Danmarks Tekniske Universitet *Side 1 af 27 sider.*

Skriftlig prøve: 28. maj 2016

Kursus navn og nr: **Introduktion til Statistik (02323, 02402 og 02593)**

Tilladte hjælpemidler: Alle

Dette sæt er besvaret af

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

Opgavesættet består af 30 spørgsmål af "multiple choice" typen fordelt på 12 opgaver. Besvarelserne af "multiple choice"spørgsmålene anføres i det i CampusNet uploadede svarark, med numrene på de svarmuligheder, du mener er de korrekte.

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	IV.1	V.1	V.2	V.3	VI.1	VI.2
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

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

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

Husk at angive dit **studienummer** på din besvarelse. Sættets sidste side er nr. 27; blad lige om og se, at den 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

I sikkerhedskontrollen i en lufthavn tjekker man præcis 10000 passagerer hver dag. Baseret på data fra en længere periode er man kommet frem til at i gennemsnit 8 per 10000 passagerer har skarpe genstande i håndbagagen. Lad X være en stokastisk variabel, som beskriver antallet af passagerer med skarpe genstande på en dag (dvs. antaget ud af præcis 10000). X antages at følge en binomialfordeling.

Spørgsmål I.1 (1)

Hvad er det forventede antal passagerer med skarpe genstande på en dag og hvad er variansen for X?

1 ☐ $E[X] = 0.0008$ og $V[X] = 10000 \cdot 0,8 \cdot 0,2 = 1600$

2 ☐ $E[X] = 0.0008 \cdot 10000 = 8$ og $V[X] = 10000 \cdot 0.0008 \cdot 0.0002 = 0.0016$

3 ☐ $E[X] = 0.0008 \cdot 10000 = 8$ og $V[X] = 10000 \cdot 0.0008 \cdot 0.9992 = 7.994$

4 ☐ $E[X] = 0.0008 \cdot 10000 = 8$ og $V[X] = 10000 \cdot 0.0008 = 8$

5 ☐ $E[X] = 0.8 \cdot 10 = 8$ og $V[X] = 10 \cdot 0.8 \cdot 0.2 = 1.6$

Spørgsmål I.2 (2)

Hvad er sandsynligheden for at finde mere end 10 passagerer med skarpe genstande på en given dag?

1 ☐ `qbinom(0.9, 10000, 0.0008)`

2 ☐ `1-dbinom(9990, 10000, 0.0008)`

3 ☐ `dbinom(10, 10000, 0.0008)`

4 ☐ `1-pbinom(9990, 10000, 0.0008)`

5 ☐ `1-pbinom(10, 10000, 0.0008)`

Fortsæt på side 3

Opgave II

En medicinalvirksomhed har lavet et studie, hvor 300 personer blev tilfældigt inddelt i 3 behandlingsgrupper med 100 patienter i hver. En gruppe fik en placebobehandling, en gruppe fik virksomhedens eget produkt, og den sidste gruppe fik en konkurrents produkt. For hver patient måltes vægtændringen over en tidsperiode, og man ender med et datamateriale med 300 observationer af vægtændring, hvor fokus er på sammenligning af den gennemsnitlige vægtændring i hver gruppe.

Spørgsmål II.1 (3)

Hvilken form for statistisk analyse er mest velegnet til dette?

1 ☐ Multiple lineær regressionsanalyse

2 ☐ Test for uafhængighed i en $r \times c$ antalstabel

3 ☐ Parret t-test

4 ☐ Tovejs variansanalyse

5 ☐ Envejs variansanalyse

Fortsæt på side 4

Opgave III

En stokastisk variabel X følger en uniform fordeling på intervallet $[0; 1]$.

Spørgsmål III.1 (4)

Middelværdien og variansen af $(X + 2) \cdot 4$ er

1 ☐ $\mu = \frac{5}{2}$ og $\sigma^2 = 4^2$

2 ☐ $\mu = 10$ og $\sigma^2 = 4^2$

3 ☐ $\mu = 8$ og $\sigma^2 = 4^2$

4 ☐ $\mu = 8$ og $\sigma^2 = \frac{1}{3}$

5 ☐ $\mu = 10$ og $\sigma^2 = \frac{4}{3}$

Fortsæt på side 5

Opgave IV

En producent af droner har stor fokus på flyvetiden mellem opladninger. Flyvetiden afhænger bl.a. af dronens vægt og dronen består basalt set af et batteri (B), et skelet (S) og fire motorer med rotorer ($M_1, \ldots, M_4$). Det antages at vægtene af de enkelte komponenter er uafhængige af hinanden og i det følgende er alle vægte i gram. Man har fundet, at de tre forskellige typer komponenters vægte kan beskrives ved følgende normalfordelinger: Batteri: $B \sim N(100, 10^2)$, skelet: $S \sim N(40, 5^2)$ og motorer med rotor: $M_i \sim N(15, 2^2),\ i = 1, \ldots, 4$. (Hver fordeling er angivet på den sædvanlige form: $N(\mu, \sigma^2)$)

Spørgsmål IV.1 (5)

Middelværdi og varians for vægten af de færdige droner findes til

1 ☐ $\mu = 200$ og $\sigma^2 = 189$

2 ☐ $\mu = 200$ og $\sigma^2 = 141$

3 ☐ $\mu = 155$ og $\sigma^2 = 189$

4 ☐ $\mu = 155$ og $\sigma^2 = 129$

5 ☐ $\mu = 170$ og $\sigma^2 = 141$

Fortsæt på side 6

Opgave V

Der findes en anbefaling om at spise 600 g frugt og grønt om dagen. Man laver med jævne mellem stikprøveudersøgelser af danskernes kostvaner for at se, om anbefalingen overholdes.

Resultaterne for det daglige indtag af frugt og grønt (i gram) for de seneste fire af denne slags kostundersøgelser (foretaget i årene 1995, 2000-2002, 2003-2004 samt 2005-2008) kan opgøres med følgende output fra R.

Undersøgelse	n	median	mean	var	std
1995	1564	259.82	290.887	28861.55	169.887
2000-2002	3043	386.057	433.817	62029.21	169.887
2003-2004	1310	404.936	453.279	74159.29	272.322
2005-2008	1983	429.132	479.285	77166.51	277.789

Undersøgelse	2.5%	5.0%	Q1	Q3	95.0%	97.5%
1995	66.102	87.062	171.209	374.303	606.609	686.361
2000-2002	98.613	129.574	257.224	555.168	928.673	1055.419
2003-2004	83.48	127.528	256.286	583.723	974.246	1180.891
2005-2008	105.348	141.81	279.359	617.371	991.09	1189.367

I samtlige spørgsmål i denne opgave kan man antage, at data fra hver af de fire undersøgelser er normalfordelt.

Spørgsmål V.1 (6)

Spørgsmålet er ikke længere en del af pensum

Fortsæt på side 7

Spørgsmål V.2 (7)

Der planlægges med udgangspunkt i den observerede variation i kostundersøgelsen 2005-2008 en ny kostundersøgelse. Hvor stor skal stikprøven være, hvis 90%-konfidensintervallet for middelindtaget af frugt og grønt ønskes at have en bredde på 20 g?

1 ☐ $n \approx 77166.51/\left(\frac{20}{1.96}\right)^2 = 741.1$

2 ☐ $n \approx \left(\frac{479.285 \cdot 1.6449}{10}\right)^2 = 6215.4$

3 ☐ $n \approx \frac{77166.51}{1.96 \cdot 20} = 1968.5$

4 ☐ $n \approx \left(\frac{1.6449 \cdot 277.789}{10}\right)^2 = 2087.9$

5 ☐ $n \approx \left(\frac{1.6449 \cdot \sqrt{1983}}{1.6456}\right)^2 = 1981.3$

Spørgsmål V.3 (8)

Bestem 95%-konfidensintervallet for middelindtaget af frugt og grønt i 2003-2004-undersøgelsen.

1 ☐ $404.936 \pm 1.9618 \cdot \sqrt{\frac{74159.29}{1310}} = [390.176; 419.697]$

2 ☐ $[83.48; 1180.891]$

3 ☐ $453.279 \pm 1.9618 \cdot 7.524 = [438.518; 468.040]$

4 ☐ $453.279 \pm 1.6460 \cdot \frac{272.322}{\sqrt{1310}} = [440.894; 465.664]$

5 ☐ $453.279 \pm 1.96 \cdot 272.322 = [-80.472; 987.030]$

Fortsæt på side 8

Opgave VI

En større arbejdsplads udtog en tilfældig stikprøve omfattende 20 medarbejdere og bestemte deres daglige indtag af frugt og grønt, og fik følgende observationer af daglig indtag (i gram):

740.59	262.28
667.96	730.55
809.33	324.19
1138.12	421.93
489.42	561.23
352.78	552.96
1309.66	130.96
259.86	440.82
896.01	955.03
481.00	257.80

I samtlige spørgsmål i denne opgave kan man antage, at data er normalfordelt.

Opsummering i R giver følgende resultater for indtaget af frugt og grønt:

```
   n      median       mean       varians       Std. dev.
  20    521.1898   589.1245     98996.08        314.6364


    2.5%        5.0%        Q1         Q3        95.0%      97.5%
 191.2095    251.4552   345.635   757.777    1146.697   1228.178
```

Spørgsmål VI.1 (9)

Hvad er 90% konfidensintervallet for variansen σ^2 for det daglige indtag af frugt og grønt af medarbejdere i virksomheden?

1 ☐ $[\frac{20 \cdot 314.636}{32.852}; \frac{20 \cdot 314.636}{8.907}] = [191.548; 706.492]$

2 ☐ $98996.08 \pm 30.144 \cdot \frac{314.636}{\sqrt{20}} = [96875.31; 101116.90] = [311.248^2; 317.989^2]$

3 ☐ $98996.08 \pm 1.7959 \cdot \frac{314.636^2}{\sqrt{20}} = [59241.79; 138750.40] = [243.396^2; 372.492^2]$

4 ☐ $[314.636^2 - 10.117 \cdot 314.636; 314.636^2 + 30.144 \cdot 314.636] = [309.537^2; 329.364^2]$

5 ☐ $[\frac{19 \cdot 314.636^2}{30.144}; \frac{19 \cdot 314.636^2}{10.117}] = [249.796^2; 431.181^2]$

Fortsæt på side 9

Spørgsmål VI.2 (10)

Faktisk består ovenstående data af 2 stikprøver, hvor venstre søjle angive indtagene for 10 mænd og højre søjle indtagene for 10 kvinder. Man ønsker at undersøge, om der er forskel i mænd og kvinders middelindtag af frugt og grønt.

Der foretages følgende kørsler i R: (hvoraf ikke alle nødvendigvis er fornuftige)

```
m <- c(740.59, 667.96, 809.33, 1138.12, 489.42, 352.78,
       1309.66, 259.86, 896.01, 481.00)
f <- c(262.28, 730.55, 324.19, 421.93, 561.23, 552.96,
       130.96, 440.82, 955.03, 257.80)
t.test(m, f, paired = TRUE)

##
##  Paired t-test
##
## data:  m and f
## t = 1.7378, df = 9, p-value = 0.1163
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -75.65101 577.04701
## sample estimates:
## mean of the differences
##                 250.698

mean(f) - mean(m)

## [1] -250.698

t.test(m, f)

##
##  Welch Two Sample t-test
##
## data:  m and f
## t = 1.9001, df = 16.481, p-value = 0.07506
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28.33599 529.73199
## sample estimates:
## mean of x mean of y
##   714.473   463.775
```

Fortsæt på side 10

```r
t.test(m, mu = median(f))
```

```
##
##  One Sample t-test
##
## data:  m
## t = 2.6577, df = 9, p-value = 0.02614
## alternative hypothesis: true mean is not equal to 431.375
## 95 percent confidence interval:
##  473.5091 955.4369
## sample estimates:
## mean of x
##   714.473
```

```r
t.test(f)
```

```
##
##  One Sample t-test
##
## data:  f
## t = 5.957, df = 9, p-value = 0.0002135
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  287.6581 639.8919
## sample estimates:
## mean of x
##   463.775
```

Fortsæt på side 11

Hvad er konklusionen på test af hypotesen (på niveau $\alpha = 0.05$)

$$H_0 : \mu_f = \mu_m$$
$$H_1 : \mu_f \neq \mu_m$$

svarende til en undersøgelse af om der er forskel i mænd og kvinders middelindtag af frugt og grønt.

1 ☐ Ja der er signifikant forskel mellem mænd og kvinders indtag af frugt og grønt pr. dag, idet den relevante p-værdi er 0.1163

2 ☐ Det fremgår at, der er en signifikant forskel på mænd og kvinders indtag af frugt og grønt pr. dag, idet $\hat{\mu}_f - \hat{\mu}_m = 463.775 - 714.473 = -250.698$ Det fremgår at mænd spiser mere frugt og grønt pr dag end kvinder.

3 ☐ Der er ikke signifikant forskel på mænd og kvinders indtag af frugt og grønt pr. dag, idet den relevante p-værdi er 0.07506

4 ☐ Ja der er signifikant forskel mellem mænd og kvinders indtag af frugt og grønt pr. dag, idet den relevante p-værdi er 0.02614

5 ☐ Nej, der er ikke signifikant forskel på indtag af frugt og grønt pr. dag for mænd og kvinder, idet den relevante p-værdi er 0.0002135

Spørgsmål VI.3 (11)

Hvad er den øvre kvartil (75% fraktil) for de 10 indtag for kvinderne, baseret på lærebogens definition af dette?

1 ☐ 561.23

2 ☐ 262.28

3 ☐ 431.375

4 ☐ 709.97

5 ☐ 246.19

Fortsæt på side 12

Opgave VII

I et studie har man undersøgt dioxinudledningen fra et dansk forbrændingsanlæg. Et udsnit af de målte variable er vist i tabellen herunder, de 3 variable er: Dioxin målt i "parts per millon", belastning af anlægget målt som relativ afvigelse fra en reference, samt indholdet af vand i den udledte gas (målt i %). Som det ses i tabellen er der i alt 23 målinger. Gennemsnit og empirisk standard afvigelse ("sample standard deviation") er angivet nederst i tabellen.

	Dioxin (ppm)	Belastning	H_20 (%)
	DIOX	NEFF	H2O
1	984.10	0.2560	13.78
2	662.00	0.3520	14.59
3	270.90	-0.0200	12.55
⋮	⋮	⋮	⋮
21	112.70	0.0490	13.84
22	94.20	0.1350	14.18
23	323.20	0.2820	12.56
$\bar{x}$	329.16	-0.0266	12.589
s	254.95	0.2105	1.980

Den primære interesse i studiet knytter sig til om dioxinudledningen kan påvirkes ved at justere belastningen. Til dette formål har man kørt nedenstående R-kode (idet dataindlæsningen dog er udeladt)

```
fit1 <- lm(DIOX ~ NEFF)
summary(fit1)

##
## Call:
## lm(formula = DIOX ~ NEFF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -348.41 -116.61  -22.98  101.19  496.16
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)    347.8       44.7   7.781 0.000000128 ***
## NEFF           702.2      215.3   3.262     0.00373 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 212.6 on 21 degrees of freedom
## Multiple R-squared:  0.3362,Adjusted R-squared:  0.3046
## F-statistic: 10.64 on 1 and 21 DF,  p-value: 0.00373
```

Fortsæt på side 13

Man undersøger altså modellen

$$\texttt{DIOX}_i = \beta_0 + \beta_1 \texttt{NEFF}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

Spørgsmål VII.1 (12)

På signifikansniveau $\alpha = 0.05$ hvad er konklusionen omkring belastningens effekt på dioxin udledningen (både konklusion og argument skal være korrekt)?

1 ☐ Der er en effekt da $1.3 \cdot 10^{-7} < 0.05$, og $\beta_1 > 0$ da $347.8 > 0$

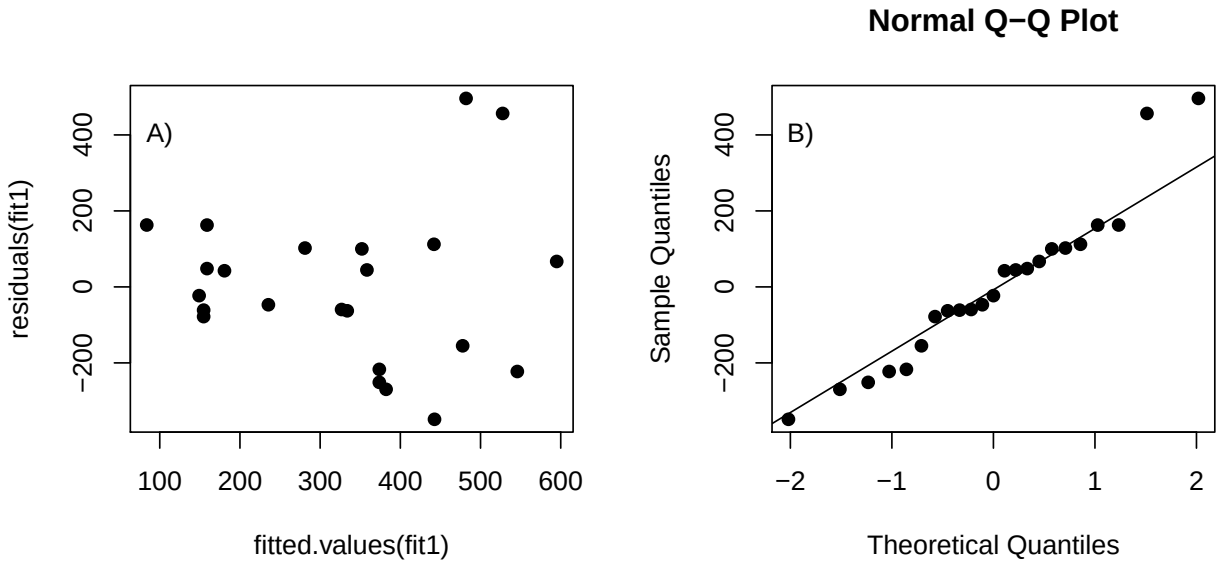
2 ☐ Der er en effekt da $702.2 > 347.2$, og $\beta_1 > 0$ da $3.26 > 0$

3 ☐ Der er en effekt da $0.0037 < 0.05$, og $\beta_1 > 0$ da $702.2 > 0$

4 ☐ Der kan ikke påvises en effekt da $3.26 < 7.78$.

5 ☐ Der kan ikke påvises en effekt da $0.0037 > \frac{0.05}{100}$.

For at undersøge om forudsætningerne for at bruge modellen er opfyldt har man lavet 2 residualplot i figuren herunder.



Fortsæt på side 14

Spørgsmål VII.2 (13)

Hvilke antagelser undersøges primært i hver de 2 plots (både antagelser og figurhenvisning skal være korrekt)?

1 ☐ Varianshomogenitet (A) og normalfordelingsantagelse (B)

2 ☐ $E(\epsilon) = 0$ (A) og $V(\epsilon) = \sigma^2$ (B)

3 ☐ Varianshomogenitet (A) og linearitetsantagelsen (B)

4 ☐ $E(\epsilon) = 0$ (A) og uafhængighed (B)

5 ☐ Uafhængighed (A) og varianshomogenitet (B)

Uanset udfaldet af forrige spørgsmål beslutter man at foretage analysen på log-tranformerede dioxin-data. Resultatet af analysen foretaget i R er vist herunder (idet en del tal dog er erstattet af bogstaver)

```
> fit2 <- lm(log(DIOX) ~ NEFF)
> summary(fit2)

Call:
lm(formula = log(DIOX) ~ NEFF)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29588 -0.44048  0.05093  0.49403  0.94119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.5927        A     B  < 2e-16 ***
NEFF          1.8416        C     D     E
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6266 on 21 degrees of freedom
Multiple R-squared:  0.2862,Adjusted R-squared:  0.2522
F-statistic:  8.42 on 1 and 21 DF,  p-value: 0.00853
```

Fortsæt på side 15

Spørgsmål VII.3 (14)

Hvad er D?

1 ☐ D= $\frac{0.623^2}{21} = 0.019$

2 ☐ D= $0.623 \cdot \sqrt{\frac{1}{22 \cdot 0.211^2}} = 0.63$

3 ☐ D= $\frac{1.84}{\mathtt{C}}$

4 ☐ D= $\frac{\mathtt{C}}{\mathtt{B}}$

5 ☐ D= $\frac{0.623}{\sqrt{22}} = 0.13$

Spørgsmål VII.4 (15)

Hvad er det sædvanlige 95% konfidensinterval for hældningen i modellen for log(DIOX)?

1 ☐ $1.84 \pm 1.72 \cdot \mathtt{B}$

2 ☐ $1.84 \pm 2.08 \cdot 0.2862$

3 ☐ $1.84 \pm 1.72 \cdot \mathtt{D}$

4 ☐ $1.84 \pm 2.08 \cdot 0.623$

5 ☐ $1.84 \pm 2.08 \cdot \mathtt{C}$

Fortsæt på side 16

Man ønsker nu at undersøge om vanddamp bør inkluderes i modellen. Til dette formål formuleres en multipel regressionsmodel

$$\log(\text{DIOX}_i) = \beta_0 + \beta_1 \text{NEFF}_i + \beta_2 \text{H2O}_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

For at undersøge modellen har man kørt nedenstående R-kode

```
fit3 <- lm(log(DIOX) ~ NEFF + H2O)
summary(fit3)

##
## Call:
## lm(formula = log(DIOX) ~ NEFF + H2O)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11709 -0.36741  0.05337  0.36192  0.90410
##
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)
## (Intercept)   7.4704     0.8098   9.225 0.0000000121 ***
## NEFF          2.1963     0.5955   3.688      0.00146 **
## H2O          -0.1484     0.0633  -2.345      0.02948 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5687 on 20 degrees of freedom
## Multiple R-squared:  0.4401,Adjusted R-squared:  0.3841
## F-statistic:  7.86 on 2 and 20 DF,  p-value: 0.003028
```

Spørgsmål VII.5 (16)

Hvad er parameterestimaterne for modellen?

1 ☐ $\hat{\beta}_0 = 7.47$, $\hat{\beta}_1 = 2.20$, $\hat{\beta}_2 = -0.148$ og $\hat{\sigma} = 0.569$

2 ☐ $\hat{\beta}_0 = 9.22$, $\hat{\beta}_1 = 3.69$, $\hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 0.4401$

3 ☐ $\hat{\beta}_0 = 7.47$, $\hat{\beta}_1 = 2.20$, $\hat{\beta}_2 = -0.148$ og $\hat{\sigma} = 0.384$

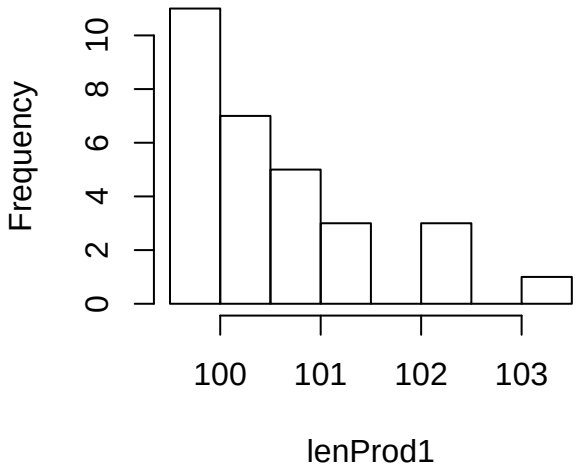
4 ☐ $\hat{\beta}_0 = 9.22$, $\hat{\beta}_1 = 3.69$, $\hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 0.569$

5 ☐ $\hat{\beta}_0 = 9.22$, $\hat{\beta}_1 = 3.69$, $\hat{\beta}_2 = -2.35$ og $\hat{\sigma} = 7.86$

Fortsæt på side 17

Opgave VIII

I en produktion regner man med at skrotte en del af de producerede elementer pga. krav om en minimumslængde. Man har regnet ud, at det økonomisk er ok, hvis man skrotter 25% af elementerne. Et eksperiment er udført med en produktionsmetode og målinger er indsamlet for længden af 50 producerede elementer. Observationerne er indlæst og gemt i vektoren `lenProd1`. Et histogram af observationerne er



Man vil udregne et konfidensinterval for den nedre kvartil (dvs. 25% fraktilen) vha. simulering uden antagelse af fordeling. Følgende R-kode er kørt:

```
## Simuler 10000 stikprøver
k = 10000
simSamples = replicate(k, sample(lenProd1, replace = TRUE))

simStat = apply(simSamples, 2, quantile, probs=0.25)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##     0.5%     2.5%       5%      95%    97.5%    99.5%
##  99.5825  99.7225  99.7600 100.0575 100.1025 100.2300

simStat = apply(simSamples, 2, quantile, probs=0.5)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##     0.5%     2.5%       5%      95%    97.5%    99.5%
##  99.9450  99.9850 100.0000 100.5900 100.6550 100.8801

simStat = apply(simSamples, 2, quantile, probs=0.75)
quantile(simStat, c(0.005,0.025,0.05,0.95,0.975,0.995))

##     0.5%     2.5%       5%      95%    97.5%    99.5%
## 100.3000 100.4625 100.5125 101.4300 101.9550 102.1450
```

Fortsæt på side 18

Bemærk, at optionen `probs` bliver "sendt videre"til `quantile` funktionen, så de tre forskellige `apply`-kald beregner en forskellig fraktil med `quantile` funktionen.

Spørgsmål VIII.1 (17)

Hvad er 95% konfidensintervallet for den nedre kvartil (dvs. 25% fraktilen) af længden?

1 ☐ $[99.72, 100.10]$

2 ☐ $[100.00, 100.59]$

3 ☐ $[99.59, 100.23]$

4 ☐ $[100.46, 101.96]$

5 ☐ $[100.49, 101.43]$

Spørgsmål VIII.2 (18)

I nedenstående angiver Q en kvartil, således at Q_1 er den nedre kvartil, Q_2 er medianen og Q_3 er den øvre kvartil. I hvilken af følgende to-sidede tests ville nulhypotesen blive afvist på niveau $\alpha = 0.01$ under de ovenstående antagelser og simuleringsresultater?

1 ☐ $H_0 : Q_1 = 100$ vs. $H_1 : Q_1 \neq 100$

2 ☐ $H_0 : Q_2 = 100$ vs. $H_1 : Q_2 \neq 100$

3 ☐ $H_0 : Q_2 = 101$ vs. $H_1 : Q_2 \neq 101$

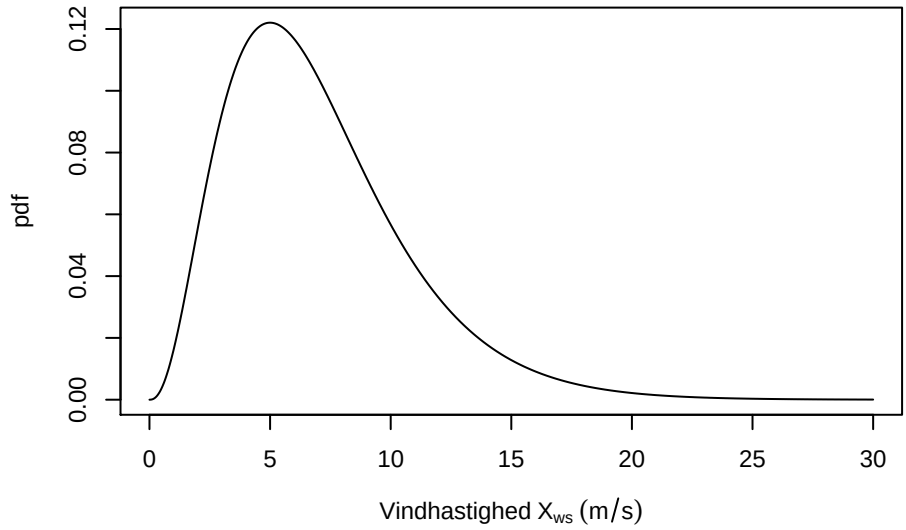
4 ☐ $H_0 : Q_3 = 101$ vs. $H_1 : Q_3 \neq 101$

5 ☐ $H_0 : Q_3 = 102$ vs. $H_1 : Q_3 \neq 102$

Fortsæt på side 19

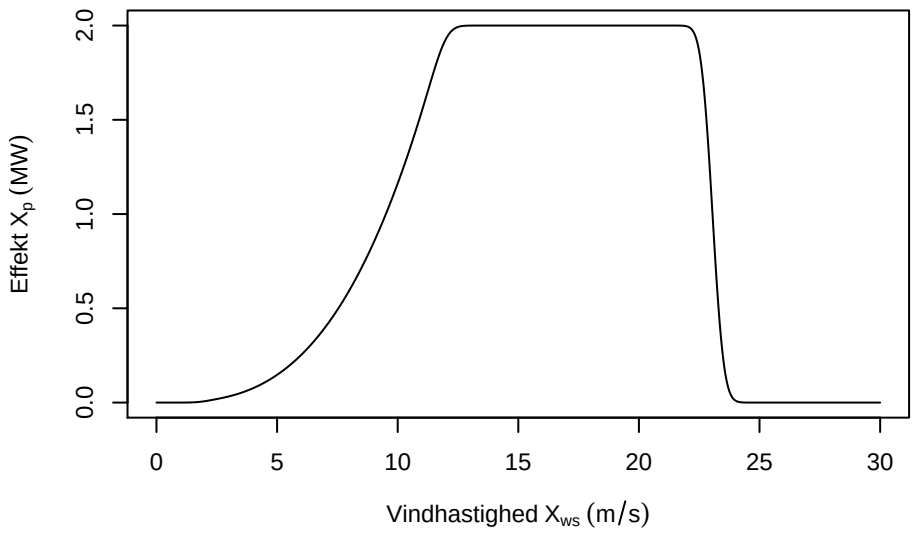
Opgave IX

Under planlægningen af opførelse af en ny vindmølle er der udført undersøgelser af vindforholdene på stedet, hvor vindmøllen skal stå. Man har fundet ud af, at den gennemsnitlige vindhastighed per time på stedet kan beskrives med nedenstående tæthedsfunktion (pdf) i figur 1:



Figur 1: *Tæthedsfunktionen (pdf) for vindhastigheden* X_{ws}.

For at undersøge hvordan vindmøllens energiproduktion vil være bruges en funktion, den såkaldte 'effektkurve', for vindmøllen. Den anvendte effektkurve er vist i figur 2:



Figur 2: *Effektkurven, dvs. funktion mellem vindhastigheden* X_{ws} *og effekten* X_{p}.

Det ses, at med en vindhastighed på 5 m/s bliver den generede effekt omkring 0.15 MW og ved 15 m/s genereres 2 MW. Denne funktion kan anvendes direkte på gennemsnitlige timeværdier af vindhastighed og giver da gennemsnitlige timeværdier af effekt.

Fortsæt på side 20

Lad X_{ws} være vindhastigheden, så bliver den generede effekt

$$X_{\text{p}} = f_{\text{effektkurve}}(X_{\text{ws}})$$

hvor $f_{\text{effektkurve}}()$ er effektkurvefunktionen vist i figur 2.

Spørgsmål IX.1 (19)

Udfra plottet af tæthedsfunktionen i figur 1, konkluder hvilket af følgende sandsynlighedsudsagn der ikke er korrekt (NB: Du skal altså markere det FALSKE udsagn - fire af udsagnene er korrekte!):

1 ☐ $P(X_{\text{ws}} > 12) \approx 0.10$

2 ☐ $P(X_{\text{ws}} < 5) \approx 0.34$

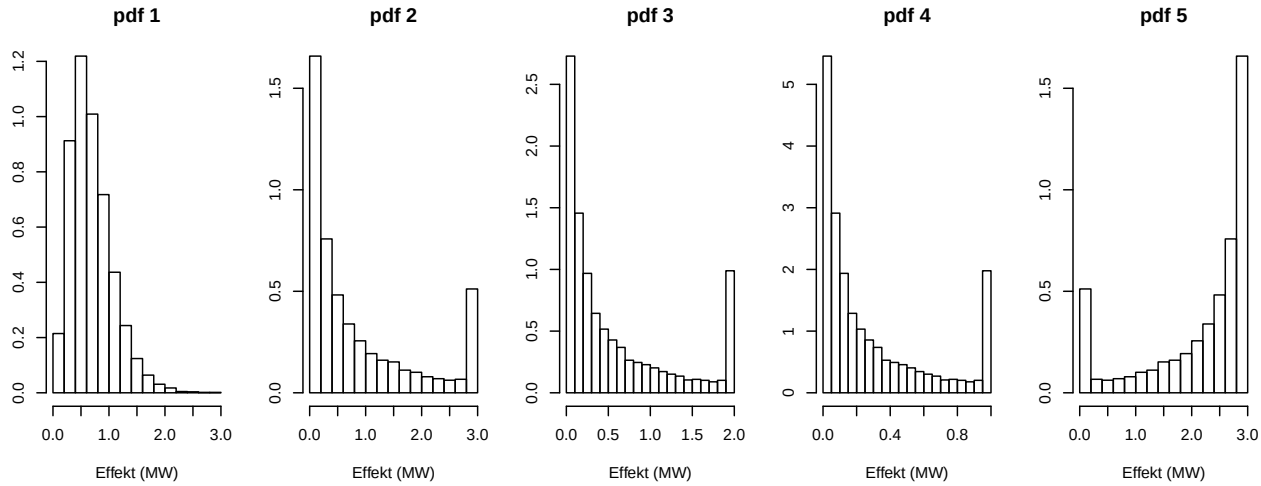
3 ☐ $P(X_{\text{ws}} > 10) \approx 0.19$

4 ☐ $P(X_{\text{ws}} > 0) \approx 1$

5 ☐ $P(X_{\text{ws}} < 15) \approx 0.04$

Spørgsmål IX.2 (20)

Tæthedsfunktionen for den generede effekt (gennemsnit per time) X_{eff} er fundet ved simulering. Hvilken af følgende tæthedsfunktioner (pdfer) kan være tæthedsfunktionen for X_{eff}?



Fortsæt på side 21

1 ☐ pdf 1

2 ☐ pdf 2

3 ☐ pdf 3

4 ☐ pdf 4

5 ☐ pdf 5

Spørgsmål IX.3 (21)

Ved beregning af prognoser af genereret effekt fra vindmøller er det meget vigtigt at inkludere usikkerheden. Den kan beskrives med variansen på effektprognosen σ^2_{eff}. I intervallet mellem 5 og 10 m/s er effektkurvefunktionen

$$f_{\text{effektkurve}}(X_{\text{ws}}) = aX^3_{\text{ws}} \quad \textit{for } 5 < X_{\text{ws}} < 10 \tag{1}$$

Desuden kendes variansen på vindprognosen i intervallet, den er betegnet med σ^2_{ws}.

Hvilken af følgende udtryk beregner en god tilnærmelse til variansen på effektprognosen i intervallet fra 5 til 10 m/s?

1 ☐ $\sigma^2_{\text{eff}} = X^3_{\text{ws}}\sigma^2_{\text{ws}}$

2 ☐ $\sigma^2_{\text{eff}} = 9a^2X^4_{\text{ws}}\sigma^2_{\text{ws}}$

3 ☐ $\sigma^2_{\text{eff}} = \int_5^{10} \sigma^2_{\text{ws}} 3ax^2 dx$

4 ☐ $\sigma^2_{\text{eff}} = \int_5^{10} \sigma^2_{\text{ws}} x^3 dx$

5 ☐ $\sigma^2_{\text{eff}} = a^2\sigma^2_{\text{ws}}$

Fortsæt på side 22

Opgave X

En supermarkedskæde vil gerne følge udviklingen i salget af økologisk kød. Derfor har de i fire år lavet en undersøgelse blandt deres kunder, hvor de bl.a. spørger, om kunderne har købt økologisk kød. Fordelingen af svar ses i tabellen nedenfor.

	2011	2012	2013	2014
Købte økologisk kød	68	72	81	90
Købte ikke-økologisk kød	432	428	419	410

Spørgsmål X.1 (22)

Supermarkedskæden ønsker at teste hypotesen om, at der er den samme andel, der køber økologisk kød hvert år. Dvs.

$$H_0 : p_1 = p_2 = p_3 = p_4$$

Hvor p_1 er andelen, der køber økologisk kød i 2011, p_2 andelen, der køber økologisk kød i 2012 osv.

Hvad er det forventede antal økologiske køb i 2014 under hypotesen om, at andelene er de samme hvert år?

1 ☐ 144.69

2 ☐ 250.00

3 ☐ 77.75

4 ☐ 43.48

5 ☐ 422.25

Fortsæt på side 23

Spørgsmål X.2 (23)

Hypotesen

$$H_0 : p_1 = p_2 = p_3 = p_4$$

testes vha. en χ^2-fordelt teststørrelse. Bestem bidraget $q_{Nej,2011}$ til teststørrelsen χ^2_{obs}, som kommer fra respondenter, der svarer at de køber ikke-økologisk kød i 2011.

1 ☐ $q_{Nej,2011} = 0.2251$

2 ☐ $q_{Nej,2011} = 1.2227$

3 ☐ $q_{Nej,2011} = 9.75$

4 ☐ $q_{Nej,2011} = 0.0231$

5 ☐ $q_{Nej,2011} = 0.2201$

Spørgsmål X.3 (24)

Supermarkedskæden udfører nu testet af hypotesen

$$H_0 : p_1 = p_2 = p_3 = p_4$$

for at se på udviklingen i salget af økologisk kød.

Den relevante teststørrelse bliver i dette tilfælde 4.3977.

Hvilken af nedenstående R kommandoer beregner p-værdien for testet?

1 ☐ `pchisq(4.3977, 3)`

2 ☐ `2*(1-pnorm(4.3977))`

3 ☐ `1-pchisq(4.3977, 3)`

4 ☐ `2*(1-pchisq(4.3977, 4))`

5 ☐ `1-pchisq(4.3977, 6)`

Fortsæt på side 24

Opgave XI

Undersøgelser har vist, at teenagepiger har en lavere livstilfredshed end drenge. Derfor har et hold studerende på første år besluttet at undersøge livstilfredsheden blandt deres medstuderende. Resultatet af deres undersøgelse ses i tabellen nedenfor.

	Høj livstilfredshed	Ikke høj livstilfredshed
Mænd	68	208
Kvinder	18	74

Spørgsmål XI.1 (25)

Hvad er det korrekte 95% konfidensinterval for forskellen mellem andelen af mænd og kvinder med høj livstilfredshed?

1 ☐ $(0.2464 - 0.1957) \pm 1.64 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.029; 0.131)$

2 ☐ $(0.2464 - 0.1957) \pm 1.96 \cdot \sqrt{(\frac{0.2464(1-0.2464)}{276})^2 + (\frac{0.1957(1-0.1957)}{92})^2} = (0.047; 0.054)$

3 ☐ $\frac{0.2464}{0.1957} \pm 1.96 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (1.16; 1.35)$

4 ☐ $(0.2464 - 0.1957) \pm 3.84 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.137; 0.238)$

5 ☐ $(0.2464 - 0.1957) \pm 1.96 \cdot \sqrt{\frac{0.2464(1-0.2464)}{276} + \frac{0.1957(1-0.1957)}{92}} = (-0.045; 0.146)$

Fortsæt på side 25

Spørgsmål XI.2 (26)

Man ønsker nu at teste hypotesen, at der er den samme andel med høj livstilfredshed blandt mænd og kvinder. Dvs. man vil teste hypotesen (på signifikansniveau $\alpha = 0.05$).

$$\begin{aligned} H_0 : p_1 &= p_2 \\ H_A : p_1 &\neq p_2 \end{aligned}$$

Her er p_1 andelen med høj livstilfredshed blandt mænd, og p_2 andelen med høj livstilfredshed blandt kvinder.

Hvad er konklusionen på dette test? (Både konklusion og argumentation skal være korrekt).

1 ☐ H_0 forkastes, idet $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276+92})}} = 2.298$ giver en p-værdi på 0.02

2 ☐ H_0 accepteres, idet teststørrelsen $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276}+\frac{1}{92})}} = 0.995$ giver en p-værdi på 0.32

3 ☐ H_0 accepteres, idet teststørrelsen $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{\frac{0.2464(1-0.2464)}{276}+\frac{0.1957(1-0.1957)}{92}}} = 1.038$ giver en p-værdi på 0.15

4 ☐ H_0 accepteres, idet $z_{obs} = \frac{(0.2464-0.1957)}{\sqrt{0.2337(1-0.2337)(\frac{1}{276+92})}} = 2.298$ giver en p-værdi på 0.02

5 ☐ H_0 forkastes, idet teststørrelsen $z_{obs} = \frac{(0.2464-0.1957)}{\frac{0.2464(1-0.2464)}{276}+\frac{0.1957(1-0.1957)}{92}} = 21.3$ giver en p-værdi < 0.0001

Fortsæt på side 26

Opgave XII

18 testpersoner har vurderet baskvaliteten i 3 forskellige hovedtelefoner, således at alle 18 personer har vurderet alle 3 hovedtelefoner, så datamaterialet består af 54 observationer af baskvalitet på en skala mellem 0 og 150. Den gennemsnitlige baskvalitet for de tre hovedtelefoner blev:

Hovedtelefon	Gennemsnit
1	53.5
2	55.5
3	97.1

Spørgsmål XII.1 (27)

Hvordan beregnes $SS(Tr)$ i 2-vejs variansanalysen, der undersøger middelbaskvaliteten for de tre hovedtelefoner? (Hvor "Tr" nu refererer til de 3 hovedtelefoner)

1 ☐ $18 \cdot (53.5 - 68.7)^2 + 18 \cdot (55.5 - 68.7)^2 + 18 \cdot (97.1 - 68.7)^2$

2 ☐ $(53.5 - 68.7)^2 + (55.5 - 68.7)^2 + (97.1 - 68.7)^2$

3 ☐ $\frac{(53.5-68.7)^2}{53.5} + \frac{(55.5-68.7)^2}{55.5} + \frac{(97.1-68.7)^2}{55.5}$

4 ☐ $\frac{(53.5-68.7)}{53.5} + \frac{(55.5-68.7)}{55.5} + \frac{(97.1-68.7)}{55.5}$

5 ☐ $3 \cdot (53.5 - 68.7)^2 + 3 \cdot (55.5 - 68.7)^2 + 3 \cdot (97.1 - 68.7)^2$

Spørgsmål XII.2 (28)

Hvis man, i tråd med ovenstående lader "personer" udgøre "blokkene", så oplyses, at $SS(Bl) = 6003.5$ og at $SSE = 7160.3$ i den 2-vejs variansanalyse. Hvad bliver F-teststørrelsen for hypotesen om, at de 18 personer har den samme middelværdi?

1 ☐ $F_{obs} = \frac{18 \cdot 6003.5}{210.6/3}$

2 ☐ $F_{obs} = \frac{3 \cdot 6003.5}{7160.3/17}$

3 ☐ $F_{obs} = \frac{6003.5/17}{7160.3/34}$

4 ☐ $F_{obs} = \frac{(6003.5-210.6)^2}{7160.3}$

5 ☐ $F_{obs} = \frac{(6003.5/18-210.6)}{\sqrt{(210.6)}}$

Fortsæt på side 27

Spørgsmål XII.3 (29)

Hypotesen om ingen forskel i middelbaskvalitet for de tre høretelefoner bliver ved det sædvanlige test vurderet ved brug af hvilken stikprøvefordeling?

1 ☐ z-fordeling (= standard normalfordeling)

2 ☐ t-fordeling med 53 frihedsgrader

3 ☐ χ^2-fordeling med 53 frihedsgrader

4 ☐ F-fordeling med frihedsgraderne 2 og 34

5 ☐ F-fordeling med frihedsgraderne 3 og 51

Spørgsmål XII.4 (30)

Hvad bliver 95% konfidensintervallet for middelforskellen mellem høretelefon 2 og 1? (Idet man kan gå ud fra, at der er tale om en såkaldt "pre-planned"sammenligning)

1 ☐ $2 \pm 2 \cdot 210.6$

2 ☐ $2 \pm 2.03 \cdot \sqrt{210.6}$

3 ☐ $2 \pm 1.96 \cdot \frac{210.6}{54}$

4 ☐ 2 ± 1.96

5 ☐ $2 \pm 2.03 \cdot \sqrt{2 \cdot 210.6 \frac{1}{18}}$

SÆTTET ER SLUT. GOD SOMMER!