Danmarks Tekniske Universitet *Side 1 af 25 sider.*

Skriftlig prøve: 14. august 2019

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

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

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

Eksamenssættet består af 25 sider.

Fortsæt på side 2

Multiple choice opgaver: *Der gøres opmærksom på, at der i hvert spørgsmål er én og kun én svarmulighed, som er rigtig. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde. Husk altid at afrunde dit eget resultat til antallet af decimaler givet i svarmulighederne før du vælger et svar.*

Opgave I

Antag at $X_1, \ldots, X_{25}$ er uafhængige stokastiske variable, som alle er normalfordelte $N(5, 2^2)$.

Spørgsmål I.1 (1)

Hvilken af nedenstående værdier har den egenskab, at sandsynligheden for, at X_1 er lavere end denne værdi, er 15% (husk at svaret kan være afrundet)?

1 ☐ -0.85

2 ☐ 0.85

3 ☐ 2.93

4 ☐ 3.93

5 ☐ 5.43

Spørgsmål I.2 (2)

Hvad er sandsynligheden for, at stikprøvegennemsnittet $\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i$ er større end 4.5?

1 ☐ $P(\bar{X} > 4.5) = 0.89$

2 ☐ $P(\bar{X} > 4.5) = 0.85$

3 ☐ $P(\bar{X} > 4.5) = 2.05 \times 10^{-10}$

4 ☐ $P(\bar{X} > 4.5) = 0.18$

5 ☐ $P(\bar{X} > 4.5) = 0.55$

Fortsæt på side 3

Opgave II

Givet Lambert Beers lov kan lysets absorption gennem en væskeopløsning beregnes ved

$$A = \gamma \cdot l \cdot c$$

hvor γ er en konstant, l er strækning gennem væsken og c er koncentrationen af opløsningen.

Spørgsmål II.1 (3)

I betragtning af at standardafvigelsen for strækning σ_l og standardafvigelsen for koncentrationen σ_c er kendt, hvilken af de følgende formler vil approksimere standardafvigelsen af absorbansen?

1 ☐ $(\frac{\partial A}{\partial c})^2\sigma_l^2 + (\frac{\partial A}{\partial l})^2\sigma_c^2$

2 ☐ $\sqrt{(\frac{\partial A}{\partial c})^2\sigma_l^2 + (\frac{\partial A}{\partial l})^2\sigma_c^2}$

3 ☐ $(\frac{\partial A}{\partial l})^2\sigma_l^2 + (\frac{\partial A}{\partial c})^2\sigma_c^2$

4 ☐ $\sqrt{(\frac{\partial A}{\partial l})^2\sigma_l^2 + (\frac{\partial A}{\partial c})^2\sigma_c^2}$

5 ☐ $\sqrt{(\frac{\partial A}{\partial c})^2\sigma_l^2} + \sqrt{(\frac{\partial A}{\partial l})^2\sigma_c^2}$

Spørgsmål II.2 (4)

I et eksperiment er den gennemsnitlige strækning bestemt til at være 1 cm med en standardafvigelse på 0.1 cm. Den gennemsnitlige koncentration bestemmes til 0.65 M med en standardafvigelse på 0.09 M. γ er angivet som 0.3 $M^{-1}cm^{-1}$. Hvilke af følgende simuleringer kan bruges til at bestemme standardafvigelsen for absorbansen?

1 ☐
```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1)
c = rnorm(k, 0.65, 0.09)
A = e*l*c
var(A)
```

2 ☐
```
e = 0.3
l = rnorm(1, 1, 0.1^2)
c = rnorm(1, 0.65, 0.09^2)
A = e*l*c
sd(A)
```

3 ☐
```
e = 0.3
l = rnorm(1, 1, 0.1^2)
c = rnorm(1, 0.65, 0.09^2)
A = e*l*c
var(A)
```

4 ☐
```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1^2)
c = rnorm(k, 0.65, 0.09^2)
A = e*l*c
sd(A)
```

5 ☐
```
k = 10000
e = 0.3
l = rnorm(k, 1, 0.1)
c = rnorm(k, 0.65, 0.09)
A = e*l*c
sd(A)
```

Spørgsmål II.3 (5)

I ovennævnte spørgsmål blev en tilfældig prøve fra en normalfordeling simuleret ved hjælp af kommandoen `rnorm`. Hvilken af kommandoerne nedenfor kan bruges til at simulere en tilfældig prøve af længde `n` fra standardnormalfordelingen?

1 ☐
```
pnorm(runif(n))
```

2 ☐
```
qnorm(runif(n))
```

3 ☐
```
dnorm(runif(n))
```

4 ☐
```
qnorm(punif(n))
```

5 ☐
```
pexp(n)
```

Fortsæt på side 5

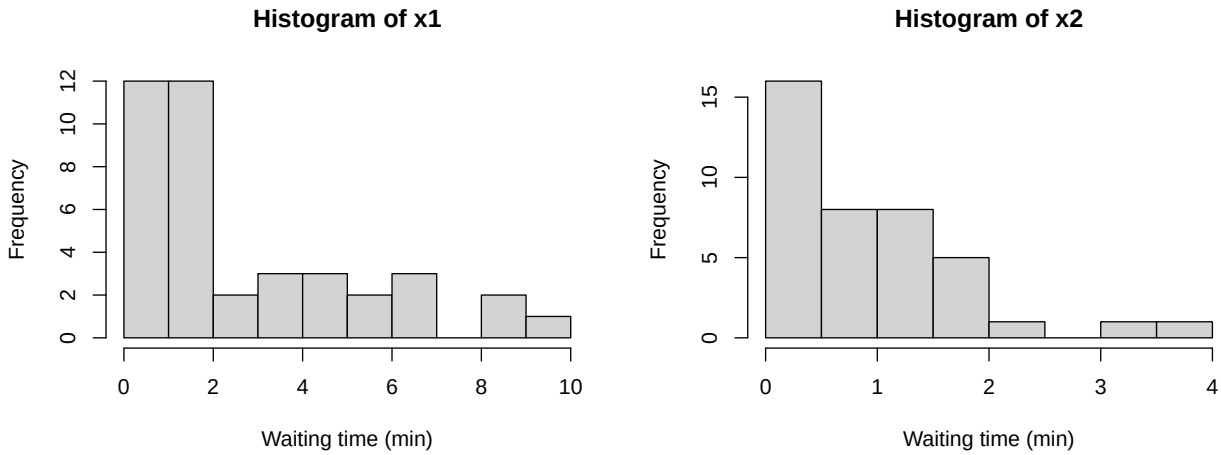
Opgave III

En supermarkedkædes personaleafdeling er interesseret i at sammenligne kundernes ventetider i to lokale butikker. Ventetiden (i minutter) er målt for 40 kunder i hver af de to butikker i løbet af en eftermiddag fra kl. 16.00 til kl. 17.00.

Lad $X_{1,i}$ repræsentere den i'te observerede ventetid i butik 1. Den kan antages at følge en eksponentielfordeling $X_{1,i} \sim Exp(\lambda_1)$ hvor $i = 1, \ldots, 40$.

Lad $X_{2,i}$ repræsentere den i'te observerede ventetid i butik 2. Den kan antages at følge en eksponentiel fordeling $X_{2,i} \sim Exp(\lambda_2)$ hvor $i = 1, \ldots, 40$.

Data fra hver prøve er gemt i R i vektorerne `x1` og `x2`, henholdsvist, og et histogram for hver stikprøve er plottet nedenfor:



De gennemsnitlige ventetider (i minutter) for de to butikker er:

```
mean(x1)

## [1] 2.76

mean(x2)

## [1] 0.897
```

Spørgsmål III.1 (6)

Giv et estimat af rateparametrene λ_1 og λ_2. Parametrene skal beregnes i kunder pr. time (h^{-1}).

1 ☐ $\hat{\lambda}_1 = 2.76\,\mathrm{h}^{-1}$ og $\hat{\lambda}_2 = 0.90\,\mathrm{h}^{-1}$

2 ☐ $\hat{\lambda}_1 = 0.36\,\mathrm{h}^{-1}$ og $\hat{\lambda}_2 = 1.11\,\mathrm{h}^{-1}$

3 ☐ $\hat{\lambda}_1 = 19.54\,\mathrm{h}^{-1}$ og $\hat{\lambda}_2 = 25.45\,\mathrm{h}^{-1}$

4 ☐ $\hat{\lambda}_1 = 21.74\,\mathrm{h}^{-1}$ og $\hat{\lambda}_2 = 66.89\,\mathrm{h}^{-1}$

5 ☐ Det er ikke muligt at estimere λ_1 og λ_2.

Spørgsmål III.2 (7)

I to andre butikker blev lignende prøver indsamlet. Parametrene blev estimeret til $\hat{\lambda}_1 = 0.23$ min^{-1} for den første butik og $\hat{\lambda}_2 = 0.39$ min^{-1} for den anden.

For at finde ud af om der var en signifikant forskel i middelventetid mellem de to butikker blev følgende beregninger udført i R:

```
k <- 10000
simX1_samples <- replicate(k, rexp(40, 0.23))
simX2_samples <- replicate(k, rexp(40, 0.39))
sim_dif_means <- apply(simX1_samples, 2, mean) - apply(simX2_samples, 2, mean)

quantile(sim_dif_means, c(0.005, 0.995))

##   0.5%  99.5%
## -0.121  3.955

quantile(sim_dif_means, c(0.025, 0.975))

##  2.5% 97.5%
##  0.30  3.42

quantile(sim_dif_means, c(0.05, 0.95))

##    5%   95%
## 0.547 3.156
```

Hvilket af følgende udsagn er korrekt?

1 ☐ Ikke-parametrisk bootstrapping blev udført. 95% konfidensintervalet er [-0.121, 3.955] og indeholder nul, derfor er middelventetiderne signifikant forskellige.

2 ☐ Ikke-parametrisk bootstrapping blev udført. 95% konfidensintervalet er [-0.121, 3.955] og indeholder nul, derfor er middelventetiderne ventetider ikke signifikant forskellige.

3 ☐ Parametrisk bootstrapping blev udført. 95% konfidensintervalet er [0.30, 3.42] og indeholder ikke nul, derfor er middelventetiderne signifikant forskellige.

4 ☐ Parametrisk bootstrapping blev udført. 95% konfidensintervalet er [0.30, 3.42] og indeholder ikke nul, derfor er middelventetiderne ikke signifikant forskellige.

5 ☐ Parametrisk bootstrapping blev udført. 95% konfidensintervalet er [0.547, 3.156] og indeholder ikke nul, derfor er middelventetiderne ikke signifikant forskellige.

Fortsæt på side 8

Opgave IV

Denne opgave handler om kvalitetskontrol i en virksomhed, der producerer harddiske til NAS ("Network Attached Storage"). Man vil undersøge sandsynligheden for, at en bestemt type harddisk går i stykker indenfor de første tre år ved "typisk brug". Virksomheden udvælger en tilfældig stikprøve med 950 harddiske fra deres produktion. De beder kunderne, der køber disse harddiske om at indrapportere, hvis en disk fejler indenfor de første tre år af dens levetid. Alle NAS-harddiskene antages at have samme sandsynlighed p for at fejle indenfor de første tre år, og diskene antages at fejle uafhængigt af hinanden.

Spørgsmål IV.1 (8)

Det blev indrapporteret at i alt 92 af harddiskene fejlede indenfor de første tre år af deres levetid. Angiv den estimerede standardafvigelse, $\hat{\sigma}_{\hat{p}}$, for den estimerede andel af harddiske, som går i stykker indenfor tre år:

1 ☐ $\hat{\sigma}_{\hat{p}} = 9.2 \cdot 10^{-5}$

2 ☐ $\hat{\sigma}_{\hat{p}} = 0.0031$

3 ☐ $\hat{\sigma}_{\hat{p}} = 0.0096$

4 ☐ $\hat{\sigma}_{\hat{p}} = 0.087$

5 ☐ $\hat{\sigma}_{\hat{p}} = 0.30$

Spørgsmål IV.2 (9)

Virksomheden har en målsætning om, at 90% af deres NAS-harddiske skal have en levetid på over tre år. Ved hjælp af et statistisk test ønsker de at undersøge, om de lever op til dette mål. Hvilken statistisk nulhypotese er da relevant at teste?

1 ☐ $H_0 : \ p = 0.1$

2 ☐ $H_0 : \ p = 0.9$

3 ☐ $H_0 : \ p \neq 0.1$

4 ☐ $H_0 : \ p \neq 0.9$

5 ☐ Ingen af ovenstående hypoteser kan benyttes.

Opgaveteksten fortsættes:
Virksomheden ønsker nu at sammenligne levetiden for deres særlige NAS-harddiske med levetiden for almindelige harddiske (når disse bruges i et NAS-setup). Til dette formål præsenterer de følgende antalstabel, hvor de også har inkluderet data for levetiden for 650 almindelige harddiske:

	NAS HDD	Alm. HDD	Total
$<$ 1 år	10	7	17
1-2 år	33	45	78
2-3 år	49	69	118
$>$ 3 år	858	529	1387
Total	950	650	1600

Denne tabel angiver, hvor mange af en bestemt type harddisk, der fejlede indenfor et bestemt aldersinterval. Ud fra denne tabel kan man f.eks. læse, at 69 ud af 650 almindelige harddiske gik i stykker efter 2-3 års brug. Disse data skal bruges i de resterende spørgsmål i denne opgave.

Spørgsmål IV.3 (10)

Virksomheden ønsker at undersøge, om de to typer harddiske har samme sandsynlighed for at fejle indenfor de første tre år af deres levetid. Hvilken af følgende R-koder udfører det relevante statistiske test?

1 ☐ `prop.test(x = c(49, 69), n = c(950, 650), correct = FALSE)`

2 ☐ `prop.test(x = c(49, 69), n = c(858, 529), correct = FALSE)`

3 ☐ `prop.test(x = c(92, 950), n = c(121, 650), correct = FALSE)`

4 ☐ `prop.test(x = c(92, 121), n = c(950, 650), correct = FALSE)`

5 ☐ Ingen af ovenstående.

Spørgsmål IV.4 (11)

Virksomheden kunne også have valgt at undersøge, om fordelingen af antallet harddiske der bryder sammen i de fire aldersintervaller er forskellig for de to typer diske. Under den tilsvarende nulhypotese H_0, hvad er så antallet af almindelige harddiske, som forventes at fejle efter 1-2 år?

1 ☐ 29

2 ☐ 33

3 ☐ 39

4 ☐ 45

5 ☐ Ingen af ovenstående tal er det korrekte svar.

Spørgsmål IV.5 (12)

Antag at virksomheden faktisk udfører et χ^2-test for at undersøge, om fordelingen af antallet af harddiske, der bryder sammen i de fire aldersintervaller, er forskellig for de to typer harddiske. Hvor mange frihedsgrader har den χ^2-fordeling, som anvendes i hypotesetestet?

1 ☐ 1 frihedsgrad

2 ☐ 2 frihedsgrader

3 ☐ 3 frihedsgrader

4 ☐ 6 frihedsgrader

5 ☐ 9 frihedsgrader

Fortsæt på side 11

Opgave V

På en lille ø ved man, at frekvensen af blackouts i det elektriske system er et blackout per uge. Definer den stokastiske variabel X, som antallet af blackouts for en tilfældigt udvalgt uge. Antallet af blackouts per uge antages at følge en poissonfordeling.

Spørgsmål V.1 (13)

Hvad er variansen af X?

1 ☐ $\sigma^2 = \frac{1}{7}$

2 ☐ $\sigma^2 = 0.368$

3 ☐ $\sigma^2 = 1$

4 ☐ $\sigma^2 = 2.72$

5 ☐ $\sigma^2 = 7$

Spørgsmål V.2 (14)

Hvad er sandsynligheden for at der ikke er noget blackout på en tilfældigt udvalgt dag?

1 ☐ 0.13

2 ☐ 0.24

3 ☐ 0.53

4 ☐ 0.76

5 ☐ 0.87

Fortsæt på side 12

Opgave VI

Man vil gerne sammenligne 5 grupper med 6 observationer i hver. Man laver derfor en envejs variansanalyse, hvor man tester hypotesen om, at alle grupper har samme middelværdi. Observationerner er lavet uafhængigt af hinanden. Teststørrelsen for testet bliver 4.30.

Spørgsmål VI.1 (15)

Hvad bliver p-værdien for testet?

1 ☐ 0.009

2 ☐ 0.0002

3 ☐ 0.03

4 ☐ 0.00001

5 ☐ 0.05

Fortsæt på side 13

Opgave VII

Man har fået følgende variansanalyseskema fra en en-vejs variansanalyse:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatm	2	1.19	A	B	C
Residuals	9	4.53	D		

Spørgsmål VII.1 (16)

Som det ses, mangler der nogle tal. Ved B burde der stå:

1 ☐ 0.26

2 ☐ 1.18

3 ☐ 1.53

4 ☐ 2.20

5 ☐ 7.40

Spørgsmål VII.2 (17)

Det oplyses, at der var lige mange observationer i hver gruppe. Hvor mange observationer var der i én af grupperne?

1 ☐ 2

2 ☐ 3

3 ☐ 4

4 ☐ 5

5 ☐ 9

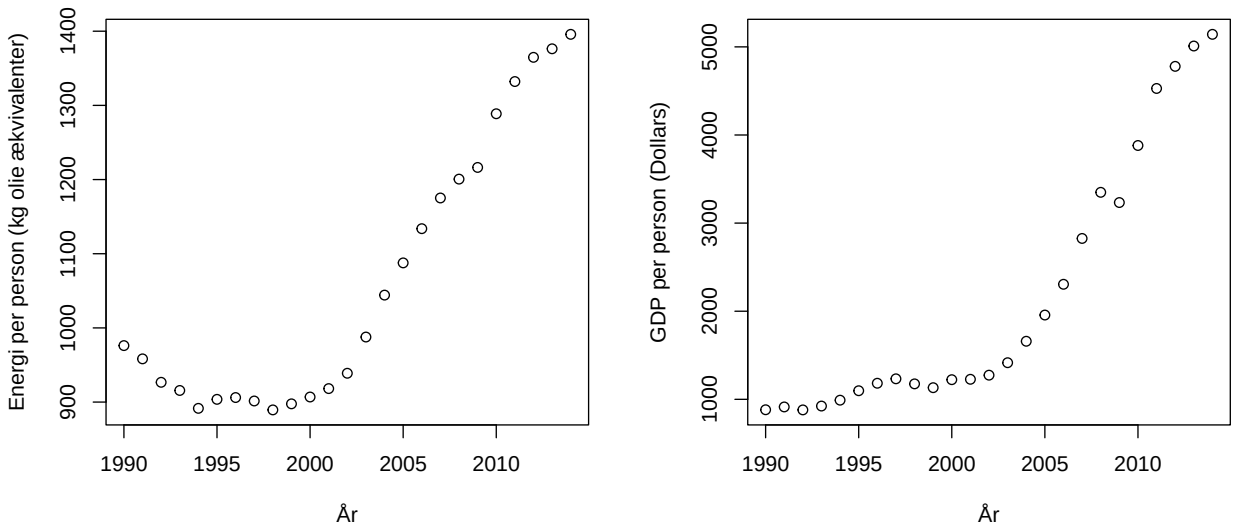
Fortsæt på side 14

Opgave VIII

Det er en ingeniørmæssig udfordring at udvikle teknologi, som kan dække verdens energiforbrug på en bæredygtig måde. Ved at bruge Verdensbankens befolkningsprognoser for 2050 kommer man frem til, at hvis alle skal have et energiforbrug, som befolkningen i de rige lande har nu, så energiforbruget om 30 år være omkring tre gange så stort som i 2014.

I denne opgave anvendes data hentet fra Verdensbanken, hvor verdens lande er inddelt i kategorierne: lav-, mellem- og højindkomstlande. Netop udviklingen i mellemindkomstlandene er meget betydende for udvikingen af verdens energiforbrug.

Følgende plot viser Energiforbruget og Brutto National Produktet (GDP) per år per person for mellemindkomstlandene i perioden 1990 til 2014:



Data er de plottede årlige værdier, som er gemt i vektorerne: `year` årstallet, `energy` energiforbrug og `gdp` er GDP. Kun dette data anvendes og alle konklusioner i opgaven gælder derfor kun mellemindkomstlandene i netop denne periode.

Først beregnes fire nøgletal:

```
c(mean(energy), mean(gdp))

## [1] 1061 2169

c(sd(energy), sd(gdp))

## [1]  179 1465
```

Derefter er to forskellige simple lineære regressionsmodeller estimeret:

```
summary(lm(energy ~ year))

##
## Call:
## lm(formula = energy ~ year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -122.49  -60.45    3.37   74.54  174.70
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42299.35    4669.35   -9.06  4.8e-09 ***
## year            21.66       2.33    9.29  3.0e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.1 on 23 degrees of freedom
## Multiple R-squared:  0.789,Adjusted R-squared:  0.78
## F-statistic: 86.2 on 1 and 23 DF,  p-value: 3.03e-09

summary(lm(energy ~ gdp))

##
## Call:
## lm(formula = energy ~ gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.82 -29.23  -9.45  27.37  69.09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.01e+02   1.35e+01    59.5   <2e-16 ***
## gdp         1.20e-01   5.18e-03    23.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.2 on 23 degrees of freedom
## Multiple R-squared:  0.959,Adjusted R-squared:  0.957
## F-statistic:  536 on 1 and 23 DF,  p-value: <2e-16
```

Spørgsmål VIII.1 (18)

Ifølge analyserne hvad er da den gennemsnitlige årlige stigning i energiforbrug i perioden estimeret til (i "kg olie ækvalent" pr. år)?

1 ☐ 0.120

2 ☐ 2.33

3 ☐ 3.37

4 ☐ 21.7

5 ☐ 801

Spørgsmål VIII.2 (19)

Hvad beregnes korrelationen mellem energiforbrug og GDP til?

1 ☐ 0.83

2 ☐ 0.93

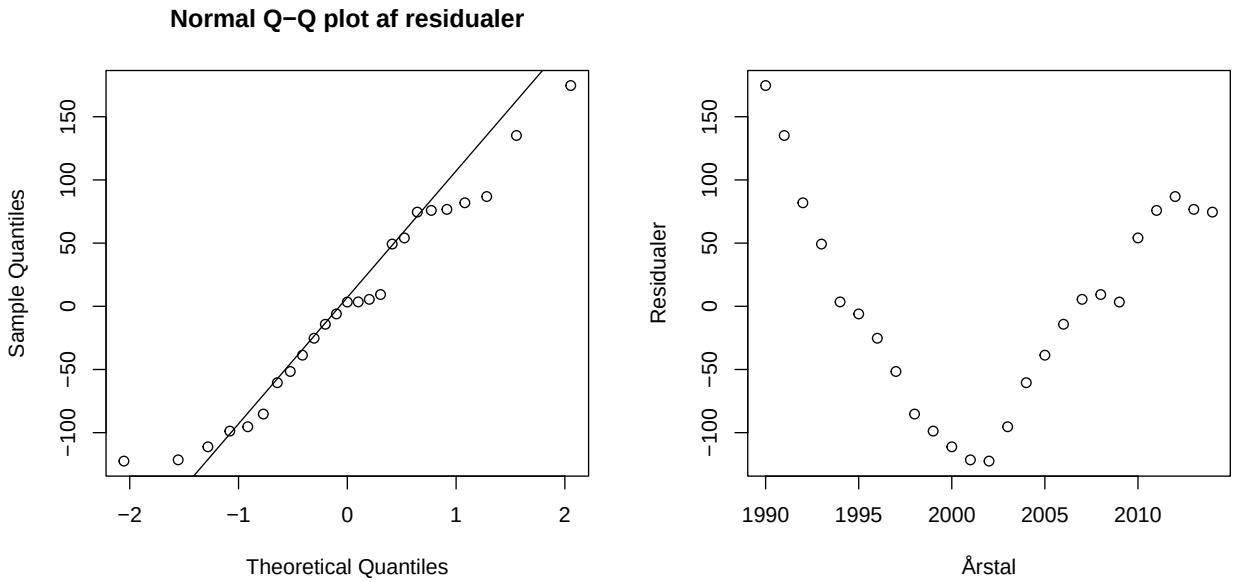
3 ☐ 0.98

4 ☐ 1.93

5 ☐ Dette kan ikke beregnes med de oplysninger, som er givet i opgaven.

Spørgsmål VIII.3 (20)

Følgende to plots er lavet til residualanalysen af den estimerede model mellem energiforbrug og årstal:



Hvilken af følgende konklusioner er mest hensigtsmæssig på baggrund af disse plots (både konklusion og argument skal være korrekt)?

1 ☐ Antagelsen om uafhængige afvigelser bør afvises, da residualernes fordeling ser ud til at være meget højreskæv.

2 ☐ Antagelsen om uafhængige afvigelser bør afvises, da residualernes fordeling ser ud til at være meget venstreskæv.

3 ☐ Antagelsen om uafhængige afvigelser bør afvises, da der tydeligt ses en lineær sammenhæng mellem residualerne og årstallene.

4 ☐ Antagelsen om uafhængige afvigelser bør afvises, da der tydeligt ses en ikke-lineær sammenhæng mellem residualerne og årstallene.

5 ☐ Ingen af ovenstående konklusioner med tilhørende argument er korrekter.

Spørgsmål VIII.4 (21)

Ifølge bogens definition er der da nogle ekstreme observationer i stikprøven, bestående af de observerede residualer fra den estimerede model mellem energiforbrug og årstal (både konklusion of argument skal være korrekt)?

1 ☐ Ja, da $-262.9 < 122.5$ og $174.7 < 277.0$.

2 ☐ Nej, da $-262.9 < 122.5$ og $174.7 < 277.0$.

3 ☐ Ja, da $135.0 < 297.2$.

4 ☐ Nej, da $135.0 < 297.2$.

5 ☐ Ja, da $0.5 < 0.789$.

Spørgsmål VIII.5 (22)

Man udvider nu modellen til en multipel lineær regressionsmodel, ved at anvende både årstallet og GDP som forklarende variabler.

Følgende resultat fås ved estimering af modellen:

```
summary(lm(energy ~ year + gdp))

##
## Call:
## lm(formula = energy ~ year + gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52.87 -29.05  -9.28  27.18  69.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.24e+02   4.98e+03    0.13     0.90
## year        8.93e-02   2.50e+00    0.04     0.97
## gdp         1.19e-01   1.26e-02    9.52    3e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38 on 22 degrees of freedom
## Multiple R-squared:  0.959,Adjusted R-squared:  0.955
## F-statistic:  256 on 2 and 22 DF,  p-value: 5.75e-16
```

Ved sammenligning af resultatet fra modellen med kun årstallet som forklarende variabel (fra opgavens start) og resultatet af modellen med både årstallet og GDP, kan følgende "absurde" konklusion drages for hypotesen om en sammenhæng mellem årstallet og energiforbruget:

Der findes meget stærke beviser for hypotesen, når årstallet alene anvendes som forklarende variabel, mens der findes få eller ingen beviser, når både årstallet og GDP anvendes.

Dette resultat er dog på ingen måde absurd statistisk set, da det ofte kan forekomme hvis følgende er tilfældet:

1 ☐ GDP er faldende i perioden.

2 ☐ Der er en relativ høj ikke-lineær sammenhæng mellem årstallet og energiforbruget i det observerede data.

3 ☐ Der er en relativ høj ikke-lineær sammenhæng mellem årstallet og GDP i det observerede data.

4 ☐ Der er en relativt høj korrelation mellem årstallet og energiforbruget i det observerede data.

5 ☐ Der er en relativt høj korrelation mellem årstallet og GDP i det observerede data.

Fortsæt på side 20

Opgave IX

I et studie af to typer grisefoder har man indelt en gruppe på 20 grise i to (mindre) grupper (x: gruppe 1 med 8 grise og y: gruppe 2 med 12 grise). De to grupper modtog fra de var 3 måneder til de var slagtemodne (6 måneder) hver sin type foder. Tabellen herunder viser grisenes slagtevægt (kg):

x	113.3	117.9	111.9	109.6	109.6	111.5	97.8	103.3				
y	110.7	108.3	110.6	106.7	109.7	107.5	105.9	111.0	99.9	110.2	99.4	103.6

Det oplyses at $\bar{x} = 109.4$, $\bar{y} = 107.0$, $s_x^2 = 6.2^2$ og $s_y^2 = 4.1^2$. Det kan antages at vægten af de slagtemodne grise følger en normalfordeling i hver gruppe. Man har desuden udregnet den sammenvægtede varians til $s_p^2 = 5.0^2$.

Spørgsmål IX.1 (23)

Hvad er 95% konfidensintervallet for middelværdien af slagtevægten af grise fra gruppe 1?

1 ☐ $[104.2, 114.6]$

2 ☐ $[105.2, 113.6]$

3 ☐ $[107.6, 111.2]$

4 ☐ $[101.7, 117.1]$

5 ☐ $[106.6, 112.2]$

Spørgsmål IX.2 (24)

Man ønsker at bestemme et 99% konfidensinterval for variansen af slagtevægten for gruppe 1. Hvordan beregnes dette korrekt?

1 ☐ $\left[\frac{7 \cdot 6.2^2}{20.3}, \frac{7 \cdot 6.2^2}{1.0}\right]$

2 ☐ $\left[\frac{8 \cdot 6.2}{20.3}, \frac{8 \cdot 6.2}{1.0}\right]$

3 ☐ $\left[\frac{9 \cdot 6.2}{20.3}, \frac{9 \cdot 6.2}{1.0}\right]$

4 ☐ $\left[\frac{8 \cdot 6.2^2}{20.3}, \frac{8 \cdot 6.2^2}{1.0}\right]$

5 ☐ $\left[\frac{7 \cdot 6.2}{20.3}, \frac{7 \cdot 6.2}{1.0}\right]$

Spørgsmål IX.3 (25)

Ved test for forskellen i slagtevægt mellem gruppe 1 og gruppe 2, hvad fåes da den sædvanligt anvendte Welch teststatistik til?

1 ☐ $|t_{\text{obs}}| = 0.96$

2 ☐ $|t_{\text{obs}}| = 1.0$

3 ☐ $|t_{\text{obs}}| = 2.6$

4 ☐ $|t_{\text{obs}}| = 49.8$

5 ☐ $|t_{\text{obs}}| = 90.8$

Spørgsmål IX.4 (26)

Hvis man i et nyt forsøg ønsker en styrke på 80% for at kunne detektere en forskel på 4 kg mellem de to grupper på konfidensniveau 99%, og bruger den sammenvægtede varians som gæt på populationens varians, hvor mange grise skal så mindst indgå i dette forsøg?

1 ☐ 22

2 ☐ 42

3 ☐ 52

4 ☐ 78

5 ☐ 104

Fortsæt på side 22

Opgave X

I forbindelse med en procedure til check af kemi i drikkevandet, tages målinger fra fem lokationer. For hver lokation laves en måling af samme stof med tre forskellige målemetoder. Alle målingerne kan antages at være taget uafhængigt af hinanden. Man ønsker at implementere en analyse, som kan afgøre om der er en signifikant forskel mellem lokationerne.

Spørgsmål X.1 (27)

Denne analyse vil bedst laves ved:

1 ☐ En en-vejs variansanalyse.

2 ☐ En to-vejs variansanalyse.

3 ☐ Et analyse baseret på t-tests.

4 ☐ Et analyse baseret på et χ^2 test.

5 ☐ En regressionsanalyse.

Fortsæt på side 23

Opgave XI

Tre forskellige maskiner bruges i en produktion af gummi. For at kontrollere om maskinerne producerer gummi af samme kvalitet, har man målt kvaliteten et antal gange for hver af de 3 maskiner. Observationerne er foretaget uafhængigt af hinanden. Man fik følgende observationer:

Maskine 1: 17.5, 16.9, 15.8, 18.6
Maskine 2: 16.4, 19.2, 17.7
Maskine 3: 20.3, 15.7, 17.8, 18.9

Spørgsmål XI.1 (28)

Testet for hypotesen af, at middelkvaliteten for de tre maskiner er ens, foretages ved at vurdere teststørrelsen i en:

1 ☐ Normalfordeling med middelværdi 0.

2 ☐ t-fordeling med 2 frihedsgrader.

3 ☐ t-fordeling med 8 frihedsgrader.

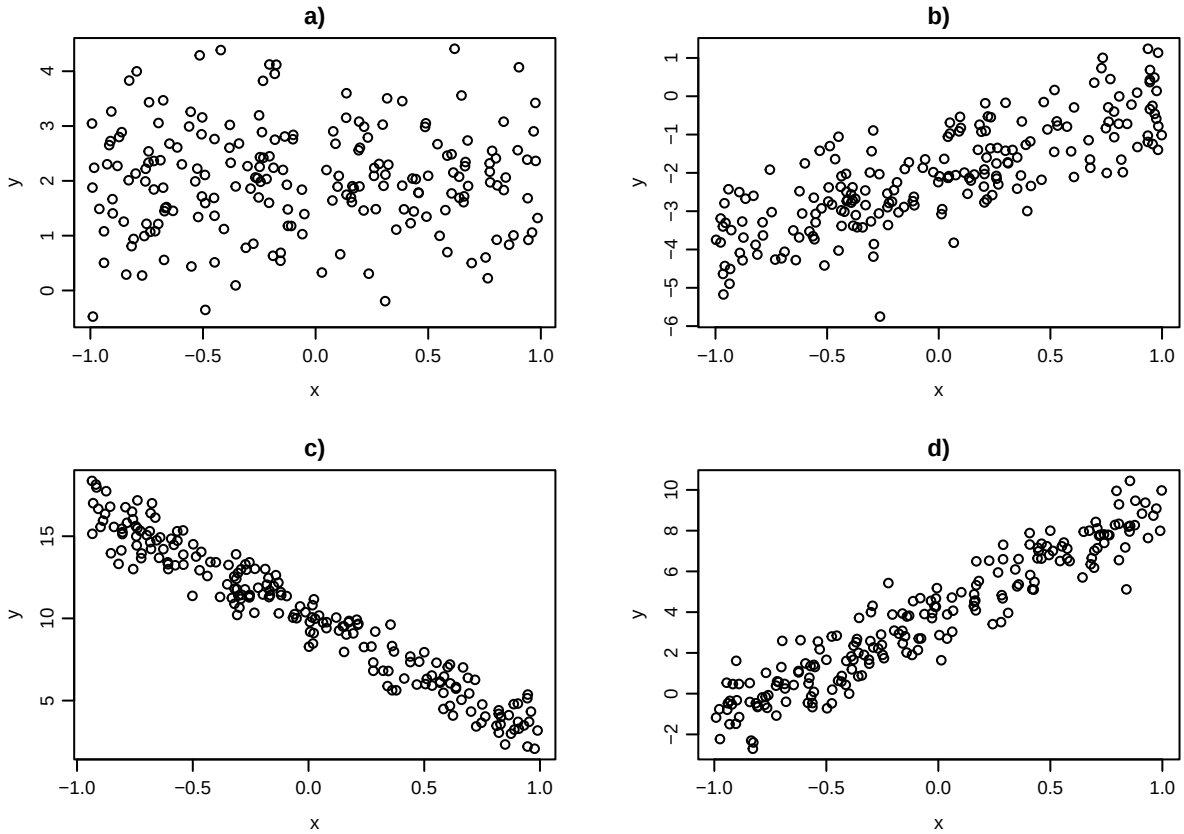
4 ☐ F-fordeling med 2 og 8 frihedsgrader.

5 ☐ Ingen af ovenstående.

Fortsæt på side 24

Opgave XII

Herunder ses fire scatterplots af y-målinger mod x-målinger:



Spørgsmål XII.1 (29)

Hvilke fire korrelationskoefficienter (i rækkefølgen: `a), b), c), d)`) passer bedst med figuren?

1 ☐ 0.02, 0.79, 0.95, -0.97

2 ☐ 0.02, 0.95, 0.79, -0.97

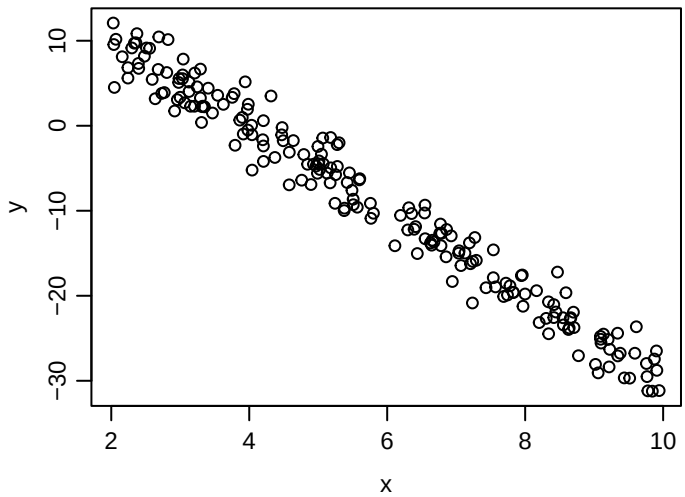
3 ☐ -0.97, 0.02, 0.79, 0.95

4 ☐ 0.02, 0.95, -0.97, 0.79

5 ☐ 0.02, 0.79, -0.97, 0.95

Spørgsmål XII.2 (30)

En anden samling x og y målinger er plotted herunder:



En lineær regression er udført på målingerne i plotted med R koden

```
summary(lm(y ~ x))
```

og resultatet for koefficientestimaterne er:

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)      A       0.0716   138.3   <2e-16 ***
## x                B       0.1236   -54.9   <2e-16 ***
```

De estimerede værdier af koefficienterne er erstattet med bogstaver.

Hvilken af følgende er svar er det eneste som ikke er meget usandsynligt?

1 ☐ A er 10 og B er -2.

2 ☐ A er 20 og B er -5.

3 ☐ A er 4 og B er -5.

4 ☐ A er 4 og B er -2.

5 ☐ A er 10 og B er -8.

SÆTTET ER SLUT. Fortsat god sommer!