Programmering for Matematikere

—

Miniprojekt: Et lufthavnskø problem

Guide til simulering af tilfældige tal

T. Kallehauge, J. J. Nielsen

Aalborg Universitet, November 12, 2020

Dette dokument er supplement til det primære dokument omhandlende miniprojektet. I dette dokument bliver der givet en guide til hvordan man kan simulere ankomsttidspunkter og landingsvarighed set fra et *statistiks* synspunkt. Da I på nuværende tidspunkt ikke har haft statistik, vil de nødvendige begreber gennemgås, men kun på overfladisk niveau. Guiden er primært matematisk. Der vil blive givet hints til hvordan tallene kan simuleres i Python, men det er i udgangspunktet op til jer at finde ud af.

1 Introduktion til sandsynlighed og fordelinger

Kort sagt er et sandsynlighedsmål P en funktion, der bestemmer sandsynligheden for hændelser A som $P(A)$. P er begrænset mellem 0 og 1. Eksempelvis er sandsynligheden for hændelsen *slå en 6'er* med en 6-sidet terning $P(\text{slå en 6'er}) = 1/6 \approx 0.17\%$.
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 $P(\text{slå en 6'er}) = P(X = 6) = p_X(6)$.

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].

2 Simulering af ankomsttidspunkter - Poisson proces

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[2], da vil $T_1, T_2, \ldots$ nødvendigvis følge en såkaldt *eksponentielfordeling* med rate λ og vi kalder punktprocessen for en *Poisson process*.

[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 $P(0 \leq X \leq 0.5) = \int_0^{0.5} 1\, dx = 0.5$.

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

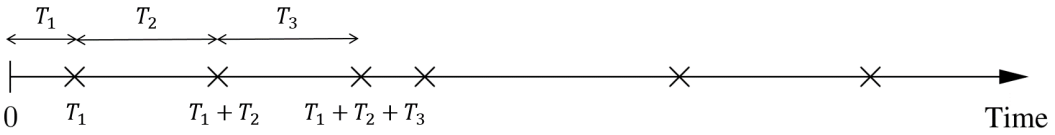

Figure 1: Punktprocess hvor $\times$ markerer hændelser i tid - eksempelvis ankomsttidspunkter af fly.

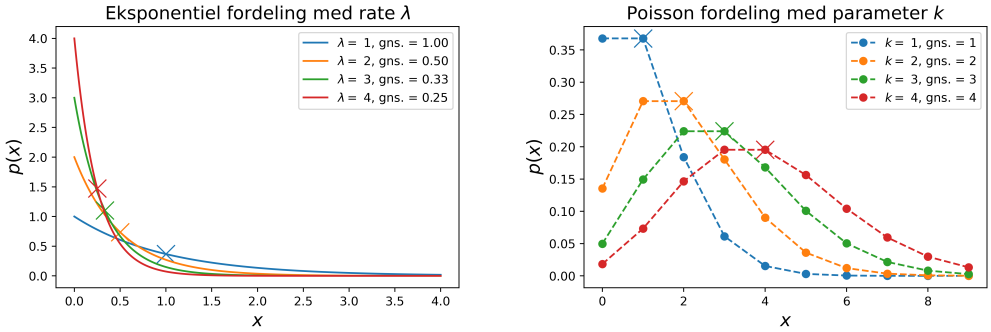

Figure 2: pdf og pmf for eksponentiel og Poisson fordeling. $\times$ markerer gennemsnit ved de forskellige parametre.

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 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[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 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:

- Hvordan kan man bestemme rate parametren λ ud fra de givne oplysninger?

[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.

- 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.

3 Ikke-parametriske fordelinger og simulering af landingsvarighed

Til landingstiderne har vi fået oplyst en række tidsintervaller og hvor mange flys landingsvarigheder, der typisk falder inden for hvert interval. I dette afsnit lærer vi teorien til hvordan disse oplysninger kan bruges til at generere tilfældige landingsvarigheder i overensstemmelse med de givne oplysninger.

Eksponentielfordelingen var et eksempel på en parametrisk fordeling hvor blot ét enkelt tal, λ, karakteriserer hele fordelingen. Fordelingen for landingsvarighed er dog en smule mere kompliceret da mange parametre såsom flytype og vejrforhold har indflydelse. Så i stedet for at prøve at finde en tilpas kompliceret fordeling, der passer på data, kan vi tilpasse en *ikke parametrisk* fordeling direkte ud fra data. Den nemmeste metode til dette er følgende. Lad intervallet $[a, b]$ være udfaldsrummet for en tilfældig variabel X og lad $I_1 = [a, x_1), I_2 = [x_1, x_2) \ldots, I_N = [x_{N-1}, b]$ være N delmængder, der dækker intervallet. Lad os nu antage at vi har en række observationer af X, og definer da størrelsen:

$$p_i = \frac{n_i}{\Delta_i M},$$

hvor n_i er antallet af observerede punkter inden for det i'te interval, Δ_i er størrelsen på hvert interval og M er det total antal observationer. Følgende funktion er da en ikke parametrisk pdf for det observerede data:

$$p_X(x) = \begin{cases} p_1 & x \in I_1 \\ p_2 & x \in I_2 \\ \vdots & \\ p_N & x \in I_N \\ 0 & x \notin [a, b] \end{cases}$$

Eksempel: Vi har observeret følgende dataset med 20 observationer:

$$\vec{x} = (2.4, 3.8, 2.8, 2.4, 1.7, 3.1, 1.7, 6.7, 9.9, 1.5, 4.7, 2.3, 2.5, 7.8, 0.2, 0.3, 0.1, 5.4, 4.5, 6.1)$$

Inddeles intervallet $[0, 10]$ i $[0, 2), [2, 4), \ldots, [8, 10]$ fås således:

$$\vec{nr} = (6, 7, 3, 3, 1), \quad \Delta = 2, \quad M = 20 \quad \text{og}$$

$$p_X(x) = \begin{cases} \frac{6}{2 \cdot 20} & x \in [0, 2) \\ \frac{7}{2 \cdot 20} & x \in [2, 4) \\ \frac{3}{2 \cdot 20} & x \in [4, 6) \\ \frac{3}{2 \cdot 20} & x \in [6, 8) \\ \frac{1}{2 \cdot 20} & x \in [8, 10] \\ 0 & x \notin [0, 10] \end{cases},$$

illustreret i figur 3.

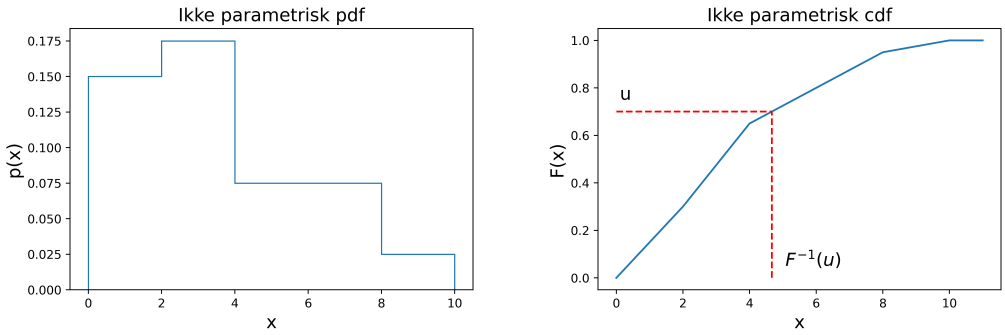

Figure 3: Ikke parametrisk pdf og cdf for eksemplet med 20 observationer.

For at simulere tal givet en pdf kan man bruge den *inverse cdf metode*. En cdf er den kumulative fordelings funktion eller på engelsk cumulative distribution function, og denne beskriver sandsynligheden for at en tilfældig variabel er mindre end eller lig en bestemt værdi. For en tilfældig variabel X noteres cdf'en typisk F_X, og der gælder at $F_X(x) = P(X \leq x)$. For kontinuerte fordelinger kan man udregne CDF'en ved at integrere pdf'en via:

$$F_X(x) = \int_{-\infty}^{x} p_X(x)\ dx$$

Givet en cdf kan man udtrække tal fra den fordeling den stammer fra ved følgende to skridt:

- Udtræk et tilfældigt tal u fra en uniform fordeling mellem 0 og 1.
- Sæt $x = F_X^{-1}(u)$ hvor F_X^{-1} er den inverse cdf for X (givet at den eksisterer). x vil da følge fordelingen for X.

Dette koncept er illustreret i figur 3.

Ovenstående kan bruges til at simulere mange fordelinger, men i vores tilfælde kan det være en smule besværligt at integrere sig frem til F_X og derefter at finde den inverse funktion F_X^{-1}. I det specifikke tilfælde hvor vi har en ikke parametrisk pdf, der er uniformt fordelt i nogle bestemte intervaller, kan vi simulere tal fra fordelingen på følgende vis:

- Vi har igen delt udfaldsrummet $[a, b]$ op i N intervaller

$$I_1 = [a, x_1),\ I_2 = [x_1, x_2),\ \ldots,\ I_N = [x_{N-1}, b]$$

- Udregn blot cdf'en i højresiden af de valgte intervaller altså $F_X(x_1), F_X(x_2), \ldots, F_X(x_N)$ hvor $x_N = b$.
  Hint: $F_X(x_1), F_X(x_2), \ldots, F_X(x_N)$ har en særlig simpel form når vi kender $p_1, \ldots, p_N$. Lav udledningen på tavlen.
- Udtræk et tilfældigt tal u uniformt fordelt mellem 0 og 1.
- Find det mindste interval I_i således $u \geq F_X(x_i)$, altså I_i hvor $i = \min\{i \,|\, u \geq F_X(x_i)\}$.
- Simuler x fra en uniform fordeling i intervallet I_i.

Med en velvalgt illustration, kan man overbevise sig selv om at ovenstående metode er ækvivalent med invers cdf metoden når vi har en pdf, der består af en række flade intervaller. Der er selvfølgelig mange andre metoder man kan bruge, som i meget gerne må undersøge, men ovenstående er tilstrækkeligt for miniprojektet til simulering af landingstider. Overvej følgende hjælpespørgsmål:

- Hvordan kan vi bruge oplysningerne i miniprojektet til at udregne p_i værdierne?
- Hvordan kan vi udregne $F_X(x_i)$ værdierne ud fra p_i værdierne?
- Hvordan kan vi simulere tilfældige landingsvarigheder givet ovenstående?

References

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