Danmarks Tekniske Universitet *Side 1 af 26 sider.*

Skriftlig prøve: 20. august 2017

Kursus navn og nr: **Introduktion til Statistik (02323 og 02402)**

Tilladte hjælpemidler: Alle

Dette sæt er besvaret af

_____	_____	_____
(studienummer)	(underskrift)	(bord nr)

Opgavesættet består af 30 spørgsmål af "multiple choice" typen fordelt på 9 opgaver. Besvarelserne af "multiple choice"spørgsmålene anføres i det i CampusNet uploadede svarark (på 6 seperate sider), med numrene på de svarmuligheder, du mener er de korrekte.

Der gives 5 point for et korrekt "multiple choice" svar og -1 for et ukorrekt svar. KUN følgende 5 svarmuligheder er gyldige: 1, 2, 3, 4 eller 5. Hvis et spørgsmål efterlades blankt eller andet type svar angives, tæller det ikke med i besvarelsen. Envidere, hvis mere end et svar angives, hvilket faktisk er teknisk muligt i online-systemet, så tæller det ikke med (dvs. giver "0 point"). Det antal point, der kræves for, at et sæt anses for tilfredstillende besvaret, afgøres endeligt ved censureringen.

Den endelige besvarelse af opgaverne gøres ved at udfylde og online-aflevere svararket via CampusNet. Skemaet her er KUN et nød-alternativ til dette (husk at angive dit studienummer på din besvarelse, hvis du afleverer skemaet).

Opgave	I.1	I.2	I.3	II.1	II.2	II.3	III.1	III.2	IV.1	IV.2
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

Opgave	IV.3	IV.4	IV.5	V.1	V.2	V.3	V.4	V.5	VI.1	VI.2
Spørgsmål	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Svar										

Opgave	VI.3	VI.4	VI.5	VII.1	VIII.1	VIII.2	VIII.3	IX.1	IX.2	IX.3
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Sættet består af 26 sider.

Fortsæt på side 2

Multiple choice opgaver: *Der gøres opmærksom på, at ideen med opgaverne er, at der er ét og kun ét rigtigt svar på de enkelte spørgsmål. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde.*

Opgave I

Et svømmeteam tager på en ugelang træningslejr med fokus på crawltræning. Før og efter laves en test hvor træneren tager hver svømmers tid på samme distance svømmet med crawl.

Tiderne er gemt i samme rækkefølge (for svømmerne) i følgende vektorer i R: `before` er tiderne før og `after` er tiderne efter træningslejren.

Følgende hypotese ønskes testet

$$\mu_{\text{efter}} - \mu_{\text{før}} = 0$$
$$\mu_{\text{efter}} - \mu_{\text{før}} \neq 0$$

hvor $\mu_{\text{før}}$ og μ_{efter} er middelværdien for hele holdets tid henholdvis før og efter træningslejren.

Spørgsmål I.1 (1)

Hvilket af følgende `R`-kald beregner korrekt p-værdien i en t-test for hypotesen?

1 ☐ `t.test(after, before, mu=0)`

2 ☐ `t.test(after, before, mu=-10)`

3 ☐ `t.test(after, before, mu=10)`

4 ☐ `t.test(after, mu=10)`

5 ☐ `t.test(after-before, mu=0)`

Spørgsmål I.2 (2)

Testens p-værdi blev beregnet til 0.00287. Kan nulhypotesen afvises på signifikansniveau $\alpha = 5\%$ (både konklusion og argument skal være korrekt)?

1 ☐ Ja, da p-værdien er under signifikansniveaet afvises nulhyposen

2 ☐ Nej, da p-værdien er under signifikansniveaet accepteres nulhyposen

3 ☐ Ja, da p-værdien er over signifikansniveaet afvises nulhyposen

4 ☐ Nej, da p-værdien er over signifikansniveaet accepteres nulhyposen

5 ☐ Der skal flere oplysninger til før konklusionen kan drages

Spørgsmål I.3 (3)

Hver dag på træningslejren trækkes der lod om hvem der skal vaske op. Der skal bruges 4 hver dag til at vaske op og der er 35 deltagere. Der er lige stor sandsynlighed for at blive trukket hver dag. Beregn sandsynligheden for at en tilfældigt udvalgt deltager ikke kommer til at vaske op på hele træningslejren, som strækker sig over 7 aftener hvor der skal vaskes op.

1 ☐ 1 - $\binom{7}{0} \cdot 0.144^0 \cdot (1 - 0.144)^{7-0} = 0.57$

2 ☐ $\binom{5}{2} \cdot 0.144^2 \cdot (1 - 0.144)^{5-2} = 0.09$

3 ☐ $\binom{7}{7} \cdot 0.798^7 \cdot (1 - 0.798)^{7-7} = 0.21$

4 ☐ $\binom{7}{7} \cdot 0.886^7 \cdot (1 - 0.886)^{7-7} = 0.43$

5 ☐ $\binom{5}{2} \cdot 0.886^2 \cdot (1 - 0.886)^{5-2} = 0.01$

Fortsæt på side 4

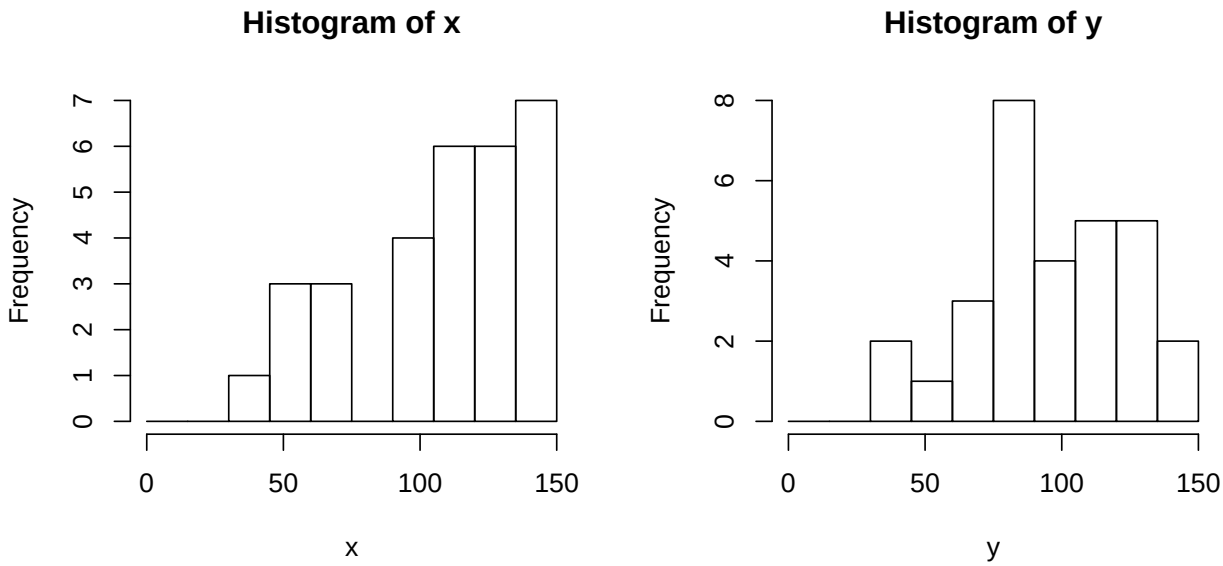
Opgave II

I forbindelse med eksamen i Introduktion til Statistisk ønskes det undersøgt om udenlandske studerende klarer sig godt. Scoren til eksamen regnes ud som et tal mellem -30 og 150, da der er 30 spørgsmål og der tælles -1 for et forkert svar og 5 point for et korrekt svar, samt der kun kan afgives et svar til hvert spørgsmål.

Der er udtaget to tilfældige stikprøver af scoren: en for udenlandske studerende (`x`) og en for danske studerende (`y`). Hver stikprøve har 30 observationer.

Spørgsmål II.1 (4)

For at vurdere den mest passende analyse, så plottes et histogram af hver stikprøve:



Hvad er den mest passende betragtning ud fra de givne oplysninger?

1 ☐ Der er ikke noget der indikerer at stikprøverne ikke kommer fra symmetriske fordelinger

2 ☐ Stikprøverne kan ikke antages at komme fra symmetriske fordelinger og dette understøttes af histogrammerne, specielt ses `x` at være højreskæv

3 ☐ Stikprøverne kan ikke antages at komme fra symmetriske fordelinger og dette understøttes af histogrammerne, specielt ses `x` at være venstreskæv

4 ☐ Populationerne hvorfra stikprøverne er taget begge antages at være exponentialfordelte

5 ☐ Populationerne hvorfra stikprøverne er taget begge antages at være normalfordelte

Spørgsmål II.2 (5)

Det besluttes at den bedste analyse er inkluderet i følgende R-kode:

```
## Number of simulations
k <- 10000
## Simulate each sample k times
simxsamples <- replicate(k, sample(x, replace=TRUE))
simysamples <- replicate(k, sample(y, replace=TRUE))
## Calculate the sample mean differences
simmeandifs <- apply(simxsamples,2,mean) - apply(simysamples,2,mean)
## Quantiles of the differences gives the CI
quantile(simmeandifs, c(0.005,0.995))

##  0.5% 99.5%
## -9.23 31.63

quantile(simmeandifs, c(0.025,0.975))

##   2.5%  97.5%
## -4.125 26.106

## CI for the median differences
simmediandifs <- apply(simxsamples,2,median) - apply(simysamples,2,median)
quantile(simmediandifs, c(0.005,0.995))

##   0.5%  99.5%
## -10.42  43.05

quantile(simmediandifs, c(0.025,0.975))

##   2.5%  97.5%
## -3.975 39.525
```

Hvilket af følgende udsagn er korrekt?

1 ☐ Der er udregnet ikke-parametriske bootstrap konfidensintervaller for forskelle mellem to populationer

2 ☐ Der er udregnet parametriske bootstrap konfidensintervaller for forskelle mellem to populationer

3 ☐ Der er udregnet konfidensintervaller for forskelle mellem to populationer med antagelse af normalfordeling

4 ☐ Der er udregnet konfidensintervaller for forskelle mellem to populationer med antagelse af eksponentielfordeling

5 ☐ Der er udregnet konfidensintervaller for forskelle mellem to populationer med antagelse af Poissonfordeling

Spørgsmål II.3 (6)

Følgende hypotese ønskes testet på signifikansniveau $\alpha = 5\%$

$$H_0 : q_{0.5,\mathrm{x}} = q_{0.5,\mathrm{y}}$$
$$H_1 : q_{0.5,\mathrm{x}} \neq q_{0.5,\mathrm{y}}$$

hvor $q_{0.5,\mathrm{x}}$ angiver 50% fraktilen for udenlandske studerende og $q_{0.5,\mathrm{y}}$ angiver 50% fraktilen for danske studerende.

Hvilket af følgende udsagn er korrekt (ikke alle udsagn giver nødvendigvis mening)?

1 ☐ H_0 afvises og det kan konkluderes at danske studerende klarer sig signifikant bedre end udenlandske på det angivne signifikansniveau

2 ☐ H_0 afvises og det kan konkluderes at udenlandske studerende klarer sig signifikant bedre end danske på det angivne signifikansniveau

3 ☐ H_0 afvises ikke og det er derfor ikke påvist at danske studerende klarer sig signikant anderledes end udenlandske på det angivne signifikansniveau

4 ☐ H_0 afvises ikke og det er derfor påvist at danske studerende klarer sig signifikant anderledes end udenlandske på det angivne signifikansniveau

5 ☐ Ingen af ovenstående udsagn er korrekte

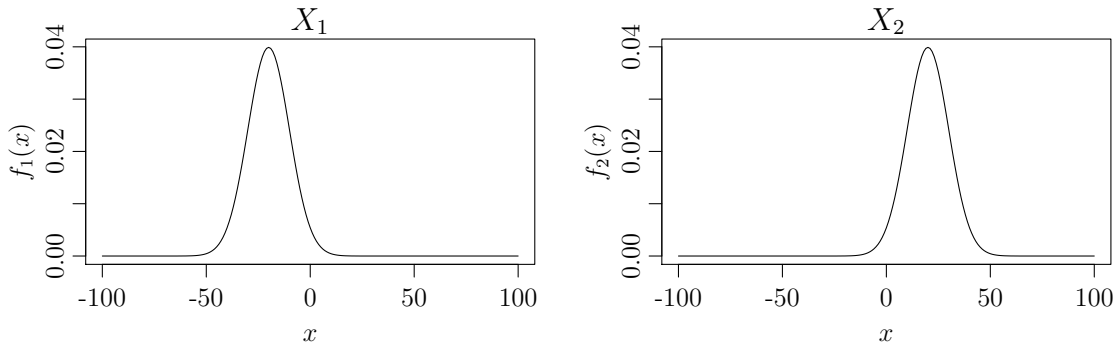
Fortsæt på side 7

Opgave III

Lad følgende to uafhængige stokastiske variabler være givet ved

$$X_1 \sim N(-20, 10^2) \quad \text{og} \quad X_2 \sim N(20, 10^2).$$

Deres tæthedsfunktioner (pdf) er da:

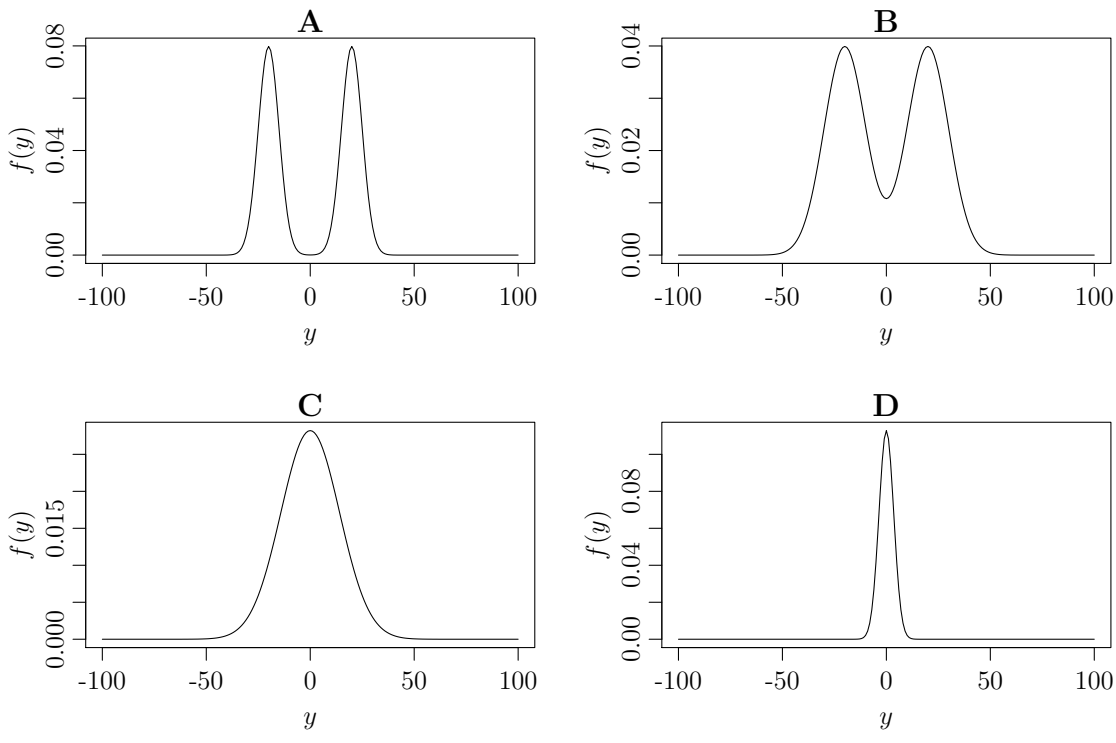


Spørgsmål III.1 (7)

Nu defineres en ny stokastisk variabel ved

$$Y = X_1 + X_2.$$

Hvilket af følgende plots er af tæthedsfunktionen (pdf) for Y?



1 ☐ Plot A

2 ☐ Plot B

3 ☐ Plot C

4 ☐ Plot D

5 ☐ Ingen af de viste plots kan være pdf for Y

Spørgsmål III.2 (8)

Hvis man antog at X_1 og X_2 repræsenterede hver deres population og man planlagde et forsøg hvor man ville teste for forskel i middelværdi med den sædvanligt anvendte ikke-parrede t-test. Hvad er den mindste stikprøvestørrelse $n = n_1 = n_2$ som skal tages fra hver population være for, på signifikansniveau $\alpha = 5\%$, at opnå en styrke af testen på mindst 99%?

1 ☐ $n = 4$ observationer i hver stikprøve

2 ☐ $n = 12$ observationer i hver stikprøve

3 ☐ $n = 38$ observationer i hver stikprøve

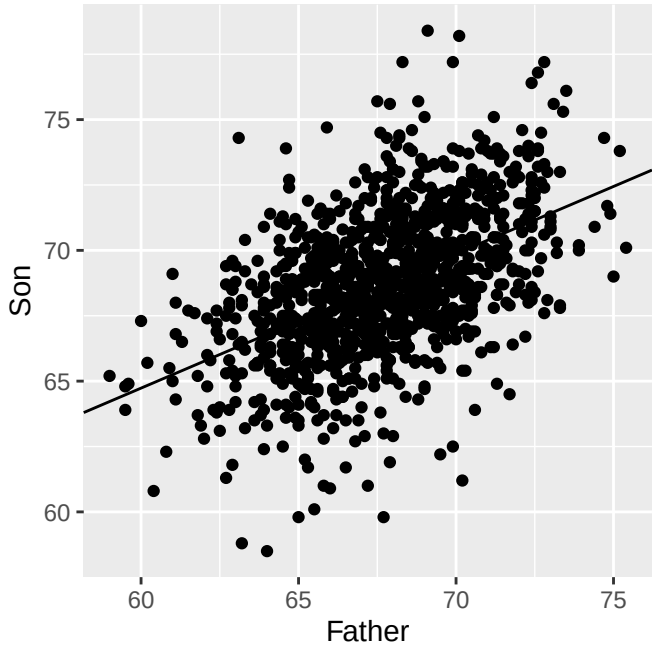
4 ☐ $n = 69$ observationer i hver stikprøve

5 ☐ $n = 248$ observationer i hver stikprøve

Fortsæt på side 9

Opgave IV

Nedenstående figur viser sammenhængen mellem højden på omkring 1000 fædre og deres sønner målt i tommer:



Den viste regressionslinje beskriver følgende model estimeret til observationerne

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.},$$

hvor Y_i er højden af den i'te søn og x_i er højden af den i'te far.

Spørgsmål IV.1 (9)

Hvilket af følgende udsagn er en korrekt beskrivelse af regressionslinjen?

1 ☐ Linjen beskriver et estimat af sønners middelhøjde som en funktion af deres fædres middelhøjde

2 ☐ Linjen beskriver et estimat af den lineære korrelation mellem den gennemsnitlige højde af far og søn

3 ☐ Linjen beskriver et estimat af fædres middelhøjde som funktion af højden af deres sønner

4 ☐ Linjen beskriver et estimat af sønners middelhøjde som en funktion af højden af deres fædre

5 ☐ Linjen beskriver et estimat af en søns højde som funktion af faderens højde

Spørgsmål IV.2 (10)

Man har valgt at analysere data med følgende R-kode, hvor `fs` er et data.frame med kolonnerne `Son` og `Father` der indeholder de observerede højder:

```
summary(fit <- lm(Son ~ Father, data=fs))
```

Hvilket giver følgende resultat hvor et par af tallene er erstattet af bogstaver:

```
Call:
lm(formula = Son ~ Father, data = fs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8910 -1.5361 -0.0092  1.6359  8.9894

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.89280          A   18.49   <2e-16 ***
Father       0.51401          B   19.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.438 on 1076 degrees of freedom
Multiple R-squared:  0.2512, Adjusted R-squared:  0.2505
F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Hvor stor en andel af variationen i højden af sønner er ikke forklaret af højden af fædrene?

1 ☐ ca. 25%

2 ☐ ca. 50%

3 ☐ ca. 75%

4 ☐ ca. 86.5%

5 ☐ ca. 66%

Spørgsmål IV.3 (11)

Hvad er estimatet af standardafvigelsen på estimatet af koefficienten for `Father`?

1 ☐ $\hat{\sigma}_{\beta_1} = 0.514/2.438 = 0.211$

2 ☐ $\hat{\sigma}_{\beta_1} = 2.438/1076 = 0.00227$

3 ☐ $\hat{\sigma}_{\beta_1} = 0.514 \cdot 19.00 = 9.77$

4 ☐ $\hat{\sigma}_{\beta_1} = 33.89/18.49 = 1.83$

5 ☐ $\hat{\sigma}_{\beta_1} = 0.514/19.0 = 0.027$

Spørgsmål IV.4 (12)

Givet følgende beregninger i R, hvad er da et 95% konfidensinterval for middelhøjden af sønner for fædre, der er 75 tommer høje?

```
mean(fs$Father); var(fs$Father)

## [1] 67.68683
## [1] 7.539566

mean(fs$Son); var(fs$Son)

## [1] 68.68423
## [1] 7.930949
```

1 ☐ $33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540 \cdot (1078-1)}}$

2 ☐ $33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540}}$

3 ☐ $33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438^2 \cdot \sqrt{\frac{1}{1078} + \frac{(75-67.687)^2}{7.540 \cdot (1078-1)}}$

4 ☐ $33.893 + 0.514 \cdot 75 \pm 1.96 \cdot 2.438 \cdot \sqrt{\frac{1}{1077} + \frac{(75-67.687)^2}{7.540 \cdot (1077-1)}}$

5 ☐ $33.893 + 0.514 \cdot 75 \pm 1.65 \cdot 2.438^2 \cdot \sqrt{\frac{1}{1077} + \frac{(75-67.687)^2}{7.540 \cdot (1077-1)}}$

Spørgsmål IV.5 (13)

Man får nu oplysninger om hver families månedlige indkomst og opstiller modellen

$$Y_i = \beta_0 + \beta_1 \cdot x_{1,i} + \beta_2 \cdot x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.},$$

hvor Y_i er højden af den i'te søn, $x_{1,i}$ er højden af den i'te far, og $x_{2,i}$ er indkomsten for den i'te familie.

Under følgende to antagelser:

- Rige familier spiser bedre og at bedre kost har en signifikant positiv effekt, der giver familiens sønner større vækst
- Der er uafhængighed mellem faderens højde og familiens indkomst

Hvad er da konsekvensen af at inddrage indkomst i modellen (ikke alle svarmuligheder er nødvendigvis meningsfulde)?

1 ☐ Inddragelse af indkomst i modellen vil bidrage til at mindske residualvariationen ($\hat{\sigma}^2$) og usikkerheden på regressionskoefficienten for faderens højde (β_1) vil blive reduceret

2 ☐ Inddragelse af indkomst i modellen vil bidrage til at mindske residualvariationen ($\hat{\sigma}^2$), men dette vil ikke have indflydelse på usikkerheden af regressionskoefficienten for faderens højde (β_1)

3 ☐ Da fædrenes højde er uafhængig af fædrenes indkomst vil inddragelse af indkomst i modellen ikke have nogen indflydelse på estimatet af β_1 eller usikkerheden på denne

4 ☐ Inddragelse af indkomst i modellen vil bruge endnu en frihedsgrad, så et konfidensinterval for β_1 må forventes at være bredere end hvis indkomst ikke var inkluderet i modellen

5 ☐ Man må forvente en høj grad af multikollinearitet imellem estimaterne af β_1 og β_2, så modellen alligevel skal reduceres til en simpel lineær regressionsmodel

Fortsæt på side 13

Opgave V

I mennesket findes en lang række forskellige genetisk bestemte blodtypesystemer. De mest kendte er nok AB0- og Rhesus-systemerne. Et andet blodtypesystem er det såkaldte MN-blodtypesystem, som bestemmes af et enkelt gen Glycophorin A (GPA). I GPA-genet findes to alleller M og N således at et menneske kan have genotypen (blodtypen) MM, MN, eller NN.

Fordelingen af blodtyper i MN-blodtypesystemet søges nu estimeret ud fra en stikprøve af frivillige studerende på to forskellige filippinske universiteter. Det ene universitet, University of the Philippines-Diliman her forkortet UPD, er landets største universitet hvor studerende kommer fra hele landet. Det andet universitet, Isabela State University her forkortet ISU, er et lille universitet hvor de studerende primært kommer fra lokalområdet. Følgende antalstabel angiver fordelingen af genotyperne blandt de studerende i stikprøverne ved de to universiteter:

Blodtype	UDP	ISU
MM	19	43
MN	15	7
NN	17	9

Spørgsmål V.1 (14)

Angiv χ^2 teststørrelsen og konklusionen på testen for sammenligningen af fordelingen fra de to universiteter (både teststørrelse og konklusion skal være korrekt).

1 ☐ Teststørrelsen er $\chi^2 = 14.15$, dens fordeling har 2 frihedsgrader og testen viser, at der er nogen evidens for en forskel på MN-blodtypefordelingen ved de to universiteter

2 ☐ Teststørrelsen er $\chi^2 = 14.15$, dens fordeling har 2 frihedsgrader og testen viser, at der er meget stærk evidens for en forskel på MN-blodtypefordelingen ved de to universiteter

3 ☐ Teststørrelsen er $\chi^2 = 3.76$, dens fordeling har 2 frihedsgrader og testen viser, at der ikke er evidens for en forskel på MN-blodtypefordelingen ved de to universiteter

4 ☐ Teststørrelsen er $\chi^2 = 4.57$, dens fordeling har 1 frihedsgrad og testen viser, at der er signifikant forskel på MN-blodtypefordelingen ved de to universiteter

5 ☐ Teststørrelsen er $\chi^2 = 3.76$, dens fordeling har 1 frihedsgrad og testen viser, at der er svag evidens for en forskel på MN-blodtypefordelingen ved de to universiteter

Spørgsmål V.2 (15)

En biologisk population siges at være i Hardy-Weinberg (HW) ligevægt hvis andelen af genotyperne kan skrives som

$$\begin{aligned} p_{\mathrm{MM}} &= p^2, \\ p_{\mathrm{MN}} &= 2pq, \\ p_{\mathrm{NN}} &= q^2. \end{aligned}$$

Hvor p og q kaldes "allelfrekvenserne" for henholdsvis M og N. De beregnes ved

$$\begin{aligned} p &= \frac{2 \cdot X_{\mathrm{MM}} + X_{\mathrm{MN}}}{2n}, \\ q &= \frac{2 \cdot X_{\mathrm{NN}} + X_{\mathrm{MN}}}{2n}, \end{aligned}$$

hvor X_{blodtype} er det observerede antal af den pågældende blodtype og n er stikprøvestørrelsen. Således bliver for UDP

$$p_{\mathrm{MN}} = 2 \cdot \frac{2 \cdot X_{\mathrm{MM}} + X_{\mathrm{MN}}}{2n} \cdot \frac{2 \cdot X_{\mathrm{NN}} + X_{\mathrm{MN}}}{2n} = 0.4992,$$

en beregnet andel af MN blodtype under HW-ligevægt.

En simpel test for om populationen på UDP ikke er i HW-ligevægt kan være af nulhypotesen

$$\begin{aligned} H_0 &: p_{\mathrm{MN,UDP}} = 0.4992 \\ H_1 &: p_{\mathrm{MN,UDP}} \neq 0.4992 \end{aligned}$$

dvs. om den observerede andel med MN blodtype på UDP ($p_{\mathrm{MN,UDP}}$) er lig andelen under HW-ligevægt.

Det ønskes testet om det kan afvises at genotyperne på UDP er i HW-ligevægt. Hvad bliver den sædvanligt anvendte teststørrelse for denne test?

1 ☐ Teststørrelsen er $\chi^2 = 2(1.99 + 4.30 + 2.32) = 17.2$

2 ☐ Teststørrelsen er $z_{\mathrm{obs}} = \frac{15-25.46}{\sqrt{25.46 \cdot (1-\frac{25.46}{51})}} = -2.93$

3 ☐ Teststørrelsen er $z_{\mathrm{obs}} = \frac{15-51}{\sqrt{51 \cdot (1-\frac{15}{51})}} = -6.00$

4 ☐ Teststørrelsen er $\chi^2 = (1.99^2 + 4.30^2 + 2.32^2)/2 = 13.9$

5 ☐ Teststørrelsen er $\chi^2 = 1.99 + 4.30 + 2.32 = 8.6$

Spørgsmål V.3 (16)

For en anden type test om HW-ligevægt fås teststørrelsen til $\chi^2 = 24.52$ og den skal følge en χ^2-fordeling med 1 frihedsgrad under nulhypotesen. Hvad bliver p-værdien og konklusionen på testen ved brug af et signifikansniveau på 0.001?

1 ☐ p-værdien er `pchisq(24.52, df=1)` ≈ 1 og hypotesen om HW-ligevægt kan ikke afvises

2 ☐ p-værdien er `1 - pchisq(24.52, df=1)` < 0.001 og hypotesen om HW-ligevægt kan ikke afvises

3 ☐ p-værdien er `1 - pchisq(24.52, df=1)`< 0.001 og hypotesen om HW-ligevægt afvises

4 ☐ p-værdien er `1 - pnorm(sqrt(24.52))` < 0.001 og hypotesen om HW-ligevægt kan ikke afvises

5 ☐ p-værdien er `1 - pnorm(sqrt(24.52))` < 0.001 og hypotesen om HW-ligevægt afvises

Spørgsmål V.4 (17)

Det er af teoretiske grunde blevet foreslået at frekvenserne af genotyperne MM og NN i den bagvedliggende population er ens og man ønsker nu på baggrund af observationerne fra UDP at belyse denne hypotese. Under antagelse af at andelene for MM og NN er uafhængige, er et 90% konfidensinterval for forskellen i andelen af MM og NN ($p_{\text{MM,UDP}} - p_{\text{NN,UDP}}$) givet ved:

1 ☐ $2/51 \pm 1.64\sqrt{\frac{19\cdot 32}{51^3} + \frac{17\cdot 34}{51^3}}$

2 ☐ $2/51 \pm 1.96\sqrt{\frac{19\cdot 32}{51^3} + \frac{17\cdot 34}{51^3}}$

3 ☐ $2/51 \pm 1.64\sqrt{\frac{19\cdot 32}{51^2} + \frac{17\cdot 34}{51^2}}$

4 ☐ $2/51 \pm 1.96\sqrt{\frac{19\cdot 32}{51^2} + \frac{17\cdot 34}{51^2}}$

5 ☐ $2/51 \pm 1.68\sqrt{\frac{19\cdot 32}{51^3} + \frac{17\cdot 34}{51^3}}$

Spørgsmål V.5 (18)

Hvad er den sædvanlige teststørrelse for testen af $p_{\text{MM,UDP}} = p_{\text{NN,UDP}}$?

1 ☐ $z_{\text{obs}} = \frac{2}{51\sqrt{\frac{19\cdot 32}{51^2} + \frac{17\cdot 34}{51^2}}}$

2 ☐ $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{19\cdot 32}{51^3} + \frac{17\cdot 34}{51^3}}}$

3 ☐ $z_{\text{obs}} = \frac{2}{51\sqrt{\frac{6}{17}\frac{11}{17}\frac{2}{52}}}$

4 ☐ $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{6}{17}\frac{6}{19}\frac{2}{52}}}$

5 ☐ $z_{\text{obs}} = \frac{2/51}{\sqrt{\frac{19\cdot 34}{51^3} + \frac{17\cdot 32}{51^3}}}$

Fortsæt på side 16

Opgave VI

En stikprøve med følgende 10 observationer er indsamlet:

```
x <- c(-1.63, -1.37, -1.21, -0.60, -0.36, -0.26, -0.18, 0.02, 0.29, 0.39)
```

Læg mærke til at stikprøven er blevet sorteret i ovenstående.

Stikprøvegennemsnittet og -standardafvigelsen er udregnet:

```
mean(x)

## [1] -0.491

sd(x)

## [1] 0.7003
```

Spørgsmål VI.1 (19)

Hvad er stikprøvevariansen?

1 ☐ $s^2 = 0.21$

2 ☐ $s^2 = 0.49$

3 ☐ $s^2 = 1.46$

4 ☐ $s^2 = 1.70$

5 ☐ $s^2 = 2.36$

Spørgsmål VI.2 (20)

Hvad er den første kvartil af stikprøven?

1 ☐ $Q_1 = -1.37$

2 ☐ $Q_1 = -1.29$

3 ☐ $Q_1 = -1.21$

4 ☐ $Q_1 = -0.91$

5 ☐ $Q_1 = -0.60$

Spørgsmål VI.3 (21)

Hvilket af følgende er et korrekt 95% konfidensinterval for middelværdien af populationen hvorfra stikprøven er taget?

1 ☐ $-0.491 \pm t_{0.975}\frac{0.490}{\sqrt{10}} = [-0.84, -0.14]$ hvor $t_{0.975} = 2.26$ er en fraktil i t-fordelingen med 9 frihedsgrader

2 ☐ $-0.491 \pm t_{0.95}\frac{0.700}{\sqrt{9}} = [-0.92, -0.64]$ hvor $t_{0.95} = 1.83$ er en fraktil i t-fordelingen med 9 frihedsgrader

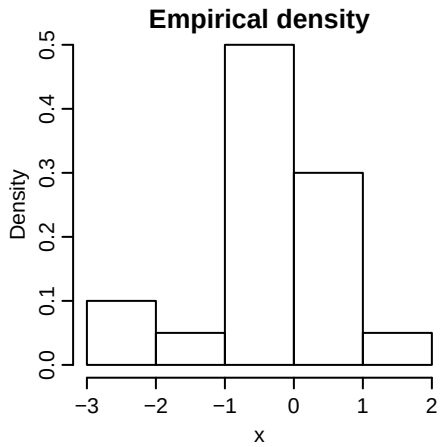
3 ☐ $-0.491 \pm t_{0.95}\frac{0.490}{\sqrt{9}} = [-0.79, -0.19]$ hvor $t_{0.95} = 1.83$ er en fraktil i t-fordelingen med 9 frihedsgrader

4 ☐ $-0.491 \pm t_{0.975}\frac{0.700}{10} = [-0.65, -0.33]$ hvor $t_{0.975} = 2.26$ er en fraktil i t-fordelingen med 9 frihedsgrader

5 ☐ $-0.491 \pm t_{0.975}\frac{0.700}{\sqrt{10}} = [-0.99, 0.01]$ hvor $t_{0.975} = 2.26$ er en fraktil i t-fordelingen med 9 frihedsgrader

Spørgsmål VI.4 (22)

En anden stikprøve er indsamlet og følgende empiriske tæthed er lavet for stikprøven:



Hvor stor er stikprøven, dvs. mange observationer n er der i stikprøven?

1 ☐ 20

2 ☐ 30

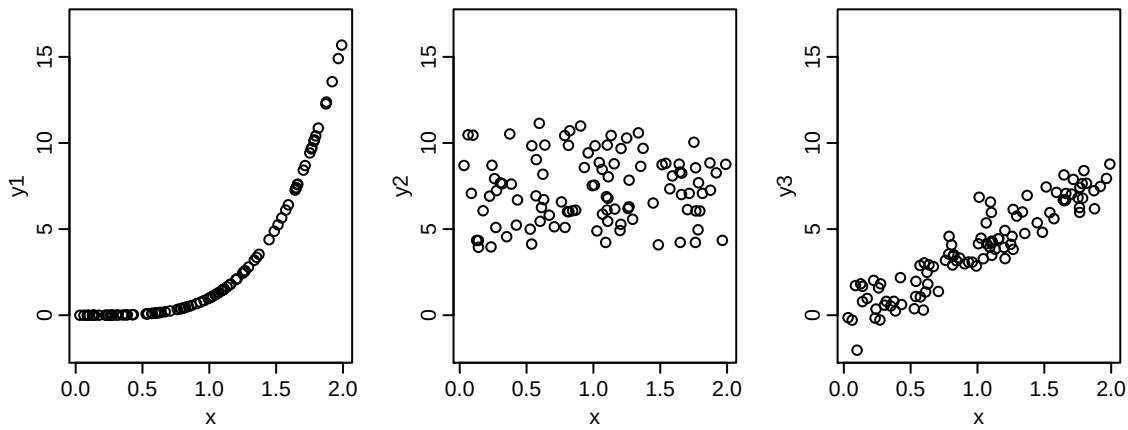
3 ☐ 100

4 ☐ 300

5 ☐ Dette spørgsmål kan ikke besvares med de givne oplysninger

Spørgsmål VI.5 (23)

Givet følgende 3 plots af sammenhørende værdier af x og y for stikprøver fra 3 forskellige populationer:



Hvilket af følgende er det eneste ikke meget usandsynlige udsagn om korrelationerne af de populationer som stikprøverne er taget fra?

1 ☐ $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0$ og $\rho_{XY_3} = 0.33$

2 ☐ $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0$ og $\rho_{XY_3} = -0.89$

3 ☐ $\rho_{XY_1} = 0$, $\rho_{XY_2} = 0.61$ og $\rho_{XY_3} = 0.91$

4 ☐ $\rho_{XY_1} = 0.87$, $\rho_{XY_2} = 0$ og $\rho_{XY_3} = 0.92$

5 ☐ $\rho_{XY_1} = 0.22$, $\rho_{XY_2} = 0$ og $\rho_{XY_3} = -0.34$

Fortsæt på side 19

Opgave VII

I en endelig population af N enheder med middelværdi $\mathrm{E}[Y] = \mu$ og varians $\mathrm{V}[Y] = \sigma^2$ betragter vi en stikprøve med n enheder Y_i, $i = 1, \ldots, n$. Hvis en stikprøven tages tilfældigt og uden tilbagelægning har gennemsnittet, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, variansen $\mathrm{V}(\bar{Y}) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}$. Interessen falder nu på totalen $\tau = \sum_{i=1}^{N} Y_i$, der kan estimeres ved $\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} Y_i$.

Spørgsmål VII.1 (24)

Hvad er variansen af estimatoren $\hat{\tau}$ dvs. $\mathrm{V}(\hat{\tau})$?

1 ☐ $\mathrm{V}(\hat{\tau}) = \frac{N^2}{n}\sigma^2$

2 ☐ $\mathrm{V}(\hat{\tau}) = N(N-n)\frac{\sigma^2}{n}$

3 ☐ $\mathrm{V}(\hat{\tau}) = \frac{N^2}{n^3}\sigma^2$

4 ☐ $\mathrm{V}(\hat{\tau}) = N^2(1-n)\sigma^2$

5 ☐ $\mathrm{V}(\hat{\tau}) = \frac{N}{n}\sigma^2$

Fortsæt på side 20

Opgave VIII

Op til 1970'erne var det i Finland kun tilladt at sælge og servere alkoholiske drikke i byerne og ikke i landområderne. Da man på dette tidspunkt ønskede at ophæve salgsbegrænsningerne på alkohol i landområderne opstod der bekymring om det ville lede til en øget rate af trafikulykker. Forud for ophævelsen udførte man derfor et pilotprojekt hvor man i fire landkommuner ekstraordinært gav lov til at sælge alkohol i butikkerne og i yderligere fire landkommuner ekstraordinært gav lov til at servere alkohol i restauranter og andre serveringssteder foruden at sælge alkohol i butikkerne. Fire andre landkommuner uden ekstraordinære tilladelser skulle agere kontrol. Data på antallet af trafikulykker fra de 12 udvalgte landkommuner igennem et år er præsenteret i nedenstående tabel:

Navn	Kontrol	Salg	SalgOgServering
	177	226	226
	225	196	229
	167	198	215
	176	206	188
Sum	745	826	858

og den valgte analyse er en ANOVA. Resultat er:

```
## Analysis of Variance Table
##
## Response: Accidents
##           Df Sum Sq Mean Sq F value Pr(>F)
## Treatment  A 1696.2  848.08     C     D
## Residuals  B 3670.7  407.86
```

Hvor `Treatment` er en factor variabel der grupperer landkommunerne i de tre grupper og `Accidents` er antallet af trafikulykker.

Spørgsmål VIII.1 (25)

For at undersøge om tilgængeligheden af alkohol har betydning for hyppigheden af trafikulykker ønsker man at sammenligne middelantallet af trafikulykker i de 3 grupper. Under antagelse af at variansen i antallet af trafikulykker er konstant mellem grupperne, hvad bliver da resultatet af testen af om der er forskel i middelantallet af trafikulykker mellem de 3 grupper på signifikansniveay $\alpha = 0.05$?

1 ☐ Teststørrelsen $F_{\text{obs}} = 1.232$ der under H_0 følger en F-fordeling med 3 og 8 frihedsgrader, hvilket giver en p-værdi på 0.360 og undersøgelsen giver derfor ikke grund til at tro at en lempelse af alkoholtilgængeligheden vil øge antallet af trafikulykker

2 ☐ Teststørrelsen $F_{\text{obs}} = 2.079$ der under H_0 følger en F-fordeling med 2 og 9 frihedsgrader, hvilket giver en p-værdi på 0.181 og undersøgelsen viser derfor at en lempelse af alkoholtilgængeligheden med sikkerhed ikke vil øge antallet af trafikulykker

3 ☐ Teststørrelsen $F_{\text{obs}} = 2.079$ der under H_0 følger en F-fordeling med 2 og 9 frihedsgrader, hvilket giver en p-værdi på 0.181 og undersøgelsen giver derfor ikke grund til at tro at en lempelse af alkoholtilgængeligheden vil øge antallet af trafikulykker

4 ☐ Teststørrelsen $F_{\text{obs}} = 4.324$ der under H_0 følger en F-fordeling med 2 og 9 frihedsgrader, hvilket giver en p-værdi på 0.0483 og undersøgelsen viser derfor at en lempelse af alkoholtilgængeligheden ændrer antallet af trafikulykker

5 ☐ Teststørrelsen $F_{\text{obs}} = 4.324$ der under H_0 følger en F-fordeling med 3 og 8 frihedsgrader, hvilket giver en p-værdi på 0.0434 og undersøgelsen viser derfor at en lempelse af alkoholtilgængeligheden ændrer antallet af trafikulykker

Spørgsmål VIII.2 (26)

Hvad er estimatet af afvigelsernes standardafvigelse?

1 ☐ $\hat{\sigma} = 1696.2/(12-1) = 154$

2 ☐ $\hat{\sigma} = \sqrt{1696.2/(3-1)} = 29.1$

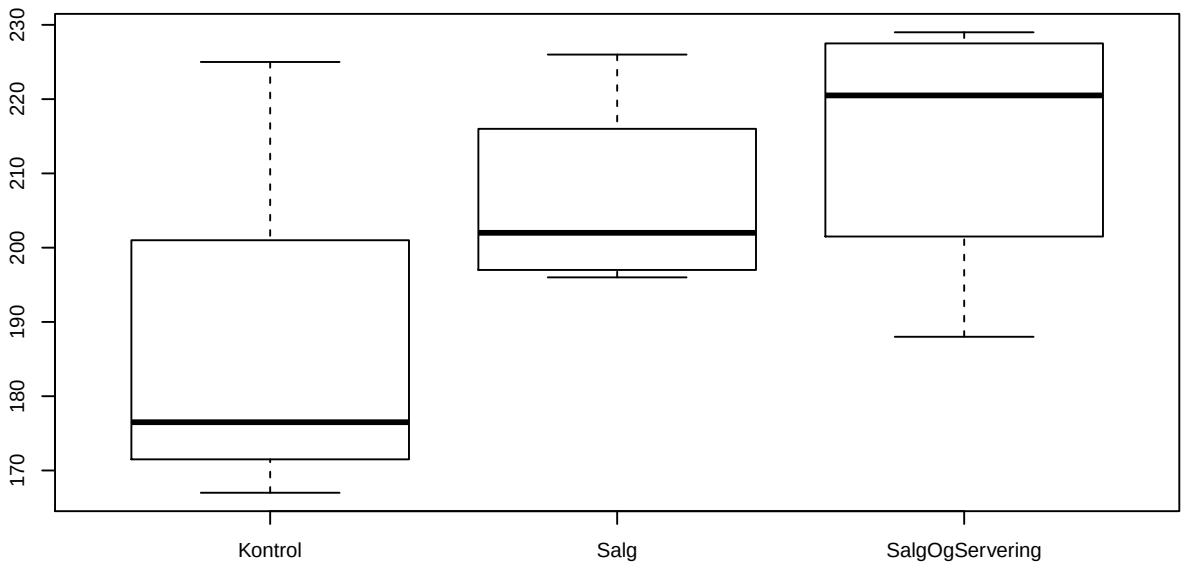
3 ☐ $\hat{\sigma} = \sqrt{1696.2/(12-1)} = 11.0$

4 ☐ $\hat{\sigma} = \sqrt{3670.7/(12-3)} = 20.2$

5 ☐ $\hat{\sigma} = 5367.1/(12-3)^2 = 66.3$

Spørgsmål VIII.3 (27)

Antagelsen om varianshomogenitet undersøges med følgende boxplots:



Hvilket af følgende udsagn er den mest korrekte slutning på denne baggrund (ikke alle udsagn er nødvendigvis meningsfulde)?

1 ☐ Taget det høje antal observationer i betragtning, så findes der ikke belæg for at antagelsen af varianshomogenitet ikke er opfyldt

2 ☐ Taget det høje antal observationer i betragtning, så findes der belæg for at antagelsen af varianshomogenitet ikke er opfyldt

3 ☐ Taget det lave antal observationer i betragtning, så findes der ikke belæg for at antagelsen af varianshomogenitet ikke er opfyldt

4 ☐ Taget det lave antal observationer i betragtning, så findes der belæg for at antagelsen af varianshomogenitet ikke er opfyldt

5 ☐ Der kan ikke tages stilling til antagelsen af varianshomogenitet på baggrund af de givne oplysninger

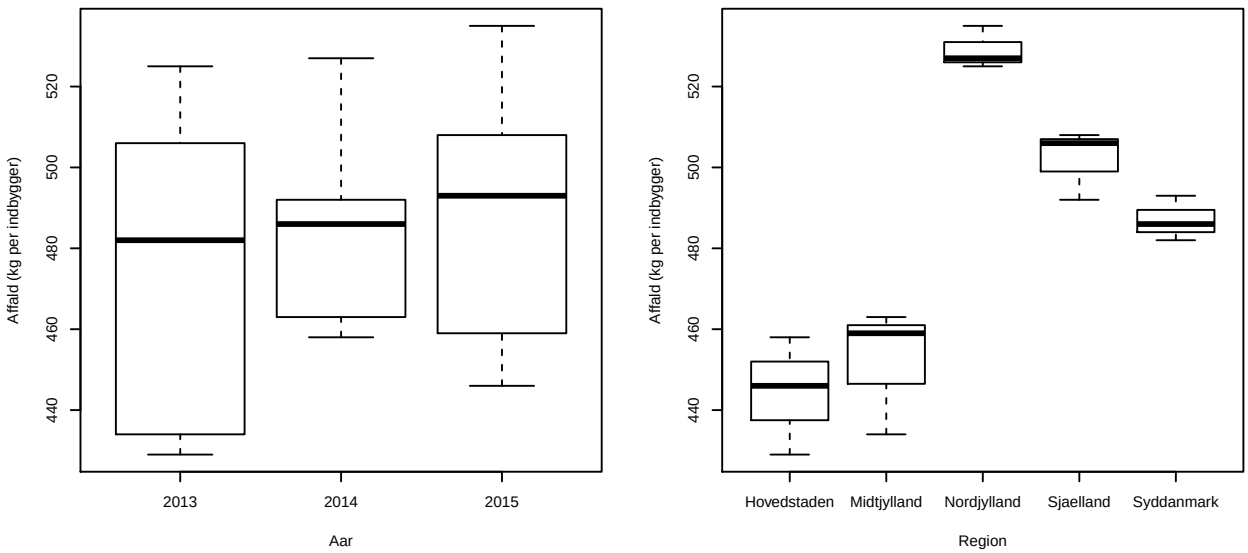
Fortsæt på side 23

Opgave IX

Miljø- og Fødevarestyrelsen indsamler hvert år data om affald i Danmark og udgiver en rapport med data og analyser. I Affaldsstatisk 2015 ([1]) kan man finde en opgørelse af affald (kg) per indbygger for årene 2013 til 2015 opdelt på regioner:

	Hovedstaden	Midtjylland	Nordjylland	Sjaelland	Syddanmark
2013	429	434	525	506	482
2014	458	463	527	492	486
2015	446	459	535	508	493

Følgende boxplot viser affald per borger opdelt på henholdsvis år og region:



Der er udført en 2-vejs ANOVA og resultatet er:

```
## Analysis of Variance Table
##
## Response: Affald
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Aar        2   463.3   231.7  2.5551    0.1386
## Region     4 14847.1  3711.8 40.9386 2.266e-05 ***
## Residuals  8   725.3    90.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[1] http://www2.mst.dk/Udgiv/publikationer/2017/05/978-87-93614-01-7.pdf

Spørgsmål IX.1 (28)

Hvilket af følgende udsagn er korrekt når der anvendes en signifikansniveau på $\alpha = 5\%$?

1 ☐ Udfra boxplottet kan det ses at der ikke kan påvises signifikant forskel i affald opdelt på år, hvilket også er konklusionen af ANOVA testen

2 ☐ Udfra boxplottet kan det ses at der ikke kan påvises signifikant forskel i affald opdelt på år, men fra ANOVA testen ses det, at der kan påvises signifikant forskel i affald opdelt på år

3 ☐ Udfra boxplottet kan man ikke afgøre om der kan påvises signifikant forskel i affald opdelt på år, men fra ANOVA testen ses det, at der ikke kan påvises signifikant forskel i affald opdelt på år

4 ☐ Udfra boxplottet kan man ikke afgøre om der kan påvises signifikant forskel i affald opdelt på år, men fra ANOVA testen ses det, at der kan påvises signifikant forskel i affald opdelt på år

5 ☐ Ingen af ovenstående udsagn er korrekte

Spørgsmål IX.2 (29)

Der er yderligere opgjort hvor stor en mængde affald per borger, der bliver sorteret i de fem regioner, og andelen af affald der er sorteret er beregnet opdelt på år og region. Der er udført en 2-vejs ANOVA på data med følgende resultat:

```
## Analysis of Variance Table
##
## Response: Andel
##           Df    Sum Sq   Mean Sq F value   Pr(>F)
## Aar        2 0.0109878 0.0054939  13.054 0.003026 **
## Region     4 0.0173773 0.0043443  10.323 0.003019 **
## Residuals  8 0.0033668 0.0004208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hvilken af følgende konklusioner er korrekt med anvendelse af et signifikansniveau på 5% (både argument og konklusion skal være korrekt)?

1 ☐ Da p-værdien > 0.05 for den relevante test, kan der ikke påvises en signifikant ændring i den sorterede andel mellem årene

2 ☐ Da $P(F > 13.054) < 0.05$ hvor F følger den relevante F-fordeling, kan der kan påvises en signifikant ændring i den sorterede andel mellem årene

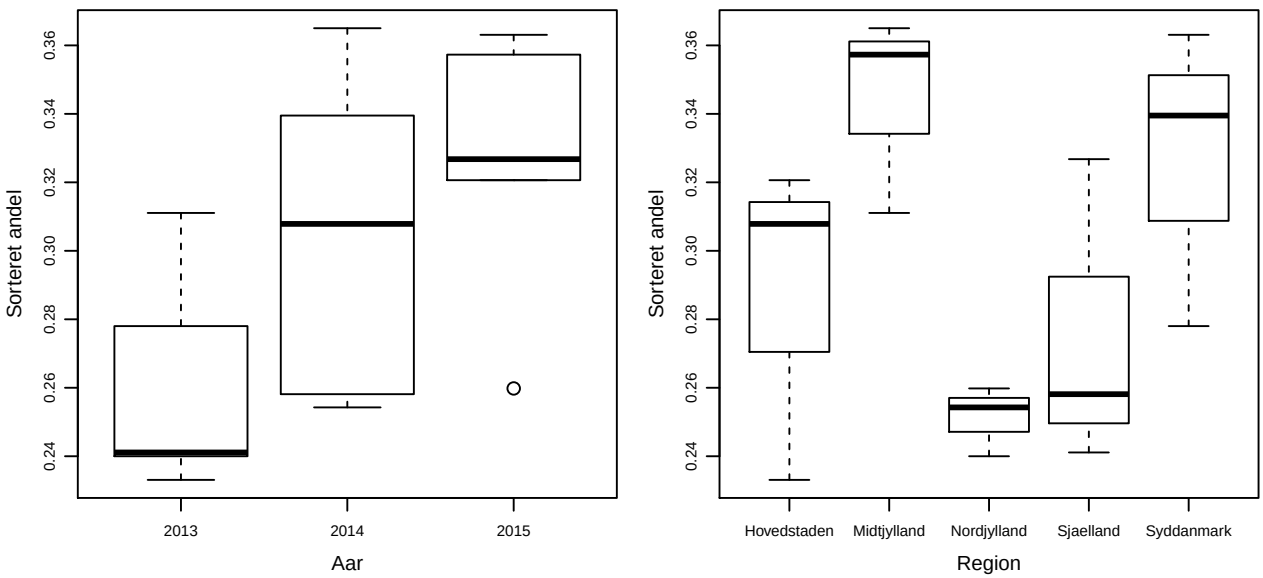
3 ☐ Da $P(T > 0.003) > 0.05$ hvor T følger den relevante t-fordeling, kan der ikke påvises en signifikant ændring i den sorterede andel mellem årene

4 ☐ Da $P(T < 10.323) < 0.05$ hvor T følger den relevante t-fordeling, kan der kan påvises en signifikant ændring i den sorterede andel mellem årene

5 ☐ Da $1 - P(T > 10.323) > 0.05$ hvor T følger den relevante t-fordeling, der kan ikke påvises en signifikant ændring i den sorterede andel mellem årene

Spørgsmål IX.3 (30)

Boxplottene der viser andelen af sorteret affald opdelt på år og opdelt på region er:



Det ses at der i 2015 er en observation som er væsentligt lavere en resten, den er ifølge det modificerede boxplot identificeret som en outlier.

Hvilket af følgende udsagn er ikke korrekt (Tip: Husk at der kun er en observation for hvert år for hver region)?

1 ☐ Den laveste observation i 2013 tilhører Hovedstaden

2 ☐ Den laveste observation i 2015 (dvs. outlieren) tilhører Nordjylland

3 ☐ Sjælland har hvert år haft en højere observation end Hovedstaden

4 ☐ Nordjylland har den laveste median

5 ☐ 75% fraktilen for 2014 er højere end 25% fraktilen for 2015

SÆTTET ER SLUT. God sensommer!