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: