Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Reeksamen, februar 2023

Fire timers skriftlig prøve. Alle hjælpemidler tilladt, herunder computer, men du må ikke tilgå internettet bortset fra i forbindelse med udlevering og aflevering af eksamensopgaven.

Der er 3 opgaver, som vægtes med henholdsvis 50 %, 25 % og 25 % i bedømmelsen.

Husk at mange delspørgsmål kan besvares uafhængigt af hinanden.

Data til opgave 1 og opgave 2 bliver gjort tilgængelige sammen med selve eksamensopgaven. Alle svar i opgave 1 og opgave 2 skal begrundes. Husk at det ikke er tilstrækkeligt at aflevere R-kode med tilhørende output. Du skal derimod svare på det du bliver spurgt om, typisk ved at du skriver svaret i almindelig tekst, hvori du inddrager relevante tal fra R-outputtet.

Opgave 3 består af multiple choice spørgsmål. For hvert multiple choice spørgsmål er der netop et korrekt svar, og din besvarelse skal udelukkende bestå af *bogstavet* ud for dit valgte svar. Du kan altså ikke begrunde svaret. Hvis et multiple choice spørgsmål besvares forkert, ikke besvares, eller flere svar bliver givet, så gives 0 point.

Din besvarelse skal afleveres elektronisk. Du skal følge de officielle instrukser du har modtaget omkring aflevering af opgaven.

Opgave 1

Denne opgave vægtes med 50 % ved bedømmelsen, og svarene skal begrundes.

Ved et forsøg har man målt respirationen i 12 jordprøver. Ved forsøgets start blev jordprøverne tilsat et plantemateriale (italiensk rajgræs). Hver af de 12 prøver blev analyseret netop en gang på dag 2, 4, 6, 7, 9 eller 12 efter forsøgets start, hvor man målte den totale (kumulerede) mængde kuldioxid (CO_2) der var blevet dannet fra forsøgets start til analysetidspunktet. Til hvert tidspunkt blev der analyseret to jordprøver.

Datafilerne `feb2023opg1.txt` og `feb2023opg1.xlsx` indeholder data fra forsøget og kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data1 <- read_excel(path = "feb2023opg1.xlsx")
```

eller

```
data1 <- read.table(file = "feb2023opg1.txt", header = T)
```

De første linjer i datasættet kan ses her

```
##   day      co2
## 1   2 1.521962
## 2   2 1.447735
## 3   4 2.773195
## 4   4 3.560853
## 5   6 5.155170
## 6   6 4.305960
```

Variablen `co2` beskriver den totale (kumulerede) mængde CO_2 i prøven. Variablen `day` angiver antal dage fra forsøgets start til målingen af CO_2 blev foretaget.

1. Opskriv (i din besvarelse) den statistiske model der fittes med koden

   ```
   mod1 <- lm(co2 ~ factor(day), data = data1)
   ```

   Angiv estimater for den forventede totale mængde CO_2 fra analyser foretaget på dag 2 og dag 12.

2. Du skal tage udgangspunkt i modellen `mod1` fra delopgave **1.1** ved besvarelse af denne delopgave.

   Angiv et 95 % - konfidensinterval for den forventede mængde CO_2 på dag 9 efter forsøgets start.

   Angiv desuden et 95 % - konfidensinterval for *forskellen* i den forventede mængde CO_2 på dag 12 og på dag 9 efter forsøgets start.

3. Benyt `R`-koden

```
linreg <- lm(log(co2) ~ day, data = data1)
```

   til at bestemme estimater for parametrene α, β og σ^2 i den lineære regressionsmodel

$$\log(\texttt{co2}_i) = \alpha + \beta \cdot \texttt{day}_i + e_i,$$

   hvor e_i'erne er uafhængige og normalfordelte $\sim \mathcal{N}(0, \sigma^2)$.

   Forklar, hvordan man skal fortolke værdien $\hat{\beta}$ af estimatet for β.

   Angiv desuden et 95 % - konfidensinterval for β.

4. Udfør et test for, om det er rimeligt at antage, at der er en lineær sammenhæng mellem forventet værdi af $\log(CO_2)$ og antallet af dage siden forsøgets start.

   Hint: Der er flere korrekte løsninger på dette spørgsmål. Det er vigtigt at du forklarer, hvilken model du tager udgangspunkt i, for at teste hypotesen.

5. Benyt modellen fra delspørgsmål **1.3** til at angive et estimat og et 95 % - konfidensinterval for indholdet af CO_2 fra analyser foretaget 10 dage efter forsøgets start.

6. Benyt modellen fra delspørgsmål **1.3** til at bestemme et 95 % - prædiktionsinterval for den totale mængde CO_2 i en måling taget 10 dage efter forsøgets start.

Opgave 2

Denne opgave vægtes med 25 % ved bedømmelsen, og svarene skal begrundes.

I denne opgave betragtes en udvidelse af datasættet fra Opgave 1, hvor man også har målt den totale (kumulerede) mængde CO_2 i 12 prøver med kontroljord (dvs. uden tilsat plantemateriale). Der er også blevet analyseret to prøver med kontroljord på hver af dagene 2, 4, 6, 7, 9 og 12 efter forsøgets start.

Datafilerne `feb2023opg2.txt` og `feb2023opg2.xlsx` indeholder data fra forsøget og kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data2 <- read_excel(path = "feb2023opg2.xlsx")
```

eller

```
data2 <- read.table(file = "feb2023opg2.txt", header = T)
```

Datasættet indeholder 24 linjer, hvoraf de første kan ses her

```
##   treat day      co2
## 1   raj   2 1.521962
## 2   raj   2 1.447735
## 3   raj   4 2.773195
## 4   raj   4 3.560853
## 5   raj   6 5.155170
## 6   raj   6 4.305960
```

Variablen `co2` beskriver den totale (kumulerede) mængde CO_2 i prøven. Variablen `treat` angiver om prøven var tilsat rajgræs (`treat = raj`) eller ej (`treat = kontrol`), mens `day` angiver antal dage fra forsøgets start til målingen af CO_2 blev foretaget.

1. Angiv `R`-koden til at fitte en model for tosidet variansanalyse med vekselvirkning, hvor både `treat` og `day` inddrages i modellen som kategoriske variable. Du bedes benytte log(`co2`) som responsvariabel i modellen.

   Fit modellen i `R` og angiv et estimat for indholdet af CO_2 i en jordprøve med rajgræs som er foretaget 9 dage efter forsøgets start.

   Angiv desuden et estimat for *forskellen* i den totale mængde CO_2 fra analyser af kontrolprøver og jordprøver med rajgræs der foretages 9 dage efter forsøgets start.

2. Undersøg om forskellen i den forventede værdi af responsen $\log(CO_2)$ mellem de to typer jordprøver er den samme, uanset hvor lang tid efter forsøgets start analyserne foretages.

3. Diskuter grundigt, om det er rimeligt at beskrive sammenhængen mellem den totale (kumulerede) mængde CO_2 og variablene `treat` samt `day` ved den *blandede model*, der kan fittes med `R`-koden

```
mod2 <- lm(log(co2) ~ treat + day, data = data2)
```

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 25 % i bedømmelsen. For hvert delspørgsmål er der netop et korrekt svar. Din besvarelse skal udelukkende bestå af bogstavet for dit valgte svar; du kan altså ikke begrunde svaret. Hvis et spørgsmål besvares forkert, ikke besvares eller flere svar bliver givet, gives 0 point. Du må naturligvis gerne bruge R til opgaven.

3.1 Den gennemsnitlige fodlængde for børn i alderen fra 9-11 år kan antages at være normalfordelt med middelværdi 19.664 cm og spredning 0.929 cm. Beregn sandsynligheden for, at et tilfældigt valgt barn i denne aldersgruppe har en fodlængde på mellem 18 og 21 cm.

A. Ca. 81.4 %

B. Ca. 92.5 %

C. Ca. 68.3 %

D. Ca. 46.3 %

E. Ca. 88.8 %

3.2 Den gennemsnitlige fodlængde for børn i alderen fra 9-11 år kan antages at være normalfordelt med middelværdi 19.664 cm og spredning 0.929 cm. Bestem værdien, L, så 5 % af børn i alderen fra 9-11 år har en fodlængde mindre end L.

A. $L \approx 21.19$ cm

B. $L \approx 21.48$ cm

C. $L \approx 19.62$ cm

D. $L \approx 17.84$ cm

E. $L \approx 18.13$ cm

3.3 For at undersøge en metode til behandling af sår, har man på baggrund af billeder målt udbredelsen/arealet af såret (i cm^2) før og efter behandling. Der indgår data fra otte personer, og datasættet kan ses her

```
my_data
```

```
##   before after
## 1   99.2  86.0
## 2   90.2  85.3
## 3   81.2  93.3
## 4   98.1  85.9
## 5   93.7  95.6
## 6  110.9  96.6
## 7   90.9  84.0
## 8  110.0  97.0
```

Hvilken af følgende R-kommandoer gør det muligt **direkte i outputtet** at aflæse en P-værdi for test af hypotesen om, at den forventede ændring af sårets areal er 0 hen over behandlingsperioden?

A. `t.test(my_data$before, my_data$after-my_data$before, paired = T)`

B. `summary(lm(after ~ before, data = my_data))`

C. `summary(lm(after - before ~ 1, data = my_data))`

D. `t.test(my_data$before, my_data$after, paired = F)`

E. `summary(lm(after - before ~ before, data = my_data))`

3.4 En terning kastes 100 gange og vi interesserer os for antallet af seksere. Det antages at der i hvert kast er sandsynlighed 1/6 for at slå en sekser.

Find sandsynligheden for, at man slår 25 seksere eller mere i de 100 kast med terningen.

A. Ca. 1.6 %

B. Ca. 2.2 %

C. Ca. 1.2 %

D. Ca. 1.0 %

E. Ca. 97.8 %

3.5 For at undersøge om en terning er skæv, har man kastet den 100 gange og hver gang noteret antallet af øjne. Resultatet fremgår af følgende tabel (fx. ses at der er observeret 17 seksere)

```
## x
##  1  2  3  4  5  6
## 21 17 12 18 15 17
```

Hvad kan man konkludere på baggrund af følgende R-output?

```
chisq.test(c(21, 17, 12, 18, 15, 17), p = c(1, 1, 1, 1, 1, 1)/6)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(21, 17, 12, 18, 15, 17)
## X-squared = 2.72, df = 5, p-value = 0.7431
```

A. Vi forkaster hypotesen om, at der er lige stor sandsynlighed for at få 1, 2, 3, 4, 5 og 6 øjne.

B. Ikke noget. Der er benyttet et forkert test. Man bør hellere lave et homogenitetstest (test for ens proportioner).

C. Vi kan ikke forkaste hypotesen om, at der er lige stor sandsynlighed for at få 1, 2, 3, 4, 5 og 6 øjne.

D. Ikke noget. Der er benyttet et forkert test. Man bør hellere lave et test for uafhængighed.

E. Der er 74.3 % sandsynlighed for, at terningen ikke er skæv.

3.6 På et kursus deltog 75 studerende i eksamen, hvoraf 15 opnåede topkarakteren 12. Det følgende år var der kommet en ny underviser på kurset, og her blev der givet karakteren 12 til 10 ud af de 30 studerende som gik til eksamen.

Angiv P-værdien samt en konklusion på baggrund af et test for, om andelen af studerende som opnår topkarakter er ens, uanset hvem der underviser på kurset. Du skal udføre testet uden kontinuitetskorrektion.

A. $P = 0.1473$ så andelen der opnår topkarakter kan *ikke* antages at være den samme uanset hvem der underviser.

B. $P = 0.1473$ så andelen der opnår topkarakter kan antages at være den samme uanset hvem der underviser.

C. $P = 0.2658$ så andelen der opnår topkarakter kan antages at være den samme uanset hvem der underviser.

D. $P = 0.2319$ så andelen der opnår topkarakter kan *ikke* antages at være den samme uanset hvem der underviser.

E. $P = 0.2319$ så andelen der opnår topkarakter kan antages at være den samme uanset hvem der underviser.