STK1100 våren 2023

Poisson-fordelingen og Poisson-prosessen

Svarer til avsnitt 3.7 i læreboka

Matematisk institutt
Universitetet i Oslo

Poisson-fordelingen er oppkalt etter den franske fysikeren og matematikeren Siméon Denis Poisson (1781-1840)

Poisson viste at hvis $n \to \infty$ og $p \to 0$ på en slik måte at $np \to \lambda > 0$ så vil (se forelesningen)

$$\binom{n}{x} p^x (1-p)^{n-x} \to \frac{\lambda^x}{x!} e^{-\lambda}$$

Poisson brukte dette til å finne en tilnærming til den binomiske fordelingen når n er stor og p er liten

Seinere fant en ut at en kunne bruke den grensen Poisson fant som punktsannsynligheten til en stokastisk variabel

Vi sier at en stokastisk variabel X er Poisson-fordelt med parameter λ hvis den har punktsannsynlighet

$$p(x;\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \text{for } x = 0,1,2,....$$

Merk at $\sum_{x=0}^{\infty} p(x;\lambda) = 1$ slik det skal være for en punktsannsynlighet (jf. forelesningen)

Poisson-fordeling med Python

Punktsannsynlighet:

$$p(x;\lambda) = P(X = x) = \frac{\lambda^x}{x!}\, e^{-\lambda}$$

Python: import scipy.stats as stats
stats.poisson.pmf(x,λ)

Kumulativ fordeling:

$$F(x;\lambda) = P(X \leq x) = \sum_{y=0}^{x} \frac{\lambda^y}{y!}\, e^{-\lambda}$$

Python: stats.poisson.cdf(x,λ)

Tabell A.2 bak i boka gir $F(x;\lambda)$ for noen verdier av λ

Vi vil ikke bry oss om disse tabellene

Sammenligner binomisk og Poisson

Vi vil sammenligne de binomiske fordelingene med $n = 20$, $p = 0.10$ og $n = 200$, $p = 0.01$ med Poisson-fordelingen med $\lambda = 2$

Python:

```
import numpy as np
import scipy.stats as stats
x=np.arange(0,10)
b20= stats.binom.pmf(x,20,0.1)
b200=stats.binom.pmf(x,200,0.01)
pois= stats.poisson.pmf(x,2)
np.column_stack((x,b20,b200,pois))
```

Resultatet er gitt på neste slide

x	$b(x;20,0.10)$	$b(x;200,0.01)$	$p(x;2)$
0	0.1216	0.1340	0.1353
1	0.2702	0.2707	0.2707
2	0.2852	0.2720	0.2707
3	0.1901	0.1814	0.1804
4	0.0898	0.0902	0.0902
5	0.0319	0.0357	0.0361
6	0.0089	0.0117	0.0120
7	0.0020	0.0033	0.0034
8	0.0004	0.0008	0.0009
9	0.0001	0.0002	0.0002

Vi kan ofte bruke Poisson-fordelingen til å beskrive forekomsten av «sjeldne begivenheter»:

- Antall tvillingfødsler i løpet av ett år på et sykehus
- Antall krefttilfeller i løpet av ett år i en kommune
- Antall ulykker i løpet av én måned på en byggeplass

Vi vil se på to konkrete eksempler om litt, men først vil vi finne forventningen og variansen til Poisson-fordelingen

Forventning og varians

Momentgenererende funksjon:

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \cdot p(x;\lambda) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x}{x!} e^{-\lambda}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t \lambda)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

Kumulantgenererende funksjon

$$R_X(t) = \ln\{M_X(t)\} = \lambda(e^t - 1)$$

Forventning

$$\mu = E(X) = R'_X(0) = \lambda$$

Varians:

$$\sigma^2 = V(X) = R''_X(0) = \lambda$$

Eksempel: Antall drepte av hestespark i det preusiske kavaleriet (1875-1894)

Eksempelet ble publisert av Ladislaus von Bortkiewicz i 1898 og var et av de første eksemplene på bruk av Poisson-fordelingen til å beskrive et datamateriale

Det er registrert antall soldater som ble drept av hestespark i 10 kavalerikorps i en periode på 20 år (totalt 200 «korps-år»)

Antall drepte	0	1	2	3	4	5+
Antall korps-år	109	65	22	3	1	0

Totalt antall drepte

$$0 \cdot 109 + 1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot 1 = 122$$

Kan dataene beskrives med Poisson-fordelingen?

La X være antall drepte i et korps-år

Hvis X er Poisson-fordelt, har vi at

$$P(X = x) = p(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

Forventet antall korps-år med x drepte er $n \cdot p(x;\lambda)$

For λ kan vi bruke gjennomsnittlig antall drepte per korps-år, som er $122/200 = 0.61$

Antall drepte	0	1	2	3	4	5+
Antall korps-år	109	65	22	3	1	0
Forventet antall	108.7	66.3	20.2	4.1	0.6	0.1

Poisson-fordelingen passer bra

Eksempel:
Forekomst av anencefali i Edinburgh 1956-1966

Anencefali er en alvorlig defekt hos fosteret som gjør at hjernen ikke utvikler seg som den skal.

Antall barn født med anencefali i Edinburgh i de 132 månedene fra 1955 til 1966 er gitt i de to første linjene av tabellen nedenfor:

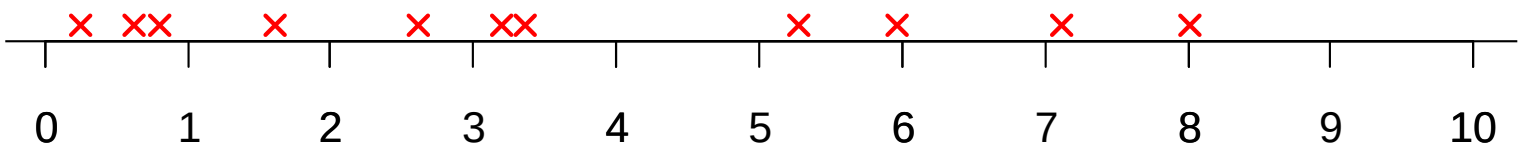
# tilfeller	0	1	2	3	4	5	6	7	8
# observert	18	42	34	18	11	6	0	2	1
# forventet	18.4	36.2	35.7	23.5	11.6	4.5	1.4	0.4	0.1

Siste linje gir forventet antall fra Poisson-fordelingen med $\lambda = 260 / 132 = 1.97$ (svarende til gjennomsnittlig antall tilfeller)

Poisson-fordelingen passer bra

Poisson-prosessen

Vi observerer begivenheter (markert med x) som hender over tid:



Vi antar at

- Sannsynligheten for at det inntreffer én begivenhet i et intervall av lengde Δt er $\alpha\Delta t + o(\Delta t)$. Her er $o(\Delta t)$ en størrelse som er slik at $o(\Delta t)/\Delta t \to 0$ når $\Delta t \to 0$
- Sannsynligheten for at det inntreffer mer enn én begivenhet i et intervall av lengde Δt er $o(\Delta t)$
- Antall begivenheter i et intervall er uavhengig av hvor mange begivenheter som har skjedd tidligere

La $P_k(t)$ være sannsynligheten for at det inntreffer k begivenheter i et intervall av lengde t

Da har vi at $P_k(t) = \frac{(\alpha t)^k}{k!} e^{-\alpha t}$ (jf. forelesningen)

Antall begivenheter i et intervall av lengde t er Poisson-fordelt med parameter $\lambda = \alpha t$

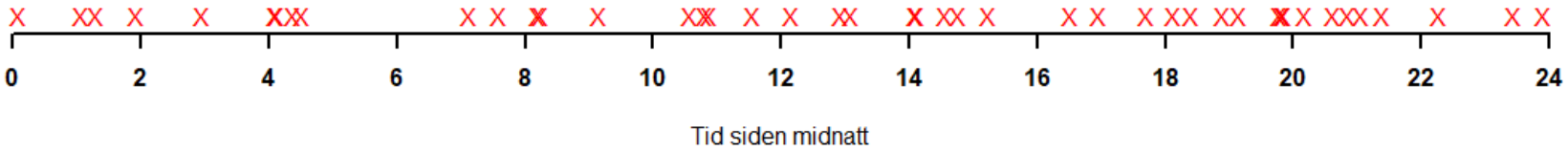
Merk at α er forventet antall begivenheter per tidsenhet, og er raten til Poisson-prosessen

Poisson-prosessen er en modell for begivenheter som skjer tilfeldig i tid

Eksempel: Tidspunkt for fødsler

18. desember 1997 ble det født 44 barn ved et sykehus i Australia (kilde: http://jse.amstat.org/datasets/babyboom.txt)

Tidspunktet for fødslene (tid etter midnatt) er vist nedenfor:



Oppsummert har vi:

Antall fødsler per time	0	1	2	3	4	5+
Antall timer med gitte antall fødsler	3	8	6	4	3	0
Forventet antall	3.8	7.0	6.4	3.9	1.8	0.9

Siste linje gir forventet antall for Poisson-fordelingen med $\lambda = \frac{44}{24} = 1.84$ (gjennomsnittlig antall fødsler per time)

En Poisson-prosess gir en god beskrivelse av fødslene

Eksempel: Fødsler ved Ahus 1999-2014

I en artikkel i Legeforeningens tidsskrift studeres fordelingen av antall ikke-planlagte fødsler per dag ved Ahus i perioden 1999-2014 (totalt vel 50 000 fødsler)
https://tidsskriftet.no/2015/12/originalartikkel/sesongjusterte-fodselsfrekvenser-er-poisson-fordelte

De fant at antall fødsler per dag er Poisson-fordelt, der forventningen λ øker over perioden 1999-2015 (siden antall fødsler ved Ahus har økt i perioden) og avhenger i noen grad av tid på året (siden det blir født flere barn om våren og sommeren enn om høsten og vinteren)

For en «gjennomsnittsmåned» i 2014 fant de $\lambda = 10$

Hvis antall fødsler per dag er Poisson-fordelt med $\lambda = 10$, hvor mange fødeplasser må sykehuset ha?

Hvis *X* er Poisson-fordelt med $\lambda = 10$ har vi

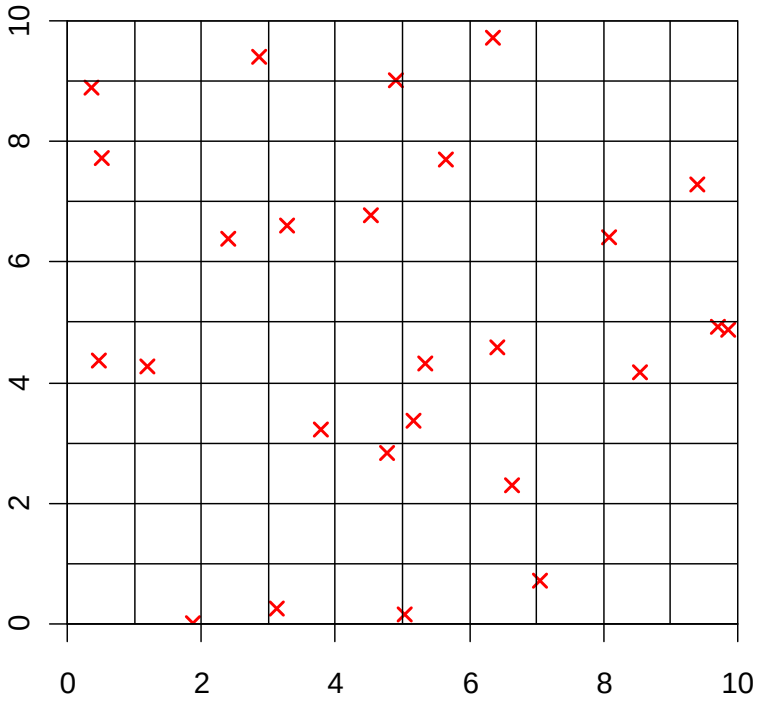
x	$P(X \leq x)$
8	0.333
9	0.458
10	0.583
11	0.697
12	0.793
13	0.865
14	0.917
15	0.951
16	0.973
17	0.986
18	0.993

Hvis sykehuset har 10 fødeplasser, er sannsynligheten 58.3% for at alle som føder en dag får plass

Hvis sykehuset har 15 fødeplasser, er sannsynligheten 95.1% for at alle som føder en dag får plass

Hvis sykehuset har 18 fødeplasser, er sannsynligheten 99.3% for at alle som føder en dag får plass

Poisson-prosess i planet



Anta at:

- forventet antall punkter per arealenhet er α
- det er ingen sammenfallende punkter
- antall punkter i disjunkte områder er uavhengige

Da har vi en Poisson-prosess i planet

Dette er en modell for punkter som er «tilfeldig fordelt»

La X være antall punkter i et område R med areal $a(R)$

Da er X Poisson-fordelt med parameter $\lambda = \alpha \cdot a(R)$

Eksempel:
Bombetreff av tyske V-1-raketter i Syd-London

Ser på antall treff av V-1-bomber i 576 små områder i Syd-London (0.25 km^2 hver) fra juni 1944 til mars 1945

Totalt antall treff var 537

# treff	0	1	2	3	4	5+
# observert	229	211	93	35	7	1
# forventet	226.7	211.4	98.5	30.6	7.1	1.6

Poisson-fordelingen med $\lambda = 0.93$ (tilsvarende gjennomsnittlig antall treff per område) passer bra

Case studie for Poisson-fordelingen: Krefttilfeller i Sømna

Dagbladet 10. januar 1993

POLITIKERE ER SKREMT

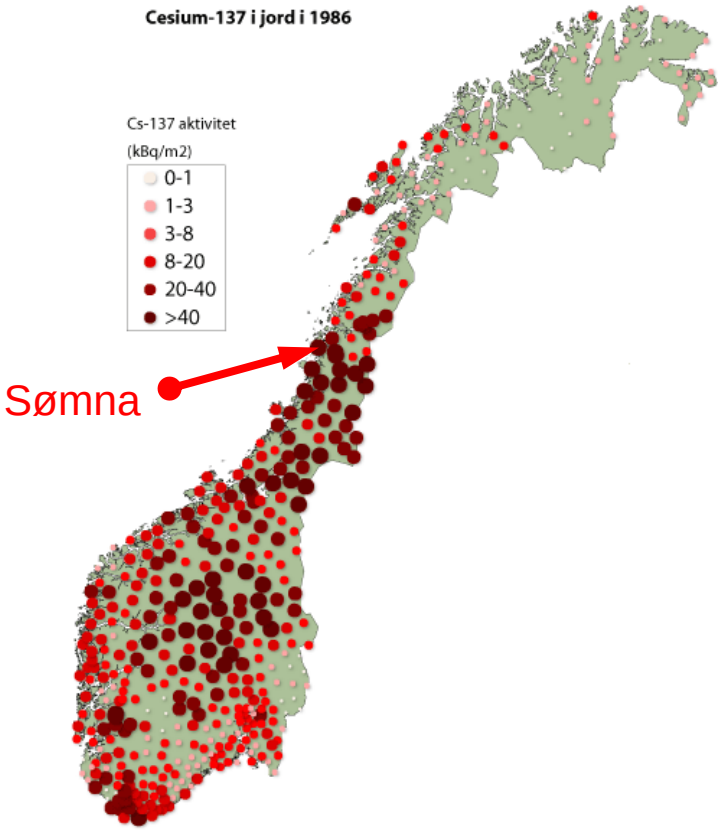
Krever Stortings-orientering

om kreftteoriene i Sømna

– Skremmende. Slik reagerer politikere Dagbladet har snakket med på reportasjen fra Sømna om frykten for kreftøkning som følge av Tsjernobyl.

Dagbladet 11. januar 1993

- Det ble observert tre tilfeller av hjernesvulst i Sømna kommune i 1992
- Det var uvanlig mange i en så liten kommune
- Saken vakte stor oppsikt i media og blant politikere
- Det ble sett i sammen-heng med Tsjernobyl-ulykken 26. april 1986



Etter Kreftregisterets statistikk vil en for en kommune av Sømnas størrelse og befolknings-sammensetning i gjennomsnitt observere ett tilfelle av hjernesvulst hvert sjette år *hvis* kreftrisikoen er som i resten av landet

Mer presist: Hvis kreftrisikoen i Sømna var som ellers i landet, vil antall tilfeller av hjernesvulst i løpet av ett år være Poisson-fordelt med $\lambda = 0.16$

Det gir: $P(\text{minst 3 tillfeller av hjernesvulst})$

$$= \sum_{k=3}^{\infty} \frac{0.16^k}{k!} e^{-0.16} = 0.0006$$

Det som skjedde i Sømna var svært usannsynlig hvis kommunen hadde samme kreftrisko som landet forøvrig

Tror at det skyldes ren tilfeldighet

— Tilfellet Sømna ser foreløpig ut som en ren tilfeldighet, sier Frøydis Langmark, Kreftregisterets leder.

I 1992 ble det registrert tre tilfeller av hjernesvulst i Sømna kommune i på Helgeland. Det normale i tidligere år har vært null til ett tilfelle. Hittil i år er det ikke registrert noen tilfeller.

Leder av Kreftregisteret, Frøydis Langmark, bruker uttrykket «cluster», eller opphopning om det som er skjedd. Hun sier dette er et ikke uvanlig fenomen, og at man sjelden finner spesielle grunner til at slikt skjer. Men Kreftregisteret vil likevel fortsette med å undersøke saken, for å finne mulige årsaker.

Aftenposten 20. oktober 1993

- Hvordan kan vi forklare at krefttilfellene i Sømna skyldes en ren tilfeldighet?
- Vi må da være oppmerksom på hvorfor krefttilfellene i Sømna vakte oppsikt
- Det skjedde nettopp fordi det ble registrert uvanlig mange krefttilfeller i denne ene kommunen i dette ene året
- Alle de kommunene og alle de årene der det ikke skjer noe oppsiktsvekkende, er det ingen som bryr seg om

- Vi kan derfor spørre: Hva er sannsynligheten for at vi en gang i blant, i en eller annen kommune, vil observere noe så påfallende som det en gjorde i Sømna i 1992, ved en ren tilfeldighet?
- Regneeksempel: Vi tenker oss at vi har 100 kommuner med samme størrelse og befolkningsstruktur som Sømna, og at vi observerer antall krefttilfeller i disse kommunene i en tiårsperiode
- Sannsynligheten for at vi i minst én av kommunene vil oppleve minst tre tilfeller av hjernesvulst i løpet av ett år ved en ren tilfeldighet er:

$$1-(1-0.0006)^{1000} = 0.45$$

- Det er sannsynlig at noe så usannsynlig som det som skjedde i Sømna vil skje en gang i blant ved en ren tilfeldighet