Skriftlig prøve: 15. august 2018

Kursus navn og nr.: **Introduktion til Statistik (02403)**

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 14 opgaver. For at besvare spørgsmålene skal du udfylde "multiple choice" svararket (6 seperate 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 gøres 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	III.1	III.2	IV.1	V.1	VI.1	VI.2	VI.3
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

Opgave	VII.1	VII.2	VIII.1	VIII.2	IX.1	IX.2	IX.3	X.1	X.2	X.3
Spørgsmål	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Svar										

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

Sættet 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

Frank arbejder i en butik, som han først må lukke, når alle kunder er gået. Tiden en kunde bruger i butikken er eksponentialfordelt med en middelværdi på 20 minutter. Der er lige ankommet en kunde og Frank har en date med sin kæreste om 25 minutter. Det tager ham 5 minutter at lukke forretningen og 5 minutter at gå til restauranten, hvor de skal mødes.

Spørgsmål I.1 (1)

Hvis det antages, at der ikke kommer flere kunder til butikken, hvad er så sandsynligheden for, at Frank kommer for sent til sin date?

1 ☐ 0.29

2 ☐ 0.37

3 ☐ 0.47

4 ☐ 0.61

5 ☐ 1

Fortsæt på side 3

Opgave II

Antag at X og Y er uafhængige og normalfordelte stokastiske variable, hhv. $X \sim N(3,2)$ og $Y \sim N(4,1)$.

Spørgsmål II.1 (2)

Hvad er da variansen af $Z = 2X - 3Y$?

1 ☐ 1

2 ☐ 5

3 ☐ 11

4 ☐ 13

5 ☐ 17

Spørgsmål II.2 (3)

Antag nu at kovariansen mellem X og Y er 1.

Hvad er variansen af $Z = 2X - 3Y$?

1 ☐ 1

2 ☐ 5

3 ☐ 11

4 ☐ 13

5 ☐ 17

Fortsæt på side 4

Opgave III

Antag at $X_i \sim N(\mu_1, \sigma_1^2)$ samt at $Y_i \sim N(\mu_2, \sigma_2^2)$. Antag desuden at alle stokastiske variable er uafhængige. Lad desuden S_1^2 og S_2^2 være de sædvanlige variansestimatorer

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$
$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

hvor $\bar{X}$ og $\bar{Y}$ som sædvanlig er gennemsnit af hhv. $(X_1, ..., X_{n_1})$ og $(Y_1, ..., Y_{n_2})$. Antag desuden at $\mu_1 = \mu_2 = 2$, $n_1 = n_2 = 9$ og $\sigma_1^2 = 2\sigma_2^2 = 2$.

Spørgsmål III.1 (4)

Hvad er så middelværdi og varians af den stokastiske variabel $Z = S_1^2/2 + S_2^2/2$?

1 ☐ $E[Z] = 2$, og $V[Z] = \frac{1}{2}$

2 ☐ $E[Z] = 1$, og $V[Z] = \frac{1}{2}$

3 ☐ $E[Z] = 1$, og $V[Z] = \frac{1}{8}$

4 ☐ $E[Z] = 2$, og $V[Z] = \frac{1}{8}$

5 ☐ $E[Z] = \frac{3}{2}$, og $V[Z] = \frac{5}{16}$

Spørgsmål III.2 (5)

Hvilket af følgende udsagn om forholdet mellem S_1^2 og S_2^2 er sandt?

1 ☐ $\frac{1}{2}\frac{S_1^2}{S_2^2} \sim F(8, 8)$

2 ☐ $2\frac{S_1^2}{S_2^2} \sim F(9, 9)$

3 ☐ $\frac{S_1^2}{S_2^2} \sim \chi^2(8)$

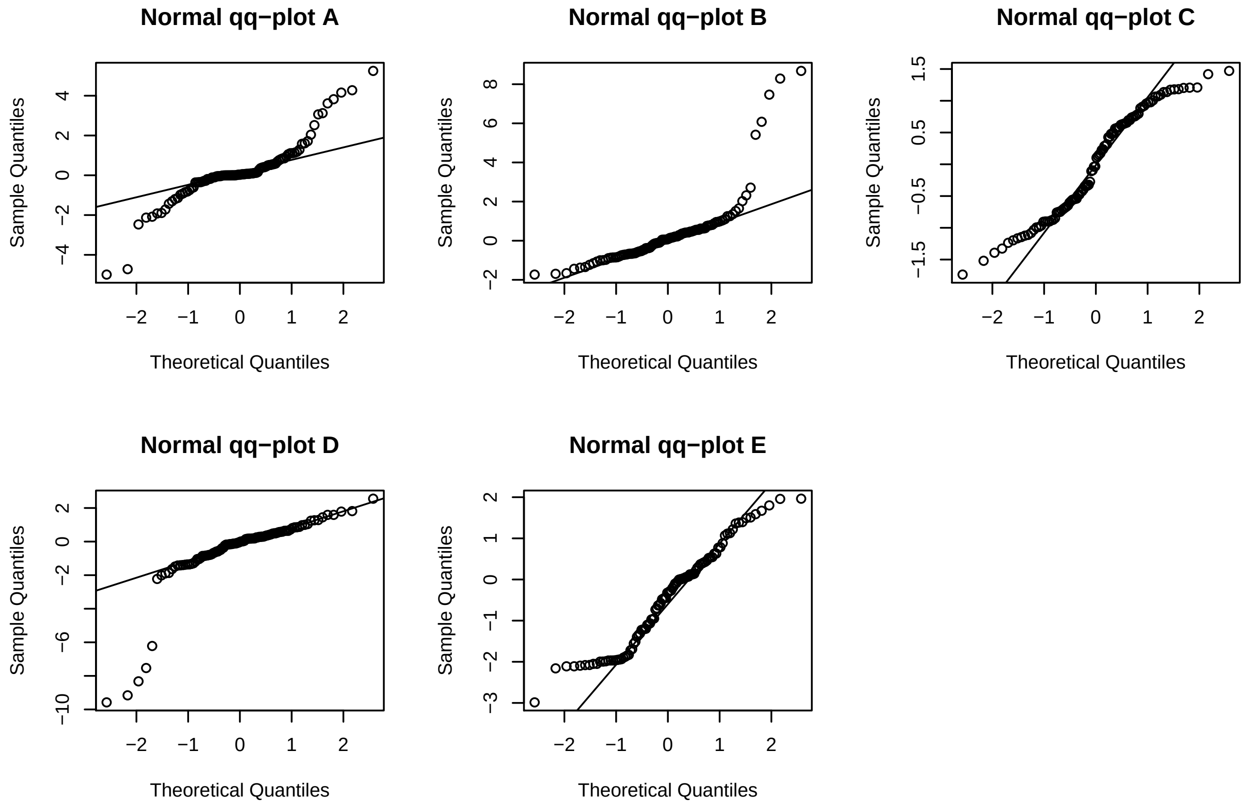
4 ☐ $2\frac{S_1^2}{S_2^2} \sim F(8, 8)$

5 ☐ $\frac{S_1^2}{S_2^2} \sim \chi^2(9)$

Fortsæt på side 5

Opgave IV

Nedenfor er vist fem qq-plots:



Spørgsmål IV.1 (6)

Hvilket af disse fem qq-plots viser, at data indeholder en række outliers med værdier under stikprøvegennemsnittet og ellers kan antages normalfordelt?

1 ☐ A

2 ☐ B

3 ☐ C

4 ☐ D

5 ☐ E

Fortsæt på side 6

Opgave V

Antag at $X_1, ..., X_{100}$ er uafhængige stokastiske variable, der alle er normalfordelte $N(\mu, \sigma^2)$. Antag desuden at de 100 stokastiske variable repræsenterer en stikprøve og at man fra en realisering af denne har beregnet $\bar{x} = 2.3$ og $s^2 = 0.25$.

Spørgsmål V.1 (7)

Hvad er et 95% konfidensinterval for $\exp(\mu)$?

1 ☐ [9.03, 11.01]

2 ☐ [7.31, 12.64]

3 ☐ [1.32, 3.28]

4 ☐ [3.74, 26.57]

5 ☐ [9.97, 12.81]

Fortsæt på side 7

Opgave VI

En stikprøve er tilfældigt udtaget fra en population og følgende analyse er kørt:

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 3.638, df = 19, p-value = 0.00175
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.141765 4.235249
## sample estimates:
## mean of x
##  2.688507
```

Spørgsmål VI.1 (8)

Ville nulhypotesen

$$H_0 : \mu_X = 0$$

blive afvist på signifikansniveau $\alpha = 0.1$ (både konklusion og argument skal være korrekt)?

1 ☐ Nej, da p-værdien for det relevante test er 0.072 og dermed større end $\alpha = 0.1$.

2 ☐ Ja, da p-værdien for det relevante test er 0.072 og dermed større end $\alpha = 0.1$.

3 ☐ Ja, da p-værdien for det relevante test er 0.035 og dermed større end $\alpha = 0.1$.

4 ☐ Nej, da p-værdien for det relevante test er 0.00175 og dermed mindre end $\alpha = 0.1$.

5 ☐ Ja, da p-værdien for det relevante test er 0.00175 og dermed mindre end $\alpha = 0.1$.

Spørgsmål VI.2 (9)

Ville nulhypotesen

$$H_0 : \mu_X = 2$$

blive afvist på signifikansniveau $\alpha = 5\%$ (både konklusion og argument skal være korrekt)?

1 ☐ Nej, da μ_0 ikke er indeholdt i 95% konfidensintervallet.

2 ☐ Nej, da μ_0 er indeholdt i 95% konfidensintervallet.

3 ☐ Ja, da μ_0 ikke er indeholdt i 95% konfidensintervallet.

4 ☐ Ja, da μ_0 er indeholdt i 95% konfidensintervallet.

5 ☐ Kan ikke afgøres med de givne oplysninger.

Spørgsmål VI.3 (10)

Hvor mange observationer består stikprøven af?

1 ☐ $n = 10$

2 ☐ $n = 18$

3 ☐ $n = 19$

4 ☐ $n = 20$

5 ☐ $n = 21$

Fortsæt på side 9

Opgave VII

I et linæert regressionsproblem har man opstillet følgende designmatrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}$$

hvor vi benævner modellens parametre $\boldsymbol{\beta} = [\beta_0, \beta_1]$, modellen er således givet ved $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ med $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

Spørgsmål VII.1 (11)

Hvilket udsagn om den resulterende model er korrekt (uafhængigt af data)?

1 ☐ $\hat{\beta}_0 = \bar{Y}$

2 ☐ $\hat{Y}_3 = \hat{Y}_1 + \hat{Y}_2$

3 ☐ $\hat{\beta}_1 = \bar{Y}$

4 ☐ $\hat{Y}_1 = \hat{Y}_2 + \hat{Y}_3$

5 ☐ $\hat{Y}_2 = \hat{Y}_1$

Spørgsmål VII.2 (12)

Hvis residualviariansen er kendt og lig med 1 (dvs. $\sigma^2 = 1$), hvad er så kovariansen mellem estimaterne $\hat{\beta}_0$ og $\hat{\beta}_1$?

1 ☐ 2

2 ☐ -0.8

3 ☐ 0.8

4 ☐ 1

5 ☐ -1

Fortsæt på side 10

Opgave VIII

Hekseprocesserne i Salem var en serie høringer og retssager i den britiske koloni Massachusetts i Amerika, mellem februar 1692 og maj 1693. Under disse processer blev 185 personer, 141 kvinder og 44 mænd, tiltalt for at være hekse. 19 af de tiltalte, 14 kvinder og 5 mænd, blev hængt.

	Tiltalte mænd	Tiltalte kvinder	Tiltalte i alt
Hængt	5	14	19
Ikke hængt	39	127	166
I alt	44	141	185

Vi ønsker at teste hypotesen

$$H_0 : p_1 = p_2,\ H_1 : p_1 \neq p_2$$

hvor p_1 er andelen af tiltalte kvinder der blev hængt, og p_2 er andelen af tiltalte mænd der blev hængt.

Spørgsmål VIII.1 (13)

Hvad bliver den sædvanlige teststørrelse, når man ønsker at teste den overstående hypotese?

1 ☐ $z_{\text{obs}} = (\frac{5}{44} + \frac{14}{141})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{44} - \frac{1}{141})} = 5.61$

2 ☐ $z_{\text{obs}} = (\frac{14}{19} + \frac{127}{166})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{19} - \frac{1}{166})} = 22.9$

3 ☐ $z_{\text{obs}} = (\frac{5}{44} - \frac{14}{141})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{44} + \frac{1}{141})} = 0.274$

4 ☐ $z_{\text{obs}} = (\frac{127}{166} - \frac{14}{19})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{19} + \frac{1}{166})} = 0.384$

5 ☐ $z_{\text{obs}} = (\frac{127}{166} - \frac{14}{19})/\sqrt{\frac{19}{185} \cdot (1 - \frac{19}{185}) \cdot (\frac{1}{19} - \frac{1}{166})} = 0.431$

Spørgsmål VIII.2 (14)

Under antagelse af, at nulhypotesen er sand, hvad er da, det forventede antal af tiltalte kvinder, der blev hængt?

1 ☐ $\frac{19 \cdot 166}{185} = 17.0$

2 ☐ $\frac{5 \cdot 141}{44} = 16.0$

3 ☐ $\frac{14 \cdot 141}{141} = 14.0$

4 ☐ $\frac{14 \cdot 141}{185} = 10.7$

5 ☐ $\frac{19 \cdot 141}{185} = 14.5$

Fortsæt på side 11

Opgave IX

En producent af rottegift ønsker, at afprøve hvilken af 4 typer rottegift, rotter er mest tilbøjelige til at spise. De 4 typer rottegift har henholdsvis neutral smag, vaniljesmør-smag, roastbeef-smag og brød-smag.

Producenten planlægger, at teste de 4 typer rottegift på et antal rotter, ved at give hver af rotterne 4 smagsprøver rottegift med hver sin smag, og for hver rotte at registrere hvilken af de 4 typer rottegift, rotten først spiste.

Spørgsmål IX.1 (15)

Hvor mange rotter skal producenten bruge, hvis hun antager at populationens andel er 0.5, og ønsker at bestemme et 95% konfidensinterval med en middelbredde på 4% for andelen af rotter, der foretrækker rottegift med neutral smag?

1 ☐ 1691

2 ☐ 601

3 ☐ 9604

4 ☐ 6764

5 ☐ 2401

Spørgsmål IX.2 (16)

Producenten beslutter sig for at give smagsprøver til 1000 rotter. Antallet af rotter, der vælger hver af de 4 typer rottegift, er angivet i nedenstående tabel:

Neutral	Vaniljesmør	Roastbeef	Brød	I alt
265	224	269	242	1000

Angiv et 95% konfidensinterval for andelen af rotter, der foretrækker rottegift med neutral smag.

1 ☐ $0.265 \pm 1.96 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{265}} = [0.212, 0.318]$

2 ☐ $0.265 \pm 1.6449 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{265}} = [0.220, 0.310]$

3 ☐ $0.265 \pm 1.96 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{735}} = [0.233, 0.297]$

4 ☐ $0.265 \pm 1.6449 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{735}} = [0.238, 0.297]$

5 ☐ $0.265 \pm 1.96 \cdot \sqrt{\frac{0.265 \cdot (1-0.265)}{1000}} = [0.238, 0.292]$

Spørgsmål IX.3 (17)

Producenten ændrer på smagen af rottegift-typen med vaniljesmør-smag, og gentager derefter forsøget. Denne gang bliver udfaldet af forsøget som følger:

Neutral	Vaniljesmør	Roastbeef	Brød	I alt
219	284	263	234	1000

Du skal undersøge, om der er en signifikant forskel på andelen af rotter, der foretrækker rottegift med vaniljesmør-smag i de to forsøg, svarende til nulhypotesen

$$H_0 : p_1 = p_2$$

hvor p_1 er andelen af rotter der foretrækker rottegift med vaniljesmør-smag i det første forsøg, og p_2 er andelen af rotter der foretrækker rottegift med vaniljesmør-smag i det andet forsøg.

Angiv hvilket af følgende kald, der i R kan bruges til dette:

1 ☐ `prop.test(x=c(224,284), n=c(2000,2000), correct=FALSE)`

2 ☐ `prop.test(x=c(0.224,0.284), n=c(2000,2000), correct=FALSE)`

3 ☐ `prop.test(x=60, n=1000, correct=FALSE)`

4 ☐ `prop.test(x=c(224,284), n=c(1000,1000), correct=FALSE)`

5 ☐ `prop.test(x=60, n=2000, correct=FALSE)`

Fortsæt på side 14

Opgave X

Der er blevet udført et forsøg med 33 cylinderformede beholdere, alle fyldt til randen med et pulvermateriale. Blandt andet blev beholdernes (indvendige) radius og højde (i cm) målt og vægten af indholdet (i g) blev bestemt. Målingerne af beholdernes radier er gemt i vektoren `radius` i R, mens vektoren `ratio` indeholder målinger af forholdet mellem vægten af hver beholders indhold og beholderens højde (i g/cm).

Følgende kode er derefter kørt i R:

```
radius2 <- radius^2
model1 <- lm(ratio ~ radius + radius2)
summary(model1)

##
## Call:
## lm(formula = ratio ~ radius + radius2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -52.269 -20.380  -2.684  15.071  64.217
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.9362    38.3505  -0.129    0.898
## radius       -0.4221     6.7042  -0.063    0.950
## radius2       2.8765     0.2645  10.876 6.25e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.19 on 30 degrees of freedom
## Multiple R-squared:  0.993,Adjusted R-squared:  0.9926
## F-statistic:  2141 on 2 and 30 DF,  p-value: < 2.2e-16
```

Spørgsmål X.1 (18)

Hvilket af følgende udsagn er det eneste, som beskriver den statistiske model svarende til `model1` korrekt?

1 ☐ Der er tale om en simpel lineær regressionsmodel. Modellen har to afhængige variable, hhv. vægt og højde. Radius er den forklarende variabel.

2 ☐ Der er tale om en simpel lineær regressionsmodel. Forholdet mellem vægt og højde udgør den afhængige variabel, mens radius er den forklarende variabel.

3 ☐ Der er tale om en multipel lineær regressionsmodel. Radius og radius kvadreret er de afhængige variable, mens forholdet mellem vægt og højde udgør den forklarende variabel.

4 ☐ Der er tale om en multipel lineær regressionsmodel. Forholdet mellem vægt og højde udgør den afhængige variabel. Der er to forklarende variable, hhv. radius og radius kvadreret.

5 ☐ Der er tale om en multipel lineær regressionsmodel. Modellen har to afhængige variable, hhv. vægt og højde, og to forklarende variable, hhv. radius og radius kvadreret.

Spørgsmål X.2 (19)

Angiv, med udgangspunkt i `model1`, et estimat for vægten af pulvermaterialet i en fyldt beholder, som har både radius og højde 10 cm.

1 ☐ Forholdet mellem vægt og højde estimeres til 271.5 g/cm, så et estimat for vægten bliver 27.15 g.

2 ☐ Forholdet mellem vægt og højde estimeres til 278.5 g/cm, så et estimat for vægten bliver 2785 g.

3 ☐ Forholdet mellem vægt og højde estimeres til 271.5 g/cm, så et estimat for vægten bliver 2715 g.

4 ☐ Forholdet mellem vægt og højde estimeres til 278.5 g/cm, så et estimat for vægten bliver 27.85 g.

5 ☐ Vægten af pulvermaterialet estimeres til 278.5 g.

Spørgsmål X.3 (20)

Sæt signifikansniveauet til $\alpha = 0.05$. Man ønsker at undersøge, om modellen givet ved `model1` kan reduceres til en simpel lineær regressionsmodel, hvor radius udelukkende indgår kvadratisk. Hvad bliver konklusionen ud fra ovenstående R-output? (Både argument og konklusion skal være rigtige).

1 ☐ Den relevante p-værdi er 0.898 og dermed større end signifikansniveauet. Modellen kan ikke reduceres som ønsket.

2 ☐ Den relevante p-værdi er 0.950 og dermed større end signifikansniveauet. Modellen kan ikke reduceres som ønsket.

3 ☐ Den relevante p-værdi er 0.950 og dermed større end signifikansniveauet. Modellen kan reduceres som ønsket.

4 ☐ Den relevante p-værdi er $6.25 \cdot 10^{-12}$ og dermed mindre end signifikansniveauet. Modellen kan ikke reduceres som ønsket.

5 ☐ Den relevante p-værdi er $6.25 \cdot 10^{-12}$ og dermed mindre end signifikansniveauet. Modellen kan reduceres som ønsket.

Spørgsmål X.4 (21)

Følgende kode er nu kørt i R (bemærk at nogle dele af outputtet er erstattet af `x`):

```
model2 <- lm(ratio ~ radius2)
summary(model2)

##
## Call:
## lm(formula = ratio ~ radius2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.639 -20.469  -2.871  14.762  63.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.277      9.264       x        x x
## radius2        2.860      0.043       x        x x
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.7 on x degrees of freedom
## Multiple R-squared:  0.993,Adjusted R-squared:  0.9928
## F-statistic:     x on 1 and  x DF,  p-value:         x
```

Der tages udgangspunkt i den statistiske model givet ved `model2`, som har formen

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,.$$

Her er Y_i forholdet mellem vægt og højde, x_i er radius kvadreret og ε_i, $i = 1, \ldots, 33$, er uafhængige og identisk $N(0, \sigma^2)$-fordelte.

Angiv et 95% konfidensinterval for modellens hældningsparameter.

1 ☐ $2.860 \pm 1.696 \cdot 0.043 = [2.79, 2.93]$

2 ☐ $2.860 \pm 2.040 \cdot 29.7 = [-57.73, 63.45]$

3 ☐ $2.860 \pm 1.696 \cdot \sqrt{29.7} = [-6.38, 12.10]$

4 ☐ $2.860 \pm 2.040 \cdot \sqrt{29.7} = [-8.26, 13.98]$

5 ☐ $2.860 \pm 2.040 \cdot 0.043 = [2.77, 2.95]$

Spørgsmål X.5 (22)

Tag igen udgangspunkt i modellen givet ved `model2`, og benyt R-outputtet fra sidste spørgsmål i det følgende. Sæt signifikansniveauet til $\alpha = 0.05$. Test (nul)hypotesen om at modellens intercept kan sættes til 0. Hvad kan konkluderes? (Både konklusion og argument skal være rigtige).

1 ☐ Modellens intercept kan ikke sættes til 0, da den relevante p-værdi er 0.78, og hypotesen dermed forkastes.

2 ☐ Modellens intercept kan ikke sættes til 0, da den relevante p-værdi er $2 \cdot 10^{-7}$, og hypotesen dermed forkastes.

3 ☐ Modellens intercept kan sættes til 0, da den relevante p-værdi er 0.44, og hypotesen dermed accepteres.

4 ☐ Modellens intercept kan sættes til 0, da den relevante p-værdi er 0.78, og hypotesen dermed accepteres.

5 ☐ Modellens intercept kan ikke sættes til 0, da den relevante p-værdi er 0.44, og hypotesen dermed forkastes.

Fortsæt på side 18

Opgave XI

Spørgsmål XI.1 (23)

Når man skal bestemme andelen af bakterier, som er resistente overfor et antibiotikum, kan man tage en prøve af bakterier, f.eks. fra en gris, og udplade det samme volumen bakterier på agarplader med og uden antibiotika. Volumenet er valgt så man forventer at tælle 20 på den plade, som er uden antibiotika.

Man estimerer andelen af resistente bakterier i prøven til det observerede forhold:

$$\text{"andel"} = \frac{\text{"Antal på plade med antibiotika"}}{\text{"Antal på plade uden antibiotika"}}$$

Dvs. hvis antibiotikaet ikke dræber alle bakterierne, så bliver "andel" over 0% og dermed er der resistens.

Man vil undersøge fordelingen af "andel", hvis den sande andel resistente bakterier er 75%. Det kan antages at udfaldet på en plade er Poissonfordelt.

Der er kørt følgende kode:

```
k <- 10000
noAntibiotica <- rpois(k, lambda = 20)
withAntibiotica1 <- rpois(k, lambda = 15)
withAntibiotica2 <- rpois(k, lambda = noAntibiotica*0.75)

quantile(noAntibiotica, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
##    12    13    28    29

quantile(withAntibiotica1, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
##     8     9    22    23

quantile(withAntibiotica1 / withAntibiotica2, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
## 0.417 0.500 2.286 2.750

quantile(withAntibiotica1 / noAntibiotica, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
## 0.353 0.407 1.333 1.500

quantile(withAntibiotica2 / noAntibiotica, c(0.025, 0.05, 0.95, 0.975))

##  2.5%    5%   95% 97.5%
## 0.385 0.435 1.100 1.167
```

Den simulerede værdi af 2.5% og 97.5% fraktilerne i fordelingen af ”andel” findes til

1 ☐ $q_{0.025} = \frac{1}{12}, q_{0.975} = 1 - \frac{1}{29}$

2 ☐ $q_{0.025} = 0.417, q_{0.975} = 2.750$

3 ☐ $q_{0.025} = 0.353, q_{0.975} = 1.500$

4 ☐ $q_{0.025} = 0.385, q_{0.975} = 1.167$

5 ☐ $q_{0.025} = 0.435, q_{0.975} = 1.100$

Spørgsmål XI.2 (24)

Der udtages en prøve fra en gris og der tælles 16 overlevende bakterier på pladen uden antibiotika og 22 på pladen med antibiotika.

Hvilken af følgende konklusionener er den eneste meningsfulde for andelen af resistente bakterier i grisen?

1 ☐ Andelen er 137.5%

2 ☐ Den mest sandsynlige andel er 80%

3 ☐ Den mest sandsynlige andel er 100%.

4 ☐ Den mest sandsynlige andel er 72.7%

5 ☐ Den eneste mulighed er, at der må være sket en fejl i laboratoriet.

Fortsæt på side 20

Opgave XII

En supermarkedskæde vil undersøge effekten af en reklamekampagne. I 15 butikker har de talt antal solgte varer i ugen før og ugen efter reklamekampagnen.

Nedstående tabel viser de indsamlede data, som er indlæst i vektorerne `before` og `after` i R:

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15
before	31	83	136	493	28	505	510	127	138	19	35	37	268	64	224
after	38	79	132	551	33	560	547	124	152	15	41	39	278	66	290

Endvidere oplyses resultaterne ved kørsel af tre forskellige kommandoer på de to vektorer:

	before	after
mean	179.9	196.3
sd	182.1	202.1
var	33160.6	40861.0

Spørgsmål XII.1 (25)

Et ikke-parametrisk bootstrappet 95% konfidensinterval for middelværdien af den relative ændring i salget findes med hvilken af følgende koder (`k` er i alle sat til 10000)?

1 ☐
```
simSamples <- replicate(k, sample((after-before)/before, replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

2 ☐
```
simSamples <- replicate(k, sample((after-before)/(before+after), replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

3 ☐
```
simSamples <- replicate(k, sample(after-before, replace = TRUE))/
              replicate(k, sample(after, replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

4 ☐
```
simSamples <- replicate(k, sample(after-before, replace = TRUE)) -
              replicate(k, sample(after, replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

5 ☐

```r
simSamples <- replicate(k, sample(after-before, replace = FALSE))/
               replicate(k, sample(after, replace = TRUE))
quantile(apply(simSamples, 2, mean), c(0.025, 0.975))
```

Spørgsmål XII.2 (26)

Under antagelse af normalfordeling findes et 95% konfidensinterval for variansen af det ugentlige salg før reklamekampagnen findes til (uagtet af, om antagelsen er realistisk):

1 ☐ $\left[\frac{33161}{26.1}, \frac{33161}{5.63}\right]$

2 ☐ $\left[\sqrt{\frac{6499468}{23.7}}, \sqrt{\frac{6499468}{6.57}}\right]$

3 ☐ $\left[\frac{464248}{26.1}, \frac{464248}{5.63}\right]$

4 ☐ $\left[\frac{2549}{26.1}, \frac{2549}{5.63}\right]$

5 ☐ $\left[\sqrt{\frac{2549}{23.7}}, \sqrt{\frac{2549}{6.57}}\right]$

Fortsæt på side 22

Opgave XIII

En papirproducent ønsker at finde ud af om, der er forskel på papirkvaliteten produceret med træ fra forskellige leverandører (`supplier`). I fremstillingen måles en bestemt variabel Y, som man ved, bestemmer kvaliteten af papiret, således at jo højere værdi som måles, des højere bliver papirets kvalitet. Følgende værdier er samlet ind fra adskilte produktionskørsler med træ fra 3 forskellige leverandører:

Supplier A	Supplier B	Supplier C
9.3	14.0	10.4
9.2	10.5	10.4
8.0	10.5	9.6
6.9	8.3	8.5

Spørgsmål XIII.1 (27)

Firmaets ingeniører har kørt følgende analyse i R. Hvad bliver konklusionen på signifikansniveau $\alpha = 5\%$ for, om der er forskel på kvaliteten af papiret produceret med træ fra de 3 leverandører (både konklusion og argument skal være korrekt)?

```
anova(lm(y ~ Supplier))

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## Supplier   2 12.302  6.1508  2.4126 0.1449
## Residuals  9 22.945  2.5494
```

1 ☐ Der <u>kan ikke</u> påvises en signifikant forskel på kvaliteten, da p-værdien er <u>under</u> signifikansniveauet.

2 ☐ Der <u>kan</u> påvises en signifikant forskel på kvaliteten, da p-værdien er <u>over</u> signifikansniveauet.

3 ☐ Der <u>kan ikke</u> påvises en signifikant forskel på kvaliteten, da p-værdien er <u>over</u> signifikansniveauet.

4 ☐ Der <u>kan</u> påvises en signifikant forskel på kvaliteten, da p-værdien er <u>under</u> signifikansniveauet.

5 ☐ Ingen af ovenstående konklusioner er korrekte.

Spørgsmål XIII.2 (28)

Der refereres igen til analysen fra forrige spørgsmål. Hvor stor en andel af den totale variation er forklaret af effekten af de tre leverandører?

1 ☐ $\frac{2.5494}{12.302+22.945+6.1508+2.5494} = 0.058\%$

2 ☐ $\frac{6.1508}{12.302+22.945+6.1508+2.5494} = 14.0\%$

3 ☐ $\frac{2.5494}{6.1508+2.5494} = 29.3\%$

4 ☐ $\frac{12.302}{12.302+22.945} = 34.9\%$

5 ☐ $\frac{6.1508}{6.1508+2.5494} = 70.7\%$

Fortsæt på side 24

Opgave XIV

Denne opgave er en fortsættelse af forrige opgave. Ingeniørerne kommer nu i tanker om, at kørslerne var lavet på 4 forskellige anlæg (`plant`). Dette var der taget højde for i forsøgsdesignet, således at der for hver leverandør blev skiftet mellem maskinerne i hver kørsel. Det er derfor muligt at tage højde for effekten af anlægget i analysen. Data sættes nu op således, at det er inddelt efter de to faktorer:

	Supplier A	Supplier B	Supplier C
Plant 1	9.3	14.0	10.4
Plant 2	9.2	10.5	10.4
Plant 3	8.0	10.5	9.6
Plant 4	6.9	8.3	8.5

og følgende analyse køres (bemærk, at nogle værdier i resultatet er erstattet af bogstaver og eventuelle * i resultatet er fjernet):

```
anova(lm(y ~ Supplier + Plant))

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Supplier   2 12.3017  6.1508    A       B
## Plant      3 17.3867  5.7956    C       D
## Residuals  6  5.5583  0.9264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Spørgsmål XIV.1 (29)

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

1 ☐ Der er både en signifikant effekt af leverandør og anlæg, da de relevante p-værdier er henholdsvis 0.030 og 0.028.

2 ☐ Der er en signifikant effekt af leverandør, men ikke af anlæg, da de relevante p-værdier er henholdsvis 0.030 og 0.056.

3 ☐ Der er ikke signifikant effekt af hverken leverandør eller anlæg, da de relevante p-værdier er henholdsvis 0.060 og 0.056.

4 ☐ Der er en ikke signifikant effekt af leverandør, men der er signifikant effekt af anlæg, da de relevante p-værdier er henholdsvis 0.060 og 0.028.

5 ☐ Der er ikke signifikant effekt af hverken leverandør eller anlæg, da de relevante p-værdier er henholdsvis 0.12 og 0.17.

Spørgsmål XIV.2 (30)

Hvilke antagelser skal valideres, før resultatet af analysen kan anvendes?

1 ☐ Man behøver ikke at validere nogen antagelser.

2 ☐ Antagelsen om varianshomogenitet af afvigelserne skal valideres, men CLT medfører at normalitetsantagelsen ikke behøver at blive valideret.

3 ☐ Man bør foretage validering, men det er ikke muligt pga. den lave antal af observationer.

4 ☐ Antagelsen om varianshomogenitet af afvigelserne behøver ikke valideres, men antagelsen om normalfordeling af afvigelserne skal valideres.

5 ☐ Både antagelsen om varianshomogenitet og normalfordeling af afvigelserne skal valideres.

SÆTTET ER SLUT. Fortsat god sommer!