Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamen, november 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 bedømmelsen. Husk at de fleste spørgsmål kan besvares uafhængigt af hinanden.

Data til opgave 1 og opgave 2 udleveres på en USB-nøgle. Navne på filer der indeholder data fremgår af opgaveteksten. Denne USB-nøgle 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-nøgle som udleveres til formålet af eksamensvagterne. Bemærk at kun pdf-format accepteres og at du skal benytte en anden USB-nøgle 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. Data er venligst stillet til rådighed af Miriam Höllmer.

Data består af målinger af volumen i venstre forkammer af hjertet for 97 hunde. Der indgår data fra 5 forskellige hunderacer i datasættet. Hundene i datasættet lider ikke af hjertesygdomme, så målingerne anses at stamme fra raske hunde med normal størrelse af hjertet.

Filerne `hunde.xlsx` og `hunde.txt` indeholder data. Der er en datalinje for hver hund og to variable: `race`, som angiver hundens race, samt `maxLA`, som angiver volumen af venstre forkammer i hjertet målt i mL.

1.1 Antag at datasættet er indlæst i R under navnet `hunde`. Begrund hvorfor modellen fittet med `R`-koden

```
model1 <- lm(maxLA ~ race, data = hunde)
```

er uegnet til at beskrive sammenhængen mellem hjertevolumen og race. Du skal besvare spørgsmålet ved at udføre modelkontrol og kommentere på relevante grafer. Du kan enten vedlægge graferne elektronisk eller lave skitser i hånden.

1.2 Opskriv `R`-koden til at fitte en ensidet variansanalysemodel (ANOVA) med logaritmen til hjertevolumen som respons og `race` som forklarende variabel.

Opskriv også den tilhørende statistiske model, fit modellen i R, og angiv estimatet for residualspredningen σ og estimatet for det forventede log-transformerede hjertevolumen for en hund af racen `Whippet`.

Til besvarelse af spørgsmål **1.3-1.5** bedes du benytte modellen fra spørgsmål **1.2**. Du behøver ikke lave modelkontrol.

1.3 Lav et hypotesetest med henblik på at undersøge, om det forventede log-transformerede hjertevolumen kan antages at være ens for alle racer.

1.4 Bestem et estimat og et 95 %-konfidensinterval for forskellen i forventet log-transformeret hjertevolumen mellem hunde af racerne `Labrador` og `Petit_Basset`.

1.5 Hundene i datasættet er som bekendt raske. Bør der for en `Labrador` med et hjertevolumen på 32 mL rejses mistanke om, at hunden lider af et forstørret venstre forkammer i hjertet?

Hint: Du kan fx. benytte `predict()`- funktionen på et nyt datasæt konstrueret med kommandoen `newdata <- data.frame(race = "Labrador")`.

Opgave 2

Denne opgave vægtes med 35 % ved bedømmelsen, og svarene skal begrundes. Data er venligst stillet til rådighed af Julie Midtgaard.

Data til opgaven stammer fra et studie, hvor man ønskede at undersøge effekten af et træningsprogram på konditionen. Data findes i filerne `training.txt` og `training.xlsx`. Der er 67 datalinjer (en for hver forsøgsperson) og følgende variable

- `age`: alder ved starten af træningsperioden (i år)
- `sex`: forsøgspersonens køn (`K`: kvinde, `M`: mand)
- `before`: grundkondition målt som maksimal iltoptagelse (liter O_2/min) inden træningsprogrammet blev påbegyndt
- `after`: maksimal iltoptagelse ved afslutningen af træningsperioden (liter O_2/min)

Ved besvarelse af spørgsmål **2.1** og **2.2** bedes du se bort fra variablene `age` og `sex`.

2.1 Angiv en metode som er velegnet til at undersøge, om træningsprogrammet forbedrer konditionen (iltoptagelsen). Begrund dit svar.

Udfør et test for om træningsprogrammet forbedrer konditionen (iltoptagelsen), og skriv en konklusion på testet. R koden skal fremgå af besvarelsen.

2.2 Angiv et estimat og et 95 % - konfidensinterval for den forventede ændring i konditionen (maksimal iltoptagelse) hen over træningsperioden.

I det følgende interesserer vi os for variablen `forskel = after - before`. Man kunne forestille sig at ændringen af forsøgspersonernes kondition (`forskel`) hen over træningsperioden afhænger af forsøgspersonens køn og alder.

2.3 Opskriv den statistiske model der fittes med R-kommandoen

```
training$forskel <- training$after - training$before
model2 <- lm(forskel ~ sex + age, data = training)
```

hvor datasættet her er indlæst i R som `training`. Angiv også estimaterne for parametrene i modellen svarende til `model2`.

2.4 Undersøg med et hypotesetest om forsøgspersonens alder har indflydelse på ændringen i konditionen.

2.5 Find et estimat for den gennemsnitlige ændring i kondition (`forskel`) for en mand på 40 år.

Du skal **ikke** angive et 95 %-konfidensinterval for estimatet.

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 Vægten af en tilfældigt valgt hund af racen Grand Danois antages at være normalfordelt med middelværdi 66.1 kg og en spredning på 7.7 kg. Hvad er sandsynligheden for at en tilfældigt valgt Grand Danois vejer mellem 60 og 70 kg?

A. 0.1000

B. 0.6937

C. 0.3063

D. 0.4796

E. 0.5204

3.2 Vægten af en tilfældigt valgt hund af racen Grand Danois antages at være normalfordelt med middelværdi 66.1 kg og en spredning på 7.7 kg. Angiv hvilken vægt en Grand Danois skal have, for at være blandt de 10 % tungeste.

A. 75.97 kg

B. 59.49 kg

C. 73.80 kg

D. 78.77 kg

E. 56.23 kg

3.3 Ved sammenligning af middelværdierne for to grupper/populationer beregnes på baggrund af to uafhængige stikprøver et 95 % - konfidensinterval for forskellen δ som går fra -0.3 til 1.7. Hvad kan vi da konkludere?

A. Nulhypotesen, $H_0 : \delta = 0$ bliver forkastet, hvis vi benytter et signifikansniveau på 5 %

B. Ved test af nulhypotesen $H_0 : \delta = 0$ fås en p-værdi på under 5 %.

C. Ved test af nulhypotesen $H_0 : \delta = 0$ fås en p-værdi på over 5 %.

D. Vi kan intet sige om hypotesen $H_0 : \delta = 0$ på baggrund af oplysningerne.

3.4 Ved et fodringsforsøg med 12 grise har man målt vægttilvæksten `wi` (enhed: pound per day) for hver gris. Desuden har man for hver gris registreret

A Indhold af antibiotika i foderet: `A = 1` svarer til 0 mg, `A = 2` svarer til 40 mg

B Indhold af B_{12}-vitamin i foderet: `B = 1` svarer til 0 mg, `B = 2` svarer til 5 mg

Alle 4 kombinationer af antibiotika (`A`) og B_{12}-vitamin (`B`) blev afprøvet med 3 grise for hver kombination.

Nedenfor ses resultatet af at køre `summary()` på en tosidet variansanalysemodel (ANOVA) med vekselvirkning mellem de to faktorer `A` og `B`.

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    1.1900000 0.03496029 34.0386144 6.062954e-10
## ANTIA2        -0.1566667 0.04944132 -3.1687394 1.321973e-02
## VITAB2         0.0300000 0.04944132  0.6067799 5.608182e-01
## ANTIA2:VITAB2  0.4800000 0.06992059  6.8649306 1.290220e-04
```

Hvilket af følgende udsagn er korrekt, på baggrund af resultaterne i R-outputtet?

A. Tilsætning af vitamin B_{12} påvirker ikke væksten (`wi`)

B. Tilsætning af vitamin B_{12} påvirker ikke væksten (`wi`), når der ikke tilsættes antibiotika (dvs. hvis `A = 1`)

C. Den forventede værdi af tilvæksten (`wi`) for `A = 2` (antibiotika 40 mg) og `B = 2` (vitamin B_{12} på 5 mg) estimeres til ca. 1.670

D. Den forventede værdi af tilvæksten (`wi`) for `A = 2` (antibiotika 40 mg) og `B = 2` (vitamin B_{12} på 5 mg) estimeres til ca. 1.063

3.5 I 2019 var 40 % af de studerende på Statistisk Dataanalyse 1 indskrevet på Biologi-Bioteknologi uddannelsen. Ved lodtrækning udvælges 10 studerende fra Statistisk Dataanalyse 1. Hvad er sandsynligheden for, at der er højst 5 (dvs. 5 eller færre) studerende fra Biologi-Bioteknologi som udvælges?

A. Ca. 0.201

B. Ca. 0.834

C. Ca. 0.167

D. Ca. 0.367

E. Ca. 0.400

3.6 Den fulde version af datasættet fra opgave 1 i eksamenssættet indeholdt oprindelig også variablen `wgt` der angiver hundenes vægt i kg. Hvis man laver lineær regression af logaritmen til hjertevolumen (`log(maxLA)`) med logaritmen til hundens vægt (`log(wgt)`) som forklarende variabel, så fås følgende R-output.

```
summary(lm(log(maxLA) ~ log(wgt), data = hunde))$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -0.1193072 0.09683692 -1.232043 2.209739e-01
## log(wgt)     0.8916347 0.03172285 28.107014 7.907221e-48
```

Angiv t-teststørrelsen til test af hypotesen $H_0 : \beta = 1$, hvor β er hældningen i den lineære regressionsmodel.

A. -1.23

B. 28.11

C. -3.42

D. -11.56

E. 1.99

3.7 I en nyligt publiceret artikel af P.B. Hansen & M. Penkowa (2017) har man optalt, at 5 patienter fik infektion ud de 13 patienter som blev behandlet med bismuth. Tilsvarende var der 6 ud af 10 patienter som fik infektion, hvis de blev behandlet med placebo.

I artiklen angives en forkert p-værdi på 0.0001 for test af hypotesen om, at der er lige stor infektionsrisiko for de to grupper patienter. Angiv en korrekt p-værdi baseret på χ^2-teststørrelsen for test af denne hypotese (når der ikke ønskes anvendt kontinuitetskorrektion).

Bemærk: Ved besvarelsen må du se bort fra den advarsel (Warning) som R kommer med, fordi nogle af de forventede *celleantal* under hypotesen er mindre end 5.

A. 0.545

B. 0.215

C. 0.305

D. 0.812

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: