# 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.