Danmarks Tekniske Universitet *Side 1 af 23 sider.*

Skriftlig prøve: 17. maj 2020

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

Varighed: 4 timer

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 10 opgaver. For at besvare spørgsmålene skal du udfylde "multiple choice" svararket (6 separate 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 laves 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	II.1	II.2	II.3	III.1	III.2	IV.1	IV.2	IV.3
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

Opgave	IV.4	IV.5	V.1	V.2	V.3	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	VII.4	VII.5	VII.6	VIII.1	VIII.2	IX.1	IX.2	IX.3	X.1
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Eksamenssættet består af 23 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 decimaler givet i svarmulighederne før du vælger et svar. Husk også, at der kan forekomme små afvigelser mellem resultatet af bogens formler og tilsvarende indbyggede funktioner i R.*

Opgave I

Elektriske komponenters egenskaber er ikke præcist som specificeret, f.eks. varierer modstanden på modstandskomponenter, således at køber man en modstandskomponent så er modstanden igennem den ikke præcist det angivne. I forbindelse med produktionen af elektriske kredsløb er det af stor interesse ikke at få for stor variation i kvaliteten af det samlede kredsløb. Som eksempel kan modstanden igennem to parallelt forbundede modstande beregnes ved

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

hvor R_1 er modstanden gennem den ene og R_2 gennem den anden modstand. Modstanden måles i ohm.

Antag at $R_1 \sim N(4, 0.2)$ og $R_2 \sim N(2, 0.2)$.

Spørgsmål I.1 (1)

Man køber 100 R_1 modstande - som kan antages uafhængige af hinanden. Hvad er sandsynligheden for at ingen af disse har en modstand under 3 ohm?

1 ☐ 1.27%

2 ☐ 2.78%

3 ☐ 13.9%

4 ☐ 27.9%

5 ☐ 42.4%

Spørgsmål I.2 (2)

Beregn et estimat af standardafvigelsen af den samlede modstand R (svaret er afrundet til to betydende cifre, tip: hvis du bruger simulering, så husk at gentage nok gange så resultat bliver stabilt)?

1 ☐ 0.026

2 ☐ 0.094

3 ☐ 0.16

4 ☐ 0.21

5 ☐ 0.44

Fortsæt på side 4

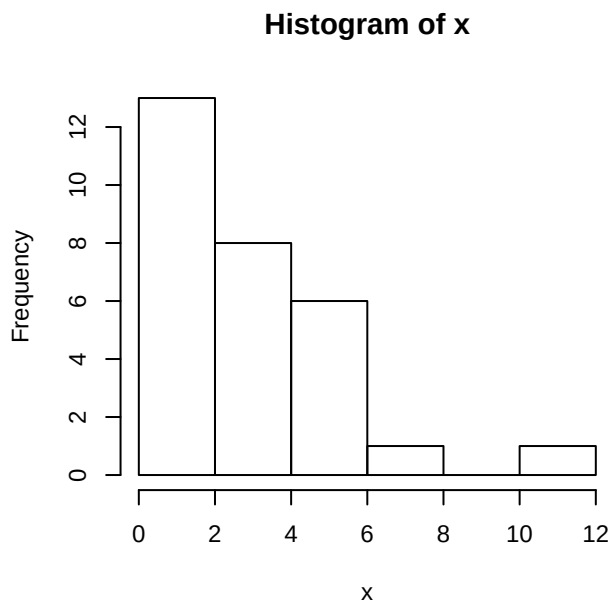
Opgave II

I et computersystem bruges en optimeringsrutine og beregningstiderne for denne rutine ønskes undersøgt. Tiderne er målt i timer og indlæst i R med følgende kode:

```
x <- c(1.6, 2, 3.4, 4, 2.1, 0.6, 0.4, 0.4, 6, 0.4, 4.9, 2, 2, 4.6, 0.5,
       3.4, 7.2, 10.5, 3.2, 1.3, 5.7, 1.9, 2.6, 2.5, 4.4, 1.8, 3.9, 6, 0.9)
```

Spørgsmål II.1 (3)

Man ønsker at vurdere hvilken fordeling udfaldene i stikprøven kunne stamme fra og man har derfor lavet nedenstående histogram af observationerne i `x`:



Vurder på baggrund af de givne oplysninger hvilken af følgende fordelinger, der, med størst rimelighed, kan antages at have genereret udfaldene i stikprøven?

1 ☐ En normal fordeling

2 ☐ En Poisson fordeling

3 ☐ En exponentiel fordeling

4 ☐ En t-fordeling

5 ☐ En binomial-fordeling

Spørgsmål II.2 (4)

Hvad er et estimatet af middelværdien og standardafvigelsen af beregningstiderne?

1 ☐ $\hat{\mu} = 2.53$ og $\hat{\sigma} = 1.66$

2 ☐ $\hat{\mu} = 3.36$ og $\hat{\sigma} = 0.48$

3 ☐ $\hat{\mu} = 3.11$ og $\hat{\sigma} = 2.37$

4 ☐ $\hat{\mu} = 1.98$ og $\hat{\sigma} = 5.63$

5 ☐ $\hat{\mu} = 3.96$ og $\hat{\sigma} = 2.81$

Spørgsmål II.3 (5)

Man ønsker at give en garanti for at beregningstiden er under et vist niveau. Man vil derfor beregne et konfidensinterval for 90% fraktilen og der defineres en funktion i R til beregning af denne:

```
q90 <- function(x){ quantile(x, prob=0.9, type=2) }
```

Hvilken at følgende R koder beregner et 95% procent ikke-parametrisk bootstrap konfidensinterval for 90% fraktilen af beregningstiderne?

1 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.05, 0.95))
```

2 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.025, 0.975))
```

3 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.05, 0.95))
```

4 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.1, 0.90))
```

5 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, q90)
quantile(simmeans, c(0.025, 0.975))
```

Fortsæt på side 6

Opgave III

En NGO har 15 telefonopringere ansat til at skaffe nye medlemmer. Lad X repræsentere det antal medlemmer, en enkelt opringer får på en arbejdsdag. Antallet af nye medlemmer, hver opringer får på en dag, kan antages at være uafhængigt af hinanden. Fra erfaring ved man, at en god model er, at X følger en binomialfordeling, hvor sandsynligheden for, at et nyt medlem fåes ved en opringning, er sat til 7%. Det antages, at opringerne kan nå 120 opkald på en dag.

Spørgsmål III.1 (6)

Hvad er sandsynligheden for, at en opkalder på en enkelt dag skaffer over 5 nye medlemmer?

1 ☐ 0.12

2 ☐ 0.85

3 ☐ 0.45

4 ☐ 0.17

5 ☐ 0.96

Spørgsmål III.2 (7)

Hvis Y angiver det samlede antal nye medlemmer de 15 opringere kan skaffe på en dag, hvad er da er middelværdi og varians for Y?

1 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 10.8$

2 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 41.9$

3 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 43.5$

4 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 102.4$

5 ☐ $\mathrm{E}(Y) = 126$ og $\mathrm{V}(Y) = 117.2$

Fortsæt på side 7

Opgave IV

I Danmark diskuterer man ofte om det har været en god eller en dårlig sommer. I tabellen herunder er angivet gennemsnitstemperaturer for månederne maj-september i årene 2014-2018:

	2014	2015	2016	2017	2018	Average
May	11.7	9.7	12.9	12.0	15.0	12.26
June	14.9	12.7	16.0	14.7	16.5	14.96
July	19.5	15.5	16.4	15.5	19.2	17.22
August	16.0	17.4	16.1	16.0	17.5	16.60
September	14.6	13.2	16.2	13.3	14.1	14.28
Average	15.34	13.70	15.52	14.30	16.46	15.01

For at undersøge om der er forskel mellem årene, har man kørt følgende R-kode, hvor `month` er månederne Maj-September, `year` er indikatorer for årene 2014-2018 og `temp` er gennemsnitstemperaturerne:

```
anova(fit <- lm(temp ~ year + month))

## Analysis of Variance Table
##
## Response: temp
##           Df Sum Sq Mean Sq F value  Pr(>F)
## year       4   23.4    5.85    3.87   0.022 *
## month      4   77.5   19.37   12.82 7.3e-05 ***
## Residuals 16   24.2    1.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Spørgsmål IV.1 (8)

På signifikansniveau $\alpha = 0.05$ hvad er konklusionen fra R-koden herover (ved forskel menes signifikant forskel i middel. Husk at alle dele af svaret skal være korrekt)?

1 ☐ Der er hverken forskel på temperaturen mellem måneder eller år, da $7.3 \cdot 10^{-5} < 0.05$ og $0.022 < 0.05$.

2 ☐ Der er både forskel på temperaturen mellem måneder og år, da $2 \cdot 7.3 \cdot 10^{-5} < 0.05$ og $2 \cdot 0.022 < 0.05$.

3 ☐ Der er forskel på temperaturen mellem måneder da $7.3 \cdot 10^{-5} < 0.05$, men der er ikke forskel mellem år, da $0.022 > 7.3 \cdot 10^{-5}$.

4 ☐ Der er forskel på temperaturen både mellem måneder og år, da $7.3 \cdot 10^{-5} < 0.05$ og $0.022 < 0.05$.

5 ☐ Der er hverken forskel temperaturen mellem måneder eller år, da $2 \cdot 7.3 \cdot 10^{-5} < 0.05$ og $2 \cdot 0.022 < 0.05$.

Spørgsmål IV.2 (9)

Modellen anvendt ved testen ovenfor kan skrives som

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \text{ og i.i.d.}$$

Hvad er estimatet af σ^2?

1 ☐ $\hat{\sigma}^2 = (23.4 + 77.5 + 24.2)/(4 + 4 + 16)$

2 ☐ $\hat{\sigma}^2 = 24.2/16$

3 ☐ $\hat{\sigma}^2 = \sqrt{24.2}$

4 ☐ $\hat{\sigma}^2 = (5.85 + 19.37 + 1.51)/(4 + 4 + 16)$

5 ☐ $\hat{\sigma}^2 = \sqrt{1.51}$

Spørgsmål IV.3 (10)

Hvad er

$$\sum_{\text{all } i,j} (y_{ij} - \bar{y})^2$$

hvor y_{ij} er de enkelte observerede temperaturer og $\bar{y}$ er gennemsnittet af alle de observerede temperature?

1 ☐ 77.4

2 ☐ 26.7

3 ☐ 24.2

4 ☐ 16.7

5 ☐ 125.1

Spørgsmål IV.4 (11)

Man ønsker nu at lave parvise sammenligninger af gennemsnits temperaturen i de enkelte år. På signifikansniveau $\alpha = 0.05$, hvad er den Bonferonni korrigerede Least Significant Difference (LSD)?

1 ☐ 1.27

2 ☐ 2.27

3 ☐ 1.61

4 ☐ 1.36

5 ☐ 2.53

Spørgsmål IV.5 (12)

Hvad ville p-værdien være, hvis man havde valgt at undersøge om der er forskel mellem år, ved en en-vejsanalyse (dvs. ignorerede forskellen mellem måneder)?

1 ☐ 0.362

2 ☐ 0.638

3 ☐ 0.499

4 ☐ 0.463

5 ☐ 0.537

Fortsæt på side 10

Opgave V

I forbindelse med et forskningsprojekt om energiforbrug på skoler, blev der undersøgt hvordan eleverne og lærerne indstiller radiatortermostaterne i klasselokalerne. På tilfældigt udvalgte dage i en kold periode, blev der på en række skoler, i tilfældigt udvalgte klasselokaler, noteret hvorledes termostaterne stod indstillet.

På forhånd var det fastsat, at en passende indstilling er mellem 2 og 3 på termostaterne på alle radiatorer i et lokale, da det ellers kan indikere under- eller overdimensionede radiatorer. Desuden er det ikke godt, hvis termostaterne er indstillet forskelligt, da det fører til ringere komfort og ringere returvandsafkøling.

Følgende observationer blev gjort i perioden:

	Skole 1	Skole 2	Skole 3	Skole 4
Ikke-passende	18	11	22	9
Passende	38	36	15	12

Dvs. på Skole 2 var der 11 ud af 47 lokaler, hvori termostaterne ikke var passende indstillet.

Spørgsmål V.1 (13)

Hvad er 95% konfidensintervallet for andelen af termostater, som ikke var indstillet passende på Skole 1 (bemærk, at resultatet fra R funktionerne og bogens formel kan være lidt forskellige, men altid tættest på det rigtige svar)?

1 ☐ $[0.07, 0.17]$

2 ☐ $[0.20, 0.44]$

3 ☐ $[0.26, 0.53]$

4 ☐ $[0.05, 0.22]$

5 ☐ $[0.18, 0.51]$

Spørgsmål V.2 (14)

Det var planlagt at sammenligne skolerne for at undersøge, om der var forskellig praksis for termostatindstilling på skolerne. Under nulhypotesen, at der ikke var forskel, hvad er da det forventede antal af ikke-passende indstillede termostater på Skole 3?

1 ☐ $e_{13} = 37 \cdot \frac{60}{161} = 13.8$

2 ☐ $e_{13} = 15 \cdot \frac{22}{37} = 8.9$

3 ☐ $e_{13} = 60 \cdot \frac{124}{161} = 46.2$

4 ☐ $e_{13} = 22 \cdot \frac{22}{37} = 13.1$

5 ☐ $e_{13} = 15 \cdot \frac{124}{161} = 11.6$

Spørgsmål V.3 (15)

Man ønsker at undersøge, om der var forskel på andelen af ikke-passende termostatindstillinger på de fire skoler. Anvend et signifikansniveau på 1%. Hvad bliver konklusionen (både konklusion og argument skal være korrekt)?

1 ☐ Der kan ikke påvises en forskel mellem skolerne, da den relevante observerede teststatistik er under den kritiske værdi på 15.4.

2 ☐ Der kan påvises en forskel mellem skolerne, da den relevante observerede teststatistik er under den kritiske værdi på 11.3.

3 ☐ Der kan ikke påvises en forskel mellem skolerne, da den relevante observerede teststatistik er over den kritiske værdi på 0.0057.

4 ☐ Der kan ikke påvises en forskel mellem skolerne, da den relevante observerede teststatistik er under den kritiske værdi på 0.0057.

5 ☐ Der kan påvises en forskel mellem skolerne, da den relevante observerede teststatistik er over den kritiske værdi på 11.3.

Fortsæt på side 12

Opgave VI

De følgende 3 spørgsmål omhandler forskellige statistiske problemer, der kan opstå ved behandling af vand.

Spørgsmål VI.1 (16)

Ved kontrol af drikkevand måles kvaliteten ved regelmæssige vandanalyser. Der er naturligvis lovgivning om kvaliteten, bl.a. krav om forskellige stoffers koncentrationer. Et af kravene er, at vandets ledningsevne ved forbrugers taphane ikke må være over 2500 μS/cm ved 20 °C. Hvis ledningsevnen er over dette niveau, så er der for høj en koncentration af salte og vandet betegnes da som aggressivt.

Man har målt ledningsevnen ved tilfældigt udvalgte forbrugeres taphaner. Lad en måling repræsenteres ved den stokastiske variabel X_i, som kan antages normalfordelt. Man har indsamlet 20 uafhængige målinger, hvor man har bestemt ledningsevnen, og de observerede værdier er gemt i vektoren `x` i R. Man ønsker nu at teste, om vandet er aggressivt i middel, og man opstiller følgende nulhypotese om middelværdien, μ, af ledningsevnen i drikkevandet ved forbrugernes taphaner

$$H_0 : \mu = 2500$$

med den alternative hypotese

$$H_1 : \mu \neq 2500$$

Følgende R kode køres:

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 4.8527, df = 19, p-value = 0.0001106
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   704.1022 1772.1090
## sample estimates:
## mean of x
##  1238.106
```

Baseret på resultatet fra R ovenover og med et 5% signifikansniveau, hvilket af følgende udsagn er da korrekt (både konklusion og begrundelse skal være korrekt)?

1 ☐ Vi kan afvise H_0, da det relevante konfidensinterval ikke indeholder 0.

2 ☐ Vi må acceptere H_0, da stikprøvegennemsnittet ligger i det relevante konfidensinterval.

3 ☐ Vi må acceptere H_0, da teststørrelsen er større end 1.96.

4 ☐ Vi kan afvise H_0, da p-værdien er 0.0001106.

5 ☐ Vi kan afvise H_0, da 2500 ikke ligger i det relevante konfidensinterval.

Spørgsmål VI.2 (17)

På et vandværk renses vandet med en af to metoder, A eller B, og restkoncentrationen af en substans måles. En måling af restkoncentration er givet i de stokastiske variable $Y_{A,i}$ og $Y_{B,i}$ for henholdsvis metode A og B. Man ønsker nu at undersøge, hvilken af de to metoder, der renser vandet bedst. $Y_{A,i}$ og $Y_{B,i}$ kan antages at være normalfordelte og deres varianser, σ_A^2 og σ_B^2, kan antages at være ens. Der udtages nu uafhængigt af hinanden to stikprøver med 20 observationer for hver metode. Nulhypotesen

$$H_0: \quad \mu_A = \mu_B$$

ønskes testes. Den alternative hypotese er

$$H_1: \quad \mu_A \neq \mu_B$$

Hvilken af følgende muligheder er da en korrekt fremgangsmåde?

1 ☐ En parret t-test med 19 frihedsgrader

2 ☐ En parret t-test med 18 frihedsgrader

3 ☐ En two-sample t-test med 38 frihedsgrader

4 ☐ En two-sample t-test med 39 frihedsgrader

5 ☐ En F-test for varianshomogenitet

Spørgsmål VI.3 (18)

Der planlægges en ny undersøgelse af koncentrationen af et stof i en drikkevandsboring. Man ønsker at opnå en styrke på 90% for at påvise et niveau i middelværdi på 2 enheder forskellig fra en givet værdi. Man har erfaring for at standardafgivelsen er på 3.5 enheder og man ønsker at udføre testen på signifikansniveau 5% (husk, at resultater fra R funktioner kan afvige en smule fra resultatet med bogens formler). Hvor mange observationer skal der tages for at de givne krav opfyldes?

1 ☐ Mindst 15 observationer

2 ☐ Mindst 35 observationer

3 ☐ Mindst 48 observationer

4 ☐ Mindst 67 observationer

5 ☐ Mindst 102 observationer

Fortsæt på side 15

Opgave VII

På 30 tilfældigt udvalgte sommerdage har man registreret sammenhørende værdier af temperaturen kl. 12, x, målt i grader celcius og antallet af solgte is, Y, hos en iskæde. Man har nu fittet følgende model i R:

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

Resultat af dette er vist nedenfor:

```
summary(fit1)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -161.24  -81.60  -46.14  103.83  249.85
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2382.001    116.620  -20.43   <2e-16 ***
## x             230.703      5.083   45.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124.7 on 28 degrees of freedom
## Multiple R-squared:  0.9866,Adjusted R-squared:  0.9861
## F-statistic:  2060 on 1 and 28 DF,  p-value: < 2.2e-16
```

Spørgsmål VII.1 (19)

Baseret på ovenstående R-output, hvad er da estimatet for variansen af afvigelserne $\hat{\sigma}^2$?

1 ☐ 116.6

2 ☐ 116.6^2

3 ☐ 124.7

4 ☐ 124.7^2

5 ☐ $(230.7/28)^2$

Spørgsmål VII.2 (20)

Baseret på ovenstående R-output, hvad er da prædiktionen af middelværdien af solgte is, $\hat{y}_{\text{new}}$, ved $x_{\text{new}} = 25°$ C?

1 ☐ 231 is.

2 ☐ 2382 is.

3 ☐ 3386 is.

4 ☐ 5768 is.

5 ☐ 11535 is.

Spørgsmål VII.3 (21)

Baseret på ovenstående R-output, hvad er da den kritiske værdi for testet

$$H_0 : \beta_1 = 0$$

og med et 1% signifikansniveau?

1 ☐ 2.76

2 ☐ 2.05

3 ☐ 1.96

4 ☐ 1.70

5 ☐ 2.56

Spørgsmål VII.4 (22)

Hvilket af følgende udsagn om et prædiktionsinterval for $Y_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}} + \varepsilon_{\text{new}}$ er ikke korrekt?

1 ☐ Prædiktionsintervallet er bredere end et tilsvarende konfidensinterval.

2 ☐ Prædiktionsintervallets bredde afhænger af stikprøvestørrelsen.

3 ☐ Prædiktionsintervallet er symmetrisk omkring den prædikterede værdi.

4 ☐ Prædiktionsintervallets bredde afhænger af værdien af x_{new}.

5 ☐ Hvis stikprøvestørrelsen bliver stor nok, så bliver bredden på prædiktionsintervallet lig 0.

Spørgsmål VII.5 (23)

Man mistænker, at den lineære model ikke er en korrekt model. Man fitter nu i stedet modellen $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$. Modellen er fittet ved (bemærk at `x2` er `x` kvadreret):

```
x2 <- x^2
fit2 <- lm(y ~ x + x2)
summary(fit2)

##
## Call:
## lm(formula = y ~ x + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -193.67  -59.32  -25.73   64.96  263.87
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -159.3055   470.1439  -0.339    0.737
## x             24.9850    42.9277   0.582    0.565
## x2             4.5715     0.9502   4.811 5.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.15 on 27 degrees of freedom
## Multiple R-squared:  0.9928,Adjusted R-squared:  0.9922
## F-statistic:  1856 on 2 and 27 DF,  p-value: < 2.2e-16
```

Baseret på dette R-output og med et 1% signifikansniveau, hvad kan nu konkluderes om sammenhængen mellem issalg og temperatur (konklusion og argument skal begge være korrekte)?

1 ☐ Sammenhængen afviger statistisk signifikant fra en lineær sammenhæng, da $\hat{\beta}_2$ er positiv.

2 ☐ Sammenhængen afviger ikke statistisk signifikant fra en lineær sammenhæng, da p-værdien for $\hat{\beta}_1$ er over 1%.

3 ☐ Sammenhængen afviger statistisk signifikant fra en lineær sammenhæng, da p-værdien for $\hat{\beta}_2$ er under 1%.

4 ☐ Vi kan ikke afvise at sammenhængen er lineær, da $\hat{\beta}_2 < \hat{\beta}_1$.

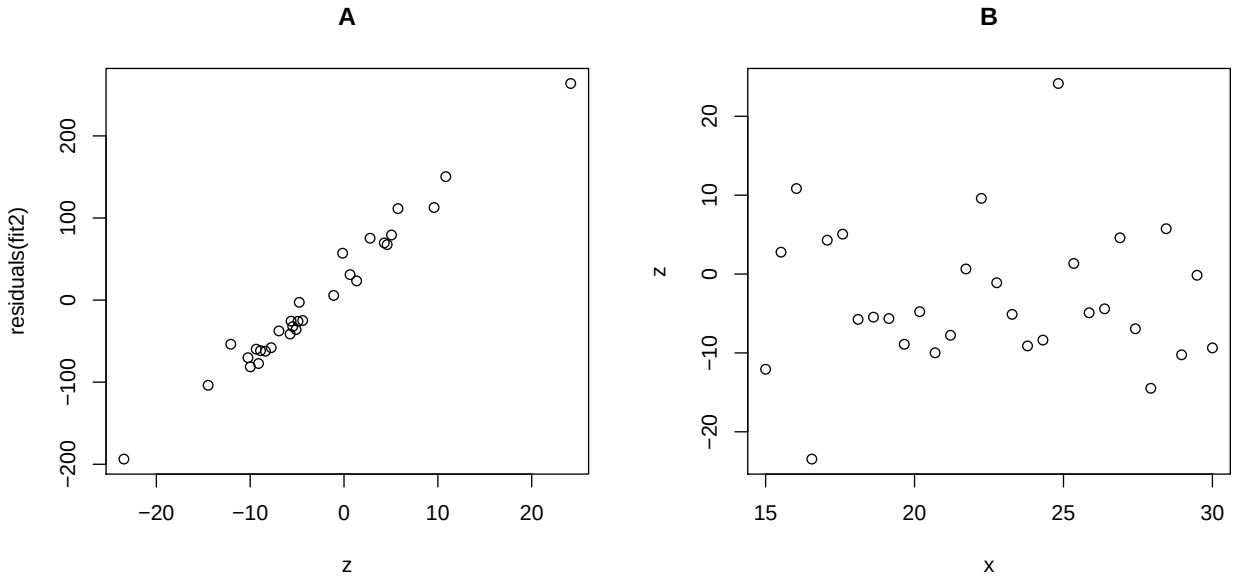
5 ☐ Vi kan ikke afvise at sammenhængen er lineær, da $R^2 \approx 1$.

Spørgsmål VII.6 (24)

I denne opgave undersøges resultatet fra fittet af modellen fra forrige spørgsmål.

Nedenfor er vist 2 plots (A og B), hvor:

- z er en ny variabel, som er målinger af mængden af solskin på en dag (enhenden er underordnet).
- x temperaturen på en dag.
- `residuals(fit2)` er residualerne fra modellen.



Hvilket af følgende udsagn er korrekt?

1 ☐ Plot B bruges til at undersøge om antagelsen om varianshomogenitet er opfyldt.

2 ☐ Plot A bruges til at undersøge om normalfordelingsantagelsen er opfyldt.

3 ☐ Plot B indikerer er en stærk sammenhæng mellem z og x.

4 ☐ Plot A kan bruges til at undersøge om z bør inkluderes i modellen.

5 ☐ Plot B kan bruges til at undersøge om sammenhængen mellem x og Y er modelleret korrekt.

Fortsæt på side 19

Opgave VIII

Antallet af stjerneskud per time, X, er givet ved $X \sim Po(3)$, altså Poissonfordelt med middelværdi 3 stjerneskud per time.

Spørgsmål VIII.1 (25)

Hvis man tæller stjerneskud i 4 timer, hvor mange stjerneskud vil man da kunne forvente at se (husk at forventningsværdien er lig middelværdien)?

1 ☐ 8

2 ☐ 9

3 ☐ 12

4 ☐ 16

5 ☐ 24

Spørgsmål VIII.2 (26)

Antag, at man netop har observeret et stjerneskud. Hvad er da sandsynligheden for at vente mere end 10 minutter på det næste stjerneskud?

1 ☐ $P(X = 0)$, $X \sim Po(3)$

2 ☐ $P(X > 0)$, $X \sim Po(3)$

3 ☐ $P(Y > 10)$, $Y \sim Exp(3)$

4 ☐ $P(Y > \frac{10}{60})$, $Y \sim Exp(3)$

5 ☐ $P(Y > \frac{10}{6})$, $Y \sim Exp(3)$

Fortsæt på side 20

Opgave IX

I en produktion af stålrør er man interesseret i rørenes diameter. Man udtager derfor en stikprøve på 30 rør og beregner stikprøvevariansen til $s^2 = 531$ mm^2.

Spørgsmål IX.1 (27)

Hvad er da 99% konfidensinterval for standardafvigelsen på rørernes diameter?

1 ☐ [18.4, 31.0]

2 ☐ [18.1, 30.6]

3 ☐ [310, 1079]

4 ☐ [17.2, 34.3]

5 ☐ [294, 1175]

Spørgsmål IX.2 (28)

Normalvis foretages målingen af stålrørenes diameter med en af to forskellige målemetoder. Man mistænker nu, at de to metoder ikke måler helt ens. Man udvælger derfor tilfældigt 11 rør. Hvert rør måles nu med begge metoder. Observationen gjort med målemetode 1 på rør i er nedenfor givet på position i i vektoren `x` og tilsvarende fra målemetode 2 på position i i voktoren `y`. Målingerne med begge metoder kan antages at være normalfordelte.

Man foretager nu følgende analyser i R:

```
t.test(x-y)

##
##  One Sample t-test
##
## data:  x - y
## t = -2.541, df = 10, p-value = 0.0293
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -2.409246 -0.158027
## sample estimates:
## mean of x
##  -1.28364
```

```r
t.test(x,y)

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -0.1353, df = 20, p-value = 0.894
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.0683  18.5010
## sample estimates:
## mean of x mean of y
##   96.9018   98.1855
```

Man ønsker at teste nulhypotesen om, at de to målemetoder har ens middelværdi

$$H_0 : \mu_X = \mu_Y$$

mod den alternative hypotese, at de er forskellige

$$H_1 : \mu_X \neq \mu_Y$$

På et 5% signifikansniveau, hvilken af følgende resultater er da korrekt (både konklusion og argument skal være korrekte)?

1 ☐ Vi kan afvise H_0, da $p < 0.05$.

2 ☐ Vi kan ikke afvise H_0, da $p = 0.89$.

3 ☐ Vi kan afvise H_0, da den alternative hypotese ikke er lig 0.

4 ☐ Vi kan ikke afvise H_0, da teststørrelsen t_{obs} er -0.14.

5 ☐ Vi kan afvise H_0, da forskellen på stikprøvegennemsnittene er større end 1.

Spørgsmål IX.3 (29)

Firmaet, der producerer stålrørene, får nu en ny og hurtigere maskine til at producere stålrør. Uanset svaret i forrige spørgsmål, så bruges nu kun målemetode 1 i det følgende. Den ønskede diameter på stålrørene er 100 mm, og man vil sikre sig, at maskinen er korrekt kalibreret. Man vil planlægge et nyt eksperiment, hvor man ønsker at teste nulhypotesen

$$H_0 : \mu_X = 100$$

mod den alternative

$$H_1 : \mu_X \neq 100$$

Man bruger $\hat{\sigma}_x^2 = 502.23$, et signifikansniveau på 5% og har kørt følgende R-kode, idet man vil bruge en stikprøvestørrelse på $n = 40$

```
power.t.test(n=40, sd=sqrt(502.23), power=0.9, type="one.sample")

##
##      One-sample t test power calculation
##
##               n = 40
##           delta = 11.8
##              sd = 22.4
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
```

Baseret på R-koden ovenfor, hvad kan nu konkluderes inden eksperiment udføres (både argument og konklusion skal være korrekt)?

1 ☐ Hvis den sande middelværdi er 88.2 eller lavere har vi mere end 90% chance for at afvise H_1.

2 ☐ Hvis den sande middelværdi er 88.2 eller lavere har vi mere end 90% chance for at afvise H_0.

3 ☐ Hvis den sande middelværdi er 88.2 eller lavere vil vi afvise H_0.

4 ☐ Hvis den sande middelværdi afviger med mindre end 11.8 accepterer vi H_0.

5 ☐ Hvis den sande middelværdi afviger med mindre end 11.8 afviser vi H_0.

Fortsæt på side 23

Opgave X

I produktionen af en bestemt type plade er der for hver plade en 20% sandsynlighed, for at pladen har en fejl. Der udtages nu en tilfældig stikprøve på 10 plader.

Spørgsmål X.1 (30)

Hvad er sandsynligheden for at højst 3 plader har en fejl i stikprøven?

1 ☐ 0.95

2 ☐ 0.32

3 ☐ 0.68

4 ☐ 0.88

5 ☐ 0.60

SÆTTET ER SLUT. God sommer!