Københavns Universitet
Det Natur- og Biovidenskabelige Fakultet

Eksamen i Statistik 1

11. april 2019

Eksamen varer 4 timer. Alle hjælpemidler er tilladt under hele eksamen, men du må ikke have internetforbindelse. Besvarelsen må gerne skrives med blyant. Eksamenssættet består af tre opgaver med i alt 12 delspørgsmål; alle delspørgsmål vægtes ens i bedømmelsen.

Opgave 1

Betragt den negative binomialfordeling med antalsparameter 2 og sandsynlighedsparameter $p \in (0,1)$, dvs. fordelingen der har tæthed

$$f_p(x) = (x+1)p^2(1-p)^x, \quad x \in \mathbb{N}_0$$

med hensyn til tællemålet på $\mathbb{N}_0$.

Det kan uden bevis benyttes at f_p faktisk er en tæthed for alle $p \in (0,1)$, og at en stokastisk variabel X med tæthed f_p har middelværdi og varians givet ved

$$\mathrm{E}_p X = \frac{2(1-p)}{p}, \quad \mathrm{V}_p X = \frac{2(1-p)}{p^2}.$$

Lad $X_1, \ldots, X_n$ være uafhængige stokastiske variable der alle har fordeling givet ved tæthed f_p med ukendt $p \in (0,1)$. Definer desuden $X_+ = \sum_{i=1}^n X_i$.

Q1: Opskriv log-likelihoodfunktionen, scorefunktionen og den observerede informationsfunktion. Bestem derefter Fisherinformationen.

Q2: Gør rede for at hvis $X_+ > 0$, så er

$$\hat{p} = \frac{2n}{X_+ + 2n} \tag{1}$$

en entydig maximum likelihood estimator for p.

Q3: Gør rede for at $\hat{p}$ også er givet ved (1) når $X_+ = 0$ hvis parameterområdet udvides til $p \in (0,1]$ og man bruger konventionen $0^0 = 1$.

I det følgende er $\hat{p}$ defineret ved (1) uanset om $X_+ > 0$ eller $X_+ = 0$.

Q4: Vis at $1/\hat{p}$ er central for $1/p$, men at $\hat{p}$ ikke er central for p.

Vink: Benyt Jensens ulighed.

Q5: Gør rede for, at $\hat{p}$ er asymptotisk normalfordelt med parametre $N\{p, p^2(1-p)/(2n)\}$.

Q6: En studerende lægger kabale hver aften i en uge. Hver aften lægger hun kabalen indtil den er gået op to gange, og skriver ned hvor mange gange kabalen *ikke* er gået op.

Hun får følgende observationer:

$$0 \quad 12 \quad 8 \quad 3 \quad 7 \quad 9 \quad 14 \tag{2}$$

Observationen 0 svarer altså til at kabalen gik op i de to første forsøg, mens observationen 12 svarer til at hun måtte lægge kabalen i alt 14 gange den pågældende aften.

Man kan argumentere for, at observationerne ovenfor kan betragtes som indbyrdes uafhængige observationer fra den betragtede negative binomialfordelingsmodel, hvor p angiver sandsynligheden for at en kabale går op når den studerende lægger den en enkelt gang.

Beregn nu et estimat og et approksimativt 95% konfidensinterval for p baseret på observationerne i (2).

Opgave 2

Betragt fordelingen P_λ der har tæthed

$$f_\lambda(x) = \frac{x}{\lambda} e^{-x^2/(2\lambda)} \cdot 1_{(0,\infty)}(x)$$

med hensyn til Lebesguemålet på $\mathbb{R}$. Fordelingen er bestemt af parameteren $\lambda > 0$.

Lad $X_1, \ldots, X_n$ være uafhængige stokastiske variable, der alle har fordeling P_λ med ukendt $\lambda > 0$.

Q1: Gør rede for at familien $\mathscr{P} = \{P_\lambda, \lambda > 0\}$ er en eksponentiel familie og bestem familiens grundmål, kanoniske stikprøvefunktion, kanoniske parameter, samt kumulantfunktion.

Q2: Bestem maksimum likelihood estimatoren for λ.

Q3: Bestem fordelingen af maksimum likelihood estimatoren for λ, og vis at den er en central estimator for λ.

Vink: Vis først at fordelingen af X_i^2 er en eksponentialfordeling med skalaparameter $1/(2\lambda)$.

Opgave 3

En sygeplejerske er mistænkt for at have forgiftet et antal patienter og for at belyse dette spørgsmål laves en opgørelse over antallet af dødsfald på den afdeling, hvor hun var tjenstgørende. Opgørelsen er delt op efter om dødsfaldene er sket på en vagt lige før hun var mødt ind, under hendes vagt, eller på vagten umiddelbart efter at hun var gået hjem. Opgørelsen er fordelt på to perioder, hvor hun var ansat på den pågældende afdeling. Resultatet af opgørelsen er angivet nedenfor.

	Inden	Under	Efter
Periode A	12	32	12
Periode B	6	18	7

Under antagelse af, at antallet af dødsfald i de forskellige vagtperioder er uafhængige og Poissonfordelte ønskes det belyst, om der er særligt mange dødsfald under den pågældendes vagt, sammenlignet med andre vagtperioder.

Betragt derfor den multiplikative Poissonmodel, altså hvor det antages, at

$$\mathrm{E}(X_{pv}) = \alpha_p \beta_v, \quad p = \mathrm{A, B}; \quad v = \mathrm{Inden, Under, Efter}.$$

hvor X_{pv} er antal dødsfald på vagten v i perioden p og parametrene $\alpha_p, \beta_v \in \mathbb{R}_+$ er ukendte.

Q1: Gør rede for, at ovennævnte model er en generaliseret lineær model med kanonisk linkfunktion.

Q2: Angiv maksimum likelihood estimatoren for $\mathrm{E}(X) = \{\mathrm{E}(X_{vp})\}$ under antagelse af ovennævnte model anvendt på det givne datamateriale.

Data kan eventuelt indlæses i `R` ved at lave en tekstfil `nurse.txt` med følgende indhold

```
Periode Vagt Antal
A Inden 12
A Under 32
A Efter 12
B Inden 6
B Under 18
B Efter 7
```

og derefter køre kommandoen

```
nurse <- read.table("nurse.txt", header=TRUE)
```

Modellen kan nu specificeres og analyseres via kommandoen

```
m <- glm(Antal ~ Vagt + Periode, family="poisson", data=nurse)
```

Q3: Undersøg nu om det kan antages, at β_v ikke afhænger af v. Gør endvidere rede for, hvad denne undersøgelse udsiger om den oprindelige problemstilling?