
Oblig 3a

Levering: 1 PDF, i rett mappe på Canvas. Lever eventuell **R**/**MatLab**/**Wolfram**-kode som kildefil i tillegg.

Førstefrist: 26. mars, 18:00

Sistefrist: 2. april, 18:00

Godkjent: 55% + (antall i gruppa) * 10%

1. (5%) Kapittel 11: oppgave 2
2. (10%) Begreper
 - (a) Hva er en *parameter*?
 - (b) Hva er en *observator*?
 - (c) Hva er en *hyperparameter*?
 - (d) Hva er en *Poisson-prosess*?
 - (e) Hva er en *Bernoulli-prosess*?
 - (f) Hva er en *posterior* fordeling?
 - (g) Hva er en *prediktiv* fordeling?

Gaussisk.

NB: Vi har i forelesningene brukt Σ_x , Σ_0 og SS_0 i stedet for S_x , S_0 og B_0

3. (5%) Kapittel 13: oppgave 1k
4. (5%) Kapittel 13: oppgave 3

Bernoulli

5. (5%) Kapittel 13: oppgave 5a
6. (5%) Kapittel 13: oppgave 6b
7. (5%) Kapittel 13: oppgave 11

Poisson

8. (5%) Kapittel 13: oppgave 14a
9. (5%) Kapittel 13: oppgave 19

10. Tre inferenser:

- (a) **Bernoulli (25%):** Du skal generere tilfeldige data fra en Bernoulli-prosess med parameter p . Bruk $p = 0.349$ for forsøket. I forsøket skal dere se om dere kan gjette p ut fra random-genererte data.
 - i. **Grunnleggende:** Hvorfor kan du bruke begge disse to kodelinjene for å generere resultatet av 30 Bernoulli-forsøk med parameter p ? Får du noe mer ut av den ene enn av den andre?

```
rbinom(30,1,p)
rbinom(1,30,p)
```

- ii. **Observasjons-versjon:** Hva viser koden under? Kjør den flere ganger, og beskriv hva som skjer.

```
p=0.349
m=50
n=10
z = rbinom(m,n,p)/n
hist(z,breaks=seq(-0.5,n+.5,1)/n)
m=mean(z)
s=sd(z)
abline(v=m,col="green",lwd=1)
abline(v=m+c(-s,s),col="pink",lwd=1)
abline(v=p,col="blue",lwd=1)
```

- iii. Endre koden ved å velge større verdier for m og n . Kjør koden. Hva er effekten av større verdi av m ? Hva er effekten av større verdi av n ? Forklar hvorfor det skjer.

- iv. **Inferens-versjon:** Versjonen over krever veldig mange forsøk. For inferens-versjonen trenger du bare generere n forsøk en eneste gang. Eller fire i denne obligen, for du skal gjøre de punkt 3, 4, 5, 6, 9 og 10 under for alle disse fire verdiene av n , ($n = 10, 100, 1000, 10000$).

- (1) La p ha prior hyperparametere a_0 og b_0 . Velg disse fra Jeffreys' nøytrale hyperparametere for Bernoulli-prosess. (Hint: se formelheftet). (Gjøre en gang)
- (2) Finn, og tegn opp, *prior* sannsynlighetsfordeling for p . (Gjøre en gang)
- (3) Simulér n forsøk, og skriv opp k =antall positive, og l =antall negative.
- (4) Finn *posterior* hyperparametre a_1 og b_1 for p .
- (5) Finn, og tegn opp, *posterior* sannsynlighetsfordeling for p .
- (6) Tegn inn $\mu_p = E[p]$ og $\mu_p \pm \sigma_p$ sammen med sannsynlighetsfordelingen.
- (7) Tegn alle sannsynlighetsfordelingene sammen i ett koordinatsystem.
- (8) Tegn de *kumulative* sannsynlighetsfordelingene sammen i ett koordinatsystem.
- (9) Prediktivt: Hvis du skulle gjøre 5 forsøk til, hva ville sannsynligheten for presis 2 suksesser være? (Bruk prediktiv fordeling for K_{+5} for hver av de fire n -verdiene.) Sammenlign svarene foran for anslått p med svaret når du bruker at du vet at $p = 0.349$.
- (10) Prediktivt: telle antall bommerter fra nå frem til du hadde 3 suksesser til, hva ville sannsynligheten for presis 4 bommerter være? (Bruk prediktiv fordeling for L_{+3} for hver av de fire n -verdiene.) Sammenlign svarene foran for anslått p med svaret når du bruker at du vet at $p = 0.349$.

- (b) **Poisson (25%):** I **R** finner vi i **library(datasets)** lista *discoveries*, som er antall store oppdagelser skjedd i årene 1860–1959. *discoveries* er antall som ble

gjort i 1860, osv. Du kan analysere dataene i hvilken software du vil, eller på hånd, men du må selvfølgelig minimum bruke **R** for å eksportere dataene.

- Lag histogram over *discoveries*, med x = antall oppdagelser i året, og la vertikal retning / areal indikere hvor mange år som hadde den oppdagelsesraten.

Hvis du bruker **R**: `hist(discoveries,breaks=k)` ... Prøv forskjellige verdier for k , og dokumentér ved å legge ved bildet du synes så best ut.

Analysere data: Tidsenheten deres t er år. Antall forekomster k er antall viktige oppdagelser i løpet av disse t årene. Dere skal analysere raten λ av viktige oppdagelser per år.

- Start med nøytrale prior hyperparametre κ_0 og τ_0 . Bare skriv ned disse.
 - Se på observasjonene for de første 3 årene. Hva er n og t for observasjonen her? Bruk disse til å finne *posterior* hyperparametre κ_1 og τ_1 . Tegn opp posterior sannsynlighetsfordeling for λ (for illustrasjonene anbefales regneverktøy som **R**, **MatLab** eller **Wolfram** sterkt fremfor frihåndtegning!).
 - Legg til observasjonene for de neste 22 årene. Hva er n og t for observasjonen her? Bruk disse til å finne *posterior* hyperparametre κ_2 og τ_2 . Tegn opp den nye posterior sannsynlighetsfordelingen for λ .
 - Legg til de resterende 75 årene med observasjoner, og finn *posterior* hyperparametre κ_3 og τ_3 . Tegn opp den nye posterior sannsynlighetsfordelingen for λ .
 - Tegn alle de tre sannsynlighetsfordelingene sammen i ett koordinatsystem.
 - La X være antall nye oppdagelser i løpet av de neste to årene etter talldataene du har analysert. Hva er $P(X = 5)$?
- (c) **Gaussisk (25%):** Les inn seigmanndataene fra oblig 1a fra *alle* gruppens .csv-filer. Se kunngjøring på Canvas om hvor dere kan laste ned filene fra, og forslag til **R**-script for å lese inn. (Dere kan bruke andre programmeringsspråk hvis dere foretrekker det.) For hver av strekktypene (Laban, Brynild, ...) gjør følgende:
- Bruk nøytrale prior hyperparametre for gaussiske prosesser. (Skriv ned!)
 - Summér opp måledataene dine i observatorer n , Σ_x , og Σ_{x^2} .
 - Finn *posterior* hyperparametre for gaussiske prosesser. Bruk μ og σ ukjent. Bruker dere .Rmd, kan dere generere dette automatisk. Anbefales!
 - Finn sannsynlighetsfordelingene for μ , τ og X_+

For kun 1 av typene, gjør som følger:

- (d) Tegn sannsynlighetsfordelingene for μ og X_+ sammen i samme koordinatsystem (for illustrasjonene anbefales regneverktøy som **R**, **MatLab** eller **Wolfram** sterkt fremfor frihåndtegning!).
- (e) Siden dere kan generere posterior med kode, så er det anbefalt å løse denne med regneverktøy: Finn posterior $\tau_{10\%}$ ved å bruke bare (cirka) 10% av dataene, og tilsvarende også $\tau_{30\%}$ og $\tau_{100\%}$. Plott de tre sannsynlighetsfordelingene sammen i samme graf, og kommenter på forskjellen.