Introduktion til Sandsynlighed og Statistik til modellering og simulering

Datalogi og Software 1. semester

Tobias Kallehauge

Aalborg Universitet, November 12, 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. 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 bestemt definition. For at måle sandsynlighed indføres sandsynlighedsmålet P, som er en funktion der bestemmer sandsynligheden for udfald i *tilfældige eksperimenter*. Hvis et tilfældigt eksperiment har mulige udfald $A_1, A_2, \ldots$ da vil $P(A_i)$ være sandsynligheden for netop 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, hvoraf den vigtigste er at den altid antager værdier mellem 0 og 1 (se den fulde definition i [1, sektion 1.3]).

For at formalisere notationen i tilfældige forsøg, eksempelvis at slå med en terning, indfører vi begrebet *tilfældige variable*. En tilfældig variabel er en funktion $X : S \to \mathbb{R}$, der får sine værdier fra et tilfældigt forsøg hvor S er mængden af mulige udfald. I forsøget med terningen vil $X(\text{slå en 6'er}) = 6$, $X(\text{slå en 1'er}) = 1$, osv. Vi kan så skrive $P(X = 6) = 1/6$. Et andet eksempel er møntkast hvor den tilfældige variabel X kan repræsentere antallet af kroner i løbet af 3 møntkast. Hver vil $X(PPK) = 1$, $X(KKP) = 2$, $X(KKK) = 3$ osv.

Med begrebet tilfældige variable kan vi indføre *sandsynlighedsmassefunktionen*, på engelsk *probability mass funktion* (pmf), som egentlig blot er en nemmere notations metode. 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, eksempelvis er $P(\text{slå en 6'er}) = P(X = 6) = p_X(6)$.

2 Fordelinger og fordelingsfunktioner

Det viser sig at mange forsøg kan kategoriseres med parametriske *fordelinger* med tilhørende pmf'er. Fra kursusgangen om tilfældige tal kender vi allerede et par fordelinger nemlig den *uniforme fordeling* og *normalfordelingen*. Forsøget med terningslag er et eksempel på en uniform fordeling hvor $p_X(x) = 1/6$ for $x = 1, \ldots, 6$, altså lige stor sandsynlighed for alle udfald. I terningslag forsøget følger X altså en uniform fordeling med udfald mellem 1 og 6 og vi skriver $X \sim \text{unif}(1, 6)$. Forsøget med terning og møntkast er eksempler på *diskrete fordelinger*, da der er et tælleligt antal udfald i forsøget. Vi kan også have *kontinuerte* fordelinger hvor man ikke kan tælle antal udfald, eksempelvis normalfordelingen hvor udfaldet kan være alle reelle tal. Ved kontinuerte fordelinger snakker man istedet for pdf'er om *sandsynlighedstæthedsfunktioner*, på engelsk *probability density function* (pdf). En uniform fordeling kan også være kontinuert hvis vi tillader alle reelle tal i et interval $[a, b]$ og vi har pdf funktionen $p_X(x) = 1/(b - a)$[1].

[1] I kontinuerte fordelinger skal man integrere for at udregne sandsynligheder. Vi har specielt at $P(a \leq X \leq b) = \int_a^b p_X(x)\, dx$. Eksempelvis har vi for en kontinuert uniform fordelingen mellem 0 og 1 at $X \sim \text{unif}(0, 1)$, $p_X(x) = 1$ og $P(0 \leq X \leq 0.5) = \int_0^{0.5} 1\, dx = 0.5$.

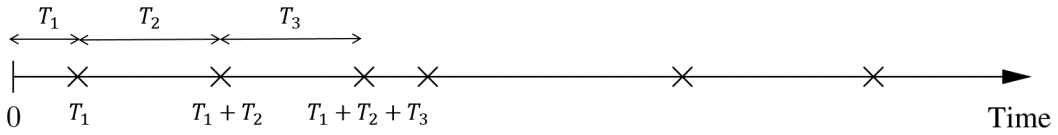

Figure 1: Punktprocess hvor × markerer hændelser i tid - eksempelvis ankomsttidspunkter af fly.

3 Forventet værdi, varians og standard afvigelse

4 Specielle fordelinger og modellering

5 Punktprocesser

De relevante oplysninger vi har fået at vide om lufthavnens trafik er:

- Der ankommer i gennemsnit 200 fly om dagen, men på en given dag kan der både ankomme færre eller flere.
- Trafikken ankommer indenfor en periode på 13 timer.
- Ankomsttidspunkterne er tilfældige inden for perioden og ikke koordineret flyene imellem.
- Trafikken forventes at stige med 5% om året fremover.

Et sådant setup er klassisk indenfor modellering af systemer hvor begivenheder sker tilfældigt i tid. Lignende eksempler kunne være ankomst af opgaver til en computerserver, kø på apoteket, tidspunkter for jordskælv, ulykker på motervej og henfald af radioaktive partikler. Alle disse eksempler kalder man for *punktprocesser* hvor tidsperioden mellem to hændelser er tilfældige variable $T_1, T_2, \ldots$ som illustreret i figur 3. Det viser sig, at når man antager uafhængighed mellem tidspunkter[2], da vil $T_1, T_2, \ldots$ nødvendigvis følge en såkaldt *eksponentielfordeling* med rate λ og vi kalder punktprocessen for en *Poisson process*.
En exponentielfordeling er en kontinuert fordeling og hvis X følger en eksponentielfordeling med rate λ da er pdf funktionen[3]:

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

og vi skriver $X \sim \exp(\lambda)$ (se figur 4). Raten λ i eksponentielfordelingen har indflydelse på hvor hurtigt vi ser et udfald og en eksponentialtfordelt X vil i gennemsnit have udfaldet $1/\lambda$. I en Poisson fordeling har vi altså at $T_i \sim \exp(\lambda)$ for $i = 1, 2 \ldots$, hvor hver T_i har gennemsnit $1/\lambda$.

Følgende resultat om Possion processer er nyttigt. Lad $X(t)$ være en tilfældig variabel der tæller antallet af punkter fra 0 til t i en Poisson process med rate λ. Det viser sig da at $X(t)$ følger den såkaldte *Possion fordeling* med parameteren $k = \lambda t$. En Poisson fordeling er en diskret fordeling og hvis X følger en Poisson fordeling med parameter k da er pmf funktionen[4]

$$p_X(x) = e^{-k}\frac{k^x}{x!}, \quad x = 0, 1, \ldots$$

Den gennemsnitlige værdi af en Poisson fordeling med parameter k er blot k (se figur 4). Lad os se på følgende eksempel: En Possson process har rate $\lambda = 2$. Ud fra resultatet vil antal punkter ved $t = 10$ sekunder være Poisson fordelt med parameter $k = 20$ og vi forventer gennemsnitligt 20 punkter ved $t = 10$.

Et sidste 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

[2] Se [1, pp. 123–127] for de præcise detaljer.

[3] Man kan simulere tal fra eksponentielfordelingen med `numpy.random.exponential` hvor man skal angive parametret `scale`. `scale` svarer til λ^{-1}

[4] Man kan simulere tal fra Poissonfordelingen med `np.random.poisson` hvor man skal angive parameteret `lamb`. `lamb` svarer til k.

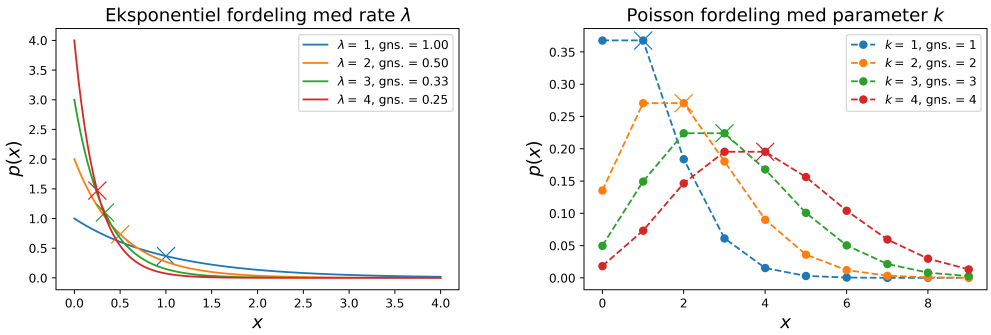

Figure 2: pdf og pmf for eksponentiel og Poisson fordeling. $\times$ markerer gennemsnit ved de forskellige parametre.

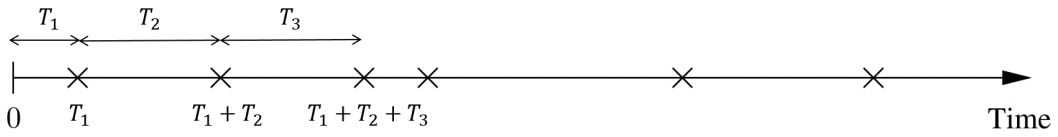

Figure 3: Punktprocess hvor $\times$ markerer hændelser i tid - eksempelvis ankomsttidspunkter af fly.

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.

I kan læse mere om eksponentielfordelinger, Poissonfordelinger og Poisson processer i [1] (online version er gratis tilgængelig på AUB). Heri finder i også beviser for resultaterne om Poisson processen. Ovenstående er dog til tilstrækkeligt til at generere de tal vi skal bruge i miniprojektet. Overvej følgende spørgsmål for at simulere ankomsttidspunkter af fly:

- Hvordan kan man bestemme rate parametren λ ud fra de givne oplysninger?
- Hvordan kan man simulere antal fly på en dag?
- Hvordan kan man simulere ankomsttidspunkter givet antal fly på en dag?
- Hvordan kan effekten af stigende trafik på 5% om året simuleres? Hint: Brug λ.

6 Generation af tilfældige tal i C

De relevante oplysninger vi har fået at vide om lufthavnens trafik er:

- Der ankommer i gennemsnit 200 fly om dagen, men på en given dag kan der både ankomme færre eller flere.
- Trafikken ankommer indenfor en periode på 13 timer.
- Ankomsttidspunkterne er tilfældige inden for perioden og ikke koordineret flyene imellem.
- Trafikken forventes at stige med 5% om året fremover.

Et sådant setup er klassisk indenfor modellering af systemer hvor begivenheder sker tilfældigt i tid. Lignende eksempler kunne være ankomst af opgaver til en computerserver, kø på apoteket, tidspunkter for jordskælv, ulykker på motervej og henfald af radioaktive partikler. Alle disse eksempler kalder man for *punktprocesser* hvor tidsperioden mellem to hændelser er tilfældige variable $T_1, T_2, \ldots$ som illustreret i figur 3. Det viser sig, at når man antager uafhængighed mellem tidspunkter[5], da vil $T_1, T_2, \ldots$ nødvendigvis følge en såkaldt *eksponentielfordeling* med rate λ og vi kalder punktprocessen for en *Poisson process*.

[5] Se [1, pp. 123–127] for de præcise detaljer.

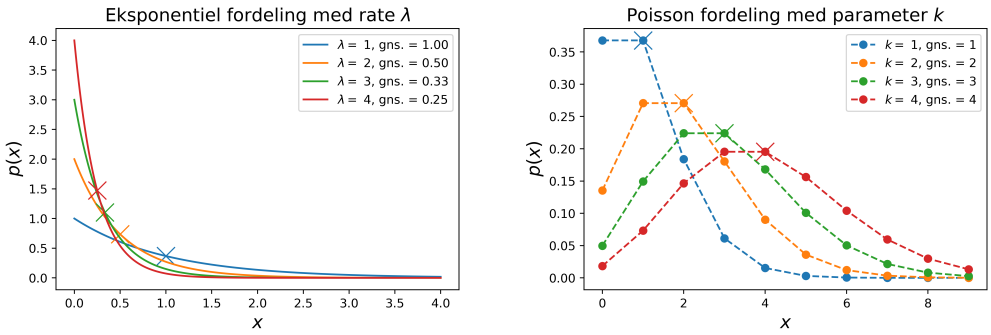

Figure 4: pdf og pmf for eksponentiel og Poisson fordeling. $\times$ markerer gennemsnit ved de forskellige parametre.

En exponentielfordeling er en kontinuert fordeling og hvis X følger en eksponentielfordeling med rate λ da er pdf funktionen[6]:

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

og vi skriver $X \sim \exp(\lambda)$ (se figur 4). Raten λ i eksponentielfordelingen har indflydelse på hvor hurtigt vi ser et udfald og en eksponentialtfordelt X vil i gennemsnit have udfaldet $1/\lambda$. I en Poisson fordeling har vi altså at $T_i \sim \exp(\lambda)$ for $i = 1, 2 \ldots$, hvor hver T_i har gennemsnit $1/\lambda$.

Følgende resultat om Possion processer er nyttigt. Lad $X(t)$ være en tilfældig variabel der tæller antallet af punkter fra 0 til t i en Poisson process med rate λ. Det viser sig da at $X(t)$ følger den såkaldte *Possion fordeling* med parameteren $k = \lambda t$. En Poisson fordeling er en diskret fordeling og hvis X følger en Poisson fordeling med parameter k da er pmf funktionen[7]

$$p_X(x) = e^{-k}\frac{k^x}{x!}, \quad x = 0, 1, \ldots$$

Den gennemsnitlige værdi af en Poisson fordeling med parameter k er blot k (se figur 4). Lad os se på følgende eksempel: En Possson process har rate $\lambda = 2$. Ud fra resultatet vil antal punkter ved $t = 10$ sekunder være Poisson fordelt med parameter $k = 20$ og vi forventer gennemsnitligt 20 punkter ved $t = 10$.

Et sidste 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.

I kan læse mere om eksponentielfordelinger, Poissonfordelinger og Poisson processer i [1] (online version er gratis tilgængelig på AUB). Heri finder i også beviser for resultaterne om Poisson processen. Ovenstående er dog til tilstrækkeligt til at generere de tal vi skal bruge i miniprojektet. Overvej følgende spørgsmål for at simulere ankomsttidspunkter af fly:

- Hvordan kan man bestemme rate parametren λ ud fra de givne oplysninger?
- Hvordan kan man simulere antal fly på en dag?
- Hvordan kan man simulere ankomsttidspunkter givet antal fly på en dag?
- Hvordan kan effekten af stigende trafik på 5% om året simuleres? Hint: Brug λ.

References

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

[6] Man kan simulere tal fra eksponentielfordelingen med `numpy.random.exponential` hvor man skal angive parametret `scale`. `scale` svarer til λ^{-1}

[7] Man kan simulere tal fra Poissonfordelingen med `np.random.poisson` hvor man skal angive parameteret `lamb`. `lamb` svarer til k.