Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 2018

Fire timers skriftlig prøve. Alle hjælpemidler er tilladt, herunder computer, men du må ikke tilgå internettet.

Der er 3 opgaver, som vægtes med henholdsvis 42%, 28% 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-ouputtet 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 42% ved bedømmelsen, og svarene skal begrundes. Data er stillet til rådighed af Lars Båstrup-Spohr fra Biologisk Institut på KU.

Data består af fosforkoncentrationen i 1242 søer i Danmark og Sydsverige. Søerne ligger i fem forskellige områder (Jylland, Østdanmark, Skåne, Småland og Blekinge), og man er interesseret i at sammenligne fosforkoncentrationen i de fem områder.

Filerne `soeer.xlsx` og `soeer.txt` indeholder data. Der er en datalinie for hver sø og to variable: `Sted`, som angiver hvilket af de fem områder søen ligger i, og `Fosfor`, som er fosforkoncentration i søen, målt i μg per liter.

1.1 Forklar kortfattet hvorfor datasættet lægger op til en ensidet ANOVA (ensidet variansanalyse).

Angiv desuden en `lm`-kommando der kan bruges til at fitte en ensidet ANOVA med fosforkoncentration som respons og område som forklarende variabel, samt `lm`-kommando der kan bruges til at fitte en ensidet ANOVA med log-transformeret fosforkoncentration som respons og område som forklarende variabel

1.2 Lav modelkontrol for hver af de to modeller fra spørgsmål 1. Kommentér herunder relevante grafer, og forklar hvorfor modellen med log-transformeret fosforkoncentration som respons er at foretrække. Du skal enten vedlægge de relevante grafer elektronisk eller lave skitser af dem i hånden.

I det følgende skal du benytte modellen med log-transformeret fosforkoncentrationen som respons.

1.3 Undersøg med et enkelt hypotesetest om den forventede log-transformerede fosforkoncentrationen kan antages at være den samme for alle fem områder.

1.4 Bestem estimater for den forventede log-transformerede fosforkoncentrationen i Blekinge og i Skåne. Bestem også de tilhørende 95% konfidensintervaller.

1.5 Bestem et estimat og et 95% konfidensinterval for forskellen i forventet log-transformeret fosforkoncentrationen mellem Blekinge og Skåne.

Bestem derefter et estimat og et 95% konfidensinterval for den faktor som forforkoncentrationen er højere i Skåne i forhold til Blekinge.

1.6 Undersøg med et enkelt hypotesetest om den forventede log-transformerede fosforkoncentrationen kan antages at være ens i de tre svenske områder (Blekinge, Skåne, Småland). For at få fuldt pointtal skal du stadig bruge modellen for alle 1242 søer.

Opgave 2

Denne opgave vægtes med 28% ved bedømmelsen, og svarene skal begrundes.

Data til denne opgave stammer fra valgkampe i forbindelse med 15 amerikanske borgmestervalg, hvor den siddende borgmester stillede op. Data er tilgængelige i filerne `elections.xlsx` og `elections.txt`. Der er 15 datalinier og følgende tre variable:

- `approval`: Den siddende borgmesters popularitet ved valgkampens begyndelse, angivet som den procentdel der i en meningsmåling angav at ville stemme på kandidaten
- `expenditures`: Udgifter til valgkampagnen for den siddende borgmester, angivet i 1000 dollars
- `performance`: Resultatet af valget for den siddende borgmester, angivet som den procentdel af stemmerne der gik til ham/hende

I det første spørgsmål skal du kun benytte variablene `approval` og `expenditures`.

2.1 Opskriv en lineær regressionsmodel, der kan benyttes til at undersøge om populariteten ved valgkampens begyndelse har betydning for hvor mange penge der benyttes til valgkampagnen.

Er der i evidens i data for at der er en sammenhæng mellem borgmesterens popularitet og hans/hendes udgifter til valgkampagnen?

Man er først og fremmest interesseret i at undersøge hvordan den siddende borgmesters udgifter til valgkampagnen påvirker hans/hendes chance for at genvinde valget. Vi vil benytte den multiple regressionsmodel der fittes med følgende kommando (hvor `elections` er navnet på datasættet med de tre variable):

```
lm(performance ~ approval + expenditures, data=elections)
```

2.2 Opskriv den estimerede sammenhæng mellem de tre variable svarende til den multiple regressionsmodel. Angiv også estimatet for residualspredningen.

2.3 Forklar kortfattet hvordan estimatet hørende til variablen `expenditures` skal fortolkes.

Undersøg med et hypotesetest om det hjælper på valgresultatet at bruge flere penge på valgkampagnen.

2.4 I en ny valgkamp har en siddende borgmester en popularitet på 40% ved valgkampens start og vælger at bruge 110000 dollars på sin kampagne. Bestem et interval som med 95% sandsynlighed vil indeholde valgresultatet for den pågældende borgmester.

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 SAT er en adgangstest til højere uddannelser i USA, og den blev taget af 1.7 millioner high school-elever i USA i 2017. Resultatet af testen kaldes for SAT-scoren. I 2017 var SAT-scoren normalfordelt med middelværdi 1060 og spredning 195.

Bestem sandsynligheden for at en tilfældig high school-elev har en SAT-score der er mindre end 880.

A. 0.741
B. 0.498
C. 0.259
D. 0.822
E. 0.178

3.2 SAT er en adgangstest til højere uddannelser i USA, og den blev taget af 1.7 millioner high school-elever i USA i 2017. Resultatet af testen kaldes for SAT-scoren. I 2017 var SAT-scoren normalfordelt med middelværdi 1060 og spredning 195.

Et college optager kun studerende der har en SAT-score blandt de 25% højeste. Hvor stor en SAT-score skal man have for at komme i betragtning?

A. 1192
B. 1442
C. 928
D. 1201
E. 1284

3.3 I en amerikansk spørgeskemaundersøgelse blev 100 personer spurgt, om de var rygere eller ikke-rygere, og om de syntes at skatten på cigaretter skulle hæves. Fordelingen af personer er vist i tabellen nedenfor.

	Ja til skattestigning	Nej til skattestigning
Ikke-ryger	44	32
Ryger	5	19

Man har udført et test for uafhængighed i tabellen.

Hvad er p-værdien og konklusionen? Bemærk at p-værdien er beregnet uden kontinuitetskorrektion, dvs. med optionen `correct=FALSE`.

A. $p = 0.013$, og ikke-rygere er mere tilbøjelige end rygere til at være for en skattestigning
B. $p = 0.013$, og der er ikke tegn på sammenhæng mellem rygevaner og holdning til skattestigning
C. $p = 0.40$, og ikke-rygere er mere tilbøjelige end rygere til at være for en skattestigning
D. $p = 0.40$, og der er ikke tegn på sammenhæng mellem rygevaner og holdning til skattestigning
E. $p = 0.0015$, og der er ikke tegn på sammenhæng mellem rygevaner og holdning til skattestigning
F. $p = 0.0015$, og ikke-rygere er mere tilbøjelige end rygere til at være for en skattestigning

3.4 10% af den danske befolkning er venstrehåndede. Udtag en stikprøve på 25 personer. Hvad er sandsynligheden for at fire eller færre af de 25 personer er venstrehåndede.

A. 0.138
B. 0.098
C. 0.764
D. 0.902
E. 0.862
F. 0.236

3.5 En ny type medicin mod infektion kan enten gives som tablet eller som dråber, og i to forskellige doser. I et studie har man fordelt 88 patienter tilfældigt i fire grupper svarende til de fire kombinationer af administration (tablet/dråber) og dosis (høj/lav). Patienterne har fået medicinen i 6 dage, og ændringen i infektionstal fra start til slut er registreret.

Man vil benytte ændringen i infektionstal som responsvariabel, og man vil gerne estimere effekten af den høje dosis i forhold til den lave samt estimere effekten af tabletformen i forhold til dråbeformen. Fra tidligere studier ved man at effekten af dosis er den samme uanset hvordan medicinen administreres, og man er ikke interesseret i at undersøge dette nærmere.

Hvilken type model bør benyttes til analysen?

A. En ensidet ANOVA med dosis som forklarende variabel
B. En tosidet ANOVA med metode og dosis samt deres vekselvirkning som forklarende variable
C. En lineær regression med dosis som forklarende variabel
D. En ensidet ANOVA med metode som forklarende variabel
E. En tosidet ANOVA med metode og dosis som forklarende variable, men uden vekselvirkning

3.6 En politisk meningsmåling gennemføres på 1000 personer. Man tæller specielt hvor mange personer der siger at de vil stemme på Alternativet; dette antal kaldes Y. Så er Y binomialfordelt, $Y \sim \text{bin}(1000, p)$, hvor p er andelen i befolkningen der vil stemme på Alternativet.

Ved folketingsvalget i 2015 fik Alternativet 4.4% af stemmerne, og man tester derfor hypotesen $H_0 : p = 0.044$ svarende til at tilslutningen til partiet er uændret.

Hvad er en type II fejl i denne situation?

A. Vi konkluderer at tilslutningen til Alternativet er uændret selvom den i virkeligheden har ændret sig.

B. Vi konkluderer at tilslutningen til Alternativet har ændret sig, når dette også er sandt i virkeligheden.

C. Vi konkluderer at tilslutningen til Alternativet er uændret, når dette også er sandt i virkeligheden.

D. Vi konkluderer at tilslutningen til Alternativet har ændret sig selvom den i virkeligheden er uændret.

3.7 Man har registreret fødselsvægten for en stikprøve på 32 nyfødte børn. Gennemsnittet viste sig at være 3487 g, og spredningen var 443 g.

Bestem et 95% konfindensinterval for den gennemsnitlige fødselsvægt.

A. $(3464\,,\,3510)$

B. $(3327\,,\,3647)$

C. $(2583\,,\,4391)$

D. $(2736\,,\,4238)$

E. $(3459\,,\,3515)$

F. $(3354\,,\,3620)$

3.8 En stikprøve bestående af 100 personer fra den voksne danske befolkning har taget en matematiktest. Resultaterne kan antages at være uafhængige og normalfordelte med middelværdi μ og spredning σ, som begge er ukendte parametre. Udfra resultaterne har man beregnet et 95% konfidensinterval for μ til $(71.5\,,\,75.3)$.

Hvad kan vi konkludere?

A. 95 personer i stikprøven havde et resultat mellem 71.5 og 75.3.

B. For 95% af befolkningen ligger reultatet mellem 71.5 og 75.3.

C. Vi er 95% sikre på at gennemsnittet for befolkningen ligger mellem 71.5 og 75.3.

D. Hvis vi valgte 100 andre personer, ville de alle have et resultat mellem 71.5 og 75.3.

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: