

Poisson-prosesser

Vi har sett tidligere at forekomsten av hendelser følger en *Poisson-prosess* dersom:

1. Antallene hendelser i ikke-overlappende intervaller er uavhengige.
2. For ethvert lite intervall $[s, s + \Delta s]$ er $P(\text{en hendelse i } [s, s + \Delta s]) \approx \lambda \Delta s$.
3. $P(\text{to eller flere hendelser i } [s, s + \Delta s]) \approx 0$.

Samme for alle s

Dvs $P(\text{to eller flere hendelser "samtidig"}) \approx 0$

Det kan da vises at $X =$ "antall hendelser i $[0, t]$ " er Poisson-fordelt med forventningsverdi $\mu = \lambda t$.

Dvs

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

Tilsvarende vil også $X =$ "antall hendelser i $[a, b]$ " være Poisson-fordelt med forventningsverdi

$$\mu = \lambda(b - a)$$

Parameteren λ kalles *intensiteten* til prosessen.



Vi har tidligere vist at tid til første hendelse i en Poisson-prosess er eksponensialfordelt med forventning $1/\lambda$:

$$\begin{aligned} P(T_1 > t) &= P(\text{ingen hendelser i } [0, t]) \\ &= \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \end{aligned}$$

Videre er også alle tider mellom hendelser eksponensialfordelte.

Siden eksponensialfordelingen er hukommelsesløs får vi at Poisson-prosesser også vil være hukommelsesløse. Dette har vi allerede vist i busseksemplet i notatene til kap. 6. Uansett når vi går inn i en Poisson-prosess er fordelingen til tid til neste hendelse eksponensial med forventning $1/\lambda$, uavhengig av hva som har skjedd tidligere.

At prosessen er hukommelsesløs følger direkte fra punkt 1 og 2 i definisjonen. Sannsynligheten for en hendelse i $[s, s + \Delta s]$ er den samme for alle s (punkt 2) og uavh. av hva som har skjedd tidligere (punkt 1).

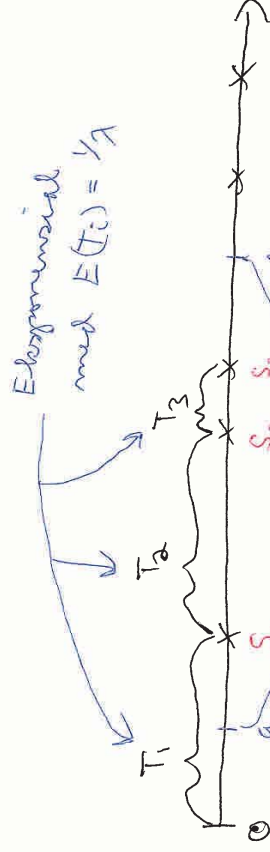
Det kan vises at en sum av k uavh. identisk eksponensialfordelte variable med forventning β er gammafordelt med $\alpha = k$ og $\beta = \beta$. Fra dette følger det direkte at tid til k te hendelse i en Poisson-prosess vil være gammafordelt med $\alpha = k$ og $\beta = 1/\lambda$ (siden tider mellom hendelser er eksponensialfordelte).

(Se öv. 4)

Kort oppsummert:

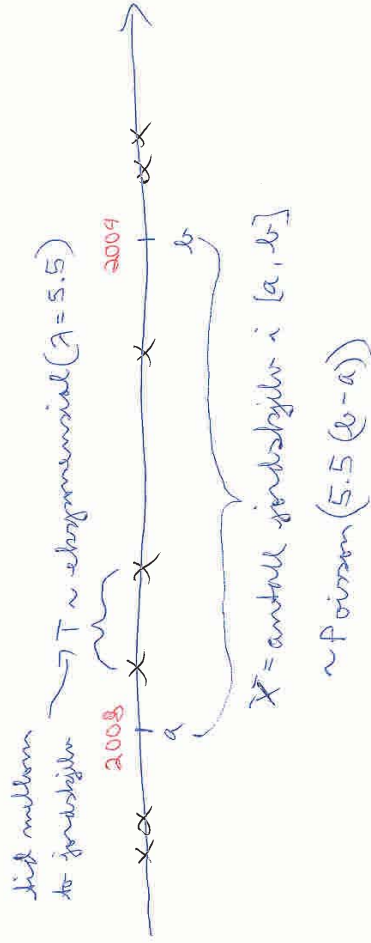
- Antall hendelser i $[a, b]$ er Poisson-fordelt med parameter $\lambda(b - a)$.
- Tid til første hendelse, og tid mellom to hendelser i en Poisson-prosess er eksponensialfordelt med forventning $1/\lambda$.
- Tid fra et vilkårlig tidspunkt i en Poisson-prosess til neste hendelse er eksponensialfordelt med forventning $1/\lambda$ (prosessen er hukommelsesløs).
- Tid til k te hendelse i en Poisson-prosess er gammafordelt med $\alpha = k$ og $\beta = 1/\lambda$.

Poisson prosess med intensitet λ :



$Y = \text{antall hendelser i } [a, b]$
 $\sim \text{Poisson}(\lambda(a-a))$
 $S_2 \sim \text{gamma}(\alpha=k, \beta=1/\lambda)$

Eksempel Forekomst av mtebær i Norge kan modelleres som en Poisson-prosess med parameter $\lambda = 5.5$ per år.



La f. eks. X = antall jordskjelv i 2008 $\sim \text{Poisson}(5.5)$

$$P(X > 5) = 1 - P(X \leq 5) = 1 - 0.5289 = \underline{\underline{0.47}}$$

Hva er sanns. for at det gir mer enn 3 mm mellom to skjelv?

$$\begin{aligned}
 P(T > \frac{1}{4}) &= \int_{\frac{1}{4}}^{\infty} 5.5 e^{-5.5t} dt = [-e^{-5.5t}]_{\frac{1}{4}}^{\infty} \\
 &= -0 - (-e^{-5.5 \cdot \frac{1}{4}}) = \underline{\underline{0.25}}
 \end{aligned}$$

3 mm = 1/4 år $\rightarrow$ eksagg. med $\lambda = 5.5$

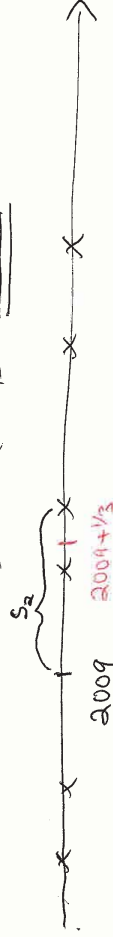


U = tid fra start 2009 til neste jordskjelv

B.g.a. at Poisson prosessen er lukket med hensyn til U , som T , eksponensialfordelt med parameter $\lambda = 5.5$

F. eks. vil $P(\text{flere jordskjelv de første mnd av 2009})$

$$= P(U > 1/4) = P(T > 1/4) = \underline{\underline{0.25}}$$



La S_2 = tid fra start 2009 til andre jordskjelv

Har da at $S_2 \sim \text{gamma}(\alpha=2, \beta=1/5.5)$

$$\begin{aligned}
 P(S_2 > 1/3) &= \int_{1/3}^{\infty} \frac{1}{(\frac{1}{5.5})^2} x^{2-1} e^{-x/\frac{1}{5.5}} dx \\
 &= \int_{1/3}^{\infty} (5.5)^2 x e^{-x \cdot 5.5} dx = \underline{\underline{0.45}}
 \end{aligned}$$

4 mnd $\rightarrow$ *deriver integrasjon*

Y = antall jordskjelv i [2009, 2009 + 1/3] $\sim \text{Poisson}(5.5 \cdot \frac{1}{3})$

$$\begin{aligned}
 P(S_2 > 1/3) &= P(Y < 2) = P(Y=0) + P(Y=1) \\
 &= \frac{(1.83)^0}{0!} e^{-1.83} + \frac{(1.83)^1}{1!} e^{-1.83} = \underline{\underline{0.45}}
 \end{aligned}$$

Estimering 1

-Punktestimering

Dekkes av kap. 8, 9.1-9.3 og 9.15/9.14.

Vi har til nå sett på en rekke forskjellige sannsynlighetsfordelinger og sett hvordan disse kan brukes til å modellere mange forskjellige typer fenomen. Vi har bl.a. sett på hvordan vi kan regne på sannsynligheter for ulike utfall, regne ut forventningsverdi, standardavvik/varians, etc.

En grunnleggende forutsetning vi har gjort hele veien er at vi på ett eller annet vis har kjent den eksakte sannsynlighetsfordelingen til fenomenet vi har sett på - dvs at vi både har kjent den eksakte typen fordeling og de aktuelle parameterverdiene. I praksis er ofte ikke dette tilfellet - spesielt vil parameterverdiene typisk være ukjente! (Vi skal fortsatt anta at vi er i stand til å angi en (noenlunde) "korrekt" type fordeling.) Estimering handler om å anslå verdien på ukjente parameterverdier ut fra observerte data.

Eksempel: Levetiden til en bestemt type kretskort antas eksponensialfordelt. Vi har observert at levetiden for seks slike kort ble hhv 612, 1009, 95, 1303, 599, og 780. Hva forteller dette oss om verdien på parameteren i eksponensialfordelingen? Dvs om forventet levetid? $\square$

Eksempel: Antall ulykker per år på en oljeplattform antas Poissonfordelt. De siste fem årene har det blitt registrert hhv 5, 8, 12, 7 og 13 ulykker. Hva forteller dette oss om parameteren λ i Poisson-fordelingen? Dvs om forventet antall ulykker per år/intensiteten? $\square$

Noen grunnbegrep

Anta at vi har observasjoner/data $X_1, \dots, X_n$ fra et fenomen som er beskrevet ved en sannsynlighetsfordeling $f(x; \theta)$. Verdien på parameteren er ukjent. Dersom observasjonene $X_1, \dots, X_n$ er uavhengige og identisk fordelte (u.i.f.) kan de brukes til å anslå den ukjente parameteren θ .

Observasjonene er identisk fordelte dersom alle har samme sannsynlighetsfordeling $f(x; \theta)$ (med samme parameter). Observasjonene er uavhengige dersom simultanfordelingen kan skrives $f(x_1, \dots, x_n) = f(x_1) \dots f(x_n)$ - vil i praksis si at informasjon om verdien på en av observasjonene ikke påvirker sannsynlighetsfordelingen til de øvrige.

At observasjoner er u.i.f. vil si det samme som at de er et tilfeldig utvalg (en tilfeldig valgt delmengde) av alle tenkelige mulige observasjoner av fenomenet som kan gjøres.

Alle tenkelige mulige observasjoner kalles *populasjonen*, og hvordan de fordeler seg er

beskrevet ved sannsynlighetsfordelingen $f(x; \theta)$. Den typiske problemstillingen er at vi basert på informasjonen i et tilfeldig utvalg, $X_1, \dots, X_n$, ønsker å si noe om hele populasjonen, f.eks. i form av å anslå parameteren θ .

Eksempel: Levetid kretskort. Populasjonen er levetiden til alle tenkelige kretskort av den aktuelle typen. Vi antar (basert på tidligere erfaring med denne typen fenomen) at denne populasjonen kan beskrives ved en eksponensialfordeling. Parameteren i eksponensialfordelingen er ukjent. Dersom observasjonene våre, $X_1, \dots, X_n$, er u.i.f., dvs et tilfeldig utvalg fra populasjonen av levetider, kan vi fra disse anslå, eller estimere parameteren.

Dersom $X_1, \dots, X_n$ ikke er u.i.f./tilfeldig utvalg, f.eks. dersom vi kun bruker observerte levetider som er større enn 1000, kan ikke observasjonene uten videre brukes til å estimere den ukjente parameterverdien. $\square$

$\lambda = \text{estimator} / \text{estimat}$

Estimatorer

Definisjon: En *estimator* er en funksjon av stokastiske variable som brukes til å anslå en ukjent parameterverdi.

Eksempel: Tre vanlig brukte estimatorer er:

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\hat{p} = \frac{X}{n}$$

for hhv forventningsverdi μ og varians σ^2 i normalfordeling, og p i binomisk fordeling. $\square$

Det er vanlig å bruke gjennomsnittet til å estimere forventningsverdier også i andre fordelinger. F.eks. dersom $T_1, \dots, T_n$ er u.i.f. eksponensial med forventning β er det rimelig å estimere β med

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^n T_i.$$

Merk: Estimatorer betegner vi med store bokstaver (=før eksperimentet er utført).

Innsatt observerte tallverdier, f.eks. $x_1, \dots, x_n$, (=etter eksperimentet er utført) kan vi regne ut et estimat, f.eks. $\hat{\mu} = \bar{x}$ eller $\hat{p} = x/n$. Estimatorer er altså stok. var. mens estimat er tallsvar.

Eksempel: Levetid kretskort. Estimatoren vi bruker til å anslå den ukjente parameteren β i eksponensialfordelingen er $\hat{\beta} = \frac{1}{n} \sum_{i=1}^n T_i$. Etter at vi har observert dataene 612, 1009, 95, 1303, 599, og 780 kan vi regne ut estimatet $\hat{\beta} = \frac{1}{6} \sum_{i=1}^6 t_i = 733$. $\square$