Danmarks Tekniske Universitet *Side 1 af 23 sider.*

Skriftlig prøve: 22. maj 2022

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

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

Opgave	VII.3	VIII.1	VIII.2	IX.1	IX.2	X.1	X.2	XI.1	XI.2	XI.3
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

En maskine, der producerer en bestemt type varer, undersøges. Når maskinen fungerer korrekt, er 3% af de producerede varer defekte. Antag, at vi tilfældigt udvælger ti varer, der er produceret på maskinen, og at vi er interesserede i antallet af defekte varer.

Spørgsmål I.1 (1)

Hvad er sandsynligheden for ikke at finde nogle defekte varer?

1 ☐ 0.0009

2 ☐ 0.0582

3 ☐ 0.4900

4 ☐ 0.737

5 ☐ 0.9127

Spørgsmål I.2 (2)

Hvad er antallet af defekter, hvor der er 98% eller højere sandsynlighed for, at opnå dette antal eller færre defekter i eksperimentet?

1 ☐ 1

2 ☐ 2

3 ☐ 3

4 ☐ 5

5 ☐ 8

Spørgsmål I.3 (3)

I et andet planlagt eksperiment er udfaldet beskrevet af den stokastiske variabel X. Tæthedsfunktionen for X er:

X	0	1	2	3
$f(x)$	0.1	0.3	0.4	0.2

Middelværdien er $\mathrm{E}(X) = 1.7$. Hvilket af følgende udtryk beregner variansen?

1 ☐ $\mathrm{V}(X) = 0.1 \cdot 0 + 0.3 \cdot 1 + 0.4 \cdot 2 + 0.2 \cdot 3$

2 ☐ $\mathrm{V}(X) = 0.1 \cdot 0 + 0.3 \cdot 1 + 0.4 \cdot 4 + 0.2 \cdot 9$

3 ☐ $\mathrm{V}(X) = 0.1 \cdot 2.89 + 0.3 \cdot 0.49 + 0.4 \cdot 0.09 + 0.2 \cdot 1.69$

4 ☐ $\mathrm{V}(X) = 0.1 \cdot 2.89 + 0.3 \cdot 7.29 + 0.4 \cdot 13.69 + 0.2 \cdot 22.09$

5 ☐ $\mathrm{V}(X) = 0.1 \cdot (-1.3) + 0.3 \cdot (-0.7) + 0.4 \cdot 0.3 + 0.2 \cdot 1.3$

Fortsæt på side 4

Opgave II

Det danske energiselskab Ørsted lavede i 2017 en undersøgelse i forskellige lande. Undersøgelsen handlede om folks mening om emner hidrørende klimaændringer. For hvert land blev der taget en tilfældigt udvalgt stikprøve, der var repræsentativ for befolkningen hvad angår alder, køn, region og indkomst.

Et af spørgsmålene var: "Hvor vigtigt mener du, det er at skabe en verden kun drevet af vedvarende energi?".

Lad andelen, der svarede ja til spørgsmålet i Kina, være p_1. Tilsvarende lad andelen, der svarede ja til spørgsmålet i Danmark, være p_2.

I Kina svarede $x_1 = 1920$ ja, ud af $n_1 = 2000$ personer, der blev spurgt, og i Danmark svarede $x_2 = 1801$ ja, ud af $n_2 = 2024$ personer, der blev spurgt.

Spørgsmål II.1 (4)

Hvad er estimatet af standardafvigelsen (standard error) på den estimerede andel, som svarede ja i Danmark?

1 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.00696$

2 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.0114$

3 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.0136$

4 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.0179$

5 ☐ $\hat{\sigma}_{\hat{p}_2} = 0.0834$

Spørgsmål II.2 (5)

Givet et signifikansniveau på $\alpha = 0.01$, hvad er konklusionen for den sædvanlige test for forskel i andel med nulhypotesen:

$$H_0 : p_1 = p_2$$

(både konklusion og argumentation skal være korrekt)?

1 ☐ Nulhypotesen afvises da $0.96 \neq 0.89$, derfor er de to andele signifikant forskellige.

2 ☐ Nulhypotesen accepteres da $0.96 - 0.89 > 0.01$, derfor er de to andele ikke signifikant forskellige.

3 ☐ Nulhypotesen afvises da $0 \notin [0.060, 0.081]$, derfor er de to andele signifikant forskellige.

4 ☐ Nulhypotesen accepteres da $0 \notin [0.060, 0.081]$, derfor er de to andele ikke signifikant forskellige.

5 ☐ Nulhypotesen afvises da $0 \notin [0.049, 0.091]$, derfor er de to andele signifikant forskellige.

Fortsæt på side 6

Opgave III

For at sammenligne to programmer til træning af industriansatte i at udføre en faglært opgave, indgik 20 ansatte i et forsøg. Heraf blev 10 udvalgt tilfældig og blev trænet efter "Metode 1", og de resterende 10 ansatte blev trænet efter "Metode 2". Efter træningen blev alle deltagere udsat for en test, der registrerer hastigheden af udførelsen af opgaven.

Følgende observationer i minutter blev målt (stikprøvegennemsnit og -standardafvigelse er inkluderet for hver stikprøve):

	1	2	3	4	5	6	7	8	9	10	Mean	Std. dev.
Metode 1	11.9	22.5	12.4	16.5	12.6	17.2	9.8	15.0	17.1	14.1	14.9	3.6
Metode 2	18.9	20.1	14.6	16.5	16.2	24.5	17.7	24.1	17.3	20.2	19.0	3.3

Spørgsmål III.1 (6)

Hvis vi antager, at de sande standardafvigelser for de to metoder er ens, hvad er da estimatet af den samlede (poolede) standardafvigelse?

1 ☐ $s_{\text{pooled}} = \frac{3.6+3.3}{2}$

2 ☐ $s_{\text{pooled}} = \sqrt{\frac{3.6^2+3.3^2}{2}}$

3 ☐ $s_{\text{pooled}} = \frac{3.6^2+3.3^2}{2}$

4 ☐ $s_{\text{pooled}} = \sqrt{\frac{3.6+3.3}{2}}$

5 ☐ Det er ikke muligt at beregne den samlede standardafvigelse på grund af at de to stikprøvevarianser ikke er ens, dvs. $s^2_{\text{Metode1}} \neq s^2_{\text{Metode2}}$.

Spørgsmål III.2 (7)

Vi kører en pooled t-test med stikprøverne i R og får outputtet:

```
##
##  Two Sample t-test
##
## data:  metode1 and metode2
## t = -2.6559, df = 18, p-value = 0.01609
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.3397176 -0.8562943
## sample estimates:
## mean of x mean of y
##  14.91241  19.01042
```

Hvilket udsagn er korrekt (begge dele af udsagnet skal være korrekt)?

1 ☐ Vi accepterer nulhypotesen om ens middelhastighed. Vores risiko for at lave en type I fejl er 95%.

2 ☐ Vi afviser nulhypotesen om ens middelhastighed. Vores risiko for at lave en type I fejl er 95%.

3 ☐ Vi accepterer nulhypotesen om ens middelhastighed. Vores risiko for at lave en type I fejl er 5%.

4 ☐ Vi afviser nulhypotesen om ens middelhastighed. Vores risiko for at lave en type I fejl er 5%.

5 ☐ Vi kan ikke anvende den pooled t-test under antagelsen om lige store populationsvarianser.

Spørgsmål III.3 (8)

Vi ønsker nu at planlægge et nyt eksperiment, hvor vi styrer testens styrke for at påvise signifikant forskel mellem metoderne, stadig med lige mange observationer i de to grupper. Vi ønsker at bruge et signifikansniveau på $\alpha = 1\%$ og vi ønsker at have en 98% sandsynlighed for at detektere en forskel i middelværdi på 5 minutter.

Uafhængigt af resultaterne i spørgsmålene ovenfor, vil vi bruge et gæt på populationens varians på 16.

Hvad er det mindste antal observationer, man skal tage fra hver gruppe for at opfylde ovenstående krav?

1 ☐ Mindst 12

2 ☐ Mindst 18

3 ☐ Mindst 30

4 ☐ Mindst 45

5 ☐ Mindst 62

Fortsæt på side 8

Opgave IV

En virksomhed samler maskiner af forskellige komponenter. Antag, at levetiden af komponenterne i en maskine kan modelleres uafhængigt og med den samme eksponentielle fordeling.

Spørgsmål IV.1 (9)

Hvis komponenternes gennemsnitlige levetid er 3 år, hvilken af følgende R-koder beregner sandsynligheden for, at en tilfældigt udvalgt komponent holder længere end et år?

1 ☐ `1 - dexp(0, rate=1/3)`

2 ☐ `pexp(1, rate=3)`

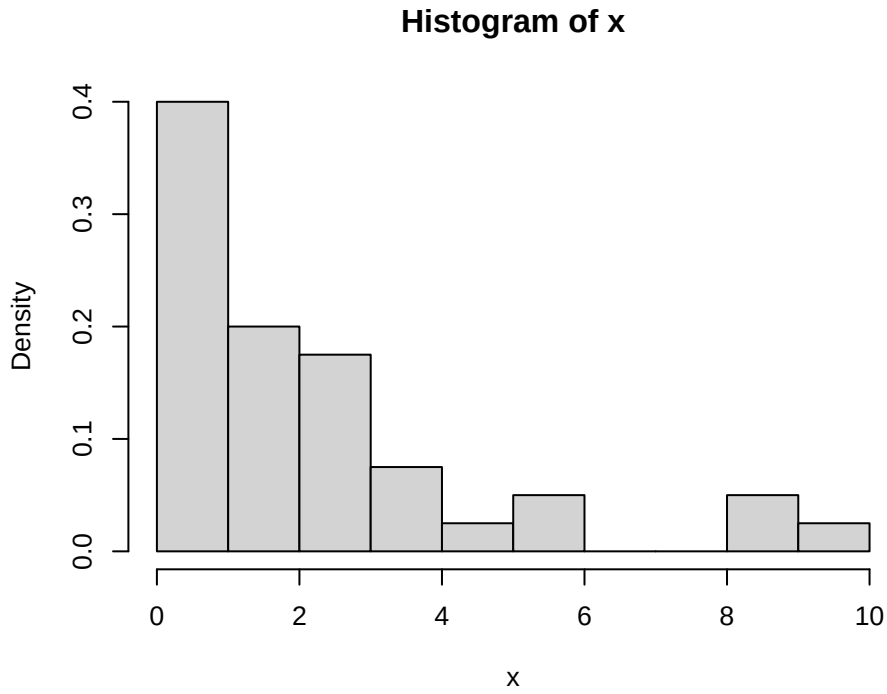
3 ☐ `1 - pexp(0, rate=1/3)`

4 ☐ `1 - pexp(1, rate=1/3)`

5 ☐ `dexp(0, rate=3)`

Spørgsmål IV.2 (10)

Et eksperiment blev udført ved at måle hvornår komponenter holdt op med at virke, når de blev udsat for et accelereret aldringsmiljø. Et tæthedshistogram af en taget stikprøve er plottet nedenfor:



Hvad er antallet af observationer i stikprøven?

1 ☐ $n = 80$

2 ☐ $n = 160$

3 ☐ $n = 320$

4 ☐ $n = 480$

5 ☐ Dette kan ikke afgøres med den givne information.

Spørgsmål IV.3 (11)

Et parametrisk bootstrappet 95% konfidensinterval for middelværdien blev beregnet med R-koden nedenfor. Den oprindelige stikprøve var blevet indlæst i vektoren `x`:

```
# Set the number of simulations:
k <- 100000
# Simulate k samples
simSamples <- replicate(k, rexp(length(x), 1/mean(x)))
# Compute the simulated means
simMean <- apply(simSamples, 2, mean)
# Quantiles for the confidence interval
quantile(simMean, c(0.025, 0.975))

##     2.5%    97.5%
## 1.561813 2.914021
```

Baseret på denne analyse, hvad er konklusionen på en test af nulhypotesen

$$H_0 : \mu = 2$$

på signifikansniveau $\alpha = 0.05$ (både argumentet og konklusionen skal være korrekt)?

1 ☐ Nulhypotesen accepteres, da $2 \in [1.56, 2.91]$, derfor konkluderer vi at middelværdien måske er 2.

2 ☐ Nulhypotesen accepteres, da $2 \in [1.56, 2.91]$, derfor konkluderer vi at middelværdien er 2.

3 ☐ Nulhypotesen afvises, da $2 \in [1.56, 2.91]$, derfor konkluderer vi at middelværdien måske er 2.

4 ☐ Nulhypotesen afvises, da $2 \in [1.56, 2.91]$, derfor konkluderer vi at middelværdien er 2.

5 ☐ Nulhypotesen afvises, da $2 \in [1.56, 2.91]$, derfor konkluderer vi at middelværdien er forskellig fra 2.

Fortsæt på side 10

Opgave V

Denne opgave handler om at beregne standardafvigelse og varians for funktioner af stokastiske variable.

Spørgsmål V.1 (12)

Beregnet med simulering, hvad er standardafvigelsen af Y tilnærmelsesvist når

$$Y = e^{X_1} + X_2^4 + X_1 \cdot X_2$$

hvor X_1 og X_2 er uafhængige og begge følger en standard normalfordeling?

1 ☐ $\sigma_Y \approx 3.3$

2 ☐ $\sigma_Y \approx 10$

3 ☐ $\sigma_Y \approx 100$

4 ☐ $\sigma_Y \approx 920$

5 ☐ $\sigma_Y \approx 9800$

Spørgsmål V.2 (13)

Lad Y være defineret ved

$$Y = X_1^3 + 5X_2$$

De to stokastiske variabler X_1 og X_2 er uafhængige og har standardafvigelserne, henholdsvis, σ_1 og σ_2. Lad x_1 og x_2 være observationer af henholdsvis X_1 og X_2

Hvad er den lineære tilnærmelse til variansen af Y, udledt ved hjælp af fejl-udbredelses metoden (propagation of error method)?

1 ☐ $\mathrm{V}(Y) \approx 9x_1^4\sigma_1 + 25\sigma_2$

2 ☐ $\mathrm{V}(Y) \approx 3x_1^2\sigma_1^2 + 5\sigma_2^2$

3 ☐ $\mathrm{V}(Y) \approx 9x_1^4\sigma_1^2 + 25\sigma_2^2$

4 ☐ $\mathrm{V}(Y) \approx 9x_1^2\sigma_1 + 25x_2\sigma_2$

5 ☐ $\mathrm{V}(Y) \approx 3x_1^4\sigma_1 + 5x_2\sigma_2$

Fortsæt på side 11

Opgave VI

12 observationer af mangan (Mn) ved seks forskellige koncentrationer blev analyseret med *inductively coupled plasma atomic emission spectroscopy* (ICP-AES).

Koncentrationerne af mangan er målt i ppb (parts per billion / miliardtedele, 10^{-9}). Data læstes ind i R ved:

```
# Mangankoncentrationer
x <- c(0, 0, 2, 2, 4, 4, 6, 6, 8, 8, 10, 10)
# ICP-AES værdier
y <- c(114, 14, 870, 1141, 2087, 2212, 3353, 2633, 3970, 4299, 4950, 5207)
```

En lineær regression udførtes

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ hvor } \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d. for } i = 1, \ldots, 12.$$

hvor Y_i er ICP-AES værdien og x_i er mangankoncentrationen for den i'te observation.

Vi antager implicit at modellens antagelser er opfyldt.

Spørgsmål VI.1 (14)

Hvad er estimatet for β_1?

1 ☐ 49.2

2 ☐ 504.3

3 ☐ 511.0

4 ☐ 520.7

5 ☐ 2570

Spørgsmål VI.2 (15)

Man vil gerne kende usikkerheden i ICP-AES værdien for en ny observation, hvor mangankoncentrationen er 5 ppb. Hvad er 95% prædiktionsintervallet ved denne koncentration?

1 ☐ [2087, 3054]

2 ☐ [2437, 2705]

3 ☐ [465, 544]

4 ☐ [2388, 2656]

5 ☐ [2038, 3005]

Spørgsmål VI.3 (16)

Vi ønsker at teste hypotesen $H_0 : \beta_0 = 0$, da dette vil kunne indikere om den forventede ICP-AES værdi er nul, når mangankoncentrationen er 0 ppb.

Hvilket af følgende udsagn er korrekt?

1 ☐ Vi accepterer nulhypotesen, da p-værdien er 0.006.

2 ☐ Vi afviser nulhypotesen, da p-værdien er 0.006.

3 ☐ Vi accepterer nulhypotesen, da $|1 - \hat{\beta}_0|$ er mindre end standardafvigelsen.

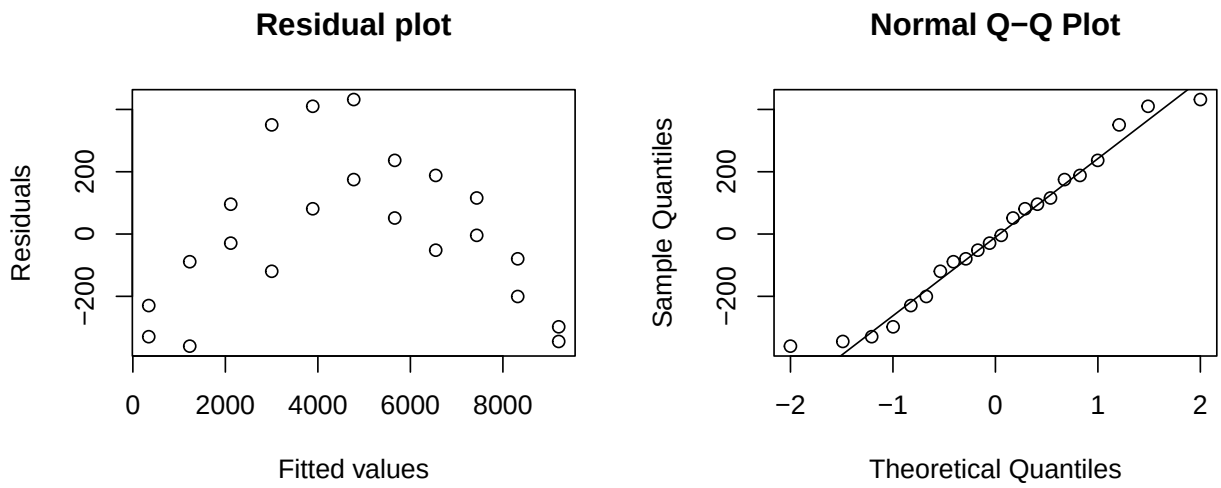
4 ☐ Vi accepterer nulhypotesen, da p-værdien er 0.655.

5 ☐ Vi afviser nulhypotesen, da p-værdien er 0.655.

Spørgsmål VI.4 (17)

Vi modtager efterfølgende data med højere mangankoncentrationer (op til 20 ppb), og en ny lineær regession $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ blev udført.

De viste plots herunder er henholdvis et residualplot og et normal q-q plot af residualerne:



Hvilket af følgende udsagn er den korrekte fortolkning af disse to plots?

1 ☐ Residualplottet er tvivlsomt. Dette indikerer et problem med normalitetsantagelsen.

2 ☐ Residualplottet er tvivlsomt. Dette indikerer et problem med antagelsen om lineær sammenhæng.

3 ☐ Residualplottet ser (rimeligt) fornuftigt ud, men q-q plottet er tvivlsomt. Dette indikerer et problem med antagelsen om lineær sammenhæng.

4 ☐ Der ses ikke en lineær tendens i residualplottet. Dette er evidens for nulhypotesen om at koncentration ikke har en signifikant effekt på ICP-AES værdien.

5 ☐ Hverken residualplottet eller q-q plottet er relateret til modellens validetet eller de tilhørende nulhypoteser.

Spørgsmål VI.5 (18)

Til sidst fittes den kurvelineære model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ på de nye data. De nye data er gemt i data.framen `Mangan2`. Resultatet er:

```
x2 <- Mangan2$x^2
fit <- lm(y ~ x + x2, data = Mangan2)
summary(fit)

##
## Call:
## lm(formula = y ~ x + x2, data = Mangan2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -251.95  -63.95   17.22   69.20  218.05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.6469    74.8293   0.169    0.868
## x           553.4783    17.4075  31.795  < 2e-16 ***
## x2           -5.5157     0.8383  -6.580 2.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138.9 on 19 degrees of freedom
## Multiple R-squared:  0.9979,Adjusted R-squared:  0.9977
## F-statistic:  4500 on 2 and 19 DF,  p-value: < 2.2e-16
```

Hvilket af de følgende udsagn er korrekt, givet et signifikansniveau på $\alpha = 1\%$?

1 ☐ Tilføjelsen af det kvadratiske led fører ikke til en signifikant forbedringen af modellen, da p-værdien for β_1 er mindre end p-værdien for β_2.

2 ☐ ICP-AES værdien forøges i gennemsnit med $(-5.5)^2 = 30.25$, når koncentrationen stiger med 1 ppb.

3 ☐ ICP-AES værdien forøges i gennemsnit med 553.5, når koncentrationen stiger med 1 ppb.

4 ☐ Hverken β_1 eller β_2 er signifikant forskellig fra nul, da de tilhørende p-værdier er mindre end 0.01.

5 ☐ Vi kan med modellen forklare mere end 99% af den observerede variation i data.

Fortsæt på side 15

Opgave VII

I overgangen til et CO_2 udledningsfrit energisystem skal fossile brændstoffer udfases. For eksempel skal opvarmning med naturgas udskiftes med andre kilder, f.eks. til fjernvarme baseret på vedvarende kilder.

Når det er besluttet, om fjernvarme skal etableres i et nyt område afholdes et informationsmøde. På mødet orienterer fjernvarmeselskabet om økonomi og mulighederne for tilslutning til fjernvarmenettet og sammenligner med andre alternativer.

Der er blevet gennemført en undersøgelse blandt husejere, der deltog i et informationmøde, for at afgøre, om mødet ændrede deres mening om at tilslutte sig fjernvarmen.

Deres meninger blev indsamlet anonymt før og efter mødet. Adspurgt om de ville tilslutte sig fjernvarmen var deres svar:

	Før	Efter	Sum
Ja	18	28	46
Nej	22	14	36
Ikke besluttet	25	17	42
Sum	65	59	124

Den sædvanlige nulhypotese, at andelen var den samme før og efter, er:

$$H_0 : p_{i,1} = p_{i,2}, \text{ for alle rækker } i = 1, 2, 3.$$

Spørgsmål VII.1 (19)

Hvad er det forventede antal personer under nulhypotesen, der svarer 'Ikke besluttet' efter informationsmødet?

1 ☐ $17/124$

2 ☐ $25 \cdot 59/124$

3 ☐ $42 \cdot 17/59$

4 ☐ $59 \cdot 42/124$

5 ☐ $17 \cdot 25/59$

Spørgsmål VII.2 (20)

Den følgende R kode blev kørt:

```
chisq.test(matrix(c(18, 28, 22, 14, 25, 17), ncol = 2, byrow = TRUE),
           correct=FALSE)

##
##  Pearson's Chi-squared test
##
## data:  matrix(c(18, 28, 22, 14, 25, 17), ncol = 2, byrow = TRUE)
## X-squared = 5.1973, df = 2, p-value = 0.07437
```

Hvad er den korrekte konklusion om nulhypotesen når der testes på signifikansniveau $\alpha = 0.05$?

1 ☐ Nulhypotesen afvises da p-værdien er over signifikansniveauet.

2 ☐ Nulhypotesen afvises da p-værdien er under signifikansniveauet.

3 ☐ Nulhypotesen accepteres da p-værdien er over signifikansniveauet.

4 ☐ Nulhypotesen accepteres da p-værdien er under signifikansniveauet.

5 ☐ Ingen af ovenstående konklusioner er korrekte, da der ikke er givet nok information til at drage en konklusion.

Spørgsmål VII.3 (21)

Hvad er det kritiske niveau, dvs. en χ^2-fraktil, for test af nulhypotesen på signifikansniveau $\alpha = 0.01$?

1 ☐ 9.21

2 ☐ 2.32

3 ☐ 1.96

4 ☐ 0.196

5 ☐ 0.103

Fortsæt på side 17

Opgave VIII

En brochure, der inviterer til at abonnere på en slankekur, fortæller, at deltagere forventes at tabe 23 pund på fem uger. Lad X være vægttab. Fra data om fem ugers vægttab for $n = 56$ deltagere er stikprøvegennemsnittet og -standardafvigelsen beregnet til at være $\bar{x} = 21.5$ og $s = 9.8$ pund.

For at undersøge påstanden om vægttab skal hypotesen

$$H_0 : \mu = 23$$

testes med de indhentede data.

Spørgsmål VIII.1 (22)

Hvilket af følgende udsagn er korrekt, når man anvender signifikansniveau $\alpha = 0.05$ (både argument og konklusion skal være korrekt)?

1 ☐ 95% konfidensintervallet er $[18.88, 24.12]$. Det hypoteserede vægttab på 23 pund er indeholdt i intervallet, hvorfor udsagnet kan underbygges.

2 ☐ 95% konfidensintervallet er $[18.88, 24.12]$. Det hypoteserede vægttab på 23 pund er indeholdt i intervallet, hvorfor udsagnet IKKE kan underbygges.

3 ☐ 95% konfidensintervallet er $[19.30, 23.69]$. Det hypoteserede vægttab på 23 pund er indeholdt i intervallet, hvorfor udsagnet kan underbygges.

4 ☐ 95% konfidensintervallet er $[19.30, 23.69]$. Det hypoteserede vægttab på 23 pund er indeholdt i intervallet, hvorfor udsagnet IKKE kan underbygges.

5 ☐ 95% konfidensintervallet er $[20.88, 22.12]$. Det hypoteserede vægttab på 23 pund er IKKE indeholdt i intervallet, hvorfor udsagnet kan underbygges.

Spørgsmål VIII.2 (23)

Hvad er værdien af teststatistikken som bruges til at teste hypotesen?

1 ☐ $t_{\text{obs}} = 1.135$

2 ☐ $t_{\text{obs}} = -1.135$

3 ☐ $t_{\text{obs}} = -1.145$

4 ☐ $t_{\text{obs}} = 16.42$

5 ☐ $t_{\text{obs}} = -16.42$

Fortsæt på side 18

Opgave IX

Mere end to millioner besøger DTU's hjemmeside hver måned, hvilket gør det muligt for DTU at tiltrække forskere, studerende og andre. Hjemmesiden har i gennemsnit syv besøgende i minuttet. Det antages, at gennemsnittet er konstant.

Spørgsmål IX.1 (24)

Hvad er sandsynligheden for, at der er to eller flere besøgende på hjemmesiden i en tilfældigt udvalgt periode på et minut?

1 ☐ 0.09

2 ☐ 0.64

3 ☐ 0.77

4 ☐ 0.97

5 ☐ 0.99

Spørgsmål IX.2 (25)

Hvad er sandsynligheden for, at der ikke er besøgende i en tilfældigt udvalgt periode på 30 sekunder?

1 ☐ 0.03

2 ☐ 0.07

3 ☐ 0.18

4 ☐ 0.43

5 ☐ 0.78

Fortsæt på side 19

Opgave X

Følgende målinger er taget for 3 grupper:

Group 1	Group 2	Group 3
1.89	3.15	1.54
2.35	2.16	2.02
1.68	2.40	2.01
2.11	2.59	2.11

Data kan læses ind i R ved at bruge følgende kommando:

```
y <- c(1.89, 2.35, 1.68, 2.11, 3.15, 2.16, 2.40, 2.59, 1.54, 2.02, 2.01, 2.11)
```

Spørgsmål X.1 (26)

Den samlede middelværdi $\bar{y}$ og middelværdien inden for hver gruppe $\bar{y}_i$ (for $i = 1, 2, 3$) er angivet nedenfor:

$\bar{y}$	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$
2.17	2.01	2.58	1.92

Vi udfører en envejs variansanalyse (ANOVA). Hvad er total sum of squares (SST), treatment sum of squares (SS(Tr)) og sum of squared errors (SSE)?

1 ☐ $SST = 0.18$, $SS(Tr) = 0.51$, $SSE = 0.11$

2 ☐ $SST = 0.31$, $SS(Tr) = 0.42$, $SSE = 0.88$

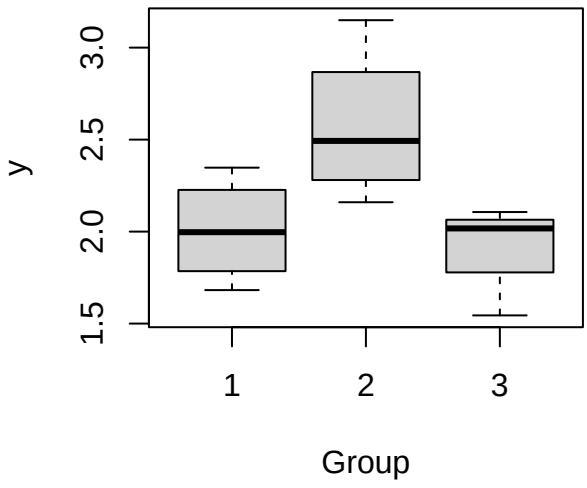
3 ☐ $SST = 2.50$, $SS(Tr) = 2.12$, $SSE = 0.38$

4 ☐ $SST = 1.99$, $SS(Tr) = 1.01$, $SSE = 0.98$

5 ☐ $SST = 4.12$, $SS(Tr) = 0.75$, $SSE = 3.37$

Spørgsmål X.2 (27)

Data præsenteret ovenfor er blevet visualiseret ved hjælp af et boxplot:



Hvilket af følgende udsagn er sandt?

1 ☐ De sorte linjer i boksene angiver gennemsnittet af hver stikprøve.

2 ☐ Medianerne er cirka 2.0, 2.5 og 2.0 for henholdsvis gruppe 1, 2 og 3.

3 ☐ Boksbredden defineres som forskellen mellem øvre og nedre kvartil, altså forskellen mellem 95. og 5. percentil.

4 ☐ Boksbredden er defineret som forskellen mellem øvre og nedre kvartil, altså forskellen mellem 90. og 10. percentil.

5 ☐ Whiskers på boxplot definerer Interquartile Range, dvs. $IQR = Q3 - Q1$.

Fortsæt på side 21

Opgave XI

Havres ernæringsmæssige kvalitet undersøges ved at underkaste 6 sorter af havrekerner uden skal en mineralanalyse. Planterne dyrkes under fire forskellige behandlinger i et randomiseret blokdesign, og målinger af protein i procent af tørvægt registreres ved høsttidspunktet.

En tovejs ANOVA-model for disse data er

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

hvor Y_{ij} er det relative proteinindhold i den i'te havrekernevariant ved den j'te behandling, og α_i og β_j repræsenterer effektstørrelser svarende til henholdsvis havrevariant og behandling.

Modelresultatet er angivet i ANOVA-tabellen nedenfor (bemærk, at nogle værdier er blevet erstattet af spørgsmålstegn):

```
## Analysis of Variance Table
## Response: protein
##            Df    Sum Sq  Mean Sq   F value   Pr(>F)
## havre      5     2.2060  0.44120   4.2367    0.01333 *
## behandling ?     0.2554  0.08513   ?         0.50410
## Residuals  15    1.5620  0.10414
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Desuden er de estimerede effektstørrelser:

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimeret effekt	-0.09	0.59	0.37	-0.22	0.53	0.35	0.24	0.37	0.30	0.09

Spørgsmål XI.1 (28)

Den samlede middelværdi $\hat{\mu} = \bar{y} = 5.6$ er blevet estimeret ud fra dette data. Givet resultaterne i ANOVA-tabellen og estimerede effektstørrelser, hvad er det forventede (forudsagte) proteinindhold for havrekerne af variant 2 ved behandlingsniveau 1 og 4, når kun signifikante effekter tages i betragtning ved signifikansniveau 5%?

1 ☐ $\hat{y}_{21} = 2 \cdot 2.2060$ og $\hat{y}_{24} = 2 \cdot 2.2060$

2 ☐ $\hat{y}_{21} = 2 \cdot 2.2060 + 1 \cdot 0.2554$ og $\hat{y}_{24} = 2 \cdot 2.2060 + 4 \cdot 0.2554$

3 ☐ $\hat{y}_{21} = 0.59$ og $\hat{y}_{24} = 0.59$

4 ☐ $\hat{y}_{21} = 5.6 + 0.24$ og $\hat{y}_{24} = 5.6 + 0.09$

5 ☐ $\hat{y}_{21} = 5.6 + 0.59$ og $\hat{y}_{24} = 5.6 + 0.59$

Spørgsmål XI.2 (29)

Nogle af elementerne i ANOVA-tabellen ovenfor er blevet erstattet af spørgsmålstegn. Hvilket af følgende udsagn er korrekt for behandlingen?

1 ☐ Frihedsgrader er 4 og den observerede teststatistik er 0.8175.

2 ☐ Frihedsgrader er 3 og den observerede teststatistik er 0.8175.

3 ☐ Frihedsgrader er 4 og den observerede teststatistik er 0.4087.

4 ☐ Frihedsgrader er 3 og den observerede teststatistik er 0.4087.

5 ☐ Frihedsgrader er 4 og den observerede teststatistik er 1.960.

Spørgsmål XI.3 (30)

Ved test af nulhypotesen $H_{0,\text{Havre}} : \alpha_i = 0$, hvor $i = 1, 2, ..., k$, hvilken implikation angivet nedenfor er korrekt på et signifikansniveau $\alpha = 0.05$?

1 ☐ Sandsynligheden for at lave en type I fejl er 1.33%.

2 ☐ Sandsynligheden for at lave en type I fejl er 95%.

3 ☐ Sandsynligheden for at lave en type II fejl er 98.67%.

4 ☐ Sandsynligheden for at teststatistikken er højere end den observerede teststatistik er 98.67% givet at nulhypotesen er falsk.

5 ☐ Sandsynligheden for at teststatistikken er højere end den observerede teststatistik er 1.33% givet at nulhypotesen er sand.

Fortsæt på side 23

SÆTTET ER SLUT. God sommer!