

Skriftlig prøve: 14. maj 2023

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 14 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.**

<b>Opgave</b>	I.1	I.2	II.1	III.1	IV.1	V.1	V.2	VI.1	VI.2	VI.3
<b>Spørgsmål</b>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
<b>Svar</b>										

<b>Opgave</b>	VII.1	VII.2	VIII.1	VIII.2	VIII.3	IX.1	X.1	X.2	X.3	XI.1
<b>Spørgsmål</b>	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
<b>Svar</b>										

<b>Opgave</b>	XI.2	XII.1	XII.2	XIII.1	XIII.2	XIV.1	XIV.2	XIV.3	XIV.4	XIV.5
<b>Spørgsmål</b>	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
<b>Svar</b>										

Eksamenssættet består af 24 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 virksomhed producerer cylinderformede vandtønder med et volumen på 100 liter og har ud fra fremstillingsprocessen bestemt usikkerhed mht. radius og højde på vandtønderne. De er nu interesserede i at bestemme usikkerheden af tøndernes volumen.

Formlen for cylindernes volumen er givet ved  $V = H \cdot \pi \cdot R^2$ , hvor  $H$  og  $R$  er hhv. højde og radius, som er antaget indbyrdes uafhængige. Følgende information er givet:

$$\mu_H = 0.6 \text{ m}, \sigma_H = 5 \cdot 10^{-3} \text{ m}$$

$$\mu_R = 0.23 \text{ m}, \sigma_R = 3 \cdot 10^{-3} \text{ m}$$

$$\mu_V = 0.1 \text{ m}^3, \sigma_V = ?$$

#### Spørgsmål I.1 (1)

Hvilken af følgende formler kan anvendes til at approksimere standardafvigelsen (usikkerheden) på tøndernes volumen,  $\sigma_V$ ? **Hint:** Brug kommandoen `pi` i dine R-udregninger til at sikre den fornødne præcision af dit resultat.

1   $\sigma_V = 0.166^2 \cdot \sigma_R^2 + 0.867^2 \cdot \sigma_H^2$

2   $\sigma_V = 0.166^2 \cdot \sigma_H^2 + 0.867^2 \cdot \sigma_R^2$

3   $\sigma_V = \sqrt{0.166^2 \cdot \sigma_H^2 + 0.867^2 \cdot \sigma_R^2}$

4   $\sigma_V = \sqrt{0.166^2 \cdot \sigma_R^2 + 0.867^2 \cdot \sigma_H^2}$

5   $\sigma_V = \sqrt{0.166^2 \cdot \sigma_R^2 + 0.434^2 \cdot \sigma_H^2}$

#### Spørgsmål I.2 (2)

Hvilken af følgende R-kommandoer simulerer produktionen af 1000 tønder og udregner volumen? Det antages at data er normalfordelt.

1  `rnorm(1000, 0.6, 0.03) * pi * rnorm(1000, 0.23, 0.05)^2`

2  `pnorm(1000, 0.6, 0.003) * pi * pnorm(1000, 0.23, 0.005)^2`

3  `rnorm(1000, 0.6, 0.005) * pi * rnorm(1000, 0.23, 0.003)^2`

4  `qnorm(1000, 0.6, 0.005) * pi * qnorm(1000, 0.23, 0.003)^2`

5  `qnorm(500, 0.6, 0.003) + pi * qnorm(500, 0.23, 0.005)^2`

Fortsæt på side 4

## Opgave II

Militære radar- og missildetektionssystemer er designet til at advare et land om en fjendes angreb. Håbet er, at detektionssystemet vil være i stand til at identificere et angreb og udstede en advarsel. Antag, at et bestemt detektionssystem har en sandsynlighed på 0.90 for at opdage et missilangreb.

### Spørgsmål II.1 (3)

Hvis to detektionssystemer er installeret i samme område og fungerer uafhængigt, hvad er sandsynligheden for, at begge systemer vil opdage angrebet?

- 1  0.9
- 2  0.81
- 3  0.92
- 4  0.99
- 5  0.999

Fortsæt på side 5

### Opgave III

Eksamensresultatet for studerende antages normalfordelt. I en 100-point afsluttende eksamen er gennemsnitsscoren 55 og standardafvigelsen 19. De bedst scorende 15% af de studerende får karakteren 12 efter den danske 7-trinsskala.

#### Spørgsmål III.1 (4)

Hvad er minimumspoint for at opnå karakteren 12 i eksamen?

- 1  70.99
- 2  74.69
- 3  76.76
- 4  78.80
- 5  82.80

Fortsæt på side 6

## Opgave IV

En undersøgelse viste, at et flertal af de adspurgte planlægger at foretage deres ferieindkøb online, fordi de ikke ønsker at bruge penge på benzin til kørsel fra butik til butik. Antag, at vi har en gruppe på 10 kunder; 7 foretrækker at foretage deres ferieindkøb online og 3 foretrækker at foretage deres ferieindkøb i butikker.

En tilfældig stikprøve på 3 af disse 10 shoppere er udvalgt til en mere dybdegående undersøgelse af, hvordan økonomien har påvirket deres indkøbsadfærd.

### Spørgsmål IV.1 (5)

Hvad er sandsynligheden for, at flertallet (enten 2 eller 3) foretrækker at handle online?

- 1  0.0006
- 2  0.0136
- 3  0.8167
- 4  0.9012
- 5  0.9984

Fortsæt på side 7

## Opgave V

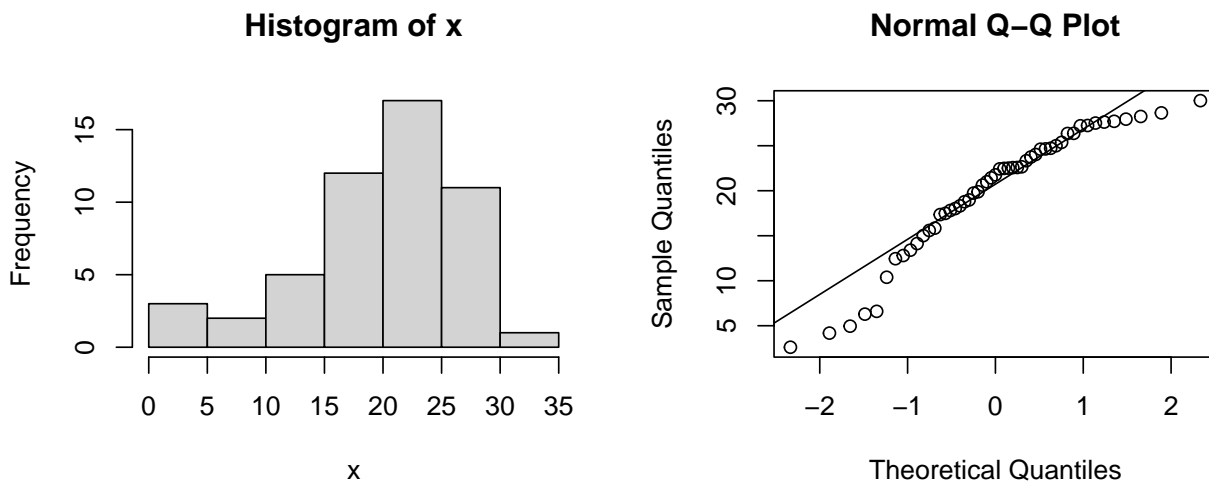
Elproduktion med solceller er afgørende for overgangen til et bæredygtigt energisystem. I en test blev to systemer af samme størrelse placeret ved siden af hinanden, så de blev udsat for nøjagtig samme forhold. Strømproduktionen fra hvert system blev målt over en periode på 50 dage. Lad  $x_i$  og  $y_i$  angive den målte energiproduktion i kWh for dag  $i$  i perioden, for henholdsvis System 1 og 2. Derfor er  $n = n_x = n_y = 50$ .

Lad  $\delta_i = x_i - y_i$  angive den daglige forskel, og  $\bar{\delta} = \frac{1}{50} \sum_{i=1}^{50} \delta_i$  og  $s_{\delta}^2 = \frac{1}{49} \sum_{i=1}^{50} (\delta_i - \bar{\delta})^2$ .

### Spørgsmål V.1 (6)

Ved beregning af et konfidensinterval for middelværdien ved hjælp af  $t$ -fordelingen kan antagelsen om normalfordeling af populationen tages i betragtning.

For de observerede værdier fra System 1 kontrolleres antagelsen med følgende histogram og normal QQ-plot:



Hvad er den passende konklusion med hensyn til gyldigheden af konfidensintervallet baseret på de to plots, samt overvejelse af stikprøvestørrelsen?

- 1  Der er ikke nogen tydelig afvigelse fra normalfordelingen, og normalfordelingsantagelsen er dermed opfyldt og konfidensintervallet er gyldigt. Stikprøvestørrelsen påvirker ikke denne konklusion.
- 2  Der er en klar afvigelse fra normalfordelingen, og normalfordelingsantagelsen er dermed ikke opfyldt og konfidensintervallet er ikke gyldigt. Stikprøvestørrelsen påvirker ikke denne konklusion.
- 3  Der er en klar afvigelse fra normalfordelingen, og normalfordelingsantagelsen er dermed ikke opfyldt, og da stikprøvestørrelse er over 30 er normalfordelingsantagelsen ifølge den centrale grænsesætning betydende og dermed er konfidensintervallet ikke gyldigt.

- 4  Der er en klar afvigelse fra normalfordelingen, og normalfordelingsantagelsen er dermed ikke opfyldt, men da stikprøvestørrelse er over 30 er normalfordelingsantagelsen, ifølge den centrale grænsesætning, ikke betydende og dermed er konfidensintervallet gyldigt.
- 5  Ingen af ovenstående er passende konklusioner.

### Spørgsmål V.2 (7)

Hvilken teststatistik er korrekt at bruge til at teste hypotesen

$$H_0 : \mu_X - \mu_Y = 0$$

dvs. middelforskel i energi genereret af de to systemer?

- 1   $t_{\text{obs}} = \frac{\bar{x} - \bar{y}}{\sqrt{s_x/n_x + s_y/n_y}}$
- 2   $t_{\text{obs}} = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_x + s_y^2/n_y}}$
- 3   $t_{\text{obs}} = \frac{\bar{\delta}}{\sqrt{s_x^2/n_x + s_y^2/n_y}}$
- 4   $t_{\text{obs}} = \frac{\bar{x} - \bar{y}}{\sqrt{s_\delta/n}}$
- 5   $t_{\text{obs}} = \frac{\bar{\delta}}{\sqrt{s_\delta^2/n}}$

Fortsæt på side 9



## Opgave VI

En kontormedarbejder fik råd om at træne mere og mindst gå 10000 skridt om dagen. Ved hjælp af en skridttæller optalte kontormedarbejderen antallet af skridt om dagen. Efter 10 dage indtastede han tallet fra hver dag afrundet til nærmeste 100 i R ved:

```
x <- c(8500, 10300, 6800, 10600, 4900, 6200, 10800, 5700, 5100, 9000)
```

Kontormedarbejderen ville nu bruge dette data til at konkludere, om han trænede tilstrækkeligt ifølge det råd han havde fået.

### Spørgsmål VI.1 (8)

Hvad er 99% konfidensintervallet for middelværdien af antallet af skridt gået pr. dag beregnet med den indsamlede stikprøve?

- 1  [1297, 14283]
- 2  [4274, 11306]
- 3  [5400, 10180]
- 4  [6126, 9454]
- 5  [6442, 9138]

### Spørgsmål VI.2 (9)

Hvad er konklusionen for test af hypotesen

$$H_0 : \mu = 10000$$

med de indsamlede data på et 5% signifikansniveau (både konklusionen og argumentet skal være korrekt)?

- 1  Nulhypotesen afvises, da  $p$ -værdien er under 5%.
- 2  Nulhypotesen afvises, da  $p$ -værdien er over 5%.
- 3  Nulhypotesen afvises ikke, da  $p$ -værdien er under 5%.
- 4  Nulhypotesen afvises ikke, da  $p$ -værdien er over 5%.
- 5  Ingen af ovenstående konklusioner og argumenter er korrekte.

### Spørgsmål VI.3 (10)

I et eksperiment ønsker kontomedarbejderen at sammenligne to perioder, hvor skridt pr. dag observeres. Hver periode har en længde på 14 dage og han ønsker at teste med et signifikansniveau på 1% og have testens styrke på 80%. Han ønsker at være i stand til at opdage en forskel på populationsmiddel på 1000 skridt. Hvad er den standardafvigelse for populationen der fører til ovenstående signifikansniveau, styrke og forskel (afrundet ned til nærmeste hele tal)?

1   $\sigma = 235$

2   $\sigma = 470$

3   $\sigma = 724$

4   $\sigma = 909$

5   $\sigma = 1061$

Fortsæt på side 11

## Opgave VII

Vi har observeret tre variable og udført en multipel lineær regression

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i \text{ hvor } \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.}$$

```
summary(lm(y~x1+x2))

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.880 -2.584  1.109  1.828  6.598
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9653     1.3833  -0.698  0.49673
## x1             1.2933     0.3481   3.715  0.00231 **
## x2            11.8911     0.6950  17.109 8.82e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.76 on 14 degrees of freedom
## Multiple R-squared:  0.956, Adjusted R-squared:  0.9497
## F-statistic: 152.1 on 2 and 14 DF, p-value: 3.196e-10
```

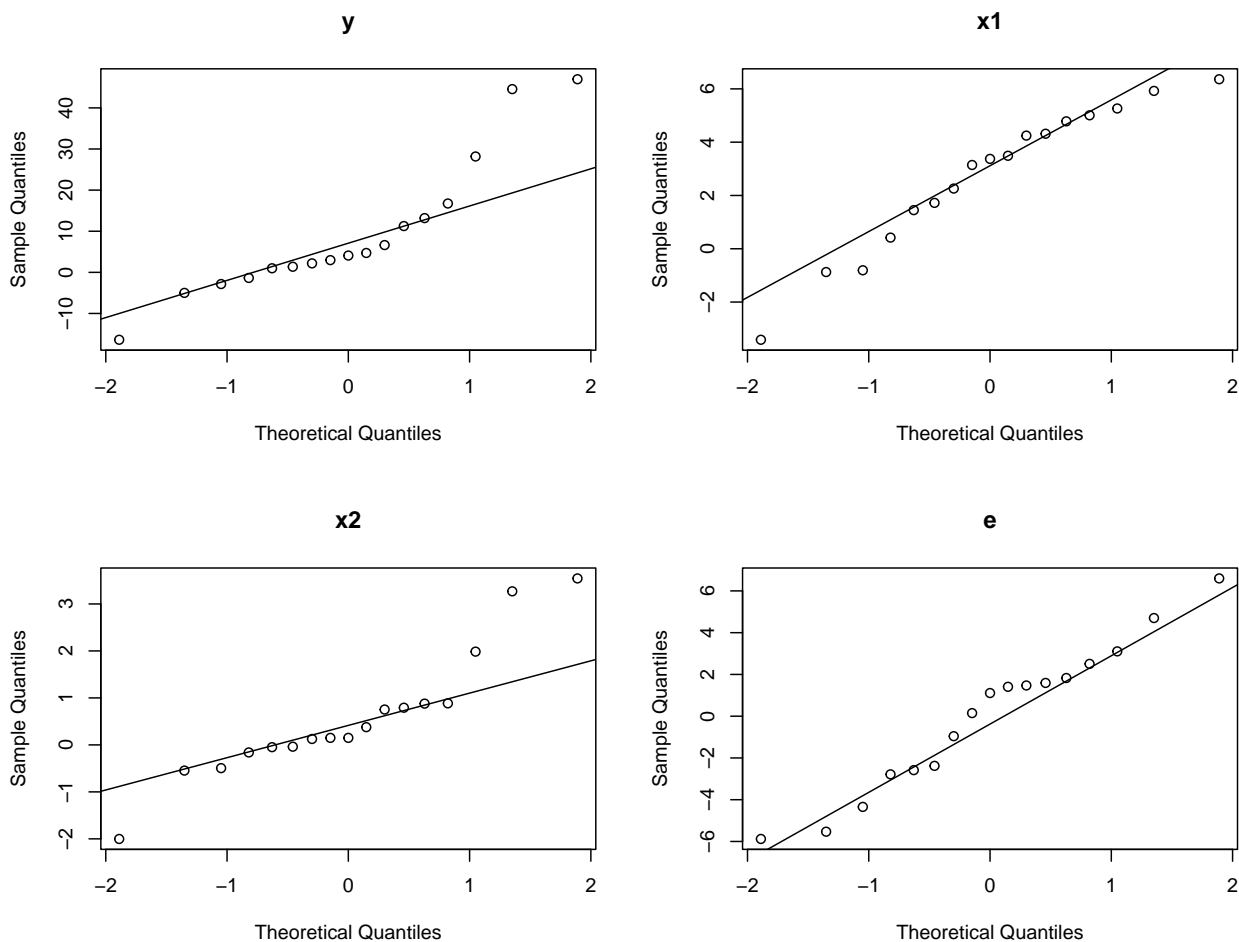
### Spørgsmål VII.1 (11)

Betragt output-linjen "x1" i lm-resultatet. Hvilket af følgende udsagn er korrekt?

- 1  'Std. Error' udtrykker usikkerheden på estimatet af hældningskoefficienten  $\beta_2$ .
- 2  'Std. Error' udtrykker usikkerheden på den forventede værdi af en observation, hvor  $x_1 = 1$  og  $x_2 = 0$ .
- 3   $t$ -værdien er et mål for modelkontrol. En lille  $t$ -værdi indikerer at modellen er valid.
- 4   $t$ -værdien kan bruges til at vurdere om der en signifikant sammenhæng mellem  $x_1$  og  $y$ .
- 5  Hverken 'Std. Error' eller  $t$ -værdien har noget at gøre med usikkerheden i modellen.

### Spørgsmål VII.2 (12)

For at vurdere normalitetsantagelsen i modellen, laves fire QQ-plots med  $y_i$ ,  $x_{1,i}$ ,  $x_{2,i}$  samt residualerne  $e_i$ :



Hvilket af disse plots bør bruges til at undersøge normalitetsantagelsen?

- 1  QQ-plottet med  $y_i$
- 2  QQ-plottet med  $x_{1,i}$
- 3  QQ-plottet med  $x_{2,i}$
- 4  QQ-plottet med  $e_i$
- 5  Ingen af ovenstående.

Fortsæt på side 13

## Opgave VIII

For at undersøge langtidsvirkningerne af miljøpolitikken på et bestemt sted, målte miljøagenturet mængden af ilt opløst i vandet (DO) mg/L, hvilket er en god indikator for vandkvaliteten.

Data indlæstes i R:

```
year <- 1990:2015
DO <- c(1.52, 2.88, 1.60, 2.24, 2.45, 1.84, 2.03, 2.33, 2.81,
        2.46, 2.36, 2.23, 2.81, 2.70, 2.63, 2.00, 2.40, 2.45,
        2.48, 2.51, 2.55, 2.77, 2.70, 2.23, 2.88, 3.09)
```

og vi har udført en lineær regression. Bemærk at noget af outputtet er blevet erstattet med et X:

```
summary(lm(DO ~ year))

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -52.916308  17.494219  -3.025   X       X
## year         0.027634   0.008736   3.163   X       X
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3341 on 24 degrees of freedom
```

### Spørgsmål VIII.1 (13)

Hvad er estimatet for den forventede/gennemsnitlige stigning i DO over fem år?

- 1  -0.028 mg/L
- 2  0.028 mg/L
- 3  0.14 mg/L
- 4  2.8%
- 5  17.5 mg/L

### Spørgsmål VIII.2 (14)

Miljøagenturet ønsker at teste nulhypotesen  $H_0 : \beta_1 = 0$  ved signifikansniveau 1%, hvor  $\beta_1$  er hældningskoefficienten i den lineære regression. Hvilket af følgende udsagn er korrekt (både konklusionen og argumentet skal være korrekt)?

- 1  De kritiske værdier for testen er  $\pm 2.06$ . Nulhypotesen afvises, da  $t_{\text{obs}} > 2.06$ .
- 2  De kritiske værdier for testen er  $\pm 2.80$ . Nulhypotesen afvises, da  $t_{\text{obs}} > 2.80$ .
- 3  De kritiske værdier for testen er  $\pm 2.06$ . Nulhypotesen accepteres, da  $t_{\text{obs}} > 2.06$ .
- 4  De kritiske værdier for testen er  $\pm 2.80$ . Nulhypotesen accepteres, da  $\hat{\beta}_1$  ligger inden for dette interval.
- 5   $p$ -værdien for testen er  $p = 0.004$ . Nulhypotesen accepteres, da  $p < 0.01$ .

### Spørgsmål VIII.3 (15)

Antag at DO værdien i 2022 var 2.21 mg/L. Hvilket af følgende udsagn er korrekt (både konklusionen og argumentet skal være korrekt)?

- 1  95% prædiktionsintervallet er  $[2.17, 3.75]$ . Observationen passer fornuftigt med modellen.
- 2  95% konfidensintervallet er  $[2.38, 3.34]$ . Observationen passer fornuftigt med modellen.
- 3  95% konfidensintervallet er  $[2.38, 3.34]$ . Observationen passer ikke så godt med modellen.
- 4  95% prædiktionsintervallet er  $[2.17, 3.75]$ . Observationen passer ikke så godt med modellen.
- 5  Ingen af ovenstående udsagn er korrekte

Fortsæt på side 15

## Opgave IX

Der opstod en afbrydelse af internettjenesten for de kunder, der bor i en bestemt by. Da kunderne ringede til internetudbyderens kontor, fortalte en optaget besked dem, at virksomheden var klar over tjenesteafbrydelsen, og at det var forventet at tjenesten ville blive genoprettet om to timer. Antag, at to timer er den gennemsnitlige tid til at udføre reparationen, og at reparationstiden har en eksponentiel sandsynlighedsfordeling.

### Spørgsmål IX.1 (16)

Hvad er sandsynligheden for, at reparationen vil tage mellem en time og to timer?

- 1  0.3935
- 2  0.5521
- 3  0.0821
- 4  0.3934
- 5  0.2387

Fortsæt på side 16

## Opgave X

Denne opgave indeholder tre spørgsmål relateret til simulering og bootstrapping.

### Spørgsmål X.1 (17)

Højder på 15 studerende blev målt i cm og aflæst i R.

Udfør parametrisk bootstrapping ved at anvende 1000 simuleringer og med antagelse af normalfordeling af højderne.

Brug følgende R-kode. Kopier ind i R og udfyld det manglende kode ved ? symbolerne:

```
heights <- c(162, 172, 178, 154, 173, 174, 166, 166,  
            166, 164, 167, 163, 165, 170, 177)  
set.seed(1234)  
k <- 1000  
sim_samples <- replicate(k, ?)  
sim_stats <- apply(sim_samples, 2, ?)  
quantile(sim_stats, c(?, ?))
```

Husk at køre `set.seed(1234)`, når du beregner resultatet.

Hvilket af følgende er det beregnede 95% konfidensintervallet for medianen af de studerendes højde?

- 1  [164.29, 171.64]
- 2  [164.01, 171.61]
- 3  [164.69, 171.15]
- 4  [164.73, 171.14]
- 5  [163.66, 170.81]

### Spørgsmål X.2 (18)

Vi ønsker at generere 10 tilfældige tal fra en log-normalfordeling med  $\alpha = 0$  and  $\beta = 1$ . Hvilken af følgende R-kommandoer er korrekt?

- 1  `runif(10)`



2  `pnorm(runif(10))`

3  `plnorm(runif(10))`

4  `qnorm(runif(10))`

5  `qlnorm(runif(10))`

### Spørgsmål X.3 (19)

Hvilken af følgende R-kommandoer må nødvendigvis indgå i en ikke-parametrisk bootstrap baseret på en stikprøve  $x$ ?

1  `sim_means = apply(x, 2, mean)`  
`quantile(sim_means, c(0.025, 0.975))`

2  `sample(rnorm(length(x), mean = mean(x), sd = sd(x)),`  
`size = 1000, replace = FALSE)`

3  `replicate(10000, sample(x, replace = FALSE))`

4  `replicate(10000, sample(x, replace = TRUE))`

5  `sample(rnorm(length(x), mean = mean(x), sd = sd(x)),`  
`size = 1000, replace = TRUE)`

Fortsæt på side 18

## Opgave XI

Flymaling påføres metaloverflader ved en af to metoder: dypning eller sprøjtning. Et team af produktionsingeniører ønsker at undersøge hvorvidt fire forskellige malinger hæfter lige godt på metaloverfladen ved dypning og sprøjtning.

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 hæftestyrken for den  $i$ 'te maling ved anvendelse af den  $j$ 'te metode.  $\alpha_i$  og  $\beta_j$  repræsenterer effekten af hhv. anvendt maling og anvendt metode.

Resultatet af modelfittet er opsummeret i nedenstående ANOVA-tabel:

```
## Analysis of Variance Table
##
## Response: adhesion
##           Df  Sum Sq Mean Sq F value Pr(>F)
## paint      3 0.93308 0.31103  1.0876 0.4733
## method     ? 1.26474      ?      ?      ?
## Residuals  3 0.85790 0.28597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Spørgsmål XI.1 (20)

Hvor mange observationer er inkluderet i undersøgelsen?

- 1  Vi har ikke tilstrækkelig information til at afgøre dette.
- 2  4
- 3  6
- 4  8
- 5  16

### Spørgsmål XI.2 (21)

Et andet team af ingeniører udførte en undersøgelse af lignende slags som den ovenfor og fik følgende resultater med fire andre malinger:

```
## Analysis of Variance Table
##
## Response: adhesion
##           Df Sum Sq Mean Sq F value    Pr(>F)
## paint      3 2.53730 0.84577 43.5483 0.005671 **
## method     1 0.00099 0.00099  0.0511 0.835659
## Residuals  3 0.05826 0.01942
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ingeniørerne har tænkt sig at at udføre parvise post-hoc sammenligninger af alle malingerne og ønsker sig en “family-wise error rate” for type I fejlen på 5%. Hvilken af følgende Bonferroni korrektioner skal anvendes?

- 1   $\alpha_{\text{Bonferroni}} = 0.025$
- 2   $\alpha_{\text{Bonferroni}} = \frac{0.05}{4}$
- 3   $\alpha_{\text{Bonferroni}} = \frac{0.05}{3}$
- 4   $\alpha_{\text{Bonferroni}} = \frac{0.05}{6}$
- 5   $\alpha_{\text{Bonferroni}} = 0.05$

Fortsæt på side 20

## Opgave XII

Følgende data repræsenterer vægten af dagligt fødeindtag (i kg) for voksne rådyr, indsamlet i forskellige perioder af året:

```
consumption <- c(5.4, 4.5, 4.6, 4.4, 4.9, 3.3, 4.1, 4.6, 4.8, 4.6,  
                5.2, 4.7, 4.4, 4.8, 4.8, 5.2, 4.9, 4.8, 5.6, 5.5)  
month <- c(rep("Feb", 5), rep("May", 5), rep("Aug", 5), rep("Nov", 5))
```

### Spørgsmål XII.1 (22)

Hvad er den totale variation (SST) for det daglige fødeindtag?

- 1  0.1843
- 2  0.2535
- 3  2.1215
- 4  2.9480
- 5  5.0695

### Spørgsmål XII.2 (23)

Indlæs data i R og test nulhypotesen om samme daglige fødeindtag i alle måneder. Anvend et signifikansniveau  $\alpha = 0.05$ . Hvilket af følgende udsagn er korrekt?

- 1  Vi analyserer data ved brug af multipel lineær regression. Vi accepterer nulhypotesen om ens fødevareindtag, da  $p = 0.3 > \alpha$ .
- 2  Vi analyserer data ved brug af en-vejs ANOVA. Vi accepterer nulhypotesen om ens fødevareindtag, da  $p = 0.3 > \alpha$ .
- 3  Vi analyserer data ved brug af multipel lineær regression. Vi afviser nulhypotesen om ens fødevareindtag, da  $p = 0.3 > \alpha$ .
- 4  Vi analyserer data ved brug af en-vejs ANOVA. Vi accepterer nulhypotesen om ens fødevareindtag, da  $p = 0.03 < \alpha$ .
- 5  Vi analyserer data ved brug af en-vejs ANOVA. Vi afviser nulhypotesen om ens fødevareindtag, da  $p = 0.03 < \alpha$ .

Fortsæt på side 21

### Opgave XIII

Telefonopkald hos Regional Airways ankommer med en rate på 48 per time ved reservations-  
skranken for Regional Airways og følger en Poisson fordeling.

#### Spørgsmål XIII.1 (24)

Lad den tilfældige variabel  $X$  angive antallet af opkald. Hvad er variansen af  $X$ ?

1  2304

2  48

3  9.6

4  4

5  2

#### Spørgsmål XIII.2 (25)

Antag, at ingen opkald i øjeblikket er i venteposition. Hvis agenten bruger 5 minutter på at  
gennemføre det aktuelle opkald, hvad er sandsynligheden for, at to eller flere opkald venter?

1  0.9817

2  0.9084

3  0.7619

4  0.5665

5  0.7760

Fortsæt på side 22

## Opgave XIV

Danmarks Statistik foretager hvert år undersøgelser i den danske befolkning. En af dem er Levevilkårsundersøgelsen (SILC), som indeholder spørgsmål vedr. økonomisk sårbarhed. Et spørgsmål stillet i spørgeskemaet er: “Hvor nemt er det for din husstand at klare sig økonomisk?”

Følgende svar er observeret for ledige personer de sidste fem år:

	2017	2018	2019	2020	2021	Sum
Meget nemt	9	8	10	11	10	48
Nemt	15	14	20	18	15	82
Hverken nemt eller svært	26	24	26	27	27	130
Svært	23	25	19	20	24	111
Meget svært	28	30	26	24	24	132
Sum	101	101	101	100	100	503

Følgende svar er observeret for beskæftigede personer de sidste fem år:

	2017	2018	2019	2020	2021	Sum
Meget nemt	19	20	23	20	24	106
Nemt	31	29	30	31	32	153
Hverken nemt eller svært	30	31	29	30	28	148
Svært	15	13	12	15	12	67
Meget svært	6	7	6	5	5	29
Sum	101	100	100	101	101	503

### Spørgsmål XIV.1 (26)

Ved kun at tage svarene fra ledige personer i 2021 i betragtning er det observeret, at 24 ud af 100 svarede “Meget svært”. Hvilket af følgende svar indeholder den bedste beregning af 95% konfidensintervallet for denne andel?

1   $0.2708333 \pm 2.58\sqrt{\frac{0.1975}{96}}$

2   $0.24 \pm 1.96\sqrt{\frac{0.1824}{100}}$

3   $0.2708333 \pm 1.96\sqrt{\frac{0.1975}{96}}$

4   $0.24 \pm 2.58\sqrt{\frac{0.1824}{100}}$

5   $0.24 \pm 1.64\sqrt{\frac{0.1824}{100}}$

### Spørgsmål XIV.2 (27)

Ved kun at tage svarene fra beskæftigede personer i 2021 er det observeret, at 5 ud af 101 svarede “Meget svært”. Hvilket af følgende svar indeholder den bedste beregning af 95% konfidensintervallet for denne andel?

1   $0.0666667 \pm 1.96\sqrt{\frac{0.0622}{105}}$

2   $0.049505 \pm 1.96\sqrt{\frac{0.0471}{101}}$

3   $0.0666667 \pm 2.58\sqrt{\frac{0.0622}{105}}$

4   $0.049505 \pm 2.58\sqrt{\frac{0.0471}{101}}$

5   $0.049505 \pm 1.64\sqrt{\frac{0.0471}{101}}$

### Spørgsmål XIV.3 (28)

Det er af interesse at konkludere, om fordelingen af svarene fra grupperne beskæftiget og ledig er forskellige. Brug svarene fra 2021 fra de to grupper for at teste, om der er en signifikant forskel. Hvad er konklusionen på et 5% signifikansniveau (både konklusion og argumentation skal være korrekt)?

1  Der er en signifikant forskel, da  $p$ -værdien er 0.0345.

2  Der er ikke en signifikant forskel, da  $p$ -værdien er 0.9655.

3  Der er ikke en signifikant forskel, da  $p$ -værdien er 0.9997.

4  Der er ikke en signifikant forskel, da  $p$ -værdien er 0.01215.

5  Der er en signifikant forskel, da  $p$ -værdien er 0.00001047.

### Spørgsmål XIV.4 (29)

Det er af interesse at undersøge, om fordelingen af svar har ændret sig over tid. I betragtning af svarene fra ledige personer. Under den sædvanlige nulhypotese, hvad er det forventede antal svar i kategorien “Nemt” i 2019?

1   $e_{23} = 3.2604$

2   $e_{23} = 4.0159$

3   $e_{23} = 13.3677$

4   $e_{23} = 16.465$

5   $e_{23} = 20$

### Spørgsmål XIV.5 (30)

Det er fortsat af interesse at undersøge, om fordelingen af svar har ændret sig over tid. Igen, i betragtning af svarene fra ledige personer, med den sædvanlige test. Hvilken af følgende er den rigtige R-kommando på beregning af  $p$ -værdien og konklusion på et 5% signifikansniveau?

1   $2*(1 - pt(4.5303, df=16)) = 0.0003414$  og dermed en signifikant ændring over tid.

2   $pnorm(4.5303) = 0.9999971$  og dermed ingen signifikant ændring over tid.

3   $1 - pchisq(4.5303, df=16) = 0.9976$  og dermed ingen signifikant ændring over tid.

4   $pf(4.5303, df1=4, df2=4) = 0.9137$  og dermed ingen signifikant ændring over tid.

5   $1 - pf(4.5303, df1=4, df2=4) = 0.08627$  og dermed ingen signifikant ændring over tid.

SÆTTET ER SLUT. God sommer!