Introduktion til Sandsynlighed og Statistik til modellering og simulering

Datalogi og Software 1. semester

Tobias Kallehauge

Aalborg Universitet, November 14, 2020

Denne note omhandler sandsynlighed, statistik og hvordan disse kan bruges i praksis specielt i sammenhæng med simuleringer, hvor det ofte er nødvendigt at udtrække tal fra bestemte sandsynlighedsfordelinger. Der vil primært fokuseres på den matematiske forståelse, men til sidst i noten gives et par eksempler på hvordan teorien kan bruges i et C-program. Formålet med noten er at give den *nødvendige* teori til software- og datalogiprojekter på 1 semester, og teorien gennemgås derfor relativt overfladisk med kun få eksempler. Den primære kilde bag noten er [1], som er gratis tilgængelig på aub.aau.dk. Heri findes også beviser for resultaterne, der er udeladt i noten. Har I brug for at henvise til teorien præsenteret her anbefales i derfor også at henvise til [1] fremfor noten her. Forslag til ændringer, stavefejl mm. kan sendes til tkal@es.aau.dk.

1 Introduktion til sandsynlighed

Sandsynlighed er et begreb der (mis)bruges i mange sammenhænge indenfor videnskab, økonomi, politik, osv., men helt grundlæggende er det en matematisk konstruktion med en præcis definition. For at måle sandsynlighed indføres sandsynlighedsmålet P, en funktion der bestemmer sandsynligheden for udfald i *tilfældige eksperimenter*. Hvis et tilfældigt eksperiment har N mulige udfald $A_1, A_2, \ldots, A_N$ da vil $P(A_i)$ være sandsynligheden for udfaldet A_i.

Eksempel 1.1. Vi kaster en fair 6-siddet terning med mulige udfald

$$A_1 = \text{“slå en 1’er”}, \quad A_2 = \text{“slå en 2’er”}, \quad \ldots \quad , A_6 = \text{“slå en 6’er”}$$

Sandsynligheden for alle udfald er lige stor og vi har eksempelvis at

$$P(A_6) = P(\text{“slå en 6’er”}) = 1/6 \approx 0.17\%.$$

Funktionen P er defineret ud fra en række matematiske egenskaber. Vigtigst er at den altid antager værdier mellem 0 og 1 således $0 \leq P(A) \leq 1$ for ethvert udfald A i et tilfældigt eksperiment. Se den fulde definition i [1, sektion 1.3].

For at formalisere notationen i tilfældige forsøg indfører vi begrebet *tilfældige variable*. En tilfældig variabel er en funktion der omsætter udfald i et tilfældigt forsøg til talværdier.

Eksempel 1.2. En tilfældig variabel X tæller antal gange en mønt lander på krone i løbet af 3 kast. Vi har så $X(PPK) = 1$, $X(KKP) = 2$, $X(KKK) = 3$ osv.

Eksempel 1.3. Lad X være en tilfældige variabel der beskriver talværdien af terningens udfald fra eksempel 1.1, Vi har således $X(\text{“slå en 1’er”}) = 1$, $X(\text{“slå en 2’er”}) = 2$, osv. Vi kan da skrive eksempelvis $P(X = 6) = 1/6$.

Hvordan man rent faktisk udregner sandsynligheder er en historie for en anden gang. Her vil vi nøjes med at se på eksempler hvor vi allerede kender sandsynlighederne for alle udfald karakteriseret ved såkaldte *fordelingsfunktioner*.

2 Fordelinger og fordelingsfunktioner

Før vi kan indføre fordelingsfunktioner skal vi kategorisere mellem *diskrete* og *kontinuerte* tilfældige variable. Eksemplerne vi har set indtil videre er diskrete da der er et *tælleligt* antal udfald. Det også muligt at have en diskret variabel med tælleligt uendelig mange udfald så længe man kan associere associere udfaldene med en tællelig mængde såsom de ikke negative heltal $\{0, 1, 2, \ldots\}$. Antallet af terningslag før der slås en 6’er et eksempel på en tællelig uendelig mænge da der ikke er en øvre grænse for antal slag. Kontinuerte tilfældige variable derimod er *utællelige* og er typisk associeret med de reelle tal $\mathbb{R}$ eller et interval heri. Eksempelvis er højden af en tilfældigt udvalgt person eller tiden det tager for et atom at henfalde radioaktivt begge kontinuerte tilfældige variable.

2.1 Diskrete fordelinger

Vi indfører nu *sandsynlighedsmassefunktionen*, på engelsk *probability mass funktion* (pmf). En pmf p_X er defineret som $p_X(x) = P(X = x)$ for en tilfældig variable X. Bemærk her at X er den tilfældige variabel mens x er et reelt tal.

Eksempel 2.1. I X fra eksempel 1.3 med terningkast er $p_X(x) = 1/6$ for $x = 1, 2, \ldots, 6$ og vi har eksempelvis $P(\text{“slå en 6'er”}) = P(X = 6) = p_X(6) = 1/6$.

Eksempel 2.2. For en tilfældig variabel X har vi givet pmf'en:

$$p_X(x) = \begin{cases} \frac{3}{6} & x = -1 \\ \frac{1}{6} & x = 0 \\ \frac{2}{6} & x = 1 \end{cases},$$

og vi har eksempelvis at $P(X = 0) = p_X(0) = \frac{1}{6}$. Selvom vi ikke ved noget om hvilket tilfældigt forsøg X stammer fra, ved vi ud fra p_X alt om hvordan X opfører sig fra et statistisk synspunkt. Vi siger derfor at p_X karakteriserer X fuldstændigt samt at X følger fordelingen for p_X.

Med en pmf kan man nemt lave beregninger, der omhandler delmængder af udfaldsrummet.

Eksempel 2.3. Givet pmf'en for X i eksempel 2.1 med terningkast har vi eksempelvis:

$$P(\text{“Slå mindst 3”}) = P(X \leq 3) = p_X(1) + p_X(2) + p_X(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$
$$P(\text{“Slå mere end 4”}) = P(X > 5) = p_X(5) + p_X(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$
$$P(\text{“Slå mellem 1 og 6”}) = P(X \in \{1, 2, 3, 4, 5, 6\}) = \sum_{x=1}^{6} p_X(x) = \sum_{x=1}^{6} \frac{1}{6} = 1$$

I det sidste eksempel udregnes sandsynligheden for hele udfaldsrummet “Slå mellem 1 og 6” til 1. Dette giver intuitivt mening, men er faktisk også en egenskab der definerer sandsynlighedsmålet.

2.2 Kontinuerte fordelinger

Ved kontinuerte fordelinger snakker man istedet for pmf'er om *sandsynlighedstæthedsfunktioner*, på engelsk *probability density function* (pdf). For at forstå forskellen ser vi på følgende eksempel.

Eksempel 2.4. En klassisk kontinuert fordelt tilfældig variabel er X fordelt mellem 0 og 1 med lige stor sandsynlighed for alle værdier. Med hvad er så sandsynligheden for en bestemt værdi i intervallet, eks. $P(X = 0.5)$? Svaret er faktisk 0, og det er fordi der er utælleligt mange mulige udfald og sandsynligheden for et specifikt udfald er altid 0. Spørger man om sandsynligheder for intervaller i stedet kan man dog få ikke-nul sandsynligheder. Intuitivt giver det eksempelvis mening at $P(0.25 \leq X \leq 0.5) = 0.25$, men for at komme frem til dette skal det bruges at X har pdf funktionen $p_X(x) = 1$ og vi får sandsynligheden ud fra følgende integrale:

$$P(0.25 \leq X \leq 0.5) = \int_{0.25}^{0.5} p_X(x)\ dx = \int_{0.25}^{0.5} 1\ dx = [x]_{x=0.5}^{0.25} = 0.5 - 0.25 = 0.25$$

For en kontinuert tilfældig variabel X med pdf p_X gælder der generelt:

$$P(a \leq X \leq b) = \int_a^b p_X(x)\ dx$$

Med denne teori har vi et grundlag for sandsynlighed og vi vil nu bevæge os over i statistik.

3 Forventet værdi, varians og standardafvigelse

I statistik forsøger vi at karakterisere fordelinger ud fra *statistikker*, der kan fortælle os om vigtige egenskaber for disse. Den *forventede værdi* er en sådan statistik, og som navnet hentyder fortæller den om det forventede udfald af et tilfældigt forsøg. For diskret tilfældig variabel X med mulige udfald $\{x_1, x_2, \ldots\}$ og pmf p_X er den forventede værdi defineret:

$$E[X] = \sum_{k=1}^{\infty} x_k \cdot p_X(x_k),$$

altså en vægtet sum over alle mulige udfald.

Eksempel 3.1. En skummel person på gaden tilbyder dig at gamble i et terningspil hvor du mister 1 krone hvis du slår mindre end 4, du vinder ingenting hvis du slår 4 og du vinder 1 krone hvis du slår 5 eller mere. Burde du spille med?
For at vurdere dette starter vi med at definere den tilfældige variabel X, der er -1 hvis du slår mindre end 4, 0 hvis du slår 4 og 1 hvis du slår 5 eller mere. Sandsynligheden for de tre udfald er henholdsvis 3/6, 1/6 og 2/6 og derfor er pmf funktionen den samme som i eksempel 2.2. Vi kan nu udregne det forventede udfald af spillet:

$$E[X] = -1 \cdot p_X(-1) + 0 \cdot p_X(0) + 1 \cdot p_X(1) = -1 \cdot \frac{3}{6} + 0 \cdot \frac{1}{6} + 1 \cdot \frac{2}{6} = -\frac{1}{6}$$

Du forventes altså at miste 1/6 krone hver gang du spiller og du anbefales herfra ikke at spille med. Bemærk som her, at det forventede udfald ikke nødvendigvis er blandt de mulige udfald.

For kontinuerte tilfældige variabel er den forventede værdi defineret ud fra et integrale, men fortolkningen er den samme. Hvis X er en kontinuert tilfældig variabel med mulige udfald i alle reelle tal og pdf p_X, da er den forventede værdi:

$$E[X] = \int_{-\infty}^{\infty} x \cdot p_X(x)\ dx$$

Eksempel 3.2. Den forventede værdi af X fra eksempel 2.4 med pdf $p_X(x) = 1$ for udfald mellem 0 og 1 er:

$$E[X] = \int_0^1 x \cdot 1\ dx = \left[\frac{1}{2}x^2\right]_{x=0}^1 = \frac{1}{2}1^2 - \frac{1}{2}0^2 = \frac{1}{2}$$

Det græske symbol μ bruges som oftest til at notere den forventede værdi altså $\mu = E[X]$. To andre meget brugbare statistikker er *varians* og *standardafvigelse*, der fortæller noget om hvor meget X afviger fra sin forventede værdi. Varians for en tilfældig variabel med forventet værdi μ er defineret som kvadratet af den forventede afvigelse fra μ:

$$\mathrm{Var}[X] = E[(X-\mu)^2] = \begin{cases} \sum_{k=1}^{\infty}(x_k - \mu)^2 p_X(x_k) & \text{for diskret } X \\ \int_{-\infty}^{\infty}(x-\mu)^2 p_X(x)\ dx & \text{for kontinuert} X \end{cases}$$

Varians er en positiv størrelse og er matematisk nem at arbejde med, men kan være lidt svær at fortolke. Hvis X eksempelvis er en vægt i gram (g) med forventet værdi $\mu = 4$ g, da er en varians på $\mathrm{Var}[X] = 4$ g^2 svær at fortolke. Derfor bruges standardafvigelsen, der fortæller om den forventede afvigelse fra middelværdien og er defineret ud fra varians:

$$\mathrm{Std}[X] = \sqrt{\mathrm{Var}[X]}$$

Eksempel 3.3. Variansen for spillet i eksempel 3.1 med $\mu = 1/6$ er:

$$\begin{aligned}\mathrm{Var}[X] &= (-1-\mu)^2 p_X(-1) + (0-\mu)^2 p_X(0) + (1-\mu)^2 p_X(1) \\ &= (-1+1/6)^2\frac{3}{6} + (1/6)^2\frac{1}{6} + (1+1/6)^2\frac{2}{6} = \frac{174}{216} \approx 0.81\end{aligned}$$

Og standard afvigelsen er $\mathrm{Std}[X] \approx 0.90$.

Eksempel 3.4. Variansen for X i eksempel 2.2 med forventet værdi $\mu = 1/2$ er:

$$\mathrm{Var}[X] = \int_0^1 (x-1/2)^2 \cdot 1\ dx = \left[\frac{1}{3}x^3 - \frac{1}{2}x^2 + \frac{1}{4}x\right]_{x=0}^1 = \frac{1}{12} \approx 0.08,$$

med standard afvigelse $\mathrm{Std}[X] = 1/\sqrt{12} \approx 0.29$.

Varians og standardafvigelse noteres typisk henholdsvis σ^2 og σ. Med forventet værdi, varians og standard afvigelse har vi de vigtigste statistikker til at forstå en lang række fordelinger. I næste afsnit skal vi se på en række specielle fordelinger.

4 Specielle fordelinger og modellering

Tobias

Indsæt kilder!

Det viser sig at mange tilfældige variable kan kategoriseres med en *parametrisk* fordeling hvor pmf/pdf'en er bestemt ud fra én eller flere parametre. Her vil vi se på nogle af de mest almindelige, hvornår disse optræder i virkeligheden og hvilke statistikker der karakteriserer dem.

4.1 Bernoulli fordelingen

En diskret tilfældig variabel med 0 og 1 som mulige udfald kaldes en Bernoulli fordeling og er karakteriseret ved sandsynligheden for udfaldet 1 ved parametren $p \in [0,1]$. Hvis X følger en Bernoulli fordeling med parametren p da er pmf'en:

$$p_X(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases},$$

og vi skriver $X \sim B(p)$ der betyder "X følger en Bernoulli fordeling med parameter p". Vi har $E[X] = p$ og $\mathrm{Var}[x] = p(1-p)$.

Eksempel 4.1. Det er givet at sandsynligheden for at en person en anden er $p = 0.25$. Hvis X("smitte overføres") $= 1$ og X("ingen smitte overføres") $= 0$, da har vi $X \sim B(0.25)$ med forventet værdi $E[X] = 0.25$ og standard afvigelse $\mathrm{Std}[X] = \sqrt{(0.25)(1-0.25)} \approx 0.43$. Man siger så at smitte overføres forventeligt 25% af gangene med en standard afvigelse på 43% altså $25\% \pm 43\%$. Se [1, pp. 111–112] for mere om Bernoulli fordelingen.

4.2 Uniform fordeling

Hvis en tilfældig variabel har lige stor sandsynlighed for alle udfald i et interval kaldes den uniform. I terning eksemplet introduceret eksempel 1.1 følger X er en diskret tilfældig variabel med en uniform fordelingen for udfaldene 1 til 6. Uniforme fordelinger ses dog typisk for kontinuerte tilfældige variable som i eksempel 2.4. Hvis X følger en uniform fordelingen inden for intervallet $[a,b]$ da gælder at:

$$p_X(x) = \frac{1}{b-a}, \quad x \in [a,b],$$

og vi skriver $X \sim \mathrm{unif}[a,b]$. Vi har $E[X] = (a+b)/2$ og $\mathrm{Var}[X] = (b-a)^2/12$. Se [1, pp. 90–92] for mere om den uniforme fordeling.

4.3 Poisson fordelingen

Poisson fordelingen er en diskret fordeling og er vigtig inden for simulering da den ofte ses for tilfældige variable der beskriver antallet af uforudsigelige hændelser inden for en tidsperiode. Typiske eksempler er antallet af jordskælv, bilulykker, antallet af stavefejl i en P1 rapport og besøg på en hjemmeside. Den originale brug af Poisson fordelingen var af Siméon Poisson, der opfandt fordelingen til at beskrive antallet af Preussiske soldater, der blev sparket ihjel af deres hest i det 19. århundrede. For at noget er Poisson fordelt er det vigtigt at der er et tilstrækkeligt tilfældigt element i udfaldet. Et eksempel som ankomst tidspunkter for busser vil derfor ikke være Poisson fordelt da tidsplanen fjerner det tilfældige element. Poisson fordelingen er karakteriseret ud fra parameteren λ og har mulige udfald i de positive heltal. Hvis X følger en Poisson fordeling med parameter λ da gælder at:

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

og vi skriver $X \sim \mathrm{Poi}(\lambda)$ (se figur 1). Poisson fordelingen har den specielle egenskab at variansen er den forventede værdi altså $E[X] = \lambda$ og $\mathrm{Var}[X] = \lambda$. Se [1, pp. 117–120] for mere om Poisson fordelingen.

Eksempel 4.2. I figur 2 ses nævnte eksempel med antallet af omkomne Preussiske soldater. Her sammenlignes et normaliseret histogram med Poisson fordelingen for $\lambda = 0.61$, og det ses at fordelingen passer godt i dette eksempel.

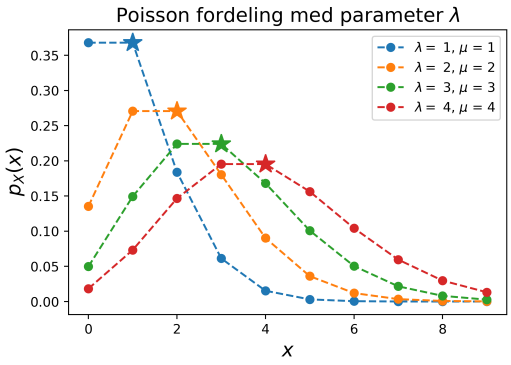

Figure 1: Poisson fordeling med forskellige parmetre λ. Stjernerne markerer forventet værdi.

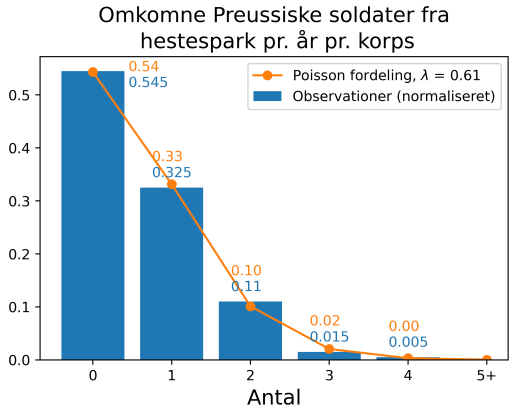

Figure 2: Data: [2]

4.4 Eksponentielfordeling

Eksponentielfordelingen er en kontinuert fordeling, som også er vigtig inden for simulering, da den kan bruges til at modellere tiden mellem tilfældige hændelser såsom jordskælv, ankomst af kunder i en butik og jobs til en server. Egenskaben der definerer at eksponentielfordelingen er at den er*hukommelsesløs* egenskab. Hvis X er en tilfældig variabel der fortæller om levealderen på eksempelvis en elektronisk genstand da kan den hukommelsesløse egenskab tolkes som at X ikke *ældes* i den forstand at sandsynligheden for at X overlever endnu et stykke tid ikke afhænger af dens nuværende alder. Formelt skrives den hukommelsesløse egenskab som:

$$P(X > x + y \text{ givet } X > y) = P(X > x)$$

Eksempel 4.3. I et vejkryds på landet ankommer der en bil en gang i timen gennemsnitligt. Dette vil være en hukommelsesløs process da sandsynligheden for at der ikke kommer en ny bil efter eksempelvis 20 minutter givet at der ikke er kommet en bil de første 10 minutter er det samme som sandsynligheden for at der ikke kommer en ny bil efter 10 minutter.

Det viser sig at tilfældige variable med den hukommelsesløse egenskab følger en eksponentielfordelingen. Hvis X følger en eksponentiel fordeling med parameter λ da er pdf'en:

$$p_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0,$$

og vi skriver $X \sim \exp(\lambda)$ (se figur 3). Vi har $E[X] = 1/\lambda$ og $\text{Var}[X] = 1/\lambda^2$, som giver en fortolkning for λ der kan refereres til som *fejlraten*. Se [1, pp. 123–127] for mere om eksponentielfordelingen.

Eksempel 4.4. Hvis T måler tiden i sekunder (s) mellem ankomsten af biler i eksempel 4.3, da vil T modelleres godt som en eksponentielfordeling med parameter $\lambda = 1/60^2$ s^{-1} og vi skriver $T \sim \exp(1/60^2 \text{ s}^{-1})$.

Den vågne læser har måske bemærket at der er en sammenhæng mellem eksponentielfordelingen, der modellerer tiden mellem tilfældige hændelser, og Poisson fordelingen der måler antallet af tilfældige hændelser inden for en tidsperiode. Denne sammenhæng vil blive belyst i sektion 5.

4.5 Normalfordeling

Sidst men ikke mindst har vi normalfordelingen. Normalfordelingen er en kontinuert tilfældig variabel, og er den mest brugte indenfor statistik. Den optræder eksempelvis ved støjen på en radiokanal, IQ for en befolkning, spændingen fra en strømforsyning og meget mere. Hvis man ikke kender til fordelingen for en tilfældig variabel vil statistikere ofte antage en normalfordeling. Fordelingen opstår når der måles en størrelse hvor mange tilfældige faktorer og støj påvirker målingen. Fordelingen er karakteriseret ud fra dens forventede værdi μ og varians σ^2, og hvis X følger en normalfordelingen med $E[X] = \mu$ og $\text{Var}[X] = \sigma^2$, da er pdf'en:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad x \in \mathbb{R},$$

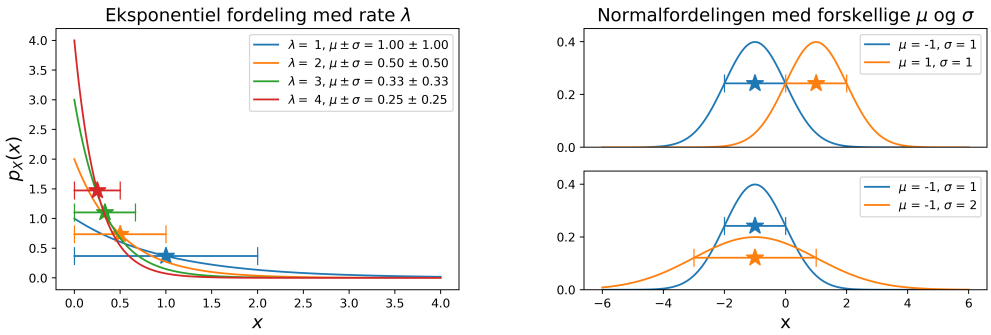

Figure 3: Eksponentiel og normal fordeling med forskellige λ. Stjernerne og tilhørende intervaller er forventet værdi $\pm$ standard afvigelse.

og vi skriver $X \sim \mathcal{N}(\mu, \sigma^2)$. Se figur 3 for hvordan μ og σ^2 har indflydelse på formen på pdf'en. Bemærk at normalfordelingen kan antage alle reelle tal $\mathbb{R}$. En process hvor negative tal ikke er mulige, såsom vindhastighed, er derfor ikke altid godt modeleret med normalfordelingen især hvis den forventede værdi er tæt på 0. Se [1, pp. 127–131] for mere om normalfordelingen.

5 Punktprocesser

I eksempel 4.4 med trafikken i et vejkryds blev tiden mellem to bilers ankomst modelleret med en eksponentielfordelt tilfældig variabel T med parametren $\lambda = 1/60^2$ s^{-1}. Lad nu $T_1, T_2, \ldots$ være forskellige tilfældige variable der modellerer tiden mellem hver bil over en tidsperiode, eksempelvis en dag, således $T_i \sim \exp(\lambda)$ for $i = 0, 1, \ldots$. Ankomsttidspunkter over en hel dag er da en såkaldt *punkt process*, hvor hvert punkt markerer ankomstentidspunktet for en bil (se figur 4).

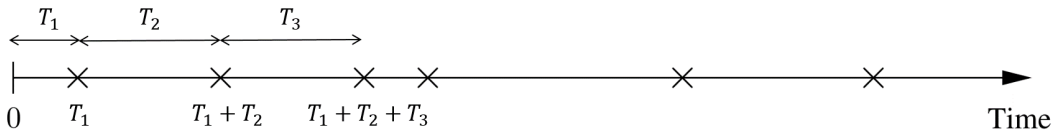

Figure 4: Punktprocess hvor $\times$ markerer hændelser i tid, eksempelvis ankomsttidspunkter af biler i et trafikkryds.

I det specielle tilfælde hvor $T_1, T_2, \ldots$ er eksponentielfordelte med parameter λ, da er punkt processen den såkaldte *Poisson process*. Poisson processen er nytting til at modellere processer hvor ankomster i tid sker tilfældigt og uafhængigt i tid. Andre eksempler kunne være ankomst af opgaver til en computerserver, kø på apoteket, tidspunkter for jordskælv, ulykker på motervej og henfald af radioaktive partikler.

Følgende resultat om Possion processer er nyttigt og giver den sammenhæng vi tidligere så mellem Possion og eksponentiel fordelingen. Lad $X(t)$ være en tilfældig variabel der tæller antallet af punkter fra 0 til t i en Poisson process med parameter λ. Det viser sig da at $X(t)$ følger en Possion fordeling med parameteren $t\lambda$. Den forventede værdi af $X(t)$ er dermed $E[X(t)] = t\lambda$. λ kan således tolkes som antal forventede hændelser per tidsenhed og kaldes af denne grund *raten*.

Eksempel 5.1. I en lufthavn ankommer hvert døgn gennemsnitligt 400 fly. Flyene ankommer tilfældigt i løbet af døgnet og er ikke koordineret flyene imellem. Ud fra disse oplysninger kan vi antage en Poisson process for ankomsttidspunkter. Hvis $X(t)$ er antallet af fly i løbet af t timer vil denne følge en Poisson fordeling hvor:

$$E[X(24)] = 24\lambda = 400 \Rightarrow \lambda = 400/24 \approx 16.7,$$

således $X(t) \sim \text{Poi}(16.7 \cdot t)$ og $T_i \sim \exp(16.7)$. Da tiden mellem to flyvere T_i er eksponentielfordelt med parameter 16.7 da vil den forventede tid være $E[T_i] = 1/16.7 = 0.06$ timer $= 3.6$ minutter med en standardafvigelse også på 3.6 minutter.

Et andet nyttigt resultat kan bruges til nemt at simulere en Poisson process. Vi har givet en Poisson process med rate λ. Givet at der er N antal punkter i tidsperioden $[0, t]$ vil fordelingen for tidspunkterne være uafhængigt uniformt fordelte tilfældige variable i intervallet $[0, t]$. Kender vi N kan man altså blot simulere N punkter uniformt fordelt i intervallet $[0, t]$ for at simulere en Poisson process. De to resultater om Poisson processen kan således bruges til at simulere dem, hvilket er opsummeret i algoritme 1. Se [1, pp. 240–247] for mere om Poisson processer.

Algorithm 1 Simulering af Poisson process med rate λ mellem 0 og t

Simuler $N \sim \text{Poi}(t\lambda)$
for $i = 0$ til $N - 1$ **do**
  Simuler $Y_i \sim \text{unif}(0, t)$ {Y_i er tidspunktet for en hændelse}
end for

6 Generation af tilfældige på computere

Der findes specielle algoritmer til at generere tilfældige tal på en computer. Dette kan bruges i eksempelvis en diskret tids simulering hvor tilfældige tal *trækkes* fra fordelinger baseret på statistiske modeller. Da de tilfældige tal genereres ud fra en deterministisk algoritme kalder vi dem for *pseudo* tilfældige, men de kan dog opføre sig statistisk ligesom ægte tilfældige tal og er derfor brugbare. Lad os se på hvordan man kan generere et tilfældigt heltal i C med `rand` funktionen fra `stdlib`:

```c
#include <stdio.h>
#include <stdlib.h>

void main() {

    int N = 5;
    unsigned int nr;

    printf("Random numbers between 0 and %u: ", RAND_MAX);

    for(int i = 0; i < N; i++) {
       nr = rand();
       printf("%u, ", nr);
    }
    printf("\n");
}
```

I konsollen får vi:

```
$ rand_tutorial1.exe
Random numbers between 0 and 32767: 41, 18467, 6334, 26500, 19169,
```

Gentages eksemplet ser vi dog noget specielt:

```
$ rand_tutorial1.exe
Random numbers between 0 and 32767: 41, 18467, 6334, 26500, 19169,
```

Vi får de samme tilfældige tal begge gange og det gør vi da tallene er genereret ud fra en deterministisk algoritme. Mere præcist baseres første kald af `rand` på et såkaldt *seed*, et ikke negativt heltal der "sætter gang" i genereringen af tilfældige tal. For `stdlib` er seed som udgangspunktet 0, og ændres dette ikke vil de samme tal altid genereres. Man kan manuelt sætte seed med `srand` funktionen fra `stdlib`. Ønsker man automatisk valg af seed kan man bruge `time` funktionen fra `time` biblioteket:

```c
#include <stdio.h>
#include <stdlib.h>
#include <time.h>

void main() {

    int N = 5;
    unsigned int nr;
    srand(time(0));

    printf("Random numbers between 0 and %u: ", RAND_MAX);

    for(int i = 0; i < N; i++) {
       nr = rand();
```

```
        printf("%u, ", nr);
    }
    printf("\n");
}
```

```
$ rand_totorial2.exe
Random numbers between 0 and 32767: 29042, 25101, 25927, 18539, 21689,
```

I videnskabeligt arbejde er det vigtigt at kontrollere seed manuelt, således resultater principielt kan genskabes. Det er også tit praktisk at kontrollere seed manuelt ved debugging af programmer. Læs mere om algoritmer til genering af tilfældige tal og seed i [1, pp. 282–283].

7 Generering af tilfældige tal i C

C er ikke ligefrem statistikerens første valg når der skal arbejdes med simuleringer og statistiske modeller, og det er derfor begrænset hvor mange biblioteker der findes til at trække tal fra forskellige fordelinger især de mere avancerede som Poisson og eksponentielfordelingen[1]. Heldigvis kan vi basere genering af tilfældige ud fra `rand` funktionen i kombination med passende algoritmer for at simulere fra den ønskede fordeling. Der gives her algoritmerne til at trække fra fordelingerne beskrevet i denne note, og det er så op til jer at implementere i C.

7.1 Tilfældigt tal mellem 0 og N

Vi har allerede set hvordan man genererer tilfældige tal mellem `0` og `RAND_MAX` som typisk er $32764 = 2^{15} - 1$. Algoritme 2 beskriver hvordan et tilfældigt tal mellem 0 og $N \leq$ `RAND_MAX` kan genereres.

Algorithm 2 Tilfældigt tal X mellem 0 og $N \leq$ `RAND_MAX`

Simuler Y mellem 0 og `RAND_MAX` med `rand`.
$X \leftarrow Y \mod (N+1)$ { mod er modulus operatoren som er `%` i C}

7.2 Uniform fordeling

Simulering af kontinuert uniformt fordelte tal mellem 0 og 1 danner base for simulering af alle andre fordelinger, algoritme 3 beskriver hvordan.

Algorithm 3 Uniform tilfældige variabel X mellem 0 og 1

Simuler Y mellem 0 og N
$X \leftarrow Y/N$

Teknisk set er X i algoritme 3 diskret uniformt fordelt i mængden $\left\{0, \frac{1}{N}, \ldots, \frac{N-1}{N}, N\right\}$, men for tilpas høj N er dette en acceptabel tilnærmelse. Algoritme 4 beskriver simulering af en uniformt tilfældig variabel i et vilkårligt interval $[a, b]$.

Algorithm 4 Uniform tilfældige variabel X mellem a og b

Simuler $U \sim \text{unif}(0, 1)$
$X \leftarrow U(b-a) + a$

7.3 Diskrete fordelinger

Givet en pmf kan alle diskrete fordelinger, herunder Bernoulli og Poisson, simuleres med samme algoritme. Bernoulli fordelingen er dog speciel simpel at simulere og fortjener sin egen algoritme - algoritme 6.

[1] I min søgen har jeg ikke kunne finde andre biblioteker end stdlib, så skriv endelig til mig hvis du kan finde et godt bibliotek.

Algorithm 5 Bernoulli fordeling med parameter p

Simuler $U \sim \text{unif}(0,1)$
if $U > p$ **then**
  $X \leftarrow 1$
else
  $X \leftarrow 0$
end if

Den generelle algoritme til simulering af tilfældige tal fra diskete fordelinger med pmf p_X og mulige udfald $x_1, x_2, \ldots$ er givet i algoritme X

Algorithm 6 Diskret tilfældig variabel X med pmf p_X og mulige udfald $x_1, x_2, \ldots$

$F_0 \leftarrow 0$ og $F_k \leftarrow \sum_{j=1}^{k} p_X(x_k)$ for $k = 1, 2, \ldots$ {Kumulativ sum af pmf}
Simuler $U \sim \text{unif}(0,1)$
Find i således $F_{i-1} < U \leq F_i$
$X \leftarrow x_i$.

I fordelinger som Poisson fordelingen hvor der ikke er nogen øvre grænse er det nødvendigt at vælge en $k_{\max}$ i udregning af F_k. Det vides at $F_k \to 1$ for store k så et fint kriterie kan være at vælge $k_{\max}$ som det mindste k således $F_k > 1 - \epsilon$ for en et lille tal $\epsilon > 0$ eksempelvis $\epsilon = 10^{-5}$. Se [1, pp. 283–285] for mere om simulering af diskrete tilfældige variable.

7.4 Eksponentiel fordelingen

Eksponentielfordelingen kan nemt simuleres ved brug af den *inverse transformationsmetode*. Teorien udelades her men algoritmen er givet i algoritme REF.

Algorithm 7 Eksponentielfordelt X med parameter λ

Simuler $U \sim \text{unif}(0,1)$
$X \leftarrow -\frac{1}{\lambda} \ln(1-U)$ {ln er den naturlige algoritme}

Se [1, pp. 285–287] for mere om den inverse transformationsmetode.

7.5 Normalfordelingen

Normalfordelingen er en smule mere avanceret at simulere og bruger den såkaldte forkastelsesmetode opsummeret i algoritme 8 til simulering af en standard normalfordeling med forventet værdi 0 og varians 1

Algorithm 8 Normalfordelt X med forventet værdi 0 og varians 1 således $X \sim \mathcal{N}(0,1)$

1: Simuler $U \sim \text{unif}(0,1)$ og $Y \sim \exp(1)$ {Uafhængigt}
2: **if** $U \leq e^{-(Y-1)^2/2}$ **then**
3: $|X| = Y$ {Absolut værdi uden fortegn}
4: **else**
5: Gå til skridt 1
6: **end if**
7: Simuler $V \sim \text{unif}(0,1)$ {Find fortegn}
8: **if** $V \leq \frac{1}{2}$ **then**
9: $X \leftarrow |X|$
10: **else**
11: $X \leftarrow -|X|$
12: **end if**

Algoritme REF beskriver simulering af en normalfordeling med vilkårlig forventet værdi og varians.

Algorithm 9 Normalfordelt X med forventet værdi μ og varians σ^2

Simuler $Y \sim \mathcal{N}(0, 1)$
$X \leftarrow Y\sigma + \mu$

Se [1, pp. 288–290] for mere om forkastelsesmetoden og simulering af normalfordelingen.

References

[1] Olofsson Peter og Andersson, Mikeal. *Probability, Statistics and Stochastic Processes*. 2nd ed. John Wiley & Sons, 2012.

[2] *The Poisson Distribution*. URL: `http://www.mun.ca/biology/scarr/smcPoisson_distributions.html`.