcal{G}_2$ för $j > N$. Då gäller

$$\mathsf{E}(Y_i) = \sum_{j=1}^{N} \mathsf{E}(\tau_{ij}).$$

Dela upp intensitetsmatrisen i block enligt

$$A = \begin{pmatrix} \tilde{A} & A_{\mathcal{G}_1\mathcal{G}_2} \\ A_{\mathcal{G}_2\mathcal{G}_1} & A_{\mathcal{G}_2\mathcal{G}_2} \end{pmatrix},$$

där $\tilde{A}$ innehåller övergångsintensiteterna mellan alla tillstånd i $\mathcal{G}_1$, $A_{\mathcal{G}_1\mathcal{G}_2}$ övergångsintensiteterna från $\mathcal{G}_1$ till $\mathcal{G}_2$ osv. Det visar sig att $\mathsf{E}(\tau_{ij})$ kan bestämmas med hjälp av $\tilde{A}$, vilket framgår av följande sats:

Sats 5.8 (Övergångstider) *Låt $\tau = (\tau_{ij})_{i,j=1}^{N}$ vara matrisen av besökstider i olika tillstånd enligt ovan. Då gäller*

$$\mathsf{E}(\tau) = \begin{pmatrix} \mathsf{E}(\tau_{11}) & \ldots & \mathsf{E}(\tau_{1N}) \\ \vdots & & \vdots \\ \mathsf{E}(\tau_{N1}) & \ldots & \mathsf{E}(\tau_{NN}) \end{pmatrix} = -\tilde{A}^{-1},$$

och speciellt blir $\mathsf{E}(Y_i)$ summan av elementen i rad nummer i hos $-\tilde{A}^{-1}$. □

Eftersom vi endast är intresserade av den i:te raden $(\mathsf{E}(\tau_{i1}), \ldots, \mathsf{E}(\tau_{iN}))$ hos $-\tilde{A}^{-1}$ för att bestämma $\mathsf{E}(Y_i)$, behöver vi inte beräkna inversen, det räcker att lösa ekvationssystemet

$$(\mathsf{E}(\tau_{i1}), \ldots, \mathsf{E}(\tau_{iN}))\tilde{A} = -(0, \ldots, 0, 1, 0, \ldots, 0),$$

där enhetsvektorn har en etta i position i.

Bevis av Sats 5.8.(*) Anta först $i \neq j$ och låt T_1 beteckna tiden då första övergången $E_i \to E_k$ sker. Enligt Sats 5.7 gäller då $T_1 \in Exp(-1/a_{ii})$ och $\mathsf{P}(E_i \to E_k = \tilde{p}_{ik})$. Genom att betinga på vart processen hoppar första gången erhålls

$$\begin{aligned} \mathsf{E}(\tau_{ij}) &= \sum_{k\geq 1} \mathsf{E}(\tau_{ij} \mid E_i \to E_k)\mathsf{P}(E_i \to E_k) = \sum_{k\geq 1} \mathsf{E}(\tau_{kj})\tilde{p}_{ik} = \sum_{k=1}^{N} \mathsf{E}(\tau_{kj})\tilde{p}_{ik} \\ &= -\sum_{k;k\neq i,k\leq N} \mathsf{E}(\tau_{kj})\,\frac{a_{ik}}{a_{ii}} \Leftrightarrow \sum_{k=1}^{N} a_{ik}\mathsf{E}(\tau_{kj}) = 0, \end{aligned}$$

där vi utnyttjat $\tau_{kj} = 0$ om $E_k \in \mathcal{G}_2$. Om $i = j$ fås analogt

$$\begin{aligned}\mathsf{E}(\tau_{ii}) &= \sum_k (\mathsf{E}(\tau_{ki}) + \mathsf{E}(T_1))\tilde{p}_{ik} = -\sum_{k;k\neq i,k\leq N} \mathsf{E}(\tau_{kj})\,\frac{a_{ik}}{a_{ii}} + \frac{-1}{a_{ii}}\sum_k \tilde{p}_{ik} \\ &\Leftrightarrow \sum_{k=1}^{N} a_{ik}\mathsf{E}(\tau_{ki}) = -1,\end{aligned}$$

där vi utnyttjade att radsummorna till $\tilde{P}$ är 1. I matrisform kan de två sista ekvationerna sammanföras till

$$\tilde{A}\mathsf{E}(\tau) = -I,$$

vilket bevisar satsen. □

Vi avslutar med en enkel tillämpning av Sats 5.8. Mer avancerade exempel kommer att ges i Kapitel 6.

Exempel 5.9 (Livslängdsprocesser, forts.) Betrakta återigen processen i Exempel 5.1, med $\mathcal{G}_1 = E_0$ och absorberande tillstånd $\mathcal{G}_2 = E_1$. Här är $\tilde{A} = -\lambda$ en skalär, eftersom $\mathcal{G}_1$ endast innehåller ett element. Medeltiden till absorption blir

$$\mathsf{E}(Y_0) = -\tilde{A}^{-1} = \lambda^{-1},$$

vilket stämmer väl överens med att vi vet att $Y_0 \in Exp(1/\lambda)$. □

5.7 Övningsuppgifter

5.1. Vilken/vilka av följande matriser är intensitetsmatris till en Markovprocess?

(a)

$$A_1 = \begin{pmatrix} -3 & 4 & 0 \\ 1 & -2 & 1 \\ 1 & 2 & -3 \end{pmatrix},$$

(b)

$$A_2 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -2 & 2 \\ 1 & 1 & -2 \end{pmatrix},$$

(c)

$$A_3 = \begin{pmatrix} -3 & 2 & 1 \\ 2 & -4 & 2 \\ 1 & 1 & -2 \end{pmatrix}.$$

5.2. Kunder anländer till en affär med ett betjäningsorgan i grupper om X, där $p_X(1) = 1/2$, $p_X(2) = p_X(3) = 1/4$. Grupperna anländer med intensitet λ och betjäningstiderna har väntevärde μ^{-1}. Inför lämpliga tillstånd och ställ upp intensitetsmatrisen. (Det förutsätts att kunderna betjänas individuellt, oavsett vilken grupp de anländer med.)

5.3. En Markovprocess $\{X(t),\ t \geq 0\}$ i kontinuerlig tid har intensitetsmatris

$$A = \begin{pmatrix} -2 & 2 \\ 1 & -1 \end{pmatrix}.$$

Den diskret samplade processen $\tilde{Y}(n) = X(nh)$, $n = 0, 1, 2, \ldots$ blir en Markovkedja. Bestäm dess övergångsmatris.

5.4. Har Markovprocessen med

$$A = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}$$

någon asymptotisk fördelning? Har den inbäddade Markovkedjan någon asymptotisk fördelning?

5.5. Existerar en asymptotisk fördelning till Markovprocessen med

$$A = \begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & -2 & 0 & 2 \\ 3 & 0 & -3 & 0 \\ 0 & 2 & 0 & -2 \end{pmatrix}?$$

5.6. Låt $\tilde{X}(n)$ beteckna den inbäddade Markovkedjan till en Markovprocess med

$$A = \begin{pmatrix} -1 & 1/2 & 1/2 \\ 2 & -3 & 1 \\ 1 & 2 & -3 \end{pmatrix},$$

och hopptidpunkter $T_1, T_2, \ldots$. Bestäm

$$\mathsf{P}(T_1 > 1,\, \tilde{X}(1) = E_2,\, \tilde{X}(2) = E_3 \,|\, \tilde{X}(0) = E_1).$$

5.7. Fyra komponenter I, II, III och IV går sönder med intensiteterna λ, 2λ, 3λ respektive 4λ. Då en trasig komponent repareras är de övriga ur drift (så att aldrig mer än en komponent är trasig samtidigt). Lagningsintensiteten för respektive komponent är μ, 2μ, 3μ respektive 4μ. Definiera en lämplig Markovprocess med tillhörande inbäddad Markovkedja, och besvara följande: Vilken är sannolikheten för respektive maskin att just den går sönder först? Hur lång tid tar det i genomsnitt tills den första maskinen är lagad? (*Ledning*: Den andra frågan besvaras lämpligen med hjälp av den första och betingning.)

5.8. Anta att en Markovprocess med intensitetsmatris $A = (a_{ij})$ har en entydig stationär fördelning $\pi = (\pi_1, \pi_2, \ldots)$. Visa att vektorn $\tilde{\pi}$, med komponenter

$$\tilde{\pi}_i = \frac{\pi_i(-a_{ii})}{\sum_k \pi_k(-a_{kk})},$$

är den stationära fördelningen till den inbäddade Markovkedjan. Räkna ut π och $\tilde{\pi}$ då

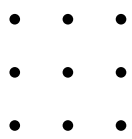
$$A = \begin{pmatrix} -6 & 3 & 3 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix}.$$

5.9. En partikels rörelse mellan tre tillstånd beskrivs av en Markovprocess med intensitetsmatris

$$A = \begin{pmatrix} -2 & 2 & 0 \\ 2 & -3 & 1 \\ 1 & 1 & -2 \end{pmatrix}.$$

Givet start i E_1, hur lång tid tar det i genomsnitt att nå E_3?

5.10. I kvadraten nedan



startar man i mitten, och kan sedan förflytta sig till alla grannar i horisontell och vertikal led med intensitet λ.

(a) Hur lång tid tar det i genomsnitt att nå en hörnpunkt? (*Ledning*: Av symmetriskäl räcker det att införa tre tillstånd.)

(b) Samma fråga, om man inte kan återvända till mittpunkten.

5.11. (*) Betrakta en födelsedödsprocess med konstanta födelse- och dödsintensiteter $\lambda_0 = \lambda_1 = \ldots = \mu_1 = \mu_2 = \ldots \equiv \lambda$. Om man startar i E_0, hur lång tid tar det i genomsnitt att nå E_N?

Kapitel 6

Tillförlitlighet

6.1 Inledning

Ett system av något slag, t.ex. ett datornät, sägs vara tillförlitligt om det fungerar när man behöver det. I *statistisk tillförlitlighetsteori* försöker man bygga upp modeller för att beräkna tillförlitligheten hos system. Det kan gälla hur variationer i kvalitet hos enskilda komponenter påverkar en telefonväxels tillförlitlighet eller hur vindstyrkan (som ett exempel på en stokastisk process) påverkar risken för kollaps hos en oljeplattform i Nordsjön. I andra sammanhang gäller det att utforma en optimal underhållsstrategi för en viss maskin: om man byter komponenter för ofta blir det dyrt och byter man för sällan kan allvarliga fel uppstå, vilket orsakar extrakostnader.

Tillförlitlighetsteori i modern statistisk mening är en relativt ung vetenskap: först under andra världskriget sköt utvecklingen fart på allvar. Ett av de första exemplen på tillförlitlighetsproblem var de tyska VI-raketerna. Trots att de enskilda komponenterna var av god kvalitet, hade raketerna ofta låg tillförlitlighet. Detta ledde till ett intresse för att bygga upp en matematisk teori som beskrev hur ett komplicerat systems funktionssannolihet beror av de enskilda komponenternas egenskaper. Andra tillämpningar som fört tillförlitlighetsteorin framåt är rymdfartsprogrammen med deras extrema krav på driftssäkerhet samt tele- och datorteknikens utveckling mot alltmer komplicerade system där kraven på perfekt funktion är höga.

Vi låter $\{X(t);\, t \geq 0\}$ vara den stokastiska process (ej nödvändigtvis Markov) som beskriver systemets utveckling i tiden. Vid varje tidpunkt gäller att

$$X(t) \in \mathcal{G} = \{E_1, E_2, \ldots\},$$

och vi tänker oss att tillståndsrummet kan delas upp i två disjunkta komponenter, $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$, där $\mathcal{G}_1$ och $\mathcal{G}_2$ svarar mot de tillstånd då systemet fungerar respektive är ur funktion. Vidare antas $X(0) \in \mathcal{G}_1$, dvs systemet fungerar vid tidpunkten 0. För ett *underhållet system* utförs reparation så snart $X(t) \in \mathcal{G}_2$ (eller även tidigare, i förebyggande syfte). För ett *icke-underhållet system* är i stället $\mathcal{G}_2$ en absorberande grupp av tillstånd. Det enklaste exemplet är livslängdsprocesser, för vilka $\mathcal{G}_1$ och $\mathcal{G}_2$ består av vardera ett tillstånd.

6.2 Icke-underhållna system

6.2.1 Funktionssannolikhet, intensitetsfunktion och förväntad livslängd

Livslängden för ett icke-underhållet system ges av

$$Y = \inf\{t \geq 0;\, X(t) \in \mathcal{G}_2\}.$$

Oftast beskrivs kontinuerliga stokastiska variabler med hjälp av sin fördelningsfunktion och täthetsfunktion. För livslängder är det ofta mer praktiskt att räkna med *funktionssannolikheten* (överlevnadsfunktionen) och *intensitetsfunktionen*, som båda finns definierade i Blom A, Avsnitt 12.3. För fullständighets skull ger vi definitionerna även här. Om livslängden beskrivs av den s.v. Y ges funktionssannolikheten av

$$R_Y(t) = \mathsf{P}(\text{"komponenten fungerar vid tiden } t\text{"}) = \mathsf{P}(Y > t) = 1 - F_Y(t)$$

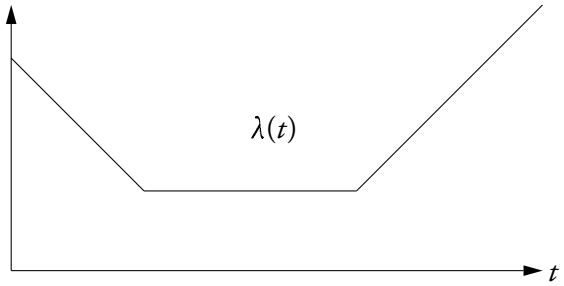
medan intensitetsfunktionen definieras som

$$\lambda(t) = \lim_{h \to 0} \frac{\mathsf{P}(Y < t + h \,|\, Y > t)}{h} = \frac{f_Y(t)}{1 - F_Y(t)} = \frac{f_Y(t)}{R_Y(t)}. \tag{6.1}$$

Omvänt kan vi uttrycka funktionssannolikheten med hjälp av intensitetsfunktionen enligt

$$R_Y(t) = \exp\left(- \int_0^t \lambda(u)\, du \right).$$

Ofta försöker man utgå från hur $\lambda(t)$ kan tänkas bero av tiden när man skall försöka föreslå en lämplig fördelning för livslängderna. Ibland kan man anta att komponenterna åldras, dvs intensitetsfunktionen växer med tiden. Sådana komponenter säges vara IFR ("Increasing Failure Rate"). På motsvarande sätt kallas komponenter med avtagande intensitetsfunktion DFR ("Decreasing Failure Rate"). En avtagande intensitetsfunktion kan synas långsökt, men uppträder oftare än man tror. Ett exempel är då systemet är en blandning av dåliga och bra komponenter: först går alla de dåliga komponenterna sönder och därefter avtar intensitetsfunktionen på grund av att endast de bra komponenterna återstår. En tredje typ av intensitetsfunktion är de som följer en s.k. badkarskurva, se Figur 6.1:



Figur 6.1: *Badkarskurva.*

Intensitetsfunktionen är först avtagande, men när barnsjukdomarna rättats till är den relativt konstant för att sedan öka då åldringsfenomen sätter in.

Vi ska börja med att ge två standardexempel på hur funktionssannolikheten för hela systemet kan beräknas om den är känd för systemets komponenter. Mer komplicerade system kan i princip erhållas som kombinationer av dessa.

Exempel 6.1 (Seriekoppling av n oberoende komponenter.) Vid seriekoppling fungerar systemet om och endast om alla komponenterna fungerar. Vi får

$$R_Y(t) = R_1(t)R_2(t) \cdot \ldots \cdot R_n(t),$$

där $R_i(t)$ är funktionssannolikheten för den i:te komponenten. Livslängden är minimum av de enskilda komponenternas livslängder och

$$\lambda(t) = \sum_{i=1}^{n} \lambda_i(t).$$

□

Exempel 6.2 (Parallellkoppling av n oberoende komponenter.) Vid parallellkoppling fungerar systemet om minst en av komponenterna fungerar. Vi får

$$1 - R_Y(t) = (1 - R_1(t))(1 - R_2(t)) \cdot \ldots \cdot (1 - R_n(t)).$$

Här är livslängden maximum av livslängden för de enskilda komponenterna. Däremot får $\lambda(t)$ inte något enkelt generellt utseende. □

Exempel 6.3 (Exponentialfördelad livslängd.) Om $Y \in Exp(1/\lambda)$, ger definitionen av exponentialfördelningen och (6.1) att

$$\lambda(t) = \frac{f_Y(t)}{R_Y(t)} = \frac{\lambda e^{-\lambda t}}{1 - (1 - e^{-\lambda t})} = \lambda.$$

Exponentialfördelningen karakteriseras alltså av konstant intensitetsfunktion och är den enda fördelning som är både IFR och DFR. □

Exempel 6.4 (Weibullfördelad livslängd.) Definitionen av Weibullfördelningen ges i Blom A, sid 70. Insättning i (6.1) ger

$$\lambda(t) = \frac{\frac{c}{a}(\frac{t}{a})^{c-1} \exp(-(t/a)^c)}{1 - (1 - \exp(-(t/a)^c))} = \frac{c}{a}(\frac{t}{a})^{c-1}.$$

Här är tydligen fördelningen IFR om formparametern $c \geq 1$ och DFR om $c \leq 1$. Exponentialfördelningen svarar mot $c = 1$. □

Att Weibullfördelningen kan beskriva en komponents åldrande oavsett om $\lambda(t)$ växer eller avtar är en av anledningarna till dess stora popularitet. En annan orsak till att den är vanlig i tillförlitlighetssammanhang är att minimum av n oberoende likafördelade Weibullvariabler också är Weibullfördelad. Detta gör att fördelningen används för att beskriva styrkan hos material, eftersom ett konstuktionselements styrka ofta beror på styrkan hos dess svagaste länk, dvs minimum av de olika komponenternas styrkor. Eftersom det är lämpligt att en del av en balk och hela balken beskrivs av samma klass av fördelningar, är Weibullfördelningen lämplig. Praktiska försök har visat att denna teori stämmer väl i praktiken för t.ex. konstruktionsändamål. Den kan däremot inte användas då principen om svagaste länk inte gäller, t.ex. för fibermaterial, där brott på en fiber bara resulterar i att de andra fibrerna övertar lasten.

Ett annat viktigt begrepp är den förväntade livslängden $\mathsf{E}(Y)$, som ofta kallas MTTF (”Mean Time To Failure”). Eftersom livslängden är en positiv s.v. kan vi använda den formel som anges för väntevärdet i Blom A, sid 118, nämligen

$$\text{MTTF} = \mathsf{E}(Y) = \int_0^\infty R_Y(t)\, dt. \tag{6.2}$$

Eftersom det i tillförlitlighetssammanhang oftast är $R_Y(t)$ som är känd, är (6.2) mycket praktisk. För Markovprocesser kan alternativt Sats 5.8 användas för bestämning av den förväntade livslängden.

6.2.2 Redundans

Ett grundläggande förfarande inom tillförlitlighetsteorin är att öka säkerheten genom att införa reserver, s.k. redundanta komponenter. Detta är viktigt i många tekniska sammanhang där hög tillförlitlighet krävs, t.ex. konstruktion av telefonväxlar, reglersystem, datorer etc. Man skiljer på reservkomponenter som är inkopplade hela tiden (t.ex. parallellkoppling), *aktiv redundans*, och sådana som kopplas in först då den ordinarie komponenten går sönder, *passiv redundans*. Det finns också många gränsfall mellan dessa, det kan ju tänkas att reserven åldras lite även när den inte är inkopplad, s.k. ljum reserv. Vi ska nedan gå igenom några principiellt intressanta exempel:

Exempel 6.5 (Aktiv redundans.) För att åstadkomma ökad driftssäkerhet ersätts en komponent med n parallellkopplade komponenter. Livslängden för varje komponent antas vara exponentialfördelad med intensitet λ. Systemets funktionssannolikhet ges då av

$$R_Y(t) = 1 - (1 - \exp(-\lambda t))^n,$$

och efter insättning i (6.2),

$$\text{MTTF} = \int_0^\infty R_Y(t)\,dt = \left(1 + \frac{1}{2} + \ldots + \frac{1}{n}\right)\frac{1}{\lambda} \approx \frac{\ln n}{\lambda}. \tag{6.3}$$

I Blom A, sid 279, anges hur man visar den sista formeln med induktionsresonemang. Eftersom Y är maximum av ett antal oberoende och exponentialfördelade s.v. kan man alternativt utnyttja representationen

$$Y = \tau_{11} + \tau_{12} + \ldots + \tau_{1n},$$

där τ_{1i} är tiden mellan att komponent nr $i-1$ och i går sönder. Man kan visa att $\{\tau_{1i}\}$ är oberoende, med $\tau_{1i} \in Exp(1/((n-i+1)\lambda))$, och då följer (6.3) lätt genom väntevärdesbildning. Orsaken till att τ_{1i} har denna fördelning är att då $i-1$ komponenter gått sönder finns det $n-i+1$ stycken fungerande kvar, och intensiteten med vilken den första av dessa går sönder är $(n-i+1)\lambda$.

Låt oss även visa hur man med hjälp av Sats 5.8 kan gå tillväga: Inför tillstånden

$$E_i : \text{"}i-1 \text{ enheter trasiga"}, \quad i = 1, \ldots, n+1.$$

Vi får en Markovprocess som startar i E_1 och har intensitetsmatrisen

$$A = \begin{pmatrix} -n\lambda & n\lambda & 0 & \ldots & 0 & 0 \\ 0 & -(n-1)\lambda & (n-1)\lambda & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda & \lambda \\ 0 & 0 & 0 & \ldots & 0 & 0 \end{pmatrix}.$$

Vi sätter $\mathcal{G}_1 = \{E_1, \ldots, E_n\}$ och $\mathcal{G}_2 = \{E_{n+1}\}$. Då ges alltså $\tilde{A}$ av den övre $n \times n$ delmatrisen av A, och enligt Sats 5.8 ges livslängden $\mathsf{E}(Y)$ av första radsumman i $-\tilde{A}^{-1}$. Enligt Sats (5.8) fås $(\mathsf{E}(\tau_{11}), \ldots, \mathsf{E}(\tau_{1n}))$ som första raden i $-\tilde{A}^{-1}$, och den bestämmer man genom att lösa

$$(\mathsf{E}(\tau_{11}), \ldots, \mathsf{E}(\tau_{1n}))\tilde{A} = (-1, 0, \ldots, 0),$$

dvs

$$\begin{cases} \mathsf{E}(\tau_{11}) = \dfrac{1}{n\lambda}, \\ \mathsf{E}(\tau_{1i}) = \dfrac{n-i+2}{n-i+1}\,\mathsf{E}(\tau_{1,i-1}), \quad i = 2, \ldots, n \end{cases} \iff \mathsf{E}(\tau_{i1}) = \frac{1}{(n-i+1)\lambda}, \quad i = 1, \ldots, n.$$

Summering $\mathsf{E}(Y) = \sum_{i=1}^n \mathsf{E}(\tau_{i1})$ ger samma värde som i (6.3). □

Exempel 6.6 (Kall reserv eller passiv redundans.) I den enklaste modellen för redundans åldras inte reserven när den ej är inkopplad, s.k. *kall reserv*. Livslängden för systemet blir då summan av de enskilda komponenternas livslängder, och då är MTTF lätt att räkna ut. I föregående exempel med exponentialfördelade livslängder fås exempelvis

$$\text{MTTF} = \sum_{i=1}^n \mathsf{E}(Y_i) = \frac{n}{\lambda},$$

där Y_i är livslängden för den i:te komponenten. Vi ser att systemets genomsnittliga livslängd är en faktor n större en de enskilda komponenternas livslängd, jämfört med (approximativt) ln n i föregående exempel. Den passiva redundansen ger i detta fall betydligt större livslängd hos systemet, till priset av det arbete (och extrakostnad) som det innebär att kontinuerligt byta ut komponenter. □

Exempel 6.7 (Ljum reserv.) Anta att en komponent har exponentialfördelad livslängd med intensiteten λ då den är inkopplad samt att den åldras med en lägre intensitet λ_0 då den inte är inkopplad. Man talar då om en *ljum reserv*. (Det innebär exempelvis att en komponent kan vara trasig då den kopplas in.) Beräkning av MTTF sker nu exempelvis med hjälp av Sats 5.8 i Kapitel 5. Låt oss illustrera beräkningarna då $n = 3$. Vi inför tillstånden

E_0 : Bägge komponenterna hela.
E_1 : En komponent hel.
E_2 : Bägge komponenterna sönder.

Det ger intensitetsmatrisen

$$A = \begin{pmatrix} -(\lambda+\lambda_0) & \lambda+\lambda_0 & 0 \\ 0 & -\lambda & \lambda \\ 0 & 0 & 0 \end{pmatrix}.$$

och

$$-\tilde{A}^{-1} = \begin{pmatrix} \dfrac{1}{\lambda+\lambda_0} & \dfrac{1}{\lambda} \\ 0 & \dfrac{1}{\lambda} \end{pmatrix}.$$

Bilda sedan första radsumman hos $-\tilde{A}^{-1}$;

$$\text{MTTF} = \frac{1}{\lambda+\lambda_0} + \frac{1}{\lambda}.$$

Observera specialfallen $\lambda_0 = 0$ (passiv redundans) och $\lambda_0 = \lambda$ (aktiv redundans). □

6.3 Underhållna system

I ett underhållet system utförs reparation då det inte fungerar. Ofta används beteckningen MTTR (”Mean Time To Repair”) för väntevärdet av reparationstiden D (eller allmännare, väntevärdet av tiden i $\mathcal{G}_2$, då ju reparationer även kan utföras då systemet fungerar). På samma sätt blir MTTF väntevärdestiden av tiden i $\mathcal{G}_1$ mellan två perioder då systemet inte fungerar. Väntevärdet av tiden mellan två fel kallas MTBF (”Mean Time Between Failures”). Vi har alltså

$$\begin{aligned} \text{MTTR} &= \mathsf{E}(D) = \int_0^\infty R_D(t)\,dt \\ \text{MTBF} &= \text{MTTF} + \text{MTTR} \end{aligned}$$

En annan intressant storhet är *tillgängligheten* MTTF/MTBF, som anger hur stor andel av tiden systemet kan förväntas fungera. Under vissa allmänna villkor gäller

$$\frac{\text{MTTF}}{\text{MTBF}} = \lim_{t\to\infty} \sum_{E_i\in\mathcal{G}_1} \mathsf{P}(X(t) = E_i) = \sum_{E_i\in\mathcal{G}_1} \pi_i, \tag{6.4}$$

där $\pi = (\pi_i)_{i\in\mathcal{G}}$ är den asymptotiska fördelningen. Denna formel motiveras i Blom A, Avsnitt 14.2, då $\mathcal{G}_1$ och $\mathcal{G}_2$ innehåller vardera ett tillstånd och de successiva gång- och reparationstiderna är oberoende.

Exempel 6.8 (Stokastisk reparationstid.) Låt livslängden hos en komponent vara Weibullfördelad. Då erhålles

$$\begin{aligned} \text{MTTF} &= \int_0^\infty R_Y(t)\,dt = \int_0^\infty \exp(-(t/a)^c)\,dt = [x = (t/a)^c] \\ &= \frac{a}{c}\int_0^\infty x^{(1/c)-1}\exp(-x)\,dx = \frac{a}{c}\,\Gamma(1/c) = a\Gamma(1+1/c). \end{aligned}$$

(Se Blom A, sid 71, för definitionen av Gammafunktionen $\Gamma(p)$.) Om reparationstiderna är exponentialfördelade med intensitet μ blir MTTR $= 1/\mu$ och tillgängligheten

$$\pi_A = \frac{a\Gamma(1+1/c)}{a\Gamma(1+1/c)+1/\mu}.$$

□

För mer komplicerade system kan $\mathcal{G}_1$ och $\mathcal{G}_2$ innehålla flera undertillstånd (ett system kan ju gå sönder på flera olika sätt eller fungera men ändå ha en eller flera trasiga komponenter). Då är i allmänhet inte de successiva livslängderna och reparationstiderna oberoende och likafördelade, t.ex. beror ju reparationstiden på vilken komponent det var som gick sönder. Uttrycket för tillgängligheten blir mer komplicerat, men för Markovsystem (dvs då livslängden hos enskilda komponenter och reparationstider är exponentialfördelade) kan den bestämmas såväl momentant vid tiden t som asymptotiskt då $t \to \infty$ med hjälp av satserna 5.4 och 5.5 i Kapitel 5. Dessa båda satser anger ju sannolikheten att systemet befinner sig i ett visst tillstånd vid tiden t respektive vid jämvikt. Sannolikheten för att systemet ska fungera fås sedan genom att summera sannolikheterna för tillstånden i $\mathcal{G}_1$ (jfr (6.4)).

Vidare kan Sats 5.8 i Kapitel 5 användas för att bestämma MTTF, man låter helt enkelt $\mathcal{G}_2$ svara mot alla de tillstånd som gör att systemet inte fungerar. Genom att betinga på vilket tillstånd i $\mathcal{G}_1$ man hoppade till efter föregående reparation fås

$$\text{MTTF} = \sum_{i\in\mathcal{G}_1} \pi_i \mathsf{E}(Y_i),$$

där $\mathsf{E}(Y_i)$ bestäms ur Sats 5.8 och $\{\pi_i\}_{i\in\mathcal{G}_1}$ är sannolikhetsfördelningen över $\mathcal{G}_1$ *precis efter ett hopp från* $\mathcal{G}_2$ (dvs *inte* jämviktsfördelningen, då ju processen även kan befinna sig i $\mathcal{G}_2$ vid jämvikt).

Exempel 6.9 En fabrik har tre maskiner som går sönder oberoende av varandra med intensiteten λ. Vid fabriken finns en reparatör som reparerar maskinerna med den konstanta reparationsintensiteten μ. Hur stor andel av tiden kommer alla maskiner att vara i drift, och hur lång tid tar det i genomsnitt tills alla maskiner går sönder, förutsatt att alla maskiner är hela då fabriken startar?

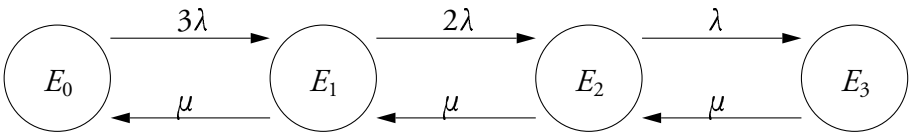
För att lösa problemet inför vi fyra tillstånd

$$E_i = \text{"}i \text{ maskiner är trasiga"}, \quad i = 0,1,2,3.$$

Övergångsintensiteten mellan E_0 och E_1 är 3λ, ty minimum av tre oberoende exponentialfördelade variabler med intensitet λ blir en ny exponentialfördelad variabel med intensitet 3λ. På samma sätt blir intensiteterna för övergångarna $E_1 \to E_2$ och $E_2 \to E_3$ 2λ respektive λ, se Figur 6.2.

Intensitetsmatrisen blir

$$A = \begin{pmatrix} -3\lambda & 3\lambda & 0 & 0 \\ \mu & -(\mu+2\lambda) & 2\lambda & 0 \\ 0 & \mu & -(\mu+\lambda) & \lambda \\ 0 & 0 & \mu & -\mu \end{pmatrix}.$$



Figur 6.2: *Tillståndsdiagram för Exempel 6.9.*

Låt π_i beteckna P("systemet befinner sig i E_i vid jämvikt"). Nu ger Sats 5.5 i Kapitel 5 att radvektorn $\pi = (\pi_0, \pi_1, \pi_2, \pi_3)$ uppfyller $\pi A = 0$. Med $\rho = \lambda/\mu$ har detta ekvationssystem lösningen

$$\pi = (\pi_0,\ 3\rho\pi_0,\ 6\rho^2\pi_0,\ 6\rho^3\pi_0).$$

Eftersom π är en sannolikhetsvektor måste $\sum_{i=0}^{3} \pi_i = 1$, vilket ger lösningen

$$\pi = \frac{(1,\ 3\rho,\ 6\rho^2,\ 6\rho^3)}{1 + 3\rho + 6\rho^2 + 6\rho^3}.$$

I själva verket är detta en födelsedödsprocess och ett specialfall av det kösystem som beskrivs i Blom A, sid 300, då det allmänna fallet med N maskiner behandlas.

Vi övergår nu till att bestämma medeltiden tills systemet går sönder och använder samma beteckningar som i Sats 5.8, Kapitel 5. Vi delar in tillstånden i två grupper $\mathcal{G}_1 = \{E_0, E_1, E_2\}$ och $\mathcal{G}_2 = \{E_3\}$. Då fås $\tilde{A}$ som den övre 3×3 delmatrisen till A. Sedan löser vi ekvationssystemet

$$(\mathsf{E}(\tau_{00}),\ \mathsf{E}(\tau_{01}), \mathsf{E}(\tau_{02}))\,\tilde{A} = (-1, 0, 0)$$

för att beräkna den första raden ($\mathsf{E}(\tau_{00})$, $\mathsf{E}(\tau_{01})$, $\mathsf{E}(\tau_{02})$) i $\tilde{A}^{-1}$. Här är τ_{0i} den tid som tillbringas i E_i innan systemet går sönder (dvs vi hamnar i E_3) givet att vi startar i E_0. Den förväntade tiden tills alla maskiner går sönder blir nu

$$\mathsf{E}(Y_0) = \mathsf{E}(\tau_{00}) + \mathsf{E}(\tau_{01}) + \mathsf{E}(\tau_{02}) = \frac{1}{\lambda} \cdot \frac{1 + 6\rho + 11\rho^2}{6\rho(1+\rho)}. \tag{6.5}$$

Notera specialfallet $\mu = 0$, som svarar mot parallellkoppling av tre komponenter (maskiner). Enligt (6.3) fås

$$\mathsf{E}(Y_0) = \frac{1}{\lambda}\left(1 + \frac{1}{2} + \frac{1}{3}\right) = \frac{11}{6\lambda},$$

vilket överensstämmer med gränsvärdet av (6.5) då $\rho \to \infty$.

Observera slutligen att $\mathsf{E}(Y_2)$, inte $\mathsf{E}(Y_0)$, ger MTTF då systemet svängt in sig i jämvikt (här avses att systemet fungerar då någon maskin är i drift). Orsaken är att när vi lämnar $\mathcal{G}_2$ hoppar vi alltid till E_2, och då måste tiden tills nästa gång vi hamnar i $\mathcal{G}_2$ att ges av Y_2. Vi överlåter åt den intresserade läsaren att bestämma $\mathsf{E}(Y_2)$ på samma sätt som (6.5). □

6.4 Övningsuppgifter

6.1. Ett system består av n komponenter varav en aktiv (felintensitet λ) och $n-1$ ljumma reserver (felintensitet λ_0, $0 < \lambda_0 < \lambda$). Så snart en aktiv komponent går sönder ersätts den av en reserv (som då blir aktiv). Vilken genomsnittlig livslängd har systemet? (*Ledning*: Inför E_i = "i komponenter trasiga", $i = 0, 1, \ldots, n$.)

6.2. En fabrik har två maskiner, som går sönder med intensiteterna λ_1 respektive λ_2. Den ende reparatören lagar maskinerna med intensitet μ, och om båda är trasiga reparerar han alltid maskin 1 först (dvs han avbryter reparation av 2 om 1 går sönder). Hur lång tid tar det i genomsnitt från det att båda maskinerna är trasiga tills det att de är hela igen?

Kapitel 7

Svar till övningsuppgifter

4.1. $f_{00}^{(n)} = pq^{n-1}$, $n = 1, 2, \ldots$,
$\mu_0 = p^{-1}$,
$\pi_i = pq^i$, $i = 0, 1, 2, \ldots$

4.2. 0.03456

4.3. Alla tillstånd är positivt beständiga.

4.4. $\mathsf{E}(Y_{11}) = \pi_1^{-1} = 4$

4.5. I (b) och (d) existerar en asymptotisk fördelning.

4.6. I (b), (c) och (d) är den stationära fördelningen entydig.

4.7. $\pi = (1/3, 1/6, 1/6, 1/3)$. Asymptotisk fördelning existerar ej.

4.10. $3/2$

4.11. 5

4.12. N^2

5.1. Endast (c) är en intensitetsmatris.

5.2. Med E_i = "i kunder i affären", $i = 0, 1, 2, \ldots$, fås $A = (a_{ij})$, där $a_{00} = -\lambda$, $a_{01} = \lambda/2$, $a_{02} = a_{03} = \lambda/4$, och för $i \geq 1$; $a_{i,i-1} = \mu$, $a_{ii} = -(\lambda + \mu)$, $a_{i,i+1} = \lambda/2$, $a_{i,i+2} = a_{i,i+3} = \lambda/4$.

5.3. $P = \begin{pmatrix} \frac{1}{3} + \frac{2}{3}\,\mathrm{e}^{-3h} & \frac{2}{3} - \frac{2}{3}\,\mathrm{e}^{-3h} \\ \frac{1}{3} - \frac{1}{3}\,\mathrm{e}^{-3h} & \frac{2}{3} + \frac{1}{3}\,\mathrm{e}^{-3h} \end{pmatrix}$

5.4. Asymptotisk fördelning existerar endast i kontinuerlig tid.

5.5. Asymptotisk fördelning existerar ej.

5.6. $\mathrm{e}^{-1}/6$

5.7. P("maskin i går sönder först") $= i/10$.
Förväntad tid tills första reparation är klar: $\dfrac{1}{10\lambda} + \dfrac{2}{5\mu}$.

5.8. $\pi = (1/7, 3/7, 3/7)$,
$\tilde{\pi} = (1/3, 1/3, 1/3)$.

5.9. $5/2$

5.10. (a) $7\lambda^{-1}/8$, (b) $3\lambda^{-1}/4$

5.11. $\lambda^{-1}N(N+1)/2$

6.1. $\displaystyle\sum_{i=0}^{n-1} \frac{1}{\lambda + i\lambda_0}$

6.2. $\dfrac{2}{\mu} + \dfrac{\lambda_1}{\mu^2}$

Våren 2001

Matematisk statistik
Matematikcentrum
Lunds universitet
Box 118, 221 00 Lund

http://www.maths.lth.se/