Binomial-fordelingen, Poisson-fordelingen og hændelser i tid og rum

October 24, 2023

Rasmus Waagepetersen

Department of Mathematical Sciences
Aalborg University

AALBORG UNIVERSITY
DENMARK

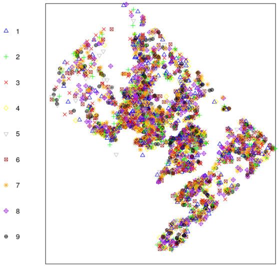
Dagens emner

- Eksempler på punktmønster-data
- Statistik for punktmønstre
- Binomial og Poisson fordeling
- Poisson som grænsefordeling for binomial
- Model for uafhængige hændelser/Poisson-processen

Crimes in Washington D.C.

Washington DC

Crime scenes in Jan-Feb 2017

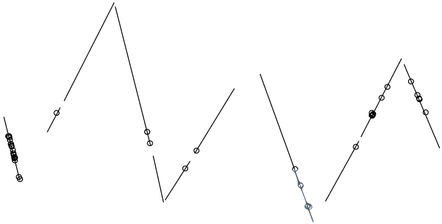


Focus on 6 types of crimes: theft, robbery, theft from car, theft of car, assault, burglary (5378 crime events)

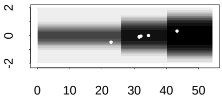
Covariate information:

- socio-economic within census tracts: income level, age, race,...
- distance to nearest police station

Whale positions



Close up:



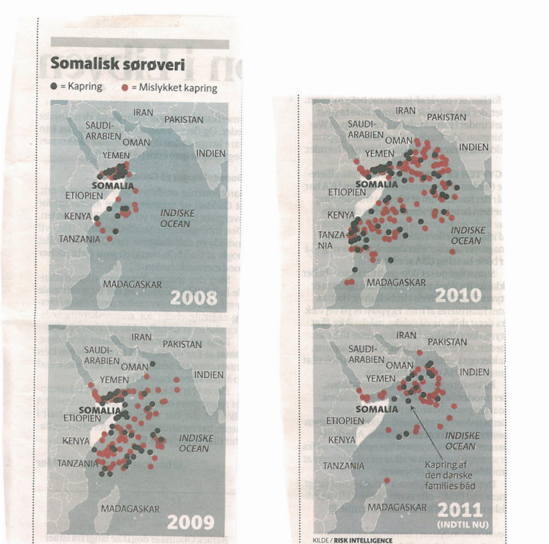
Aim: estimate whale intensity λ

Observation window W = narrow strips around transect lines

Varying detection probability: inhomogeneity (thinning)

Variation in prey intensity: clustering

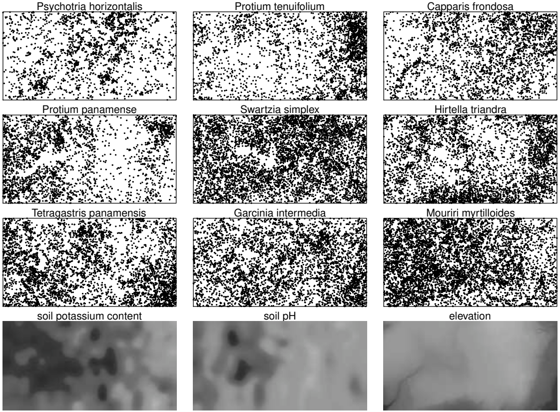
Somalian pirates - two-type space-time



Tropical rain forest

Barro Colorado Island data set: $> 300,000$ locations of trees belonging to > 300 species !

Example:
9 species and some covariates



Hændelser over tid

- opkald til en service-desk
- henfald af atomer i radioaktivt materiale
- biler, der passerer et vejkryds

Typiske spørgsmål i forbindelse med en statistisk analyse

1. Afhænger forekomsten af punkter (træer, forbrydelser) af de geografiske variable ?
2. Forekommer de enkelte punkter uafhængigt af hinanden, eller er der interaktion mellem punkterne ?
3. Er der interaktion mellem forskellige typer af punkter ?

Vi vil i dag se på spørgsmål 2. !

Hvordan undersøger vi uafhængighed ?

Ide: tæl antal nabopunkter omkring hvert punkt indenfor en given radius og sammenlign med *forventede antal*, hvis punkterne optræder uafhængigt af hinanden.

Hvis observerede antal er større eller mindre end det forventede antal tvivler vi på uafhængighed

? hvad vil det matematisk sige at punkterne er “uafhængige af hinanden”

? hvordan udregner vi den ovennævnte forventede værdi

Vi vil konstruere en model for uafhængige punkter/hændelser baseret på binomial og Poisson-fordelingerne

Binomial-fordeling

Antag et eksperiment har to udfald A og B med sandsynligheder henholdsvis p og $1 - p$ hvor $0 \leq p \leq 1$.

Gentag eksperimentet n gange og lad

$$I_i = \begin{cases} 1 & \text{hvis udfald A} \\ 0 & \text{hvis udfald B} \end{cases}$$

i $i$te eksperiment.

Lad

$$X = I_1 + I_2 + \cdots + I_n$$

være antal gange udfald A forekom.

Hvis eksperimenterne er uafhængige, er X *binomialfordelt* $bin(n, p)$.

Binomial-fordeling fortsat

X kan antage værdierne 0,1,2,...,n.

Sandsynligheden for at X er lig k er

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Her er $n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1$

Poisson-fordelingen

Hvis X har en *Poisson-fordeling* kan X antage alle ikke-negative heltallige værdier $0, 1, 2, 3, \ldots$

Hvis X har forventet værdi μ, er Poisson-sandsynlighederne givet ved

$$P(X = k) = \exp(-\mu)\frac{\mu^k}{k!}$$

Opgaver

1. Antag X er binomialfordelt $bin(3, 0.5)$. Hvad er sandsynlighederne for, at X antager værdierne 0,1,2,3 ?
2. Antag X er Poissonfordelt med $\mu = 1$. Hvad er sandsynlighederne for, at X antager værdierne 0,1,2,3 ?

Poisson-fordeling som grænsefordeling af binomialfordelingen

Resultat: kig på en følge af binomialfordelte variable $X_1, X_2, X_3, \ldots$ hvor X_n er $bin(n, \mu/n)$.

Dvs. antalsparameteren n vokser, men sandsynlighedsparameteren $p_n = \mu/n$ bliver mindre. Imidlertid er middelværdien af X_n:

$$\mathbb{E}X_n = np_n = n\mu/n = \mu$$

konstant.

Da gælder, at X_n kommer tættere og tættere på en Poissonfordeling med middelværdi μ når n bliver større og større.

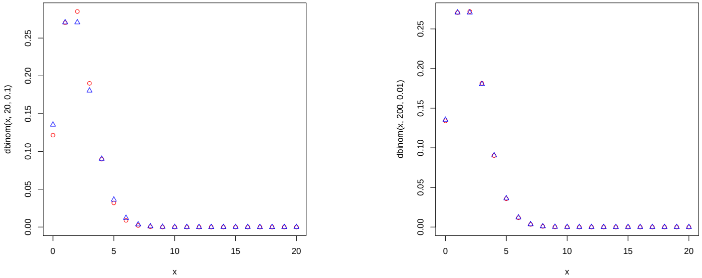
Sagt på en anden måde:

$$P(X_n = k) = \binom{n}{k}(\mu/n)^k(1-\mu/n)^{n-k} \to \exp(-\mu)\frac{\mu^k}{k!}$$

når n går mod uendelig.

Sammenligning af binomial og Poisson med samme middelværdi

bin$(20, 0.1)$ og Poisson $\mu = 2$ *bin*$(200, 0.01)$ og Poisson $\mu = 2$



Klassisk grænseværdi-resultat

Når n går mod uendelig gælder

$$(1 - x/n)^n \to \exp(-x) \tag{1}$$

Vi vil bruge dette til at vise

$$P(X_n = k) = \binom{n}{k}(\mu/n)^k(1 - \mu/n)^{n-k} \to \exp(-\mu)\frac{\mu^k}{k!} \tag{2}$$

Opgaver vedr. (1)

1. Undersøg resultatet (1) for $x = 1$ ved at udregne $(1 - 1/n)^n$ for store værdier af n og sammenlign med $\exp(-1) = 1/\exp(1)$
2. Vis, at (1) er ækvivalent med

$$n \ln(1 - x/n) \to -x$$

   (vink: se evt. regneregler vedr. $\exp$ og $\ln$ på sidste slide)
3. Vis at tangenten t til funktionen $f(z) = \ln(1 - z)$ i punktet $z = 0$ har forskrift

$$t(z) = -z$$

   (prøv evt. at tegne i Maple)
4. I fortsættelse af foregående spørgsmål gælder, at

$$\ln(1 - x/n) = -\frac{x}{n} + b_n$$

   hvor nb_n går mod nul (dette er en konsekvens af Taylors formel). Brug nu de ovenstående resultater til at konkludere (1).

Udledning af Poisson grænseresultat

Se på binomialsandsynligheden

$$\binom{n}{k}(\mu/n)^k(1-\mu/n)^{n-k} = \frac{n!}{k!(n-k)!}(\mu/n)^k(1-\mu/n)^{n-k}$$

Vis, at den kan omskrives til

$$(1-\mu/n)^n\frac{\mu^k}{k!}\frac{n(n-1)(n-2)\cdots(n-k+1)}{\underbrace{n\cdot n\cdots n}_{k \text{ gange}}}(1-\mu/n)^{-k}$$

Vink: prøv evt. først med et konkret eksempel, f.eks. $n=5$.

Udledning fortsat

Argumenter for, at

$$(1 - \mu/n)^{-k} \to 1$$

og

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{\underbrace{n \cdot n \cdots n}_{k \text{ gange}}} \to 1$$

(vink: kig på faktoren $(n-k+1)/n$ for $k = 4$ og lad $n = 10, 100, 1000$)

Dermed følger (2) af (1) og resultaterne på denne og forrige slide.

Model for uafhængige hændelser i en delmængde af planen

Betragt et kvadratisk område W med sidelængde 1 (bare for at være specifik - W kunne være en arbitrær delmængde med positivt areal)

Inddel W i $n = m^2$ delkvadrater $W_1, \ldots, W_n$, hver med sidelængde $1/m$ og areal $1/n$.

Antag, at der i W_l optræder et punkt med sandsynlighed μ/n, og at der ikke er et punkt med sandsynlighed $1 - \mu/n$. Hvis der er et punkt i W_l, placeres det på en tilfældig position i W_l. Antag endeligt, at hændelserne, at der forekommer punkter i $W_1, \ldots, W_n$ er uafhænige.

Dermed har vi opnået en model for uafhængige hændelser. Modellen afhænger dog af den konkrete opdeling i små kvadrater.

Poisson-processen

Vi kan nu frigøre os fra valget af underopdeling i små kvadrater ved at lade m og dermed n gå mod uendelig.

Der gælder så, at i grænsen vil det totale antal punkter være Poisson fordelt med middelværdi μ.

Endvidere: hvis vi ser på to disjunkte delmængder A og B af W, vil antal punkter i A og B være uafhængige.

Endeligt, antal punkter i A vil være Poisson-fordelt med middelværdi $\mu|A|$, hvor $|A|$ angiver arealet af A.

Dermed har vi opnået en model for uafhængige hændelser (som kaldes Poisson-processen).

Udledning vedr. Poisson-processen

Betragt to disjunkte delmængder A og B af W.

Bemærk: disse mængder er ikke nødvendigvis præcist givet som foreninger af små kvadrater W_l, men forskellen mellem f.eks. A og foreningen af kvadrater W_l, som er indeholdt i A kan ignoreres, når n går mod uendelig. Dermed kan vi f.eks. benytte, at antal kvadrater i A kan udregnes som $A|/(1/n)$.

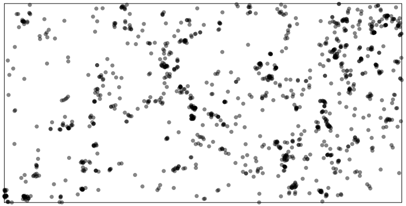
Argumenter for, at antal punkter i A og B er uafhængige for ethvert n.

Argumenter for, at for alle n er antal punkter i A binomialfordelt $bin(n|A|, \mu/n)$ (vink: jf. bemærkningen ovenfor kan vi antage, at $n|A|$ er et heltal)

Konkluder endeligt, at når n går mod uendelig, bliver antallet af punkter i A Poissonfordelt med middelværdi $\mu|A|$.

Forekommer regnskovstræerne uafhængigt af hinanden ?

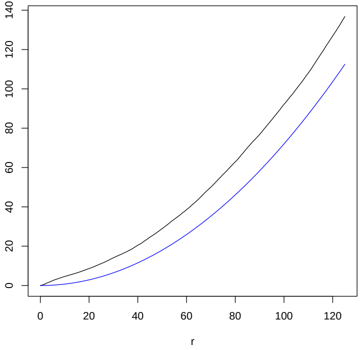
Acalypha Diversifolia observeret i 1000m$\times$500m område:



Vi estimerer μ ved det totale antal træer divideret med arealet: 1146/(500*1000)=0.002292.

For forskellige afstande $r > 0$ udregner vi gennemsnittet af antal nabopunkter indenfor afstand r for hvert punkt og sammenligner med $0.002292\pi r^2$ (hvorfor ???)

Gennemsnitligt antal naboer for data og forventet værdi hvis uafhænighed



Sort: data, blå: forventet værdi. Konklusion ?

ln og exp

Lidt genopfriskning:

$$\ln(a^b) = b\ln(a) \quad \ln(ab) = \ln(a) + \ln(b)$$

ln og exp er hinandens inverse, dvs.

$$\ln(\exp(x)) = x \text{ og } \exp(\ln(x)) = x$$

Tak for denne gang !

AALBORG UNIVERSITY
DENMARK