# 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 %