Danmarks Tekniske Universitet *Side 1 af 26 sider.*

Skriftlig prøve: 27. maj 2018

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

Tilladte hjælpemidler: Alle

Dette sæt er besvaret af

_____ _____ _____

(studienummer) (underskrift) (bord nr.)

Opgavesættet består af 30 spørgsmål af "multiple choice" typen, som er fordelt på 11 opgaver. For at besvare spørgsmålene skal du udfylde "multiple choice" svararket (6 seperate sider) på CampusNet med numrene på de svarmuligheder, som du mener er de rigtige.

Der gives 5 point for et korrekt "multiple choice" svar og -1 point for et forkert svar. KUN følgende 5 svarmuligheder er gyldige: 1, 2, 3, 4 eller 5. Hvis et spørgsmål efterlades blankt eller et ugyldigt svar angives, gives der 0 point for spørgsmålet. Endvidere, hvis mere end et svar angives til det samme spørgsmål, hvilket faktisk er teknisk muligt i online-systemet, gives der 0 point for spørgsmålet. Det antal point der kræves, for at opnå en bestemt karakter eller for at bestå eksamen afgøres endeligt ved censureringen.

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

Opgave	I.1	I.2	I.3	I.4	II.1	II.2	II.3	III.1	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	VI.1	VI.2	VI.3	VII.1	VII.2
Spørgsmål	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Svar										

Opgave	VII.3	VIII.1	IX.1	IX.2	X.1	X.2	X.3	XI.1	XI.2	XI.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 der i hvert spørgsmål er én og kun én svarmulighed, som er rigtig. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde. Husk altid at afrunde dit eget resultat til antallet af decimalerne givet i svarmulighederne før du vælger et svar.*

Opgave I

For at undersøge downloadhastigheden ved en arbejdsstation har man målt downloadtiderne (i sekunder) for 53 filer af forskellige størrelser (målt i MB). Downloadtiderne er gemt i vektoren `time` i R, mens de tilsvarende filstørrelser er gemt i vektoren `size`. Desuden er følgende kode blevet kørt i R:

```
logtime <- log(time)
model1 <- lm(logtime ~ size)
summary(model1)

##
## Call:
## lm(formula = logtime ~ size)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.571 -0.673  0.132  0.745  2.190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.109      0.482   -0.23     0.82
## size           0.127      0.023    5.50  1.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.979 on 51 degrees of freedom
## Multiple R-squared:  0.372,Adjusted R-squared:  0.36
## F-statistic: 30.2 on 1 and 51 DF,  p-value: 1.25e-06
```

Den statistiske model givet ved `model1` er en lineær regressionsmodel af formen

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,,$$

hvor ε_i, $i = 1, \ldots, 53$, antages uafhængige og identisk $N(0, \sigma^2)$-fordelte.

Spørgsmål I.1 (1)

Angiv den afhængige variabel og den forklarende variabel i modellen givet ved `model1`.

1 ☐ Downloadtid er den afhængige variabel. Filstørrelse er den forklarende variabel.

2 ☐ Filstørrelse er den afhængige variabel. Logaritmen til downloadtid er den forklarende variabel.

3 ☐ Downloadtid er den forklarende variabel. Filstørrelse er den afhængige variabel.

4 ☐ Filstørrelse er den forklarende variabel. Logaritmen til downloadtid er den afhængige variabel.

5 ☐ Da (logaritmen til) downloadtid afhænger af filstørrelse, er både filstørrelse og logaritmen til downloadtid afhængige variable. Der er ingen forklarende variabel i modellen.

Spørgsmål I.2 (2)

Angiv estimater for parametrene i modellen givet ved `model1`.

1 ☐ $\hat{\beta}_0 = 0.127$, $\hat{\beta}_1 = -0.109$, $\hat{\sigma} = 0.979$

2 ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.979^2$

3 ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.372$

4 ☐ $\hat{\beta}_0 = 0.127$, $\hat{\beta}_1 = -0.109$, $\hat{\sigma} = 0.979^2$

5 ☐ $\hat{\beta}_0 = -0.109$, $\hat{\beta}_1 = 0.127$, $\hat{\sigma} = 0.979$

Spørgsmål I.3 (3)

Følgende kode er også blevet kørt i R:

```
mean(size)

## [1] 20.088

(53-1)*var(size)

## [1] 1807.6
```

Angiv med udgangspunkt i modellen givet ved `model1` et 90% prædiktionsinterval for logaritmen til downloadtiden for en fil på 17 MB.

1 ☐ $-0.109 + 0.127 \cdot 17 \pm 2.0076 \cdot 0.979 \cdot \sqrt{1 + \frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

2 ☐ $17 \cdot 0.127 - 0.109 \pm 0.979 \cdot 1.6753 \cdot \sqrt{1 + \frac{1}{53} + \frac{(20.088-17)^2}{1807.6}}$

3 ☐ $-0.109 + 0.127 \cdot 17 \pm 1.6753 \cdot \sqrt{0.979} \cdot \sqrt{1 + \frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

4 ☐ $-0.109 + 17 \cdot 0.127 \pm 1.6753 \cdot 0.979 \cdot \sqrt{\frac{1}{53} + \frac{(17-20.088)^2}{1807.6}}$

5 ☐ $17 \cdot 0.127 + 0.109 \pm 0.979 \cdot 2.0076 \cdot \sqrt{1 + \frac{1}{53} + \frac{(20.088-17)^2}{1807.6}}$

Spørgsmål I.4 (4)

Med udgangspunkt i modellen givet ved `model1` ønsker man at undersøge, om der er en signifikant lineær sammenhæng mellem logaritmen til downloadtid og filstørrelse. Formuler den tilsvarende statistiske nulhypotese om at der ikke er sammenhæng mellem de to variable.

1 ☐ $H_0 : \hat{\beta}_1 = 0$

2 ☐ $H_0 : \beta_1 \neq \beta_0$

3 ☐ $H_0 : \beta_1 = 0$

4 ☐ $H_0 : \beta_1 \neq 0$

5 ☐ $H_0 : \hat{\beta}_1 \neq 0$

Fortsæt på side 5

Opgave II

Togvogne i en mine skal lastes med store klippestykker, som er sprængt fri. En maskine har først sorteret stykkerne i to bunker ud fra deres størrelse og vægt. Vægtene af klippestykkerne antages uafhængige og normalfordelte, således at vægten af et tilfældigt udtrukket stykke fra bunke 1 kan repræsenteres ved $X_1 \sim N(20, 5^2)$ kg og fra bunke 2 ved $X_2 \sim N(50, 10^2)$ kg.

Spørgsmål II.1 (5)

Hvad er sandsynligheden for, at et tilfældigt udtrukket stykke fra bunke 1 vejer over 25 kg?

1 ☐ 15.9 %

2 ☐ 84.1 %

3 ☐ 42.1 %

4 ☐ 57.9 %

5 ☐ Ingen af de angivne værdier.

Spørgsmål II.2 (6)

Vælg et korrekt udsagn:

Der er 20% sandsynlighed for, at et tilfældigt udtrukket stykke fra bunke 2 er tungere end

1 ☐ 41.6 kg.

2 ☐ 52.5 kg.

3 ☐ 58.4 kg.

4 ☐ 67.4 kg.

5 ☐ 134 kg.

Spørgsmål II.3 (7)

Hvis togvognene bliver lastet for tungt, er man nødt til at stoppe driften. En manuel kran skal da fjerne klippestykker fra den overlastede vogn, og dette er dyrt.

Der er to robotkraner, som laster togvognene. Den ene tager fra bunke 1, og den anden tager fra bunke 2. Hvis hver kran tager 10 stykker fra sin bunke, og alle 20 stykker bliver lagt i den samme (tomme) togvogn, hvad er så sandsynligheden for, at den samlede last i denne togvogn nu overstiger 800 kg?

1 ☐ Den samlede vægt er $Y \sim N(700, 15625)$, så $P(Y > 800) = 21.2\%$.

2 ☐ Den samlede vægt er $Y \sim N(700, 12500)$, så $P(Y > 800) = 18.6\%$.

3 ☐ Den samlede vægt er $Y \sim N(700, 2500)$, så $P(Y > 800) = 2.28\%$.

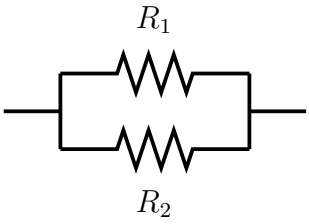
4 ☐ Den samlede vægt er $Y \sim N(700, 1250)$, så $P(Y > 800) = 0.234\%$.

5 ☐ Den samlede vægt er $Y \sim N(700, 225)$, så $P(Y > 800) = 1.31 \cdot 10^{-9}\%$.

Fortsæt på side 7

Opgave III

Modstandene i det elektriske kredsløb



er estimeret til $\hat{R}_1 = 2\ \Omega$ og $\hat{R}_2 = 3\ \Omega$, med estimeret standardafvigelse for estimatorerne (*standard error*) henholdsvis $\hat{\sigma}_{\hat{R}_1} = 0.2$ og $\hat{\sigma}_{\hat{R}_2} = 0.5$. R_1 og R_2 kan antages at være uafhængige og normalfordelte.

Den samlede modstand igennem kredsløbet er givet ved

$$R = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2}}$$

Spørgsmål III.1 (8)

Bestem ved hjælp af simulation et 95% konfidensinterval for den samlede modstand R. Følgende R-kode skal benyttes for at få det angivne resultat (og efter denne kode er kørt, kan man nøjes med ét yderligere funktionskald i R for at få resultatet):

```
set.seed(7643)
k <- 10000
R1 <- rnorm(k, mean = 2, sd = 0.2)
R2 <- rnorm(k, mean = 3, sd = 0.5)
R <- 1/(1/R1 + 1/R2)
```

1 ☐ [1.11, 1.29]

2 ☐ [0.96, 1.40]

3 ☐ [0.92, 1.47]

4 ☐ [0.82, 1.58]

5 ☐ [0.72, 1.68]

Fortsæt på side 8

Opgave IV

Studerendes valg af videregående uddannelse har stor politisk opmærksomhed. Nedenstående tabel er baseret på tal fra Universiteternes Statistiske Beredskab og indeholder antal optagne fordelt på hovedområder for udvalgte år (samt række- og søjlesummer i kursiv).

	Y2012	Y2016	Y2017	*Sum*
Hum	7966	7297	6691	*21954*
Samf	10173	10253	10006	*30432*
Sund	2789	3137	3157	*9083*
TekNat	8551	10130	10339	*29020*
Sum	*29479*	*30817*	*30193*	*90489*

Spørgsmål IV.1 (9)

Baseret på tallene fra 2017 ønsker man at teste en hypotese om, at den andel optagne som har valgt en teknisk eller naturvidenskabelig bacheloruddannelse (TekNat) er 32.0%. Den relevante teststørrelse for denne hypotese, som antages approksimativt at følge en standardnormalfordeling, er

1 ☐ $(10253 - 0.68 \cdot 30817)/\sqrt{30817 \cdot 0.32 \cdot 0.68} = -130.70$

2 ☐ $(10339 - 0.32 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.32} = 12.18$

3 ☐ $(10130 - 0.32 \cdot 30817)/\sqrt{30817 \cdot 0.32 \cdot 0.68} = 3.28$

4 ☐ $(10339 - 0.32 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.68} = 8.36$

5 ☐ $(10339 - 0.68 \cdot 30193)/\sqrt{30193 \cdot 0.32 \cdot 0.32} = -183.30$

Spørgsmål IV.2 (10)

Det ønskes også undersøgt, om der fra 2016 til 2017 er sket en ændring i den andel, som optages på de humanistiske uddannelser (Hum).

Fire forskellige test er udført i R med følgende kode:

```
  prop.test(x = 6691, n = 30193, p = 7297/30817, correct = FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  6691 out of 30193, null probability 7297/30817
## X-squared = 38.5, df = 1, p-value = 5.5e-10
## alternative hypothesis: true p is not equal to 0.23678
## 95 percent confidence interval:
```

```
##  0.21696 0.22633
## sample estimates:
##       p
## 0.22161
```

```
prop.test(x = c(7297, 6691), c(30817, 30193), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(7297, 6691) out of c(30817, 30193)
## X-squared = 19.9, df = 1, p-value = 8.2e-06
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.0085083 0.0218461
## sample estimates:
##  prop 1  prop 2
## 0.23678 0.22161
```

```
binom.test(x = 6691, n = 30193, p = 7297/30817)
```

```
##
##  Exact binomial test
##
## data:  6691 and 30193
## number of successes = 6691, number of trials = 30193, p-value = 4.3e-10
## alternative hypothesis: true probability of success is not equal to 0.23678
## 95 percent confidence interval:
##  0.21693 0.22634
## sample estimates:
## probability of success
##                0.22161
```

```
binom.test(x = 6691, n = 30193, p = (6691+7297)/(30193+30817))
```

```
##
##  Exact binomial test
##
## data:  6691 and 30193
## number of successes = 6691, number of trials = 30193, p-value = 0.0015
## alternative hypothesis: true probability of success is not equal to 0.22927
## 95 percent confidence interval:
##  0.21693 0.22634
```

```
## sample estimates:
## probability of success
##               0.22161
```

Hvilken kodelinje udfører det ønskede test?

1 ☐ `binom.test(x = 6691, n = 30193, p = 7297/30817)`

2 ☐ `prop.test(x = c(7297, 6691), c(30817, 30193), correct = FALSE)`

3 ☐ `binom.test(x = 6691, n = 30193, p = (6691+7297)/(30193+30817))`

4 ☐ `prop.test(x = 6691, n = 30193, p = 7297/30817, correct = FALSE)`

5 ☐ Ingen af de fire linjers kode undersøger den ønskede hypotese.

Spørgsmål IV.3 (11)

Det ønskes også undersøgt ved et hypotesetest, om fordelingen af optagne mellem de forskellige hovedområder har ændret sig henover de tre årgange, for hvilke der er data. Antallet af frihedsgrader i den relevante fordeling af teststørrelsen er:

1 ☐ 3

2 ☐ 4

3 ☐ 6

4 ☐ 12

5 ☐ 20

Spørgsmål IV.4 (12)

Under antagelse om uafhængighed mellem årgang og hovedområde estimeres det forventede antal, som blev optaget på TekNat i 2017, til

1 ☐ 10339

2 ☐ $29020 \cdot 30193/90489 = 9683$

3 ☐ $(8551 + 10130 + 10339)/3 = 9673$

4 ☐ $10339 \cdot 29020/30193 = 9937$

5 ☐ $10339 \cdot 30193/29020 = 10757$

Spørgsmål IV.5 (13)

Som det næste skridt i at teste om fordelingen mellem hovedområder har ændret sig over tid oplyses følgende: Teststørrelsen er beregnet til 314.5. Der benyttes et signifikansniveau på $\alpha = 0.05$. I den fordeling, der benyttes til at vurdere teststørrelsen, er 0.95 og 0.975 fraktilerne henholdsvis 12.59 og 14.45. Hvad kan da konkluderes? (Både konklusion og argumentation skal være korrekt).

1 ☐ På basis af de oplyste tal kan det ikke bedømmes statistisk, om fordelingen mellem hovedområder har ændret sig.

2 ☐ Fordelingen mellem hovedområder har ikke ændret sig signifikant, da teststørrelsen er større end den oplyste 0.95 fraktil.

3 ☐ Fordelingen mellem hovedområder har ikke ændret sig signifikant, da teststørrelsen er større end den oplyste 0.975 fraktil.

4 ☐ Fordelingen mellem hovedområder har ændret sig signifikant, da der under nulhypotesen er 95% sandsynlighed for at få en teststørrelse, der er større end 12.59.

5 ☐ Fordelingen mellem hovedområder har ændret sig signifikant, da teststørrelsen er større end den oplyste 0.95 fraktil.

Fortsæt på side 12

Opgave V

Et forsøg er blevet udført med henblik på at undersøge holdbarheden af en bestemt type medicin. I alt 26 identiske, uåbnede flasker medicin med samme produktionsdato blev benyttet i forsøget. Halvdelen af flaskerne blev opbevaret ved stuetemperatur (21 °C), den anden halvdel ved køleskabstemperatur (5 °C). Efter 90 dage blev flaskerne åbnet, og indholdet af det aktive lægemiddelstof (i mg/ml) i hver flaske blev målt. Resultaterne blev indlæst i R i to vektorer, `hightemp` (målingerne fra flaskerne opbevaret ved 21 °C) og `lowtemp` (målingerne fra flaskerne opbevaret ved 5 °C).

Desuden er følgende kode blevet kørt i R:

```
t.test(lowtemp, hightemp)

##
##  Welch Two Sample t-test
##
## data:  lowtemp and hightemp
## t = 5.3, df = 24, p-value = 2e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.4316 3.2592
## sample estimates:
## mean of x mean of y
##    7.4397    5.0943
```

Spørgsmål V.1 (14)

Hvad kan konkluderes, når der benyttes et signifikansniveau på $\alpha = 0.05$?

1 ☐ Middelindhold af det aktive stof er signifikant større ved opbevaring ved køleskabstemperatur end ved stuetemperatur. Forskellen estimeres til at være 2.35 mg/ml.

2 ☐ Det ses ud fra p-værdien, at der ikke er signifikant forskel på middelindhold af det aktive stof mellem flasker opbevaret ved hhv. stue- og køleskabstemperatur.

3 ☐ Middelindhold af det aktive stof er signifikant større ved opbevaring ved køleskabstemperatur end ved stuetemperatur. Forskellen estimeres til at være 5.30 mg/ml.

4 ☐ Middelindhold af det aktive stof er signifikant mindre ved opbevaring ved køleskabstemperatur end ved stuetemperatur. Forskellen estimeres til at være 2.35 mg/ml.

5 ☐ 95% konfidensintervallet indeholder 2.35. Derfor er der ikke signifikant forskel på middelindhold af det aktive stof mellem flasker opbevaret ved hhv. stue- og køleskabstemperatur.

Spørgsmål V.2 (15)

Et 99% konfidensinterval for forskellen i middelindhold af det aktive stof mellem flasker opbevaret ved køleskabs- og stuetemperatur kan bestemmes på følgende vis:

1 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.7969} \cdot 2.0639 = [1.67, 3.02]$

2 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.4922} \cdot 2.7969 = [1.32, 3.37]$

3 ☐ $\left[1.4316 \cdot \frac{2.7969}{2.0639}, 3.2592 \cdot \frac{2.7969}{2.0639}\right] = [1.94, 4.42]$

4 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.0639} \cdot 2.7969 = [1.11, 3.58]$

5 ☐ $2.3454 \pm \frac{3.2592-2.3454}{2.0639} \cdot 2.4922 = [1.24, 3.45]$

Fortsæt på side 14

Opgave VI

I en produktionsproces går det nogle gange galt, og komponenten må kasseres efter en nærmere undersøgelse. Erfaringsmæssigt ved man, at der vil være 20% sandsynlighed for at en komponent må kasseres. Vurderingen ("behold" eller "kassér") for en given komponent er uafhængig af vurderingen for de andre komponenter. Man kan regne med, at der hver dag produceres 20 komponenter.

Spørgsmål VI.1 (16)

Hvad er sandsynligheden for, at der på en tilfældigt udvalgt dag ikke er nogen komponenter, som skal kasseres?

1 ☐ Antallet af komponenter, som må kasseres, er hypergeometrisk fordelt, og sandsynligheden er 0.

2 ☐ Antallet af komponenter, som må kasseres, er binomialfordelt, og sandsynligheden er 0.0115.

3 ☐ Antallet af komponenter, som må kasseres, er binomialfordelt, og sandsynligheden er 0.0576.

4 ☐ Antallet af komponenter, som må kasseres, er hypergeometrisk fordelt, og sandsynligheden er 0.0692.

5 ☐ Antallet af komponenter, som må kasseres, er binomialfordelt, og sandsynligheden er 0.630.

Spørgsmål VI.2 (17)

En simulering af antallet af kasserede komponenter per dag er blevet kørt. Lad X_i betegne antal kasserede komponenter på en tilfældigt udvalgt dag i. Der er simuleret en stikprøve på $n = 20$ værdier, som betegnes med x_i, $i = 1, \ldots, 20$. Hvilket af følgende udsagn er det eneste, som kan være korrekt?

1 ☐ Det forventede antal komponenter, som skal kasseres i løbet af én dag, er $\mu_X = 4.5$. Stikprøvegennemsnittet for de simulerede værdier var $\hat{\mu}_X = 4.3$.

2 ☐ Det forventede antal komponenter, som skal kasseres i løbet af én dag, er $\mu_X = 4$. Stikprøvegennemsnittet for de simulerede værdier var $\hat{\mu}_X = 3.7$.

3 ☐ Det forventede antal komponenter, som skal kasseres i løbet af én dag, er $\mu_X = 5$. Stikprøvegennemsnittet for de simulerede værdier var $\hat{\mu}_X = 4.9$.

4 ☐ Det forventede antal komponenter, som skal kasseres i løbet af én dag, er $\mu_X = 2$. Stikprøvegennemsnittet for de simulerede værdier var $\hat{\mu}_X = 2.2$.

5 ☐ Det forventede antal komponenter, som skal kasseres i løbet af én dag, er $\mu_X = 4.25$. Stikprøvegennemsnittet for de simulerede værdier var $\hat{\mu}_X = 4.25$.

Spørgsmål VI.3 (18)

Der planlægges et forsøg for at estimere andelen af komponenter, som skal kasseres.

Hvor mange dage skal eksperimentet køre for at opnå et 95% konfidensinterval for andelen af komponenter, der skal kasseres, som har en forventet bredde på 5%-point?

1 ☐ $n = (1.96 \cdot 0.2/0.05)^2 = 61.5$, dvs. 4 dage.

2 ☐ $n = 0.16 \cdot (1.96/0.05)^2 = 246$, dvs. 13 dage.

3 ☐ $n = 0.16 \cdot (1.96/0.025)^2 = 983$, dvs. 50 dage.

4 ☐ $n = 0.2 \cdot (1.96/0.025)^2 = 1229$, dvs. 62 dage.

5 ☐ $n = (1.96 \cdot 0.2/0.025)^2 = 3934$, dvs. 197 dage.

Fortsæt på side 16

Opgave VII

På en fabrik der producerer slagtøjsinstrumenter og tilbehør produceres der også trommestikker. Med henblik på kvalitetskontrol har man målt længden (i cm) af 20 tilfældigt udvalgte trommestikker. Disse længder er indlæst i vektoren `length` i R. Længderne kan antages at være uafhængige og normalfordelte med middelværdi μ og varians σ^2.

Spørgsmål VII.1 (19)

Følgende kommandoer er kørt i R:

```
sum(length)

## [1] 793.1

var(length)

## [1] 0.01099
```

Angiv et udtryk for et 95% konfidensinterval for trommestikkernes middellængde.

1 ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{0.01099}{\sqrt{20}}$

2 ☐ $793.1 \pm 2.086 \cdot \frac{0.01099}{\sqrt{20}}$

3 ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{\sqrt{0.01099}}{\sqrt{20}}$

4 ☐ $\frac{793.1}{20} \pm 2.086 \cdot \frac{\sqrt{0.01099}}{\sqrt{20}}$

5 ☐ $\frac{793.1}{20} \pm 2.093 \cdot \frac{0.01099}{\sqrt{19}}$

Spørgsmål VII.2 (20)

Der udføres et t-test med henblik på at undersøge, om middellængden af trommestikkerne kan antages at være 39.60 cm. Det oplyses, at den observerede t-teststørrelse bliver $t_{\text{obs}} = 2.38$. Hvilken af følgende konklusioner er den eneste rigtige?

1 ☐ Middellængden er signifikant forskellig fra 39.60 cm, når der benyttes et signifikansniveau på $\alpha = 0.01$.

2 ☐ Da p-værdien er større end 0.05, forkastes nulhypotesen om at middellængden er 39.60 cm, uanset om signifikansniveauet $\alpha = 0.05$ eller $\alpha = 0.01$ benyttes.

3 ☐ Da p-værdien er mindre end 0.05, accepteres nulhypotesen om at middellængden er 39.60 cm ved et signifikansniveau på $\alpha = 0.05$.

4 ☐ Middellængden er signifikant forskellig fra 39.60 cm, når der benyttes et signifikansniveau på $\alpha = 0.05$.

5 ☐ Da p-værdien er større end 0.05, accepteres nulhypotesen om at middellængden er 39.60 cm, uanset om signifikansniveauet $\alpha = 0.05$ eller $\alpha = 0.01$ benyttes.

Spørgsmål VII.3 (21)

I en senere kvalitetskontrol ønsker man at udtage en ny stikprøve. Denne skal være så stor, at en forskel på 0.5 mm mellem trommestikkernes middellængde og den ønskede middellængde på 39.60 cm kan opdages med styrke 90%, ved brug af signifikansniveau $\alpha = 0.01$. Her benyttes 0.11 som bud på populationens standardafvigelse. Benyt `power.t.test`-funktionen i R til at bestemme den nødvendige stikprøvestørrelse.

1 ☐ $n = 20$

2 ☐ $n = 22$

3 ☐ $n = 146$

4 ☐ $n = 53$

5 ☐ $n = 76$

Fortsæt på side 18

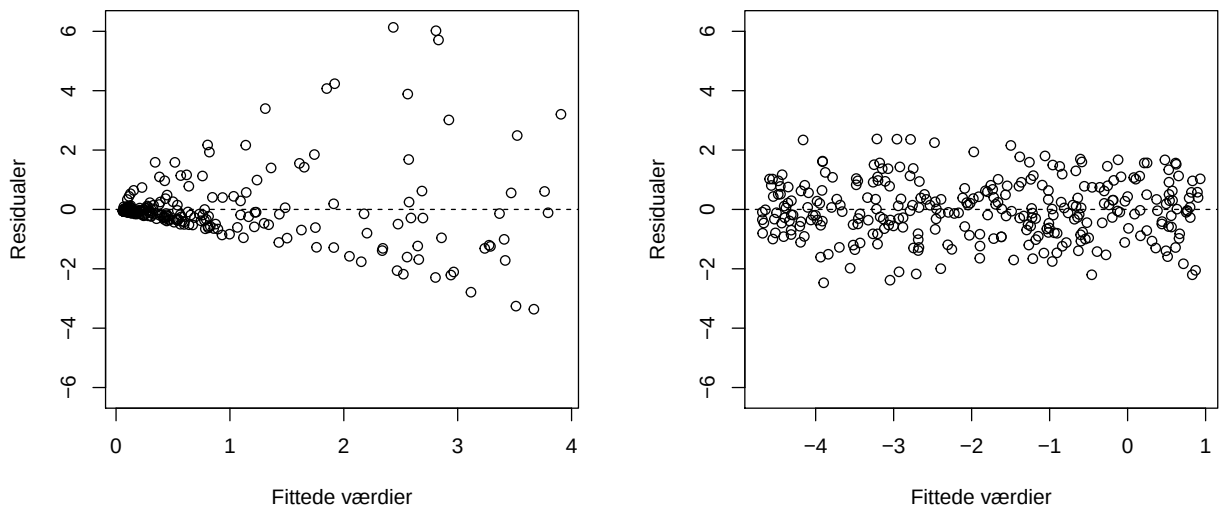
Opgave VIII

To numeriske variable er indlæst i R som `x` og `y`. Man ønsker at beskrive sammenhængen mellem disse to variable ved hjælp af en lineær regressionsmodel. Til dette formål er to forskellige lineære regressionsmodeller blevet estimeret i R, se R-koden nedenfor.

```
logx = log(x)
logy = log(y)
model1 <- lm(y ~ x)
model2 <- lm(logy ~ logx)
```

Spørgsmål VIII.1 (22)

Nedenfor ses plots af residualerne mod de fittede værdier for henholdsvis `model1` (til venstre) og `model2` (til højre). Vil man på baggrund af disse plots foretrække at analysere data ved hjælp af den statistiske model givet ved `model1` eller den givet ved `model2`? (Både argument og konklusion skal være rigtige).



1 ☐ Der er tydelige lineære sammenhænge mellem residualerne og de fittede værdier i plottet til venstre, mens der ikke ses nogen lineær sammenhæng i plottet til højre. Man vil derfor foretrække `model1`.

2 ☐ Antagelsen om varianshomogenitet er tydeligvis ikke opfyldt for `model2`, mens antagelsen virker rimelig for `model1`. Man vil derfor foretrække `model1`.

3 ☐ I plottet til venstre ligger rigtig mange fittede værdier i intervallet [0,1], mens de er spredt bedre ud over hele x-aksen i plottet til højre. Man vil derfor foretrække `model2`.

4 ☐ Antagelsen om varianshomogenitet er tydeligvis ikke opfyldt for `model1`, mens antagelsen virker rimelig for `model2`. Man vil derfor foretrække `model2`.

5 ☐ Residualerne i figuren til højre er tydeligvis ligefordelte på et interval omkring 0 (og således ikke normalfordelte), mens residualerne i figuren til venstre godt kunne være normalfordelte. Man vil derfor foretrække `model1`.

Fortsæt på side 20

Opgave IX

Der er mange faktorer, som påvirker indeklimaet i en bygning. Et af de mest anvendte mål for kvaliteten af indeklimaet er niveauet af CO_2. Hvis der ikke er tilstrækkelig ventilation, så bliver CO_2-niveauet for højt, og det går ud over bl.a. koncentrationsevnen. I nye bygninger med undervisningslokaler må CO_2-niveauet ikke overstige 1000 ppm – i udeluft er der omkring 400 ppm (før den industrielle revolution var det omkring 280 ppm!).

I en undersøgelse af indeklima i undervisningslokaler blev stikprøver af CO_2-niveau målt i to forskellige lokaler. Begge stikprøver består af én-times gennemsnitsværdier, som er målt henover en periode på 2 måneder. Kun værdier, hvor der er mennesker til stede i lokalerne, er taget med i stikprøverne. Observationerne for henholdsvis lokale 1 og lokale 2 er indlæst i R i vektorerne `room1CO2` og `room2CO2`.

Spørgsmål IX.1 (23)

Følgende kode er kørt i R:

```
length(room1CO2)

## [1] 304

length(room2CO2)

## [1] 252

sum((room1CO2 - mean(room1CO2))^2)

## [1] 131606104

(length(room2CO2)-1)*var(room2CO2)

## [1] 12775276
```

Bestem stikprøvestandardafvigelsen for henholdsvis lokale 1 (s_1) og lokale 2 (s_2).

1 ☐ $s_1 = 659.0475$ og $s_2 = 225.6048$

2 ☐ $s_1 = 434343.6$ og $s_2 = 50897.51$

3 ☐ $s_1 = 657.9626$ og $s_2 = 225.1567$

4 ☐ $s_1 = 11471.97$ og $s_2 = 3574.252$

5 ☐ Ingen af de fire svarmuligheder ovenfor kan være rigtig.

Spørgsmål IX.2 (24)

Desuden er følgende kode blevet kørt i R:

```
Q3 <- function(x){ quantile(x, 0.75) }

simSamples1 <- replicate(10000, sample(room1CO2, replace = TRUE))
simSamples2 <- replicate(10000, sample(room2CO2, replace = TRUE))

simQ3s1 <- apply(simSamples1, 2, Q3)
simQ3s2 <- apply(simSamples2, 2, Q3)
simQ3sdiff <- simQ3s1 - simQ3s2

quantile(simQ3s1, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##       0%     2.5%       5%      95%    97.5%     100%
## 1417.896 1562.332 1583.146 1833.104 1838.021 1953.792

quantile(simQ3s2, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##       0%     2.5%       5%      95%    97.5%     100%
## 772.0833 827.5000 831.1042 916.4583 920.5000 966.6667

quantile(simQ3sdiff, c(0, 0.025, 0.05, 0.95, 0.975, 1))

##        0%      2.5%        5%       95%     97.5%      100%
##  534.6458  685.8297  712.4562  976.1042  991.6250 1093.4167
```

Bestem ud fra dette R-output et 95% konfidensinterval for forskellen mellem 0.75 fraktilerne for CO_2-niveauet i lokale 1 og 2.

1 ☐ $[916, 1833]$

2 ☐ $[828, 921]$

3 ☐ $[686, 992]$

4 ☐ $[1556 - 828, 1838 - 921] = [728, 917]$

5 ☐ $[828, 1838]$

Fortsæt på side 22

Opgave X

Tabellen herunder viser det gennemsnitlige høstudbytte (målt i hkg/ha) for 5 afgrøder (Crop 1-5) i Danmark i årene 2014-2017.

	2014	2015	2016	2017	*Average*
Crop 1	79	80	73	83	*78.75*
Crop 2	46	48	47	52	*48.25*
Crop 3	64	63	57	66	*62.50*
Crop 4	66	68	62	68	*66.00*
Crop 5	57	60	55	58	*57.50*
Average	*62.40*	*63.80*	*58.80*	*65.40*	*62.60*

Udover række- og søjlegennemsnittene angivet i tabellen (i kursiv) oplyses det, at $SS(\text{Year}) = 118.8$ og $SST = 2172.8$.

I denne opgave betragtes de 20 gennemsnitlige høstudbytter i tabellen som observationer fra 20 forskellige tilfældigt udvalgte marker. Det antages, at der indenfor hvert år ikke er forskel på det forventede udbytte fra de fem afgrøder. Analysen skal derfor udføres, som om at der er sået samme afgrøde på alle markerne (og informationen om afgrødetype skal således ikke bruges i opgaven). Vi benytter en model af formen

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

med $\sum \alpha_i = 0$, hvor $\varepsilon_{ij} \sim N(0, \sigma^2)$ og uafhængige.

Spørgsmål X.1 (25)

Lad α_1 beskrive effekten af år 2014 på det forventede høstudbytte. Angiv estimatet for α_1.

1 ☐ $\hat{\alpha}_1 = 78.75 - 62.40 = 16.35$

2 ☐ $\hat{\alpha}_1 = 62.40 - 62.60 = -0.20$

3 ☐ $\hat{\alpha}_1 = 78.75$

4 ☐ $\hat{\alpha}_1 = 62.60$

5 ☐ $\hat{\alpha}_1 = 62.40$

Spørgsmål X.2 (26)

Sæt signifikansniveauet til $\alpha = 0.05$. Angiv den kritiske værdi for det sædvanlige test der udføres for at undersøge, om det forventede høstudbytte er forskelligt mellem årene.

1 ☐ 3.73

2 ☐ 3.24

3 ☐ 26.30

4 ☐ 3.49

5 ☐ 2.96

Spørgsmål X.3 (27)

Angiv estimatet for σ^2.

1 ☐ $\hat{\sigma}^2 = 2054$

2 ☐ $\hat{\sigma}^2 = 3.058$

3 ☐ $\hat{\sigma}^2 = 36.7$

4 ☐ $\hat{\sigma}^2 = 29.7$

5 ☐ $\hat{\sigma}^2 = 128.375$

Fortsæt på side 24

Opgave XI

Datasættet brugt i denne opgave er det samme som i den forrige opgave. Her skal det dog benyttes, at data beskriver det gennemsnitlige høstudbytte (målt i hkg/ha) for 5 forskellige afgrøder (Crop 1-5) i Danmark i årene 2014-2017.

	2014	2015	2016	2017	*Average*
Crop 1	79	80	73	83	*78.75*
Crop 2	46	48	47	52	*48.25*
Crop 3	64	63	57	66	*62.50*
Crop 4	66	68	62	68	*66.00*
Crop 5	57	60	55	58	*57.50*
Average	*62.40*	*63.80*	*58.80*	*65.40*	*62.60*

Udover række- og søjlegennemsnittene angivet i tabellen (i kursiv) oplyses det, at $SS(\text{Crop}) = 2017.3$, $SS(\text{Year}) = 118.8$ og $SST = 2172.8$. De gennemsnitlige høstudbytter antages at være udfald af normalfordelte stokastiske variable.

Spørgsmål XI.1 (28)

Et centralt spørgsmål er, om der er sket en udvikling i udbyttet over tid. For at undersøge dette ønsker man at teste, om der er statistisk signifikant forskel mellem de fire år, idet der tages hensyn til variationen mellem de forskellige afgrøder. Angiv p-værdi og konklusionen på signifikansniveau $\alpha = 0.05$ for det sædvanlige test.

1 ☐ Der er ikke signifikant forskel på udbyttet mellem årene, da $p = 0.82 > 0.05$.

2 ☐ Der er signifikant forskel på udbyttet mellem årene, da $p = 0.00045 < 0.05$.

3 ☐ Der er ikke signifikant forskel på udbyttet mellem årene, da $p = 2 \cdot 10^{-10} < 0.05$.

4 ☐ Der er signifikant forskel på udbyttet mellem årene, da $p = 0.82 > 0.05$.

5 ☐ Der er signifikant forskel på udbyttet mellem årene, da $p = 8 \cdot 10^{-7} < 0.05$.

Spørgsmål XI.2 (29)

Man har på forhånd besluttet, at forskellen mellem udbyttet i år 2014 og år 2017 skal ses som et udtryk for den overordnede udvikling. Benyt signifikansniveauet $\alpha = 0.05$ og test hypotesen om, at der ikke er forskel mellem de to år. Hvad er konklusionen? (Både konklusion og argument skal være korrekt).

1 ☐ Der er signifikant forskel på udbyttet i de to år, da $\frac{65.4-62.4}{62.6} < 0.05$.

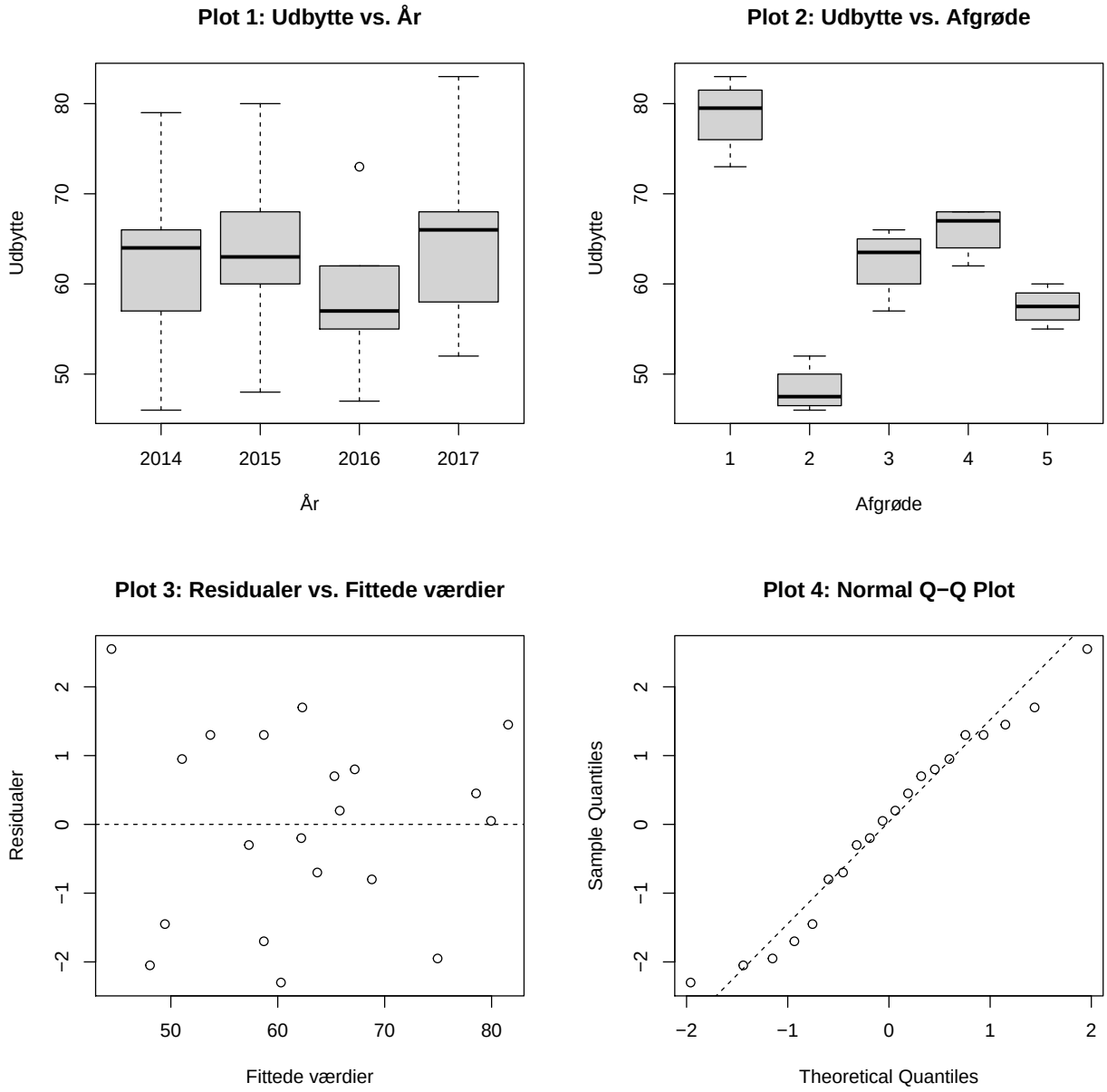
2 ☐ Der er ikke signifikant forskel på udbyttet i de to år, da $\frac{118.8}{2017.3} > 0.05$.

3 ☐ Der er ikke signifikant forskel på udbyttet i de to år, da $\frac{3}{\sqrt{36.7}} < 2.18$.

4 ☐ Der er signifikant forskel på udbyttet i de to år, da $\frac{-3}{1.11} < -2.18$.

5 ☐ Der er signifikant forskel på udbyttet i de to år, da $\frac{3}{36.7} > 0.05$.

Som en del af modelkontrollen har man lavet forskellige plots i figuren herunder.



Spørgsmål XI.3 (30)

Baseret på figuren, hvilket af følgende udsagn er da korrekt?

1 ☐ Da variationen indenfor år er meget større end variationen indenfor afgrøder (Plot 1 og

Plot 2), er antagelsen om varianshomogenitet tydeligvis ikke opfyldt.

2 ☐ Normalfordelingsantagelsen er tydeligvis ikke opfyldt (Plot 3), mens antagelsen om varianshomogenitet tydeligvis er opfyldt (Plot 4).

3 ☐ Normalfordelingsantagelsen ser ud til at være opfyldt (Plot 4), og det samme gælder antagelsen om varianshomogenitet (Plot 3).

4 ☐ Udbyttet kan ikke være forskelligt for de fem afgrøder, da residualerne varierer tilfældigt omkring den vandrette linje gennem $y = 0$ (Plot 2 og Plot 3).

5 ☐ Der kan umuligt være forskel på udbyttet i de forskellige år (Plot 1), men normalfordelingsantagelsen er tydeligvis opfyldt (Plot 3).

Eksamenssættet er slut. God sommer!