Danmarks Tekniske Universitet *Side 1 af 28 sider.*

Skriftlig prøve: 28. maj 2017

Kursus navn og nr: **Introduktion til Statistik (02323 og 02402)**

Tilladte hjælpemidler: Alle

Dette sæt er besvaret af

_____	_____	_____
(studienummer)	(underskrift)	(bord nr)

Opgavesættet består af 30 spørgsmål af "multiple choice" typen fordelt på 11 opgaver. Besvarelserne af "multiple choice"spørgsmålene anføres i det i CampusNet uploadede svarark (på 6 seperate sider), med numrene på de svarmuligheder, du mener er de korrekte. **Der er et og kun et korrekt svar til hvert spørgsmål.**

Der gives 5 point for et korrekt "multiple choice" svar og -1 for et ukorrekt svar. KUN følgende 5 svarmuligheder er gyldige: 1, 2, 3, 4 eller 5. Hvis et spørgsmål efterlades blankt eller andet type svar angives, tæller det ikke med i besvarelsen. Envidere, hvis mere end et svar angives, hvilket faktisk er teknisk muligt i online-systemet, så tæller det heller ikke med (dvs. giver "0 point"). Det antal point, der kræves for, at et sæt anses for tilfredstillende besvaret, afgøres endeligt ved censureringen af sættene.

Den endelige besvarelse af opgaverne gøres ved at udfylde og online-aflevere svararket via CampusNet. Skemaet her er KUN et nød-alternativ til dette.

Opgave	I.1	I.2	II.1	III.1	III.2	III.3	III.4	III.5	III.6	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	VII.1	VII.2	VII.3
Spørgsmål	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Svar										

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

I "nødstilfælde": Husk at angive dit **studienummer** på din besvarelse. Sættets sidste side er nr. 28; check lige, at de alle er der.

Fortsæt på side 2

Multiple choice opgaver: *Der gøres opmærksom på, at ideen med opgaverne er, at der er ét og kun ét rigtigt svar på de enkelte spørgsmål. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde.*

Opgave I

Møn er Nordens første officielle Dark Sky område. En mand var taget dertil for at se på stjerner og stjerneskud. Det var en klar aften, og han havde læst, at han skulle forvente at kunne se 2 stjerneskud per minut.

Det kan antages at intensiteten af stjerneskud er konstant og at de kommer uafhængigt af hinanden.

Spørgsmål I.1 (1)

Hvilken fordeling beskriver bedst ventetiden i minutter mellem to stjerneskud?

1 ☐ En eksponentialfordeling med $\lambda = 2$

2 ☐ En Poisson-fordeling med $\lambda = 2$

3 ☐ En normalfordeling med $\mu = 1$ og $\sigma = \sqrt{2}$

4 ☐ En uniform fordeling med $\alpha = 0$ og $\beta = 2$

5 ☐ En Poisson-fordeling med $\lambda = 1/2$

Spørgsmål I.2 (2)

Hvad er sandsynligheden for, at der ikke kommer et stjerneskud i løbet af 4 minutter?

1 ☐ 0.8824969

2 ☐ 0.1353353

3 ☐ 0.9996645

4 ☐ 3.3546263×10^{-4}

5 ☐ Der skal være kommet mindst ét stjerneskud på 4 minutter.

Fortsæt på side 3

Opgave II

Ved syn af biler undersøger man blandt andet forbrændingen ved at måle på udstødningsgasserne. Antag at en bils sande udledning af partikler er 0.12 g/km, hvor EURO2 normens grænse er 0.08 g/km. Målingen har en standardafvigelse på 0.02 g/km. Det antages at målingerne er normalfordelte omkring den sande udledning. En bil bliver testet med en enkelt måling.

Spørgsmål II.1 (3)

Hvad er sandsynligheden for, at bilen bliver testet som havende for høj udledning af partikler?

1 ☐ 0.6113513

2 ☐ 0.0227501

3 ☐ 0.3886487

4 ☐ 0.9772499

5 ☐ 0.6113513

Fortsæt på side 4

Opgave III

I et oprensningsforsøg har man målt det såkaldte udbytte, når man har doseret en vis mængde enzym. Responsvariablen var det procentvise udbytte i forhold til det teoretisk højst opnåelige (X). Data fra 10 forskellige prøver fra forsøget var:

x_i
74.7
74.2
74.1
69.6
75.4
76.3
76.7
75.6
72.0
74.3
$\bar{x} = 74.29$
$s = 2.115$

Spørgsmål III.1 (4)

Hvad er 80%-fraktilen for disse data ved brug af bogens definition af fraktiler?

1 ☐ 74.10

2 ☐ 74.50

3 ☐ 75.95

4 ☐ 74.29

5 ☐ 75.60

Fortsæt på side 5

Spørgsmål III.2 (5)

Hvis man antager at $X \sim N(\mu, \sigma^2)$, og anvender de sædvanligt estimerede parametre ($\mu = \bar{x}$ og $\sigma = s$), hvilket er så det eneste udsagn, der kan være korrekt:

1 ☐ Mere end 99% af populationen er inden for $[72.18, 76.40]$ (I R: `mean(x) + c(-1, 1) * sd(x)`)

2 ☐ Mere end 99% af populationen er inden for $[70.06, 78.52]$ (I R: `mean(x) + c(-1, 1) * 2 *sd(x)`)

3 ☐ Omtrent 95% af populationen er inden for $[72.18, 76.40]$ (I R: `mean(x) + c(-1, 1) * sd(x)`)

4 ☐ Mere end 99% af populationen er inden for $[67.95, 80.63]$ (I R: `mean(x) + c(-1, 1) * 3 *sd(x)`)

5 ☐ Mindre end 95% af populationen er inden for $[67.95, 80.63]$ (I R: `mean(x) + c(-1, 1) * 3 *sd(x)`)

Spørgsmål III.3 (6)

Hvad er 95%-konfidensintervallet for middelværdien?

1 ☐ $2.262 \pm 74.29\frac{2.115}{\sqrt{10}} = [-47.42, 51.95]$

2 ☐ $74.29 \pm 1.812\frac{2.115}{\sqrt{9}} = [73.01, 75.57]$

3 ☐ $74.29 \pm 1.96\frac{2.115}{\sqrt{9}} = [72.91, 75.67]$

4 ☐ $74.29 \pm 2.262\frac{2.115}{\sqrt{10}} = [72.78, 75.80]$

5 ☐ $74.29 \pm 1.96\frac{2.115^2}{\sqrt{10}} = [71.52, 77.06]$

Fortsæt på side 6

Spørgsmål III.4 (7)

Hvis man skal lave et 95%-konfidensinterval for standardafvigelsen, hvilke fraktiler skal man så bruge?

1 ☐ `qnorm(0.025)` og `qnorm(0.975)`

2 ☐ `qchisq(0.025, 9)` og `qchisq(0.975, 9)`

3 ☐ `qf(0.025, 9, 9)` og `qf(0.975, 9, 9)`

4 ☐ `qt(0.025, 9)` og `qt(0.975, 9)`

5 ☐ `qunif(0.025)` og `qunif(0.975)`

Spørgsmål III.5 (8)

p-værdien for hypotesetestet af $H_0 : \mu = 70$ er:

1 ☐ `2*(1-pt(4.29/(2.115/sqrt(10)), 9))`

2 ☐ `2*(1-pnorm(70/(2.115/sqrt(10))))`

3 ☐ `(1-pt(-4.29/(2.115/sqrt(9)), 10))`

4 ☐ `1-pnorm(-4.29/(2.115/sqrt(10)))`

5 ☐ `1-qt(2.115/4.29, 9)`

Spørgsmål III.6 (9)

I et nyt forsøg, der planlægges, ønskes et 95% konfidensinterval for middelværdien med en forventet bredde på omkring 1. Antag et standardafvigelsen er 2.115. Hvor stor en stikprøve kræver det omtrent for at opnå denne ønskede præcision?

1 ☐ 5

2 ☐ 1230

3 ☐ 4

4 ☐ 100

5 ☐ 69

Fortsæt på side 7

Opgave IV

I et oprensningsforsøg har man undersøgt 2 forskellige doseringer af et enzym, benævnt d_1 og d_2, med formålet at undersøge en eventuel effekt for udbyttet. Responsvariablen var det procentvise udbytte i forhold til det teoretisk højst opnåelige. Data fra 19 forskellige prøver fra forsøget var:

Dosis d_1	Dosis d_2
74.7	79.6
74.2	77.5
74.1	82.5
69.6	76.7
75.4	78.2
76.3	76.7
76.7	76.6
75.6	78.1
72.0	79.2
74.3	
$\bar{x}_1 = 74.29$	$\bar{x}_2 = 78.34$
$s_1 = 2.115$	$s_2 = 1.898$

Følgende blev kørt i R:

```
x1 <- c(74.7, 74.2, 74.1, 69.6, 75.4, 76.3, 76.7, 75.6, 72.0, 74.3)
x2 <- c(79.6, 77.5, 82.5, 76.7, 78.2, 76.7, 76.6, 78.1, 79.2)
t.test(x2, x1)

##
##  Welch Two Sample t-test
##
## data:  x2 and x1
## t = 4.4041, df = 17, p-value = 0.0003878
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.112129 5.996760
## sample estimates:
## mean of x mean of y
##  78.34444  74.29000
```

Fortsæt på side 8

Spørgsmål IV.1 (10)

Hvad er 99%-konfidensintervallet for forskellen i middelværdien for dosis d_2 og dosis d_1?

1 ☐ $[2.45, 5.66]$

2 ☐ $[1.39, 6.72]$

3 ☐ $[1.97, 6.14]$

4 ☐ $[2.11, 6.00]$

5 ☐ $[74.29, 78.34]$

Spørgsmål IV.2 (11)

Konklusionen på det sædvanlige t-test (baseret på $\alpha = 0.05$) for denne situation er (både konklusion og argument skal være korrekt):

1 ☐ De to varianser er signifikant forskellige idet p-værdien er lille

2 ☐ De to middelværdier er signifikant forskellige idet p-værdien er stor

3 ☐ De to middelværdier er omtrentlig ens idet p-værdien er stor

4 ☐ De to middelværdier er omtrentlig ens idet p-værdien er lille

5 ☐ De to middelværdier er signifikant forskellige idet p-værdien er lille

Fortsæt på side 9

Spørgsmål IV.3 (12)

Et nyt studie med 2 doser planlægges. Der antages at standardafvigelsen inden for hver gruppe er 2, og at der skal udføres et t-test på niveau $\alpha = 0.05$. Følgende ting køres i R:

```
power.t.test(n=30, delta=2, sd=2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 30
##           delta = 2
##              sd = 2
##       sig.level = 0.05
##           power = 0.9677083
##     alternative = two.sided
##
## NOTE: n is number in *each* group

power.t.test(power=0.80, delta=1, sd=2, sig.level=0.05)

##
##      Two-sample t test power calculation
##
##               n = 63.76576
##           delta = 1
##              sd = 2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Kun et af følgende udsagn er sandt. Hvilket?

1 ☐ Risikoen for at et studie med $n = 30$ i hver gruppe ikke finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 1, er ca. 97%

2 ☐ Chancen for at et studie med $n = 30$ i hver gruppe finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 1, er ca. 97%

3 ☐ Chancen for at et studie med $n = 64$ i hver gruppe finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 1, er ca. 80%

4 ☐ Risikoen for at et studie med $n = 64$ i hver gruppe ikke finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 1, er ca. 80%

5 ☐ Risikoen for at et studie med $n = 30$ i hver gruppe ikke finder en signifikant forskel på middelværdierne, hvis den reelle forskel er 2, er ca. 97%

Fortsæt på side 10

Opgave V

På danske kraftværker bliver materiale afbrændt for at generere elektricitet og i forbrændingen udledes CO2. Det er en gas, som forstærker drivhuseffekten og derfor medvirker til opvarmning af atmosfæren. Dette fører mange negative konsekvenser med sig, og man er derfor interesseret i at nedbringe CO2 udledningerne. Dette gøres ved at indføre mere vind- og solenergiproduktion i systemet.

Hver dag beregnes CO2 udledningen (i gram af CO2 ækvivalent gas) per producerede kWh elektricitet i Danmark på baggrund af data fra ENTSO-E om elproduktionen.

I kolonne 1 i tabellen ses datoen. I kolonne 2 er gennemsnitsværdier af CO2 angivet for de 15 dage med højst vindenergiproduktion i perioden fra 1. dec. 2016 til 1. april 2017. Kolonne 3 i tabellen angiver elproduktion med kul. Denne kolonne anvendes ikke i de to første spørgsmål.

t	co2intensity (gCO2eq/kWh)	coal (MW)
2016-12-02	230	1016
2016-12-25	205	817
2016-12-26	203	746
2017-01-01	212	948
2017-01-05	292	1448
2017-01-12	260	1398
2017-02-08	317	1409
2017-02-12	321	1578
2017-02-21	235	1102
2017-02-22	268	1325
2017-02-23	233	1187
2017-03-01	253	1195
2017-03-02	260	1093
2017-03-16	212	976
2017-03-22	250	1095

Dette er indlæst i R i data.table `X` og følgende er kørt i R:

```
## Put observations in x
x <- X$co2intensity
## Number of simulated samples
k <- 100000
n <- length(x)

## Simulation
simsamples <- replicate(k, sample(x, replace = TRUE))
## Calculate the mean of each simulated sample
simmeans <- apply(simsamples, 2, mean)
```

Fortsæt på side 11

```r
## Quantiles of the differences gives the CI
quantile(simmeans, c(0.005, 0.995))

##     0.5%    99.5%
## 227.2667 275.2000

quantile(simmeans, c(0.01, 0.99))

##       1%      99%
## 229.3333 272.8000

quantile(simmeans, c(0.025, 0.975))

##     2.5%    97.5%
## 232.3333 269.2667

quantile(simmeans, c(0.05, 0.95))

##       5%      95%
## 235.0000 265.9333

quantile(simmeans, c(0.1, 0.9))

##      10%      90%
## 238.0667 262.4000
```

Spørgsmål V.1 (13)

Der ønskes et 99% bootstrap konfidensinterval for middelværdien af CO2-intensiteten uden antagelse af fordeling. Hvad bliver det korrekte interval?

1 ☐ $[227, 275]$

2 ☐ $[229, 273]$

3 ☐ $[232, 269]$

4 ☐ $[235, 266]$

5 ☐ $[238, 262]$

Fortsæt på side 12

Spørgsmål V.2 (14)

Hvilket et af følgende udsagn er ikke korrekt? Hvert udsagn omhandler en nulhypotese fremsat om middelniveauet af CO2-intensiteten μ_{CO2} ved høj vindenergiproduktion, og konklusionen drages baseret på resultaterne fra den kørte R kode? (Bemærk igen: der er 4 sande og 1 usandt udsagn - man skal udpege det usande!)

1 ☐ Nulhypotesen $H_0 : \mu_{CO2} = 200$ ville være afvist på 10% signifikansniveau

2 ☐ Nulhypotesen $H_0 : \mu_{CO2} = 220$ ville være afvist på 5% signifikansniveau

3 ☐ Nulhypotesen $H_0 : \mu_{CO2} = 230$ ville være afvist på 5% signifikansniveau

4 ☐ Nulhypotesen $H_0 : \mu_{CO2} = 270$ ville være afvist på 5% signifikansniveau

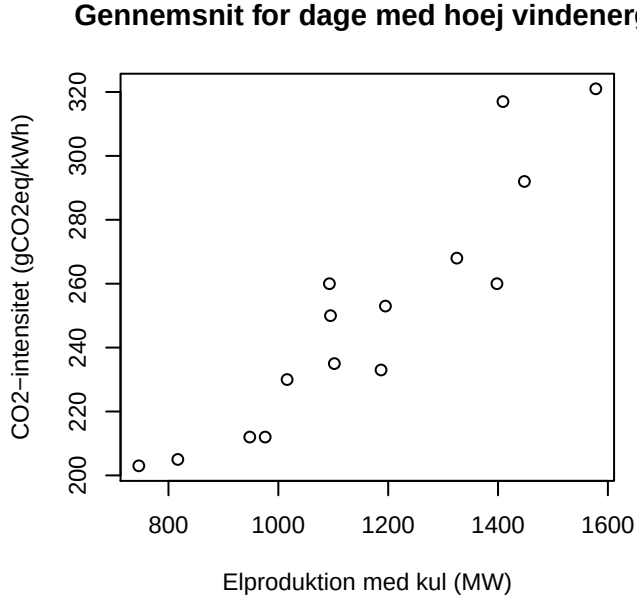
5 ☐ Nulhypotesen $H_0 : \mu_{CO2} = 270$ ville være afvist på 1% signifikansniveau

Fortsæt på side 13

Spørgsmål V.3 (15)

I dette spørgsmål undersøges sammenhængen mellem den kulfyrede elproduktion (kolonne 3) og CO2-intensiteten ved høj vindenergiproduktion (kolonne 2).

For at visualisere sammenhængen er følgende scatterplot genereret af værdierne:



Udfra en betragtning af plottet, hvilket af følgende udsagn er den mest korrekte betragtning?

1 ☐ Der er en korrelation omkring -1.2

2 ☐ Der er en korrelation omkring 0.1

3 ☐ Der er en korrelation omkring 0

4 ☐ Der er en korrelation omkring 0.9

5 ☐ Der er en korrelation omkring 1.2

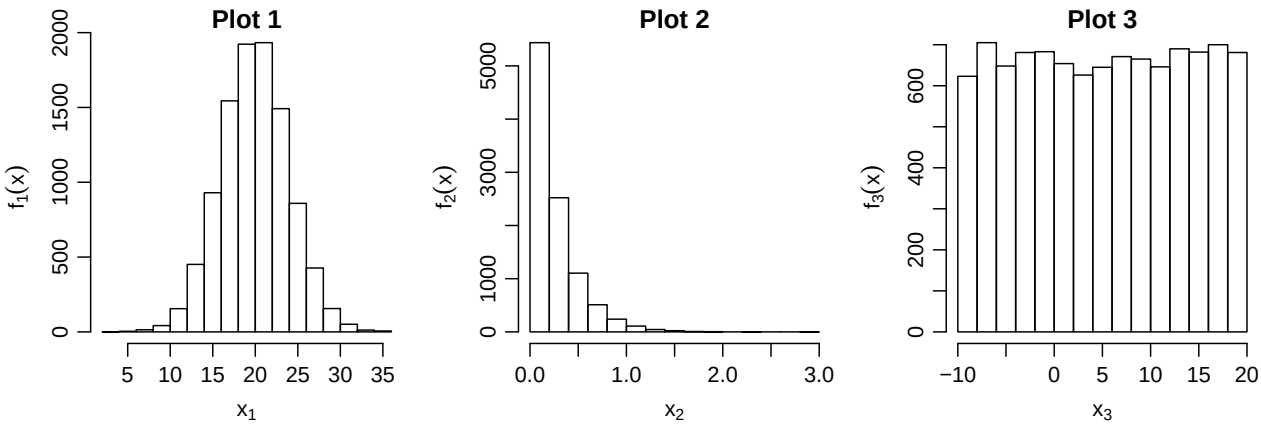
Fortsæt på side 14

Opgave VI

Denne opgave omhandler stokastiske variable og simulering.

Spørgsmål VI.1 (16)

Følgende tre histogrammer er af simuleringer af $n = 10000$ observationer fra tre fordelinger:



Hvilke af følgende fordelinger er de simulerede (rækkefølgen og parameterværdierne skal også være korrekt)?

1 ☐ Plot 1: $X_1 \sim N(10, 4^2)$, Plot 2: $X_2 \sim \text{Exp}(4)$ og Plot 3: $X_3 \sim U(5, 20)$

2 ☐ Plot 1: $X_1 \sim U(5, 20)$, Plot 2: $X_2 \sim \text{Exp}(1)$ og Plot 3: $X_3 \sim N(20, 4^2)$

3 ☐ Plot 1: $X_1 \sim U(-10, 20)$, Plot 2: $X_2 \sim \text{Exp}(1)$ og Plot 3: $X_3 \sim N(20, 4^2)$

4 ☐ Plot 1: $X_1 \sim N(20, 4^2)$, Plot 2: $X_2 \sim \text{Exp}(4)$ og Plot 3: $X_3 \sim U(-10, 20)$

5 ☐ Plot 1: $X_2 \sim \text{Exp}(4)$, Plot 2: $X_1 \sim N(10, 4^2)$ og Plot 3: $X_3 \sim U(5, 20)$

Fortsæt på side 15

Spørgsmål VI.2 (17)

Lad de stokastiske variable $X_i \sim N(2, 4^2)$ for $i = 1, \ldots, 20$ være i.i.d. og definer derefter følgende stokastiske variable som funktion af disse

$$\bar{X} = \frac{1}{20} \sum_{i=1}^{20} X_i,$$

$$S = \sqrt{\frac{1}{19} \sum_{i=1}^{20} (X_i - \bar{X})^2},$$

$$Y = \frac{\bar{X} - 2}{S}.$$

Hvilken fordeling følger den stokastiske variabel Y?

1 ☐ Y følger normalfordelingen $N(0, 1^2)$

2 ☐ Y følger normalfordelingen $N(0, 4^2)$

3 ☐ Y følger χ^2-fordelingen med 20 frihedsgrader

4 ☐ Y følger t-fordelingen med 20 frihedsgrader

5 ☐ Y følger t-fordelingen med 19 frihedsgrader

Fortsæt på side 16

Opgave VII

En motionsløber ønsker at måle effekten af sin træning. Til dette formål har han målt sammenhængende værdier af gennemsnitspuls (slag per minut), uger i træningsprogrammet og hastighed (km/h), for en bestemt strækning han løber ofte.

Motionisten har besluttet at måle effekten ved at undersøge om den gennemsnitlige hastighed stiger i løbet af tiden (ugerne).

Dataindlæsningen i R er:

```
week <- c(1, 1, 1, 3, 3, 4, 5, 5, 5, 5, 6, 6, 7, 7, 8, 9, 9, 10, 12, 12, 13, 13,
          15, 15, 15, 16, 16)

pulse <- c(137.6, 140.1, 143.0, 148.6, 135.6, 139.0, 155.8, 135.0, 149.0, 133.0,
           135.3, 139.8, 137.2, 137.9, 136.8, 134.6, 152.3, 131.9, 137.2, 160.3,
           130.9, 130.9, 131.8, 131.4, 135.6, 138.6, 136.3)

speed <- c(10.01, 10.02, 10.39, 11.86,  9.65, 10.40, 12.60,  9.80, 11.52,  9.59,
           10.26, 10.42, 10.05, 10.48, 10.03, 10.29, 12.22, 10.27, 10.80, 13.79,
           10.40,  9.49, 10.09, 10.34, 11.18, 11.33, 11.34)
```

Indledningvis estimeres parametre i modellen

$$\text{speed}_i = \beta_0 + \beta_1 \text{week}_i + \epsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

Det antages desuden at ε_i'erne er uafhængige.

Resultatet af estimationen i R er:

```
summary(lm(speed~week))

##
## Call:
## lm(formula = speed ~ week)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3909 -0.6058 -0.3407  0.2741  2.9492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.36053    0.38514  26.901   <2e-16 ***
## week         0.04003    0.04046   0.989    0.332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 25 degrees of freedom
## Multiple R-squared:  0.03766,Adjusted R-squared:  -0.0008306
## F-statistic: 0.9784 on 1 and 25 DF,  p-value: 0.3321
```

Fortsæt på side 17

Spørgsmål VII.1 (18)

På signifikansniveau $\alpha = 0.05$ hvad er konklusionen i forhold til øget hastighed (både argument og konklusion skal være korrekt)?

1 ☐ Der kan ikke påvises en signifikant effekt af træningen, da $0.0377 < 0.05$

2 ☐ Der er en signifikant effekt af træningen, da $26.9 > t_{0.975}$ (hvor frihedsgrader for t-fordelingen er 25)

3 ☐ Der kan ikke påvises en signifikant effekt af træningen, da $0.332 > 0.05$

4 ☐ Der er en signifikant effekt af træningen, da $0.04 < 0.05$

5 ☐ Der er en signifikant effekt af træningen, da $0.989 > 0.95$

Spørgsmål VII.2 (19)

Med modellen ovenfor hvad er da 95%-konfidensintervallet for β_0?

1 ☐ $10.36 \pm 1.008 \cdot 1.96$

2 ☐ $10.36 \pm \frac{0.385}{\sqrt{25}} \cdot 2.06$

3 ☐ $0.040 \pm 0.0405 \cdot 1.96$

4 ☐ $0.040 \pm \frac{0.989}{25} \cdot 1.96$

5 ☐ $10.36 \pm 0.385 \cdot 2.06$

Fortsæt på side 18

Motionsløberen beslutter nu at undersøge en model for sammenhængen mellem puls og hastighed (og ignorere uger).

Resultatet af estimationen i R er:

```
 summary(lm(speed ~ pulse))

##
## Call:
## lm(formula = speed ~ pulse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78371 -0.40486 -0.00015  0.35978  0.96661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.06134    1.82173  -2.778   0.0102 *
## pulse        0.11324    0.01308   8.659 5.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5139 on 25 degrees of freedom
## Multiple R-squared:  0.7499,Adjusted R-squared:  0.7399
## F-statistic: 74.98 on 1 and 25 DF,  p-value: 5.388e-09
```

Baseret på ovenstående model ønsker motionsløberen et usikkerhedsinterval for en ny løbetur. Han bruger som forudsætning, at han kan holde en gennemsnitspuls på 160 slag i minuttet.

Som hjælp til opgaven har han udregnet følgende tal:

```
length(week)

## [1] 27

c(mean(week), var(week))

## [1]  8.222222 23.871795

c(mean(pulse), var(pulse))

## [1] 139.09259  59.38225

c(mean(speed), var(speed))

## [1] 10.689630  1.015396
```

Fortsæt på side 19

Spørgsmål VII.3 (20)

Hvad er 95% prædiktionsintervallet for hastigheden?

1 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{27 \cdot 59.38}}$

2 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{1 + \frac{1}{27} + \frac{(139.1-160)^2}{26 \cdot 1.02}}$

3 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{1 + \frac{1}{27} + \frac{(139.1-160)^2}{26 \cdot 59.38}}$

4 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51^2 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{59.38}}$

5 ☐ $-5.06 + 0.113 \cdot 160 \pm 2.06 \cdot 0.51 \cdot \sqrt{\frac{1}{27} + \frac{(139.1-160)^2}{59.38^2}}$

Motionsløberen beslutter nu at estimere en multiple regressionsmodel, der både indeholder uger og puls:

$$\text{speed}_i = \beta_0 + \beta_1 \cdot \text{week}_i + \beta_2 \cdot \text{pulse}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

Resultatet fra R bliver:

```
summary(lm(speed ~ week + pulse))

##
## Call:
## lm(formula = speed ~ week + pulse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59154 -0.13508 -0.00055  0.15562  0.43438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.033531   0.945936  -8.493 1.08e-08 ***
## week         0.094014   0.010414   9.028 3.48e-09 ***
## pulse        0.129052   0.006603  19.545 3.02e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2501 on 24 degrees of freedom
## Multiple R-squared:  0.9431,Adjusted R-squared:  0.9384
## F-statistic:   199 on 2 and 24 DF,  p-value: 1.148e-15
```

Fortsæt på side 20

Spørgsmål VII.4 (21)

På niveau $\alpha = 0.05$ hvilket af følgende udsagn er korrekt (både konklusion og argument skal være korrekt)?

1 ☐ Hverken effekten af uger eller effekten af puls er signifikant da $0.094 > 0.05$ og $0.129 > 0.05$

2 ☐ Både effekten af uger og effekten af puls er signifikant da $0.01 < 0.05$ og $0.0066 < 0.05$

3 ☐ Da $0.094 < 0.129$ er effekten af uger signifikant, mens effekten af puls ikke er signifikant

4 ☐ Både effekten af uger og effekten af puls er signifikant da $3.5 \cdot 10^{-9} < 0.05$ og $3.0 \cdot 10^{-16} < 0.05$

5 ☐ Hverken effekten af uger eller effekten af puls er signifikant da $3.5 \cdot 10^{-9} < 0.05$ og $3.0 \cdot 10^{-16} < 0.05$

Spørgsmål VII.5 (22)

Hvilket udsagn om fortolkning af modellen er korrekt?

1 ☐ Når der lægges 0.094 til uger stiger pulsen med 0.129

2 ☐ For en given puls stiger den forventede hastighed med 0.094km/h om ugen

3 ☐ Da $-8.03 < 0$ og hastigheden skal være positiv er modellen meningsløs

4 ☐ Da forklaringsgraden er ca. 0.25 har modellen forklaret omkring 3/4 af variationen

5 ☐ Da alle parametre er signifikante kan man konstatere at alle modelantagelser er opfyldt

Fortsæt på side 21

Opgave VIII

En person har 2 gange vurderet skarpheden (`Sharpness`) for hver af 12 forskellige setups (`Treat`) af billeder på computerskærme, dvs. 24 observationer af skarphed i alt fordelt på 12 setups. Skalaen er en kontinuert skala fra 0 til 15, i praksis udført ved at markere værdien på en linie.

Resultatet af en sædvanlig variansanalyse af disse data gav følgende R-output, hvor en del af værdierne dog er erstattet af bogstaverne A,. . . ,F:

```
Analysis of Variance Table

Response: Sharpness
          Df Sum Sq Mean Sq F value Pr(>F)
Treat      A   93.7     C       E      F
Residuals  B   51.9     D
```

Spørgsmål VIII.1 (23)

Hvad er værdierne for A og B?

1 ☐ A=11 og B=12

2 ☐ A=1 og B=22

3 ☐ A=93.7/51.9 og B= 51.9/22

4 ☐ A=12 og B=24

5 ☐ A=11 og B=23

Fortsæt på side 22

Spørgsmål VIII.2 (24)

En anden person har på tilsvarende vis et antal gange vurderet skarpheden (`Sharpness`) for hvert af nogle forskellige setups (`Treat`) af billeder på computerskærme. Resultatet af en sædvanlig variansanalyse af disse data gav følgende R-output, hvor en del af værdierne dog er erstattet af bogstaverne G,. . . ,J:

```
Analysis of Variance Table

Response: Sharpness
          Df Sum Sq Mean Sq F value Pr(>F)
Treat      7  111.5     G       I      J
Residuals 32   88.4     H
```

Hvad er værdien for J?

1 ☐ 1.261

2 ☐ 5.766

3 ☐ 0.0002

4 ☐ 0.2188

5 ☐ 0.3002

Fortsæt på side 23

Opgave IX

Otte eksperter har hver vurderet skarpheden (`Sharpness`) for hver af 12 forskellige setups (`Treat`) af billeder på computerskærme, dvs. 96 observationer af skarphed i alt fordelt på 12 setups. Skalaen er en kontinuert skala fra 0 til 15, i praksis udført ved at markere værdien på en linie:

	Setup 1	Setup 2	Setup 3	Setup 4	Setup 5	Setup 6	Setup 7	Setup 8	Setup 9	Setup 10	Setup 11	Setup 12
Person 1	9.30	4.70	6.60	8.80	5.90	7.20	7.60	5.50	8.10	8.20	6.40	7.40
Person 2	10.20	7.00	8.80	10.70	9.80	7.00	9.20	9.60	8.00	11.80	8.90	10.20
Person 3	11.50	9.50	8.00	12.90	10.00	8.20	11.50	6.40	8.60	11.20	7.70	11.00
Person 4	11.90	6.60	8.20	12.70	5.40	9.00	4.90	8.10	10.10	12.90	8.20	8.70
Person 5	10.70	4.20	5.40	11.40	8.30	7.10	6.80	3.80	9.60	8.60	3.80	10.80
Person 6	10.90	9.10	7.10	11.40	8.60	5.90	8.50	10.50	6.40	11.70	9.50	7.10
Person 7	8.50	5.00	6.30	10.80	6.80	4.60	4.70	8.80	6.70	10.00	5.50	7.50
Person 8	12.60	8.90	10.70	13.50	11.40	8.90	9.50	8.60	7.40	13.50	8.70	9.80

Resultatet af en sædvanlig tosidet variansanalyse af disse data gav følgende R-output:

```
## Analysis of Variance Table
##
## Response: Sharpness
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Person     7 122.42 17.4881  8.4596 1.212e-07 ***
## Setup     11 224.28 20.3894  9.8630 6.864e-11 ***
## Residuals 77 159.18  2.0673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Spørgsmål IX.1 (25)

Hvad er konklusionerne af de sædvanlige hypotesetests for en sådan analyse? (Både konklusioner og argumenter skal være korrekte)

1 ☐ Der er forskel på middelskarpheden for personer, men ikke for setups

2 ☐ Der er forskel på middelskarpheden for setups, men ikke for personer

3 ☐ Der er forskel på varianserne for personer men ikke for setups

4 ☐ Der er forskel på middelskarpheden for såvel setups som for personer

5 ☐ Der er forskel på varianserne for setups men ikke for personer

Fortsæt på side 24

Spørgsmål IX.2 (26)

Hvilken sandsynlighedsfordeling har været anvendt for at finde p-værdien angivet ud for `Setup` i outputtet?

1 ☐ z-fordelingen (= standard normalfordeling)

2 ☐ t-fordeling med 159 frihedsgrader

3 ☐ χ^2-fordeling med 159 frihedsgrader

4 ☐ F-fordeling med frihedsgraderne 7 og 11

5 ☐ F-fordeling med frihedsgraderne 11 og 77

Fortsæt på side 25

Opgave X

I en spørgeskemaundersøgelse under en Introduktion til Statistik undervisning blev de studerende spurgt om forskellige emner. Denne opgave vil omhandle analysen af svarene fra et af spørgsmålene. Der var 32 respondenter ialt.

Det stillede spørgsmål var: "Er du bekymret for, at vi ikke gør nok for at stoppe klimaforandringerne?"og hertil svarede de studerende følgende:

Svar	Antal
Ja	27
Nej	5

Spørgsmål X.1 (27)

Ved at benytte "Plus 2" korrektionen ved beregningen af det sædvanlige 95% konfidensinterval for andelen af studerende, som er bekymrede for klimaet (dvs. svarer ja), får man:

1 ☐ $0.844 \pm 2.04 \cdot \frac{0.844}{36} = [0.796, 0.892]$

2 ☐ $0.806 \pm 1.69 \cdot \frac{0.806}{36} = [0.768, 0.844]$

3 ☐ $0.806 \pm 1.96\sqrt{\frac{0.806 \cdot 0.194}{36}} = [0.677, 0.935]$

4 ☐ $0.844 \pm 1.69\sqrt{\frac{0.844 \cdot 0.156}{32}} = [0.736, 0.952]$

5 ☐ $0.844 \pm 1.96\sqrt{\frac{0.844 \cdot 0.156}{36}} = [0.725, 0.963]$

Fortsæt på side 26

Opgave XI

I en spørgeskemaundersøgelse er 114 respondenter blevet adspurgt om deres trafikpræferencer. To spørgsmål som blev stillet var med følgende tre identiske svarmuligheder: "bil", "cykel" og "tog eller bus".

De to spørgsmål var: "Hvis du har 10 km til DTU fra dit hjem, hvilken transportform vil du så foretrække om sommeren (det tager cirka lige lang tid)?", samt: "Hvis du har 10 km til DTU fra dit hjem, hvilken transportform vil du så foretrække om vinteren (det tager cirka lige lang tid)?".

Følgende fordeling af svar blev observeret:

		Vinter		
		Bil	Cykel	Tog el. Bus
Sommer	Bil	27	2	4
	Cykel	20	22	11
	Tog el. bus	13	3	12

Følgende er kørt i R:

```
## The data table
tbl <- matrix(c(27, 20, 13, 2, 22, 3, 4, 11, 12), nrow = 3)
rownames(tbl) <- c("Bil", "Cykel", "TogElBus")
colnames(tbl) <- c("Bil", "Cykel", "TogElBus")
tbl

##           Bil Cykel TogElBus
## Bil        27     2        4
## Cykel      20    22       11
## TogElBus   13     3       12

## Row sums (distribution for summer)
margin.table(tbl, 1)

##      Bil    Cykel TogElBus
##       33       53       28

## Column sums (distribution for winter)
margin.table(tbl, 2)

##      Bil    Cykel TogElBus
##       60       27       27

## Chi2-test
chisq.test(tbl, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 27.608, df = 4, p-value = 1.498e-05
```

Fortsæt på side 27

Spørgsmål XI.1 (28)

Hvad bliver det forventede antal, der foretrækker at cykle om sommeren og køre i bil om vinteren, under nulhypotesen: der er uafhængighed mellem trafikpræferencer om sommeren og om vinteren i den adspurgte population?

1 ☐ $e_{11} = 114 \cdot \frac{33}{114} \cdot \frac{60}{114} = 17.37$

2 ☐ $e_{23} = 114 \cdot \frac{53}{114} \cdot \frac{27}{114} = 12.55$

3 ☐ $e_{12} = 114 \cdot \frac{33}{114} \cdot \frac{27}{114} = 7.816$

4 ☐ $e_{33} = 114 \cdot \frac{28}{114} \cdot \frac{27}{114} = 6.877$

5 ☐ $e_{21} = 114 \cdot \frac{53}{114} \cdot \frac{60}{114} = 27.89$

Spørgsmål XI.2 (29)

Hvad bliver konklusionen på signifikansniveau 1% for den udførte test om afhængighed mellem trafikpræferencerne om sommer og vinter (både konklusion og argument skal være korrekt)?

1 ☐ Der er ikke påvist afhængighed mellem trafikpræferencer om sommeren og vinteren da p-værdien > 0.01

2 ☐ Der er påvist afhængighed mellem trafikpræferencer om sommeren og vinteren da p-værdien < 0.01

3 ☐ Der er ikke påvist afhængighed mellem trafikpræferencer om sommeren og vinteren da p-værdien < 0.01

4 ☐ Der er påvist afhængighed mellem trafikpræferencer om sommeren og vinteren da p-værdien > 0.01

5 ☐ Dette spørgsmål kan ikke besvares med de givne oplysninger

Fortsæt på side 28

Spørgsmål XI.3 (30)

Der er 60 ud af 114 som foretrækker at køre bil om vinteren. Ville følgende nulhypotese

$$H_0 : p_{\text{bil,vinter}} = 50\%,$$

blive afvist på 5% signifikansniveau med den normalt anvendte test (både konklusion og p-værdien skal være korrekt)?

1 ☐ Ja, da p-værdien er $0.024 < 0.05$

2 ☐ Nej, da p-værdien er $0.57 > 0.05$

3 ☐ Nej, da p-værdien er $0.40 > 0.05$

4 ☐ Ja, da p-værdien er $0.089 > 0.05$

5 ☐ Nej, da p-værdien er $0.21 > 0.05$

SÆTTET ER SLUT. GOD SOMMER!