Københavns Universitet
Det Naturvidenskabelige Fakultet

Eksamen i Statistik 1

6. april 2017

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 fire opgaver med i alt 14 delspørgsmål. Alle delspørgsmål vægtes ens. Data til opgave 3 ligger i filen `shootheight.txt` på en USB-stick. Sticken skal afleveres tilbage når eksamen slutter, men udelukkende for at den kan genbruges. Den kan altså ikke indgå som del af besvarelsen.

Opgave 1

Lad $X_1,\ldots,X_n,Y_1,\ldots,Y_n$ være uafhængige stokastiske variable hvor alle X_i er eksponentialfordelte med middelværdi θ og alle Y_i er eksponentialfordelte med middelværdi $1/\theta$. Her er $\theta > 0$ en ukendt parameter. Bemærk at den samme parameter indgår i fordelingen af både X_i'er og Y_i'er.

1. Opskriv likelihoodfunktionen for en observation $(x,y) = (x_1,\ldots,x_n,y_1,\ldots,y_n)$, og vis at scorefunktionen er givet ved

   $$S_{x,y}(\theta) = -\frac{1}{\theta^2}S_x + S_y$$

   hvor $S_x = \sum_{i=1}^n x_i$ og $S_y = \sum_{i=1}^n y_i$. Bestem desuden Fisherinformationen.

2. Bestem maksimum likelihood estimatoren og angiv dens asymptotiske fordeling.

3. Bestem fordelingerne af $S_X = \sum_{i=1}^n X_i$ og $S_Y = \sum_{i=1}^n Y_i$, og vis derefter at $Z = \theta^2 \frac{S_Y}{S_X}$ er F-fordelt med $(2n,2n)$ frihedsgrader under P_θ. Benyt resultatet til at bestemme et eksakt 95% konfidensinterval for θ.

   Vink: Du kan uden bevis benytte at hvis U og V er uafhængige og begge gammafordelte med formparameter λ og skalaparameter β, så er $\frac{V}{U}$ F-fordelt med $(2\lambda,2\lambda)$ frihedsgrader.

4. Betragt datasættet bestående af følgende observationer (for $n = 7$):

   ```
   > x
   [1] 1.90 3.64 5.79 4.36 4.15 0.04 1.93
   > y
   [1] 1.40 0.74 0.53 0.41 0.32 0.25 0.80
   ```

   Bestem det eksakte 95% konfidensinterval for θ for disse data. Bestem også et approksimativt 95% konfidensinterval baseret på den „falske" Waldteststørrelse. Kommentér forskellen på de to konfidensintervaller.

   Du kan uden bevis bruge den asymptotiske fordeling af Waldteststørrelsen.

Opgave 2

Lad X være normalfordelt på $\mathbb{R}^3$ med middelværdi 0 og varians Σ, altså $X \sim N(0,\Sigma)$, hvor

$$\Sigma = \begin{pmatrix} 1 & \varphi & 0 \\ \varphi & 4 & 0 \\ 0 & 0 & \varphi \end{pmatrix}.$$

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

1. For hvilke værdier af φ er Σ en lovlig variansmatrix? For hvilke værdier af φ er fordelingen af X en regulær hhv. singulær normalfordeling?
2. Bestem fordelingen af $\begin{pmatrix} X_1 + X_2 \\ X_3 \end{pmatrix}$.
3. Definér to potentielle estimatorer for φ:

$$\tilde{\varphi} = \frac{(X_1 + X_2)^2 - 5}{2} \quad \text{og} \quad \hat{\varphi} = X_3^2.$$

   Gør rede for at både $\tilde{\varphi}$ og $\hat{\varphi}$ er centrale for φ og bestem deres varianser. Antag endelig at $\varphi = 1$, og bestem sandsynlighederne $P(0 < \tilde{\varphi} < 2)$ og $P(0 < \hat{\varphi} < 2)$. Kommentér resultaterne i forhold til hvilken estimator du foretrækker.

   Vink: Du kan uden bevis benytte at $\mathrm{V}(Z^2) = 2\sigma^4$ hvis $Z \sim N(0,\sigma^2)$.

Opgave 3

Bladlus lever af planter og er derfor et problem for økologisk landbrug hvor der ikke må sprøjtes. Det er blevet foreslået at behandle såsæden (frøene der sås) med en bestemt svampetype inden såning. Håbet er at svampebehandlingen gør det mindre attraktivt for bladlusene at leve på planterne, dog uden at påvirke plantevæksten negativt. I denne opgave undersøges effekten på vækst af majsplanter, mere præcist højden af skuddene på majsplanter 10 dage efter såning.

Der er foretaget to eksperimenter under sammenlignelige, men ikke identiske omstændigheder. I hvert eksperiment blev 36 potter delt tilfældigt i to grupper. Halvdelen af potterne blev tilsået med et svampebehandlet frø, den anden halvdel var kontroller tilsået med et ubehandlet frø. Der er således i alt fire grupper af observationer. I eksperiment 1 var der to svampebehandlede frø der ikke spirede, så der er kun er 16 planter med svampebehandling.

Data er tilgængelige i filen `shootheight.txt` på den vedlagte USB-stick. Der er to variable:

- `group`: En faktor med fire niveauer, nemlig `control1`, `control2`, `fungus1`, `fungus2`. For eksempel angiver `fungus1` at observationen stammer fra en potte med svampebehandlet frø fra eksperiment 1.
- `sh`: En numerisk variable med højden af majplanternes skud (i mm).

Det kan være nyttigt at tegne data inden analysen, men det er ikke en del af besvarelsen. Brug fx kommandoen `stripchart(sh~group, data=dataNavn)` hvor `dataNavn` er navnet på dit R-datasæt.

1. Angiv en statistisk model der kan bruges til at undersøge om der er forskel mellem skudhøjden i de fire grupper. Fit modellen i R og udfør modelkontrol. Du skal skitsere og kommentere de figurer du laver.

2. Undersøg med et enkelt hypotesetest om den forventede skudhøjde er den samme i de fire grupper.

Lad δ_1 hhv. δ_2 være forskellen i forventet skudhøjde mellem svampebehandlede planter og kontroller i ekperiment 1 hhv. eksperiment 2.

3. Brug din model fra spørgsmål 1 til at bestemme estimater og 95% konfidensintervaller for δ_1 og δ_2.

4. Lad $\bar{\delta} = \frac{1}{2}(\delta_1 + \delta_2)$ være den gennemsnitlige forskel (over eksperiment) i skudhøjde mellem svampebehandlede planter og kontroller. Bestem et estimat og et 95% konfidensinterval for $\bar{\delta}$. Husk at man gerne vil undgå at svampebehandlingen påvirker plantevæksten negativt; hvad kan man konkludere om dette udfra data?

   Vink: Hvordan kan $\bar{\delta}$ skrives som funktion af parametrene i modellen?

Opgave 4

Lad $X_1, \ldots, X_n$ være uafhængige stokastiske variable hvor alle $X_i \sim N(0, \sigma^2)$ og variansen $\sigma^2 > 0$ er en ukendt parameter. Med andre ord: $X \sim N(0, \sigma^2 I)$. Bemærk at middelværdien er kendt.

Definér $SS_x = x^T x = \sum_{i=1}^n x_i^2$ henholdsvis $SS_X = X^T X = \sum_{i=1}^n X_i^2$.

Opgaven handler om test for hypotesen $H : \sigma^2 = 1$.

1. Gør rede for at maksimum likelihood estimatoren for σ^2 i modellen er $\hat{\sigma}^2 = \frac{1}{n} SS_X$. Vis derefter at likelihood ratio teststørrelsen for hypotesen er givet ved

$$LR(x) = -2 \log Q(x) = SS_x - n \log(SS_x) + n \log(n) - n.$$

Som alternativ til likelihood ratio testet kan vi bruge $\hat{\sigma}^2$ som teststørrelse og afvise hypotesen hvis $n\hat{\sigma}^2 \notin (z_1\,, z_2)$, hvor z_1 og z_2 er 2.5% og 97.5% fraktilen i χ^2 fordelingen med n frihedsgrader. Vi definerer altså det kritiske område hhv. acceptområdet som

$$\mathscr{K} = \{x \in \mathbb{R}^n \,|\, SS_x < z_1 \text{ eller } SS_x > z_2\}$$
$$\mathscr{A} = \{x \in \mathbb{R}^n \,|\, z_1 < SS_x < z_2\}$$

Da $SS_X = X^T X$ er χ^2 fordelt med n frihedsgrader under hypotesen, er dette et test på niveau 5%.

2. Betragt observationen bestående af følgende værdier ($n = 10$):

```
> x
 [1]  1.74  2.07 -1.20 -0.40 -1.97 -1.87 -2.16  1.48 -1.00 -1.02
```

   Angiv $\hat{\sigma}^2$ for observationen og udfør likelihood ratio testet for hypotesen $H : \sigma^2 = 1$. Du kan benytte det sædvanlige asymptotiske resultat vedr. $LR(X) = -2 \log Q(X)$ uden bevis.

   Angiv også om $x \in \mathscr{K}$ eller $x \in \mathscr{A}$ for det alternative test.

3. I dette spørgsmål skal du sammenligne niveau og styrke for de to test. Det er hele tiden hypotesen $H : \sigma^2 = 1$ der testes. Mere præcist skal du for $n = 10$ og tre forskellige sande værdier af σ^2 (se tabellen nedenfor) gøre følgende:

   - Simulere et udfald af $x = (x_1, \ldots, x_n)$.
   - Udføre likelihood ratio testet for hypotesen $H : \sigma^2 = 1$ for det simulerede x. Du kan bruge det sædvanlige asymptotiske resultat vedr. $LR(X) = -2\log Q(X)$ uden bevis.
   - Udføre det alternative test for det simulerede x, altså undersøge om $x \in \mathscr{K}$ eller $x \in \mathscr{A}$.
   - Gentage dette mindst 2000 gange og for hver testmetode beregne med hvilken relativ hyppighed hypotesen forkastes.

   Besvarelsen af spørgsmålet skal bestå af en udfyldt version af skemaet nedenfor samt kommentarer til resultaterne.

	Relativ hyppighed hvormed hypotesen forkastes	
Sand værdi af σ^2	LR test	Alternativt test
1	**	**
0.5	**	**
1.5	**	**