Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1

12. april 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 fem opgaver og dækker både pensum fra 2017 og 2018. Hvis man ønsker at blive bedømt i forhold til gammelt pensum (2017) afleverer man besvarelser af opgaverne 3, 4, og 5. Hvis man ønsker at blive bedømt i forhold til pensum fra 2018 afleverer man besvarelser af opgave 1, 2, og 3.

Hvis man ønsker at blive bedømt efter gammelt pensum skal det angives tydeligt på første side i besvarelsen.

Hvis intet er angivet, vil besvarelsen blive bedømt i forhold til pensum fra 2018 og kun besvarelser af opgave 1, 2, og 3 vil indgå i bedømmelsen. Angives, at besvarelsen ønskes bedømt i forhold til gammelt pensum, indgår kun besvarelser af opgaverne 3, 4, og 5 i bedømmelsen.

I begge tilfælde vil der være i alt 14 delspørgsmål, som skal bedømmes; alle delspørgsmål vægtes ens i bedømmelsen.

Data til Opgave 1 og 5 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

Ved opsendelsen af rumfærgen Challenger den 28. januar 1986 omkom syv astronauter i forbindelse med en eksplosion. En undersøgelseskommission blev nedsat og fastslog at årsagen til ulykken var en såkaldt O-ring, som gik i stykker i forbindelse med opsendelsen, sandsynligvis forårsaget af ekstremt koldt vejr. Filen `challenger.txt` indeholder data fra 23 tidligere opsendelser af rumfærgen. Kolonnen `temp` angiver temperaturen i Fahrenheit ved opsendelsen og kolonnen `critical` angiver om man i forbindelse med opsendelsen har observeret en kritisk tilstand for en af rumfærgens seks O-ringe, idet 1 angiver at man har observeret et problem og 0 at man ikke har.

1. Opstil en passende generaliseret lineær model til beskrivelse af afhængigheden mellem opsendelsestemperaturen og en kritisk begivenhed og angiv maximum likelihood estimatet for de indgående parametre.

2. Brug modellen og de angivne data til at afgøre om opsendelsestemperaturen har betydning for sandsynligheden for en kritisk hændelse.

3. Da rumfærgen blev sendt op var temperaturen 31 grader Fahrenheit. Estimer sandsynligheden for en kritisk hændelse med en O-ring ved denne temperatur.

4. Find et approksimativt 95% konfidensinterval for den samme sandsynlighed som estimeret under punkt 3. *Vink:* Brug for eksempel deltametoden på funktionen $f(\alpha,\beta) = \alpha + 31\beta$.

Opgave 2

Paretofordelingen anvendes for eksempel til at beskrive fordelingen af formuer over en givet tærskelværdi c og den har tæthedsfunktion

$$f_\theta^c(x) = \frac{\theta c^\theta}{x^{\theta+1}}, \quad \text{for } x > c,$$

hvor $c > 0$ er fast og kendt mens $\theta > 0$ er en ukendt parameter som kaldes fordelingens *index*.

1. Gør rede for, at familien af Paretofordelinger med fast tærskel c udgør en eksponentiel familie; angiv familiens grundmål.
2. Angiv familiens dimension, den kanoniske parameter, den kanoniske stikprøvefunktion, og kumulantfunktionen.
3. Lad nu $X_1, \ldots, X_n$ være uafhængige og Paretofordelte med kendt tærskel c og ukendt index θ. Find maximum likelihood estimatoren for θ og angiv dens asymptotiske fordeling.

Tidsskriftet *Forbes magazine* angiver hvert år en liste over verdens største personlige formuer. Nedenfor ses for 2018 størrelsen af alle personlige formuer over 50 milliarder US dollars som angivet af Forbes magazine.

År	Formue i milliarder US dollars								
2018	112	90	84	72	71	70	67	60	58.5

4. Under antagelse af at disse observationer følger en Paretofordeling med tærskel $c = 50$ og ukendt indexparameter, ønskes værdien af maximum likelihood estimatoren $\hat{\theta}$ samt et approksimativt 95% konfidensinterval for indexparameteren θ.

Opgave 3

Lad $X_1, \ldots, X_n$ være uafhængige og identisk gammafordelte med med samme skala- og formparameter, d.v.s. deres fordeling har tæthed

$$f_\theta(x) = \frac{x^{\theta-1} e^{-x/\theta}}{\Gamma(\theta)\theta^\theta}, \quad x > 0$$

hvor $\theta > 0$ er ukendt.

I det følgende indgår *digamma*funktionen ψ og *trigamma*funktionen ψ', hvor

$$\psi(y) = D\log\Gamma(y) = \frac{\Gamma'(y)}{\Gamma(y)}, \quad \psi'(y) = D^2\log\Gamma(y).$$

Begge funktioner er implementeret som standard i `R` og kaldes som `digamma()` og `trigamma()`.

Antag nu, at der foreligger en observation $x = (x_1, \ldots, x_n)$.

1. Bestem log-likelihoodfunktionen og scorefunktionen.
2. Bestem informationsfunktionen og Fisherinformationen.

3. Scoreligningen kan ikke løses explicit. Vis, at scoreligningen har en entydig løsning og at denne løsning $\hat{\theta}$ er maximum likelihood estimator for θ; det kan uden bevis benyttes, at

$$\psi'(y) = \sum_{k=0}^{\infty} \frac{1}{(y+k)^2}.$$

4. Angiv maximum likelihood estimatorens asymptotiske fordeling.

5. En alternativ estimator for θ er givet som

$$\tilde{\theta} = \sqrt{\frac{\sum_i x_i}{n}} = \sqrt{\bar{x}}.$$

   Gør rede for, at denne estimator er en momentestimator; er estimatoren central?

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

```
> x
 [1] 0.69 3.64 0.96 1.28 1.79 0.12 4.82 0.68 0.55 0.52
```

Værdien af maksimum likelihood estimatoren for θ baseret på den angivne stikprøve kan beregnes til $\hat{\theta} = 1.2278$.

6. Betragt nu hypotesen $H_0 : \theta = 1$, svarende til, at observationerne stammer fra en eksponentialfordeling; beregn en approksimativ p-værdi for likelihood ratio testet for den pågældende hypotese. Kunne observationerne være eksponentialfordelte?

Opgave 4

Lad (X,Y) være normalfordelt på $\mathbb{R}^2$ med middelværdi 0 og varians Σ, altså $(X,Y) \sim N(0,\Sigma)$, hvor

$$\Sigma = \begin{pmatrix} \alpha & \frac{\alpha}{2} \\ \frac{\alpha}{2} & \alpha \end{pmatrix}.$$

Her er α en konstant der opfylder visse betingelser, se spørgsmål 1.

1. For hvilke værdier af α er Σ en lovlig variansmatrix? For hvilke værdier af α er fordelingen af X en regulær hhv. singulær normalfordeling?

2. Bestem fordelingen af $\begin{pmatrix} X+Y \\ X-Y \end{pmatrix}$, og vis derved at $X+Y$ og $X-Y$ er uafhængige, $X+Y \sim N(0,3\alpha)$ og $X-Y \sim N(0,\alpha)$.

I det følgende kan du uden bevis benytte omskrivningen

$$XY = \frac{1}{4}(X+Y)^2 - \frac{1}{4}(X-Y)^2$$

samt at $\mathbf{E}(Z^4) = 3\sigma^4$ hvis $Z \sim N(0,\sigma^2)$.

3. Definer estimatoren

$$\hat{\alpha} = 2XY$$

Vis at $\hat{\alpha}$ er en central estimator for α og bestem variansen $\mathbf{V}(\hat{\alpha})$. Gør desuden rede for at $P(\hat{\alpha} \geq 0) < 1$ hvis $\alpha > 0$.

4. Definer estimatoren

$$\tilde{\alpha} = \frac{1}{2}(X^2 + Y^2)$$

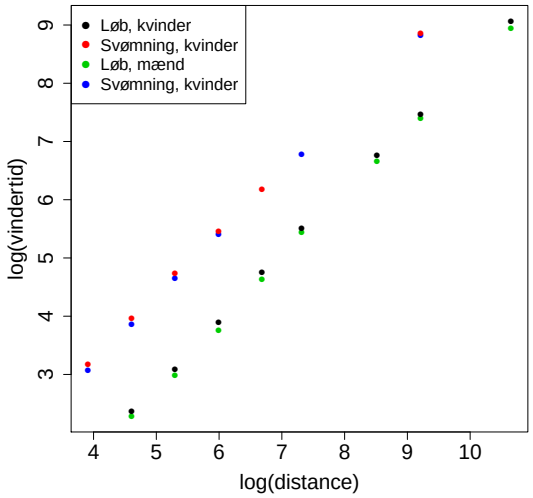
Vis at $\tilde{\alpha}$ er en central estimator for α og at variansen er $\mathbf{V}(\tilde{\alpha}) = 5\alpha^2/4$. Diskuter kortfattet om du foretrækker $\tilde{\alpha}$ eller $\hat{\alpha}$ som estimator for α.

Opgave 5

Data til denne opgave består af vindertiderne i løbe- og svømmedisciplinerne ved OL i Rio de Janeiro, 2016. Data er tilgængelige i filen `rio2016.txt` på den vedlagte USB-nøgle. Der er følgende variable:

- `koen`: Køn, med værdien 0 for mænd og 1 for kvinder
- `type`: Typen af disciplin, med værdien 0 for løb og 1 for svømning
- `distance`: Distancen for disciplinen
- `vindertid`: Vindertiden i den pågældende disciplin, målt i sekunder

Figuren nedenfor viser data, hvor både distance og vindertid er log-transformeret og punkterne er farvet efter kombinationen af køn og typen af disciplin.



Du skal først betragte den multiple regressionsmodel hvor `log(vindertid)` benyttes som responsvariabel og `log(distance)`, `type` og `koen` benyttes som forklarende variable. Hvis data er indlæst i R som `rio2016`, så kan modellen fittes med kommandoen

```
reg <- lm(log(vindertid) ~ type + koen + log(distance), data=rio2016)
```

1. Opskriv den statistiske model svarende til `reg` (med papir og blyant). Fit modellen i R, og udfør modelkontrol. Svaret vedrørende modelkontrol skal indeholde skitser af relevante figurer og kommentarer til figurerne.

Uanset hvad du konkluderede vedrørende modelkontrol i spørgsmål 1, så skal du fortsætte med modellen i spørgsmål 2–4.

Husk desuden at både `distance` og `vindertid` indgår log-transformeret i modellen.

2. Kvinderne svømmer 800 m og mændene svømmer 1500 m, men ikke omvendt. Brug modellen til at bestemme et estimat for vindertiden for kvinder på 1500 m og et estimat for vindertiden for mænd på 800 m (hvis disse discipliner fandtes).

3. Betragt to distancer hvor den ene er dobbelt så lang som den anden. Gør rede for at modellen antager at den forventede forskel i `log(vindertid)` er den samme for alle fire kombinationer af køn og disciplintype, og bestem et estimat for den fælles forventede forskel. Bestem derefter et estimat for den faktor, som vindertiden forøges med, når distancen fordobles (uanset køn og disciplintype).

4. Gør rede for at modellen antager at den forventede forskel i `log(vindertid)` mellem svømning og løb er den samme for alle distancer og begge køn, og bestem et estimat for den fælles forventede forskel. Bestem derefter et estimat for den faktor, som vindertiden er længere ved svømning end ved løb (uanset køn og distance).