Danmarks Tekniske Universitet *Side 1 af 26 sider.*

Skriftlig prøve: 25. juni 2020

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

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

Opgave	VI.4	VI.5	VII.1	VII.2	VIII.1	VIII.2	VIII.3	IX.1	IX.2	IX.3
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Eksamenssættet består af 26 sider.

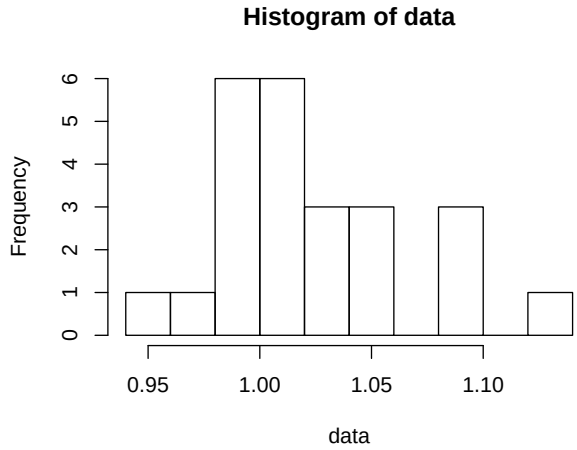
Fortsæt på side 2

Multiple choice opgaver: *Der gøres opmærksom på, at der i hvert spørgsmål er én og kun én svarmulighed, som er rigtig. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde. Husk altid at afrunde dit eget resultat til antallet af decimaler givet i svarmulighederne før du vælger et svar. Husk også, at der kan forekomme små afvigelser mellem resultatet af bogens formler og tilsvarende indbyggede funktioner i R.*

Opgave I

DNA-prøver er blevet udtaget fra 24 forskellige vævstyper. Fordelingen af nukleotidbaser er blevet målt, og for hver prøve har man udregnet forholdet mellem guanin og cytosin, i det følgende kaldet GC-forholdet.

Det antages at GC-forholdet følger en normalfordeling, et histogram af data ses nedenfor.



Spørgsmål I.1 (1)

Hvilken af følgende værdier kan være den estimerede middelværdi $\hat{\mu}$ for data?

1 ☐ 0.031

2 ☐ 0.901

3 ☐ 0.968

4 ☐ 1.027

5 ☐ 1.139

Fortsæt på side 3

Spørgsmål I.2 (2)

Hvilken af følgende værdier kan være den estimerede spredning $\hat{\sigma}$ i data?

1 ☐ 0.042

2 ☐ 0.180

3 ☐ 0.956

4 ☐ 1.017

5 ☐ 1.139

Spørgsmål I.3 (3)

Man ønsker at undersøge om GC-forholdet er signifikant forskelligt fra 1 (nul-hypotesen).

Følgende formel er blevet anvendt,

$$[\hat{\mu} - \hat{\sigma}/\sqrt{24} \cdot t_{0.975}, \hat{\mu} + \hat{\sigma}/\sqrt{24} \cdot t_{0.975}] = [1.0087, 1.0444]$$

hvor $t_{0.975}$ er 97.5%-fraktilen i en t-fordeling med 23 frihedsgrader.

Hvilket af følgende udsagn er korrekt?

1 ☐ Vi afviser nulhypotesen på et 5% signifikansniveau, da 1 ikke er indeholdt i 95% konfidensintervallet.

2 ☐ Vi afviser nulhypotesen på et 2.5% signifikansniveau, da 1 ikke er indeholdt i 97.5% konfidensintervallet.

3 ☐ Vi afviser nulhypotesen på et 2.5% signifikansniveau, da 0 ikke er indeholdt i 97.5% konfidensintervallet.

4 ☐ Vi accepterer nulhypotesen, da $|1 - \hat{\mu}|$ er mindre end standardafvigelsen.

5 ☐ Vi har ikke tilstrækkelig information til at acceptere eller afvise nulhypotesen.

Fortsæt på side 4

Opgave II

I denne opgave undersøges en model til prædiktion af elprisen på engros elmarkedet, det såkaldte Day-ahead marked. Som forklarende variable benyttes prognoser af sol- og vindkraften.

Data til estimation består af dagsgennemsnit for 28 dage i februar 2020 fra elmarkedets område "DK1", som hovedsageligt dækker Jylland.

Følgende 3 variable benyttes i modellen:

- Elprisen (EUR/MWh)
- Solenergiprognosen (MW)
- Vindenergiprognosen (MW)

Data er loaded i 3 vektorer: `price`, `solar` og `wind`, og følgende kode er kørt:

```
  mean(price)

## [1] 17.4

  mean(solar)

## [1] 30.4

  mean(wind)

## [1] 2687

  range(price)

## [1] -1.39 38.33

  range(solar)

## [1]  6.08 64.00

  range(wind)

## [1]  508 4135
```

Fortsæt på side 5

Derefter er koefficienterne i en lineær regressionsmodel estimeret med:

```
  fit <- lm(price ~ solar + wind)
  summary(fit)

##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.730 -2.785  0.861  2.368  7.453
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.566077   2.874305   13.42  6.3e-13 ***
## solar        0.063822   0.050260    1.27     A     B
## wind        -0.008606   0.000722  -11.91     C     D
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.01 on 25 degrees of freedom
## Multiple R-squared:  0.867,Adjusted R-squared:  0.857
## F-statistic: 81.7 on 2 and 25 DF,  p-value: 1.09e-11
```

Bemærk, at 4 felter er udskiftet med bogstaver.

Spørgsmål II.1 (4)

Hvad bliver resultatet af et skridt i backward selection proceduren med signifikansniveau $\alpha = 0.05$ på denne model?

1 ☐ `solar` variablen skal tages ud af modellen

2 ☐ `wind` variablen skal tages ud af modellen

3 ☐ Ingen forklarende variabler skal tages ud af modellen

4 ☐ Både `solar` og `wind` variablerne skal tages ud af modellen

5 ☐ Der er ikke tilstrækkelige informationer til at afgøre dette

Fortsæt på side 6

Spørgsmål II.2 (5)

Med den estimerede model, hvad bliver da prædiktionen af middelforskellen i elpris fra dagen i perioden med den laveste vindkraft til dagen med den højeste vindkraft, hvis solkraften antages at være konstant?

1 ☐ 2.98 EUR/MWh

2 ☐ 11.8 EUR/MWh

3 ☐ 31.2 EUR/MWh

4 ☐ 34.2 EUR/MWh

5 ☐ 38.6 EUR/MWh

Spørgsmål II.3 (6)

Hvilket af følgende kan ikke konkluderes for den estimerede model?

1 ☐ Prædiktionsintervallets bredde er mindst ved 30.4 MW solkraft og 2687 MW vindkraft

2 ☐ Den maksimale værdi af residualerne er 7.453

3 ☐ Modellen kan anvendes til at konkludere om der er signifikante ikke-lineære sammenhænge mellem variablerne

4 ☐ Estimatet af variansen for fejlen ϵ_i er $\hat{\sigma}^2 = 16.1$

5 ☐ Modellen har forklaret 86.7% af variansen

Fortsæt på side 7

Spørgsmål II.4 (7)

Det undersøges i forbindelse med modelvalideringen om en ny variabel `load`, som angiver det daglige elforbrug i de pågældende 28 dage, bør inkluderes i modellen. Igen benyttes signifikansniveau $\alpha = 5\%$ og der estimeres en simpel lineær regressions model med denne og residualerne fra den estimerede model:

```
summary(lm(fit$residuals ~ load))

##
## Call:
## lm(formula = fit$residuals ~ load)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.90  -1.35  -0.09   1.75   6.07
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.57812    6.61879   -5.83  3.8e-06 ***
## load          0.01465    0.00251    5.84  3.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.59 on 26 degrees of freedom
## Multiple R-squared:  0.568,Adjusted R-squared:  0.551
## F-statistic: 34.2 on 1 and 26 DF,  p-value: 3.68e-06
```

Hvilken af følgende konklusioner bør drages baseret på dette resultat (både argument og konklusion skal være korrekt)?

1 ☐ Der bør laves en model til prædiktion af elprisen med `load` som den eneste forklarende variabel

2 ☐ Da korrelation mellem load og residualerne påvises signifikant forskellig fra nul, bør `load` inkluderes i modeludvælgelsen

3 ☐ Da koefficienten for `load` er positiv, bør `load` ikke inkluderes i modeludvælgelsen

4 ☐ Da koefficienten for `load` er positiv, bør `load` inkluderes i modeludvælgelsen

5 ☐ Baseret på de givne informationer kan ingen af ovenstående konklusioner drages

Fortsæt på side 8

Uanset udfaldet af overstående estimeres nu en model der inkluderer `load`. Dvs. at man benytter modellen

$$\texttt{price}_i = \beta_0 + \texttt{load}_i\beta_1 + \texttt{solar}\beta_2 + \texttt{wind}_i\beta_3 + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

eller opskrevet ved matrix notation

$$\texttt{price} = \boldsymbol{X}\boldsymbol{\beta}; \quad \epsilon_i \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}) \tag{2}$$

Til hjælp for de videre udregninger, er følgende R-kode givet (der kan kopieres til R), hvor `XX` er matricen $\boldsymbol{X}^T\boldsymbol{X}$

```
XX <- c(28, 73743, 851, 75239, 73743, 195282372, 2267307, 197399368,  851,
        2267307, 32733, 2159250  ,  75239, 197399368, 2159250, 235403829)
XX <- matrix(XX, ncol=4)
XX

##       [,1]      [,2]    [,3]      [,4]
## [1,]    28     73743     851     75239
## [2,] 73743 195282372 2267307 197399368
## [3,]   851   2267307   32733   2159250
## [4,] 75239 197399368 2159250 235403829
```

Spørgsmål II.5 (8)

Varians estimatet er udregnet til $\hat{\sigma}^2$, hvad er standard fejlen (standard error), for parameteren β_0?

1 ☐ $\hat{\sigma}^2\sqrt{\frac{28}{24}}$

2 ☐ $\sqrt{7.23\hat{\sigma}^2}$

3 ☐ $2.87\hat{\sigma}$

4 ☐ $\frac{\hat{\sigma}}{\sqrt{28}}$

5 ☐ $\sqrt{28\hat{\sigma}^2}$

Fortsæt på side 9

Opgave III

Ved udviklingen af en ny medicin blev en række kandidatstoffer testet i et laboratorieforsøg. Forsøget var designet således, at højere værdier af scoren `y` indikerer, at stoffet virker bedre. Forsøget blev udført på 7 forskellige stoffer (A til G, i R angivet i faktoren `drug`) og med 5 gentagelser. Nedenstående værdier af scoren blev observeret:

A	B	C	D	E	F	G
9.0	9.5	9.7	11.3	12.0	13.4	9.5
11.3	11.8	10.4	11.3	13.2	12.4	11.5
10.1	10.1	10.8	11.3	11.9	11.1	9.2
9.9	8.6	8.8	9.5	11.8	11.2	10.3
8.9	8.8	10.2	11.3	12.1	13.0	9.6

En ANOVA analyse blev udført af medicinudviklerne og det kan antages at de nødvendige forudsætninger for denne er opfyldt. Resultat af analysen er:

```
anova(lm(y ~ drug))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## drug       6 36.107  6.0179  6.9115 0.0001397 ***
## Residuals 28 24.380  0.8707
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Spørgsmål III.1 (9)

Hvad bliver konklusionen på signifikansniveau $\alpha = 5\%$ på grundlag af analyseresultat (både argument og konklusion skal være korrekt)?

1 ☐ Ingen af nedenstående konklusioner er korrekte

2 ☐ Der kan påvises en signifikant forskel på virkningen, da p-værdien er over signifikansniveauet

3 ☐ Der kan ikke påvises en signifikant forskel på virkningen, da p-værdien er under signifikansniveauet

4 ☐ Der kan påvises en signifikant forskel på virkningen, da p-værdien er under signifikansniveauet

5 ☐ Der kan ikke påvises en signifikant forskel på virkningen, da p-værdien er over signifikansniveauet

Spørgsmål III.2 (10)

Hvad er den kritiske værdi, ved signifikansniveau $\alpha = 0.05$ for testen i den udførte ANOVA analyse?

1 ☐ 2.445

2 ☐ 0.262

3 ☐ 0.913

4 ☐ 1.691

5 ☐ 1.701

Fortsæt på side 11

Opgave IV

I brætspillet Risk angriber spillerne hinanden ved at rykke hære (der består af armeer) over grænser på et landkort, hvorefter der udkæmpes et slag mellem de to hære. Reglerne er således at den angribende part slår med 3 terninger og den forsvarende part slår med 2 terninger. Angriberen bruger kun de to terninger med det højeste antal øjne. Der benyttes sædvanlige 6-sidede terninger.

Spørgsmål IV.1 (11)

En spiller skal til at angribe og ønsker at udregne sandsynligheden for at slå mindst 5 med mindst 2 af 3 terninger, hvilken af følgende R-komandoer udregner den ønskede sandsynlighed

1 ☐ `dbinom(2, size = 2, prob = 1/3)`

2 ☐ `1 - pbinom(1, size = 3, prob = 1/3)`

3 ☐ `1 - dbinom(2, size = 3, prob = 1/6)`

4 ☐ `pbinom(2, size = 2, prob = 1/3)`

5 ☐ `1 - dbinom(1, size = 2, prob = 1/6)`

Spørgsmål IV.2 (12)

En spiller skal til at angribe og ønsker at udregne sandsynligheden for at slå mindst 4 med mindst 2 af 3 terninger, samtidig med at forsvaren slår højst 3 med hver sine 2 terninger, hvad er den ønskede sandsynlighed?

1 ☐ 0.25

2 ☐ 0.125

3 ☐ 0.5

4 ☐ 0.0625

5 ☐ 0.75

Fortsæt på side 12

Opgave V

Tabellen nedenfor viser antallet af piger og drenge født i Danmark i år 2016 af kvinder på hhv. 20, 25, 30 og 35 år:

	20	25	30	35	Total
Piger	281	1490	2333	1430	5534
Drenge	282	1539	2536	1518	5875
Total	563	3029	4869	2948	11409

Dvs. at der blev født 281 piger af 20-årige mødre i år 2016.

Spørgsmål V.1 (13)

Hvad er 95%-konfidensintervallet for sandsynligheden for, at en gravid kvinde på 30 år føder en pige?

1 ☐ $[0.48, 0.52]$

2 ☐ $[0.47, 0.49]$

3 ☐ $[0.49, 0.51]$

4 ☐ $[0.46, 0.50]$

5 ☐ $[0.44, 0.48]$

Fortsæt på side 13

Spørgsmål V.2 (14)

Vi betragter nu alt vores data og ønsker at undersøge, om der kan påvises en signifikant forskel på sandsynligheden for at få en pige og sandsynligheden for at få en dreng. Hvad er p-værdien samt konklusionen for den tilhørende test, når vi benytter signifikansniveau $\alpha = 0.05$?

1 ☐ p-værdi $= 0.49$, og den relevante nulhypotese afvises.

2 ☐ p-værdi $= 0.94$, og den relevante nulhypotese kan ikke afvises.

3 ☐ p-værdi $= 0.0014$, og den relevante nulhypotese afvises.

4 ☐ p-værdi $= 0.06$, og den relevante nulhypotese afvises.

5 ☐ p-værdi $= 0.51$, og den relevante nulhypotese kan ikke afvises.

Spørgsmål V.3 (15)

Vi vil gerne undersøge, om sandsynligheden for at føde en pige kan antages ens for kvinder i de 4 undersøgte aldersgrupper. Under den tilhørende nulhypotese, hvad er det forventede antal af piger født af 25-årige kvinder?

1 ☐ 1514

2 ☐ 1469

3 ☐ 1490

4 ☐ 1383

5 ☐ 1560

Fortsæt på side 14

Spørgsmål V.4 (16)

Vi vil stadig undersøge, om sandsynligheden for at føde en pige kan antages ens for kvinder i de 4 undersøgte aldersgrupper. Hvad er antallet af frihedsgrader i den relevante χ^2-test?

1 ☐ 4

2 ☐ 1

3 ☐ 8

4 ☐ 7

5 ☐ 3

Spørgsmål V.5 (17)

Teststørrelsen, når vi undersøger, om sandsynligheden for at føde en pige er ens for kvinder i de 4 undersøgte aldersgrupper, er beregnet til 1.6943. Hvilket kald i R giver den tilhørende p-værdi? Antag at det rigtige antal frihedsgrader er gemt i variablen `x`.

1 ☐ `1-pchisq(1.6943, df=x, lower.tail = FALSE)`

2 ☐ `1-pchisq(sqrt(1.6943), df=x, lower.tail = FALSE)`

3 ☐ `pchisq(1.6943^2, df=x, lower.tail = FALSE)`

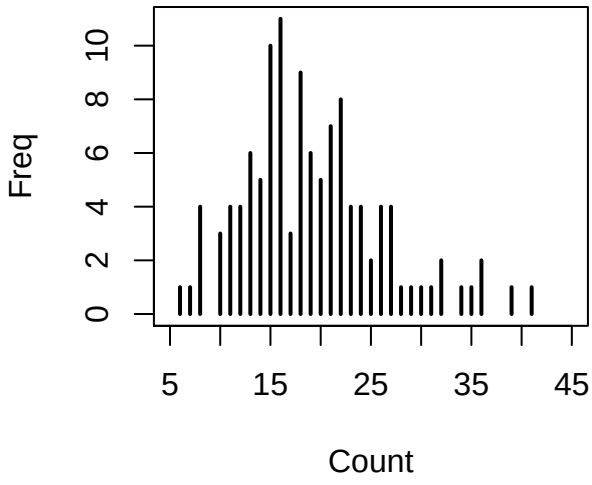
4 ☐ `pchisq(1.6943^2, df=x)`

5 ☐ `1-pchisq(1.6943, df=x)`

Fortsæt på side 15

Opgave VI

Figuren herunder viser det årlige antal såkaldte "Major earthquakes" (store jordskælv), globalt set i perioden 1900-2016.



Der er ialt 117 observationer (1900-2016), det gennemsnitlige antal er 19.2/år, med standardafvigelsen 6.97.

Spørgsmål VI.1 (18)

Hvis man antager at antallet af store jordskælv pr. år er Poisson fordelt med middelværdi 19.2, hvilket af følgende udsagn er da sandt?

1 ☐ Tiden mellem store jordskælv følger en Eksponentialfordeling med middelværdi 19.2

2 ☐ Tiden mellem store jordskælv følger en Binomialfordeling med middelværdi 19.2

3 ☐ Tiden mellem store jordskælv følger en Eksponentialfordeling med middelværdi 1/19.2

4 ☐ Tiden mellem store jordskælv følger en Poissonfordeling med middelværdi 19.2

5 ☐ Tiden mellem store jordskælv følger en Poissonfordeling med middelværdi 1/19.2

Fortsæt på side 16

Spørgsmål VI.2 (19)

Med samme antagelse som i forrige spørgsmål, hvad er da sandsyligheden for at observere mere end 40 store jordskælv i en toårig periode?

1 ☐ 0.581

2 ☐ 0.358

3 ☐ 0.419

4 ☐ 0.5

5 ☐ 0.642

Spørgsmål VI.3 (20)

Man ønsker nu et 95% konfidensinterval for parameteren λ (middelværdien) baseret på parametrisk bootstrap. For at belyse spørgsmålet har man kørt nedenstående R-kode, hvor `earthquakes` er en vektor med det årlige antal jordskælv

```
k <- 1000
n <- length(earthquakes)

sim <- replicate(k, rpois(n, lambda = mean(earthquakes)))
my.means <- apply(sim, 2, mean)
quantile(my.means, prob = c(0.025, 0.975))

##     2.5%    97.5%
## 18.40171 19.94038

mean(my.means) + sd(my.means) * c(-1, 1) * 1.96 / sqrt(n)

## [1] 19.11887 19.26401

mean(earthquakes) + sd(earthquakes) * c(-1, 1) * 1.96 / sqrt(n)

## [1] 17.91720 20.44178

sim2 <- replicate(k,sample(earthquakes, replace=TRUE))
my.means2 <- apply(sim2, 2, mean)
quantile(my.means2, prob = c(0.025, 0.975))

##     2.5%    97.5%
## 17.94850 20.45363

mean(my.means2) + sd(my.means2) * c(-1, 1) * 1.96

## [1] 17.88962 20.45184
```

Hvilket af følgende er et 95% konfidensinterval for middelværdien baseret på parametrisk bootstrap?

1 ☐ [17.95, 20.45]

2 ☐ [19.12, 19.26]

3 ☐ [17.89, 20.45]

4 ☐ [17.92, 20.44]

5 ☐ [18.40, 19.94]

Spørgsmål VI.4 (21)

Uanset svaret i foregående spørgsmål finder man fordelingsantagelsen tvivlsom og beslutter derfor at basere den videre analyse på ikke-parametrisk bootstrap. Man ønsker et estimat (inklusiv 95% konfidensinterval) for 75% fraktilen af årlige store jordskælv. Til dette formål har man kørt nedenstående R-kode

```
k <- 1000

sim1 <- replicate(k, sample(earthquakes, replace = TRUE))
Q1 <- function(x){ quantile(x, 0.25)}
samp1 <- apply(sim1, 2, Q1)
quantile(samp1, prob = c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    13    16
```

```
sim2 <- replicate(k, sample(earthquakes, replace = TRUE))
samp2 <- apply(sim2, 2, quantile, prob = 0.75)
quantile(samp2, prob = c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    21    25
```

```
sim3 <- replicate(k, rpois(length(earthquakes), lambda = mean(earthquakes)))
Q3 <- function(x){ quantile(x, 0.75)}
samp3 <- apply(sim3, 2, Q3)
quantile(samp3, prob = c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    21    23
```

```
sim4 <- replicate(k, rpois(length(earthquakes), lambda = mean(earthquakes)))
Q1 <- function(x){ quantile(x, 0.25)}
samp4 <- apply(sim4, 2, Q1)
quantile(samp4, prob = c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    15    17
```

```
sim5 <- replicate(k, sample(earthquakes))
quantile(sim5, prob = c(0.025, 0.975))
```

```
##  2.5% 97.5%
##     8    36
```

Hvad er 95% konfidensintervallet for 75% fraktilen?

1 ☐ [21, 23]

2 ☐ [8, 36]

3 ☐ [21, 25]

4 ☐ [15, 17]

5 ☐ [13, 16]

Spørgsmål VI.5 (22)

Man ønsker nu at undersøge om der er sket en udvikling i antallet af jordskælv i perioden, som en første undersøgelse ønskes et 95% ikke parametrisk bootstrap konfidensinterval for forskellen i middelværdi i to perioder (kaldet `earthquakes1` og `earthquakes2`), R-koden herunder er kørt for at belyse spørgsmålet

```
k <- 10000
n <- length(earthquakes)
n1 <- length(earthquakes1)
n2 <- length(earthquakes2)

sample1 <- replicate(k, sample(earthquakes1 - earthquakes2, replace = TRUE))
quantile(apply(sample1, 2, mean), probs = c(0.025, 0.975))
```

```
##   2.5%  97.5%
## 3.1995 7.9200
```

```
sample1 <- replicate(k, sample(earthquakes1, replace = TRUE))
sample2 <- replicate(k, sample(earthquakes2, replace = TRUE))
quantile(apply(sample1, 2, mean) -
         apply(sample2, 2, mean), probs = c(0.025, 0.975))
```

```
##   2.5%  97.5%
## 3.2200 8.1605
```

```r
sample1 <- replicate(k,sample(sort(earthquakes1) -
                              sort(earthquakes2), replace = TRUE))
quantile(apply(sample1, 2, mean), probs = c(0.025, 0.975))
```

```
##  2.5% 97.5%
##   5.0   6.3
```

```r
mean(earthquakes1 - earthquakes2) + c(-1, 1) * sd(earthquakes) * 1.96 / sqrt(n)
```

```
## [1] 4.377711 6.902289
```

```r
mean(earthquakes1 - earthquakes2) + c(-1, 1) * 1.96 *
    sqrt(var(earthquakes1) + var(earthquakes2)) * sqrt(1/n1 + 1/n2)
```

```
## [1] 2.086492 9.193508
```

Hvilket af følgende intervaller er et ikke-parametrisk bootstrap 95% konfidensinterval for forskellen i middelværdi?

1 ☐ [4.38, 6.90]

2 ☐ [3.22, 8.16]

3 ☐ [2.09, 9.19]

4 ☐ [3.20, 7.92]

5 ☐ [5.00, 6.30]

Fortsæt på side 20

Opgave VII

Et plantelaug har designet et spiringssanlæg til husstande, som placeres hos mange af deres medlemmer. De er nu på jagt efter den rette jordblanding, og derfor har de lavet et forsøg der har kørt hos flere af deres medlemmer hvor der testes fem forskellige jordblandinger. Hver deltager har lavet fem spiringer samtidigt - en med hver af de fem blandinger.

Spiretiden i timer er registreret for alle vha. et simpelt kamerasystem og følgende observationer er indsamlet:

	Blanding 1	Blanding 2	Blanding 3	Blanding 4	Blanding 5
Deltager 1	9.9	8.7	6.8	10.0	10.0
Deltager 2	5.2	3.1	8.8	10.0	5.0
Deltager 3	10.9	8.1	10.3	10.1	9.9
Deltager 4	6.1	10.6	7.6	7.1	9.8
Deltager 5	8.9	8.9	3.8	8.0	8.8
Deltager 6	5.9	6.4	5.9	6.0	7.7
Deltager 7	8.9	8.9	8.3	10.2	7.5
Deltager 8	10.3	13.4	9.3	9.6	9.7

En analyse er kørt (bemærk at enkelte værdier erstattet af et bogstav og eventuelle stjerner er fjernet):

```
anova(lm(Time ~ Blanding + Deltager))

## Analysis of Variance Table
##
## Response: Time
##           Df Sum Sq Mean Sq F value Pr(>F)
## Blanding   4    7.3    1.82    A     B
## Deltager   7   77.5   11.07    C     D
## Residuals 28   89.1    3.18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fortsæt på side 21

Spørgsmål VII.1 (23)

Hvilken konklusionen kan drages på signifikansniveau $\alpha = 5\%$ ud fra denne analyse (både konklusion og argument skal være korrekt)?

1 ☐ Der er en ikke signifikant effekt af blanding, men der er signifikant effekt af deltager, da de relevante p-værdier er henholdsvis 0.68 og 0.0083

2 ☐ Der er ikke signifikant effekt af hverken blanding eller deltager, da de relevante p-værdier er henholdsvis 0.093 og 0.17

3 ☐ Der er både en signifikant effekt af blanding og deltager, da de relevante p-værdier er henholdsvis 0.0034 og 0.014

4 ☐ Der er både en signifikant effekt af blanding og deltager, da de relevante p-værdier er henholdsvis 0.023 og 0.57

5 ☐ Der er en signifikant effekt af blanding, men ikke af deltager, da de relevante p-værdier er henholdsvis 0.0045 og 0.85

Spørgsmål VII.2 (24)

Hvor stor en del af den samlede variation er forklaret med den anvendte model?

1 ☐ $\frac{4+7}{28} = 0.393$

2 ☐ $\frac{7.3-1.82+77.5-11.07}{89.1-3.18} = 0.837$

3 ☐ $\frac{7.3+77.5}{89.1} = 0.952$

4 ☐ $\frac{7.3+77.5}{7.3+77.5+89.1} = 0.488$

5 ☐ $\frac{1.82+11.07+3.18}{7.3+77.5+89.1} = 0.0924$

Fortsæt på side 22

Opgave VIII

Som en del af et UX-designstudie blev brugere på en webside tilfældigt præsenteret for et af to websidelayouts (A og B). Brugerne skulle udføre en bestemt handling, og den brugte tid i sekunder blev målt.

Følgende kode indlæser data i R, `A` er brugere præsenteret for layout A, mens `B` er andre brugere der er præsenteret for layout B:

```
A <- c(8.360, 6.377,  7.385, 6.245, 8.766, 6.848, 6.074, 6.310, 5.946, 8.270)
B <- c(8.806, 6.513, 10.209, 5.495, 6.513, 8.529, 8.354, 5.681, 7.553, 6.834)
```

Spørgsmål VIII.1 (25)

Hvad er 99% konfidensintervallet for forskellen i middelværdi mellem de to grupper A og B?

1 ☐ [6.674, 7.783]

2 ☐ [-2.093, 1.312]

3 ☐ [-1.409, 0.628]

4 ☐ [-1.627, 0.846]

5 ☐ [-2.116, 1.335]

Spørgsmål VIII.2 (26)

Det sædvanlige t-test gav en p-værdi 0.513. Find den rette konklusion på signifikansniveau $\alpha = 0.05$ (både konklusion og argument skal være korrekt):

1 ☐ De to varianser er signifikant forskellige, da p-værdien er stor.

2 ☐ De to middelværdier er signifikant forskellige, da p-værdien er stor.

3 ☐ De to middelværdier kan rimeligvis antages at være ens, da p-værdien er stor.

4 ☐ De to middelværdier kan rimeligvis antages at være ens, da p-værdien er lille.

5 ☐ De to middelværdier er signifikant forskellige, da p-værdien er lille.

Fortsæt på side 23

Spørgsmål VIII.3 (27)

Websidesignerne vil nu forsøge sig med flere personer. Det antages at standardafvigelsen inden-for hver gruppe er 1.2, og at man skal bruge en t-test med signifikansniveau $\alpha = 0.05$.

Følgende kode er kørt i R:

```
power.t.test(n=40, delta=0.5, sd=1.2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 40
##           delta = 0.5
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.4524276
##     alternative = two.sided
##
## NOTE: n is number in *each* group

power.t.test(power=0.80, delta=0.4, sd=1.2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 142.2466
##           delta = 0.4
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Fortsæt på side 24

Hvilket af følgnede udsagn er korrekt?

1 ☐ Chancen for at et studie med $n = 40$ i hver gruppe finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 0.5, er ca. 45%.

2 ☐ Chancen for at et studie med $n = 142$ i hver gruppe finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 0.5, er ca. 80%.

3 ☐ Risikoen for at et studie med $n = 40$ i hver gruppe ikke finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 0.5, er ca. 45%

4 ☐ Risikoen for at et studie med $n = 142$ i hver gruppe ikke finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 0.5, er ca. 80%

5 ☐ Sandsynligheden for at $\mu_A \neq \mu_B$ er ca. 45%, når $n = 40$, og forøges til 80% når $n = 142$.

Fortsæt på side 25

Opgave IX

Lad $X_i \sim N(\mu_1, \sigma_1^2)$ og $Y_i \sim N(\mu_2, \sigma_2^2)$ være uafhængige stokastiske variable, antag desuden at man tager stikprøver fra hver af de to populationer af størrelsen hhv. n_1 og n_2. Lad S_p^2 betegne den sædvalige sammenvægtede varians estimator.

Spørgsmål IX.1 (28)

Hvad er $V(S_p^2)$?

1 ☐ $\frac{1}{2}(\sigma_1^2 + \sigma_2^2)$

2 ☐ $2\frac{(n_1-1)\sigma_1^4+(n_2-1)\sigma_2^4}{(n_1+n_2-2)^2}$

3 ☐ $\frac{\sigma_1^4}{(n_1-1)^2} + \frac{\sigma_2^4}{(n_2-1)^2}$

4 ☐ $2\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}$

5 ☐ $(\sigma_1^4 + \sigma_2^4)$

Definer nu

$$Q = \frac{n_1(\bar{X} - \mu_1)^2 + n_2(\bar{Y} - \mu_2)^2}{S_p^2} = \frac{D_1 + D_2}{S_p^2} \tag{3}$$

Spørgsmål IX.2 (29)

Idet vi ser Q som en funktion af D_1, D_2 og S_p^2. Hvad er fejlophobningslovens udtryk for variansen af Q, udtrykt ved σ_1^2, σ_2^2, $\mu_s = E[S_p^2]$ og $\sigma_s^2 = V[S_p^2]$, idet det oplyses at $E[D_1] = \sigma_1^2$, $E[D_2] = \sigma_2^2$, $V[D_1] = 2\sigma_1^4$ og $V[D_2] = 2\sigma_2^4$?

1 ☐ $\frac{1}{\mu_s^2}\left(2\sigma_1^2 + 2\sigma_2^2 + \frac{(\sigma_1^2+\sigma_2^2)^2}{\mu_s}\sigma_s^2\right)$

2 ☐ $\frac{1}{\mu_s^2}\left(2\sigma_1^2 + 2\sigma_2^2 + \frac{\sigma_1^2+\sigma_2^2}{\mu_s^4}\sigma_s\right)$

3 ☐ $\frac{1}{\mu_s}\left(2\sigma_1^2 + 2\sigma_2^2 + \frac{(\sigma_1^2+\sigma_2^2)}{\mu_s}\sigma_s^2\right)$

4 ☐ $\frac{1}{\mu_s^2}\left(2\sigma_1^4 + 2\sigma_2^4 + \frac{(\sigma_1^2+\sigma_2^2)^2}{\mu_s^2}\sigma_s^2\right)$

5 ☐ $\frac{1}{(\sigma_1^2+\sigma_2^2)^2}\left(2\sigma_1^4 + 2\sigma_2^4 + \frac{(\sigma_1^2+\sigma_2^2)^2}{\mu_s}\sigma_s^2\right)$

Spørgsmål IX.3 (30)

Antag nu at $\sigma_1^2 = \sigma_2^2 = \sigma^2$, hvad er det eksakte udtryk for $E(Q)$?

1 ☐ $2\frac{n_1+n_2-2}{n_1+n_2-4}$

2 ☐ $4\frac{n_1+n_2}{n_1+n_2-2}\sigma^2$

3 ☐ $4\frac{n_1+n_2}{n_1+n_2-2}$

4 ☐ $\frac{n_1+n_2-2}{n_1+n_2-4}\sigma^2$

5 ☐ $2\frac{n_1+n_2}{n_1+n_2-2}\sigma^2$

SÆTTET ER SLUT. God sommer!