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.

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?

   Den første er central for λ mens vi for $\tilde{\lambda}$ har

   $$\mathbf{E}(\tilde{\lambda}) = \mathbf{E}(\sqrt{X})\mathbf{E}(\sqrt{Y})/\sqrt{3}$$

   og

   $$\mathbf{E}(\sqrt{X}) = \frac{1}{\lambda}\int_0^\infty \sqrt{x} e^{-x/\lambda}\, dx = \sqrt{\lambda}\,\Gamma(1.5) = \sqrt{\lambda}\frac{\sqrt{\pi}}{2}$$

   og tilsvarende for Y

   $$\mathbf{E}(\sqrt{Y}) = \sqrt{3\lambda}\frac{\sqrt{\pi}}{2}$$

   hvoraf

   $$\mathbf{E}(\tilde{\lambda}) = \lambda\frac{\pi}{4} < \lambda$$

   så $\tilde{\lambda}$ er ikke central.

2. Beregn variansen for begge estimatorer.

   Idet variansen i en eksponentialfordeling med middelværdi λ er λ^2 har vi

   $$\mathbf{V}(\hat{\lambda}) = \lambda^2(9+9)/36 = \frac{1}{2}\lambda^2.$$

For at finde variansen på $\tilde{\lambda}$ fås

$$\mathbf{E}(\tilde{\lambda}^2) = \mathbf{E}(X)\mathbf{E}(Y)/3 = \lambda^2$$

så

$$\mathbf{V}(\tilde{\lambda}) = \lambda^2 - \left(\frac{\pi}{4}\lambda\right)^2 = \frac{16-\pi^2}{16}\lambda^2.$$

3. Sammenlign varianserne med Cramér–Raos nedre grænse og kommenter resultatet.

   Vi finder log-likelihoodfunktionen pånær irrelevante konstantled

$$\ell_{X,Y}(\lambda) = \frac{X}{\lambda} + \frac{Y}{3\lambda} + 2\log\lambda$$

   og videre score- og information

$$S_{X,Y}(\lambda) = \frac{\partial \ell_{X,Y}}{\partial \lambda} = -\frac{X}{\lambda^2} - \frac{Y}{3\lambda} + \frac{2}{\lambda}$$

$$I_{X,Y}(\lambda) = \frac{\partial^2 \ell_{X,Y}}{\partial \lambda^2} = 2\frac{X}{\lambda^3} + 2\frac{Y}{3\lambda^3} - \frac{2}{\lambda^2}$$

   som giver Fisherinformationen

$$i(\lambda) = \mathbf{E}\{I_{X,Y}(\lambda)\} = \frac{2}{\lambda^2}$$

   og dermed er Cramér–Rao grænsen for en central estimator $\lambda^2/2$.

   Vi ser, at $\mathbf{V}(\hat{\lambda})$ netop har denne mindste varians.

   Den nedre grænse for den ikke-centrale estimator $\tilde{\lambda}$ er bestemt som

$$\mathbf{V}(\tilde{\lambda}) \geq \frac{\{\mathbf{E}'(\tilde{\lambda})\}^2}{i(\lambda)} = \frac{(\pi/4)^2}{2/\lambda^2} = \frac{\pi^2\lambda^2}{32} = v_{\min}.$$

   Da

$$\mathbf{V}(\tilde{\lambda}) - v_{\min} = \frac{32-3\pi^2}{32} > 0$$

   idet $3\pi^2 = 29.60881\ldots$ antager $\mathbf{V}(\tilde{\lambda})$ ikke den nedre grænse.

4. Hvilken estimator har mindst kvadratisk middelfejl (mean square error)?

   For den centrale estimator er den kvadratiske middelfejl lig med variansen. For den anden estimator fås

$$\mathrm{MSE}(\tilde{\lambda}) = \mathbf{V}(\tilde{\lambda}) + \{\mathbf{E}(\tilde{\lambda}) - \lambda\}^2 = \left\{\frac{16-\pi^2}{16} + \frac{(4-\pi)^2}{16}\right\}\lambda^2 = \frac{4-\pi}{2}\lambda^2 < \lambda^2/2.$$

   Så $\tilde{\lambda}$ har den mindste kvadratiske middelfejl.

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.

   Scorefunktionen for en enkelt observation er

   $$S(x,\mu) = \frac{\partial \ell_x(\mu)}{\partial \mu} = -\frac{1}{\mu} - \frac{(x-\mu)}{X} = -\frac{1}{\mu} - 1 + \frac{\mu}{x}$$

   og informationsfunktionen

   $$I(x,\mu) = \frac{\partial^2 \ell_x(\mu)}{\partial \mu^2} = \frac{\partial S(x,\mu)}{\partial \mu} = \frac{1}{\mu^2} + \frac{1}{x}.$$

   Idet scorefunktionen har middelværdi 0, fås

   $$\mathbf{E}(1/X) = \frac{1}{\mu} + \frac{1}{\mu^2} \tag{1}$$

   og dermed

   $$i(\mu) = \mathbf{E}\{I(X,\mu)\} = \frac{1}{\mu} + \frac{2}{\mu^2} = \frac{\mu+2}{\mu^2}.$$

   Fisherinformation for n observationer fås ved at gange med antallet af observationer. Score og informationsfunktionerne fås ved at lægge størrelserne sammen

   $$S_n(\mu) = -\frac{n}{\mu} - n + \sum_i \frac{\mu}{x_i}, \quad I_n(\mu) = \frac{n}{\mu^2} + \sum_i \frac{1}{x_i}, \quad i_n(\mu) = ni(\mu).$$

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

   Vi omskriver tætheden ved at udvikle kvadratet: $(x-\mu)^2/(2x) = x/2 + \mu^2/(2x) - \mu$ så

   $$f_\mu(x) = e^{\frac{\mu^2}{2}(-x^{-1}) + (\mu + \log\mu)} \frac{1}{\sqrt{2\pi x^3}} e^{-\frac{x}{2}} \mathbf{1}_{(0,\infty)}(x).$$

   Herat ser vi, at der er tale om en en-dimensional eksponentiel familie med grundmål

   $$\nu(dx) = \frac{1}{\sqrt{2\pi x^3}} e^{-\frac{x}{2}} \mathbf{1}_{(0,\infty)}(x) \cdot dx.$$

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

   Den kanoniske parameter er $\theta = \mu^2/2$, den kanoniske stikprøvefunktion $t(x) = -1/x$, og idet $\mu = \sqrt{2\theta}$ er kumulantfunktionen

   $$\psi(\theta) = -\mu - \log\mu = -\sqrt{2\theta} - \frac{1}{2}\log\theta - \frac{\log 2}{2}.$$

   Identiteten (1) kan naturligvis også fås ved at beregne $\tau(\theta) = \psi'(\theta) = \mathbf{E}_\theta(1/X)$.

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 $\bar{T} = \frac{1}{n}\sum_{i=1}^n \frac{1}{X_i}$.

   I en eksponentiel familie er maximum likelihood estimatoren bestemt ved at sætte den kanoniske stikprøvefunktion lig sin middelværdi, altså

   $$\frac{1}{n}\sum_{i=1}^n \frac{1}{X_i} = \mathbf{E}_\mu(1/X) = \frac{1}{\mu} + \frac{1}{\mu^2}.$$

   Sættes $\bar{T} = \frac{1}{n}\sum_{i=1}^n \frac{1}{X_i}$ fører det til andengradsligningen

   $$\bar{T}\mu^2 - \mu - 1 = 0$$

   som har præcis en positiv rod

   $$\hat{\mu} = \frac{1+\sqrt{1+4\bar{T}}}{2\bar{T}}.$$

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

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

   At estimatoren er asymptotisk normalfordelt følger af, at der er tale om en eksponentiel familie og $\mu = \sqrt{2\theta}$ er en differentiabel funktion af θ. Den asymptotiske varians er

   $$\frac{1}{ni(\mu)} = \frac{\mu^2}{n(\mu+2)},$$

   idet Fisherinformationen $i(\mu)$ blev fundet ovenfor.

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.

   Stikprøvefunktionen $\bar{t}$ beregnes til

```
> tbar = mean(1/x)
> tbar
[1] 1.227484
```

og dermed fås

```
> hatmu=(1+sqrt(1+4*tbar))/(2*tbar)
> hatmu
[1] 1.397589
```

med den asymptotiske varians og standardafvigelse

```
> asvar= hatmu^2/(length(x)*(2+hatmu))
> asvar
[1] 0.05748945
> sqrt(asvar)
[1] 0.2397696
```

hvilket giver et konfidensinterval $(0.93, 1.87)$;

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

   Da 1 ligger i konfidensintervallet er observationerne i overensstemmelse med hypotesen.

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 Poisson-fordelte 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.

   Data indlæses

   ```
   > cancer <- read.table("data/cancer.txt", header=TRUE)
   > cancer
      Age       Town Population Freq
   1    A Fredericia       3059   11
   2    A      Vejle       2520    5
   3    B Fredericia        800   11
   4    B      Vejle        878    7
   5    C Fredericia        710   11
   6    C      Vejle        839   10
   7    D Fredericia        581   11
   8    D      Vejle        631   14
   9    E Fredericia        509   11
   10   E      Vejle        539    8
   11   F Fredericia        605   10
   12   F      Vejle        619    7
   ```

   Og modellen specificeres som følger

   ```
   > p1<-glm(Freq~Age+Town, offset=log(Population),family="poisson",data=cancer)
   ```

   Hvilket giver flg. output

   ```
   Coefficients:
               Estimate Std. Error z value Pr(>|z|)
   (Intercept)  -5.7328     0.2600 -22.051  < 2e-16 ***
   AgeB          1.3398     0.3438   3.897 9.75e-05 ***
   AgeC          1.5794     0.3323   4.754 2.00e-06 ***
   AgeD          1.9929     0.3204   6.220 4.97e-10 ***
   AgeE          1.8620     0.3395   5.485 4.14e-08 ***
   AgeF          1.5931     0.3485   4.572 4.83e-06 ***
   TownVejle    -0.2918     0.1873  -1.558    0.119
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

   (Dispersion parameter for poisson family taken to be 1)

       Null deviance: 63.1074  on 11  degrees of freedom
   Residual deviance:  1.9361  on 5  degrees of freedom
   ```

   Med en residualdevians på 1.9361 på 5 frihedsgrader giver modellen en fin tilpasning.

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.

   Koefficienten som angiver hvor forskellig Vejle er fra Fredericia, er estimeret til -0.2918 med en standardafvigelse på 0.1873, så den er ikke signifikant forskellig fra 0 på noget rimeligt signifikansniveau. Med andre ord er der ingen grund til at antage at hyppigheden er forskellig i de to byer.

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 $\beta_b \equiv 1$.

   Vi specificerer en model, hvor byen ikke indgår, som følger.

```
> p2<- glm(Freq ~ Age, offset=log(Population), family="poisson", data=cancer)
> summary(p2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.8542     0.2500 -23.417  < 2e-16 ***
AgeB          1.3192     0.3436   3.839 0.000123 ***
AgeC          1.5533     0.3318   4.681 2.86e-06 ***
AgeD          1.9730     0.3202   6.163 7.15e-10 ***
AgeE          1.8440     0.3393   5.434 5.50e-08 ***
AgeF          1.5775     0.3483   4.529 5.93e-06 ***
---
Residual deviance:  4.3831  on  6  degrees of freedom
```

   Nu regnes effekten om ved at tage eksponentialfunktionen af koefficienterne

```
> eff<-exp(p2$coefficients)
> round(eff,4)
(Intercept)        AgeB        AgeC        AgeD        AgeE        AgeF
     0.0029      3.7404      4.7272      7.1924      6.3216      4.8429
```

   Altså er antallet af lungekræfttilfælde 2.9 pr tusind indbyggere i aldersgruppen 40–54; de øvrige faktorer angiver den faktor, som raten bliver forøget med i de andre aldersgrupper. For eksempel er der omtrent 21 tilfælde pr tusind indbyggere i aldersgruppen 65–69. Den samlede effekt i de øvrige aldersgrupper kan for eksempel beregnes således

```
> toteff<- eff[1]*eff[2:6]
> round(toteff,4)
  AgeB   AgeC   AgeD   AgeE   AgeF
0.0107 0.0136 0.0206 0.0181 0.0139
```