Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Reeksamen, februar 2022

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 25 %, 50 % 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 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 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. Der udleveres også et datasæt som kan benyttes ved besvarelse af Opgave 3.

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

Opgave 1

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

I et spiringsforsøg undersøgte man 3 græssorter af arten rødsvingel, nemlig `napoli`, `smirna` og `symphony`. For hver art (`type`) blev der taget 16 prøver af 100 frø. De 100 frø i hver prøve blev sået i et spiringskammer og antal spirede frø blev dernæst observeret en til to gange dagligt i omkring 20 dage. På baggrund af disse observationer blev en middelspiretid for hver prøve beregnet.

Datafilerne `feb2022opg1.txt` og `feb2022opg1.xlsx` indeholder data fra forsøget og kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data1 <- read_excel(path = "feb2022opg1.xlsx")
```

eller

```
data1 <- read.table(file = "feb2022opg1.txt", header = T)
```

De første linjer i datasættet kan ses her

```
##     type  tid
## 1 napoli 5.77
## 2 napoli 5.19
## 3 napoli 5.52
## 4 napoli 5.13
## 5 napoli 5.48
## 6 napoli 5.73
```

Variablen `tid` angiver middelspiretid for de 100 frø i en prøve. Der er således totalt 48 målinger (16 fra hver art).

1.1 Angiv `R`-koden til at fitte en ensidet variansanalysemodel med middelspiringstiden (`tid`) som respons og `type` som forklarende variabel.

Angiv estimatet for residualspredningen og estimatet for den forventede middelspiringstid for arten `symphony`.

1.2 Lav et hypotesetest med henblik på at undersøge, om den forventede middelspiringstid kan antages at være ens for alle tre arter.

1.3 Angiv et estimat for forskellen i den forventede middelspringstid for frø af arterne `smirna` og `symphony`.

Diskuter om der er forskel på middelspiringstiden for arterne `smirna` og `symphony`.

Opgave 2

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

Med henblik på at undersøge formen af gulerødder blev der i sommeren 2010 udført et dyrkningsforsøg. I datasættet indgår sammenhørende værdier af `omkreds` målt i den tykke ende og længde (`length`) for 67 gulerødder (-alle mål angivet i cm). De 67 gulerødder fordeler sig på 3 forskellige sorter givet ved variablen `variety` med 3 niveauer `gul`, `orange` og `roed`.

Datafilerne `feb2022opg2.txt` og `feb2022opg2.xlsx` indeholder data fra forsøget, og kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data2 <- read_excel(path = "feb2022opg2.xlsx")
```

eller

```
data2 <- read.table(file = "feb2022opg2.txt", header = T)
```

De første fire linjer i datasættet ses her

```
##   variety omkreds length
## 1     gul     8.1   12.5
## 2     gul    10.6   11.0
## 3     gul     8.0    7.5
## 4     gul     8.8   11.5
```

Ved besvarelsen af delopgaverne **2.1**-**2.5** skal du ikke benytte variablen `variety`.

2.1 Opskriv (i din besvarelse) den statistiske model der fittes med `R`-koden

```
mod1 <- lm(length ~ omkreds, data = data2)
```

og angiv estimater for samtlige parametre i modellen.

Et udpluk af et `summary()` af modellen `mod1` kan ses her

```
summary(mod1)
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.9602306  0.6525882 -1.471419 1.460029e-01
## omkreds      1.1904812  0.1007707 11.813764 7.789958e-18
```

2.2 Angiv den forventede længde af en gulerod med en diameter på 2 cm (svarende til en `omkreds` på $\pi \cdot 2 \approx 3.14 \cdot 2 = 6.28$ cm).

Diskuter ved at se på et relevant prædiktionsinterval, om det vil være usædvanligt, at en gulerod med en diameter på 2 cm har en længde på 8 cm.

2.3 En tommelfingerregel siger, at længden af en gulerod er ca. 3 gange diameteren. Dette kan udtrykkes som en hypotese om, at den forventede længde er givet som

$$E(\texttt{length}) = 3 \cdot \texttt{diameter} = 3/\pi \cdot \texttt{omkreds}.$$

Se på relevante konfindensintervaller eller udfør et (eller flere) hypotesetest med henblik på at undersøge om datasættet understøtter tommelfingerreglen.

2.4 Tag udgangspunkt i den statistiske model som fittes med `R`-koden

```
mod2 <- lm(length ~ omkreds + I(omkreds^2), data = data2)
```

Angiv estimaterne for modellens parametre og forklar, hvordan de skal fortolkes.

Giv et forslag til, hvordan man kan bruge `mod2` til at undersøge, om den forventede længde af en gulerod er en lineær funktion af omkredsen.

2.5 En alternativ statistisk model til analyse af data kunne være `mod3` givet ved

$$\texttt{mod3}: \quad \log(\text{length}_i) = \alpha + \beta \cdot \log(\text{omkreds}_i) + e_i,$$

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

Diskuter grundigt (bl.a. ved at inddrage relevante figurer i din besvarelse) om der er grund til at foretrække `mod3` fremfor `mod1`.

Datasættet indeholder desuden sorten af gulerødderne angivet ved variablen `variety`.

2.6 Giv et forslag til en statistisk model, hvor man inddrager både `omkreds` og sort (`variety`) som forklarende variable.

For at besvare delopgaven fuldstændigt bedes du både

- opskrive modellen i din besvarelse
- angive `R`-koden til at fitte modellen
- angive estimater og forklare hvordan estimaterne fra modellen skal fortolkes

2.7 Undersøge om sorten af guleroden (`variety`) har betydning for sammenhængen mellem længde og omkreds af en gulerod. Husk at forklare din metode.

Hint: Der er flere korrekte løsninger på dette delspørgsmål. I forbindelse med din løsning bør du kommentere på udvalgte estimater, konfidensintervaller og evt. hypotesetest fra relevante statistiske modeller.

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.

På baggrund af en stikprøve af omkredsen af 67 gulerødder har man vurderet, at omkredsen kan beskrives ved en normalfordeling med middelværdi 6.31 cm og spredning 1.47 cm. Benyt disse oplysninger til at besvare delopgaverne **3.1-3.2** nedenfor.

3.1 Beregn sandsynligheden for at en tilfældig valgt gulerod har en omkreds på mellem 3 og 8 cm.

A. Ca. 98.8 %

B. Ca. 12.5 %.

C. Ca. 72.0 %.

D. Ca. 87.5 %.

E. Ca. 86.3 %.

3.2 Hvilket af følgende udsagn er korrekt?

A. 10 % af gulerødderne har en omkreds på under 8.73 cm.

B. 5 % af gulerødderne har en omkreds på over 9.19 cm.

C. 10 % af gulerødderne har en omkreds på over 8.73 cm.

D. 30 % af gulerødderne har en omkreds på over 7.08 cm.

E. 30 % af gulerødderne har en omkreds på under 7.08 cm.

3.3 Vi kaster en mønt 56 gange og observerer 20 plat. Vi lader q betegne sandsynligheden for, at mønten viser plat. Vi ønsker at teste hypotesen om at $q = 1/2$, og der skal benyttes et signifikansniveau på 5 %. Hvad kan vi konkludere?

Hint: For at løse opgaven skal du selv udføre testet i `R`. Uanset hvilken af metoderne fra kurset, som du benytter, så vil du få samme svar blandt mulighederne A-E.

A. P-værdien er mellem 5 % og 10 %, så vi forkaster hypotesen om, at $q = 1/2$

B. P-værdien er mellem 5 % og 10 %, så vi kan ikke forkaste hypotesen om, at $q = 1/2$

C. P-værdien er under 5 %, så vi forkaster hypotesen om, at $q = 1/2$

D. P-værdien er over 10 %, så vi kan ikke forkaste hypotesen om, at $q = 1/2$

E. P-værdien er under 5 % , så vi kan ikke forkaste hypotesen om, at $q = 1/2$

3.4 Ved den første forelæsning i Statistisk Dataanalyse 1 i både 2020 og 2021 har de studerende svaret på spørgsmålet:

Har du set frem til kurset Statistisk Dataanalyse 1?

De studerendes svar fremgår af følgende tabel

```
##                  Ja Ved_ikke Nej
## SD1 årgang 2021 115       47  15
## SD1 årgang 2020  86       62  29
```

Man har udført et homogenitetstest på baggrund af tabellen.

Hvad er P-værdien og konklusionen?

A. $P = 0.0047$, og de studerende har mere negative forventninger til SD1 i 2021 end i 2020.

B. $P = 0.0047$, og de studerende har mere positive forventninger til SD1 i 2021 end i 2020.

C. $P = 0.0047$, og de studerendes forventninger til SD1 er ikke forskellig i 2020 og i 2021.

D. $P = 0.0089$, og de studerende har mere positive forventninger til SD1 i 2021 end i 2020.

E. $P = 0.0089$, og de studerende har mere negative forventninger til SD1 i 2021 end i 2020.

3.5 Ved et fiktivt forsøg inddeles 30 personer (`subject`) tilfældigt i to lige store behandlingsgrupper `treat = A` eller `treat = B`. Der foretages målinger af den samme responsvariabel både før (`x`) og efter (`y`) forsøget. Strukturen af de første linjer i det tilhørende datasæt (her kaldet `data3`) er organiseret som vist her

```
head(data3)
```

```
##   subject treat     x     y
## 1       1     A 5.900 7.018
## 2       2     B 3.827 5.659
## 3       3     A 4.103 3.544
## 4       4     B 3.555 4.334
## 5       5     A 4.669 4.852
## 6       6     B 2.099 4.176
```

Hvilken af følgende R-koder vil gøre det muligt **direkte i outputtet** at aflæse en P-værdi for test af hypotesen om, at den forventede ændring i responsen er ens i de to behandlingsgrupper?

Hint: Datafilerne `feb2022opg3.txt` og `feb2022opg3.xlsx` indeholder data, så du har mulighed for at prøve at køre R-koderne selv, men det burde ikke være nødvendigt for at besvare opgaven.

A. `summary(lm(y ~ treat + x, data = data3))`

B. `summary(lm(y - x ~ treat, data = data3))`

C. `summary(lm(y ~ x, data = data3))`

D. `summary(lm(y - x ~ treat - 1, data = data3))`

E. `t.test(data3$x, data3$y, paired = T)`

3.6 I et forskningsprojekt blev kræftpatienter tilfældigt allokeret til en af to behandlinger (`treat`). Der indgår patienter med to diagnoser i projektet (givet ved variablen `diagnose`). Variablen `y` er et mål for patienternes effekt af behandlingen.

Man har kørt følgende `R`-kode

```
m1 <- lm(y ~ treat * diagnose, data = data)
m2 <- lm(y ~ treat + diagnose, data = data)
anova(m2, m1)
```

og observerer en F-teststørrelse på 2.718 med tilhørende P-værdi på 0.115. Der skal anvendes et signfikansniveau på 5 % ved fortolkningen af resultatet.

Hvad kan man konkludere på baggrund af testet?

Den forventede forskel mellem effekten af de to behandlinger er ...

A. ... 0 i begge de to diagnosegrupper.

B. ... ens i de to diagnosegrupper, men forskellig fra 0.

C. ... ens i de to diagnosegrupper.

D. ... forskellig fra 0 i mindst en af de to diagnosegrupper.

E. ... forskellig i de to diagnosegrupper.