Introduktion til Sandsynlighed og Statistik til modellering og simulering

Datalogi og Software 1. semester

Tobias Kallehauge

Aalborg Universitet, November 13, 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 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 $\{x_1, x_2, \dots\}$ 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 $\text{Var}[x] = p(1-p)$.

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

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 $X \sim \text{unif}[a,b]$. Vi har $E[X] = (a+b)/2$ og $\text{Var}[X] = (b-a)^2/12$.

4.3 Poisson fordelingen

Poisson fordelingen 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 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. [**oflofsson2012**].

4.4 Eksponentielfordeling

4.5 Normalfordeling

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 1. Det viser sig, at når man antager uafhængighed mellem tidspunkter[1], da vil

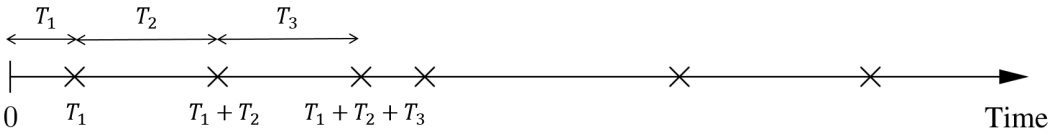

Figure 1: Punktprocess hvor $\times$ markerer hændelser i tid - eksempelvis ankomsttidspunkter af fly.

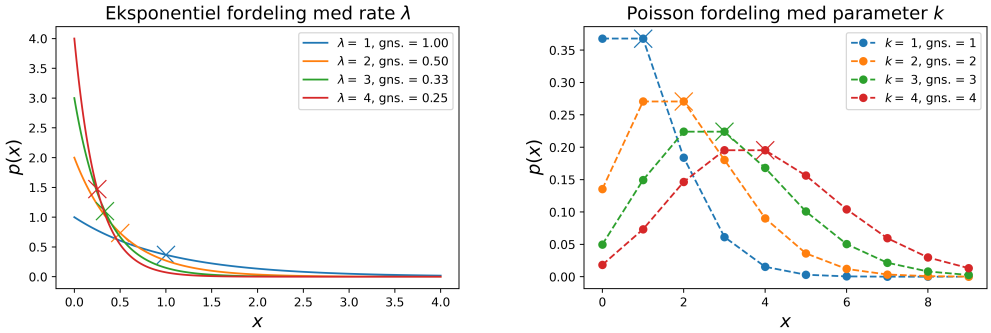

Figure 2: pdf og pmf for eksponentiel og Poisson fordeling. $\times$ markerer gennemsnit ved de forskellige parametre.

$T_1, T_2, \ldots$ 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[2]:

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

og vi skriver $X \sim \exp(\lambda)$ (se figur 2). 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[3]

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

Den gennemsnitlige værdi af en Poisson fordeling med parameter k er blot k (se figur 2). 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:

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

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

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

- 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

References

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