Københavns Universitet
Det Natur- og Biovidenskabelige Fakultet

Eksamen i Matematisk Statistik
20. juni 2019

4 timers skriftlig prøve. Alle hjælpemidler tilladt (inkl. computer uden netforbindelse). Besvarelsen må skrives med blyant. Opgavesættet består af 4 opgaver med i alt 16 delopgaver. Ved bedømmelsen vægtes alle delopgaver ens. Data til Opgave 3 findes på den udleverede USB-nøgle i filen `MatStatJuni2019.txt`*. USB-nøglen skal returneres efter eksamen, men udelukkende for at den kan genbruges. Filer på denne USB-nøgle vil således ikke kunne indgå som en del af besvarelsen.*

Opgave 1

Den inverse normalfordeling anvendes til at beskrive fordelingen af visse typer ventetider. Den har tæthedsfunktion

$$f_{\mu,\lambda}(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0.$$

Det kan uden bevis benyttes at $\int_0^\infty f_{\mu,\lambda}(x)\,dx = 1$ for alle $(\mu,\lambda) \in (0,\infty)^2$.

Lad nu $X_1,\ldots,X_n$ være uafhængige og invers normalfordelte med ukendt $(\mu,\lambda) \in (0,\infty)^2$ og lad

$$\mathcal{P} = \{P_{\mu,\lambda}, \mu > 0, \lambda > 0\}$$

hvor $P_{\mu,\lambda}$ har tæthedsfunktion $f_{\mu,\lambda}$ som ovenfor.

1. Betragt først delfamilien $\mathcal{P}_0 \subseteq \mathcal{P}$ bestemt ved restriktionen $\mu = 1$. Gør rede for at denne kan repræsenteres som en eksponentiel familie af dimension 1 med kanonisk stikprøvefunktion $t(x) = -(x + 1/x)/2$.

2. Vis at maximum likelihood estimatoren $\hat{\lambda}_n$ for λ i delfamilien $\mathcal{P}_0$ er givet som

$$\hat{\lambda}_n = \frac{1}{\bar{X}_n + \bar{Y}_n - 2},$$

hvor

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad Y_i = \frac{1}{X_i}, \quad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n \frac{1}{X_i}.$$

3. Vis at informationen $i_0(\lambda)$ for λ i delfamilien $\mathcal{P}_0$ baseret på en enkelt observation er bestemt som
$$i_0(\lambda) = \frac{1}{2\lambda^2}$$
og angiv den asymptotiske fordeling af $\hat{\lambda}_n$.

4. Vis at familien $\mathcal{P}$ kan repræsenteres som en eksponentiel familie af dimension 2 med
$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \lambda/\mu^2 \\ \lambda \end{pmatrix}, \quad t(x) = \begin{pmatrix} -x/2 \\ -1/(2x) \end{pmatrix}$$
som kanonisk parameter og kanonisk stikprøvefunktion, samt kumulantfunktion
$$\psi(\theta) = -\sqrt{\theta_1\theta_2} - \frac{1}{2}\log\theta_2;$$
angiv familiens grundmål i denne repræsentation. Det kan uden bevis anvendes at denne repræsentation er minimal og regulær.

5. Vis, at middelværdi og varians i fordelingen er bestemt som
$$\mathbf{E}_{\mu,\lambda}(X) = \mu, \quad \mathbf{V}_{\mu,\lambda}(X) = \frac{\mu^3}{\lambda}.$$

6. Vis at maximum likelihood estimatoren $(\tilde{\mu}_n, \tilde{\lambda}_n)$ for (μ, λ) i familien $\mathcal{P}$ er bestemt som
$$\tilde{\mu}_n = \bar{X}_n, \quad \tilde{\lambda}_n = \frac{\bar{X}_n}{\bar{X}_n\bar{Y}_n - 1}$$
med samme notation som under spørgsmål 2.

7. Betragt nu hypotesen $H_0 : \mu = 1$ i modellen bestemt ved $\mathcal{P}$. Angiv den asymptotiske fordeling af kvotientteststørrelsen for H_0. Begrund dit svar.

Opgave 2

Lad $X = (X_1, X_2, X_3)^T$ være normalfordelt på $\mathbb{R}^3$ med middelværdi og varians
$$\xi = EX = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{og} \quad \Sigma = \mathrm{Var}(X) = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

1. Afgør om X er singulært eller regulært normalfordelt på $\mathbb{R}^3$.
2. Vis at $Y_1 = X_1 - X_2$ og $Y_2 = X_3 - X_2$ er uafhængige og standardnormalfordelte, og angiv (de marginale) fordelinger af Y_2^2 og $Y_1/\sqrt{Y_2^2}$.

Opgave 3

Ved et interventionsforsøg med 188 børn ønsker man at undersøge, om det er muligt at påvirke børn til bedre at kunne lide grøntsager, hvis man eksponerer dem for snackbarer (her: stænger af dadelmasse), som er blevet tilsat rå grøntsager. Ved forsøgets start blev de 188 børn inddelt i 5 eksponeringsgrupper og over en periode på to uger blev de otte gange tilbudt en snackbar. Typen af snackbar var forskellig i de 5 eksponeringsgrupper. Inden eksponeringsperiodens start og efter eksponeringsperioden blev hvert barn bedt om at smage på det samme grøntsagsprodukt og angive hvor godt de kunne lide produktet på en *liking*-skala med værdierne $1, 2, \ldots, 7$ (højere værdi angiver at børnene bedre kan lide produktet). Der er således to observationer fra hvert barn. Vi opfatter *liking* som en (diskretiseret version af en) kontinuert variabel der med rimelighed kan analyseres med normalfordelingsmodeller.

På vedlagte USB-nøgle findes et datasæt `MatStatJuni2019.txt` som indeholder responsvariablen (`liking`), eksponeringsgruppe (`G`), tidspunkt (`time`: before / after) samt variablen `subj` der identificerer målinger hørende til samme barn. Data er venligst stillet til rådighed af Annemarie Olsen fra Institut for Fødevarevidenskab på Københavns Universitet.

Datasættet kan indlæses i R med følgende kommando

```
data <- read.table(file = "MatStatJuni2019.txt", header = T)
```

og de første linjer i datafilen ser ud som følger

```
head(data)

##   subj G   time liking
## 1 S101 N before      4
## 2 S101 N  after      2
## 3 S102 N before      6
## 4 S102 N  after      6
## 5 S103 N before      1
## 6 S103 N  after      5
```

De 5 eksponeringgrupper (=labels for faktoren `G`) er

- `B/C`: beetroot/carrot (dansk: rødbede/gulerod) snackbar
- `Control`: kontrolgruppe - fik ikke nogen snackbarer i eksponeringsperioden
- `N`: neutral bar - som ikke var tilsat grøntssager
- `P/S`: pumpkin/sweet potato (dansk: græskar/sød kartoffel) snackbar
- `S/J`: spinach/jerusalem artichoke (dansk: spinat/jordskok) snackbar

Indlæs data fra filen `MatStatJuni2019.txt` i R og besvar følgende.

1. Betragt følgende design (dvs. samling af faktorer)

$$\mathbb{G} = \{\texttt{subj}, \texttt{G} \times \texttt{time}, \texttt{G}, \texttt{time}, 1\}.$$

   Argumenter for at faktorerne `G` og `time` er geometrisk ortogonale og find minimum af faktorerne `subj` og `G`.

2. Opstil en passende varianskomponentmodel der kan tages som udgangspunkt for en analyse af, hvordan børnenes `liking` af grøntsagsproduktet påvirkes af interventionen. Angiv variansmatricen for de to målinger målinger, der hører til barnet givet ved `subj = S101`.

3. Fit modellen fra delspørgsmål 2. i R og angiv variansestimaterne samt estimater for den forventede/gennemsnitlige `liking` både før og efter eksponeringsperioden for børn i eksponeringsgruppe `G = B/C`.

4. Foretag en passende analyse der skal belyse om børnenes `liking` ændres hen over interventionsforsøget, herunder om ændringen påvirkes af, hvilken eksponeringsgruppe børnene kommer i. Husk at skrive en konklusion på din analyse.

   Hint: Der er flere mulige løsninger på dette delspørgsmål. Du kan enten fortolke på baggrund af estimaterne fra modellen i forrige delspørgsmål, eller du kan udføre et likelihoodratiotest for en eller flere relevante hypoteser.

Opgave 4

Nogle hjertesygdomme hos hunde viser sig blandt andet ved, at venstre forkammer i hjertet begynder at vokse. I forbindelse med diagnose af visse hjertesygdomme er der derfor behov for at kende normalværdierne for hjertevolumen hos raske hunde.

I denne opgave betragtes et datasæt bestående af målinger af hjertevolumen af 97 raske hunde (dvs. uden hjertesygdom). Datasættet indholder udover volumen i mL (`maxLA`) også hundes vægt i kg (`wgt`) og oplysninger om hundes `race`. De første linjer i datasættet kan ses her sammen med en optælling af antal observationer fordelt på `race`.

```
##             race  wgt maxLA
## 1 Border_Terrier  9.2  6.07
## 2 Border_Terrier  6.9  4.67
## 3 Border_Terrier  7.7  4.40
## 4 Border_Terrier 11.0  4.48
##
## Border_Terrier   Grand_Danois       Labrador   Petit_Basset        Whippet
##             24             19             17             20             17
```

Datasættet er venligst stillet til rådighed af Miriam Höllmer. Der indgår 5 forskellige racer i datasættet. Ved besvarelsen af opgaven skal du benytte R udskriften sidst i opgaven.

Lad X_i betegne målingen (=responsen) hørende til den i-te hund i datasættet. Vi antager, at X_i'erne er uafhængige og normalfordelte $\sim N(\xi_i, \sigma^2)$

1. Angiv middelværdiunderrummet for den lineære normale model svarende til `model1` i R udskriften. Angiv dimensionen af middelværdiunderrummet og værdien af maksimaliseringsestimatoren (MLE) samt dennes fordeling.

For at kunne udtale sig om hjertevolumen for andre hunderacer end dem der indgår i datasættet, så ønsker man at lave en regressionsmodel, hvor hundens vægt (`wgt`) benyttes til at forklare hjertevolumen.

2. Opskriv en ligning for middelværdien svarende til den af regressionsmodellerne `modA`, `modB`, `modC` som du bedst mener opfylder antagelserne bag en lineær normal model. Begrund dit svar.

3. Tag udgangspunkt i dit valg af model blandt `modA, modB, modC` fra 2. Benyt modellen til at lave et 95 % prædiktionsområde for hjertevolumen for en hund på 25 kg.

R-udskrift til brug ved besvarelse af Opgave 4

```
## read in data from file

data <- read.table(file = "data/hunde.txt", header = T, sep = "\t")
```

```
model1 <- lm(maxLA ~ race, data = data)
summary(model1)
```

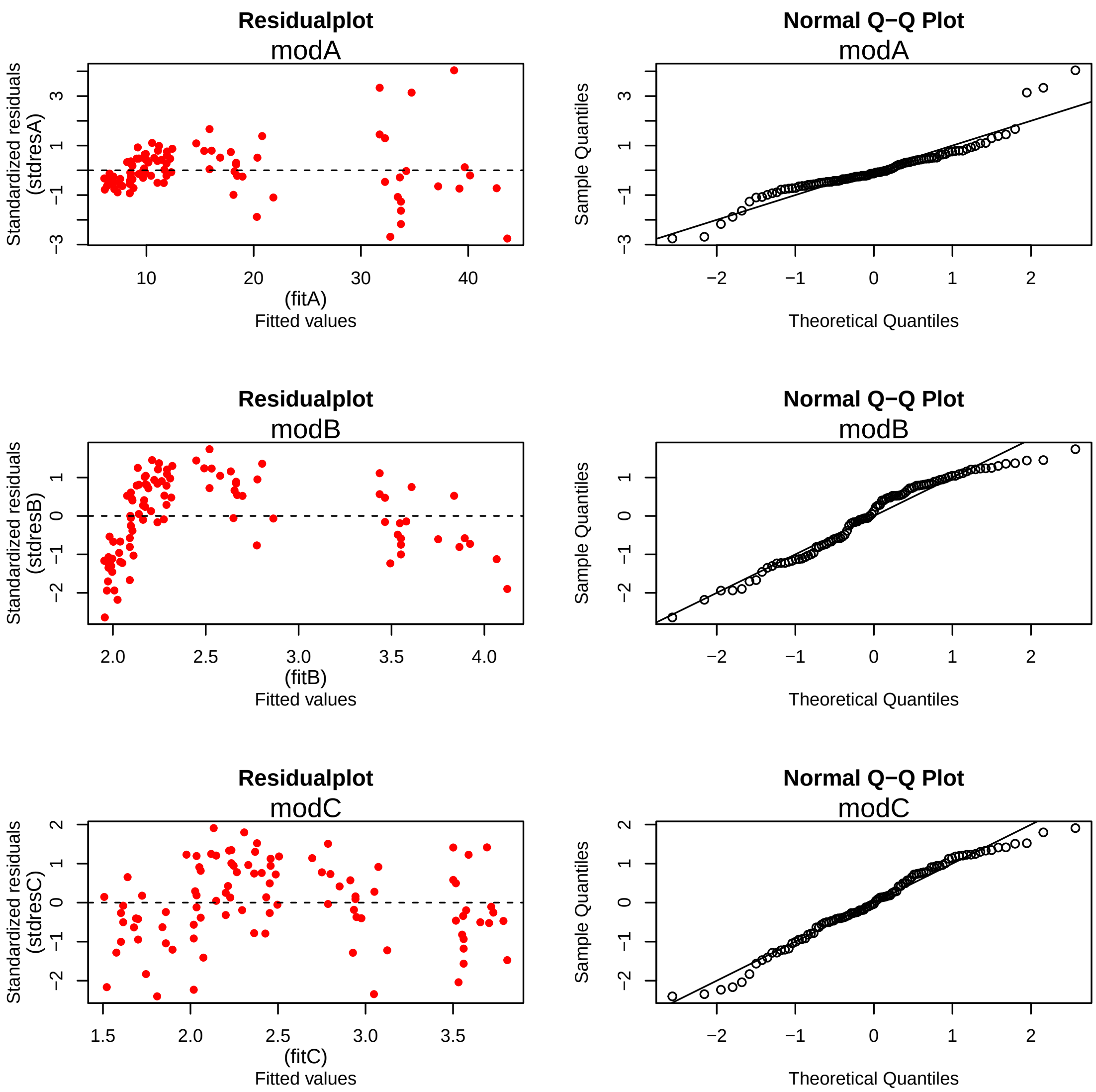
```
##
## Call:
## lm(formula = maxLA ~ race, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1179  -1.4241  -0.1113   1.2688  20.1621
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.0812     0.8784   5.784 9.94e-08 ***
## raceGrand_Danois  30.3866     1.3215  22.994  < 2e-16 ***
## raceLabrador      13.8429     1.3642  10.147  < 2e-16 ***
## racePetit_Basset   7.0097     1.3030   5.380 5.64e-07 ***
## raceWhippet        5.4599     1.3642   4.002 0.000127 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.304 on 92 degrees of freedom
## Multiple R-squared:  0.8661,Adjusted R-squared:  0.8602
## F-statistic: 148.7 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
###  fit af regressionsmodeller

modA <- lm(maxLA ~ wgt, data = data)
modB <- lm(log(maxLA) ~ wgt, data = data)
modC <- lm(log(maxLA) ~ log(wgt), data = data)

fitA <- fitted(modA)
stdresA <- rstandard(modA)
fitB <- fitted(modB)
stdresB <- rstandard(modB)
fitC <- fitted(modC)
stdresC <- rstandard(modC)
```

```
### residualplot og QQ-plot mod standardnormalfordeling
```



```r
### estimater

summary(modA)

##
## Call:
## lm(formula = maxLA ~ wgt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.416  -2.166  -0.557   2.028  16.942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.99463    0.68875   4.348 3.45e-05 ***
## wgt          0.49575    0.02046  24.227  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.319 on 95 degrees of freedom
## Multiple R-squared:  0.8607,Adjusted R-squared:  0.8592
## F-statistic:   587 on 1 and 95 DF,  p-value: < 2.2e-16

summary(modB)

##
## Call:
## lm(formula = log(maxLA) ~ wgt, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93398 -0.27200  0.04478  0.29211  0.61691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.776098   0.056943   31.19   <2e-16 ***
## wgt         0.028624   0.001692   16.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3571 on 95 degrees of freedom
## Multiple R-squared:  0.7508,Adjusted R-squared:  0.7482
## F-statistic: 286.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

```r
### estimater

summary(modC)
```

```
##
## Call:
## lm(formula = log(maxLA) ~ log(wgt), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55682 -0.13167 -0.00815  0.18163  0.44446
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.11931    0.09684  -1.232    0.221
## log(wgt)     0.89163    0.03172  28.107   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2344 on 95 degrees of freedom
## Multiple R-squared:  0.8927,Adjusted R-squared:  0.8915
## F-statistic:   790 on 1 and 95 DF,  p-value: < 2.2e-16
```

```r
### den inverse til (A^T * A) hvor A er designmatricen for parametriseringen af modA
designA <- model.matrix(modA)
solve(t(designA) %*% designA)
```

```
##           [,1]      [,2]
## [1,]  0.025429 -0.000583
## [2,] -0.000583  0.000022
```

```r
### den inverse til (A^T * A) hvor A er designmatricen for parametriseringen af modB
designB <- model.matrix(modB)
solve(t(designB) %*% designB)
```

```
##           [,1]      [,2]
## [1,]  0.025429 -0.000583
## [2,] -0.000583  0.000022
```

```r
### den inverse til (A^T * A) hvor A er designmatricen for parametriseringen af modC
designC <- model.matrix(modC)
solve(t(designC) %*% designC)
```

```
##           [,1]      [,2]
## [1,]  0.170702 -0.054205
## [2,] -0.054205  0.018319
```