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.