Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1

28. juni 2018

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 14 delspørgsmål; alle delspørgsmål vægtes ens i bedømmelsen. Data til Opgave 3 ligger på en USB-nøgle. Nøglen skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den skal altså ikke indgå som del af besvarelsen. I denne version af opgaven er der rettet et par trykfejl.

Opgave 1

Betragt X og Y uafhængige og eksponentialfordelte med $\mathbf{E}(X) = \lambda$ og $\mathbf{E}(Y) = 3\lambda$. Betragt følgende to estimatorer af λ:

$$\hat{\lambda} = (3X + Y)/6, \quad \tilde{\lambda} = \sqrt{XY/3}.$$

I det følgende kan det benyttes uden bevis at $\Gamma(1.5) = \Gamma(0.5)/2 = \sqrt{\pi}/2$, hvor

$$\Gamma(y) = \int_0^\infty u^{y-1} e^{-u}\, du$$

er gammafunktionen.

1. Hvilke af disse estimatorer er centrale for λ?
2. Beregn variansen for begge estimatorer.
3. Sammenlign varianserne med Cramér–Raos nedre grænse og kommenter resultatet.
4. Hvilken estimator har mindst kvadratisk middelfejl (mean square error)?

Opgave 2

Den inverse normalfordeling anvendes til at beskrive fordelingen af visse typer ventetider. I det specialtilfælde, hvor middelværdi og varians er ens siges fordelingen at være *standardiseret* og i så fald har den tæthedsfunktion

$$f_\mu(x) = \frac{\mu}{\sqrt{2\pi x^3}} e^{-\frac{(x-\mu)^2}{2x}}, \quad x > 0,$$

Det kan uden bevis benyttes at $\int_0^\infty f_\mu(x)\, dx = 1$ for alle $\mu > 0$ og at $\mathbf{E}(X) = \mathbf{V}(X) = \mu > 0$.

Lad nu $X_1, \ldots, X_n$ være uafhængige og standardiseret invers normalfordelte med ukendt $\mu > 0$ som ovenfor.

1. Bestem scorefunktionen $S(x,\mu)$, informationsfunktionen $I(x,\mu)$, og Fisherinformationen $i(\mu)$. *Vink:* Benyt at scorefunktionen har middelværdi 0.

2. Gør rede for, at familien af standardiserede inverse normalfordelinger med ukendt middelværdi $\mu > 0$ udgør en eksponentiel familie og angiv familiens grundmål.

3. Angiv den kanoniske parameter, den kanoniske stikprøvefunktion, samt kumulantfunktionen.

4. Vis at maximum likelihood estimatoren $\hat{\mu}$ for μ er givet som $\hat{\mu} = (1+\sqrt{1+4\bar{T}})/(2\bar{T})$, hvor $\bar{T} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}$.

5. Gør rede for, at $\hat{\mu}$ er asymptotisk normalfordelt med parametre

$$\hat{\mu} \overset{\text{as}}{\sim} N\left(\mu, \frac{\mu^2}{n(\mu+2)}\right).$$

Nedenfor er angivet et eksempel på en stikprøve som anført ovenfor med $n = 10$:

```
> x
 [1] 0.60 0.80 2.65 0.78 0.27 0.96 1.59 1.28 1.62 1.08
```

6. Under antagelse af at disse observationer følger en standard invers normalfordeling ønskes et approximativt 95% konfidensinterval for middelværdien μ.

7. Er observationerne i overensstemmelse med hypotesen $H_0 : \mu = 1$?

Opgave 3

I Ugeskrift for Læger kunne man i 1974 læse, at der var mistanke om et særligt højt antal tilfælde af lungekræft i byen Fredericia, sammenlignet med det observerede antal i nabobyerne. For eksempel var der 64 tilfælde af lungekræft blandt mænd i Fredericia i perioden 1968–1971, mens der i Vejle kun var 41 tilfælde.

Filen `cancer.txt` indeholder data som omhandler antal tilfælde af lungekræft (`Freq`) i perioden 1968–1971 hos mænd i Vejle og Fredericia i forskellige aldersgrupper samt det omtrentlige antal mænd (`Population`) i de samme aldersgrupper, bosiddende i disse byer. Aldersgrupperne er kodet som følger

	A	B	C	D	E	F
Alder	40–54	55–59	60–64	65–69	70–74	>74

For at undersøge om der kunne være tale om tilfældigheder, kunne man betragte antallet af lungekræfttilfælde X_{ab} i en given aldersgruppe a og en given by b som uafhængige og Poissonfordelte med en middelværdi af formen

$$\mathbf{E}(X_{ab}) = \alpha_a \beta_b N_{ab}$$

hvor N_{ab} angiver antallet af mandlige personer i aldersgruppe a i byen b og betragtes som fast og kendt, mens α_a og β_b er ukendte parametre, som beskriver variationen i hyppighed over aldersgrupper og lokaliteter. Modellen kan for eksempel specificeres som følger

```
glm(Freq ~ Age + Town, offset=log(Population), family="poisson", data=cancer)
```

1. Undersøg om en model af den angivne form kan beskrive data og angiv estimaterne for modellens parametre.
2. Brug modellen og analysen til at afgøre, om hyppigheden af lungekræft er forskellig i de to byer udover, hvad der kan forklares af en forskellig aldersfordeling.
3. Angiv estimater for morbiditetsraterne α_a i aldersgrupperne 40–54 og 65–69 under antagelse af at disse er ens i de to byer, altså $\beta_b \equiv 1$.