Danmarks Tekniske Universitet *Side 1 af 26 sider.*

Skriftlig prøve: 24. June 2021

Kursus navn og nr.: **Introduktion til Mathematisk Statistik (02403)**

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

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

Eksamenssættet 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 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

I en bestemt by angives den gennemsnitlige elektricitetsregning for toværelses lejligheder til at være 550 DKK. Man udtager en stikprøve på 36 toværelses lejligheder. Gennemsnittet for de 36 lejligheder er $\bar{x} = 562$ DKK med en stikprøvestandardafvigelse på $s = 40$ DKK. Nulhypotesen er $H_0 : \mu = 550$ DKK, som ønskes testet mod det tosidede alternativ.

Spørgsmål I.1 (1)

Lad t_{obs} være den sædvanlige teststørrelse i et tosidet t-test. Ved signifikansniveau α, hvad er da kriteriet for at afvise H_0?

1 ☐ Vi afviser H_0 hvis $t_{\text{obs}} \leq t_{1-\alpha}$, hvor $t_{1-\alpha}$ er $1-\alpha$ fraktilen i en t-fordeling med 34 frihedsgrader.

2 ☐ Vi afviser H_0 hvis $t_{\text{obs}} \leq t_{1-\alpha}$, hvor $t_{1-\alpha}$ er $1-\alpha$ fraktilen i en t-fordeling med 35 frihedsgrader.

3 ☐ Vi afviser H_0 hvis $t_{\text{obs}} \geq t_{1-\alpha}$, hvor $t_{1-\alpha}$ er $1-\alpha$ fraktilen i en t-fordeling med 35 frihedsgrader.

4 ☐ Vi afviser H_0 hvis $t_{\text{obs}} \leq t_{\alpha/2}$ eller $t_{\text{obs}} \geq t_{1-\alpha/2}$, hvor $t_{\alpha/2}$ og $t_{1-\alpha/2}$ er hhv. $\alpha/2$ or $1-\alpha/2$ fraktilerne i en t-fordeling med 35 frihedsgrader.

5 ☐ Vi afviser H_0 hvis $t_{\text{obs}} \leq t_{\alpha/2}$ eller $t_{\text{obs}} \geq t_{1-\alpha/2}$, hvor $t_{\alpha/2}$ og $t_{1-\alpha/2}$ er hhv. $\alpha/2$ or $1-\alpha/2$ fraktilerne i en t-fordeling med 34 frihedsgrader.

Spørgsmål I.2 (2)

Hvad er værdien af teststørrelsen (t_{obs})?

1 ☐ $t_{\text{obs}} = 1.80$.

2 ☐ $t_{\text{obs}} = 1.96$.

3 ☐ $t_{\text{obs}} = 2.03$.

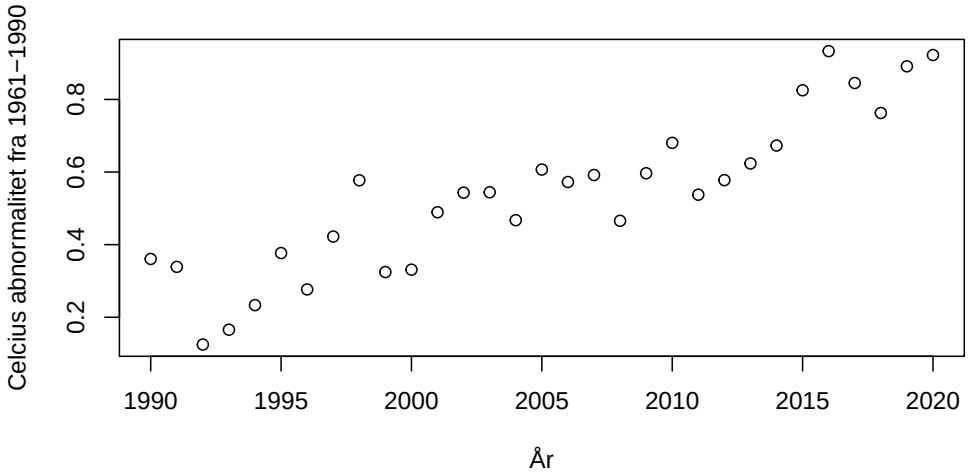
4 ☐ $t_{\text{obs}} = 11.38$.

5 ☐ $t_{\text{obs}} = 11.22$.

Fortsæt på side 3

Opgave II

Historiske værdier for den globale gennemsnitstemperatur kan hentes fra FN's IPCC-websted. Den årlige temperatur de sidste 31 år er blevet hentet og er plottet nedenfor. Temperaturen er den såkaldte temperaturabnormalitet, som er forskellen fra gennemsnittet i perioden 1961 til 1990.



Den i'te observation af temperatur er benævnt ved T_i og år y_i, således at $i = 0, \ldots, 30$.

For at analysere temperaturudviklingen over tid igennem de 31 år blev følgende model anvendt

$$T_i = \beta_0 + \beta_1 y_i + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.}$$

Fortsæt på side 4

Resultatet vises ved udskriften af summary fra R:

```
##
## Call:
## lm(formula = temperatur ~ aar)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13710 -0.07617  0.01296  0.05885  0.18984
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.622490   3.736256  -11.41 3.07e-12 ***
## aar           0.021526   0.001863   11.55 2.27e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0928 on 29 degrees of freedom
## Multiple R-squared:  0.8215,Adjusted R-squared:  0.8153
## F-statistic: 133.4 on 1 and 29 DF,  p-value: 2.271e-12
```

Spørgsmål II.1 (3)

Ifølge formuleringerne fra bevisfortolkningstabellen i bogen, hvad er konklusionen for nulhypotesen: Der er ingen ændring i gennemsnitstemperaturen i perioden (både konklusion og argument skal være korrekt)?

1 ☐ Der er meget stærke beviser mod H_0 da p-værdien er $3.07 \cdot 10^{-12}$.

2 ☐ Der er meget stærke beviser mod H_0 da p-værdien er $2.27 \cdot 10^{-12}$.

3 ☐ Der er stærke beviser mod H_0 da p-værdien er $3.07 \cdot 10^{-12}$.

4 ☐ Der er stærke beviser mod H_0 da p-værdien er $2.27 \cdot 10^{-12}$.

5 ☐ Der er svage beviser mod H_0 da p-værdien er 0.0928.

Fortsæt på side 5

Spørgsmål II.2 (4)

Hvad er 99% konfidensintervallet for gennemsnitsstigningen i temperatur per år i denne periode?

1 ☐ [-52.9, -32.3]

2 ☐ [0.0184, 0.0247]

3 ☐ [0.421, 0.431]

4 ☐ [0.0164, 0.0267]

5 ☐ [0.323, 0.529]

Spørgsmål II.3 (5)

Hvad er den empiriske korrelation koefficient mellem temperatur og år i perioden?

1 ☐ 0.178

2 ☐ 0.305

3 ☐ 0.695

4 ☐ 0.822

5 ☐ 0.906

Fortsæt på side 6

Opgave III

Kortspillet Casino spilles med et almindeligt spil kort, dvs. 52 kort, 13 i hver farve (klør, hjerter, ruder og spar) og 4 af hver værdi (f.eks. 4 esser). Spillet starter med at der placeres 4 kort med billedsiden opad på bordet.

Spørgsmål III.1 (6)

Hvad er sandsynligheden for, at mindst 3 af de 4 åbningskort er esser?

1 ☐ 0.00057

2 ☐ $2.7 \cdot 10^{-6}$

3 ☐ 0.025

4 ☐ 0.00071

5 ☐ $3.7 \cdot 10^{-6}$

Spørgsmål III.2 (7)

Hvad er sandsynligheden for, at alle de 4 åbningskort er spar i præcis 1 ud af 4 spil?

1 ☐ 0.010

2 ☐ 0.39

3 ☐ 0.42

4 ☐ 0.44

5 ☐ 0.015

Fortsæt på side 7

Opgave IV

Som følge af COVID-19 restriktioner, har man været nødt til at gøre brug af onlineværktøjer til gruppearbejde i undervisningen. Nedenstående tabel opsummerer svarene i en undersøgelse om de studerendes foretrukne onlineplatform til gruppearbejde. Undersøgelsen blev udført i tre på hinanden følgende semestre ("F"står for forår og "E"for efterår)

	F2020	E2020	F2021	*Sum*
Teams	796	729	669	*2194*
Messenger	1017	1025	1000	*3042*
Discord	278	313	315	*906*
Slack	855	1013	1033	*2901*
Sum	*2946*	*3080*	*3017*	*9043*

Spørgsmål IV.1 (8)

Vi ønsker at teste nulhypotesen om at andelen af studerende i foråret 2021, som foretrak slack, var 34.0%. Den sædvanlige teststørrelse (brugt ved store stikprøver) for denne hypotese er:

1 ☐ $(1025 - 0.66 \cdot 3017)/\sqrt{3017 \cdot 0.34 \cdot 0.66} = -36.83$

2 ☐ $(1033 - 0.34 \cdot 3017)/\sqrt{3017 \cdot 0.34 \cdot 0.34} = 0.38$

3 ☐ $(1013 - 0.34 \cdot 3017)/\sqrt{9043 \cdot 0.34 \cdot 0.66} = 0.16$

4 ☐ $(1033 - 0.34 \cdot 3017)/\sqrt{3017 \cdot 0.34 \cdot 0.66} = 0.28$

5 ☐ $(1033 - 0.66 \cdot 3017)/\sqrt{3017 \cdot 0.34 \cdot 0.34} = -51.31$

Spørgsmål IV.2 (9)

Vi vil nu teste om der er en signifikant forskel mellem efteråret 2020 og foråret 2021 ift. andelen af studerende, som foretrak Teams.

Hvilket af nedenstående stykker kode udfører den rette test i R?

1 ☐
```
binom.test(x = 669, n = 3017, p = 729/3080)
```

2 ☐
```
prop.test(x = c(729, 669), c(3080, 3017), correct = FALSE)
```

3 ☐
```
binom.test(x = 669, n = 3017, p = (669+729)/(3017+3080))
```

4 ☐
```
prop.test(x = 669, n = 3017, p = 729/3080, correct = FALSE)
```

5 ☐ Ingen af ovenstående.

Spørgsmål IV.3 (10)

Man kan også med en hypotesetest undersøge om fordelingen af foretrukken digital platform har ændret sig i løbet af de tre semestre i undersøgelsen. Antallet af frihedsgrader i teststørrelsens fordeling er da:

1 ☐ 4

2 ☐ 8

3 ☐ 6

4 ☐ 10

5 ☐ 12

Spørgsmål IV.4 (11)

Antag uafhængighed mellem semestre og foretrukken onlineplatform. Det forventede antal af studerende, som bruger Messenger i foråret 2021, estimeres så til at være:

1 ☐ 1000

2 ☐ $3017 \cdot 3042/9043 = 1015$

3 ☐ $(1017 + 1025 + 1000)/3 = 1014$

4 ☐ $(1017 + 1025)/2 = 1021$

5 ☐ $1000 \cdot 3017/3042 = 992$

Fortsæt på side 9

Spørgsmål IV.5 (12)

For at kunne teste om der er sket en signifikant ændring i foretrukken digital platform i løbet af de tre semestre, bliver følgende oplyst: Teststørrelsen er udregnet til 31.316. Signifikansniveauet sættes til $\alpha = 0.05$. For den fordeling, som teststørrelsen vurderes op imod, er 95% og 97.5% fraktilerne givet ved hhv. 12.59 og 14.45. Hvad kan konkluderes ud fra dette (både konklusion og argument skal være korrekt)?

1 ☐ De studerendes præferencer for digital platform har ændret sig signifikant, idet teststørrelsen er større end den givne 95% fraktil.

2 ☐ Fordelingen på tværs af platform og semester har ændret sig signifikant, idet der under nulhypotesen er 97.5% sandsynlighed for at observere en teststørrelse større end 14.45.

3 ☐ Fordelingen på tværs af platform og semester har ikke ændret sig signifikant, idet teststørrelsen er større end den givne 97.5% fraktil.

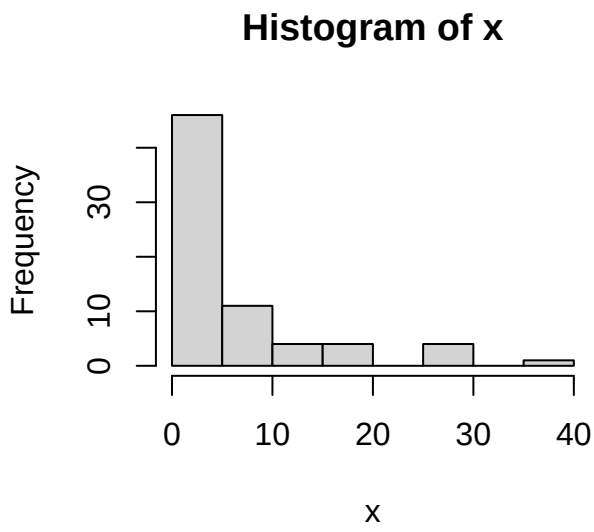
4 ☐ Fordelingen på tværs af platform og semester har ændret sig signifikant, idet der under nulhypotesen er 95% sandsynlighed for at observere en teststørrelse større end 12.59.

5 ☐ De oplyste talværdier kan ikke bruges som statistisk argument for hvorvidt fordelingen på tværs af platforme og semestre har ændret sig eller ej.

Fortsæt på side 10

Opgave V

I en undersøgelse af brugen af persienner, måles tiden (målt i timer) mellem interaktioner med persiennen. Data er vist i histogrammet nedenfor.



Nogle nøgletal er angivet i nedenstående R-output (`x` er data vist i histogrammet)

```
m <- mean(x)
s <- sd(x)
n <- length(x)
c(m, s, n)

## [1]  5.996429  8.243370 70.000000
```

Spørgsmål V.1 (13)

Hvis vi antager, at ventetiderne (tiden mellem interaktioner med persienner) følger uafhængige eksponentialfordelinger med middelværdi lig med det empiriske gennemsnit, hvad er sandsynligheden for at observere en ventetid større end 10 så?

1 ☐ 0.81

2 ☐ 0.68

3 ☐ 0.031

4 ☐ 0.97

5 ☐ 0.19

Som en hjælp til de næste spørgsmål er nedenstående R-kode blevet evalueret

```
k <- 10000

t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 6, df = 69, p-value = 6e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.03 7.96
## sample estimates:
## mean of x
##         6

X <- matrix(rexp(n * k, m), ncol=n)
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))

##    1%  2.5%    5%   10%   90%   95% 97.5%   99%
## 0.124 0.130 0.135 0.142 0.193 0.201 0.207 0.215

X <- t(replicate(k,sample(x, replace = TRUE)))
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))

##    1%  2.5%    5%   10%   90%   95% 97.5%   99%
##  3.90  4.19  4.44  4.76  7.29  7.68  8.03  8.49

X <- matrix(rexp(n * k, 1 / m), ncol = n)
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))

##    1%  2.5%    5%   10%   90%   95% 97.5%   99%
##  4.43  4.64  4.84  5.09  6.92  7.22  7.49  7.78

X <- matrix(rnorm(n * k, m, s), ncol=n)
quantile(apply(X, 1, mean),c(0.01,0.025,0.05,0.1,0.9,0.95,0.975,0.99))

##    1%  2.5%    5%   10%   90%   95% 97.5%   99%
##  3.70  4.08  4.40  4.74  7.25  7.64  7.97  8.36
```

Fortsæt på side 12

Spørgsmål V.2 (14)

Igen under antagelse af eksponentialfordelingen og baseret på R-koden ovenfor, hvad er 95 % konfidensintervallet for den gennemsnitlige ventetid, baseret på parametrisk bootstrap?

1 ☐ $[0.13, 0.21]$

2 ☐ $[4.03, 7.96]$

3 ☐ $[4.64, 7.49]$

4 ☐ $[4.19, 8.03]$

5 ☐ $[4.08, 7.97]$

Spørgsmål V.3 (15)

Hvad ville 98% konfidensintervallet være, hvis beregningerne ikke var baseret på eksponentialfordelings antagelsen (dvs. ikke-parametrisk bootstrap)?

1 ☐ $[4.03, 7.96]$

2 ☐ $[4.43, 7.78]$

3 ☐ $[3.90, 8.49]$

4 ☐ $[0.12, 0.22]$

5 ☐ $[3.7, 8.36]$

Fortsæt på side 13

Spørgsmål V.4 (16)

Data om brugen af persienner kommer fra to forskellige persienner, og det ønskes at teste, om der er forskel i den gennemsnitlige ventetid for de to persienner. I R-koden nedenfor betegner `x1`, `n1`, `m1`, `s1` henholdsvis observerede ventetider, antal observationer, gennemsnitlig ventetid og empirisk standardafvigelse for persienne et, mens `x2`, `n2`, `m2`, `s2` betegner de samme tal for persienne to. Hvilke af følgende stykker R-kode beregner et ikke-parametrisk 95% konfidensinterval for forskellen i den forventede ventetid mellem de to persienner?

1 ☐
```
D <- matrix(rexp((n1 + n2) * k, 1 / m1 - 1 / m2), ncol = n1 + n2)
quantile(apply(D, 1, mean),c(0.025, 0.975))
```

2 ☐
```
D <- t(replicate(k, sample(x1, replace = TRUE))) -
    t(replicate(k,sample(x2, replace = TRUE)))
quantile(apply(D, 1, mean),c(0.025, 0.975))
```

3 ☐
```
t.test(x1,x2)
```

4 ☐
```
D <- matrix(rexp(n1 * k, 1 / m1), ncol = n1) -
    matrix(rexp(n2 * k, 1 / m2), ncol = n2)
quantile(apply(X, 1, mean),c(0.05, 0.95))
```

5 ☐
```
D <- t(replicate(k,sample(x1 - x2, replace = TRUE)))
quantile(apply(D, 1, mean),c(0.025, 0.975))
```

Fortsæt på side 14

Opgave VI

Et forsøg designet til at undersøge hvorvidt brugen af optagede videoforelæsninger og gruppeøvelser påvirker de studerendes eksamensresultater er udført. 9 studerende (benævnt `x`) benyttede ingen af de to læringsværktøjer, mens 9 andre studerende (benævnt `y`) benyttede begge læringsværktøjer. Nedenstående tabel viser de studerendes eksamensresultater:

Studerende	1	2	3	4	5	6	7	8	9
x	54	54	38	47	48	62	54	55	52
y	67	70	71	60	68	65	68	67	64

Under antagelse af at de to grupper er uafhængige, køres følgende kode i R for at teste om forskellen på eksamenresultatet i de to grupper kan antages at være nul ($H_0 : \delta = 0$):

```
x <- c(54, 54, 38, 47, 48, 62, 54, 55, 52)
y <- c(67, 70, 71, 60, 68, 65, 68, 67, 64)
```

Outputtet fra den sædvanlige statistiske analyse er givet nedenfor. Bemærk at nogle af tallene i det sædvanlige output er blevet erstattet af bogstaverne A, B og C.

```
t = -6.0836, df = A , p-value = B
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.537160  C
sample estimates:
mean of x mean of y
 51.55556  66.66667
```

Fortsæt på side 15

Spørgsmål VI.1 (17)

Under antagelse af at data er normalfordelt, hvad er da konklusionen ved signifikansniveau $\alpha = 0.05$ (alle dele af svaret skal være korrekt)?

Brugen af læringsværktøjer:

1 ☐ påvirker eksamensresultatet, idet gennemsnitsresultatet for x og y er forskelligt.

2 ☐ har ikke en signifikant effekt på eksamensresultatet, idet den øvre grænse for konfidensintervallet er 9.685062.

3 ☐ har ikke en signifikant effekt på eksamensresultatet, idet den nedre grænse for konfidensintervallet er -20.537160.

4 ☐ har en signifikant effekt på eksamensresultatet, idet p-værdien er $6.039 \cdot 10^{-5}$.

5 ☐ har en signifikant effekt på eksamensresultatet, idet p-værdien er $4.996 \cdot 10^{-4}$.

Fortsæt på side 16

Opgave VII

I en undersøgelse af træernes vækst måles omkredsen [mm] af træerne som en funktion af alder for et antal træer. Ved hver alder (et tidspunkt) måles omkredsen for alle træer. Som en indledende analyse testes, ved hjælp af en envejs-anova-analyse, om der er forskel i omkreds som funktion af alder. Resultatet ses i R-output nedenfor (hvor nogle tal er erstattet af bogstaver).

```
> anova(lm(circumference ~ age))
Analysis of Variance Table

Response: circumference
          Df Sum Sq Mean Sq F value    Pr(>F)
age        6  A     16008.4       D  1.661e-10 ***
Residuals 28  B           C
```

Spørgsmål VII.1 (18)

Hvad er det samlede antal observationer i undersøgelsen?

1 ☐ 28

2 ☐ 33

3 ☐ 35

4 ☐ 29

5 ☐ 34

Fortsæt på side 17

Den sædvanlige model er

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{1}$$

hvor ϵ_{ij} er iid. stokastiske variable.

Spørgsmål VII.2 (19)

Hvad er det sædvanlige estimat for σ?

1 ☐ 126.5

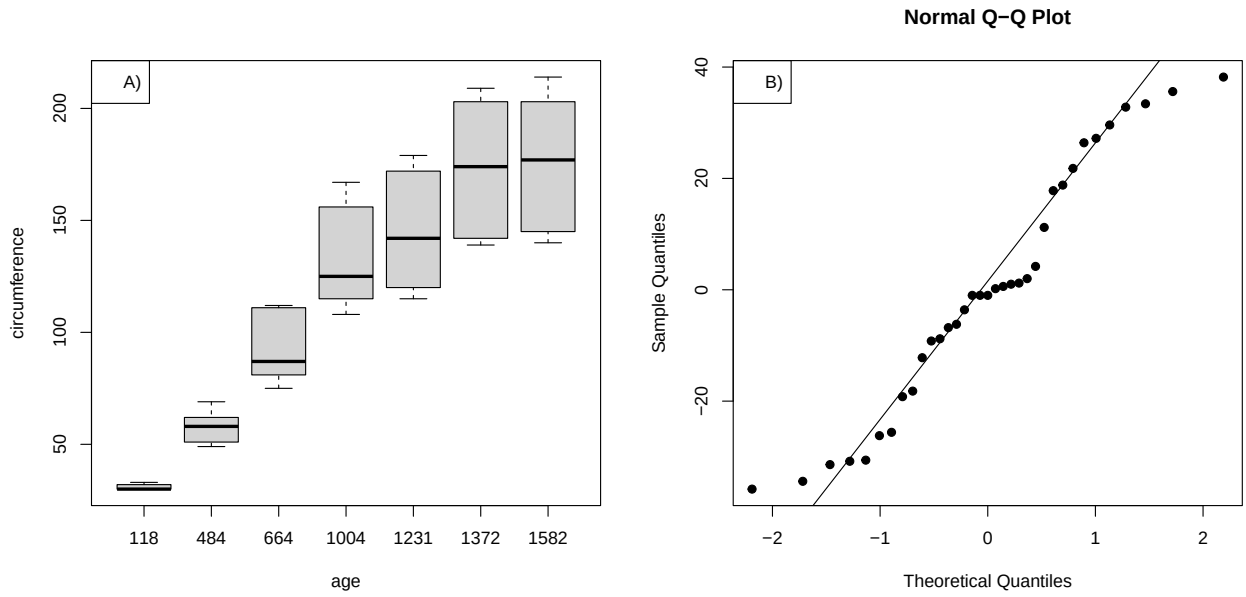
2 ☐ 27.4

3 ☐ 51.7

4 ☐ 24.1

5 ☐ 457.4

Som en del af modelkontrolen plottes nedenstående figur, der viser box-plot af omkredsen som en funktion af alder og et qq-normal plot af residualerne fra modellen.



Fortsæt på side 18

Spørgsmål VII.3 (20)

Hvilket af de følgende udsagn er mest passende ud fra figuren ovenfor (inklusiv henvisning til plot)?

1 ☐ Antagelsen om uafhængighed er klart ikke opfyldt (plot B).

2 ☐ Alle de sædvanlige antagelser ser ud til at være opfyldt (baseret på begge plot).

3 ☐ Normalfordelings antagelsen for residualerne er klart ikke opfyldt (plot A).

4 ☐ Antagelsen om varians-homogenitet er klart ikke opfyldt (plot A).

5 ☐ Antagelsen om uafhængighed er klart opfyldt (plot B).

Uanset svaret på forrige spørgsmål, besluttes det at teste, om der er en signifikant forskel mellem træer ved hjælp af modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{2}$$

hvor α_i angiver effekt of alder (tid) og β_j angiver effekten af træ. Residual variansen er estimeret til $\hat{\sigma}^2 = 186.4$, bemærk også, at alle træer måles en gang ved hver alder (dvs. 5 træer).

Spørgsmål VII.4 (21)

På signifikansniveau $\alpha = 0.05$ og baseret på 2-vejs anova-analysen, hvilket af følgende udsagn er da korrekt (alle dele af svaret skal være korrekt)?

1 ☐ Der er ikke en signifikant effekt af alder, da $15.9 > 4.54$

2 ☐ Der er en signifikant effekt af alder, da $85.9 > 2.51$

3 ☐ Der er en signifikant effekt af alder, da $1.66 \cdot 10^{-10} < 0.05$

4 ☐ Der er ikke en signifikant effekt af alder, da $1.66 \cdot 10^{-6} < 0.05$

5 ☐ Der er en signifikant effekt af alder, da $15.9 > 2.78$

Fortsæt på side 19

Spørgsmål VII.5 (22)

Hvad er den, Bonferoni-korrigeret, 95% mindst signifikant forskel (LSD) for parvise sammenligninger af alle aldersgrupper?

1 ☐ 9.46

2 ☐ 49.2

3 ☐ 24.4

4 ☐ 10.90

5 ☐ 29.3

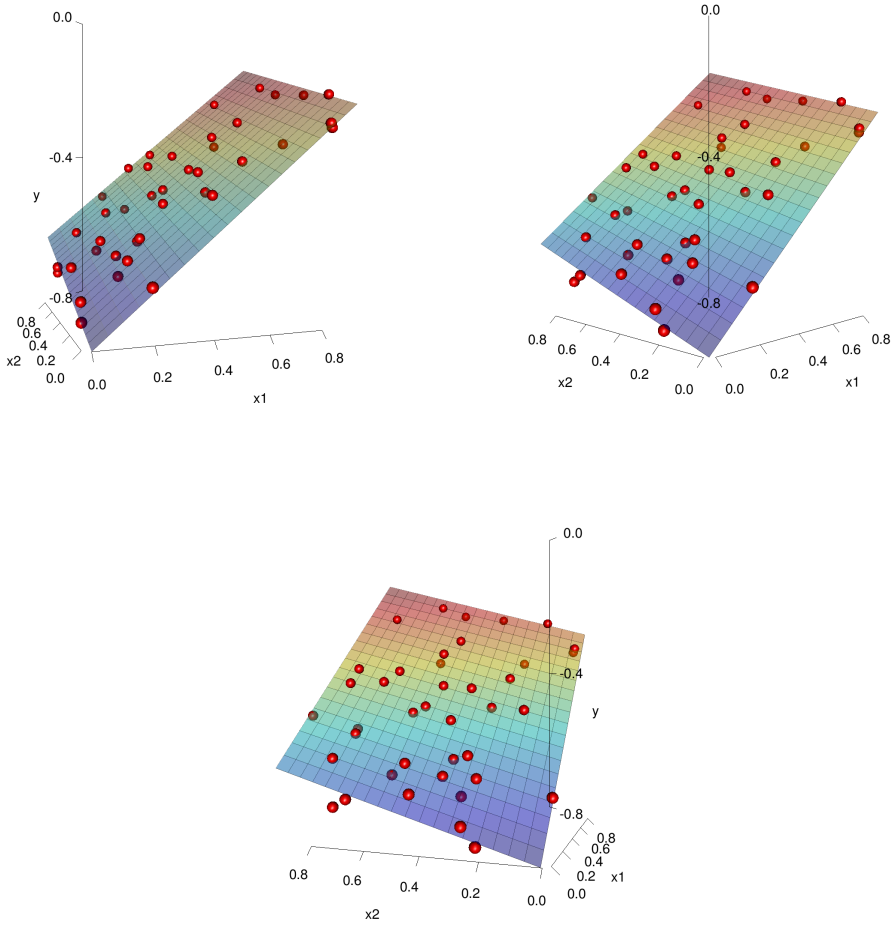
Fortsæt på side 20

Opgave VIII

Den multiple lineære regressions model

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.}$$

er anvendt på data med $i = 1, 2, \ldots, n$, og resultatet visualiseres nedenfor. De tre plots nedenfor viser punkterne og det estimerede plan i forskellige rotationer, således at det kan ses fra forskellige vinkler. Bemærk, at planet er farvet rødt for høje og blåt for lave y-værdier, og husk at kontrollere værdierne på akserne for deres retning:



Fortsæt på side 21

Spørgsmål VIII.1 (23)

Hvad er den korrekte rækkefølge af de estimerede koefficienter?

1 ☐ $\hat{\beta}_0 < \hat{\beta}_1 < \hat{\beta}_2$

2 ☐ $\hat{\beta}_0 < \hat{\beta}_2 < \hat{\beta}_1$

3 ☐ $\hat{\beta}_1 < \hat{\beta}_0 < \hat{\beta}_2$

4 ☐ $\hat{\beta}_1 < \hat{\beta}_2 < \hat{\beta}_0$

5 ☐ $\hat{\beta}_2 < \hat{\beta}_1 < \hat{\beta}_0$

Spørgsmål VIII.2 (24)

Fra plottene er det klart, at den estimerede standardafvigelse af fejlene, dvs. $\hat{\sigma}$, kun kan være en af følgende værdier, hvilken?

1 ☐ $\hat{\sigma} = 0.03$

2 ☐ $\hat{\sigma} = 2.1$

3 ☐ $\hat{\sigma} = 4.7$

4 ☐ $\hat{\sigma} = 357$

5 ☐ $\hat{\sigma} = 10028$

Det multiple lineære regressionsproblem kan skrives i matrix-vektor notation som

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

hvor $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$.

Fortsæt på side 22

Spørgsmål VIII.3 (25)

I R-koden nedenfor bruges `X` for $\boldsymbol{X}$, `beta` betegner $\hat{\boldsymbol{\beta}}$, `sigma` betegner $\hat{\sigma}$, og `n` angiver antallet af observationer. Hvilke af følgende stykker R-kode beregner den øvre grænse i et 95% prædiktions interval for en ny observation ved $x_1 = x_2 = 0$?

1 ☐
```
x <- matrix(c(0, 0, 0),ncol=3)
x %*% beta + sigma * qt(0.975, df = n - 3) *
    sqrt(1 + x %*% solve(t(X) %*% X) %*% t(x))
```

2 ☐
```
x <- matrix(c(1, 0, 0),ncol=3)
x %*% beta + sigma * qt(0.975, df = n - 3) *
    sqrt(1 + x %*% solve(t(X) %*% X) %*% t(x))
```

3 ☐
```
x <- matrix(c(1, 0, 0),ncol=3)
X %*% beta + sigma * qt(0.975, df = n) *
    sqrt(x %*% solve(t(X) %*% X) %*% t(x))
```

4 ☐
```
x <- matrix(c(0, 0, 0),ncol=3)
x %*% beta + sigma / sqrt(n) * qt(0.975, df = n - 3) *
    sqrt(1 + x %*% solve(t(X) %*% X) %*% t(x))
```

5 ☐
```
x <- matrix(c(1, 0, 0),ncol=3)
x %*% beta + sigma / sqrt(n) * qt(0.975, df = n) *
    sqrt(1 + x %*% solve(t(X) %*% X) %*% t(x))
```

Fortsæt på side 23

Opgave IX

Ved modellering af sandsynligheder anvendes probit-transformationen ofte. For et specifikt problem til modellering af sandsynligheden for at åbne vinduer foreslås følgende model

$$p(U) = \Phi\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + U + \psi\right)\right) \tag{3}$$

hvor Φ er CDF for en standard normalfordelt stokastisk variabel, h er tid på dagen (målt i timer), og β_0, β_1 og ψ er fastholdte parametre. Yderligere er U en uniformt fordelt stokastisk variabel på intervallet $[0, 2\pi]$.

Spørgsmål IX.1 (26)

Hvad er fejlophobningslovens approksimation af middelværdien og variansen for p (ϕ er tæthedsfunktionen (pdf) for en standard normalfordelt stokastisk variabel)?

1 ☐ $\mathrm{E}[p] \approx p(\pi)$, and $\mathrm{V}[p] \approx \Phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1^2 \sin^2\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{12}$

2 ☐ $\mathrm{E}[p] \approx \Phi(\pi)$, and $\mathrm{V}[p] \approx \phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1^2 \sin^2\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{12}$

3 ☐ $\mathrm{E}[p] \approx 0$, and $\mathrm{V}[p] \approx \phi^2(\pi)\, \frac{\pi^2}{3}$

4 ☐ $\mathrm{E}[p] \approx \Phi(\pi)$, and $\mathrm{V}[p] \approx \Phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1 \sin\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{3}$

5 ☐ $\mathrm{E}[p] \approx p(\pi)$, and $\mathrm{V}[p] \approx \phi^2\left(\beta_0 + \beta_1 \cos\left(\frac{2\pi h}{24} + \pi + \psi\right)\right) \beta_1^2 \sin^2\left(\frac{2\pi h}{24} + \pi + \psi\right) \frac{\pi^2}{3}$

Fortsæt på side 24

Spørgsmål IX.2 (27)

Med $h = 12$, $\beta_0 = \psi = 0$ og $\beta_1 = 1$, hvilken R-kode approksimere middelværdi og varians for p?

1 ☐
```
u <- runif(10000)
c(mean(pnorm(cos(pi + u))), var(pnorm(cos(pi + u))))
```

2 ☐
```
u <- rnorm(10000, 0, 2 * pi)
c(mean(dnorm(cos(pi + u))), var(dnorm(cos(pi + u))))
```

3 ☐
```
u <- runif(10000, 0, 2 * pi)
c(mean(pnorm(cos(pi + u))), var(pnorm(cos(pi + u))))
```

4 ☐
```
u <- runif(10000, 0, 2 * pi)
c(mean(rnorm(cos(pi + u))), var(rnorm(cos(pi + u))))
```

5 ☐
```
u <- rnorm(10000)
c(mean(pnorm(cos(pi + u), 0, 2 * pi)), var(pnorm(cos(pi + u), 0, 2 * pi)))
```

Fortsæt på side 25

Opgave X

Serien af A-papir er defineret ved, længde = $\sqrt{2}k$ gange bredde. En maskine skærer en A-serie papir. Antag, at maskinens nøjagtighed kan udtrykkes som

$$X \sim N(k, \sigma^2)$$
$$Y \sim N(\sqrt{2}k, \sigma^2)$$

hvor X er bredden og Y er længden, det kan yderligere antages, at X og Y er uafhængige.

Spørgsmål X.1 (28)

Med X og Y som defineret ovenfor, hvad er $E[X^2 + Y^2]$?

1 ☐ $3k^2$

2 ☐ $3k^2 + 2k\sigma^2$

3 ☐ $3k^2 + k\sigma^2$

4 ☐ $3k^2 + 2\sigma^2$

5 ☐ $3k^2 + 2k^2\sigma^2$

Spørgsmål X.2 (29)

Igen med X og Y som defineret ovenfor, hvad er $P\left(\frac{(X-k)^2}{(Y-\sqrt{2}k)^2} < 2\right)$?

1 ☐ 0.39

2 ☐ 0.50

3 ☐ 0.55

4 ☐ 0.67

5 ☐ 0.61

Fortsæt på side 26

Spørgsmål X.3 (30)

Hvad er $P\left(\frac{X-k}{|Y-\sqrt{2}k|} < -1\right)$? (svaret skal gælde for alle valg af σ og k).

1 ☐ 0.33

2 ☐ $P(T < -\sqrt{2}k)$, hvor $T \sim t(1)$

3 ☐ $P(T < -\sigma^2)$, hvor $T \sim t(1)$

4 ☐ 0.25

5 ☐ $P(T < -\frac{1}{\sigma^2})$, hvor $T \sim t(1)$

SÆTTET ER SLUT. God sommer!