Danmarks Tekniske Universitet *Side 1 af 26 sider.*

Skriftlig prøve: 30. maj 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	II.1	II.2	II.3	III.1	III.2	III.3	III.4	III.5	IV.1
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

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

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

Forekomsten af en sygdom i et erhverv er sådan, at arbejderne har 20% chance for at komme til at lide af den.

Spørgsmål I.1 (1)

Hvad er sandsynligheden for, at ud af 6 tilfældigt udvalgte arbejdere vil 4 eller flere få sygdommen?

1 ☐ 0.000064

2 ☐ 0.01536

3 ☐ 0.01696

4 ☐ 0.90112

5 ☐ 0.9984

Fortsæt på side 3

Opgave II

I en undersøgelse blev tre nye produkter testet for at sammenligne oplevelsen af dem. Produkterne fik navnene "A", "B"og "C". Prototyper af hvert produkt blev sendt til tilfældigt udvalgte testere, der rapporterede om deres erfaring med produktet gennem et interview. Deres svar blev bedømt efter hvor meget de kunne lide produktet og optalt i en af tre kategorier: "Lav", "Medium"eller "Høj". Resultaterne blev indlæst i R med:

```
mat <- matrix(c(24, 21, 14,
                12, 15, 22,
                15, 26, 24), ncol = 3, byrow = TRUE)
colnames(mat) <- c("Lav","Medium","Hoej")
rownames(mat) <- c("A","B","C")
```

og præsenteret i en tabel:

	Lav	Medium	Hoej
A	24	21	14
B	12	15	22
C	15	26	24

Forskere ønsker at teste, om oplevelsen af de tre produkter blev vurderet signifikant forskelligt. Derfor skal følgende nulhypotese testes

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

hvor $p_{i,j}$ angiver andelen i række i og kolonne j.

Spørgsmål II.1 (2)

Hvad er forventningsværdien af antallet i kategorien "Medium" bedømmelse for produkt "B" under nulhypotesen?

1 ☐ 0.087

2 ☐ 15.0

3 ☐ 16.3

4 ☐ 17.6

5 ☐ 19.1

Spørgsmål II.2 (3)

Hvad er resultatet af testen af nulhypotesen på et 5% signifikansniveau (både konklusionen og argumentet skal være korrekt)?

1 ☐ p-værdien er under signifikansniveauet, derfor må nulhypotesen accepteres.

2 ☐ p-værdien er under signifikansniveauet, derfor må nulhypotesen afvises.

3 ☐ p-værdien er over signifikansniveauet, derfor må nulhypotesen accepteres.

4 ☐ p-værdien er over signifikansniveauet, derfor må nulhypotesen afvises.

5 ☐ Der er ikke givet nok information til at beregne p-værdien og drage en konklusion.

Spørgsmål II.3 (4)

I dette spørgsmål anvendes kun observationerne for produkt "A":

Lav	Medium	Hoej
24	21	14

Hvad er 98% konfidensintervallet for andelen af "Lav" bedømmelse for produkt "A" (bemærk, at resultatet fra den relevante R-funktion er lidt forskelligt fra det korrekte svar, når det afrundes er det indenfor ± 0.01)?

1 ☐ $[0.26, 0.56]$

2 ☐ $[0.32, 0.81]$

3 ☐ $[0.37, 0.76]$

4 ☐ $[0.49, 0.83]$

5 ☐ $[0.52, 0.81]$

Fortsæt på side 5

Opgave III

Under hærdningsprocessen af beton øges temperaturen inde i betonen hurtigt i en periode på grund af kemiske processer. Efter temperaturforøgelsen, falder temperaturen indtil temperaturen i omgivelserne er nået. Betonens styrke kan vurderes ud fra temperaturprofilen (dvs. ud fra hvordan temperaturen steg og faldt).

I R-koden nedenfor indeholder `x1` en stikprøve af forskellen i temperatur inde i betonen fra starten til enden af hver time i løbet af Dag 3 efter at betonen er hældt ud (dvs. målinger af ændring i temperatur):

```
t.test(x1)

##
##  One Sample t-test
##
## data:  x1
## t = -4.1246, df = 21, p-value = 0.0004823
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.10939584 -0.03605871
## sample estimates:
##   mean of x
## -0.07272727
```

I nedenstående spørgsmål kan det antages, at observationerne er fra en normalfordeling med forventning μ_1 og varians σ_1^2, endvidere kan det antages, at observationerne er uafhængige.

Spørgsmål III.1 (5)

Baseret på ovenstående R-output og ved signifikansniveau $\alpha = 0.05$, kan det så konkluderes, at temperaturen er faldende, svarende til $\mu_1 < 0$ (både konklusionen og argumentet skal være korrekt)?

1 ☐ Ja, da $0.00048 < 0.05$

2 ☐ Ja, da $-0.073 < 0$

3 ☐ Nej, da $0.073 > 0.05$

4 ☐ Ja, da $-0.036 < 0$

5 ☐ Nej, da $-4.12 < 0$

I R-koden nedenfor repræsenterer `x2` forskellen mellem temperaturen indenfor hver time på Dag 4 efter betonen er hældt ud:

```
mean(x2)

## [1] -0.1181818

sd(x2)

## [1] 0.05884899

length(x2)

## [1] 22
```

Det kan antages, at observationerne er normale og uafhængige med middelværdi μ_2 og varians σ_2^2.

Spørgsmål III.2 (6)

μ_2 repræsenterer den forventede værdi på Dag 4, hvad er 95% konfidensintervallet for μ_2?

1 ☐ $[-0.144, -0.092]$

2 ☐ $[-0.140, -0.0966]$

3 ☐ $[-0.135, -0.101]$

4 ☐ $[-0.124, -0.113]$

5 ☐ $[-0.120, -0.117]$

Spørgsmål III.3 (7)

σ_2^2 repræsenterer variansen på Dag 4, hvad er 95% konfidensintervallet for standardafvigelsen σ_2?

1 ☐ $[0.0472, 0.0792]$

2 ☐ $[0.0453, 0.0841]$

3 ☐ $[0.0348, 0.120]$

4 ☐ $[0.00266, 0.00495]$

5 ☐ $[0.00205, 0.00707]$

Spørgsmål III.4 (8)

Hvis vi antager at der er ens varians i de to grupper, hvad er værdien af den sædvanlige teststørrelse for testen $H_0 : \mu_1 = \mu_2$ mod det tosidede alternativ?

1 ☐ 3.62

2 ☐ 1.81

3 ☐ 2.58

4 ☐ 1.78

5 ☐ 2.10

Spørgsmål III.5 (9)

Hvis vi bruger standardafvigelsen fra Dag 4 og signifikansniveau $\alpha = 0.05$ hvor mange observationer er da nødvendige for at detektere en middelværdi på -0.05 (når man bruger nulhypotesen om, at hældningen er nul), hvis den krævede styrke (power) er 0.9 (det korrekte svar er beregnet med formlen i bogen)?

1 ☐ 15

2 ☐ 8

3 ☐ 26

4 ☐ 5

5 ☐ 12

Fortsæt på side 8

Opgave IV

Man har indsamlet temperaturer (målt i °C) fra et område af Norditalien i årene 1984-2005. Data indlæstes i R:

```
temperature <- c(8.43, 7.89, 8.28, 7.84, 9.62, 9.41, 9.40, 8.22, 9.18, 9.17,
                 9.25, 9.68, 8.49, 8.53, 9.30, 8.94, 9.46, 9.69, 9.37, 9.42,
                 9.13, 9.18)
year <- 1984:2005
```

og vi har udført en lineær regression:

```
##
## Call:
## lm(formula = temperature ~ year)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80872 -0.31761  0.03158  0.29517  0.92517
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -82.97094   33.74565  -2.459   0.0232 *
## year          0.04611    0.01692   2.725   0.0130 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5035 on 20 degrees of freedom
## Multiple R-squared:  0.2708,Adjusted R-squared:  0.2343
## F-statistic: 7.427 on 1 and 20 DF,  p-value: 0.01304
```

Spørgsmål IV.1 (10)

Hvad er estimatet for den forventede/gennemsnitlige temperaturstigning over 10 år?

1 ☐ 0.046 °C

2 ☐ -8.2 °C

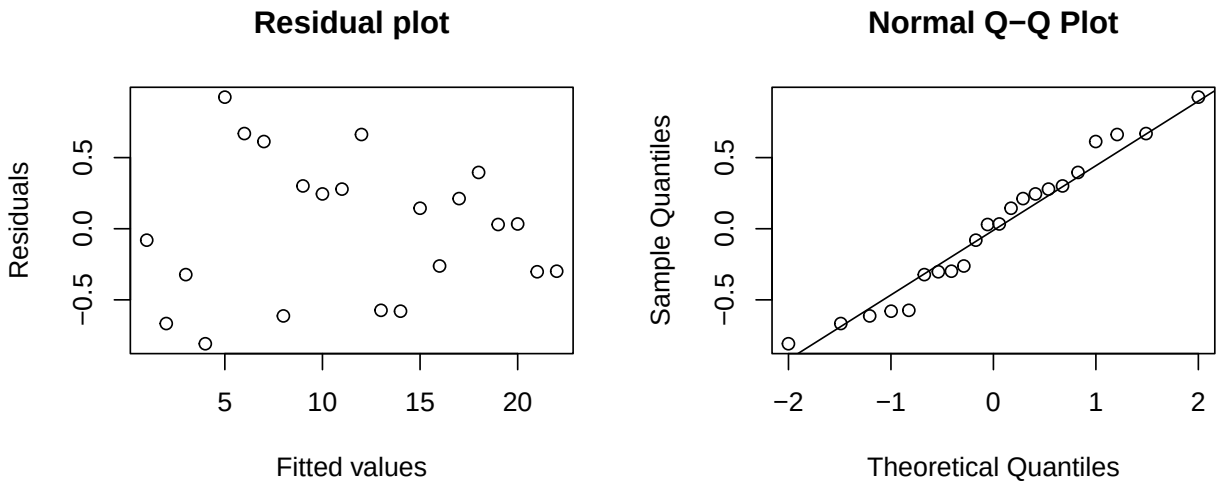
3 ☐ 0.017 °C

4 ☐ 2.7 %

5 ☐ 0.46 °C

Spørgsmål IV.2 (11)

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



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

1 ☐ Der ses ikke en lineær tendens i residualplottet. Dette er evidens for nulhypotesen om at tid ikke har en signifikant effekt på temperatur.

2 ☐ Residualplottet ser (rimeligt) fornuftigt ud, men q-q plottet er tvivlsomt. Dette indikerer et problem med normalitetsantagelsen.

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

4 ☐ Begge plots er tvivlsomme. Dette indikerer problemer både med antagelsen om lineær sammenhæng og med normalitetsantagelsen.

5 ☐ Både residualplottet og q-q plottet ser (rimeligt) fornuftige ud. Dette bekræfter modellens validitet.

Spørgsmål IV.3 (12)

Antag at temperaturen i samme område i 2017 var 10.91 °C. Hvilket af følgende udsagn er korrekt (både konklusionen og argumentet skal være korrekt)?

1 ☐ 95% konfidensintervallet er $[8.21, 9.86]$. Observationen passer fornuftigt med modellen.

2 ☐ 95% konfidensintervallet er $[8.21, 9.86]$. Observationen passer dårligt med modellen.

3 ☐ 95% prædiktionsintervallet er $[8.70, 11.37]$. Observationen passer fornuftigt med modellen.

4 ☐ 95% prædiktionsintervallet er $[8.70, 11.37]$. Observationen passer dårligt med modellen.

5 ☐ Ingen af ovenstående udsagn er korrekte.

Fortsæt på side 11

Opgave V

For at forstå hvorfor det kan være svært at bo i Danmark om vinteren, hvor der er lange mørke perioder, ønskes en analyse foretaget. Fra Dansk Meteorologisk Institut blev målinger af sollys i Isenvad, som ligger midt i Jylland, hentet for 10 års vintre. Ud fra målingerne er den længste periode uden registreret sollys beregnet for hver vinter:

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Periodelænge i dage	3.7	1.9	4.8	11.7	2.8	4.7	2.7	4.9	6.7	3.8

Spørgsmål V.1 (13)

Hvad er medianen af stikprøven?

1 ☐ 3.8

2 ☐ 4.25

3 ☐ 4.7

4 ☐ 4.77

5 ☐ 4.875

Spørgsmål V.2 (14)

Man ønsker at estimere et 90% konfidensinterval for gennemsnittet af den længste periode uden solskin i Isenvad for disse år. Stikprøven gemmes i vektoren `x`. Hvilke af følgende kodestykker beregner konfidensintervallet uden antagelse om fordeling?

1 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
quantile(apply(simsamples, 2, mean), c(0.025, 0.975))
```

2 ☐
```
simsamples <- replicate(10000, sample(x, replace = TRUE))
quantile(apply(simsamples, 2, mean), c(0.05, 0.95))
```

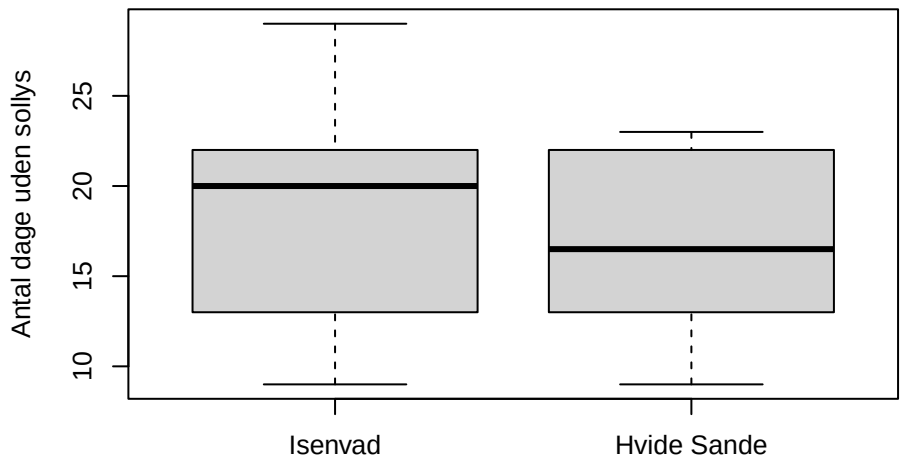
3 ☐
```
simsamples <- replicate(10000, sample(x, replace = FALSE))
quantile(apply(simsamples, 2, median), c(0.025, 0.975))
```

4 ☐
```
t.test(x, conf.level=0.9)
```

5 ☐
```
t.test(x, conf.level=0.95)
```

Spørgsmål V.3 (15)

Desuden ønskes en analyse, hvor sollys i Isenvad midt i Jylland, sammenlignes med sollys i Hvide Sande ved Jyllands vestkyst. Antallet af dage helt uden sollys i december og januar på hvert sted, for hver af de samme 10 år, er beregnet. Observationerne er opsummeret i følgende boxplots:



Observationerne er sorteret på årene og gemt i henholdsvis `x` for Isenvad og `y` for Hvide Sande:

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
x	9	13	18	29	19	22	13	22	26	21
y	9	12	13	23	17	22	15	17	22	16

Den følgende R-kode blev kørt:

```
mean(x)

## [1] 19.2

mean(y)

## [1] 16.6

k <- 10000
simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
sim1 <- apply(simxsamples, 2, mean) - apply(simysamples, 2, mean)
```

```r
simsamples <- replicate(10000, sample(x-y, replace = TRUE))
sim2 <- apply(simsamples, 2, mean)

quantile(sim1, c(0.005, 0.995))

##  0.5% 99.5%
##  -3.5   8.7

quantile(sim2, c(0.005, 0.995))

##  0.5% 99.5%
##   0.4   4.6
```

Hvilket af følgende udsagn er korrekt (både konklusion og argument skal være korrekt)?

1 ☐ To parametriske boostrapping 99% konfidensintervaller blev beregnet.

2 ☐ På 5% signifikansniveau kan det ikke konkluderes, at der er signifikant forskel mellem antallet af dage uden sollys på de to lokationer.

3 ☐ På 5% signifikansniveau kan det konkluderes, at der er signifikant forskel mellem antallet af dage uden sollys på de to lokationer.

4 ☐ Stikprøvegennemsnittet for Isenvad er lavere end for Hvide Sande.

5 ☐ Ingen af ovenstående udsagn er korrekte.

Fortsæt på side 14

Opgave VI

En familie er på udkig efter et sommerhus. De elsker virkeligt sommer og solskin, og vil for det meste bruge sommerhuset i juli. Derfor downloadede de data om solskinstimer observeret ved Hvide Sande ved Jyllands vestkyst og tilsvarende ved Hammer Odde på Bornholm. De har taget forskellen i solskinstimer, mellem de to lokationer, for hver dag i løbet af de sidste 10 år i juli.

Lad den i'te observerede forskel i solskinstimer repræsenteres af x_i, således at $x_i > 0$ indebærer, at der var mere sollys på Bornholm sammenlignet med på Jyllands vestkyst.

De beslutter at begrænse deres søgning efter et sommerhus til lokationen med flest solskinstimer, hvis en statistisk test kan vise forskel mellem lokaliteterne på signifikansniveau 5%. Værdierne gemmes i vektoren `x` i R, og følgende analyseresultat er udregnet:

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 4.722, df = 278, p-value = 3.708e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.150853 5.226247
```

Spørgsmål VI.1 (16)

Hvor mange observationer er inkluderet i den analyserede stikprøve?

1 ☐ $n = 9$

2 ☐ $n = 10$

3 ☐ $n = 11$

4 ☐ $n = 278$

5 ☐ $n = 279$

Spørgsmål VI.2 (17)

Hvad konkluderer familien baseret på ovenstående resultat (både argumentet og beslutningen om valg af lokation skal være korrekt)?

1 ☐ Der er stærk evidens mod $H_0 : \mu_X = 0$, hvilket leder dem til kun at søge efter sommerhus på Bornholm.

2 ☐ Der er stærk evidens mod $H_0 : \mu_X = 0$, hvilket leder dem til kun at søge efter sommerhus på Jyllands vestkyst.

3 ☐ Der er svag evidens mod $H_0 : \mu_X = 0$, hvilket leder dem til kun at søge efter sommerhus på Bornholm.

4 ☐ Der er svag evidens mod $H_0 : \mu_X = 0$, hvilket leder dem til kun at søge efter sommerhus på Jyllands vestkyst.

5 ☐ Der er lille eller ingen evidens mod $H_0 : \mu_X = 0$, hvilket leder dem til at søge efter sommerhus på begge lokationer.

Spørgsmål VI.3 (18)

Hvilket af følgende udsagn angående stikprøvegennemsnittet af `x` er korrekt?

1 ☐ Stikprøvegennemsnittet er 2.613.

2 ☐ Stikprøvegennemsnittet er 3.075.

3 ☐ Stikprøvegennemsnittet er 3.689.

4 ☐ Stikprøvegennemsnittet er 4.722.

5 ☐ Der er ikke givet nok information til at beregne stikprøvegennemsnittet.

Fortsæt på side 16

Opgave VII

Kunder i en bank ankommer tilfældigt og uafhængigt: sandsynligheden for en ankomst i en periode på 1 minut, er den samme som sandsynligheden for en ankomst i enhver anden periode på 1 minut. Besvar følgende spørgsmål under forudsætning af en gennemsnitlig ankomstrate på tre kunder i minuttet.

Spørgsmål VII.1 (19)

Hvad er sandsynligheden for, at nøjagtigt tre kunder ankommer i en tilfældigt valgt periode på 1 minut?

1 ☐ 0.2240

2 ☐ 0.4232

3 ☐ 0.5768

4 ☐ 0.6472

5 ☐ 0.7760

Spørgsmål VII.2 (20)

Hvad er sandsynligheden for at observere mindst tre ankomster i en tilfældigt valgt periode på 1 minut?

1 ☐ 0.2240

2 ☐ 0.4232

3 ☐ 0.5768

4 ☐ 0.6472

5 ☐ 0.7760

Fortsæt på side 17

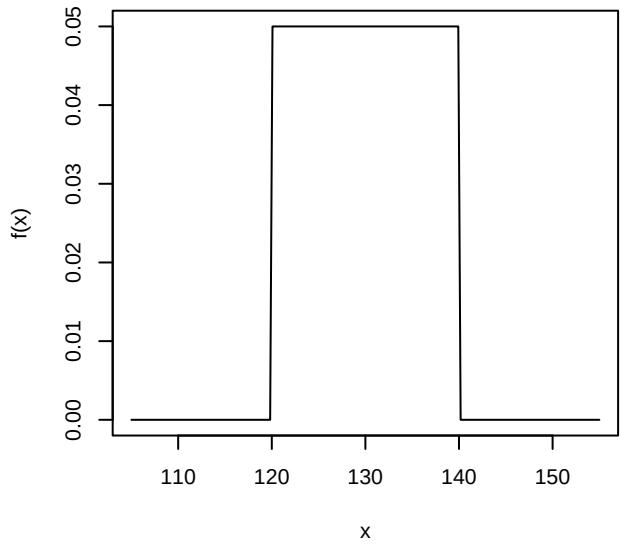
Opgave VIII

Lad den stokastiske variabel X repræsentere flyvetiden (i minutter) for et fly, der rejser fra Chicago til New York.

Tæthedsfunktionen for X er

$$f(x) = \begin{cases} 1/20 & 120 \leq x \leq 140 \\ 0 & \text{otherwise} \end{cases}$$

som er afbildet nedenfor:



Spørgsmål VIII.1 (21)

Hvad er sandsynligheden for en flyvetid mellem 120 og 140 minutter?

1 ☐ 0.2

2 ☐ 0.5

3 ☐ 0.8

4 ☐ 0.9

5 ☐ 1.0

Spørgsmål VIII.2 (22)

Hvad er standardafvigelsen på X?

1 ☐ 1.67

2 ☐ 3.33

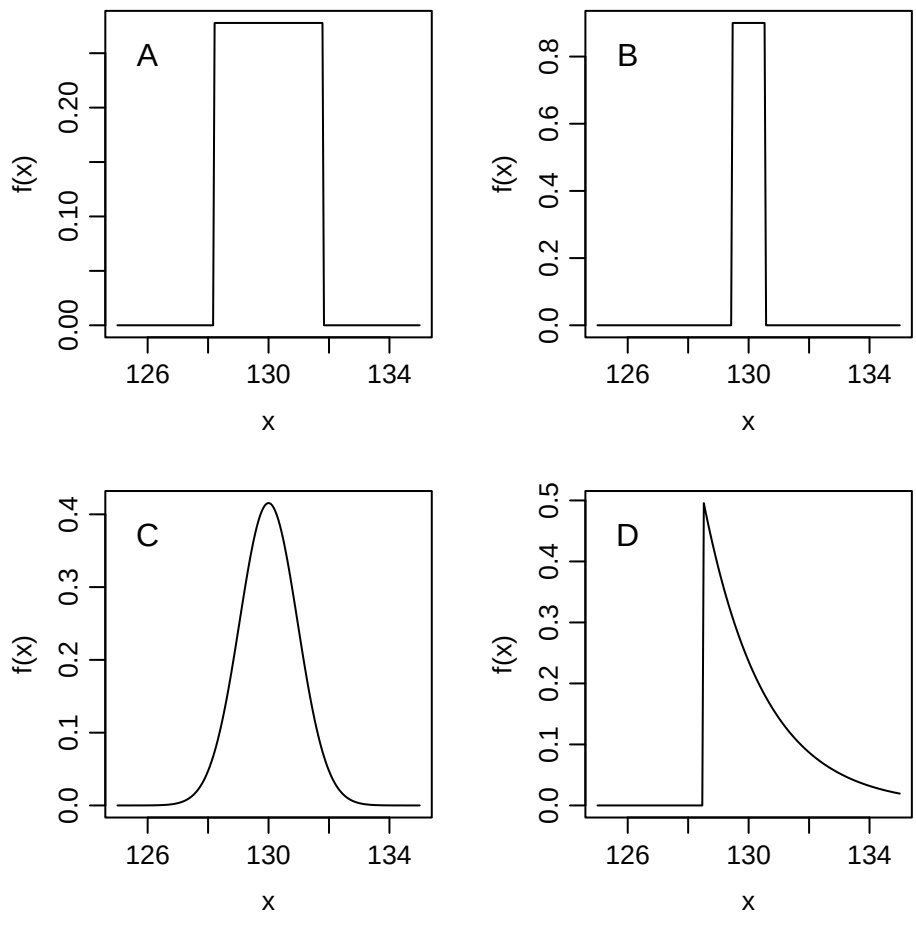
3 ☐ 4.47

4 ☐ 5.77

5 ☐ 33.33

Spørgsmål VIII.3 (23)

Hvis der blev taget en tilfældig stikprøve på $n = 36$ observationer af flyvetiderne, hvilket af følgende plot vil da repræsentere en god tilnærmelse af tæthedsfunktionen (pdf) for stikprøvegennemsnittet $\bar{X}$?



1 ☐ Plot A

2 ☐ Plot B

3 ☐ Plot C

4 ☐ Plot D

5 ☐ Ingen af de plottede tæthedsfunktioner kan være en god approximation til tæthedsfunktionen for stikprøvegennemsnittet $\bar{X}$.

Fortsæt på side 20

Opgave IX

Vi har observeret to variable, y og x_1, og har udført en lineær regression:

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

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44978 -0.20443 -0.12711  0.00835  1.11002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5178     0.1456  -10.43 6.21e-06 ***
## x1            0.4161     0.1547    2.69   0.0275 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4563 on 8 degrees of freedom
## Multiple R-squared:  0.4749,Adjusted R-squared:  0.4092
## F-statistic: 7.234 on 1 and 8 DF,  p-value: 0.02751
```

Spørgsmål IX.1 (24)

Betragt output-linjen `(Intercept)` i lm-resultatet. Hvilket af følgende udsagn er korrekt?

1 ☐ 'Std. Error' udtrykker usikkerheden på hældningsestimatet.

2 ☐ 'Std. Error' udtrykker usikkerheden på den forventede værdi/middelværdien til en observation, hvor $x_1 = 0$.

3 ☐ t-værdien kan bruges til at vurdere om der en en signifikant sammenhæng mellem x_1 og y.

4 ☐ t-værdien er et mål for modelkontrol. En lille t-værdi indikerer at modellen er valid.

5 ☐ Hverken 'Std. Error' eller t-værdien har at gøre med usikkerheden i modellen.

Spørgsmål IX.2 (25)

Antag at vi tilføjer yderligere en variabel i modellen, dvs. $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$.

Hvilket af følgende udsagn er korrekt?

1 ☐ Den forklarede varians r^2 (`Multiple R-squared` i lm-resultatet) aftager sammenlignet med modellen der kun har x_1 med.

2 ☐ Hvis man foretager baglæns selektering (backward selection), så vil man skulle fjerne x_2, hvis den tilhørende p-værdi i lm-resultatet er 0.0275.

3 ☐ Højst en af x_1 og x_2 er signifikant på et 5% signifikansniveau.

4 ☐ Mindst en af x_1 og x_2 er signifikant på et 5% signifikansniveau.

5 ☐ Hvert af de ovenstående udsagen er enten falsk, eller vi har for lidt information til at afgøre om det er.

Fortsæt på side 22

Opgave X

Som en del af en multilab-undersøgelse blev tre stoffer testet for brændbarhed ved National Bureau of Standards. Et stykke papir blev antændt på kanten af en kjole lavet af hvert stof og følgende forbrændingstider i minutter blev registreret:

Stof1	Stof2	Stof3
3.11	3.43	2.56
3.09	4.03	3.14
2.67	3.54	3.11
2.66	3.24	1.69
2.16	3.77	1.91
3.22	3.86	2.62
3.28	3.39	3.25

En envejs ANOVA blev udført. Den resulterende ANOVA tabel kan ses nedenfor (nogle elementer er erstattet med spørgsmålstegn):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Stof	?	3.71	?	8.91	0.0020
Residuals	?	3.75	?		

Spørgsmål X.1 (26)

Hvilket af følgende udsagn er korrekt?

1 ☐ Residuals: Df = 21 og Mean Sq = 0.18

2 ☐ Residuals: Df = 20 og Mean Sq = 0.19

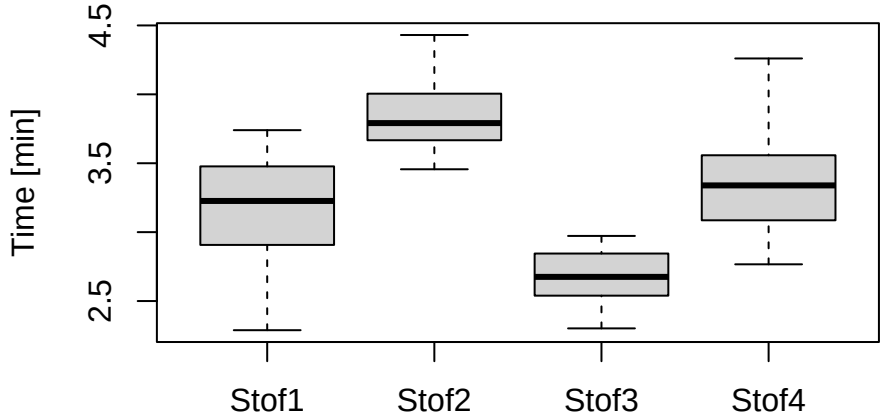
3 ☐ Residuals: Df = 19 og Mean Sq = 0.20

4 ☐ Residuals: Df = 18 og Mean Sq = 0.21

5 ☐ Stof: Df = 3 og Mean Sq = 1.237

Spørgsmål X.2 (27)

Brændbarhedsundersøgelsen gentages nu for fire andre stoftyper. Resultaterne vises i boxplot:



En envejs ANOVA blev udført. Resultatet er præsenteret i ANOVA-tabellen nedenfor (værdierne er afrundede):

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Stof	3	13.82	4.61	43.20	0.0000
Residuals	76	8.10	0.11		

Hvad ville estimatet for variansen i forbrændingstid være, hvis alle data blev betragtet som en prøve fra en enkelt population?

1 ☐ $\hat{\sigma}^2 = \frac{13.82}{3} + \frac{8.10}{76}$

2 ☐ $\hat{\sigma}^2 = \frac{13.82+8.10}{79}$

3 ☐ $\hat{\sigma}^2 = 4.61 + 0.11$

4 ☐ $\hat{\sigma}^2 = \frac{4.61}{3} + \frac{0.11}{76}$

5 ☐ Vi har ikke tilstrækkelig information til at beregne variansen.

Spørgsmål X.3 (28)

Se på ANOVA-tabellen fra det forrige spørgsmål. Vi vil teste følgende hypotese:

$$H_0 : \mu_{\text{Stof1}} = \mu_{\text{Stof2}} = \mu_{\text{Stof3}} = \mu_{\text{Stof4}} = \mu$$

Med et signifikansniveau $\alpha = 0.05$, hvilket R-kommando resulterer i korrekt kritisk værdi i den F-fordeling, der skal bruges til hypotesetesten?

1 ☐ `pf(0.05, 3, 76)`

2 ☐ `pf(0.95, 3, 76)`

3 ☐ `pf(0.975, 3, 79)`

4 ☐ `qf(0.95, 3, 76)`

5 ☐ `qf(0.975, 3, 76)`

Fortsæt på side 25

Opgave XI

Man er interesseret i at undersøge kvaliteten af fire forskellige undervisningsmetoder (A-D) med hensyn til studerendes præstationer. Et randomiseret blokdesign er blevet anvendt, hvilket betyder at tre studerende gennemgik alle fire undervisningsmetoder inklusive de tilsvarende examer i randomiseret rækkefølge. Følgende data er blevet indsamlet. Den bedst mulige eksamenspræstation er 100 (procent):

	Student1	Student2	Student3
A	84	89	91
B	85	87	91
C	85	88	89
D	86	90	96

En tovejs ANOVA med signifikansniveau $\alpha = 0.05$ blev udført for at undersøge om undervisningsmetoderne havde en signifikant forskellig effekt på elevernes præstationer. ANOVA-tabellen kan ses nedenfor, hvor nogle elementer er erstattet af spørgsmålstegn:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Student	2	91.17	45.583	21.312	0.002
Method	3	20.92	?	?	?
Residuals	6	12.83	2.139		

Spørgsmål XI.1 (29)

Hvilket af følgende udsagn er korrekt?

1 ☐ p-værdien for metode er 0.56 og der er ikke signifikant effekt af undervisningsmetoden.

2 ☐ p-værdien for metode er 0.18 og der er signifikant effekt af undervisningsmetoden.

3 ☐ p-værdien for metode er 0.18 og der er ikke signifikant effekt af undervisningsmetoden.

4 ☐ p-værdien for metode er 0.10 og der er ikke signifikant effekt af undervisningsmetoden.

5 ☐ p-værdien for metode er 0.10 og der er signifikant effekt af undervisningsmetoden.

Spørgsmål XI.2 (30)

ANOVA-tabellen ovenfor angiver, at der er en signifikant forskel mellem studerendes præstationer. Vi planlægger nu post-hoc-tests til parvis sammenligning af middelværdierne af elevernes præstation. Vi vil rette vores signifikansniveau α ved hjælp af Bonferroni-korrektion for ikke at øge risikoen for at lave en Type-I fejl.

Hvilket af følgende udsagn er korrekt?

1 ☐ Vi udfører 12 post-hoc test, derfor skal vi dividere α med 12.

2 ☐ Vi udfører 12 post-hoc test, derfor skal vi dividere α med 11.

3 ☐ Vi udfører 4 post-hoc test, derfor skal vi dividere α med 4.

4 ☐ Vi udfører 3 post-hoc test, derfor skal vi dividere α med 2.

5 ☐ Vi udfører 3 post-hoc test, derfor skal vi dividere α med 3.

SÆTTET ER SLUT. God sommer!