Danmarks Tekniske Universitet *Side 1 af 26 sider.*

Skriftlig prøve: 22. Juni 2023

Kursus navn og nr.: **Introduktion til Matematisk 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" siderne på eksamen.dtu.dk.

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

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

Opgave	VII.5	VII.6	VIII.1	VIII.2	VIII.3	IX.1	IX.2	IX.3	X.1	X.2
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

En forsker er interesseret i at sammenligne den gennemsnitlige vægtøgning (i gram) for tre forskellige grupper af mus fodret med tre forskellige diæter. Data er angivet nedenfor.

```
group1 <- c(27, 22, 18, 26, 24)
group2 <- c(32, 22, 32, 25, 25)
group3 <- c(29, 25, 30, 30, 24)
```

Spørgsmål I.1 (1)

Udfør en envejs ANOVA og test den sædvanlige nulhypotese om samme behandlingsmiddelværdier på signifikansniveau $\alpha = 0.05$. Er der en signifikant forskel i vægtøgning blandt de tre grupper?

1 ☐ p-værdien er 0.03. Forskellen mellem gruppemiddelværdier er ikke signifikant, fordi p-værdien er mindre end 0.05.

2 ☐ p-værdien er 0.1879. Forskellen mellem gruppemiddelværdier er ikke signifikant, fordi p-værdien er større end 0.05.

3 ☐ p-værdien er 0.1879. Forskellen mellem gruppemiddelværdier er signifikant, fordi p-værdien er større end 0.05.

4 ☐ p-værdien er 0.3758. Forskellen mellem gruppemiddelværdier er signifikant, fordi p-værdien er større end 0.05.

5 ☐ p-værdien er 0.03. Forskellen mellem gruppemiddelværdier er signifikant, fordi p-værdien er mindre end 0.05.

Fortsæt på side 3

Spørgsmål I.2 (2)

Eksperimentet beskrevet ovenfor blev gentaget (samme antal mus) af en anden forsker, som indsamlede et andet datasæt. Igen blev envejs ANOVA brugt til at teste for signifikant forskel mellem behandlingsmiddelværdier. Følgende ANOVA-tabel blev udregnet. Bemærk at nogle elementer er blevet erstattet af spørgsmålstegn.

```
## Analysis of Variance Table

## Response: weight_gain
##           Df Sum Sq Mean Sq F value Pr(>F)
## group      2  78.53  39.267   1.069 0.3739
## Residuals  ? 440.80       ?
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hvilket af følgende udsagn er korrekt?

1 ☐ `Df(Residuals)` = 14 og `Mean Sq(Residuals)` = 31.486.

2 ☐ `Df(Residuals)` = 15 og `Mean Sq(Residuals)` = 29.387.

3 ☐ `Df(Residuals)` = 12 og `Mean Sq(Residuals)` = 36.733.

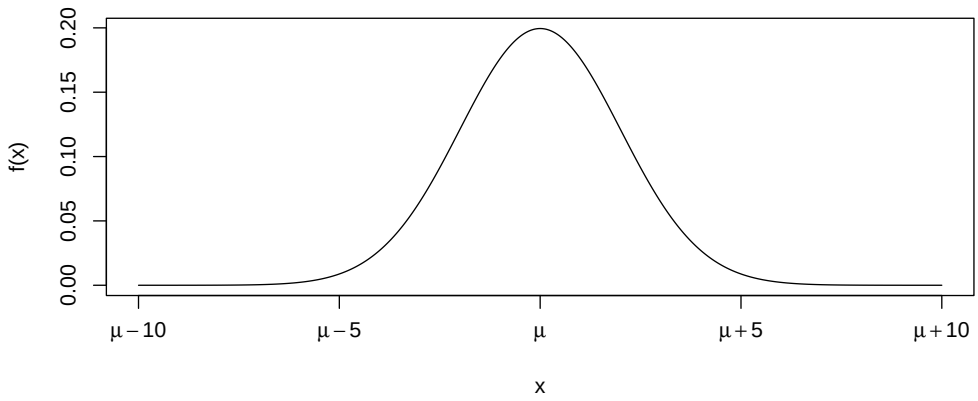
4 ☐ `Df(Residuals)` = 14 og `Mean Sq(Residuals)` = 2.805.

5 ☐ `Df(Residuals)` = 13 og `Mean Sq(Residuals)` = 3.021.

Fortsæt på side 4

Opgave II

Lad den stokastiske variabel X være normalfordelt med middelværdi μ og standardafvigelse $\sigma = 2$, dvs. $X \sim N(\mu, 2^2)$. Dens tæthedsfunktion er:



Spørgsmål II.1 (3)

Lad en anden stokastisk variabel være defineret ved funktionen

$$Y_1 = a_1 + b_1 \cdot X + b_2 \cdot X$$

Hvad er middelværdien og variansen af Y_1?

1 ☐ $\mathrm{E}(Y_1) = a_1 + (b_1 + b_2)\mu$ og $\mathrm{V}(Y_1) = (b_1 + b_2)^2 \cdot 4$

2 ☐ $\mathrm{E}(Y_1) = a_1 + b_1 + b_2$ og $\mathrm{V}(Y_1) = b_1^2 + b_2^2$

3 ☐ $\mathrm{E}(Y_1) = a_1 + b_1 + b_2$ og $\mathrm{V}(Y_1) = b_1 + b_2$

4 ☐ $\mathrm{E}(Y_1) = 0$ og $\mathrm{V}(Y_1) = b_1^2 + b_2^2$

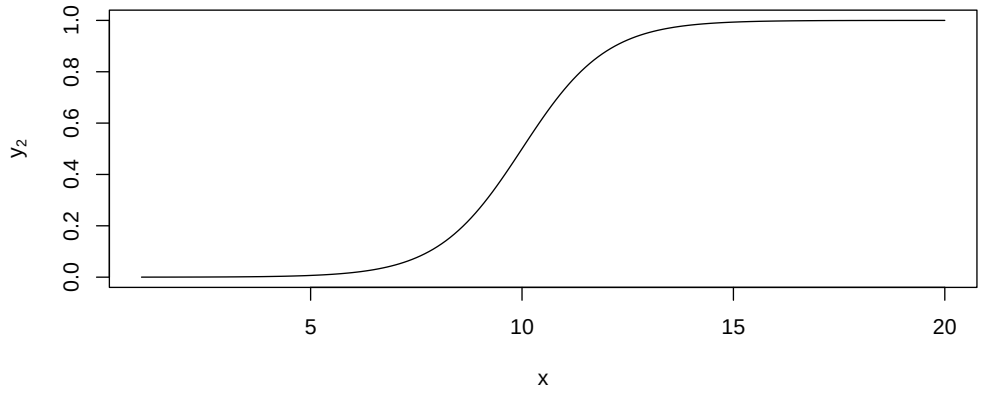
5 ☐ $\mathrm{E}(Y_1) = 0$ og $\mathrm{V}(Y_1) = b_1 + b_2$

Spørgsmål II.2 (4)

Lad en anden stokastisk variabel være defineret ved funktionen

$$Y_2 = \frac{1}{1 + \exp(a_2 + b_2 \cdot X)}$$

hvor $a_2 = 10$ og $b_2 = -1$. Et plot af denne funktion er:



Denne funktion kaldes den logistiske funktion (eller Sigmoid-funktionen).

Notationen $\mathrm{V}(Y_2|\mu = \mu_0)$ betyder variansen af Y_2, når μ er lig med μ_0. F.eks. er $\mathrm{V}(Y_2|\mu = 0)$ variansen af Y_2, når μ er lig med 0.

Hvilket af følgende udsagn er korrekt?

1 ☐ $\mathrm{V}(Y_2|\mu = 0) < \mathrm{V}(Y_2|\mu = 10) < \mathrm{V}(Y_2|\mu = 20)$

2 ☐ $\mathrm{V}(Y_2|\mu = 0) < \mathrm{V}(Y_2|\mu = 20) < \mathrm{V}(Y_2|\mu = 10)$

3 ☐ $\mathrm{V}(Y_2|\mu = 0) = \mathrm{V}(Y_2|\mu = 20) = \mathrm{V}(Y_2|\mu = 10)$

4 ☐ $\mathrm{V}(Y_2|\mu = 0) = \mathrm{V}(Y_2|\mu = 20) < \mathrm{V}(Y_2|\mu = 10)$

5 ☐ $\mathrm{V}(Y_2|\mu = 20) < \mathrm{V}(Y_2|\mu = 10) < \mathrm{V}(Y_2|\mu = 0)$

Fortsæt på side 6

Opgave III

I et statistikkursus med 589 studerende skal alle studerende til en eksamen med 2 dele, der hver varer 2 timer. De kursusansvarlige er interesseret i at vurdere, om de to dele har været lige svære ved at sammenligne pointgennemsnittene for de to dele.

Spørgsmål III.1 (5)

Hvilken test skal de kursusansvarlige anvende for at vurdere, om pointgennemsnittene for de to dele er ens?

1 ☐ En ensidet variansanalyse

2 ☐ En F-test med 2 og 589 frihedsgrader

3 ☐ En t-test med to stikprøver, hvor varianserne antages at være ens

4 ☐ En t-test med to stikprøver med en sammenvægtet varians

5 ☐ En parret t-test

Spørgsmål III.2 (6)

En af de kursusansvarlige giver det samme kursus (med den samme eksamen) på et andet universitet, hvor 240 studerende er indskrevet på kurset og tager eksamen. Nogle deskriptive nøgletal for eksamensresultaterne på de to universiteter kan findes i den nedenstående tabel:

	Universitet A	Universitet B
Studerende	589	240
Pointgennemsnit	736.4	769.9
Varians af point	169.1	402.7

Når den kursusansvarlige skal udregne 90% konfidensintervallet for forskellen i pointgennemsnit (uden at antage varianserne på de to universiteter er ens), skal den kursusansvarlige bruge en fraktil fra en t-fordeling. Hvilken fraktil og hvilket antal frihedsgrader i t-fordelingen skal den kursusansvarlige bruge i sine beregninger?

1 ☐ 10% fraktilen fra t-fordelingen med 323.93 frihedsgrader

2 ☐ 90% fraktilen fra t-fordelingen med 829 frihedsgrader

3 ☐ 90% fraktilen fra t-fordelingen med 323.93 frihedsgrader

4 ☐ 95% fraktilen fra t-fordelingen med 829 frihedsgrader

5 ☐ 95% fraktilen fra t-fordelingen med 323.93 frihedsgrader

Fortsæt på side 7

Opgave IV

Den 14. april 1912 ramte passagerskibet Titanic et isbjerg og sank den følgende dag. Nedenstående tabel viser antallet af overlevende og det samlede antal passagerer fordelt på forskellige passagerkategorier.

Klasse	1.	2.	3.	Ansatte	Total
Overlevende	202	117	178	212	709
Total	325	285	706	885	2201

Spørgsmål IV.1 (7)

Baseret på tabellen ovenfor, hvad er et 95% konfidensinterval for sandsynligheden for overlevelse (uanset passagerkategori) givet data?

1 ☐ [0.66, 0.70]

2 ☐ [0.30, 0.34]

3 ☐ [0.45, 0.50]

4 ☐ [0.46, 0.49]

5 ☐ [0.66, 0.69]

Spørgsmål IV.2 (8)

Er der en statistisk signifikant forskel i overlevelsessandsynligheden mellem besætningen og 3. klasses passagerer ved brug af et 5% signifikansniveau (både argumentet og konklusionen skal være korrekt)?

1 ☐ Ja, da teststørrelsen for den relevante test er -1.67

2 ☐ Nej, da teststørrelsen for den relevante test er 0.66

3 ☐ Ja, da teststørrelsen for den relevante test er 1.67

4 ☐ Nej, da p-værdien for den relevante test er 0.41

5 ☐ Nej, da p-værdien for den relevante test er 0.56

Fortsæt på side 9

Spørgsmål IV.3 (9)

Vi ser nu på hele tabellen. Hvad er den relevante observerede teststørrelse (q), kritiske værdi (CV) og konklusion for en test af hypotesen om, at overlevelsessandsynligheden er den samme på tværs af alle klasser, ved brug af signifikansniveau $\alpha = 0.05$?

1 ☐ q=187.1, CV=7.8, der er derfor en signifikant forskel

2 ☐ q=84.37, CV=15.5, der er derfor en signifikant forskel

3 ☐ q=84.37, CV=7.8, der er derfor en signifikant forskel

4 ☐ q=187.1, CV=15.5, der er derfor ikke en signifikant forskel

5 ☐ q=84.37, CV=7.8, der er derfor ikke en signifikant forskel

Spørgsmål IV.4 (10)

Man ønsker at teste, om 1. klasses passagerers overlevelsessandsynlighed afviger med mere end 20 procentpoint i forhold til gennemsnittet af alle andre passagerer. Hvilket af følgende udsagn er korrekt (ved brug af signifikansniveau $\alpha = 0.05$)?

1 ☐ Da $\hat{p}_{første} - \hat{p}_{andre} = 0.35$ er der en signifikant forskel og den er større end 0.2

2 ☐ Det relevante konfidensinterval er $[0.29, 0.41]$, og dermed er overlevelsessandsynligheden for 1. klasses passagerer mindst 20 procentpoint højere end andre passagerers overlevelsessandsynlighed.

3 ☐ 0.2 er ikke inkluderet i det relevante konfidensinterval, som er $[0.29, 0.41]$, og der er derfor ikke en signifikant forskel

4 ☐ Det relevante konfidensinterval er $[0.33, 0.37]$, og dermed er overlevelsessandsynligheden for 1. klasse passagerer mindst 20 procentpoint højere end andre passagerers overlevelsessandsynlighed

5 ☐ 0.2 er ikke inkluderet i det relevante konfidensinterval, som er $[0.33, 0.37]$, og der er derfor ikke en signifikant forskel

Fortsæt på side 10

Opgave V

En skoleklasse med 20 børn samler affald på en strand. Det antages, at middelværdien af det indsamlede affald er 1 kg/barn med en standardafvigelse på 0.2 kg/barn.

Spørgsmål V.1 (11)

Hvis mængden af affald, der indsamles af hvert barn, antages at være uafhængigt, hvad er da standardafvigelsen (σ) for alt det indsamlede affald?

1 ☐ $\sigma = 4.0$ kg

2 ☐ $\sigma = 0.8$ kg

3 ☐ $\sigma = 0.18$ kg

4 ☐ $\sigma = 2.0$ kg

5 ☐ $\sigma = 0.89$ kg

Spørgsmål V.2 (12)

Det besluttes nu, at børnene skal gå i par af to. Standardafvigelsen antages stadig at være 0.2 kg/barn, men nu antages det yderligere, at korrelationen mellem mængden af affald, der samles af de to elever i samme par er 0.5. Parrene antages uafhængige af hinanden. Hvad er standardafvigelsen (σ_{par}) af den samlede mængde indsamlede affald?

1 ☐ $\sigma_{par} = 1.10$ kg

2 ☐ $\sigma_{par} = 0.84$ kg

3 ☐ $\sigma_{par} = 3.60$ kg

4 ☐ $\sigma_{par} = 1.8$ kg

5 ☐ $\sigma_{par} = 1.89$ kg

Fortsæt på side 11

Spørgsmål V.3 (13)

Efter at de kom tilbage, havde et af børnene samlet 21 genstande, hvoraf 6 var lavet af plastik. Hun bliver nu bedt om at vælge 5 genstande tilfældigt, der skal diskuteres. Hvad er sandsynligheden for, at 3 af dem er lavet af plastik?

1 ☐ 0.103

2 ☐ 0.247

3 ☐ 0.119

4 ☐ 0.023

5 ☐ 0.052

Fortsæt på side 12

Opgave VI

Afdelingen for kvalitetssikring og -kontrol på en slikfabrik har udtaget en stikprøve på 26 chokoladebarer af et bestemt mærke. Efter at have vejet alle chokoladebarerne i stikprøven har man beregnet, at gennemsnitsvægten er 200.3 gram, mens den observerede standardafvigelse er 0.75 gram.

Spørgsmål VI.1 (14)

Hvad er 95% konfidensintervallet for standardafvigelsen?

1 ☐ $[0.346, 1.072]$

2 ☐ $[0.588, 1.035]$

3 ☐ $[0.611, 0.981]$

4 ☐ $[0.447, 1.053]$

5 ☐ $[0.462, 1.038]$

Spørgsmål VI.2 (15)

Slikfabrikken ønsker at teste nulhypotesen $\mathcal{H}_0 : \mu = 200$ gram (mod en tosidet alternativ hypotese) med en t-test. Hvilket af de følgende udsagn er korrekt? (Både forklaringen og konklusionen skal være korrekt)

1 ☐ Nulhypotesen forkastes på et 5% signifikansniveau, da teststørrelsen er større end $t_{0.975}(26)$

2 ☐ Nulhypotesen accepteres på et 5% signifikansniveau, da teststørrelsen er større end $t_{0.975}(26)$

3 ☐ Nulhypotesen forkastes på et 5% signifikansniveau, da teststørrelsen er større end $t_{0.975}(25)$

4 ☐ Nulhypotesen accepteres på et 10% signifikansniveau, da teststørrelsen er større end $t_{0.95}(25)$

5 ☐ Nulhypotesen forkastes på et 10% signifikansniveau, da teststørrelsen er større end $t_{0.95}(25)$

Fortsæt på side 13

Spørgsmål VI.3 (16)

Slikfabrikken planlægger et andet eksperiment for at undersøge gennemsnitsvægten yderligere. Afdelingen for kvalitetssikring og -kontrol ønsker at kunne opdage en forskel i gennemsnitsvægt på 0.3 gram (mod en tosidet alternativ hypotese), når afdelingen bruger 0.75 gram som et estimat for standardafvigelsen. Endvidere ønsker afdelingen, at både type I and type II fejlraterne ikke må være højere end 5%. Hvad er det mindste antal chokoladebarer, som skal inkluderes i eksperimentet, for at overholde afdelingens krav?

1 ☐ 10 eller 12 afhængigt af om man benytter normalfordelingsapproksimationen

2 ☐ 68 eller 70 afhængigt af om man benytter normalfordelingsapproksimationen

3 ☐ 82 eller 84 afhængigt af om man benytter normalfordelingsapproksimationen

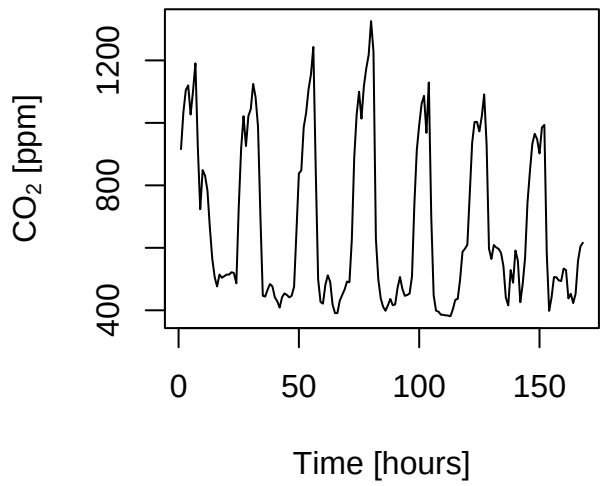
4 ☐ 97 eller 98 afhængigt af om man benytter normalfordelingsapproksimationen

5 ☐ 162 eller 164 afhængigt af om man benytter normalfordelingsapproksimationen

Fortsæt på side 14

Opgave VII

CO_2 koncentrationen er en vigtig faktor for trivsel i indemiljøet, figuren nedenfor viser timeværdier af CO_2 koncentrationen [ppm] i løbet af en uge i et rum i en bolig. Variansen af den naturlige logaritme af CO_2-koncentrationen er 0.137.



Som en indledende analyse modelleres CO_2 koncentrationen som en funktion af tidspunktet på dagen ved hjælp af modellen

$$Y_i = \beta_0 + x_{1,i}\beta_1 + x_{2,i}\beta_2 + \epsilon_i,$$

hvor Y_i er den naturlige logaritme af CO_2 koncentrationen på tidspunktet i, $\epsilon_i \sim N(0, \sigma^2)$ og iid., og

$$x_{1,i} = \sin\left(2\pi\frac{h_i}{24}\right)$$
$$x_{2,i} = \cos\left(2\pi\frac{h_i}{24}\right),$$

hvor h_i er klokkeslæt (målt i timer) for observation i.

Modellen er estimeret, og resultatet er givet nedenfor (nogle tal er erstattet af bogstaver);

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59619 -0.09527  0.03135  0.12898  0.42424
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.43959    0.01468  t1       pv1
x1           0.40303    0.02076  t2       pv2
x2           0.20019    0.02076  t3       pv3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: Sig on 165 degrees of freedom
Multiple R-squared:  R2,  Adjusted R-squared:  0.7369
F-statistic: 234.9 on 2 and 165 DF,  p-value: < 2.2e-16
```

Spørgsmål VII.1 (17)

Hvad er det samlede antal observationer brugt til estimationen?

1 ☐ 165

2 ☐ 166

3 ☐ 164

4 ☐ 167

5 ☐ 168

Spørgsmål VII.2 (18)

Hvad er rækkefølgen af p-værdierne (`pv1`, `pv2` og `pv3`) i "R-summary'et" ovenfor?

1 ☐ `pv2<pv3<pv1`

2 ☐ `pv1<pv2<pv3`

3 ☐ `pv1<pv3<pv2`

4 ☐ `pv3<pv1<pv2`

5 ☐ `pv1<pv2=pv3`

Fortsæt på side 16

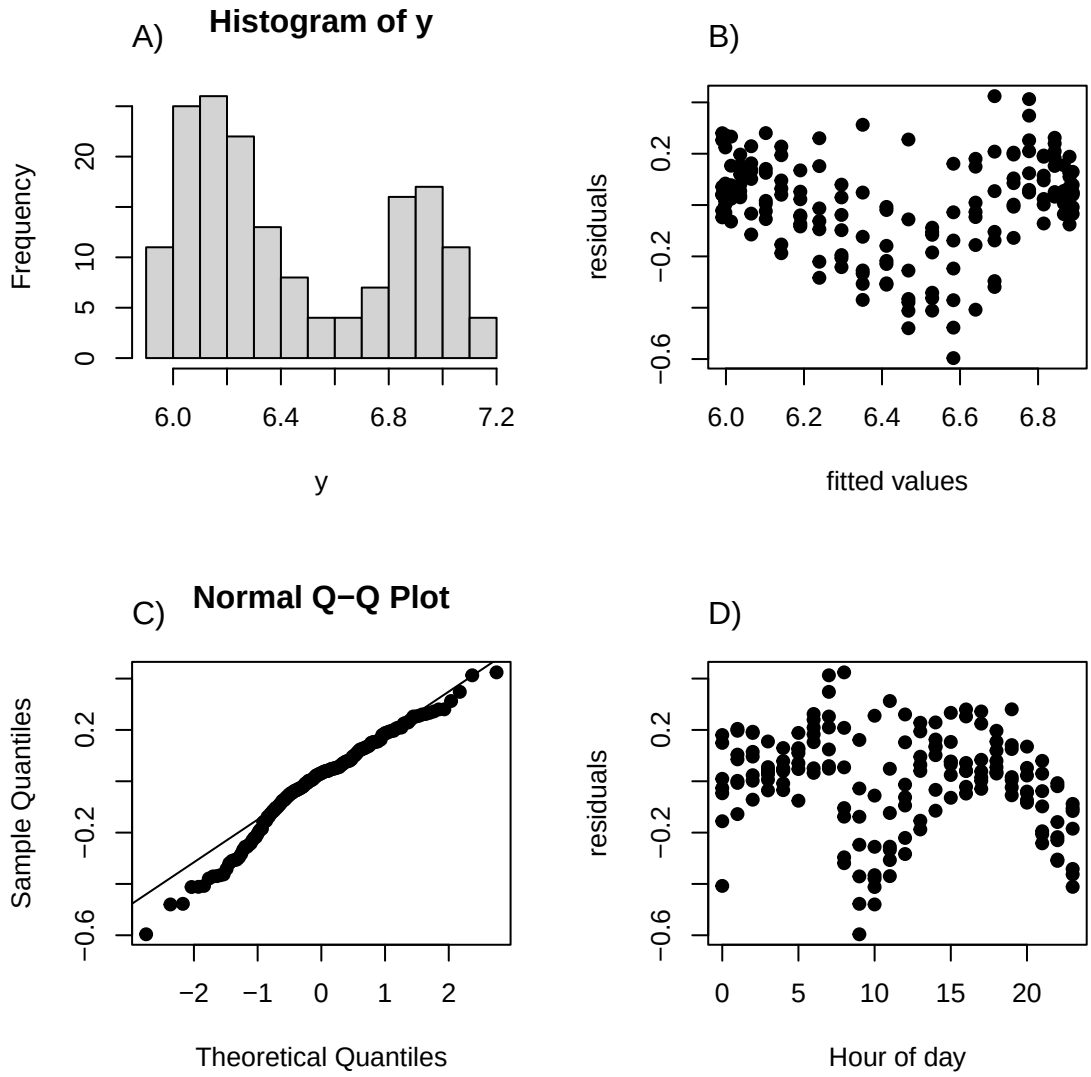
Som en del af modelvalideringen laves nedenstående figur. Figuren viser

A) Histogram of y (log-CO_2 koncentrationen)

B) Residualer som funktion af de fittedede værdier ved brug af modellen

C) Normal fraktil-fraktil plot af residualerne fra modellen

D) Residualer fra modellen som funktion af klokkeslæt



Fortsæt på side 17

Spørgsmål VII.3 (19)

Baseret på graferne i figuren, hvilket af følgende udsagn er korrekt (både udsagnet og figurhenvisningen skal være korrekte)?

1 ☐ Baseret på figur A bør vi overveje at log-transformere den afhængige variabel

2 ☐ Residualerne ser ud til at være uafhængige (figur C)

3 ☐ Normalitetsantagelsen er klart overtrådt (figur A)

4 ☐ Residualerne ser ud til at være normalfordelte (figur B)

5 ☐ Der er stadig systematiske effekter relateret til klokkeslæt (figur D)

Uanset konklusionerne fra det foregående spørgsmål, besluttes det at fortsætte undersøgelsen baseret på den udviklede model. Som en hjælp til de næste spørgsmål gives følgende relationer

$$\sum_{h=1}^{24} \sin\left(2\pi\frac{h}{24}\right) = \sum_{h=1}^{24} \cos\left(2\pi\frac{h}{24}\right) = 0$$

$$\sum_{h=1}^{24} \sin\left(2\pi\frac{h}{24}\right)\cos\left(2\pi\frac{h}{24}\right) = 0$$

$$\sum_{h=1}^{24} \sin^2\left(2\pi\frac{h}{24}\right) = \sum_{h=1}^{24} \cos^2\left(2\pi\frac{h}{24}\right) = 12$$

og som sagt er observationerne foretaget over 7 hele dage.

Spørgsmål VII.4 (20)

Med henvisning til "summary" tabellen ovenfor, hvad er `Sig`?

1 ☐ 0.27

2 ☐ 0.033

3 ☐ 0.021

4 ☐ 0.19

5 ☐ 0.056

Fortsæt på side 18

Spørgsmål VII.5 (21)

Lad nu $\hat{\sigma}$ angive den estimerede standardafvigelse, hvad er det sædvanlige 95% konfidensinterval for log-CO_2 koncentration middag ($h = 12$)?

1 ☐ $6.04 \pm 0.58\hat{\sigma}$

2 ☐ $6.24 \pm 1.97\hat{\sigma}$

3 ☐ $6.24 \pm 0.26\hat{\sigma}$

4 ☐ $6.04 \pm 0.15\hat{\sigma}$

5 ☐ $6.04 \pm 0.85\hat{\sigma}$

Spørgsmål VII.6 (22)

Hvis `x1` og `x2` blev fjernet fra modellen (altså en konstant middel-model), hvad ville standard error relateret til estimatet af β_0 så være (hint: variansen af den afhængige er angivet ovenfor)?

1 ☐ 0.0147

2 ☐ 0.00990

3 ☐ 0.00734

4 ☐ 0.0208

5 ☐ 0.0106

Fortsæt på side 19

Opgave VIII

En forsker er interesseret i at undersøge virkningerne af gødning og vandingsfrekvens på plantevækst. En tovejs ANOVA-model for disse data er:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \text{ hvor } \varepsilon_{ij} \sim N(0, \sigma^2),$$

hvor Y_{ij} er plantevæksten ved brug af det i'et gødning og det j'et vandingsfrekvens ("Dagligt", "To gange om ugen"eller "Ugentligt").

Spørgsmål VIII.1 (23)

I den statistiske model ovenfor, hvilket af følgende udsagn vedrørende α_i er korrekt (bemærk, at udsagnene er om den underliggende model ikke om statistiske test)?

1 ☐ α_i angiver effektstørrelsen for vandingsfrekvens. $\alpha_i \neq 0$ betyder, at forventet plantevækst afhænger af vandingsfrekvensen.

2 ☐ α_i angiver effektstørrelsen for vandingsfrekvens. Dette udtryk bør udelades, når plantevækst afhænger af gødningstype.

3 ☐ α_i angiver effektstørrelsen for gødning. $\alpha_i \neq 0$ indebærer, at forventet plantevækst afhænger af gødningstype.

4 ☐ α_i angiver effektstørrelsen for gødning. Dette udtryk bør udelades, når plantevækst afhænger af gødningstype.

5 ☐ α_i angiver middelværdien af den i-et gødning.

Spørgsmål VIII.2 (24)

En tovejs ANOVA blev udført. Den resulterende ANOVA-tabel er vist nedenfor. Bemærk, at p-værdierne er erstattet af spørgsmålstegn.

```
## Analysis of Variance Table

## Response: Plant_Growth
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## Fertilizer          1 8.4017  8.4017  78.766       ?
## Watering_Frequency  2 4.0133  2.0067  18.812       ?
## Residuals           2 0.2133  0.1067
```

Beregn den kritiske F-værdi for gødning og test hypotesen om ens plantevækst blandt gødninger ($\alpha = 0.05$). Hvilket af følgende udsagn er det rigtige?

1 ☐ $F_{crit} = 38.51$. Vi afviser nulhypotesen om samme plantevækst blandt gødninger, fordi $F_{obs} > F_{crit}$

2 ☐ $F_{crit} = 19$. Vi accepterer nulhypotesen om samme plantevækst blandt gødninger, fordi $F_{obs} < F_{crit}$

3 ☐ $F_{crit} = 18.51$. Vi accepterer nulhypotesen om samme plantevækst blandt gødninger, fordi $F_{obs} > F_{crit}$

4 ☐ $F_{crit} = 19$. Vi afviser nulhypotesen om samme plantevækst blandt gødninger, fordi $F_{obs} > F_{crit}$

5 ☐ $F_{crit} = 18.51$. Vi afviser nulhypotesen om samme plantevækst blandt gødninger, fordi $F_{obs} > F_{crit}$

Spørgsmål VIII.3 (25)

Hvilken af følgende kommandoer kan bruges til at vurdere, om antagelsen om normalitet er opfyldt?

1 ☐
```
lm1 <- lm(Plant_Growth~Fertilizer+Watering_Frequency, data)
qqnorm(lm1$residuals)
qqline(lm1$residuals)
```

2 ☐
```
lm1 <- lm(Plant_Growth~Fertilizer+Watering_Frequency, data)
lm1 <- anova(lm1)
qqnorm(lm1$residuals)
qqline(lm1$residuals)
```

3 ☐
```
qqnorm(data$Plant_Growth)
qqline(data$Plant_Growth)
```

4 ☐
```
qqnorm(data$Plant_Growth[data$Fertilizer=="Type 1"])
qqline(data$Plant_Growth[data$Fertilizer=="Type 1"])
```

5 ☐
```
qqnorm(rnorm(length(data)))
qqline(rnorm(length(data)))
```

Fortsæt på side 21

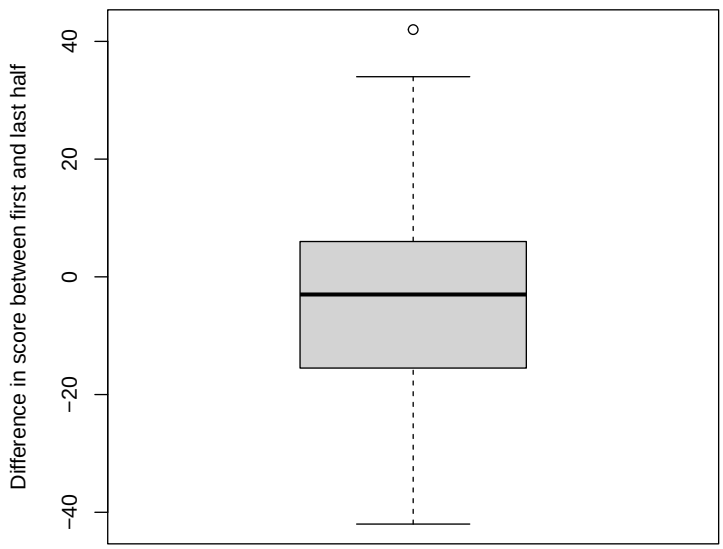
Opgave IX

Efter en multiple-choice eksamen i det indledende statistikkursus på DTU ønskede underviserne at undersøge scoren af forskellige grupper.

Et spørgsmål, de gerne ville besvare, var: Var eleverne bedre til at besvare første halvdel af eksamen (dvs. Spørgsmål 1 til 15) end sidste halvdel (Spørgsmål 16 til 30).

Lad `xfirst` være en vektor med elevernes score i første halvdel af eksamen og tilsvarende `xlast` elevernes score for anden halvdel af eksamen. De observerede forskelle i scoren mellem sidste og første halvdel for alle bestående elever er beregnet og vist med et boxplot ved:

```
x <- xlast - xfirst
boxplot(x, ylab="Difference in score between first and last half")
```



Fortsæt på side 22

Spørgsmål IX.1 (26)

Hvilken af følgende konklusioner er forkert ud fra informationen i box-plottet?

1 ☐ Mere end halvdelen af de studerende i stikprøven havde en negativ forskel i score.

2 ☐ Mere en 20% af de studerende i stikprøven havde en positiv forskel i score.

3 ☐ Mindst én studerende i stikprøven havde en forskel større end 40 point i score.

4 ☐ 60% af de studerende i stikprøven havde en positiv forskel i score.

5 ☐ Ingen studerende i stikprøven havde en forskel i score større end 50 point.

Spørgsmål IX.2 (27)

Underviserne ønsker at teste nulhypotesen

$$H_0 : \mu = 0,$$

hvor μ er middelværdien af forskellen i score mellem første og sidste del. De ønsker at teste uden brug af fordelingsantagelser om den population, hvor stikprøven er taget fra.

Følgende kode blev kørt:

```
k <- 10000
simsamples <- replicate(k, sample(x, replace = TRUE))

quantile(apply(simsamples, 2, mean), c(0.05, 0.95))

##    5%   95%
## -6.54 -1.21

quantile(apply(simsamples, 2, mean), c(0.025, 0.975))

##  2.5% 97.5%
## -7.05 -0.77

quantile(apply(simsamples, 2, mean), c(0.005, 0.995))

##  0.5% 99.5%
## -8.09  0.23
```

Hvilket af følgende svar er korrekt?

1 ☐ På et signifikansniveau $\alpha = 0.1$ detekteres en signifikant forskel i score mellem første og sidste halvdel.

2 ☐ På et signifikansniveau $\alpha = 0.025$ detekteres en signifikant forskel i score mellem første og sidste halvdel.

3 ☐ På et signifikansniveau $\alpha = 0.01$ detekteres en signifikant forskel i score mellem første og sidste halvdel.

4 ☐ Der kan ikke drages nogen konklusion, udregningerne opfylder ikke kravene, da der i beregningerne kræves at normalfordeling er antaget.

5 ☐ Ingen af ovenstående svar er korrekte.

Spørgsmål IX.3 (28)

Underviserne ønskede at undersøge, om forskellen i score mellem første og anden del af eksamen er uafhængig af den enkelte studerendes samlede score. For at undersøge dette blev de studerende opdelt i to grupper: en gruppe, der havde en lav total score, og en anden gruppe, der havde en høj total score.

Scoreforskellene for elever med lav score blev gemt i `xlow` og for højt scorende elever i `xhigh`.

Følgende kode blev kørt:

```
k <- 10000
sim.xlow.samples <- replicate(k, sample(xlow, replace = TRUE))
sim.xhigh.samples <- replicate(k, sample(xhigh, replace = TRUE))

sim.xlow.means <- apply(sim.xlow.samples, 2, mean)
sim.xhigh.means <- apply(sim.xhigh.samples, 2, mean)
sim.dif.means <- apply(sim.xhigh.samples, 2, mean) -
    apply(sim.xlow.samples, 2, mean)

quantile(sim.xlow.means, c(0.025, 0.975))

##  2.5% 97.5%
## -9.23 -2.94

quantile(sim.xhigh.means, c(0.025, 0.975))

##  2.5% 97.5%
## -3.11  1.92

quantile(sim.dif.means, c(0.025, 0.975))

##  2.5% 97.5%
##  1.41  9.56
```

Hvilken af følgende konklusioner, om de to gruppers forskel i middelværdi, er korrekt på signifikansniveau $\alpha = 0.05$ (både konklusion og argument skal være korrekt)?

1 ☐ En signifikant forskel mellem de to grupper er ikke detekteret, da deres en-stikprøve konfidensintervaller overlapper hinanden.

2 ☐ En signifikant forskel mellem de to grupper er detekteret, da deres en-stikprøve konfidensintervaller overlapper hinanden.

3 ☐ En signifikant forskel mellem de to grupper er detekteret, da det individuelle konfidensinterval for den ene gruppe inkluderer nul, men ikke for den anden gruppe.

4 ☐ En signifikant forskel mellem de to grupper er detekteret, da konfidensintervallet for forskellen i middelværdi ikke inkluderer nul.

5 ☐ Ingen af ovenstående konklusioner er korrekte.

Fortsæt på side 25

Opgave X

I en produktion af müsli tilsættes rosiner til de øvrige ingredienser i en bestemt mængde, der er velkendt. Blandingen lægges i pakker klar til salg, derfor er den forventede værdi af mængden af rosiner i hver pakke velkendt (f.eks. hvis 10 kg rosiner fordeles i 1000 pakker, så er den forventede værdi 10 g i hver pakke). Produktionsingeniøren er dog bekymret for variationen i mængden af rosiner mellem færdige pakker. Derfor udtages en prøve på 10 pakker til inspektion, og den gennemsnitlige sum af kvadrerede afvigelser mellem den kendte middelværdi og den individuelt observerede mængde af rosiner beregnes. Det antages, at mængden af rosiner i hver pakke er iid. og normalfordelt.

Den gennemsnitlige sum af kvadrerede afvigelser er defineret ved

$$\frac{1}{N_1}\sum_{i=1}^{N_1}(y_i-\mu)^2,$$

hvor, i dette tilfælde, $N_1 = 10$, y_i er mængden af rosiner i pakke i, og μ er den kendte middelværdi.

Spørgsmål X.1 (29)

Hvad er sandsynligheden for, at den observerede gennemsnitlige sum af kvadrerede afvigelser fra den (kendte) middelværdi er mindre end halvdelen af den sande varians?

1 ☐ 0.69

2 ☐ 0.11

3 ☐ 0.5

4 ☐ 0.17

5 ☐ 0.31

Fortsæt på side 26

Spørgsmål X.2 (30)

Betragt nu et andet eksperiment, der undersøger 20 pakker müsli. Det antages, at variansen er den samme i de to forsøg.

Hvis antagelsen er korrekt, hvad er så sandsynligheden for, at den observerede sum af kvadrerede afvigelser:

$$\sum_{i=1}^{N}(y_i - \mu)^2$$

i eksperiment 1 (dvs. $N = N_1 = 10$) er større end den observerede sum af kvadrerede afvigelser i eksperiment 2 (dvs. $N = N_2 = 20$)?

1 ☐ 0.090

2 ☐ 0.023

3 ☐ 0.5

4 ☐ 0.66

5 ☐ 0.097

SÆTTET ER SLUT. God sommer!