Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, januar 2020

Fire timers skriftlig prøve. Alle hjælpemidler er tilladt, herunder computer, men du må ikke tilgå internettet.

Der er 3 opgaver, som vægtes med henholdsvis 40 %, 32 % og 28 % i bedømmelsen. Husk at de fleste spørgsmål kan besvares uafhængigt af hinanden.

Data til opgave 1 og opgave 2 udleveres på en USB-nøgle. Navne på filer der indeholder data fremgår af opgaveteksten. Denne USB-nøgle skal afleveres efter eksamen, så den kan genbruges. 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, dvs. udtrække de relevante tal fra R-outputtet og svare i almindelig tekst.

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.

Du kan vælge at aflevere hele eller dele af besvarelsen som pdf-fil på den USB-nøgle som udleveres til formålet af eksamensvagterne. Bemærk at kun pdf-format accepteres og at du skal benytte en anden USB-nøgle end den som data udleveres på. Den håndskrevne del af besvarelsen må gerne skrives med blyant. Besvarelsen af multiple choice spørgsmålene, dvs. de valgte svarmuligheder, kan med fordel skrives ind på den sidste side af opgavesættet, og denne side kan vedlægges besvarelsen.

Opgave 1

Denne opgave vægtes med 40 % ved bedømmelsen, og svarene skal begrundes.

Ved et dyrkningsforsøg undersøgtes effekten af tilsætning af nitrat på tørstofproduktionen af salat. Salaten blev dyrket i 16 forskellige kar, hvor hvert kar blev tilført en nøje afmålt mængde nitrat. I forsøget anvendtes fire tilsætningsmængder af nitrat (0.5, 1.0, 2.0 og 3.0 i passende enheder). Resultaterne (gram tørstof for hvert af de 16 kar) kan findes i filerne `salat.txt` og `salat.xlsx`. Datasættet består af 16 datalinjer (en for hvert kar), og de første linjer kan ses nedenfor

```
##   nitrat  stof
## 1    0.5 23.95
## 2    0.5 26.18
## 3    0.5 25.02
## 4    0.5 22.71
## 5    1.0 34.24
## 6    1.0 31.59
```

1.1 Opskriv den statistiske model som fittes med R-koden

```
model1 <- lm(stof ~ factor(nitrat), data = salat)
```

hvor datasættet er indlæst i R under navnet `salat`.

Angiv et estimat for den forventede mængde tørstof, hvis der tilføres nitratmængder på henholdsvis 0.5 og 3.0. Angiv også estimatet for residualspredningen.

1.2 Udfør et hypotesetest med henblik på at undersøge, om tørstofmængden kan antages at være ens, uanset hvilken mængde nitrat der tilsættes til dyrkningskaret.

1.3 Angiv et estimat for den forventede mængde tørstof ved tilsætning af en nitratmængde på 1.0, når du tager udgangspunkt i modellen `model2` som fittes med R-koden

```
model2 <- lm(stof ~ nitrat, data = salat)
```

Angiv desuden (ligeledes baseret på `model2`) et 95 %-prædiktionsinterval for mængden af tørstof svarende til en måling fra et kar, der har fået tilsat nitratmængden 1.0.

1.4 Diskuter grundigt om det er rimelig at beskrive sammenhængen mellem tilsat nitratmængde og tørstofproduktion vha. af `model2`. Det kan være relevant at vedlægge relevante grafer elektronisk eller lave skitser af dem i hånden.

1.5 Udfør et statistisk test, hvor du sammenligner modellerne `model1` og `model2`. Forklar hvad man kan konkludere på baggrund af testet.

Opgave 2

Denne opgave vægtes med 32 % ved bedømmelsen, og svarene skal begrundes. Data er venligst stillet til rådighed af Nora Badawi.

Udvaskning af pesticider fra golfbane-arealer udgør en væsentlig miljøbelastning. I denne opgave betragtes et datasæt, hvor man har målt herbicidet MCPAs evne til at binde sig til jorden (også kaldet sorptionsevnen). Datafilerne `pestgolf.txt` og `pestgolf.xlsx` består af målinger af sorptionsevnen (sorptionskoefficienten `Kd` målt i mL/g) fra 18 jordprøver. Der indgår jordprøver fra tre forskellige steder givet ved variablen `Lokation` med niveauer `KNY, HONE` og `DYR`. Desuden er der for hver jordprøve anvendt et af to produkter, som indeholder herbicidet MCPA. I datasættet angives de to produkter ved `Treat = T04` og `Treat = T05`. Der er i alt 18 målinger: tre målinger (replikater) for hver af de seks kombinationer af `Lokation` og `Treat`.

2.1 Forklar kortfattet hvorfor det er naturligt at benytte en tosidet variansanalyse til disse data.

Angiv en `R`-kommando der kan bruges til at estimere den tosidede variansanalysemodel med vekselvirkning, hvor du bruger variablen `Kd` som responsvariabel og de andre variable som forklarende variable.

Angiv et estimat for den forventede sorption i en jordprøve fra `Lokation = HONE`, hvorpå der er anvendt produktet `Treat = T05`.

2.2 Undersøg med et hypotesetest, om der er vekselvirkning mellem `Lokation` og `Treat`, og forklar kortfattet hvad resultatet betyder.

Ved besvarelsen af de følgende delopgaver **2.3** og **2.4** skal du tage udgangspunkt i en additiv model for tosidet variansanalyse. Hvis datasættet er indlæst i `R` under navnet `pestgolf`, så kan modellen fittes med `R`-koden

```
modelAdd <- lm(Kd ~ Treat + Lokation, data = pestgolf)
```

2.3 Angiv et estimat og et 95 % konfidensinterval for den forventede forskel i sorptionen i jordprøver fra stederne omtalt som DYR og HONE.

2.4 Undersøg med et hypotesetest om der er forskel på sorptionen i jord fra KNY og HONE.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 28 % 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. Svar på multiple choice spørgsmål kan med fordel afleveres ved at indføre svarene på sidste side af opgaven og aflevere siden. Du må naturligvis gerne bruge R til opgaven.

3.1 I forbindelse med en patologisk undersøgelse af sorte han-mink, som har været brugt til avl, har man foretaget målinger af testikellængden. Det antages, at testikellængden er normalfordelt med middelværdi 24.1 mm og spredning 2.0 mm. Beregn sandsynligheden for at en tilfældigt valgt sort han-mink har en testikellængde på mellem 20 og 25 mm.

A. 0.653

B. 0.980

C. 0.436

D. 0.674

E. 0.736

3.2 På baggrund af en tilfældig stikprøve angiver 20 ud af 100 adspurgte personer, at de primært spiser plantebaseret kost. På baggrund af en velkendt formel bestemmes et simpelt 95 %-konfidensinterval for andelen der spiser plantebaseret kost til 0.200 ± 0.078 dvs. (0.122-0.278).

Hvad bliver det tilsvarende konfidensinterval, hvis undersøgelsen i stedet var baseret på svar fra 400 personer, hvoraf 80 spiser plantebaseret kost?

A. $(0.061 - 0.556)$

B. $(0.122 - 0.278)$

C. $(0.043 - 0.357)$

D. $(0.180 - 0.220)$

E. $(0.161 - 0.239)$

3.3 På baggrund af en stikprøve bestående af 79 sorte han-mink fandt man ud af at gennemsnitsvægten var 3043 g og spredningen var 233 g.

Angiv et 95 % konfidensinterval for den gennemsnitlige vægt af sorte han-mink, når vi antager at vægten for en sort han-mink kan beskrives ved en normalfordeling.

A. (2586-3500) g

B. (3037-3049) g

C. (3017-3069) g

D. (2579-3507) g

E. (2991-3095) g

3.4 For 366 han-mink som har været brugt til avl har man klassificeret

- parringsvilligheden i grupperne: `ingen, mellem, hoj`
- størrelsen (længden) af testiklerne i grupperne: `lille, stor`

Når de 366 mink inddeles efter de to inddelingskriterier fås følgende tabel

```
tab1
```

```
##
##         hoj ingen mellem
##   lille 142    17     31
##   stor  159     9      8
```

Hvad kan vi konkludere på baggrund af følgende R-output

```
chisq.test(tab1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 16.474, df = 2, p-value = 0.0002646
```

A. Parringsvilligheden er ikke uafhængig af testikelstørrelsen.

B. En tosidet variansanalyse viser, at der ikke er vekselvirkning mellem testikelstørrelse og parringsvillighed.

C. Der er uafhængighed mellem de to inddelingskriterier (testikelstørrelse og parringsvillighed).

D. Der er ikke lige mange hanmink i datasættet, som har store og små testikler.

E. En tosidet variansanalyse viser, at der er vekselvirkning mellem testikelstørrelse og parringsvillighed.

3.5 Ved kast med en almindelig terning er der sandsynlighed 1/6 for at slå en sekser. Hvor mange terninger skal man mindst slå med for at sikre, at der er mere end 80 % sandsynlighed for, at man slår (mindst) en sekser.

A. 6

B. 7

C. 8

D. 9

E. 10

3.6 For at undersøge effekten af en ny plantebaseret diæt har man tilfældigt udvalgt 20 personer som afprøver diæten over en periode på 6 uger. For alle personer har man foretaget vægtmålinger før og efter interventionsperioden. Hvilken af følgende statistiske metoder er velegnet til at undersøge, om diæten fører til et vægttab?

A. Lineær regression af slutvægten på startvægten, hvor man tester hypotesen om at hældningen er lig med 1.

B. Undersøg om et 95 % - konfidensinterval for slutvægten indeholder 0.

C. Sammenligning af startvægten og slutvægten opfattet som to uafhængige stikprøver.

D. Sammenligning af startvægten og slutvægten opfattet som to parrede stikprøver.

E. Lineær regression af slutvægten på startvægten, hvor man tester hypotesen om at hældningen er lig med 0.

3.7 Vi betragter et (fiktivt) datasæt (her kaldet `minkdata`) med følgende tre variable

$\texttt{wgt}_i$ kropsvægt i gram for den i-te mink

$\texttt{lgt}_i$ kropslængde i cm for den i-te mink

$\texttt{farve}_i$ pelsfarve for den i-te mink; kan være `Brun` eller `Sort`

I `R` har man fittet følgende model

```
model0 <- lm(wgt ~ lgt + farve, data = minkdata)
```

Hvordan bør man opskrive den tilhørende statistiske model (svarende til `model0`)?

A. $\texttt{wgt}_i = \alpha(\texttt{lgt}_i) + \beta(\texttt{farve}_i) + e_i$

B. $\texttt{wgt}_i = \alpha \cdot \texttt{lgt}_i + \beta(\texttt{farve}_i) + e_i$

C. $\texttt{wgt}_i = \alpha(\texttt{lgt}_i) + \beta \cdot \texttt{farve}_i + e_i$

D. $\texttt{wgt}_i = \alpha \cdot \texttt{lgt}_i + \beta \cdot \texttt{farve}_i + \delta + e_i$

E. $\texttt{wgt}_i = \alpha \cdot \texttt{lgt}_i + \beta \cdot \texttt{farve}_i + e_i$

Vi antager for alle modellerne ovenfor, at $e_1, e_2, \ldots$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$.

Besvarelse af multiple choice spørgsmål

Denne side kan med fordel afleveres sammen med den øvrige besvarelse. Anfør bogstavet hørende til dit valgte svar udfor hvert spørgsmål. Husk at du kun må skrive et bogstav til hvert spørgsmål, og at svaret ikke kan begrundes.

Opgave 3

3.1:

3.2:

3.3:

3.4:

3.5:

3.6:

3.7: