Københavns Universitet
Det Natur- og Biovidenskabelige Fakultet

Eksamen i Statistik 1 — Vejledende besvarelse

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 13 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.

For $x \in \mathbb{N}_0^n$ og $x. = \sum x_i$ er de ønskede funktioner givet ved (unødvendige multiplikative eller additive konstanter er udeladt):

$$L_x(p) = p^{2n}(1-p)^{x.}$$
$$l_x(p) = -2n \log p - x. \log(1-p)$$
$$Dl_x(p) = -\frac{2n}{p} + \frac{x.}{1-p}$$
$$D^2 l_x(p) = \frac{2n}{p^2} + \frac{x.}{(1-p)^2}$$

Derefter fås Fisherinformationen

$$i(p) = E_p D^2 l_X(p) = \frac{2n}{p^2} + \frac{EX.}{(1-p)^2} = \frac{2n}{p^2} + \frac{2n(1-p)}{p(1-p)^2} = \frac{2n}{p^2(1-p)}$$

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. Gør desuden rede for at formlen for $\hat{p}$ også giver mening når $X_+ = 0$.

Vink til sidste del: Hvilken fordeling svarer værdien $p = 1$ til (når 0^0 defineres til 1)?

Vi løser scoreligningen:

$$Dl_x(p) = 0 \Leftrightarrow \frac{2n}{p} = \frac{x.}{1-p} \Leftrightarrow p = \frac{2n}{x. + 2n}$$

Når $x. > 0$ ligger dette p i mængden $(0,1)$. Da $D^2 l_x(p) > 0$ for alle $p \in (0,1)$ og $(0,1)$ er en åben mængde, får vi derfor at at l_x har entydigt minimum for dette p når $x. > 0$. Altså: ML estimatoren er

$$p = \frac{2n}{X. + 2n}$$

når $X. > 0$.

Hvis $x. = 0$ giver formlen at $p = 1$ der strengt taget ligger udenfor parameterområdet. Men det giver god mening fordi fordelingen svarende til $p = 1$ er etpunktmålet (den udartede fordeling) i 0: Hvis alle x_i er 0, er det et fornuftigt bud at fordelingen er denne udartede fordeling, altså et godt bud at $p = 1$. Faktisk maksimerer $p = 1$ likelihoodfunktionen over intervallet $(0,1]$ i tilfældet $x. = 0$, thi så er $L_x(p) = p^2$.

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

Q3: Vis at $1/\hat{p}$ er central for $1/p$, men at $\hat{p}$ ikke er central for p.

Vink: Benyt Jensens ulighed.

Vi har $1/\hat{p} = (X. + 2n)/(2n)$ så

$$E_p\left(\frac{1}{\hat{p}}\right) = \frac{E_p X. + 2n}{2n} = \frac{E_p X.}{2n} + 1 = \frac{2n(1-p)}{2np} + 1 = \frac{1}{p}$$

for alle $p \in (0,1)$, så $1/\hat{p}$ er central for $1/p$.

Funktionen $t : x \to 1/x$ er strengt konveks på $(0,\infty)$ thi $t''(x) = 2x^{-3} > 0$ for alle $x > 0$. Endvidere har $1/\hat{p}$ middelværdi (se ovenfor) og $\hat{p}$ har middelværdi da den ligger i et begrænset interval. Jensens ulighed giver derfor at

$$E_p\hat{p} = E_p t(1/\hat{p}) \geq t(E_p(1/\hat{p})) = t(1/p) = p$$

med lighedstegn hvis og kun hvis fordelingen af $1/\hat{p}$ er udartet — men det er den ikke når $p \in (0,1)$. Således er $\hat{p}$ ikke central for p.

Q4: Benyt "den falske Wald-teststørrelse" til at bestemme et approksimativt 95% konfidens-interval for p. Du behøver ikke at gøre rede for forudsætningerne for den asymptotiske fordeling af Wald-teststørrelsen.

Den falske Wald-teststørrelse er givet ved

$$W = (\hat{p} - p)\, i(\hat{p})\, (\hat{p} - p) = (\hat{p} - p)^2 \frac{2n}{\hat{p}^2(1-\hat{p})}$$

og er approksimativt χ^2_1 fordelt når n er stor.

Et approksimativt 95% konfidensinterval er derfor givet ved

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}^2(1-\hat{p})}{2n}}.$$

Q5: 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 kabelen gik op i de to første forsøg, mens observationen 12 svarer til at hun måtte lægge kabalen i alt 14 gange den pågældende aften.

Gør rede for at situationen naturligt kan beskrives med den statistiske model fra denne opgave, hvor p er sandsynligheden for at kabalen går op når den studerende lægger den en enkelt gang. Beregn derefter et estimat og et approksimativt 95% konfidensinterval for p baseret på observationerne i (2).

Lad $X_1,\dots,X_7$ være de stokastiske variable svarende til de syv dage. Hvis den studerende ikke lærer med tiden, kan de antages at være uafhængige og identisk fordelte.

Hver dag antages de enkelte kabaler at gå op uafhængigt af hinanden og med samme sandsynlighed, p. At $X_i = x$ betyder at den studerende har måttet lægge kabalen $x+2$ gange. Heraf gik kabalen op sidste gang, og desuden netop en ud af de første $x+1$ gange. Der er $x+1$ forskellige muligheder for dette, som alle har sandsynlighed $p^2(1-p)^x$. Vi får derfor $P(X_i = x) = (x+1)p^2(1-p)^x$ som ønsket.

Vi indsætter således blot i formlerne fra spørgsmål 2 og 4, og får: $x. = 53$, $\hat{p} = 0.209$ og det approksimative 95% konfidensinterval $(0.112, 0.306)$.

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 $X_1,\dots,X_n$ være uafhængige reelle 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, og kanoniske parameter.

Vi sætter $\theta = 1/\lambda$ og får

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

som identificerer familien som en eksponentiel familie med grundmål $\mu(dx) x \cdot \eta(dx)$ hvor η er Lebesguemål på $(0,\infty)$, kanonisk parameter $\theta = 1/\lambda$, kanonisk stikprøvefunktion $t(x) = -x^2/2$, og kumulantfunktion $\psi(\theta) = -\log\theta$.

Q2: Bestem maksimum likelihood estimatoren for λ.

I en eksponentiel familie bestemmes maksimum likelihood estimatoren ved at sætte den kanoniske stikprøvefunktion lig med sin middelværdi. Vi får for middelværdien

$$E_\lambda(-X^2/2) = \psi'(\theta) = -1/\theta = -\lambda$$

og derfor er

$$\hat{\lambda} = \frac{1}{2n}\sum_i X_i^2.$$

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

Vink: Find først fordelingen af X_i^2.

Lad $Y_i = X_i^2$, således at

$$\hat{\lambda} = \frac{1}{2n}\sum_{i=1}^n Y_i = \frac{1}{2}\bar{Y}$$

Transformationssætningen for endimensionale transformationer giver at Y_i har tæthed

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

mht. Lebesguemålet på $\mathbb{R}$.

Vi får derfor:

- *$Y_1, \dots, Y_n$ er uafhængige og identisk exponentialfordelte med skalaparameter 2λ, dvs. gammafordelte med formparameter 1 og skalaparameter 2λ.*
- *$\sum_{i=1}^n Y_i$ er gammafordelt med formparameter n og skalaparameter 2λ jf. foldningsegenskaben for gammafordelingen*
- *$\hat{\lambda}$ gammafordelt med formparameter n og skalaparameter λ/n.*

Specielt er

$$E_\lambda\hat{\lambda} = n\frac{\lambda}{n} = \lambda$$

så $\hat{\lambda}$ er en central estimator for λ.

Q4: Betragt hypotesen $H : \lambda = \lambda_0$ for en given værdi $\lambda_0 > 0$. Opskriv kvotientteststørrelsen, $q(x_1, \dots, x_n)$, for hypotesen. Vis derefter at fordelingen af $Q = q(X_1, \dots, X_n)$ under hypotesen ikke afhænger af den specifikke værdi af λ_0.

Kvotientteststørrelsen er givet ved

$$\begin{aligned} q(x) = \frac{L_x(\lambda_0)}{L_x(\hat{\lambda})} &= \frac{\hat{\lambda}^n \exp\left(-\frac{1}{2\lambda_0}\sum x_i^2\right)}{\lambda_0^n \exp\left(-\frac{1}{2\hat{\lambda}}\sum x_i^2\right)} \\ &= \left(\frac{\hat{\lambda}}{\lambda_0}\right)^n \exp\left(-\frac{1}{2\lambda_0}\sum x_i^2 + n\right) \\ &= \left(\frac{\sum x_i^2}{2n\lambda_0}\right)^n \exp\left(-\frac{1}{2\lambda_0}\sum x_i^2 + n\right) \end{aligned}$$

Den stokastiske version er

$$Q = q(X) = \left(\frac{\sum X_i^2}{2n\lambda_0}\right)^n \exp\left(-\frac{1}{2\lambda_0}\sum X_i^2 + n\right)$$

og viser at Q kun afhænger af (X, λ_0) gennem $\frac{1}{2\lambda_0}\sum_{i=1}^n X_i^2$.

Under hypotesen er $\sum_{i=1}^n X_i^2$ gammafordelt med formparameter n og skalaparameter $2\lambda_0$, så $\frac{1}{2\lambda_0}\sum_{i=1}^n X_i^2$ er gammafordelt med formparameter n og skalaparameter 1. Altså afhænger fordelingen af Q ikke af den specifikke værdi af λ_0.

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 var 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

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

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

Q1: Gør rede for, at ovennævnte model er en generaliseret lineær model og angiv den tilhørende linkfunktion.

Poissonfordelingen er en velkendt dispersionsfamilie og modellen er anvender det kanoniske link $g(\mu) = \log \mu$ hvorefter log-middelværdien specificeres til at tilhøre det lineære underrum

$$\log \mu_{pv} = \log \alpha_p + \log \beta_v = \eta_p + \gamma_v.$$

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

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)
```

Efter kommandoen

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

fås følgende fittede værdier

```
m$fitted.values
        1         2         3         4         5         6
11.586207 32.183908 12.229885  6.413793 17.816092  6.770115
```

hvilket i tabelform svarer til

	Inden	*Under*	*Efter*
Periode A	*11.59*	*32.18*	*12.23*
Periode B	*6.41*	*17.82*	*6.77*

Q3: Kan det antages, at β_v ikke afhænger af vagten v?

Kommandoen

```
summary(m)
```

giver blandt andet følgende output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.50388    0.24289  10.309  < 2e-16 ***
VagtInden   -0.05407    0.32892  -0.164  0.86943
VagtUnder    0.96758    0.26950   3.590  0.00033 ***
PeriodeB    -0.59136    0.22386  -2.642  0.00825 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 28.577991  on 5  degrees of freedom
Residual deviance:  0.056908  on 2  degrees of freedom
```

Den meget lille residualdevians giver anledning til at have tillid til den multiplikative model. Det ses at koefficienten til "VagtUnder"er stærkt signifikant, så man kan ikke antage, at dødeligheden er ens på de tre vagttyper.

Q4: Hvad siger ovennævnte undersøgelse om den oprindelige problemstilling?

Der er konstateret en klart forøget mortalitet under den pågældendes vagter og dette skyldes ikke tilfældigheder, men må have en anden forklaring.