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%.

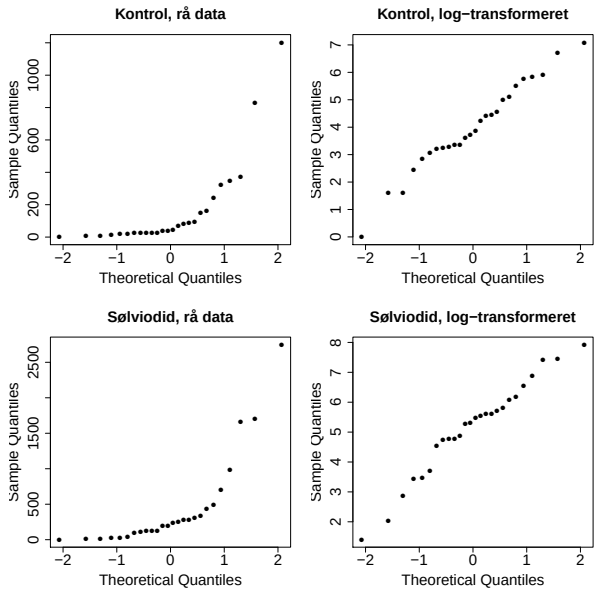
4. Tyder data på at der er en effekt af restriktive love på selvmordsraten, når man tager højde for våbenudbredelsen?

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.