Danmarks Tekniske Universitet *Side 1 af 23 sider.*

Skriftlig prøve: 27. Juni 2019

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

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

Opgave	V.4	V.5	VI.1	VI.2	VI.3	VII.1	VII.2	VII.3	VIII.1	VIII.2
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.*

Opgave I

En løber der har en bestemt tur der løbes ofte, beslutter at bruge denne tur til at måle om der sker fremgang i hans fitness. Som mål for fitness beslutter han at bruge gennemsnits hastigheden på denne tur.

I R-outputtet herunder ses resultatet af en simpel lineær regression med tiden målt i uger (`time`), siden starten af et træningsprogram, som forklarende variabel og hastigheden målt i km/t (`speed`).

```
##
## Call:
## lm(formula = speed ~ time, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.686 -0.646 -0.375  0.696  2.705
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.4929     0.4191   25.04   <2e-16 ***
## time          0.0223     0.0138    1.62     0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.11 on 38 degrees of freedom
## Multiple R-squared:  0.0647,Adjusted R-squared:  0.0401
## F-statistic: 2.63 on 1 and 38 DF,  p-value: 0.113
```

Fortsæt på side 3

Spørgsmål I.1 (1)

På signifikansniveau $\alpha = 0.05$ hvad kan der konkluderes om udviklingen i løberens fitness (alle dele af svaret skal være korrekt)?

1 ☐ Der kan ikke påvises en signifikant udvikling over tid, da $0.113 > 0.05$

2 ☐ Der kan påvises en signifikant udvikling over tid da $0.0223 < 0.05$

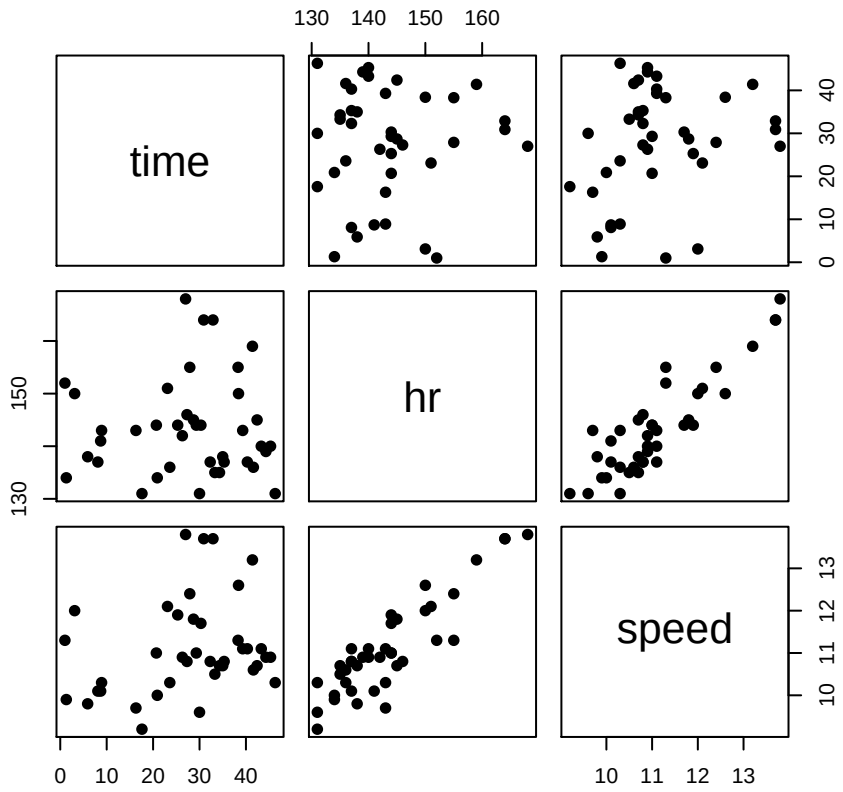
3 ☐ Der kan ikke påvises signifikant en udvikling over tid, da $0.0647 > 0.05$

4 ☐ Der kan påvises en signifikant udvikling over tid da $0.04 < 0.05$

5 ☐ Der kan påvises en signifikant udvikling over tid da $0.014 < 0.05$

Løberen har udover hastigheden på de enkelte ture også målt sin gennemsnitlige puls (slag/min) under løbeturen (`hr`), og ønsker at inkludere denne i modellen, således at hastigheden modelleres som funktion af tid og puls. Som en del af den eksplorative analyse er der vist parvise scatterplots i figur 1.

Figur 1



Fortsæt på side 4

Spørgsmål I.2 (2)

På baggrund af figur 1 kan man konkludere at:

1 ☐ Der er en klar positiv korrelation mellem tid og puls, og der er derfor muligvis et kollinaritets problem

2 ☐ Korrelationen mellem puls og hastighed er høj, og der er derfor klart et kollinaritets problem

3 ☐ Der er en klar negativ korrelation mellem tid og puls, og der er derfor ikke et kollinaritets problem

4 ☐ Der er en klar negativ korrelation mellem tid og hastighed, og der er derfor ikke et kollinaritets problem

5 ☐ Korrelationen mellem puls og hastighed er høj, men der er ikke kollinaritets problemer

Løberen har nu estimeret en model med både puls og tid som forklarende variable, resultatet ses i R-outputtet herunder.

```
summary(lm(speed ~ hr + time, data = dat))

##
## Call:
## lm(formula = speed ~ hr + time, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2849 -0.2066  0.0397  0.2738  0.7800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.00857    1.12243   -4.46  7.3e-05 ***
## hr           0.10831    0.00775   13.97  2.5e-16 ***
## time         0.02104    0.00557    3.78  0.00056 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.449 on 37 degrees of freedom
## Multiple R-squared:  0.851,Adjusted R-squared:  0.843
## F-statistic:  106 on 2 and 37 DF,  p-value: 5.08e-16
```

Fortsæt på side 5

Modellen kan skrives som

$$Y_i = \beta_0 + \texttt{hr}_i\beta_1 + \texttt{time}_i\beta_2 + \epsilon_i \tag{1}$$

hvor ϵ_i antages at opfylde de sædvanlige beingelser for den linære regression model.

Spørgsmål I.3 (3)

Hvad er det sædvanlige 95% konfidensinterval for parametreren β_1?

1 ☐ $[0.106; 0.111]$

2 ☐ $[-0.036; 0.252]$

3 ☐ $[0.093; 0.124]$

4 ☐ $[0.080; 0.137]$

5 ☐ $[0.101; 0.116]$

Modellen kan skrives i matrix-vektor notation som

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

Spørgsmål I.4 (4)

Hvor mange rækker og søjler har $\boldsymbol{X}$?

1 ☐ $\boldsymbol{X}$ har 37 rækker og 3 søjler.

2 ☐ $\boldsymbol{X}$ har 2 rækker og 2 søjler.

3 ☐ $\boldsymbol{X}$ har 40 rækker og 3 søjler.

4 ☐ $\boldsymbol{X}$ har 37 rækker og 2 søjler.

5 ☐ $\boldsymbol{X}$ har 3 rækker og 3 søjler.

Fortsæt på side 6

Spørgsmål I.5 (5)

For at forudsige hastigheden ved et løb ved tiden `time`=55uger, vil løberen gerne give et estimat af hastigheden hvis gennemsnitspulsen holdes på 160 slag/min, hvad er forudsigelsen af hastigheden under disse forhold (når den model der tager højde for både puls og tid tages i anvendelse)?

1 ☐ 10.49 km/t

2 ☐ 4.31 km/t

3 ☐ 13.48 km/t

4 ☐ 13.97 km/t

5 ☐ 11.7 km/t

Løberen ønsker også et konfidensinterval for ovenstående prædiktion. Et konfidensinterval kan skrives på formen

$$CI = \hat{Y}_{new} \pm C \tag{3}$$

Til hjælp for den videre udregning er matricen $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ udregnet til

```
XX.Inv <- matrix(c(6.2,-4.3e-2,-3.7e-3,
                   -4.3e-2,3e-4,-3.5e-6,
                   -3.7e-4,-3.5e-6,1.5e-4),
                 ncol=3)
colnames(XX.Inv) <- rownames(XX.Inv) <- c("(Intercept)", "hr", "time")
XX.Inv
```

```
##             (Intercept)       hr     time
## (Intercept)      6.2000 -4.3e-02 -3.7e-04
## hr              -0.0430  3.0e-04 -3.5e-06
## time            -0.0037 -3.5e-06  1.5e-04
```

Fortsæt på side 7

Spørgsmål I.6 (6)

Hvis der vælges et 95% konfidensinterval hvad er så C?

1 ☐ 1.09

2 ☐ 0.24

3 ☐ 0.69

4 ☐ 1.40

5 ☐ 0.49

Fortsæt på side 8

Opgave II

I en kontorbygning har man målt sammenhængen mellem en såkaldt "termal sensation vote" (TSV) kategoriseret som 0, 1 og 2 og den hastighed (FanSpeed) medarbejderen har indstillet en vifte til (også kategoriseret som 0, 1, 2). Resultatet er vist i tabellen herunder

	FanSpeed=0	FanSpeed=1	FanSpeed=2	Sum
TSV=0	97	20	20	137
TSV=1	40	24	10	74
TSV=2	8	8	10	26
Sum	145	52	40	237

Spørgsmål II.1 (7)

Vi betragter nu kun personer der har angivet TSV til 0. Hvad er det sædvanlige 95% konfidensinterval for sandsynligheden for at personer der angiver en TSV på 0 også vælger en fanspeed på 0?

1 ☐ [0.76;0.90]

2 ☐ [0.63;0.78]

3 ☐ [0.67;0.75]

4 ☐ [0.79;0.86]

5 ☐ [0.70;0.71]

Spørgsmål II.2 (8)

Vi betrager igen kun en gruppe (dvs. een TSV) og kun sandsynligheden for at vælge FanSpeed=0 eller FanSpeed forskellig fra 0. Man planlægger et nyt studie og ønsker i den forbindelse at undersøge hvor mange personer der skal være i studiet hvis man vil have en margin of error på 0.1 og antager at sandsynligheden for at vælge FanSpeed 0 er 0.3 og bruger signifikansniveau $\alpha = 0.05$?

1 ☐ 90

2 ☐ 322

3 ☐ 277

4 ☐ 42

5 ☐ 81

Fortsæt på side 9

Spørgsmål II.3 (9)

Hvad er 95% konfidensintervallet for forskellen mellem sandsynligheden for at sætte FanSpeed på 0 mellem grupperne TSV 0 og TSV 1, og hvilken konklusion fører det til?

1 ☐ Der er signifikant forskel da $0 \notin [0.51, 0.78]$

2 ☐ Der er ikke signifikant forskel da $0 \in [0.51, 0.78]$

3 ☐ Der er signifikant forskel da $0 \notin [0.03, 0.3]$

4 ☐ Der er ikke signifikant forskel da $0 \notin [0.09, 0.24]$

5 ☐ Der er signifikant forskel da $0 \notin [0.09, 0.24]$

Man ønsker nu at teste hele fordelingen, dvs. at vi betragter hele tabellen ovenfor. For at udføre test for hele fordelingen skal man udregne det forventede antal i hver celle, under hypotesen om uafhængighed i indelingskriterierne.

Spørgsmål II.4 (10)

Hvad er det forventede antal ved TSV=0 og FanSpeed=2?

1 ☐ 20

2 ☐ 5.7

3 ☐ 15.9

4 ☐ 23.1

5 ☐ 8

Spørgsmål II.5 (11)

Den sædvanlige teststørrelse for om fordelingen er ens i de 3 grupper er udregnet til 22.72, hvad er konklusion på signifikansniveau $\alpha = 0.05$ (alle dele af svaret skal være korrekt)?

1 ☐ Der er ikke signifikant forskel da $4.6 \cdot 10^{-5} < 0.05$

2 ☐ Der er ikke signifikant forskel da $0.042 < 0.05$

3 ☐ Der er signifikant forskel da $0.00014 < 0.05$

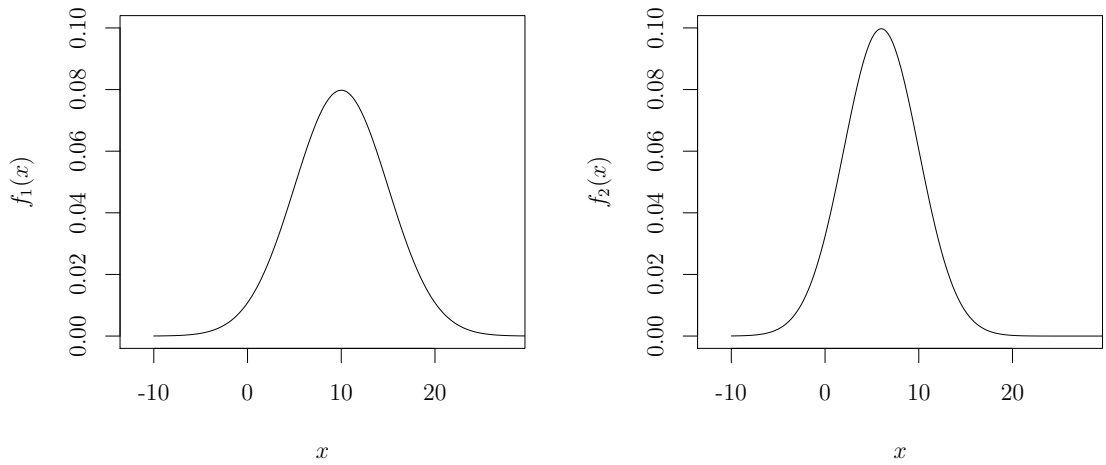
4 ☐ Der er signifikant forskel da $0.042 < 0.05$

5 ☐ Der er signifikant forskel da $4.6 \cdot 10^{-5} < 0.05$

Fortsæt på side 10

Opgave III

Lad to uafhængige stokastiske variable være $X_1 \sim N(10, 5^2)$ og $X_2 \sim N(6, 4^2)$. I nedenstående figur er de to variables tæthedsfunktioner plottet: X_1 til venstre og X_2 til højre:



Spørgsmål III.1 (12)

Hvilket af følgende udsagn er ikke sandt:

1 ☐ $\mathrm{E}(X_1) > \mathrm{E}(X_2)$

2 ☐ $\mathrm{V}(X_1) > \mathrm{V}(X_2)$

3 ☐ $P(X_1 < 0) \neq P(X_2 < 0)$

4 ☐ $P(X_1 > 10) < P(X_2 > 10)$

5 ☐ $P(X_1 > 10) = P(X_2 < 6)$

Spørgsmål III.2 (13)

Hvad er sandsynligheden for at $X_1 > 3 \cdot X_2$?

1 ☐ 0.27

2 ☐ 0.48

3 ☐ 0.19

4 ☐ 0.52

5 ☐ 0.055

Fortsæt på side 11

Spørgsmål III.3 (14)

Hvad er sandsynligheden for at $X_1 - 10 > \sqrt{(X_2 - 6)^2} = |X_2 - 6|$?

1 ☐ 0.32

2 ☐ 0.29

3 ☐ 0.45

4 ☐ 0.5

5 ☐ 0.96

Fortsæt på side 12

Opgave IV

I en hyggespilleklub for ingeniører bliver der spillet forskellige kort- og terningespil. For at have en chance for at vinde er det naturligvis altid vigtigt at have styr på sandsynlighedsregning, men specielt i denne spilleklub for ingeniører hvor alle medlemmer har indgående kendskab til statistik og sandsynlighedsregning.

Spørgsmål IV.1 (15)

I det populære kortspil casino får hver spiller uddelt 4 kort i hver runde. I dette og det næste spørgsmål kan det antages, at kortspillet er blandet således at rækkefølgen af kortene er fuldstændig tilfældig i bunken der deles fra. Spillet er uden jokere, dvs. der er 52 kort og 13 af hver kulør.

Inden spillets start, ønsker et klubmedlem at lave en beregning på følgende: Hvad er sandsynligheden for, at spilleren får uddelt en hånd i første runde, hvor alle 4 kort er spar?

1 ☐ 0.11%

2 ☐ 0.26%

3 ☐ 1.24%

4 ☐ 5.23%

5 ☐ 9.79%

Spørgsmål IV.2 (16)

I løbet af et casinospil med 2 spillere, som beskrevet i forrige spørgsmål, har en spiller husket, at der har været uddelt 10 spar. Der er 16 kort tilbage i bunken. Hvilket af følgende R kald beregner sandsynligheden for, at spilleren får uddelt en hånd hvor ingen af de 4 kort er spar?

1 ☐ `13/16 * 12/16 * 11/16 * 10/16`

2 ☐ `dbinom(13, size = 4, prob = 13/16)`

3 ☐ `dbinom(0, size = 4, prob = 3/16)`

4 ☐ `dhyper(0, m = 4, n = 16, k = 3)`

5 ☐ `dhyper(4, m = 13, n = 3, k = 4)`

Fortsæt på side 13

Spørgsmål IV.3 (17)

Der spilles også yatzy med 6 terninger i spilleklubben. Hver terning kan antages at være fair, således at der er præcis 1/6 sandsynlighed for hvert udfald. I hver runde skal en spiller slå op til 3 gange, men uanset hvad, så er der et "første slag" i hver runde for spilleren således, at der slås med alle 6 terninger samtidigt. Der er 18 runder i et yatzyspil.

Dvs. hver spiller skal 18 gange slå et "første slag" med alle 6 terninger samtidigt.

En yatsy er når alle 6 terninger viser samme tal.

Hvis kun en enkelt spiller betragtes, hvad er så sandsynligheden for at få mindst en yatsy på et "første slag"i løbet af et spil?

1 ☐ 1.4%

2 ☐ 0.69%

3 ☐ 0.23%

4 ☐ 0.15%

5 ☐ 0.0038%

Fortsæt på side 14

Opgave V

En fabrik producerer chokoladepåskeæg med tre forskellige slags fyld: chokolade, marcipan eller knas. I det følgende betragter vi et datasæt, der indeholder vægten af i alt 9 forskellige æg, tre med hver slags fyld. Data indlæses i R i to vektorer:

```
vaegt <- c(3.02, 2.98, 2.95, 3.13, 3.06, 3.12, 2.88, 2.92, 2.86)
fyld <- c("chokolade", "chokolade", "chokolade", "marcipan", "marcipan",
          "marcipan", "knas", "knas", "knas")
```

Lad Y_{ij} betegne vægten af det j'te æg med fyld i (i = chokolade, marcipan eller knas, $j = 1$, 2, eller 3). En statistisk model af følgende form benyttes til at beskrive data: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, hvor ε_{ij}'erne er uafhængige og $N(0, \sigma^2)$-fordelte, og μ, α_i'erne og σ^2 er ukendte parametre.

Spørgsmål V.1 (18)

Hvilken af følgende fortolkninger af den statistiske model er korrekt?

1 ☐ Modellen antager, at alle æg har samme middelvægt.

2 ☐ Modellen beskriver den lineære sammenhæng mellem typen af fyld og middelvægten af et æg.

3 ☐ Modellen siger, at vægtforskelle mellem æg udelukkende skyldes tilfældig støj.

4 ☐ Modellen tillader, at middelvægten af et æg afhænger af dets fyld.

5 ☐ Ingen af de ovenstående fortolkninger er rigtige.

Spørgsmål V.2 (19)

Angiv det sædvanlige estimat for μ.

1 ☐ $\hat{\mu} = 2.89$

2 ☐ $\hat{\mu} = 2.95$

3 ☐ $\hat{\mu} = 2.99$

4 ☐ $\hat{\mu} = 3.04$

5 ☐ $\hat{\mu} = 3.10$

Fortsæt på side 15

Spørgsmål V.3 (20)

Følgende kode er blevet kørt i R:

```
anova(lm(vaegt ~ fyld))

## Analysis of Variance Table
##
## Response: vaegt
##           Df Sum Sq Mean Sq F value  Pr(>F)
## fyld       2 0.0707  0.0353    29.4 0.00079 ***
## Residuals  6 0.0072  0.0012
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hvilken af følgende svarmuligheder er korrekt? (Både argument og konklusion skal være rigtige).

1 ☐ Da 29.454 er mindre end 0.99 fraktilen i en F-fordeling med 6 og 2 frihedsgrader, er der ved 1% signifikansniveau signifikant forskel på vægten af æg med forskellige fyld.

2 ☐ Da 29.454 er større end 0.95, er der ved 5% signifikansniveau signifikant forskel på vægten af æg med forskellige fyld.

3 ☐ Da 29.454 er mindre end 0.99 fraktilen i en F-fordeling med 6 og 2 frihedsgrader, er der ved 1% signifikansniveau ikke signifikant forskel på vægten af æg med forskellige fyld.

4 ☐ Da 0.0007899 er mindre end 0.05, er der ved 5% signifikansniveau ikke signifikant forskel på vægten af æg med forskellige fyld.

5 ☐ Ingen af ovenstående svarmuligheder er korrekte.

Spørgsmål V.4 (21)

Angiv et post-hoc parvist 95% konfidensinterval for den forventede vægtforskel mellem æg med chokolade- og marcipanfyld (dvs. for $\mu_{marcipan} - \mu_{chokolade}$).

1 ☐ [0.05; 0.19]

2 ☐ [0.03; 0.21]

3 ☐ [-0.05; 0.29]

4 ☐ [0.11; 0.13]

5 ☐ Ingen af ovenstående invervaller er rigtige.

Fortsæt på side 16

Det oplyses nu, at æggene også har forskellige farver: rød, blå eller grøn. For hver type fyld er der præcis ét æg af hver farve. Information om farven på de ni æg i datasættet indlæses i R i en tredje vektor kaldet `farve`.

I det næste spørgsmål benyttes en statistisk model af følgende form til at beskrive data: $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, hvor ε_{ij}'erne er uafhængige og $N(0, \sigma^2)$-fordelte, og μ, α_i'erne, β_i'erne og σ^2 er de ukendte parametre. Følgende kode er blevet kørt i R og dele af outputtet erstattet med bogstaver:

```
anova(lm(vaegt ~ fyld + farve))

Analysis of Variance Table

Response: vaegt
          Df   Sum Sq  Mean Sq F value   Pr(>F)
fyld       2 0.070689 0.035344 25.9673 0.005114 **
farve      2 0.001756 0.000878       A        B
Residuals  4 0.005444 0.001361
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Spørgsmål V.5 (22)

Bestem de korrekte værdier af A og B i R-outputtet ovenfor.

1 ☐ $A = 0.645$, $B = 0.428$

2 ☐ $A = 1.550$, $B = 0.428$

3 ☐ $A = 0.645$, $B = 0.683$

4 ☐ $A = 1.550$, $B = 0.317$

5 ☐ $A = 0.645$, $B = 0.572$

Fortsæt på side 17

Opgave VI

Koncentrationen af glukose før (c_0) og efter (c_1) en enzymatisk reaktion er blevet målt 10 gange ved hjælp af en reducerende sukkeranalyse. For hver af de 10 eksperimenter er forskellen (c_{diff}) blevet bestemt, således at $c_{diff} = c_1 - c_0$. Koncentrationsforskellen kan formodes at være normalt fordelt. Koncentrationerne og forskellen er vist i tabellen herunder sammen med deres gennemsnit og standardafvigelser.

	1	2	3	4	5	6	7	8	9	10	$\bar{c}_i$	s_i
c_0	22.2	18.5	15.1	22.7	16.2	22.2	17.7	21.5	15.5	19.0	19.06	2.94
c_1	13.3	18.9	10.6	14.7	14.9	15.1	16.6	8.7	13.0	7.2	13.30	3.59
c_{diff}	-8.9	0.4	-4.5	-8.0	-1.3	-7.1	-1.1	-12.8	-2.5	-11.8	-5.76	4.65

Data kan indlæses i R med koden vist herunder

```
c0 <- c(22.2, 18.5, 15.1, 22.7, 16.2, 22.2, 17.7, 21.5, 15.5, 19.0)
c1 <- c(13.3, 18.9, 10.6, 14.7, 14.9, 15.1, 16.6, 8.7, 13.0, 7.2)
dif <- c(-8.9, 0.4, -4.5, -8.0, -1.3, -7.1, -1.1, -12.8, -2.5, -11.8)
```

Spørgsmål VI.1 (23)

Baseret på ovenstående data, hvad er 95% konfidensintervallet for den gennemsnitlige koncentrationsforskel?

1 ☐ $-5.76 \pm 2.23 \cdot \frac{4.65}{\sqrt{10}}$

2 ☐ $-5.76 \pm 2.26 \cdot \frac{4.65}{\sqrt{10}}$

3 ☐ $-5.76 \pm 2.26 \cdot \frac{4.65}{\sqrt{9}}$

4 ☐ $-5.76 \pm 1.81 \cdot \frac{21.61}{\sqrt{10}}$

5 ☐ $-5.76 \pm 2.26 \cdot \frac{21.61}{\sqrt{20}}$

Fortsæt på side 18

Spørgsmål VI.2 (24)

En one-sample t-test med en tosidet alternativ hypotese er blevet udført for at finde ud af om den gennemsnitlige koncentrationsforskel er signifikant forskellig fra 0, når der benyttes signifikansniveau $\alpha = 0.05$. Når den sædvanlige teststørrelse er udregnet til 3.918, hvilket af følgende udsagn er da korrekt (alle dele af svaret skal være korrekt)?

1 ☐ Der er signifikant forskel da $0.0035 < 0.05$.

2 ☐ Der er signifikant forskel da $0.0044 < 0.05$.

3 ☐ Der er ikke signifikant forskel da $0.0044 < 0.05$.

4 ☐ Der er signifikant forskel da $0.0028 < 0.05$.

5 ☐ Der er ikke signifikant forskel da $0.0028 < 0.05$.

Man beslutter nu at fordelingsantagelsen er tvivlsom og beslutter derfor at konstruere et konfidensinterval for forskellen baseret på ikke parametrisk bootstrap, til dette formål har man kørt følgende R-kode

```
set.seed(1342)
k <- 10000

simsamples <- replicate(k, sample(dif, replace = TRUE))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))

##    2.5%   97.5%
## -8.5203 -3.0898

quantile(simmeans, c(0.05, 0.95))

##    5%   95%
## -8.12 -3.49

simsamples <- replicate(k, sample(dif, replace=TRUE))
simmeans <- apply(simsamples, 2, median)
quantile(simmeans, c(0.025, 0.975))

##  2.5% 97.5%
## -9.45 -1.30

simsamples0 <- replicate(k, sample(c0, replace=TRUE))
simsamples1 <- replicate(k, sample(c1, replace=TRUE))
simmeans <- apply(simsamples1, 2, mean) - apply(simsamples0, 2, mean)
quantile(simmeans, c(0.025, 0.975))
```

```
##    2.5%   97.5%
## -8.4800 -3.0198

quantile(simmeans, c(0.05, 0.95))

##    5%   95%
## -8.05 -3.49
```

Spørgsmål VI.3 (25)

Hvad er 95% konfidensintervallet baseret på ikke parametrisk bootstrap?

1 ☐ [-8.52; -3.09]

2 ☐ [-8.12; -3.49]

3 ☐ [-9.45; -1.3]

4 ☐ [-8.48; -3.02]

5 ☐ [-8.05; -3.49]

Fortsæt på side 20

Opgave VII

Under en sportskonkurrence på en folkeskole udfordrer en ambitiøs gruppe drenge fra 5. klasse en gruppe piger fra 7. klasse. De 23 drenge (n_1) hævder, at de i gennemsnit er hurtigere i sprint over en afstand på 100m i forhold til de 26 piger (n_2). Den gennemsnitlige tid for drengene og pigerne viser sig at være henholdsvis 15.06s og 15.40s. Antag, at dataene i de to grupper er normalfordelt.

Spørgsmål VII.1 (26)

Baseret på R-output'et nedenfor, hvor variablene `boys` og `girls` indeholder alle løbetider for hhv. drenge og piger, hvilket af følgende udsagn er da korrekt?

```
t.test(x = boys, y = girls)

##
##  Welch Two Sample t-test
##
## data:  boys and girls
## t = -1.25, df = 42, p-value = 0.22
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.89847  0.21151
## sample estimates:
## mean of x mean of y
##    15.057    15.400
```

1 ☐ 95% konfidensintervallet indeholder nul, derfor er drengene signifikant hurtigere end pigerne.

2 ☐ P-værdien er større end signifikansniveauet $\alpha = 0.05$, derfor er drengene signifikant hurtigere end pigerne.

3 ☐ 95% konfidensintervallet indeholder nul, derfor er drengene ikke signifikant hurtigere end pigerne.

4 ☐ P-værdien er større end signifikansniveauet $\alpha = 0.05$, derfor er drengene signifikant langsommere end pigerne.

5 ☐ Two-sample t-testen kan ikke anvendes i dette tilfælde, fordi $n_1 \neq n_2$.

Fortsæt på side 21

Spørgsmål VII.2 (27)

Angiv nul og tosidet alternativ hypotese for two-sample t-testen ovenfor (i svarmulighederne er μ_1 er middelværdien for drenge og μ_2 er middelværdien for piger).

1 ☐ $H_0 : \mu_1 = 0$ og $H_1 : \mu_1 \neq 0$.

2 ☐ $H_0 : \mu_1 = \mu_2$ og $H_1 : \mu_1 < \mu_2$.

3 ☐ $H_0 : \hat{\mu}_1 = \hat{\mu}_2$ og $H_1 : \hat{\mu}_1 < \hat{\mu}_2$.

4 ☐ $H_0 : \hat{\mu}_1 = \hat{\mu}_2$ og $H_1 : \hat{\mu}_1 \neq \hat{\mu}_2$.

5 ☐ $H_0 : \mu_1 = \mu_2$ og $H_1 : \mu_1 \neq \mu_2$.

Nedenfor er standard afvigelsen for hhv. drenge og piger udregnet

```
sd(boys)

## [1] 1.0548

sd(girls)

## [1] 0.8419
```

Spørgsmål VII.3 (28)

Hvis vi antager at den sande standard afvigelse for drenge og piger er den samme (dvs. $X_{1,i} \sim N(\mu_1, \sigma^2)$ og $X_{2,i} \sim N(\mu_2, \sigma^2)$), vi definerer desuden S_1^2 og S_2^2 på den sædvanlige måde (med $n_1 = 23$, og $n_2 = 26$), hvad er så $P\left(\frac{S_1}{S_2} > \frac{1.05}{0.84}\right)$?

1 ☐ 0.704

2 ☐ 0.141

3 ☐ 0.196

4 ☐ 0.293

5 ☐ 0.853

Fortsæt på side 22

Opgave VIII

Ved forudsigelse af el produceret ved vindkraft er den såkaldte "power curve" en vigtigt komponent. Power curven beskriver forholdet mellem vindhastighed og produceret effekt. En simpel model for power curven er

$$p(W) = \frac{1}{1 + e^{-\gamma(W-\alpha)}} \tag{4}$$

hvor p er den relative effekt (dvs. $p \in (0, 1)$) for en givet vindmølle eller vindmøllepark og W er vindhastigheden. Vindhastigheden er behæftet med en vis usikkerhed enten på grund af måleusikkerhed eller på grund af usikkerheder i forudsigelser af vindhastigheden.

Spørgsmål VIII.1 (29)

Hvis vi antager at vindhastigheden for en givet forudsigelse har middelværdi μ_w og varians σ_w^2, hvad er så approximationen af standardafvigelsen for produktionen (p) baseret på "non-linear" error propagation?

1 ☐ $\gamma^2 \sqrt{\frac{e^{-2\gamma(\mu_w-\alpha)}}{\left(1+e^{-\gamma(\mu_w-\alpha)}\right)^4}\sigma_w^2}$

2 ☐ $\gamma e^{-\gamma(\mu_w-\alpha)} p(\mu_w)^2 \sigma_w$

3 ☐ $-\gamma \sqrt{\frac{e^{-2\gamma(\mu_w-\alpha)}}{\left(1+e^{-\gamma(\mu_w-\alpha)}\right)^4}\sigma_w^2}$

4 ☐ $\gamma^2 e^{-\gamma(\mu_w-\alpha)} p(\mu_w) \sigma_w$

5 ☐ $\gamma \sqrt{\frac{e^{-\gamma(\mu_w-\alpha)}}{\left(1+e^{-\gamma(\mu_w-\alpha)}\right)^2}\sigma_w^2}$

Hvis vi antager en konkret fordeling for vindhastigheden så er error propagation ved simulation et alternativ til resultatet ovenfor. En almindelig antagelse er at vindhastigheden følger en log-normal fordeling. Man antager nu at $W \sim LN(\log(7), 0.1)$, samt at $\gamma = 1$ og $\alpha = 5$, og har kørt følgende R-kode

```
set.seed(321)
k <- 10000
w <- rnorm(k, log(7), 0.1)
p <- 1/(1 + exp(-(w - 5)))
sqrt(1/(k - 1) * var(p))

## [1] 4.3063e-05

sd(p)
```

```
## [1] 0.0043061

mean(p)^2

## [1] 0.0020478

w <- rlnorm(k, log(7), 0.1)
p <- 1/(1 + exp(-(w - 5)))
sqrt(1/(k - 1) * var(p))

## [1] 0.00078496

sd(p)

## [1] 0.078492

mean(p)^2

## [1] 0.74701
```

Spørgsmål VIII.2 (30)

Baseret på ovenstående hvad er så estimatet for standard afvigelsen af p?

1 ☐ $4.3 \cdot 10^{-5}$

2 ☐ 0.0020

3 ☐ 0.00078

4 ☐ 0.078

5 ☐ 0.75

SÆTTET ER SLUT. God sommer!