**Danmarks Tekniske Universitet** *Side 1 af 25 sider.*

*Skriftlig prøve*: 18. december 2021

*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" 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 | I.3 | II.1 | II.2 | II.3 | III.1 | 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 | VI.1 | VI.2 | VI.3 | VII.1 | VII.2 | VII.3 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Spørgsmål** | (11) | (12) | (13) | (14) | (15) | (16) | (17) | (18) | (19) | (20) |
| **Svar** | | | | | | | | | | |

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

Eksamenssættet består af 25 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

Af forskellige årsager kan det være interessant at analysere fødselsdata - for eksempel hvis man er interesseret i, om antallet af fødsler afhænger af årstiden. Fødselsdata fra Danmark er tilgængelig fra Danmarks Statistik. Faktisk kan man ved simpel visuel inspektion af de tilgængelige fødselsdata let se, at der er flere fødsler om sommeren end om vinteren. Det er imidlertid ikke klart, om der er en forskel mellem forår og efterår.

For at undersøge denne forskel, blev antallet af fødsler hvert forår og efterår i perioden 2007 til 2020 downloadet. Forskellene hvert år er anført nedenfor (en positiv forskel betyder flere fødsler om efteråret):

| 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 826 | 435 | -504 | 247 | -211 | 357 | -570 | 601 | 1459 | 770 | 830 | -156 | 748 | 309 |

Stikprøvegennemsnittet er $\bar{x} = 367.2$ og stikprøvestandardafvigelsen er $s = 571.5$.

### Spørgsmål I.1 (1)

Hvordan beregnes 95% konfidensintervallet for middelforskellen mellem forår og efterår?

1 ☐ $367.2 \pm 2.14\sqrt{\frac{571.5}{14}}$

2 ☐ $367.2 \pm 2.16\sqrt{\frac{571.5^2}{14}}$

3 ☐ $571.5 \pm 2.14\sqrt{\frac{367.2}{14}}$

4 ☐ $571.5 \pm 1.96\sqrt{\frac{367.2^2}{14}}$

5 ☐ $367.2 \pm 1.96\sqrt{\frac{571.5}{14}}$

### Spørgsmål I.2 (2)

En test for ingen forskel i middelværdi mellem forår og efterår skal udføres på signifikansniveau 5%. Hvilken konklusion kan drages af denne test baseret på de givne oplysninger (både konklusion og argument skal være korrekt)?

1 ☐ Da $p$-værdien er mindre end 5% kan der påvises en signifikant højere middelværdi i efteråret.

2 ☐ Da $p$-værdien er større end 5% kan der ikke påvises en signifikant forskel i middelværdi.

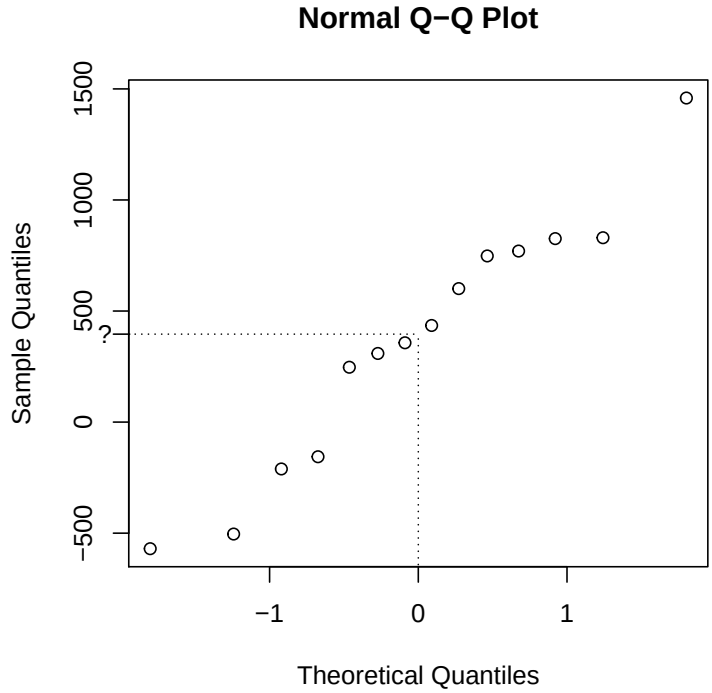
3 ☐ Da $p$-værdien er mindre end 5% kan der påvises en signifikant mindre middelværdi i efteråret.

4 ☐ Da $p$-værdien er større end 5% kan der påvises en signifikant forskel i middelværdi.

5 ☐ Da $p$-værdien er mindre end 1% kan der påvises en signifikant mindre middelværdi i efteråret.

**Spørgsmål I.3 (3)**

I modelvalideringen blev følgende normal q-q plot af stikprøven lavet til kontrol af antagelsen om normalfordeling af populationen.




På y-aksen er en værdi markeret med "?". Det er den værdi, som er angivet med de stiplede linjer, er midt imellem de to punkter, der ligger lige på hver sin side af nul på x-aksen.

Hvad kaldes denne værdi?

1 ☐ Stikprøvens første kvartil.

2 ☐ Stikprøvens tredje kvartil.

3 ☐ Stikprøvegennemsnittet.

4 ☐ Stikprøvens median.

5 ☐ Stikprøvens middelfraktilbredde (Inter Quartile Range (IQR)).

Fortsæt på side 5

## Opgave II

En stikprøve blev taget, hvor 0 angiver en ikke-succes og 1 angiver en succes. Stikprøven består af 14 observationer af værdi 0 og 18 observationer af værdi 1.

### Spørgsmål II.1 (4)

Hvad er stikprøvestandardafvigelsen?

1 ☐ $s = 0.254$

2 ☐ $s = 0.496$

3 ☐ $s = 0.504$

4 ☐ $s = 15.6$

5 ☐ $s = 16.1$

### Spørgsmål II.2 (5)

Stikprøven var fra et binomialeksperiment. Parameteren $p$ er sandsynligheden for en succes.

Hvad er estimatet af $p$?

1 ☐ 0.3164

2 ☐ 0.1914

3 ☐ 0.5625

4 ☐ 0.4375

5 ☐ 18

### Spørgsmål II.3 (6)

Givet $p = 0.5$, hvad er sandsynligheden for at observere 18 eller flere succeser i en ny stikprøve af samme størrelse?

1 ☐ 0.298

2 ☐ 0.811

3 ☐ 0.702

4 ☐ 0.189

5 ☐ 0.110

Fortsæt på side 6

**Opgave III**

## Spørgsmål III.1 (7)

Man ønsker at simulere 35 udtrækninger fra en normalfordeling med middelværdi = -2 og varians = 4. Hvilket af følgende stykker kode gør ikke dette?

1 ☐ `rnorm(n = 35, mean = -2, sd = 2)`

2 ☐ `-2 + rnorm(n = 35, mean = 0, sd = 2)`

3 ☐ `rnorm(n = 35, mean = -2, sd = 1) * 4`

4 ☐ `-2 + rnorm(n = 35, mean = 0, sd = 1) * (-2)`

5 ☐ `-2 + rnorm(n = 35, mean = 0, sd = 4) / 2`

Fortsæt på side 8

## Opgave IV

I et forsøg undersøgte forskere 20 planter af samme toårige art. De målte diameteren af plantens grundstamme (i mm) ved rodtoppen, samt mængden af frugt.

Data er gemt i R som `root`, og en lineær regression udførtes, $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

### Spørgsmål IV.1 (8)

Som det første udregner forskerne stikprøvekorrelationen til at være $r = 0.953$. Hvilket af følgende udsagn er korrekt (både konklusion og argument skal være korrekt)?

1 ☐ Der er en stærk positiv sammenhæng mellem grundstammens størrelse og frugtmængden. Et test af hypotesen $\beta_1 = 0$ vil kunne bekræfte om denne sammenhæng er signifikant eller ej.

2 ☐ Der er ikke en sammenhæng mellem grundstammens størrelse og frugtmængden, da $r$ er mellem $\pm 1.96$, hvor 1.96 er 97.5%-fraktilen i en standard-normalfordeling, $N(0, 1)$.

3 ☐ Der er en stærk sammenhæng mellem grundstammens størrelse og frugtmængden, men vi kan ikke sige om denne sammenhæng er positiv eller negativ.

4 ☐ Der er en negativ sammenhæng mellem grundstammens størrelse og frugtmængden. Derfor er både x og y normalfordelte.

5 ☐ Stikprøvekorrelationen er ikke en indikator for evt. sammenhæng mellem grundstammens størrelse og frugtmængden.

Som det næste køres følgende R-kode. Noget af outputtet er blevet erstattet med `x`:

```
summary(lm(fruit ~ diameter, data = root))

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -125.28      14.56       x        x       x
## diameter       23.25       1.74       x        x       x
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.761 on 18 degrees of freedom
```

### Spørgsmål IV.2 (9)

Den gennemsnitlige grundstammediameter var $\bar{x} = 8.31$ mm. Udregn den gennemsnitlige frugtmængde, $\bar{y}$.

1 ☐ 29.0 mg

2 ☐ 31.6 mg

3 ☐ 67.9 mg

4 ☐ 193.2 mg

5 ☐ 318.5 mg

**Spørgsmål IV.3 (10)**

Udregn 95% konfidensintervallet for $\beta_1$. Du kan også benytte at $S_{\text{xx}} = 19.90$.

1 ☐ [19.59, 26.91]

2 ☐ [19.68, 26.82]

3 ☐ [19.84, 26.66]

4 ☐ [20.23, 26.27]

5 ☐ [22.43, 24.07]

**Spørgsmål IV.4 (11)**

I dette forsøg måltes grundstammerne ved deres diameter. Hvis vi antager at stammerne er cirkulære, så er forholdet mellem areal og diameter givet ved følgende formel:

$$\text{areal} = \tfrac{\pi}{4} \cdot \text{diameter}^2$$

hvor $\pi \approx 3.1416$. Diameteren af en grundstamme måltes til at være 9.60 mm. Standardafvigelsen på målingen er $\sigma = 0.05$ mm. Ved brug af reglen for fejludvikling (error propagation rule), approksimér da standardafvigelsen på målet af denne grundstammes areal:

1 ☐ 0.05 $\text{mm}^2$

2 ☐ 0.754 $\text{mm}^2$

3 ☐ 0.960 $\text{mm}^2$

4 ☐ 1.61 $\text{mm}^2$

5 ☐ 3.72 $\text{mm}^2$

Forskerne udvidede derefter deres forsøg til også at inkludere effekten af græssende dyr, hvilket her måles på skala fra 0 til 1. Data er gemt i `root2`, og den følgende R-kode blev kørt:

```r
summary(lm(fruit ~ diameter + grazing, data = root2))
```

```
##
## Call:
## lm(formula = fruit ~ diameter + grazing, data = root2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1964  -2.8268   0.3196   3.9146  17.3264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -91.774      7.118  -12.89 2.95e-15 ***
## diameter      23.568      1.149   20.51  < 2e-16 ***
## grazing      -36.114      3.358  -10.75 6.10e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.749 on 37 degrees of freedom
## Multiple R-squared:  0.9291,Adjusted R-squared:  0.9252
## F-statistic: 242.3 on 2 and 37 DF,  p-value: < 2.2e-16
```

### Spørgsmål IV.5 (12)

Betragt R-outputtet ovenfor. Hvilket af de følgende udsagn er korrekt, givet et signifikansniveau på $\alpha = 1\%$?

1 ☐ Grundstammens størrelse ser ud til at have en signifikant effekt på frugtmængde, mens græsning ikke gør.

2 ☐ Grundstammens størrelse er ikke signifikant, da $p$-værdien er større end 0.01.

3 ☐ Både grundstammens størrelse og græsning er signifikante, da $p$-værdierne er mindre end 0.01.

4 ☐ Hverken grundstammens størrelse eller græsning er signifikant, da p-værdierne er mindre end 0.05.

5 ☐ Ingen af de involverede variable er normalfordelt, da alle $p$-værdier er mindre end 0.01.

Fortsæt på side 11

## Opgave V

Denne opgave indeholder to uafhængige spørgsmål.

### Spørgsmål V.1 (13)

I den følgende R-kode simuleres to stikprøver af forskellig størrelse:

```
x1 <- rnorm(10, mean=4, sd=5)
x2 <- rnorm(20, mean=4, sd=5)
t.test(x1, x2)
```

Hvis denne kode køres, hvad er sandsynligheden for, at den opnåede $p$-værdi i resultatet af `t.test`-funktionen er under et givet $\alpha \in [0, 1]$?

1 ☐ $\alpha \cdot \frac{2+10}{10+20}$

2 ☐ $\alpha \cdot \sqrt{\frac{1}{10} + \frac{1}{20}}$

3 ☐ $\alpha \cdot (\frac{1}{10} + \frac{1}{20})$

4 ☐ $1 - \alpha$

5 ☐ $\alpha$

### Spørgsmål V.2 (14)

Der er lavet en plan for et nyt eksperiment til test af forskellen i gennemsnit mellem to populationer. Testens styrke skal beregnes på signifikansniveau 5%. Den mindste forskel i gennemsnit, der skal detekteres, er sat til 1, og stikprøvestørrelsen er 30 i hver stikprøve.

Standardafvigelsen for de to populationer antages at være ens. For at få en værdi, skal det samlede (pooled) estimat fra et tidligere lignende eksperiment anvendes. Den forrige stikprøve fra den første population havde standardafvigelsen $s_1 = 1.8$ og stikprøvestørrelsen $n_1 = 20$, og fra den anden population havde den forrige stikprøve standardafvigelsen $s_2 = 1.4$ og stikprøvestørrelsen $n_2 = 30$.

Hvad er testens styrke med dette planlagte eksperiment?

1 ☐ 0.921

2 ☐ 0.339

3 ☐ 0.998

4 ☐ 0.227

5 ☐ 0.679

Fortsæt på side 12

## Opgave VI

En gruppe forskere ønsker at sammenligne nedbøren i april mellem to områder "A"og "B" i et bestemt land. Fra hvert område er der taget 20 observationer, som alle kan antages at være uafhængige. Data er målt i mm.

Antag at data for region A er gemt i `rainA` og data for region B er gemt i `rainB`.

### Spørgsmål VI.1 (15)

Følgende kode blev kørt:

```
sum(rainA)

## [1] 610.2105

quantile(rainA, probs = c(0.025, 0.975))

##      2.5%     97.5%
##  8.278595 66.521274
```

Hvilket af følgende udsagn kan konkluderes om stikprøvedataene fra område A?:

1 ☐ Stikprøvegennemsnittet er 30.5.

2 ☐ Stikprøvemedianen er 37.4.

3 ☐ Stikprøvevariansen er 14.9.

4 ☐ Stikprøvestandardafvigelsen er 14.9.

5 ☐ Ingen af ovenstående.

### Spørgsmål VI.2 (16)

Forskerne beslutter sig for at sammenligne medianerne af de to stikprøver.

Hvilket af følgende kodeuddrag udregner på korrekt vis et 95% konfidensinterval for medianernes forskel ved brug af ikke-parametrisk bootstrapping?

1 ☐
```
sim_median_diff <- replicate(1000,
                             median(sample(rainA, 20, replace = TRUE)) -
                             median(sample(rainB, 20, replace = TRUE)))
quantile(sim_median_diff, c(0.025, 0.975))
```

2 ☐
```
sim_median_diff <- replicate(1000,
                    median(sample(rainA - rainB, 20, replace = TRUE)))
quantile(sim_median_diff, c(0.025, 0.975))
```

3 ☐
```
t.test(rainA, rainB, paired = FALSE, conf.level = 0.95)$conf.int
```

4 ☐
```
t.test(rainA, rainB, paired = TRUE, conf.level = 0.95)$conf.int
```

5 ☐
```
t.test(rainA, rainB, paired = TRUE, conf.level = 0.975)$conf.int
```

### Spørgsmål VI.3 (17)

Som resultat af forrige spørgsmål fik forskerne konfidensintervallet [-16.3, 4.09].

Hvilket af følgende udsagn kan vi konkludere?

1 ☐ Mediannedbøren i område A er signifikant mindre end mediannedbøren i område B på et 5% signifikansniveau.

2 ☐ Mediannedbøren i område B er signifikant mindre end mediannedbøren i område A på et 5% signifikansniveau.

3 ☐ Der er ikke en signifikant forskel på mediannedbøren i område A og område B på et 5% signifikansniveau.

4 ☐ Der er en lineær sammenhæng mellem nedbøren i område A og område B.

5 ☐ Ingen af ovenstående.

Fortsæt på side 15

## Opgave VII

Udviklere af skovforvaltningsteknikker vil vide, hvordan de kan øge biodiversiten. De har udført eksperimenter, hvor forskellige teknikker til skovforvaltning blev anvendt på tilfældigt udvalgte områder i en skov. På hvert område blev en af fire forskellige skovbrugsteknikker blev anvendt.

Efter fem år blev biodiversiteten beregnet på hvert sted med Shannons biodiversitetsindeks, som måler artsrigdommen.

De observerede værdier for de fire teknikker navngivet A til D er:

| A | B | C | D |
|---|---|---|---|
| 0.9 | 1.8 | 1.5 | 0.9 |
| 1.5 | 1.8 | 1.7 | 1.6 |
| 1.4 | 2.1 | 1.8 | 1.2 |
| 1.1 | 1.7 | 1.0 | 1.6 |
| 2.0 | 2.3 | 1.9 | 1.5 |

En ANOVA-analyse blev udført for at afgøre, om der er en signifikant forskel i gennemsnit i biodiversiten efter anvendelsen af de forskellige teknikker. Det kan antages, at de nødvendige antagelser for brug af ANOVA er opfyldt. ANOVA-resultatet er:

```
anova(lm(y ~ technique))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## technique  3 1.0855 0.36183    X       X      X
## Residuals 16 1.8400 0.11500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bemærk, at nogle af værdierne er blevet erstattet med et `X`.

### Spørgsmål VII.1 (18)

Hvad er den totale varians ($SST$)?

1 ☐ 0.36183

2 ☐ 1.0855

3 ☐ 1.1783

4 ☐ 1.8400

5 ☐ 2.9255

**Spørgsmål VII.2 (19)**

Beregn $F$ teststørrelsen for ANOVA. Hvad er konklusionen af testen for forskel i teknikker på signifikansniveau $\alpha = 5\%$ (både konklusion og argument skal være korrekt)?

1 ☐ Nulhypotesen bliver ikke afvist da $F = 3.146 < 3.239$.

2 ☐ Nulhypotesen bliver afvist da $F = 3.146 < 4.077$.

3 ☐ Nulhypotesen bliver ikke afvist da $F = 3.239 < 4.077$.

4 ☐ Nulhypotesen bliver ikke afvist da $F = 0.0542 < 4.077$.

5 ☐ Nulhypotesen bliver afvist da $F = 0.0542 > 0.05$.

**Spørgsmål VII.3 (20)**

Det var forudplanlagt at beregne 95% konfidensintervallet for forskellen i middelværdi mellem to specificerede teknikker. Hvilke af de følgende R-koder beregner bredden af dette interval?

1 ☐ `2 * qt(0.95, df=16) * sqrt(0.115 * 1/25)`

2 ☐ `qt(0.95, df=16) * sqrt(1.84 * 1/5)`

3 ☐ `2 * qt(0.975, df=16) * sqrt(0.115 * 2/5)`

4 ☐ `2 * qt(0.975, df=20) * sqrt(0.115 * 2/5)`

5 ☐ `qt(0.975, df=20) * sqrt(1.84 * 1/5)`

Fortsæt på side 17

## Opgave VIII

### Spørgsmål VIII.1 (21)

To studerende har som en del af deres bachelorprojekt udført eksperimenter og indsamlet data, og ønsker nu at undersøge sammenhængen mellem to variable, "X" og "Y".

Men de kan ikke komme til enighed om hvordan man korrekt tjekker modelantagelserne bag lineær regression. Kun ét af nedenstående udsagn er korrekt. Hvilket?

1 ☐ Et QQ-plot af "Y"værdierne vil kunne afsløre om normalitetsantagelsen er opfyldt.

2 ☐ Et QQ-plot af "X"værdierne vil kunne afsløre om normalitetsantagelsen er opfyldt.

3 ☐ Et boxplot af "X" og "Y"værdierne vil kunne afsløre om antagelsen om varianshomogenitet er opfyldt.

4 ☐ Et residualplot med fittede værdier på den ene akse og residualer på den anden akse, vil kunne afsløre om linearitetsantagelsen er opfyldt.

5 ☐ Parametrisk boostrapping vil kunne afsløre om linearitetsantagelsen er opfyldt.

Fortsæt på side 18

## Opgave IX

Data fra et randomiseret blokdesignet eksperiment blev indlæst i R:

```
y <- c(3.5, 3.0, 5.4, 7.2,
       7.7, 9.0, 7.0, 6.0,
       0.4, 1.1, 1.0, 1.8)

treatm <- as.factor(c(1, 1, 1, 1,
                      2, 2, 2, 2,
                      3, 3, 3, 3))

block <- as.factor(c(1, 2, 3, 4,
                     1, 2, 3, 4,
                     1, 2, 3, 4))
```

En model for data fra et sådant eksperiment er

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

hvor $i = 1, \ldots, k$ indekserer behandlingen ($k = 3$) og $j = 1, \ldots, l$ indekserer index blokken ($l = 4$).

Ifølge kursets bog estimerer vi $\mu$ med det samlede gennemsnit, $\hat{\mu} = \bar{\bar{y}}$.

### Spørgsmål IX.1 (22)

Givet antagelserne bag modellen er opfyldt, hvad er estimatet af virkningen fra blok 1?

1 ☐ $\hat{\beta}_1 = 4.425$

2 ☐ $\hat{\beta}_1 = 4.775$

3 ☐ $\hat{\beta}_1 = 3.867$

4 ☐ $\hat{\beta}_1 = -0.558$

5 ☐ $\hat{\beta}_1 = 2.988$

### Spørgsmål IX.2 (23)

Hvad er estimatet af $\sigma$?

1 ☐ $\hat{\sigma} = 0.65$

2 ☐ $\hat{\sigma} = 1.57$

3 ☐ $\hat{\sigma} = 14.9$

4 ☐ $\hat{\sigma} = 8.93$

5 ☐ $\hat{\sigma} = 3.85$

Fortsæt på side 20

**Opgave X**

Nedenstående tabel viser antallet af dræbte trafikken i Danmark i nogle kategorier (dvs. ikke alle trafikdødsfald er inkluderet) for årene 2016-2019.

| År | 2016 | 2017 | 2018 | 2019 | Total |
|---|---|---|---|---|---|
| Almindelig personbil | 96 | 99 | 62 | 87 | 344 |
| Motorcykel | 26 | 11 | 21 | 27 | 85 |
| Cykel | 31 | 27 | 28 | 31 | 117 |
| Fodgænger | 36 | 20 | 30 | 30 | 116 |
| Total | 189 | 157 | 141 | 175 | 662 |

I det følgende defineres "blød trafikant" som enten en cyklist eller fodgænger.

**Spørgsmål X.1 (24)**

Hvad er det sædvanlige 95 % konfidensinterval for andelen af "bløde trafikanter" dræbt i trafikken baseret på ovenstående data (dvs. totalen over de 4 år)?

1 ☐ $[0.316, 0.388]$

2 ☐ $[0.496, 0.590]$

3 ☐ $[0.321, 0.382]$

4 ☐ $[0.326, 0.378]$

5 ☐ $[0.504, 0.583]$

**Spørgsmål X.2 (25)**

Som en del af analysen ønsker man at teste, om der er en statistisk signifikant forskel i andelen af bløde trafikbrugere, der blev dræbt i årene 2016 og 2019 ($H_0 : p_{2016} - p_{2019} = 0$). Hvad er konklusionen når vi bruger den sædvanlige test og signifikansniveau $\alpha = 5\%$ (både konklusion og argument skal være korrekt)?

1 ☐ Teststørrelsen bliver $Z = 0.118$. Der er ikke en signifikant forskel, da $Z < 0.95$

2 ☐ Teststørrelsen bliver $Z = 0.237$. Der er en signifikant forskel, da $Z > 0.05$.

3 ☐ Teststørrelsen bliver $Z = 0.237$. Der er ikke en signifikant forskel, da $Z < 1.96$.

4 ☐ Teststørrelsen bliver $Z = 0.237$. Der er ikke en signifikant forskel, da $Z > 0.05$.

5 ☐ Teststørrelsen bliver $Z = 0.118$. Der er ikke en signifikant forskel, da $Z < 1.96$.

**Spørgsmål X.3 (26)**

Hvad er det sædvanlige 95% konfidensinterval for forskellen i andelen af motorcyklister, der blev dræbt i 2016 og 2017 ($p_{2016} - p_{2017}$)?

1 ☐ $[0.070, 0.138]$

2 ☐ $[0.013, 0.122]$

3 ☐ $[0.022, 0.113]$

4 ☐ $[0.004, 0.131]$

5 ☐ $[0.044, 0.091]$

Som hjælp til de næste spørgsmål gives følgende resultat fra R (nogle tal er erstattet af bogstaver), `dat` er et passende udsnit af tabellen ovenfor (dvs. eksklusive totalerne)

```
> chisq.test(dat)

 chisq.test(dat)

Pearson's Chi-squared test

data:  dat
X-squared = 15.356, df = A, p-value = B
```

**Spørgsmål X.4 (27)**

På signifikansniveau $\alpha = 5\%$, hvad er konklusionen om hele fordelingen over de 4 år (både konklusion og argument skal være korrekt)?

1 ☐ Der er en signifikant ændring over årene, da $15.36 > 12.59$, hvor 12.59 er den kritiske værdi for $\chi^2$-testen.

2 ☐ Det kan ikke afvises, at fordelingen er uændret, da $p$-værdien er $0.08 > 0.05$.

3 ☐ Det kan ikke afvises, at fordelingen er uændret, da $p$-værdien er $0.50 > 0.05$

4 ☐ Der er en signifikant ændring gennem årene, da $15.36 > 9$, hvor 9 er den kritiske værdi for $\chi^2$ -testen.

5 ☐ Det kan ikke afvises, at fordelingen er uændret, da $15.35 < 26.27$, hvor 26.27 er den kritiske værdi fra $\chi^2$ -testen.

**Spørgsmål X.5 (28)**

Hvad er bidraget til teststørrelsen fra cellen "almindelig personbil" i 2016?

1 ☐ 0.050

2 ☐ 0.508

3 ☐ 0.279

4 ☐ 0.520

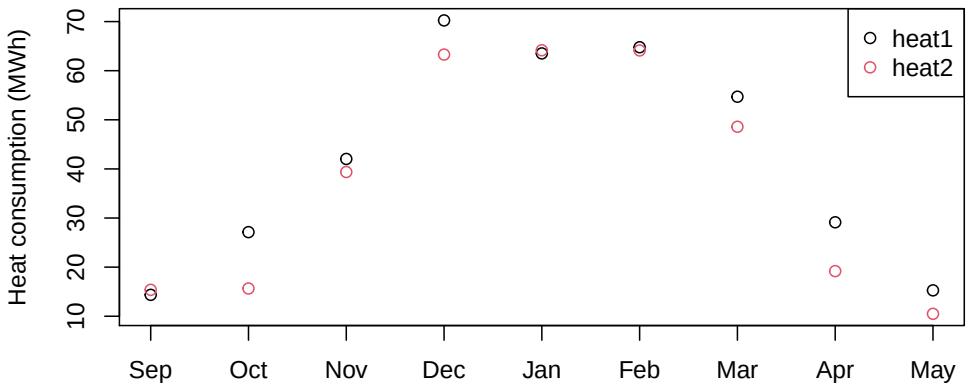
5 ☐ 0.285

Fortsæt på side 23

## Opgave XI

Den nuværende lovgivning kræver at varmeforbruget i bygninger skal reduceres i de kommende år. Et eksperiment blev udført i to identiske lejlighedsbygninger. Beboerne i den ene af de to bygninger fik givet energispareråd igennem en opvarmningssæson (september-maj).

Det ugentlige varmeforbrug, der dækker perioden, er indlæst i to vektorer i R, en for hver bygning: `heat1` og `heat2`. De er plottet herunder.




### Spørgsmål XI.1 (29)

En test for forskel i middelvarmeforbrug mellem bygningerne skal udføres. Hvilken af følgende R-koder beregner det korrekte resultat af sådan en test?

1 ☐ `summary(lm(heat1 ~ heat2))`

2 ☐ `prop.test(heat1 > heat2, length(heat1))`

3 ☐ `t.test(heat1, heat2)`

4 ☐ `t.test(heat1-heat2)`

5 ☐ `1 - pt(mean(heat1)-mean(heat2), df=length(heat1))`

### Spørgsmål XI.2 (30)

Testens $p$-værdi var 1.6%. Ifølge kursets bog, hvordan skal man formidle dette resultat?

1 ☐ Der er lidt eller ingen evidens for, at der er signifikant forskel i varmeforbruget.

2 ☐ Der er lidt eller ingen evidens for, at der ikke er signifikant forskel i varmeforbruget.

3 ☐ Der er nogen evidens for, at der er signifikant forskel i varmeforbruget.

4 ☐ Der er nogen evidens for, at der ikke er signifikant forskel i varmeforbruget.

5 ☐ Der er stærk evidens for, at der ikke er signifikant forskel i varmeforbruget.

Fortsæt på side 25

SÆTTET ER SLUT. God jul!