Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 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 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.

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

Opgave 1

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

173 studerende på Statistisk Dataanalyse 1 i 2020/2021 har angivet deres transporttid i minutter fra bopæl til campus. Desuden har man registreret studieretning og alder (i år). Der er kun inddraget data fra studerende som læser jordbrugsøkonomi (JE), husdyrvidenskab (HV), biologi-bioteknologi (BB) eller naturressourcer (NR).

Filerne `nov2020opg1.txt` og `nov2020opg1.xlsx` indeholder data. Der er en datalinje for hver af de 173 studerende og tre variable: `studie` som angiver studieretning, `alder` angivet i år, samt `transporttid` fra bopæl til Frederiksberg Campus angivet i minutter.

Datasættet kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data1 <- read_excel(path = "nov2020opg1.xlsx")
```

eller

```
data1 <- read.table(file = "nov2020opg1.txt", header = T)
```

De første seks linjer i datasættet ses her

```
##   studie alder transporttid
## 1     HV    31           20
## 2     NR    23           25
## 3     JE    19           14
## 4     BB    27           10
## 5     NR    23           11
## 6     HV    24           35
```

1.1 Antag at datasættet er indlæst i `R` under navnet `data1`. Opskriv den statistiske model som fittes med `R`-koden

```
model1 <- lm(log(transporttid) ~ studie, data = data1)
```

Angiv et estimat for den forventede log-transformerede transporttid for husdyrvidenskabstuderende (`HV`) og for studerende som læser biologi-bioteknologi (`BB`). Angiv desuden estimatet for residualspredningen σ.

1.2 Undersøg med et hypotesetest om middelværdien af logaritmen til transporttiden kan antages at være ens for studerende på de fire studieretninger.

1.3 Angiv et estimat og et 95 %-konfidensinterval for, hvor meget højere den forventede værdi af logaritmen til transporttiden er for studerende på husdyrvidensskab end for studerende på jordbrugsøkonomi.

Regn tilbage og angiv også et estimat samt et 95 %-konfidensinterval på den oprindelige skala (dvs. ikke log-transformeret), og forklar i ord hvordan resultatet skal fortolkes.

Hint: Som en del af din løsning kan du fx. bruge følgende `R`-kommando

```
model1ny <- lm(log(transporttid) ~ relevel(factor(studie), ref = "HV"),
               data = data1)
```

1.4 Benyt datasættet til at argumentere for, at husdyrvidenskabstuderende (`HV`) har længere transporttid til campus end studerende fra de øvrige studieretninger. Du bør underbygge din diskussion med et eller flere relevante hypotesetest.

Hint: Ved besvarelsen kan du fx. anvende `model1ny` (ovenfor) eller `model2` som kan fittes med `R`-koden nedenfor

```
data1$studie_hv <- data1$studie == "HV"
head(data1)
```

```
##   studie alder transporttid studie_hv
## 1     HV    31           20      TRUE
## 2     NR    23           25     FALSE
## 3     JE    19           14     FALSE
## 4     BB    27           10     FALSE
## 5     NR    23           11     FALSE
## 6     HV    24           35      TRUE
```

```
model2 <- lm(log(transporttid) ~ studie_hv, data = data1)
```

1.5 Man har en formodning om, at `alder` også kan have en sammenhæng med transporttiden. Opskriv den statistiske model som fittes med koden

```
model3 <- lm(log(transporttid) ~ studie + alder, data = data1)
```

Benyt `model3` til at diskutere, om `alder` har en sammenhæng med transporttiden til campus.

Opgave 2

Denne opgave vægtes med 32 % ved bedømmelsen, og svarene skal begrundes.

Fra hjemmesiden www.worldometers.info/coronavirus/ har man den 27/10 kl. 14:30 udtrukket information om det totale antal Corona tilfælde (`cases`) samt antallet af dødsfald (`deaths`) som er relateret til Corona virus. Der indgår data fra 100 forskellige lande.

Datasættet kan fx. indlæses med en af følgende `R`-kommandoer

```
data2 <- read.table(file = "nov2020opg2.txt", header = T)
```

eller

```
library(readxl)
data2 <- read_excel(path = "nov2020opg2.xlsx")
```

Hver linje i datasættet indholder oplysninger for et af de 100 lande.

2.1 Afgør hvilken af følgende to statistiske modeller som er mest velegnet til at beskrive sammenhængen mellem antal af dødsfald (`deaths`) og det totale antal Corona tilfælde (`cases`)

$$\begin{aligned} \texttt{deaths}_i &= \alpha + \beta \cdot \texttt{cases}_i + e_i && \textbf{Model A} \\ \log(\texttt{deaths}_i) &= \alpha + \beta \cdot \log(\texttt{cases}_i) + e_i && \textbf{Model B}. \end{aligned}$$

Her er e_i'erne uafhængige og normalfordelte $\sim N(0, \sigma^2)$. Du skal besvare spørgsmålet ved at udføre modelkontrol og kommentere på relevante grafer.

Opgaverne **2.2**-**2.4** nedenfor kan delvist besvares ud fra følgende `summary()` af **Model B**. Bemærk dog, at du også selv er nødt til at køre nogle analyser i `R`, for at lave en fuldstændig besvarelse af **2.2**-**2.4**.

```
## summary(modelB)

##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -5.768445  0.6472034 -8.912878 2.761516e-14
## log(cases)   1.150162  0.0549998 20.912119 6.550979e-38
```

2.2 Opskriv `R`-koden til at fitte **Model B**. Angiv desuden et estimat og et 95 %-konfidensinterval for regressionsparameteren β i **Model B**.

2.3 Benyt **Model B** til at lave et 95 %-prædiktionsinterval for antallet af døde i et nyt land med 10.000 Corona-tilfælde.

Hint: Du kan fx. benytte `predict()`-funktionen på et nyt datasæt konstrueret med `R`-kommandoen

```
newdata <- data.frame(cases = 10000)
```

2.4 Fortolkningen af **Model B** er, at *medianen* for antallet af døde (`deaths`) er givet som

$$\exp(\alpha) \cdot \texttt{cases}^{\beta}.$$

Hvis procentdelen af smittede som dør er ens i alle lande, så bør β være lig med 1. Brug data til at undersøge om dette er en rimelig antagelse.

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. Du må naturligvis gerne bruge R til opgaven.

3.1 En stikprøve viser, at den maksimale tid studerende kan koncentrere sig ved online undervisning kan beskrives ved en normalfordeling med middelværdi 39.4 minutter og spredning 16.0 minutter.

Beregn sandsynligheden for at en tilfældigt valgt studerende har en maksimal koncentrationstid på under 30 minutter ved online undervisning.

A. Ca. 22.1 %

B. Ca. 72.2 %

C. Ca. 31.0 %

D. Ca. 27.8 %

E. Ca. 30.0 %

3.2 En stikprøve viser, at den maksimale tid studerende kan koncentrere sig ved online undervisning kan beskrives ved en normalfordeling med middelværdi 39.4 minutter og spredning 16.0 minutter.

Hvor mange minutter bør en online forelæsning vare, hvis man vil sikre sig, at der højst er 10 % af de studerende, som mister koncentrationen under forelæsningen?

A. Ca. 18.9 minutter

B. Ca. 59.9 minutter

C. Ca. 55.4 minutter

D. Ca. 23.4 minutter

E. Ca. 65.7 minutter

3.3 En undersøgelse viser at 64.7 % af de studerende på Statistisk Dataanalyse i 2020/2021 foretrækker bagværk uden rosiner. Ved et socialt arrangement med 10 studerende fra Statistisk Dataanalyse 1 serveres boller med og uden rosiner. Hvad er sandsynligheden for, at der er mere end 8 (dvs. mindst 9!) af de indbudte studerende, som gerne vil have boller uden rosiner?

A. Ca. 17.2 %

B. Ca. 91.7 %

C. Ca. 1.3 %

D. Ca. 7.0 %

E. Ca. 8.3 %

3.4 Man ønsker at undersøge effekten af en behandling på en kvantitativ / numerisk responsvariabel. På de samme 10 forsøgspersoner foretages målinger af responsen både før behandling (`x`) og efter behandling (`y`). Data er indtastet i `R` som to variable/vektorer `x` og `y` i rækkefølge efter forsøgspersonernes nummer.

```
diff <- y - x
mean(x)
## [1] 4.896037
sd(x)
## [1] 1.290613
mean(y)
## [1] 5.295264
sd(y)
## [1] 2.01098
mean(diff)
## [1] 0.3992268
sd(diff)
## [1] 1.621325
```

Angiv en t-teststørrelse og en tilhørende p-værdi for et test af hypotesen om, at behandlingen påvirker værdien af responsen.

A. $T = 0.2462, P = 0.8105$

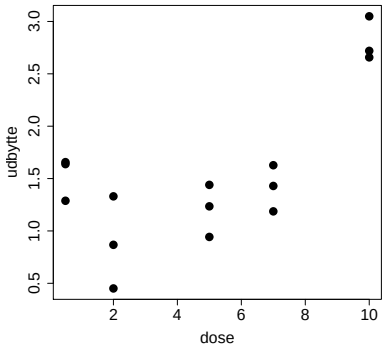
B. $T = 0.7387, P = 0.4789$

C. $T = 0.7787, P = 0.2281$

D. $T = 0.7787, P = 0.4562$

E. $T = 0.2462, P = 0.8110$

3.5 Ved et dyrkningsforsøg har man målt udbyttet på 15 forskellige marker. Der blev anvendt 5 forskellige doser af gødning på markerne, således hver dosis (0.5, 2, 5, 7 eller 10 enheder) blev afprøvet på præcis 3 marker. I datasættet er dosisgruppen angivet ved variablen `grp` og den tilhørende dosis ved variablen `dose`.




```
my_data <- data.frame(grp, dose, udbytte)
head(my_data, 8)
```

```
##   grp dose   udbytte
## 1   I  0.5 1.6567430
## 2   I  0.5 1.6378871
## 3   I  0.5 1.2879942
## 4  II  2.0 0.8673985
## 5  II  2.0 0.4508664
## 6  II  2.0 1.3301434
## 7 III  5.0 1.2347803
## 8 III  5.0 0.9426556
```

Dernæst er udført et test i `R`

```
mod1 <- lm(udbytte ~ grp, data = my_data)
mod2 <- lm(udbytte ~ dose, data = my_data)
anova(mod2, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: udbytte ~ dose
## Model 2: udbytte ~ grp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     13 3.8688
## 2     10 0.7842  3    3.0847 13.112 0.0008442 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hvad kan man konkludere på baggrund af testet?

A. At der ikke er nogen sammenhæng mellem dosis (`dose`) og `udbytte`.

B. At der ikke er en lineær sammenhæng mellem dosis (`dose`) og `udbytte`.

C. At sammenhængen mellem dosis (`dose`) og `udbytte` bør beskrives med et 2.gradspolynomium.

D. At der er en lineær sammenhæng mellem dosis (`dose`) og `udbytte`.

E. At der er en lineær sammenhæng mellem dosis (`dose`) og `udbytte`, men at hældningen er lig med 0.

3.6 173 studerende på Statistisk Dataanalyse i 2020/2021 er blevet spurgt om de foretrækker fysisk undervisning (`uv = fysisk`), online undervisning (`uv = online`) eller en blanding af de to undervisningsformer (`uv = blanding`). Samtidig har de studerende svaret på om de er positive (`inkl = Ja`) eller negative (`inkl = Nej`) over for at inkludere en person de ikke kender til at samarbejde online om øvelsesopgaverne.

Nedenfor er resultaterne af undersøgelsen opgjort ligesom der er udført et statistisk test. Ved besvarelsen må du se bort fra den advarsel (`Warning`) som R kommer med, fordi nogle af de forventede celleantal under hypotesen er mindre end 5.

```
my_table
```

```
##      uv
## inkl  online blanding fysisk
##   Ja      15       49     82
##   Nej      3       13     11
```

```
chisq.test(my_table, correct = FALSE)
```

```
## Warning in chisq.test(my_table, correct = FALSE): Chi-squared approximation
may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  my_table
## X-squared = 2.3765, df = 2, p-value = 0.3048
```

Hvad kan man konkludere ud fra testet?

A. At der er sammenhæng mellem den foretrukne undervisningsform og villigheden til at inkludere andre studerende.

B. At studerende som foretrækker fysisk undervisning er mere villige til at inkludere andre studerende.

C. At villigheden til at inkludere andre studerende afhænger af, hvilken undervisningsform man foretrækker.

D. At der ikke er sammenhæng mellem den foretrukne undervisningsform og villigheden til at inkludere andre studerende.

E. Ingenting, for data bør analyseres med en tosidet variansanalysemodel.

3.7 Ved et eksperiment ønsker man at undersøge, om et bestemt konserveringsmiddel kan forøge holdbarheden af bundter af roser.

Konserveringsmidlet kan tilsættes enten hos blomsterhandleren eller hos kunden eller begge steder. Ved forsøget indgår 24 bundter af roser inddelt i fire grupper angivet med de kategoriske variable `handler` og `kunde`, som hver kan have værdierne `tilsat` eller `ikke-tilsat`.

Der er seks bundter i hver af de fire grupper som anført i følgende tabel.

```
## # A tibble: 4 x 3
##   handler     kunde       `antal bundter`
##   <chr>       <chr>                 <int>
## 1 ikke_tilsat ikke_tilsat               6
## 2 ikke_tilsat tilsat                    6
## 3 tilsat      ikke_tilsat               6
## 4 tilsat      tilsat                    6
```

Ved eksperimentet måles den gennemsnitlige holdbarhed (`tid`) i dage for roserne i hver af de 24 bundter.

Til analyse af data er benyttet en tosidet variansanalysemodel med vekselvirkning, hvor holdbarheden (`tid`) anvendes som responsvariabel.

```
model1 <- lm(tid ~ handler * kunde, data = roser)
## summary(model1) # et udpluk af summary() for modellen
```

```
##                             Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                9.9000000  0.6174994 16.0324033 7.008949e-13
## handlertilsat              0.7333333  0.8732761  0.8397497 4.109750e-01
## kundetilsat                1.0833333  0.8732761  1.2405393 2.291323e-01
## handlertilsat:kundetilsat 2.1333333  1.2349989  1.7273970 9.950883e-02
```

Angiv et estimat for middelværdien af holdbarheden for et bundt roser, som har fået `tilsat` konserveringsmidlet hos både blomster`handler`en og hos `kunden`.

A. Ca. 9.900 dage

B. Ca. 10.983 dage

C. Ca. 10.633 dage

D. Ca. 13.850 dage

E. Ca. 3.950 dage