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.

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Reeksamen, februar 2021

Fire timers skriftlig prøve. Alle hjælpemidler er tilladt, men du må ikke kommunikere med andre under eksamen.

Der er 3 opgaver, som vægtes med henholdsvis 40 %, 40 % og 20 % 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.

COVID-19 smitte kan påvises under et aktivt sygdomsforløb vha. en PCR-test. Data til denne opgave består af opgørelser af antallet af påviste COVID-19 tilfælde i landets 98 kommuner opgjort for månederne december 2020 og januar 2021. Datasættet er hentet fra Statens Serum Instituts hjemmeside med overvågningsdata for COVID-19.

Filerne `feb2021opg1.txt` og `feb2021opg1.xlsx` indeholder data. Der er en datalinje for hver af landets 98 kommuner. Variablen `kommune_id` er en kode for de enkelte kommuner (skal ikke bruges her), og `region` angiver i hvilken af landets 5 regioner kommunen hører hjemme. Variablene `dec` og `jan` angiver det totale antal påviste COVID-19 tilfælde i december 2020 og i januar 2021. I resten af opgaven omtaler vi `dec` og `jan` som incidensen af smittede, og vi gør opmærksom på, at disse tal er opgjort med enheden *per 1000 indbyggere*.

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

```
library(readxl)
data1 <- read_excel(path = "feb2021opg1.xlsx")
```

eller

```
data1 <- read.table(file = "feb2021opg1.txt", header = T)
```

De første seks linjer i datasættet ses her

```
##        region kommune_id   dec  jan
## 1  Syddanmark        580  5.68 3.57
## 2 Nordjylland        851 11.22 4.28
## 3 Midtjylland        751 17.32 5.23
## 4  Syddanmark        492  1.34 2.68
## 5 Hovedstaden        165 27.66 9.74
## 6 Hovedstaden        201 15.60 5.27
```

I første linje betyder værdien `5.68` ud for variablen `dec` altså, at incidensen af smittede i december 2020 var 5.68 nye konstaterede COVID-19 tilfælde for hver 1000 borgere i kommunen.

1. Udfør et test for om der er sket et fald i antallet af konstaterede COVID-19 tilfælde fra december 2020 til januar 2021. Du bedes angive din R-kode i besvarelsen.

I resten af opgaven benyttes forskellige metoder til at undersøge, om ændringen (dvs. faldet) i incidensen af nye smittede kan forklares ud fra øvrige variable i datasættet.

1.2 Opskriv en lineær regressionsmodel, hvor faldet i incidensen (`fald = dec - jan`) beskrives som en lineær funktion af incidensen i december (`dec`).

Fit modellen i `R` og angiv estimater for samtlige parametre i modellen.

1.3 Benyt den lineære regressionsmodel til at diskutere, om faldet (fra december til januar) i antallet af nye smittede er større i kommuner, hvor der var mange nye smittede i december.

1.4 Bestem et estimat og et 95 % - prædiktionsinterval for faldet (fra december 2020 til januar 2021) i antallet af nye COVID-19 tilfælde for en kommune med 1000 indbyggere, hvor incidensen var 10 tilfælde per 1000 indbyggere i december 2020.

1.5 Man kunne forestille sig, at der er regionale forhold som har betydning for, hvor hurtigt restriktionerne indført omkring nytår kan aflæses i smittetallene.

Opskriv den statistiske model som fittes med koden

```
data1$fald <- data1$dec - data1$jan
model1 <- lm(fald ~ region + dec, data = data1)
```

Undersøg (fx. med udgangspunkt i `model1`), om der lader til at være regionale forskelle på, hvor stort et fald der sker i smittetallene fra december 2020 til januar 2021.

Bemærk: Der kan være flere løsninger på dette delspørgsmål, men kun en metode bedes angivet.

Opgave 2

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

Ved et markforsøg ønsker man at undersøge effekten af sprøjtning med bayleton på udbyttet af fire forskellige bygsorter (`Lami`, `Lofa`, `Salka`, `Zita`). Desuden har man målt udbyttet på nogle plots (=jordlodder), hvor der er anvendt en `Blanding` af de fire sorter. Der indgår i alt 40 plots i forsøget (8 for hver sort samt 8 for blandingen).

Filerne `feb2021opg2.txt` og `feb2021opg2.xlsx` indeholder data. Der er en datalinje for hver af de 40 plots og tre variable: `bayleton` som angiver om plottet er sprøjtet med bayleton eller ej, `variety` som angiver bygsorten, samt `udbyttet` på plottet.

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

```
library(readxl)
data2 <- read_excel(path = "feb2021opg2.xlsx")
```

eller

```
data2 <- read.table(file = "feb2021opg2.txt", header = T)
```

De første seks linjer i datasættet ses her

```
##   bayleton  variety udbytte
## 1       Ja Blanding 49.4397
## 2       Ja Blanding 50.8523
## 3       Ja Blanding 51.5586
## 4       Ja Blanding 53.6774
## 5      Nej Blanding 51.9692
## 6      Nej Blanding 52.6715
```

Ved besvarelsen af delopgave 2.1-2.3 skal du tage udgangspunkt i en ensidet variansanalysemodel, hvor `udbytte` alene antages at afhænge af bygsorten (`variety`).

2.1 Angiv et estimat for det forventede udbytte på et plot, hvor man anvender blandingssorten (`variety = Blanding`). Angiv et estimat for residualspredningen.

2.2 Undersøg med et hypotesetest om middelværdien af udbyttet kan antages at være ens for alle sorterne.

2.3 Angiv et estimat og et 95 % - konfidensinterval for forskellen i det forventede udbytte på to plots hvor man anvender sorterne `Salka` og `Zita`.

Diskuter om analysen giver anledning til at konkludere, at der er samme udbytte for sorterne `Salka` og `Zita`.

I den resterende del af opgaven arbejder vi med en model, som inkluderer alle variable i datasættet.

2.4 Antag at datasættet er indlæst i R under navnet `data2`, og at vi har fittet følgende model til data

```
modelA <- lm(udbytte ~ variety * bayleton, data = data2)
```

Argumenter for at en tosidet variansanalysemodel med vekselvirkning (`modelA`) er velegnet til at analysere sammenhængen mellem `udbytte` og de øvrige variable i datasættet.

Benyt `modelA` til at angive et estimat for det forventede udbytte for et plot med sorten `Lofa`, der ikke er blevet sprøjtet med bayleton (`bayleton = Nej`).

2.5 Udfør et hypotesetest med henblik på at undersøge, om der er vekselvirkning mellem sort og behandling/sprøjtning med bayleton. Forklar i ord, hvad man kan konkludere på baggrund af testet.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 20 % 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 Vægten af 9-årige drenge kan tilnærmelsesvis antages at være normalfordelt med middelværdi 31.0 kg og spredning 5.2 kg. Beregn sandsynligheden for, at en tilfældigt valgt dreng på 9 år vejer mellem 25 og 35 kg.

A. Ca. 77.9 %

B. Ca. 65.5 %

C. Ca. 90.3 %

D. Ca. 34.5 %

E. Ca. 22.1 %

3.2 Vægten af 9-årige drenge kan tilnærmelsesvis antages at være normalfordelt med middelværdi 31.0 kg og spredning 5.2 kg. Hvor meget skal en 9-årig dreng mindst veje for at være blandt de 5 % tungeste?

A. Ca. 37.7 kg

B. Ca. 22.4 kg

C. Ca. 41.2 kg

D. Ca. 39.6 kg

E. Ca. 41.4 kg

3.3 Det gennemsnitlige daglige antal skridt for en stikprøve af 63 børn fra 3. klasse udregnes til 10158. Stikprøvespredningen er 2736 skridt. Bestem et 95 % - konfidensinterval for det gennemsnitlige antal daglige skridt for børn i 3. klasse.

A. $(4689 - 15627)$

B. $(10145 - 10171)$

C. $(4686 - 15630)$

D. $(9469 - 10847)$

E. $(9813 - 10503)$

3.4 Den 12. januar 2021 kl. 14:00 har man optalt det totale antal indlagte med COVID-19 i Region Nordjylland og i Region Sjælland. Desuden har man registreret, hvor mange af patienterne med COVID-19, som var indlagt på en intensivafdeling.

```
##                            Nordjylland Sjælland
## Totalt                              59      163
## Heraf indlagt på intensiv           13       20
```

Angiv P-værdien samt en konklusion på baggrund af et test for, om andelen af indlagte på intensivafdelinger er den samme i Region Nordjylland og i Region Sjælland. Du skal udføre testet med kontinuitetskorrektion.

A. Der er samme andel indlagte på intensiv i de to regioner ($P = 0.1872$)

B. Der er *ikke* samme andel indlagte på intensiv i de to regioner ($P = 0.1872$)

C. Der er samme andel indlagte på intensiv i de to regioner ($P = 0.1112$)

D. Der er *ikke* samme andel indlagte på intensiv i de to regioner ($P = 0.1112$)

E. Der er samme andel indlagte på intensiv i de to regioner ($P = 0.07084$)

3.5 Ved et dyrkningsforsøg med kål ønsker man at undersøge sammenhængen mellem den tilsatte mængde kalk og udbyttet. Høstudbyttet (`udbytte`) måles på 48 jordlodder. Der indgår 8 forskellige doser af kalk i forsøget, således at der er 6 jordlodder som modtager hver dosis. Variablen `dosis` er en numerisk variabel, som angiver den anvendte dosis (målt i en passende enhed).

Hvilken af følgende R-koder kan benyttes til at teste, om der er en linæer sammenhæng mellem den tilsatte mængde kalk og udbyttet?

(I R-koden forudsættes det, at datasættet er indlæst under navnet `data3`)

A.
```
model1 <- lm(udbytte ~ factor(dosis), data = data3)
model3 <- lm(udbytte ~ 1, data = data3)
anova(model3, model1)
```

B.
```
model2 <- lm(udbytte ~ dosis, data = data3)
model3 <- lm(udbytte ~ 1, data = data3)
anova(model3, model2)
```

C.
```
model2 <- lm(udbytte ~ dosis, data = data3)
summary(model2)
```

D.
```
model1 <- lm(udbytte ~ factor(dosis), data = data3)
anova(model1)
```

E.
```
model1 <- lm(udbytte ~ factor(dosis), data = data3)
model2 <- lm(udbytte ~ dosis, data = data3)
anova(model2, model1)
```

3.6 I september 2020 blev der solgt 17197 biler i Danmark. Heraf var 4700 (svarende til 27.33 %) såkaldte *grønne* biler (dvs. enten elbiler eller hybridbiler).

Der udtrækkes tilfældigt oplysninger om 10 bilsalg fra september 2020. Beregn sandsynligheden for, at der i denne stikprøve var præcis 5 *grønne* biler (dvs. elbiler eller hybridbiler).

A. Ca. 7.8 %.

B. Ca. 10.8 %

C. Ca. 92.2 %

D. Ca. 27.3 %

E. Ca. 13.7 %

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, januar 2018

Fire timers skriftlig prøve. Alle hjælpemidler er tilladt, herunder lommeregner og computer (fx brug af R), og besvarelsen må gerne skrives med blyant. Du kan *ikke aflevere elektronisk*, heller ikke på vedlagte USB-stick.

Der er 3 opgaver med i alt 13 delspørgsmål. Alle delspørgsmål indgår med samme vægt i bedømmelsen. Husk at de fleste spørgsmål kan besvares uafhængigt af hinanden. Alle svar skal begrundes. Data til opgave 1 og opgave 2 udleveres på en USB-stick. Filnavnene fremgår af opgaveteksten. USB-sticken skal afleveres efter eksamen, men kun for at den kan genbruges.

Opgave 1

På kurset *Sandsynlighedsregning og Statistik* i 2017/18 bad underviseren de studerende om at gætte på antallet af punkter i tre figurer. Denne opgave handler om den ene figur (figur 2) og der er gæt fra 182 studerende. Data er tilgængelige på den vedlagte USB-stick som `ss2017-18.txt` og `ss2017-18.xlsx`. Der er en linie per studerende og følgende variable.

- `studie`: Det studie som den studerende er indskrevet på. De mulige værdier er `Matematik`, `MatOk` (matematik-økonomi) og `Aktuar` (aktuar/forsikringsmatematik)
- `kon`: Den studerendes køn, enten `Mand` eller `Kvinde`
- `figur2`: Den studerendes gæt på antal punkter i figuren

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

   Angiv R-kode der kan bruges til at estimere følgende to modeller, og angiv residualspredningen (Residual standard error) for begge modeller:

   - En tosidet variansanalyse *med vekselvirkning* hvor du bruger variablen `figur2` som responsvariabel og de andre variable som forklarende variable.
   - En tosidet variansanalyse *med vekselvirkning* hvor du bruger variablen `log(figur2)` som responsvariabel og de andre variable som forklarende variable. Husk til senere brug at `log` er den naturlige logaritme.

2. Udfør modelkontrol for hver af de to modeller fra spørgsmål 1. Besvarelsen skal bestå af skitser af de relevante figurer og kommentarer til figurerne, herunder argumenter for at modellen med `log(figur2)` som responsvariabel er at foretrække.

3. Undersøg med et hypotesetest om der er vekselvirkning mellem køn og studie, og forklar kortfattet hvad resultatet betyder. Du skal benytte `log(figur2)` som responsvariabel.

I de næste spørgsmål skal du benytte modellen for tosidet ANOVA *uden vekselvirkning* uanset hvad du har svaret i spørgsmål 3. Du skal benytte `log(figur2)` som responsvariabel.

4. Angiv et estimat og et 95% konfidensinterval for den forventede værdi af `log(figur2)` for kvindelige aktuarstuderende.

   Det sande antal punkter i figuren er 142. Tyder data på at kvindelige aktuarstuderede (som population) gætter for højt, for lavt, eller ingen af delene? Svaret skal begrundes, og du kan benytte at $\log(142) = 4.956$.

5. Angiv et estimat for forskellen mellem kvinder og mænd i forventet værdi af `log(figur2)`.

   Angiv derefter et estimat for den faktor som kvinders gæt er højere end mænds gæt. Er der signifikant forskel på mænd og kvinder?

6. Undersøg med *et enkelt* hypotesetest om studerende fra de tre forskellige studier (som populationer) gætter forskelligt på antallet af punkter i figuren.

Opgave 2

Data til denne opgave består af kropsmålinger fra 243 mænd. For hver mand har man blandt andet målt omkredsen ved hofte og mave, begge dele i cm. Desuden har man bestemt mændenes fedtprocent med en præcis målemetode baseret på opdriften ved undervandsvejning. Man er interesseret i at kunne prædiktere fedtprocenten ved hjælp af hofte- og/eller maveomkreds.

Data er tilgængelige i filerne `johnson-fatpct.txt` og `johnson-fatpct.xlsx` på den vedlagte USB-stick. Der er en linie per person og følgende variable:

- `bodyfat`: Fedtprocent
- `hip`: Omkreds ved hofte, målt i cm
- `abdomen`: Omkreds ved mave, målt i cm

Du skal i hele opgaven bruge variablen `bodyfat` som responsvariabel.

1. Lav en figur der illustrerer sammenhængen mellem maveomkreds og fedtprocent. Der skal være en skitse af figuren i besvarelsen.

   Angiv på baggrund af figuren en statistisk model der gør det muligt at estimere sammenhængen.

2. Angiv estimater for samtlige parametre i modellen fra spørgsmål 1.

   Betragt to mænd der har maveomkreds på henholdsvis 100 cm og 110 cm. Bestem et estimat for forskellen i forventet fedtprocent mellem de to mænd.

3. Fit den lineære regressionsmodel hvor du bruger hofteomkredsen som den eneste forklarende variabel, og angiv estimatet for regressionskoefficienten hørende til hofteomkreds.

   Fit derefter den multiple regressionsmodel hvor du inddrager både maveomkreds og hofteomkreds som forklarende variable, og angiv estimatet for regressionskoefficienten hørende til hofteomkreds.

   Forklar kortfattet hvad forskellen på de to angivne estimater kan skyldes.

4. I dette spørgsmål skal du bruge den multiple lineære regression fra spørgsmål 3. Bestem et 95% konfidensinterval og et 95% prædiktionsinterval for en mand med maveomkreds på 85 cm og hofteomkreds på 98 cm.

   Er det usædvanligt for en mand med maveomkreds på 85 cm og hofteomkreds på 98 cm at have en fedtprocent på 17? Svaret skal begrundes.

Opgave 3

Forskere i New England har udvalgt 73 sjældne plantearter tilfældigt og vurderet deres udbredelse med fem års mellemrum, nemlig i 2012 og 2017. Den samme metode og den samme skala er brugt begge år, og alle arter er vurderet begge år.

Forskellen i udbredelse mellem 2012 og 2017 er beregnet for hver af de 73 arter som

$$\text{forskel} = \text{udbredelse i 2017} - \text{udbredelse i 2012}$$

Forskellen er indlæst i R som variablen `forskel` og du kan benytte følgende værdier:

```
> mean(forskel)
[1] 3.353576
> sd(forskel)
[1] 8.303946
```

1. Forklar kortfattet hvorfor data fra de to år skal opfattes som to parrede stikprøver snarede end som to uparrede (uafhængige) stikprøver.

   Angiv et estimat og et 95% konfidensinterval for den forventede forskel i udbredelse mellem 2012 og 2017.

Man vil gerne undersøge om fredning er et effektivt redskab til at forbedre sjældne planters udbredelse. For hver af de 73 arter har man derfor koblet deres fredningsstatus med ændringen i udbredelse fra 2012 til 2017. Man har valgt ikke at bruge de specifikke værdier i variablen `forskel`, men udelukkende fortegnet, dvs. om artens udbredelse er vokset eller aftaget.

Data fremgår af tabellen nedenfor, og i resten af opgaven skal du kun bruge tallene fra tabellen.

	Ikke fredet	Fredet
Udbredelse aftaget	18	8
Udbredelse vokset	15	32

2. Undersøg med et hypotesetest om der er sammenhæng mellem fredningsstatus og ændring i udbredelse.

3. Angiv et estimat for den betingede sandsynlighed for at en plantearts udbredelse vokser givet at den fredet, samt den betingede sandsynlighed for at en plantearts udbredelse vokser givet at den ikke er fredet.

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, januar 2019

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

Der er 3 opgaver, som vægtes med henholdsvis 35%, 35% og 30% i bedømmelse. Indenfor hver opgave indgår alle spørgsmål med samme vægt. Husk at de fleste spørgsmål kan besvares uafhængigt af hinanden.

Data til opgave 1 og opgave 2 udleveres på en USB-stick. Filnavnene fremgår af opgaveteksten. Denne USB-stick 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-stick som udleveres til formålet af eksamensvagterne. Bemærk at kun pdf-format accepteres og at du skal benytte en anden USB-stick 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 35% ved bedømmelsen, og svarene skal begrundes.

På kurset *Sandsynlighedsregning og Statistik* i 2018/19 løste 85 studerende en sudoko på tid. Heraf løste 80 sudokoen korrekt, og data til denne opgaver stammer fra disse 80 studerende. Data er tilgængelige på den vedlagte USB-stick i filerne `ss_gr.xlsx` og `ss_gr.txt`. Der er en linje per studerende og følgende variable.

- `Studie`: Angiver studiet som den studerende er indskrevet på. De mulige værdier er `Matematik`, `MatOk` (matematik-økonomi) og `Aktuar` (aktuar/forsikringsmatematik)
- `SidsteSudoko`: Angiver hvornår den studerende sidst har løst en sudoko. De mulige værdier er `ForNylig` eller `LaengeSiden`
- `Tid`: Angiver antallet af sekunder som den studerende brugte til at løse sudokoen

1.1 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 `Tid` som responsvariabel og de andre variable som forklarende variable.

Angiv desuden residualspredningen i modellen.

1.2 Undersøg med et hypotesetest om der er vekselvirkning mellem studieretning og hvornår man sidst har løst sudokoer, og forklar kortfattet hvad resultatet betyder.

I resten af opgaven skal du benytte den tosidede variansanalysemodel *uden vekselvirkning*, hvor du bruger variablen `Tid` som responsvariabel og de andre variable som forklarende variable.

1.3 Angiv et estimat og et 95% konfidensinterval for forskellen i forventet løsningstid mellem en studerende som sidst har løst sudokoer for længe siden og en studerende som har løst sudokoer for nylig. Er forskellen signifikant?

1.4 Hvilken studieretning har det mindste estimat for den forventede løsningstid?

Angiv et estimat og et 95% konfidensinterval for forskellen i forventet løsningstid for en matematikstuderende og en matematik-økonomistuderende.

1.5 Undersøg med et samlet hypotesetest om den forventede løsningstid er ens for de tre studieretninger. Du skal stadig tage højde for hvornår de studerende sidst har løst sudokoer.

Opgave 2

Denne opgave vægtes med 35% ved bedømmelsen, og svarene skal begrundes.

Data til denne opgave består af information om 100 solgte huse i Gainesville, Florida i 2006. Data er tilgængelige i filerne `florida.xlsx` og `florida.txt`. Der er 100 datalinjer (en linje per hus) og følgende tre variable

- `Size`: Størrelsen af huset, angivet i square feet (1 square feet svarer til 0.0929 kvadratmeter)
- `Price`: Salgsprisen for huset, angivet i dollars
- `Baths`: Antal badeværelser i huset

I de første spørgsmål skal du kun benytte variablene `Size` og `Price`. Begge variable er kvantitative, så det er naturligt at benytte lineær regression.

Betragt følgende fire modelfit, hvor det er antages at data er indlæst i R-datasættet `florida`:

```
linreg1 <- lm(Price ~ Size, data=florida)
linreg2 <- lm(log(Price) ~ log(Size), data=florida)
linreg3 <- lm(Size ~ Price, data=florida)
linreg4 <- lm(log(Size) ~ log(Price), data=florida)
```

2.1 Forklar hvorfor `linreg2` er den mest relevante og velegnede model til at beskrive data. Du skal både argumentere for hvad der skal bruges som responsvariabel hhv. forklarende variabel, og lave modelkontrol for at afgøre om variablene bør transformeres eller ej.

I de følgende spørgsmål skal du benytte `linreg2`.

2.2 Gør rede for antagelserne i den statistiske model svarende til `linreg2`.

Angiv estimatet og et 95% konfidensinterval for hældningsparameteren i modellen.

2.3 Betragt to huse hvor det lille er på 1000 square feet, mens det store er på 2000 square feet. Bestem et estimat for forskellen mellem den forventede log-transformede salgspris for det store og det lille hus.

Bestem et estimat for den faktor som det store hus er dyrere i forhold til det lille hus.

2.4 Et hus på 3000 square feet i samme område blev også solgt i 2006. Dette hus er ikke med i datasættet. Bestem en prædiktion for salgsprisen for huset.

Huset blev solgt for 215000 dollars. Er dette en usædvanlig pris i forhold til de 100 huse i datasættet?

Man kan også inkludere antal badeværelser i modellen og fx bruge den multiple regressionsmodel der fittes med følgende kommando:

```
multipel <- lm(log(Price) ~ log(Size) + Baths, data=florida)
```

2.5 Opskriv den ligning der angiver sammenhængen mellem `log(Price)`, `log(Size)` og `Baths` svarende til modelfittet `multipel`.

Undersøg med et hypotesetest om antallet af badeværelser har signifikant betydning for salgsprisen, når der tages hensyn til husets størrelse.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 30% 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 Danske Spil laver en skrabekalender hver jul. Det er angivet på kalenderen at der er gevinst på hver tredje kalender, således at sandsynligheden for at vinde på en tilfældig kalender er 1/3.

En familie køber fem kalendere. Hvor stor er sandsynligheden for at der er gevinst på præcis tre af de fem kalendere?

A. 0.333
B. 0.165
C. 0.600
D. 0.211
E. 0.954

3.2 Danske Spil laver en skrabekalender hver jul. Det er angivet på kalenderen at der er gevinst på hver tredje kalender, således at sandsynligheden for at vinde på en tilfældig kalender er 1/3.

Hvor mange kalendere skal man købe for at sansynligheden for at der er gevinst på en eller flere kalendere, er mindst 95%.

A. Fire eller flere
B. Fem eller flere
C. Seks eller flere kalendere
D. Otte eller flere kalendere
E. Ti eller flere

3.3 Man kan blive optaget i foreningen Mensa hvis man scorer højt nok i deres intelligenstest. Ifølge foreningens hjemmeside er testen skaleret således at resultatet for hele befolkningen er normalfordelt med middelværdi 100 og spredning 15, og man optager personer der scorer over 130 i testen.

Hvor stor en andel af befolkningen kan optages i Mensa (hvis oplysningerne om befolkningen er korrekte)?

A. Cirka 1%
B. Cirka 2%
C. Cirka 3%
D. Cirka 4%
E. Cirka 5%

3.4 Hvert år deltager cirka 5000 personer i undersøgelsen *Danskernes rygevaner*. Resultaterne for 2017 og 2018 ses nedenfor.

	Ryger (dagligt eller lejlighedsvis)	Ryger ikke
2017	1106	4018
2018	1158	3859

Man har udført et test for homogenitet i tabellen (uden at justere for andre variable) for at undersøge om andelen af rygere i befolkningen har ændret sig fra 2017 til 2018.

Hvad er p-værdien og konklusionen? Bemærk at p-værdien er beregnet uden kontinuitetskorrektion, dvs. med optionen `correct=FALSE`.

A. $p = 0.07$, og stigningen fra 2017 til 2018 er ikke signifikant

B. $p = 0.14$, og stigningen fra 2017 til 2018 er ikke signifikant

C. $p = 0.07$, og der er evidens for at andelen af rygere er steget fra 2017 til 2018

D. $p = 0.14$, og der er evidens for at andelen af rygere er steget fra 2017 til 2018

E. $p = 0.035$, og der er evidens for at andelen af rygere er steget fra 2017 til 2018

F. $p = 0.035$, og stigningen fra 2017 til 2018 er ikke signifikant

3.5 I et forskningsprojekt har man undersøgt 80 hunde for slidgigt, nemlig 40 hunde af to forskellige racer (A og B). Hypotesen er at hunde fra de to racer har lige stor tilbøjelighed til at udvikle slidgigt, altså $H_0 : p_1 = p_2$ hvor p_1 hhv. p_2 angiver sandsynlighederne for at en tilfældig hund fra race A hhv. B har slidgigt.

Hvad er en type I fejl i denne situation?

A. Vi konkluderer at der *ikke* er forskel mellem racerne selvom der i virkeligheden er forskel

B. Vi drager den forkerte konklusion vedrørende sammenhængen mellem race og forekomst af slidgigt

C. Vi konkluderer at der *er* forskel på racerne selvom der i virkeligheden ikke er forskel

D. Vi konkluderer at der *er* forskel på racerne når dette også er sandt i virkeligheden

E. Vi konkluderer at der *ikke* er forskel mellem racerne når dette også er sandt i virkeligheden

3.6 I en undersøgelse om nakkeskader blev hovedomkredsen målt for 30 unge mænd der spiller amerikansk fodbold på college. Gennemsnittet viste sig at være 57.38 cm og spredningen var 3.16 cm.

Bestem et 95% konfidensinterval for den gennemsnitlige hovedomkreds blandt collegespillere i amerikansk fodbold.

A. $(56.40, 58.36)$

B. $(50.92, 63.84)$

C. $(56.20, 58.56)$

D. $(57.16, 57.60)$

E. $(54.22, 60.54)$

3.7 I et planteeksperiment har man målt længden af den længste rod på forskellige tidspunker i vækstprocessen. Hver plante fik kun målt rødder en gang, og tidspunktet varierede fra 1 uger efter spiring til 10 uger efter spiring. I en lineær regression med længden af længste rod (i cm) som responsvariabel og *antal uger* efter spiring fik man et estimat for hældningen på $\hat{\beta} = 2.95$ og residualspredning på $s = 5.13$.

Hvilket estimat for hældning og residualspredning ville man have fået hvis man i stedet havde benyttet *antal dage* efter spiring som forklarende variabel?

A. Hældningsestimat 2.95, residualspredning 0.733
B. Hældningsestimat 2.95, residualspredning 5.13
C. Hældningsestimat 0.421, residualspredning 0.733
D. Hældningsestimat 0.421, residualspredning 5.13
E. Hældningsestimat 0.421, residualspredning 1.94
F. Hældningsestimat 2.95, residualspredning 1.94

3.8 Inden en standardoperation har man bedt 60 partienter angive hvordan de opfatter deres egen smertetærskel (lav/mellem/høj). Samtlige 60 patienter var mænd. Efterfølgende har man registreret hvor meget smertestillende medicin de 60 mænd modtog den første uge efter operationen. Formålet er et undersøge om mængden af medicin afhænger af den selvopfattede smertetærskel.

Medicinmængden skal benyttes som responsvariabel, men hvilken type analyse vil du bruge til analysen?

A. Tosidet variansanalyse med smertetærskel og køn samt deres vekselvirkning som forklarende variable
B. Sammenligning at to parrede stikprøver
C. Tosidet variansanalyse med smertetærskel og køn som forklarende variable, men uden vekselvirkning
D. Sammenligning at to uafhængige stikprøver
E. Ensidet variansanalyse med smertetærskel som forklarende variabel
F. Lineær regression med smertetærskel som forklarende variabel

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:

3.8:

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:

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 2017

Fire timers skriftlig prøve. Alle hjælpemidler er tilladt, herunder lommeregner og computer (fx brug af R), og besvarelsen må gerne skrives med blyant. Du kan *ikke aflevere elektronisk*, heller ikke på vedlagte USB-stick.

Der er 4 opgaver med i alt 13 delspørgsmål. Alle delspørgsmål indgår med samme vægt i bedømmelsen. Husk at de fleste spørgsmål kan besvares uafhængigt af hinanden.

Data til opgave 1 og opgave 3 udleveres på en USB-stick. Filnavnene fremgår af opgaveteksten. USB-sticken skal afleveres efter eksamen, men udelukkende for at den kan genbruges. Den kan ikke indgå som en del af besvarelsen. Der er R-kode og R-output til opgave 2.

Opgave 1

Det bliver med jævne mellemrum diskuteret i USA om den liberale våbenlovgivning har negative konsekvenser, fx om den lette adgang til våben fører til flere selvmord. For at undersøge dette har man indsamlet oplysninger fra hver af de 50 amerikanske stater. Data er tilgængelige på den vedlagte USB-stick som `guns.txt` og `guns.xlsx`. Der er en linie per stat og følgende variable:

- `State`: Navnet på staten
- `GunOwnerPct`: Udbredelsen af skydevåben, målt som procentdelen af husholdninger der ejer mindst et skydevåben
- `SuicideRate`: Antal selvmord per 100000 indbyggere
- `Law`: Har værdien `Yes` eller `No` afhængig af om staten har love (mindst en) der lægger restriktioner på våbensalg udover det der gælder i hele USA

I de første to spørgsmål skal du kun bruge variablene `GunOwnerPct` og `SuicideRate`.

1. Lav en figur der illustrerer sammenhængen mellem udbredelsen af skydevåben og selvmordsraten. Der skal være en skitse af figuren i besvarelsen.

   Angiv på baggrund af figuren en statistisk model der gør det muligt at estimere sammenhængen, og angiv estimater for samtlige parametre i modellen.

2. Brug modellen til at undersøge om der er sammenhæng mellem våbenudbredelse og selvmordsrate.

   Bestem et estimat og et 95% konfidensinterval for den forventede selvmordsrate for en fiktiv stat med en våbenudbredelse på 45%.

Man må formode at love der sætter begrænsninger for våbensalget, begrænser udbredelsen af våben. Det er derfor fornuftigt at inddrage variablen `Law` i analysen, og de sidste spørgsmål skal derfor besvares på baggrund af følgende modelfit, hvor `guns` er navnet på det indlæste datasæt:

```
lm(SuicideRate ~ GunOwnerPct + Law, data=guns)
```

Gennemsnittet af `GunOwnerPct` er 27.15% for stater der har mindst en restriktiv lov (`Law=Yes`) og 45.19% for stater der ikke har en restriktiv love (`Law=No`).

3. Bestem estimater for den forventede selvmordsrate for to fiktive stater: en stat med mindst en restriktiv lov og en våbenudbredelse på 27.15%; og en stat uden restriktive love og en våbenudbredelse på 45.19%.

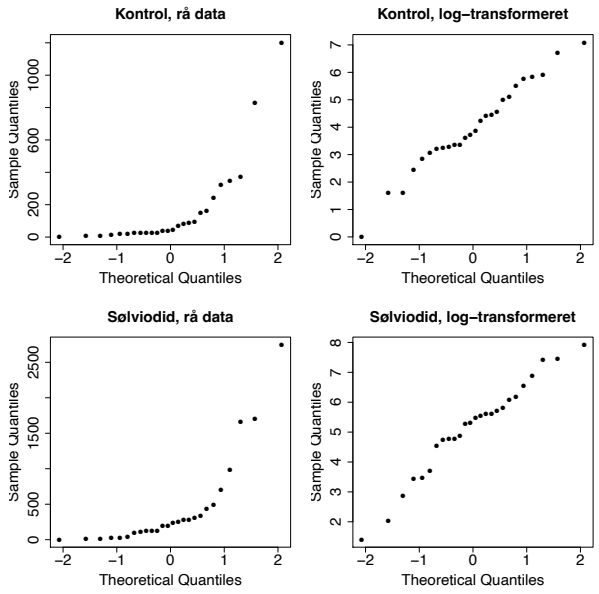
4. Tyder data på at der er en effekt af restriktive love på selvmordsraten, når man tager højde for våbenudbredelsen?

Opgave 2

Data til denne opgave består af regnmængder fra 52 skyer. Halvdelen af skyerne blev behandlet med sølviodid i eksperimentet, mens den anden halvdel var kontrolskyer som ikke blev behandlet. Regnmængden er målt i *acre-feet*, som er den mængde vand der kræves for at dække et areal på 1 acre (4047 m^2) i en højde på 1 fod (0.305 m).

Data er indlæst i datasættet `regnData` i R med to variable: `behandling` der enten er `kontrol` eller `solvIodid`, og `regn` der angiver den observerede regnmængde.

Figuren nedenfor viser fire QQ-plots: De øverste plots er for kontrolskyerne, de nederste er for skyerne behandlet med sølviodid. Til venstre er QQ-plots lavet for de rå (ikke-transformerede) data, til højre for log-transformerede værdier.



Første spørgsmål vedrører kun de 26 kontrolskyer.

1. Forklar kortfattet hvorfor det er mere fornuftigt at analysere de log-transformerede værdier fra kontrolskyerne end de ikke-transformerede værdier som en normalfordelt stikprøve.

   Bestem et 95% konfidensinterval for middelværdien af log-transformeret regnmængde for kontrolskyer. Du kan benytte at gennemsnittet for de 26 værdier af `log(regn)` fra kontrolskyer er 3.990, og at stikprøvespredningen er $s = 1.642$.

Vi skal nu interessere os for effekten af sølviodidbehandlingen.

Der er R-kode og R-output sidst i opgaven som kan benyttes ved besvarelsen. Dele af outputtet er erstattet af `XXXX`. Det er med vilje og værdierne kan beregnes udfra den givne information og en smule R-kode.

2. Angiv en statistisk model der kan bruges til at sammenligne regnmængden for kontrolskyer og skyer behandlet med sølviodid. Udfør et hypotesetest der belyser om sølviodidbehandlingen har en effekt på regnmængden.

3. Angiv et estimat og et 95% konfidensinterval for forskellen mellem de forventede værdier af logaritmen til regnmængden for de to grupper af skyer.

   Angiv derefter et estimat og et 95% konfidensinterval for den procentvise forøgelse af regnmængden når skyer tilføres sølviodid.

Uddrag af R-kørsel. Dele af outputtet er erstattet af `XXXX`. Det er med vilje og værdierne kan beregnes udfra den givne information og en smule R-kode.

```
> model <- lm(log(regn) ~ behandling, data=regnData)
> summary(model)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.9904     0.3179    XXXX    XXXX
behandlingsolvIodid   1.1438     0.4495    XXXX    XXXX
---
Residual standard error: 1.621 on 50 degrees of freedom

> confint(model)
                        2.5 %   97.5 %
(Intercept)         3.351948 4.628864
behandlingsolvIodid 0.240865 2.046697
```

Opgave 3

Data til denne opgave stammer fra et eksperiment, hvor rodlængden blev målt for 40 planter 10 dage efter såning. Planterne blev dyrket enkeltvis i potter, som stod to forskellige steder (sted A og B, 20 planter per sted). Jorden i potterne var præpareret med gødning, men på fire forskellige måder (se nedenfor). Data er tilgængelige på den vedlagte USB-stick som `rodlaengde.txt` og `rodlaengde.xlsx`. Der er en linie per plante og følgende variable:

- `dosis`: Behandlingsvariabel. Har værdierne `lav` svarende til lav dosis, `mellem1` svarende til mellemstor dosis givet som en enkelt behandling, `mellem2` svarende til mellemstor dosis givet som to behandlinger, `hoej` svarende til høj dosis.
- `sted`: Stedet hvor planten er dyrket. Har værdierne `A` og `B`.
- `lgd`: Den målte rodlængde i cm.

1. Fit modellen for tosidet variansanalyse (tosidet ANOVA) *med vekselvirkning* med rodlængden som responsvariabel, og udfør modelkontrol. Besvarelsen skal bestå af en linie R-kode med `lm`-kommandoen, skitser af de relevante figurer og kommentarer til figurerne.

2. Undersøg om effekten af dosis på rodlængden er forskellig på sted A og sted B.

I de næste spørgsmål skal du benytte modellen for tosidet ANOVA *uden vekselvirkning* med rodlængden som responsvariabel—uanset hvad du har svaret i spørgsmål 1 og 2.

3. Bestem estimatet for forventet rodlængde for en plante fra sted B, som har fået mellemstor dosis gødning givet som to behandlinger (`mellem2`).

   For hvilken af de otte kombinationer af dosis og sted er den forventede rodlængde størst? Svaret skal naturligvis begrundes.

4. Angiv estimat og 95% konfidensinterval for *forskellen* i forventet rodlængde mellem planter der har fået høj dosis gødning og planter der har fået lav dosis gødning.

   Undersøg med et hypotesetest om der er forskel mellem forventet rodlængde på sted A og sted B.

Opgave 4

En ingrediens i kosmetikprodukter mistænkes for at øge risikoen for en ellers sjælden hudsygdom. For at undersøge sammenhængen, har man udvalgt 223 kvinder med sygdommen (cases) og desuden 446 kvinder uden sygdommen (controls). Ved udvælgelsen ved man ikke om kvinderne har været eksponeret for ingrediensen, men det kan afgøres med en blodprøve.

Resultatet fremgår af tabellen nedenfor.

	Eksponeret	Ikke eksponeret	Total
Har sygdommen	54	169	223
Har ikke sygdommen	76	370	446

1. Undersøg med et hypotesetest om sandsynligheden for at en kvinde har været eksponeret, afhænger af om hun har sygdommen eller ej. Hvad er konklusionen i forhold til sammenhængen mellem ingrediensen og sygdommen?

Studiet er konstrueret således at en tredjedel af de undersøgte kvinder har sygdommen, men forekomsten af sygdommen i befolkningen er kun 0.5%. Hvis man udtager en kvinde tilfældigt fra befolkningen, er sandsynligheden altså 0.5% for at vælge en der har sygdommen. Antag desuden at sandsynlighederne for at kvinder med og uden sygdommen har været eksponeret til ingrediensen, er 0.242 og 0.170. Dette svarer til tabellen ovenfor.

2. Bestem sandsynligheden for at en tilfældig kvinde har været eksponeret. Bestem derefter den betingede sandsynlighed for at en kvinde der har været eksponeret, har sygdommen.

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 2018

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

Der er 3 opgaver, som vægtes med henholdsvis 42%, 28% og 30% i bedømmelse. Indenfor hver opgave indgår alle spørgsmål med samme vægt. Husk at de fleste spørgsmål kan besvares uafhængigt af hinanden.

Data til opgave 1 og opgave 2 udleveres på en USB-stick. Filnavnene fremgår af opgaveteksten. Denne USB-stick 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-ouputtet 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-stick som udleveres til formålet af eksamensvagterne. Bemærk at kun pdf-format accepteres og at du skal benytte en anden USB-stick 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 42% ved bedømmelsen, og svarene skal begrundes. Data er stillet til rådighed af Lars Båstrup-Spohr fra Biologisk Institut på KU.

Data består af fosforkoncentrationen i 1242 søer i Danmark og Sydsverige. Søerne ligger i fem forskellige områder (Jylland, Østdanmark, Skåne, Småland og Blekinge), og man er interesseret i at sammenligne fosforkoncentrationen i de fem områder.

Filerne `soeer.xlsx` og `soeer.txt` indeholder data. Der er en datalinie for hver sø og to variable: `Sted`, som angiver hvilket af de fem områder søen ligger i, og `Fosfor`, som er fosforkoncentration i søen, målt i μg per liter.

1.1 Forklar kortfattet hvorfor datasættet lægger op til en ensidet ANOVA (ensidet variansanalyse).

Angiv desuden en `lm`-kommando der kan bruges til at fitte en ensidet ANOVA med fosforkoncentration som respons og område som forklarende variabel, samt `lm`-kommando der kan bruges til at fitte en ensidet ANOVA med log-transformeret fosforkoncentration som respons og område som forklarende variabel

1.2 Lav modelkontrol for hver af de to modeller fra spørgsmål 1. Kommentér herunder relevante grafer, og forklar hvorfor modellen med log-transformeret fosforkoncentration som respons er at foretrække. Du skal enten vedlægge de relevante grafer elektronisk eller lave skitser af dem i hånden.

I det følgende skal du benytte modellen med log-transformeret fosforkoncentrationen som respons.

1.3 Undersøg med et enkelt hypotesetest om den forventede log-transformerede fosforkoncentrationen kan antages at være den samme for alle fem områder.

1.4 Bestem estimater for den forventede log-transformerede fosforkoncentrationen i Blekinge og i Skåne. Bestem også de tilhørende 95% konfidensintervaller.

1.5 Bestem et estimat og et 95% konfidensinterval for forskellen i forventet log-transformeret fosforkoncentrationen mellem Blekinge og Skåne.

Bestem derefter et estimat og et 95% konfidensinterval for den faktor som forforkoncentrationen er højere i Skåne i forhold til Blekinge.

1.6 Undersøg med et enkelt hypotesetest om den forventede log-transformerede fosforkoncentrationen kan antages at være ens i de tre svenske områder (Blekinge, Skåne, Småland). For at få fuldt pointtal skal du stadig bruge modellen for alle 1242 søer.

Opgave 2

Denne opgave vægtes med 28% ved bedømmelsen, og svarene skal begrundes.

Data til denne opgave stammer fra valgkampe i forbindelse med 15 amerikanske borgmestervalg, hvor den siddende borgmester stillede op. Data er tilgængelige i filerne `elections.xlsx` og `elections.txt`. Der er 15 datalinier og følgende tre variable:

- `approval`: Den siddende borgmesters popularitet ved valgkampens begyndelse, angivet som den procentdel der i en meningsmåling angav at ville stemme på kandidaten
- `expenditures`: Udgifter til valgkampagnen for den siddende borgmester, angivet i 1000 dollars
- `performance`: Resultatet af valget for den siddende borgmester, angivet som den procentdel af stemmerne der gik til ham/hende

I det første spørgsmål skal du kun benytte variablene `approval` og `expenditures`.

2.1 Opskriv en lineær regressionsmodel, der kan benyttes til at undersøge om populariteten ved valgkampens begyndelse har betydning for hvor mange penge der benyttes til valgkampagnen.

Er der i evidens i data for at der er en sammenhæng mellem borgmesterens popularitet og hans/hendes udgifter til valgkampagnen?

Man er først og fremmest interesseret i at undersøge hvordan den siddende borgmesters udgifter til valgkampagnen påvirker hans/hendes chance for at genvinde valget. Vi vil benytte den multiple regressionsmodel der fittes med følgende kommando (hvor `elections` er navnet på datasættet med de tre variable):

```
lm(performance ~ approval + expenditures, data=elections)
```

2.2 Opskriv den estimerede sammenhæng mellem de tre variable svarende til den multiple regressionsmodel. Angiv også estimatet for residualspredningen.

2.3 Forklar kortfattet hvordan estimatet hørende til variablen `expenditures` skal fortolkes.

Undersøg med et hypotesetest om det hjælper på valgresultatet at bruge flere penge på valgkampagnen.

2.4 I en ny valgkamp har en siddende borgmester en popularitet på 40% ved valgkampens start og vælger at bruge 110000 dollars på sin kampagne. Bestem et interval som med 95% sandsynlighed vil indeholde valgresultatet for den pågældende borgmester.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 30% 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 SAT er en adgangstest til højere uddannelser i USA, og den blev taget af 1.7 millioner high school-elever i USA i 2017. Resultatet af testen kaldes for SAT-scoren. I 2017 var SAT-scoren normalfordelt med middelværdi 1060 og spredning 195.

Bestem sandsynligheden for at en tilfældig high school-elev har en SAT-score der er mindre end 880.

A. 0.741
B. 0.498
C. 0.259
D. 0.822
E. 0.178

3.2 SAT er en adgangstest til højere uddannelser i USA, og den blev taget af 1.7 millioner high school-elever i USA i 2017. Resultatet af testen kaldes for SAT-scoren. I 2017 var SAT-scoren normalfordelt med middelværdi 1060 og spredning 195.

Et college optager kun studerende der har en SAT-score blandt de 25% højeste. Hvor stor en SAT-score skal man have for at komme i betragtning?

A. 1192
B. 1442
C. 928
D. 1201
E. 1284

3.3 I en amerikansk spørgeskemaundersøgelse blev 100 personer spurgt, om de var rygere eller ikke-rygere, og om de syntes at skatten på cigaretter skulle hæves. Fordelingen af personer er vist i tabellen nedenfor.

	Ja til skattestigning	Nej til skattestigning
Ikke-ryger	44	32
Ryger	5	19

Man har udført et test for uafhængighed i tabellen.

Hvad er p-værdien og konklusionen? Bemærk at p-værdien er beregnet uden kontinuitetskorrektion, dvs. med optionen `correct=FALSE`.

A. $p = 0.013$, og ikke-rygere er mere tilbøjelige end rygere til at være for en skattestigning
B. $p = 0.013$, og der er ikke tegn på sammenhæng mellem rygevaner og holdning til skattestigning
C. $p = 0.40$, og ikke-rygere er mere tilbøjelige end rygere til at være for en skattestigning
D. $p = 0.40$, og der er ikke tegn på sammenhæng mellem rygevaner og holdning til skattestigning
E. $p = 0.0015$, og der er ikke tegn på sammenhæng mellem rygevaner og holdning til skattestigning
F. $p = 0.0015$, og ikke-rygere er mere tilbøjelige end rygere til at være for en skattestigning

3.4 10% af den danske befolkning er venstrehåndede. Udtag en stikprøve på 25 personer. Hvad er sandsynligheden for at fire eller færre af de 25 personer er venstrehåndede.

A. 0.138
B. 0.098
C. 0.764
D. 0.902
E. 0.862
F. 0.236

3.5 En ny type medicin mod infektion kan enten gives som tablet eller som dråber, og i to forskellige doser. I et studie har man fordelt 88 patienter tilfældigt i fire grupper svarende til de fire kombinationer af administration (tablet/dråber) og dosis (høj/lav). Patienterne har fået medicinen i 6 dage, og ændringen i infektionstal fra start til slut er registreret.

Man vil benytte ændringen i infektionstal som responsvariabel, og man vil gerne estimere effekten af den høje dosis i forhold til den lave samt estimere effekten af tabletformen i forhold til dråbeformen. Fra tidligere studier ved man at effekten af dosis er den samme uanset hvordan medicinen administreres, og man er ikke interesseret i at undersøge dette nærmere.

Hvilken type model bør benyttes til analysen?

A. En ensidet ANOVA med dosis som forklarende variabel
B. En tosidet ANOVA med metode og dosis samt deres vekselvirkning som forklarende variable
C. En lineær regression med dosis som forklarende variabel
D. En ensidet ANOVA med metode som forklarende variabel
E. En tosidet ANOVA med metode og dosis som forklarende variable, men uden vekselvirkning

3.6 En politisk meningsmåling gennemføres på 1000 personer. Man tæller specielt hvor mange personer der siger at de vil stemme på Alternativet; dette antal kaldes Y. Så er Y binomialfordelt, $Y \sim \text{bin}(1000, p)$, hvor p er andelen i befolkningen der vil stemme på Alternativet.

Ved folketingsvalget i 2015 fik Alternativet 4.4% af stemmerne, og man tester derfor hypotesen $H_0 : p = 0.044$ svarende til at tilslutningen til partiet er uændret.

Hvad er en type II fejl i denne situation?

A. Vi konkluderer at tilslutningen til Alternativet er uændret selvom den i virkeligheden har ændret sig.

B. Vi konkluderer at tilslutningen til Alternativet har ændret sig, når dette også er sandt i virkeligheden.

C. Vi konkluderer at tilslutningen til Alternativet er uændret, når dette også er sandt i virkeligheden.

D. Vi konkluderer at tilslutningen til Alternativet har ændret sig selvom den i virkeligheden er uændret.

3.7 Man har registreret fødselsvægten for en stikprøve på 32 nyfødte børn. Gennemsnittet viste sig at være 3487 g, og spredningen var 443 g.

Bestem et 95% konfindensinterval for den gennemsnitlige fødselsvægt.

A. $(3464\,,\,3510)$

B. $(3327\,,\,3647)$

C. $(2583\,,\,4391)$

D. $(2736\,,\,4238)$

E. $(3459\,,\,3515)$

F. $(3354\,,\,3620)$

3.8 En stikprøve bestående af 100 personer fra den voksne danske befolkning har taget en matematiktest. Resultaterne kan antages at være uafhængige og normalfordelte med middelværdi μ og spredning σ, som begge er ukendte parametre. Udfra resultaterne har man beregnet et 95% konfidensinterval for μ til $(71.5\,,\,75.3)$.

Hvad kan vi konkludere?

A. 95 personer i stikprøven havde et resultat mellem 71.5 og 75.3.

B. For 95% af befolkningen ligger reultatet mellem 71.5 og 75.3.

C. Vi er 95% sikre på at gennemsnittet for befolkningen ligger mellem 71.5 og 75.3.

D. Hvis vi valgte 100 andre personer, ville de alle have et resultat mellem 71.5 og 75.3.

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:

3.8:

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 2019

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

Der er 3 opgaver, som vægtes med henholdsvis 35%, 35% og 30% 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 35 % ved bedømmelsen, og svarene skal begrundes. Data er venligst stillet til rådighed af Miriam Höllmer.

Data består af målinger af volumen i venstre forkammer af hjertet for 97 hunde. Der indgår data fra 5 forskellige hunderacer i datasættet. Hundene i datasættet lider ikke af hjertesygdomme, så målingerne anses at stamme fra raske hunde med normal størrelse af hjertet.

Filerne `hunde.xlsx` og `hunde.txt` indeholder data. Der er en datalinje for hver hund og to variable: `race`, som angiver hundens race, samt `maxLA`, som angiver volumen af venstre forkammer i hjertet målt i mL.

1.1 Antag at datasættet er indlæst i R under navnet `hunde`. Begrund hvorfor modellen fittet med `R`-koden

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

er uegnet til at beskrive sammenhængen mellem hjertevolumen og race. Du skal besvare spørgsmålet ved at udføre modelkontrol og kommentere på relevante grafer. Du kan enten vedlægge graferne elektronisk eller lave skitser i hånden.

1.2 Opskriv `R`-koden til at fitte en ensidet variansanalysemodel (ANOVA) med logaritmen til hjertevolumen som respons og `race` som forklarende variabel.

Opskriv også den tilhørende statistiske model, fit modellen i R, og angiv estimatet for residualspredningen σ og estimatet for det forventede log-transformerede hjertevolumen for en hund af racen `Whippet`.

Til besvarelse af spørgsmål **1.3-1.5** bedes du benytte modellen fra spørgsmål **1.2**. Du behøver ikke lave modelkontrol.

1.3 Lav et hypotesetest med henblik på at undersøge, om det forventede log-transformerede hjertevolumen kan antages at være ens for alle racer.

1.4 Bestem et estimat og et 95 %-konfidensinterval for forskellen i forventet log-transformeret hjertevolumen mellem hunde af racerne `Labrador` og `Petit_Basset`.

1.5 Hundene i datasættet er som bekendt raske. Bør der for en `Labrador` med et hjertevolumen på 32 mL rejses mistanke om, at hunden lider af et forstørret venstre forkammer i hjertet?

Hint: Du kan fx. benytte `predict()`- funktionen på et nyt datasæt konstrueret med kommandoen `newdata <- data.frame(race = "Labrador")`.

Opgave 2

Denne opgave vægtes med 35 % ved bedømmelsen, og svarene skal begrundes. Data er venligst stillet til rådighed af Julie Midtgaard.

Data til opgaven stammer fra et studie, hvor man ønskede at undersøge effekten af et træningsprogram på konditionen. Data findes i filerne `training.txt` og `training.xlsx`. Der er 67 datalinjer (en for hver forsøgsperson) og følgende variable

- `age`: alder ved starten af træningsperioden (i år)
- `sex`: forsøgspersonens køn (`K`: kvinde, `M`: mand)
- `before`: grundkondition målt som maksimal iltoptagelse (liter O_2/min) inden træningsprogrammet blev påbegyndt
- `after`: maksimal iltoptagelse ved afslutningen af træningsperioden (liter O_2/min)

Ved besvarelse af spørgsmål **2.1** og **2.2** bedes du se bort fra variablene `age` og `sex`.

2.1 Angiv en metode som er velegnet til at undersøge, om træningsprogrammet forbedrer konditionen (iltoptagelsen). Begrund dit svar.

Udfør et test for om træningsprogrammet forbedrer konditionen (iltoptagelsen), og skriv en konklusion på testet. R koden skal fremgå af besvarelsen.

2.2 Angiv et estimat og et 95 % - konfidensinterval for den forventede ændring i konditionen (maksimal iltoptagelse) hen over træningsperioden.

I det følgende interesserer vi os for variablen `forskel = after - before`. Man kunne forestille sig at ændringen af forsøgspersonernes kondition (`forskel`) hen over træningsperioden afhænger af forsøgspersonens køn og alder.

2.3 Opskriv den statistiske model der fittes med R-kommandoen

```
training$forskel <- training$after - training$before
model2 <- lm(forskel ~ sex + age, data = training)
```

hvor datasættet her er indlæst i R som `training`. Angiv også estimaterne for parametrene i modellen svarende til `model2`.

2.4 Undersøg med et hypotesetest om forsøgspersonens alder har indflydelse på ændringen i konditionen.

2.5 Find et estimat for den gennemsnitlige ændring i kondition (`forskel`) for en mand på 40 år.

Du skal **ikke** angive et 95 %-konfidensinterval for estimatet.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 30 % 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 Vægten af en tilfældigt valgt hund af racen Grand Danois antages at være normalfordelt med middelværdi 66.1 kg og en spredning på 7.7 kg. Hvad er sandsynligheden for at en tilfældigt valgt Grand Danois vejer mellem 60 og 70 kg?

A. 0.1000

B. 0.6937

C. 0.3063

D. 0.4796

E. 0.5204

3.2 Vægten af en tilfældigt valgt hund af racen Grand Danois antages at være normalfordelt med middelværdi 66.1 kg og en spredning på 7.7 kg. Angiv hvilken vægt en Grand Danois skal have, for at være blandt de 10 % tungeste.

A. 75.97 kg

B. 59.49 kg

C. 73.80 kg

D. 78.77 kg

E. 56.23 kg

3.3 Ved sammenligning af middelværdierne for to grupper/populationer beregnes på baggrund af to uafhængige stikprøver et 95 % - konfidensinterval for forskellen δ som går fra -0.3 til 1.7. Hvad kan vi da konkludere?

A. Nulhypotesen, $H_0 : \delta = 0$ bliver forkastet, hvis vi benytter et signifikansniveau på 5 %

B. Ved test af nulhypotesen $H_0 : \delta = 0$ fås en p-værdi på under 5 %.

C. Ved test af nulhypotesen $H_0 : \delta = 0$ fås en p-værdi på over 5 %.

D. Vi kan intet sige om hypotesen $H_0 : \delta = 0$ på baggrund af oplysningerne.

3.4 Ved et fodringsforsøg med 12 grise har man målt vægttilvæksten `wi` (enhed: pound per day) for hver gris. Desuden har man for hver gris registreret

A Indhold af antibiotika i foderet: `A = 1` svarer til 0 mg, `A = 2` svarer til 40 mg

B Indhold af B_{12}-vitamin i foderet: `B = 1` svarer til 0 mg, `B = 2` svarer til 5 mg

Alle 4 kombinationer af antibiotika (`A`) og B_{12}-vitamin (`B`) blev afprøvet med 3 grise for hver kombination.

Nedenfor ses resultatet af at køre `summary()` på en tosidet variansanalysemodel (ANOVA) med vekselvirkning mellem de to faktorer `A` og `B`.

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    1.1900000 0.03496029 34.0386144 6.062954e-10
## ANTIA2        -0.1566667 0.04944132 -3.1687394 1.321973e-02
## VITAB2         0.0300000 0.04944132  0.6067799 5.608182e-01
## ANTIA2:VITAB2  0.4800000 0.06992059  6.8649306 1.290220e-04
```

Hvilket af følgende udsagn er korrekt, på baggrund af resultaterne i R-outputtet?

A. Tilsætning af vitamin B_{12} påvirker ikke væksten (`wi`)

B. Tilsætning af vitamin B_{12} påvirker ikke væksten (`wi`), når der ikke tilsættes antibiotika (dvs. hvis `A = 1`)

C. Den forventede værdi af tilvæksten (`wi`) for `A = 2` (antibiotika 40 mg) og `B = 2` (vitamin B_{12} på 5 mg) estimeres til ca. 1.670

D. Den forventede værdi af tilvæksten (`wi`) for `A = 2` (antibiotika 40 mg) og `B = 2` (vitamin B_{12} på 5 mg) estimeres til ca. 1.063

3.5 I 2019 var 40 % af de studerende på Statistisk Dataanalyse 1 indskrevet på Biologi-Bioteknologi uddannelsen. Ved lodtrækning udvælges 10 studerende fra Statistisk Dataanalyse 1. Hvad er sandsynligheden for, at der er højst 5 (dvs. 5 eller færre) studerende fra Biologi-Bioteknologi som udvælges?

A. Ca. 0.201

B. Ca. 0.834

C. Ca. 0.167

D. Ca. 0.367

E. Ca. 0.400

3.6 Den fulde version af datasættet fra opgave 1 i eksamenssættet indeholdt oprindelig også variablen `wgt` der angiver hundenes vægt i kg. Hvis man laver lineær regression af logaritmen til hjertevolumen (`log(maxLA)`) med logaritmen til hundens vægt (`log(wgt)`) som forklarende variabel, så fås følgende R-output.

```
summary(lm(log(maxLA) ~ log(wgt), data = hunde))$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.1193072 0.09683692 -1.232043 2.209739e-01
## log(wgt)     0.8916347 0.03172285 28.107014 7.907221e-48
```

Angiv t-teststørrelsen til test af hypotesen $H_0 : \beta = 1$, hvor β er hældningen i den lineære regressionsmodel.

A. -1.23

B. 28.11

C. -3.42

D. -11.56

E. 1.99

3.7 I en nyligt publiceret artikel af P.B. Hansen & M. Penkowa (2017) har man optalt, at 5 patienter fik infektion ud de 13 patienter som blev behandlet med bismuth. Tilsvarende var der 6 ud af 10 patienter som fik infektion, hvis de blev behandlet med placebo.

I artiklen angives en forkert p-værdi på 0.0001 for test af hypotesen om, at der er lige stor infektionsrisiko for de to grupper patienter. Angiv en korrekt p-værdi baseret på χ^2-teststørrelsen for test af denne hypotese (når der ikke ønskes anvendt kontinuitetskorrektion).

Bemærk: 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.

A. 0.545

B. 0.215

C. 0.305

D. 0.812

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:

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. $T = 0.2462, P = 0.8105$

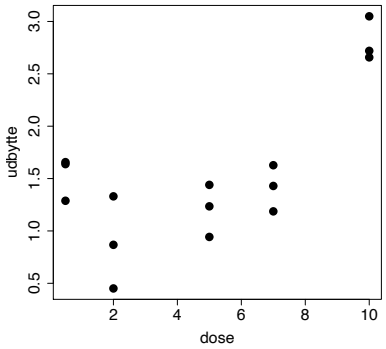
B. $T = 0.7387, P = 0.4789$

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

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 2021

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 40 %, 36 % og 24 % 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.

Formålet med denne opgave er at undersøge sammenhængen mellem `puls` og omgangstider (`tid`) på løbeture i Parc Montsouris i Paris. Datafilerne `nov2021opg1.txt` og `nov2021opg1.xlsx` består af 81 datalinjer, og kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data1 <- read_excel(path = "nov2021opg1.xlsx")
```

eller

```
data1 <- read.table(file = "nov2021opg1.txt", header = T)
```

De første seks linjer i datasættet ses her

```
##   tid_paa_dag puls tid
## 1   formiddag  143 445
## 2   formiddag  156 431
## 3   formiddag  156 428
## 4   formiddag  165 383
## 5   formiddag  163 401
## 6   formiddag  154 429
```

Hver datalinje indeholder bl.a. omgangstiden i sekunder (`tid`) på en løberute på ca. 1460 meter og den gennemsnitlige `puls` (enhed: slag per minut) på omgangen.

Ved besvarelsen af delopgave **1.1-1.4** skal du tage udgangspunkt i modellen `mod1`, der kan fittes med `R`-koden

```
mod1 <- lm(tid ~ puls, data = data1)
```

1.1 Opskriv den statistiske model svarende til modellen `mod1`, 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) 819.18562 36.1136836  22.68352 2.437125e-36
## puls         -2.54362  0.2178002 -11.67869 6.944555e-19
```

1.2 Benyt modellen `mod1` til at undersøge

- om der er evidens for, at der er en sammenhæng mellem puls og omgangstid.
- om datasættet understøtter en tommelfingerregel om, at hvis man øger `puls`en med 1 slag, så falder kilometertiden med 2 sekunder

Hint: Du kan enten se på relevante konfidensintervaller, eller du kan udføre t-test for relevante hypoteser.

1.3 Benyt modellen `mod1` til at finde et estimat og et 95 %-prædiktionsinterval for den forventede omgangstid, hvis der løbes en omgang med en `puls` på 160 (slag per minut).

1.4 Undersøg grundigt om modellen `mod1` er velegnet til at beskrive sammenhængen mellem puls og omgangstid. Du bedes både inddrage og kommentere på relevante grafer, og diskutere om det er rimeligt at beskrive sammenhængen mellem de to variable `puls` og `tid` ved en lineær funktion.

I datasættet har man desuden registreret variablen `tid_paa_dag`, som angiver om løbeturen fandt sted om `formiddag`en (før kl. 12) eller om `eftermiddag`en (efter kl. 12). Ved besvarelsen af delopgave **1.5-1.6** skal både `puls` og `tid_paa_dag` inddrages i modellen.

1.5 Opskriv den statistiske model der fittes med `R`-koden

```
mod2 <- lm(tid ~ puls + tid_paa_dag, data = data1)
```

Benyt resultaterne fra modellen `mod2` til at diskutere, om løberen er hurtigere til at løbe om formiddagen end om eftermiddagen.

1.6 Benyt `mod2` til at bestemme et estimat for den forventede løbetid på en omgang, hvis der løbes om `formiddag`en med en `puls` på 160.

Du behøver **ikke** angive et 95 %-konfidensinterval for estimatet.

Opgave 2

Denne opgave vægtes med 36 % ved bedømmelsen, og svarene skal begrundes.

Ældre kræftpatienter vil ofte opleve en hurtig tilbagegang i deres fysiske formåen under deres behandlingsforløb, som forstærker de negative effekter af selve kræftsygdommen. Ved et interventionsforsøg er en gruppe ældre kræftpatienter over 65 år ved lodtrækning blevet allokeret til enten standardbehandling (`control`) eller til en `intervention`, som bl.a. omfatter tilbud om 12 ugers fysisk træning. Formålet er at undersøge, om træning kan forhindre eller bremse tilbagegangen i patienternes fysiske formåen. I denne opgave fokuserer vi på håndgrebsstyrke (målt i kg) som mål for patienternes fysiske formåen.

Datafilerne `nov2021opg2.txt` og `nov2021opg2.xlsx` indeholder et udpluk af data fra det fulde forsøg. Der er data fra 58 patienter, og starten af datasættet kan ses her

```
##   week0 week12        treat      diagnose
## 1  27.3   27.8 intervention bugspytkirtel
## 2  24.7   25.1      control bugspytkirtel
## 3  41.8   38.6 intervention bugspytkirtel
## 4  30.3   28.6 intervention bugspytkirtel
## 5  30.1   25.0      control bugspytkirtel
## 6  30.8   26.4      control         lunge
```

Hver datalinje repræsenterer data fra en patient og indeholder variablene

- `week0` (håndgrebsstyrke målt i kg ved forsøgets start)
- `week12` (håndgrebsstyrke målt i kg ved forsøgets afslutning efter 12 uger)
- `treat` (behandlingsgruppe: `control/intervention`)
- `diagnose` (kræftsygdom, to diagnosegrupper: `lunge / bugspytkirtel`).

Data er venligst stillet til rådighed af Marta Kramer Mikkelsen. Data kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data2 <- read_excel(path = "nov2021opg2.xlsx")
```

eller

```
data2 <- read.table(file = "nov2021opg2.txt", header = T)
```

I hele opgaven benyttes ændringen i håndgrebsstyrke `week12-week0` som responsvariabel. Ved besvarelse af delopgaverne **2.1-2.2** skal du se bort fra variablen `diagnose`.

2.1 Fit modellen

```
m1 <- lm(week12 - week0 ~ treat, data = data2)
```

i `R` og angiv et estimat for residualspredningen.

Angiv desuden et estimat for den forventede ændring i håndgrebsstyrken for en patient som modtager standardbehandlingen.

2.2 Undersøg ved et hypotesetest om interventionen har en effekt på håndgrebsstyrken.

Patienterne i forsøget havde to forskellige kræftdiagnoser angivet ved variablen `diagnose` i datasættet. Man er særligt interesseret i at undersøge, om effekten af behandlingen er den ens for patienter i de to diagnosegrupper.

2.3 Forklar kortfattet hvorfor det er naturligt at benytte en tosidet variansanalysemodel (ANOVA) med vekselvirkning til at analysere datasættet.

Undersøg ved et hypotesetest om effekten af behandlingen er den samme for patienter i de to diagnosegrupper.

2.4 Tag udgangspunkt i en tosidet ANOVA uden vekselvirkning som nedenfor

```
m4 <- lm(week12 - week0 ~ treat + diagnose, data = data2)
summary(m4)$coef
```

```
##                    Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)       -1.015524   0.863508 -1.176045 0.24464226
## treatintervention  2.310833   1.036433  2.229603 0.02987625
## diagnoselunge      1.174799   1.033332  1.136903 0.26050781
```

Angiv et estimat for den forventede ændring i håndgrebsstyrken for personer fra interventionsgruppen for hver af de to diagnosegrupper.

2.5 Forklar hvordan man kan undersøge om der overhovedet sker noget med håndgrebsstyrken henover de 12 uger som forsøget varer. Du skal besvare spørgsmålet for begge behandlingsgrupper. Der lægges både vægt på, at du forklarer din fremgangsmåde, og at du fortolker resultaterne fra relevante statistiske modeller korrekt.

Hint: Der flere fornuftige løsninger på denne opgave. Din løsning bør indeholde estimater, konfidensintervaller og evt. hypotesetest fra relevante statistiske modeller. Du kan tage udgangspunkt i nogle af modellerne fra din besvarelse af delopgave **2.1-2.4**, eller du kan vælge at lave en ny model, der blot fokuserer på ændringerne over tid.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 24 % 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.

Ved forelæsningen i Statistisk Dataanalyse 1 d. 8/9-2021 blev de studerende (uden brug af lineal) bedt om at afsætte to punkter på et stykke papir med en afstand på ca. 8 cm. På baggrund af en stikprøve bestående af 20 målinger vurderes det, at afstandsmålingerne kan beskrives ved en normalfordeling med middelværdi 8.07 cm og spredning 2.10 cm. Benyt disse oplysninger til at besvare delopgaverne **3.1-3.2** nedenfor.

3.1 Beregn sandsynligheden for at en tilfældig studerende afsætter de to punkter med en indbyrdes afstand på mellem 7 og 9 centimeter.

A. Ca. 17.9 %

B. Ca. 36.6 %

C. Ca. 67.1 %

D. Ca. 30.5 %

E. Ca. 58.4 %

3.2 Angiv en afstand, L, så vi kan være sikre på, at 90 % af de studerende afsætter punkterne med kortere afstand end denne længde L.

A. $L \approx 9.8$ cm.

B. $L \approx 10.8$ cm.

C. $L \approx 12.3$ cm.

D. $L \approx 11.5$ cm.

E. $L \approx 5.4$ cm.

Ved forelæsningen d. 8/9-2021 på StatData 1 gættede 179 studerende på antallet af `Punkter` på en figur. Datasættet `data3` (som du ikke har adgang til) indeholder også oplysninger om `Studie`retning. Antal studerende fra de forskellige studieretninger er: `bioteknologi` 48, `husdyrvidenskab` 33, `jordbrugsøkonomi` 29, `naturressourcer` 69. Der er lavet en ensidet ANOVA med logaritmen til gæt på antal punkter som respons. Det oplyses her, at det korrekte antal punkter på figuren var 666.

```
model1 <- lm(log(Punkter) ~ Studie, data = data3)
```

```
summary(model1)$coef # et udpluk af output vises ...
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.8858     0.1131 52.0573   0.0000
## Studiehusdyrvidenskab      0.0874     0.1765  0.4953   0.6210
## Studiejordbrugsøkonomi     0.2340     0.1818  1.2869   0.1999
## Studienaturressourcer      0.2214     0.1464  1.5120   0.1324
```

Brug ovenstående `R`-output til at finde det korrekte svar i delopgaverne **3.3** og **3.4**.

3.3 Et 95 % - konfidensinterval for medianværdien for de 48 `bioteknologi`-studerendes gæt på antallet af punkter bliver

A. [297.7-435.1] punkter

B. [298.5-433.9] punkter

C. [287.9-449.9] punkter

D. [5.66-6.11] punkter

E. [286.7-451.8] punkter

3.4 Hvad fortæller tallet 0.2340 fra `R`-outputtet om *gæt på antal punkter* for studerende på `jordbrugsøkonomi` (JØ) og på `bioteknologi` (BB)?

A. At middelværdien estimeres til at være exp(0.2340) gange højere for JØ-studerende end for BB-studerende.

B. At middelværdien estimeres til at være 23.4 % højere for JØ-studerende end for BB-studerende.

C. At P-værdien er 0.2340 for test af hypotesen om, at der er forskel på gæt på antal punkter for JØ-studerende og for BB-studerende

D. At medianværdien estimeres til at være 23.4 % højere for JØ-studerende end for BB-studerende.

E. At medianværdien estimeres til at være exp(0.2340) gange højere for JØ-studerende end for BB-studerende.

3.5 En studerende beslutter sig for at gætte på svarene på alle 6 opgaver i en multiple choice prøve. Der er 5 svarmuligheder for hver opgave, hvoraf kun et svar er korrekt, så sandsynligheden for at svare rigtigt på hver opgave er 1/5 .

Find sandsynligheden for, at den studerende højst gætter rigtigt på to opgaver?

A. Ca. 24.6 %

B. Ca. 65.5 %

C. Ca. 90.1 %

D. Ca. 34.5 %

E. Ca. 9.9 %

3.6 Ved en forelæsning har 179 studerende bl.a. svaret på, om de tror at forelæseren kan lide at hækle, og om han kan lide at plukke kantareller. Resultaterne er opsummeret i en antalstabel, og man har kørt følgende R-kode

```
my_tab
```

```
##         haekle
## kantarel FALSE TRUE
##    FALSE    74    7
##    TRUE     75   23
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  my_tab
## X-squared = 6.9886, df = 1, p-value = 0.008203
```

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

A. Der er ingen sammenhæng mellem om studerende gætter på at forelæser kan lide at hækle og at plukke kantareller.

B. Andelen af studerende som gætter på at forelæser kan lide at hækle og at plukke kantareller er ikke lig med 50 %.

C. Andelen af studerende som gætter på at forelæser kan lide at hækle og at plukke kantareller er ens.

D. Der er en sammenhæng mellem om studerende gætter på at forelæser kan lide at hækle og at plukke kantareller.

E. Andelen af studerende som gætter på at forelæser kan lide at hækle og at plukke kantareller er ikke ens.

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 2021

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 40 %, 36 % og 24 % 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.

Formålet med denne opgave er at undersøge sammenhængen mellem `puls` og omgangstider (`tid`) på løbeture i Parc Montsouris i Paris. Datafilerne `nov2021opg1.txt` og `nov2021opg1.xlsx` består af 81 datalinjer, og kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data1 <- read_excel(path = "nov2021opg1.xlsx")
```

eller

```
data1 <- read.table(file = "nov2021opg1.txt", header = T)
```

De første seks linjer i datasættet ses her

```
##   tid_paa_dag puls tid
## 1   formiddag  143 445
## 2   formiddag  156 431
## 3   formiddag  156 428
## 4   formiddag  165 383
## 5   formiddag  163 401
## 6   formiddag  154 429
```

Hver datalinje indeholder bl.a. omgangstiden i sekunder (`tid`) på en løberute på ca. 1460 meter og den gennemsnitlige `puls` (enhed: slag per minut) på omgangen.

Ved besvarelsen af delopgave **1.1-1.4** skal du tage udgangspunkt i modellen `mod1`, der kan fittes med `R`-koden

```
mod1 <- lm(tid ~ puls, data = data1)
```

1.1 Opskriv den statistiske model svarende til modellen `mod1`, 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) 819.18562 36.1136836  22.68352 2.437125e-36
## puls         -2.54362  0.2178002 -11.67869 6.944555e-19
```

1.2 Benyt modellen `mod1` til at undersøge

- om der er evidens for, at der er en sammenhæng mellem puls og omgangstid.
- om datasættet understøtter en tommelfingerregel om, at hvis man øger `puls`en med 1 slag, så falder kilometertiden med 2 sekunder

Hint: Du kan enten se på relevante konfidensintervaller, eller du kan udføre t-test for relevante hypoteser.

1.3 Benyt modellen `mod1` til at finde et estimat og et 95 %-prædiktionsinterval for den forventede omgangstid, hvis der løbes en omgang med en `puls` på 160 (slag per minut).

1.4 Undersøg grundigt om modellen `mod1` er velegnet til at beskrive sammenhængen mellem puls og omgangstid. Du bedes både inddrage og kommentere på relevante grafer, og diskutere om det er rimeligt at beskrive sammenhængen mellem de to variable `puls` og `tid` ved en lineær funktion.

I datasættet har man desuden registreret variablen `tid_paa_dag`, som angiver om løbeturen fandt sted om `formiddag`en (før kl. 12) eller om `eftermiddag`en (efter kl. 12). Ved besvarelsen af delopgave **1.5-1.6** skal både `puls` og `tid_paa_dag` inddrages i modellen.

1.5 Opskriv den statistiske model der fittes med `R`-koden

```
mod2 <- lm(tid ~ puls + tid_paa_dag, data = data1)
```

Benyt resultaterne fra modellen `mod2` til at diskutere, om løberen er hurtigere til at løbe om formiddagen end om eftermiddagen.

1.6 Benyt `mod2` til at bestemme et estimat for den forventede løbetid på en omgang, hvis der løbes om `formiddag`en med en `puls` på 160.

Du behøver **ikke** angive et 95 %-konfidensinterval for estimatet.

Opgave 2

Denne opgave vægtes med 36 % ved bedømmelsen, og svarene skal begrundes.

Ældre kræftpatienter vil ofte opleve en hurtig tilbagegang i deres fysiske formåen under deres behandlingsforløb, som forstærker de negative effekter af selve kræftsygdommen. Ved et interventionsforsøg er en gruppe ældre kræftpatienter over 65 år ved lodtrækning blevet allokeret til enten standardbehandling (`control`) eller til en `intervention`, som bl.a. omfatter tilbud om 12 ugers fysisk træning. Formålet er at undersøge, om træning kan forhindre eller bremse tilbagegangen i patienternes fysiske formåen. I denne opgave fokuserer vi på håndgrebsstyrke (målt i kg) som mål for patienternes fysiske formåen.

Datafilerne `nov2021opg2.txt` og `nov2021opg2.xlsx` indeholder et udpluk af data fra det fulde forsøg. Der er data fra 58 patienter, og starten af datasættet kan ses her

```
##   week0 week12        treat      diagnose
## 1  27.3   27.8 intervention bugspytkirtel
## 2  24.7   25.1      control bugspytkirtel
## 3  41.8   38.6 intervention bugspytkirtel
## 4  30.3   28.6 intervention bugspytkirtel
## 5  30.1   25.0      control bugspytkirtel
## 6  30.8   26.4      control         lunge
```

Hver datalinje repræsenterer data fra en patient og indeholder variablene

- `week0` (håndgrebsstyrke målt i kg ved forsøgets start)
- `week12` (håndgrebsstyrke målt i kg ved forsøgets afslutning efter 12 uger)
- `treat` (behandlingsgruppe: `control/intervention`)
- `diagnose` (kræftsygdom, to diagnosegrupper: `lunge / bugspytkirtel`).

Data er venligst stillet til rådighed af Marta Kramer Mikkelsen. Data kan fx. indlæses med en af følgende `R`-kommandoer

```
library(readxl)
data2 <- read_excel(path = "nov2021opg2.xlsx")
```

eller

```
data2 <- read.table(file = "nov2021opg2.txt", header = T)
```

I hele opgaven benyttes ændringen i håndgrebsstyrke `week12-week0` som responsvariabel. Ved besvarelse af delopgaverne **2.1-2.2** skal du se bort fra variablen `diagnose`.

2.1 Fit modellen

```
m1 <- lm(week12 - week0 ~ treat, data = data2)
```

i `R` og angiv et estimat for residualspredningen.

Angiv desuden et estimat for den forventede ændring i håndgrebsstyrken for en patient som modtager standardbehandlingen.

2.2 Undersøg ved et hypotesetest om interventionen har en effekt på håndgrebsstyrken.

Patienterne i forsøget havde to forskellige kræftdiagnoser angivet ved variablen `diagnose` i datasættet. Man er særligt interesseret i at undersøge, om effekten af behandlingen er den ens for patienter i de to diagnosegrupper.

2.3 Forklar kortfattet hvorfor det er naturligt at benytte en tosidet variansanalysemodel (ANOVA) med vekselvirkning til at analysere datasættet.

Undersøg ved et hypotesetest om effekten af behandlingen er den samme for patienter i de to diagnosegrupper.

2.4 Tag udgangspunkt i en tosidet ANOVA uden vekselvirkning som nedenfor

```
m4 <- lm(week12 - week0 ~ treat + diagnose, data = data2)
summary(m4)$coef
```

```
##                    Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)       -1.015524   0.863508 -1.176045 0.24464226
## treatintervention  2.310833   1.036433  2.229603 0.02987625
## diagnoselunge      1.174799   1.033332  1.136903 0.26050781
```

Angiv et estimat for den forventede ændring i håndgrebsstyrken for personer fra interventionsgruppen for hver af de to diagnosegrupper.

2.5 Forklar hvordan man kan undersøge om der overhovedet sker noget med håndgrebsstyrken henover de 12 uger som forsøget varer. Du skal besvare spørgsmålet for begge behandlingsgrupper. Der lægges både vægt på, at du forklarer din fremgangsmåde, og at du fortolker resultaterne fra relevante statistiske modeller korrekt.

Hint: Der flere fornuftige løsninger på denne opgave. Din løsning bør indeholde estimater, konfidensintervaller og evt. hypotesetest fra relevante statistiske modeller. Du kan tage udgangspunkt i nogle af modellerne fra din besvarelse af delopgave **2.1-2.4**, eller du kan vælge at lave en ny model, der blot fokuserer på ændringerne over tid.

Opgave 3 (quizspørgsmål)

Denne opgave vægtes med 24 % 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.

Ved forelæsningen i Statistisk Dataanalyse 1 d. 8/9-2021 blev de studerende (uden brug af lineal) bedt om at afsætte to punkter på et stykke papir med en afstand på ca. 8 cm. På baggrund af en stikprøve bestående af 20 målinger vurderes det, at afstandsmålingerne kan beskrives ved en normalfordeling med middelværdi 8.07 cm og spredning 2.10 cm. Benyt disse oplysninger til at besvare delopgaverne **3.1-3.2** nedenfor.

3.1 Beregn sandsynligheden for at en tilfældig studerende afsætter de to punkter med en indbyrdes afstand på mellem 7 og 9 centimeter.

A. Ca. 17.9 %

B. Ca. 36.6 %

C. Ca. 67.1 %

D. Ca. 30.5 %

E. Ca. 58.4 %

3.2 Angiv en afstand, L, så vi kan være sikre på, at 90 % af de studerende afsætter punkterne med kortere afstand end denne længde L.

A. $L \approx 9.8$ cm.

B. $L \approx 10.8$ cm.

C. $L \approx 12.3$ cm.

D. $L \approx 11.5$ cm.

E. $L \approx 5.4$ cm.

Ved forelæsningen d. 8/9-2021 på StatData 1 gættede 179 studerende på antallet af `Punkter` på en figur. Datasættet `data3` (som du ikke har adgang til) indeholder også oplysninger om `Studie`retning. Antal studerende fra de forskellige studieretninger er: `bioteknologi` 48, `husdyrvidenskab` 33, `jordbrugsøkonomi` 29, `naturressourcer` 69. Der er lavet en ensidet ANOVA med logaritmen til gæt på antal punkter som respons. Det oplyses her, at det korrekte antal punkter på figuren var 666.

```
model1 <- lm(log(Punkter) ~ Studie, data = data3)
```

```
summary(model1)$coef # et udpluk af output vises ...
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.8858     0.1131 52.0573   0.0000
## Studiehusdyrvidenskab      0.0874     0.1765  0.4953   0.6210
## Studiejordbrugsøkonomi     0.2340     0.1818  1.2869   0.1999
## Studienaturressourcer      0.2214     0.1464  1.5120   0.1324
```

Brug ovenstående `R`-output til at finde det korrekte svar i delopgaverne **3.3** og **3.4**.

3.3 Et 95 % - konfidensinterval for medianværdien for de 48 `bioteknologi`-studerendes gæt på antallet af punkter bliver

A. [297.7-435.1] punkter

B. [298.5-433.9] punkter

C. [287.9-449.9] punkter

D. [5.66-6.11] punkter

E. [286.7-451.8] punkter

3.4 Hvad fortæller tallet 0.2340 fra `R`-outputtet om *gæt på antal punkter* for studerende på `jordbrugsøkonomi` (JØ) og på `bioteknologi` (BB)?

A. At middelværdien estimeres til at være exp(0.2340) gange højere for JØ-studerende end for BB-studerende.

B. At middelværdien estimeres til at være 23.4 % højere for JØ-studerende end for BB-studerende.

C. At P-værdien er 0.2340 for test af hypotesen om, at der er forskel på gæt på antal punkter for JØ-studerende og for BB-studerende

D. At medianværdien estimeres til at være 23.4 % højere for JØ-studerende end for BB-studerende.

E. At medianværdien estimeres til at være exp(0.2340) gange højere for JØ-studerende end for BB-studerende.

3.5 En studerende beslutter sig for at gætte på svarene på alle 6 opgaver i en multiple choice prøve. Der er 5 svarmuligheder for hver opgave, hvoraf kun et svar er korrekt, så sandsynligheden for at svare rigtigt på hver opgave er 1/5 .

Find sandsynligheden for, at den studerende højst gætter rigtigt på to opgaver?

A. Ca. 24.6 %

B. Ca. 65.5 %

C. Ca. 90.1 %

D. Ca. 34.5 %

E. Ca. 9.9 %

3.6 Ved en forelæsning har 179 studerende bl.a. svaret på, om de tror at forelæseren kan lide at hækle, og om han kan lide at plukke kantareller. Resultaterne er opsummeret i en antalstabel, og man har kørt følgende R-kode

```
my_tab
```

```
##         haekle
## kantarel FALSE TRUE
##    FALSE    74    7
##    TRUE     75   23
```

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

```
##
##  Pearson's Chi-squared test
##
## data:  my_tab
## X-squared = 6.9886, df = 1, p-value = 0.008203
```

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

A. Der er ingen sammenhæng mellem om studerende gætter på at forelæser kan lide at hækle og at plukke kantareller.

B. Andelen af studerende som gætter på at forelæser kan lide at hækle og at plukke kantareller er ikke lig med 50 %.

C. Andelen af studerende som gætter på at forelæser kan lide at hækle og at plukke kantareller er ens.

D. Der er en sammenhæng mellem om studerende gætter på at forelæser kan lide at hækle og at plukke kantareller.

E. Andelen af studerende som gætter på at forelæser kan lide at hækle og at plukke kantareller er ikke ens.