Danmarks Tekniske Universitet *Side 1 af 27 sider.*

Skriftlig prøve: 14. august 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 11 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 ikke med (dvs. giver "0 point"). Det antal point, der kræves for, at et sæt anses for tilfredstillende besvaret, afgøres endeligt ved censureringen.

Den endelige besvarelse af opgaverne gøres ved at udfylde og online-aflevere svararket via CampusNet. Skemaet her er KUN et nød-alternativ til dette (husk at angive dit studienummer på din besvarelse, hvis du afleverer skemaet).

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

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

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

Sættet består af 27 sider.

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

Et elselskab har udviklet en app, der kan hjælpe deres forbrugere med at analysere og mindske deres elforbrug. Det skal nu testes, om brugere af appen har mindsket deres elforbrug efter de har installeret den. Elforbruget er målt i kWh. Lad X betegne forskellen i elforbrug mellem måneden før de installerede appen ($X_{\text{før}}$) og elforbrug måneden efter de installerede appen (X_{efter}), således at

$$X = X_{\text{efter}} - X_{\text{før}}$$

Forskellen er registreret for 40 tilfældigt udvalgte brugere, som har installeret appen på forskellige tider på året. Stikprøvegennemsnittet og -standardafvigelsen er udregnet til

$$\bar{x} = -22.6$$
$$s_X = 45.5$$

Spørgsmål I.1 (1)

Beregn et 95% konfidensinterval for middelforskellen μ_X i elforbrug fra før og til efter appen blev installeret

1 ☐ $-11.3 \pm 2.02 \cdot \frac{45.5}{39} = [-13.7, -8.94]$

2 ☐ $-22.6 \pm 2.02 \cdot \frac{45.5}{39} = [-25.0, -20.2]$

3 ☐ $-22.6 \pm 2.02 \cdot \frac{45.5}{6.32} = [-37.1, -8.06]$

4 ☐ $-22.6 \pm 2.02 \cdot \frac{2070}{6.32} = [-684, 639]$

5 ☐ $-22.6 \pm 2.02 \cdot \frac{2070}{39} = [-130, 84.6]$

Spørgsmål I.2 (2)

Kan der påvises et fald på et 5% signifikansniveau i elforbruget fra måneden før til måneden efter installation af appen (både konklusion og argumentation (p-værdien) skal være korrekt)?

1 ☐ Ja, der kan påvises et signifikant fald, da p-værdien for det oplagte to-sidet test er 0.027

2 ☐ Nej, der kan ikke påvises et signifikant fald da p-værdien for det oplagte to-sidet test er 0.027

3 ☐ Ja, der kan påvises et signifikant fald, da p-værdien for det oplagte to-sidet test er 0.0032

4 ☐ Nej, der kan ikke påvises et signifikant fald, da p-værdien for det oplagte to-sidet test er 0.0032

5 ☐ Nej, der kan ikke påvises et signifikant fald, da p-værdien for det oplagte to-sidet test er 0.21

Spørgsmål I.3 (3)

Man har fundet ud af, at en del forbrugere, som installerer appen, ikke kommer igang med at bruge den lige efter installationen. Derfor har man redesignet appens onboarding (processen brugeren skal igennem første gang appen åbnes efter installation) for at få brugerne hurtigere igang. Hvis man sætter sandsynligheden for, at en bruger ikke kommer igang lige efter installationen til $p = 0.20$, samt registrererer 100 nye brugere og lader X betegne antallet af dem, som ikke kommer igang lige efter installationen. Hvilket af følgende udtryk beregner da sandsynligheden for at få mindre end 10 nye brugere som ikke kommer igang, dvs. $P(X < 10)$ i R?

1 ☐ `phyper(q=1, m=20, n=80, k=10)`

2 ☐ `pbinom(q=9, size=100, prob=0.2)`

3 ☐ `dbinom(x=10, size=100, prob=0.2)`

4 ☐ `1 - pbinom(q=10, size=100, prob=0.2)`

5 ☐ `dbinom(x=10, size=100, prob=0.2)`

Fortsæt på side 4

Opgave II

I en række forsøg har man undersøgt hvordan trykstyrken af beton afhænger af sammensætningen af betonen. De registrede forklarende variable er mængden af cement, vand og sand målt som kg/m^3. Og trykstyrken er målt i MPa. En oversigt over data er givet i nedenstående tabel:

	Cement	Water	Fine	Strength
Min.	200.0	146.0	594.0	12.25
1st Qu.	289.0	185.0	754.0	22.49
Median	339.0	186.0	781.0	31.35
Mean	344.9	188.5	776.4	31.83
3rd Qu.	393.0	192.0	809.0	37.42
Max.	540.0	228.0	945.0	74.99

Spørgsmål II.1 (4)

Man starter med at lave en simpel lineær regressionsmodel med mængden af cement som forklarende variabel og får følgende R output (hvor et par tal er erstattet af bogstaver):

```
## Call:
## lm(formula = Strength ~ Cement, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.5512  -0.6858   0.6280   1.4791  20.8302
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.736438   3.389030     A   1.96e-05 ***
## Cement        0.137930   0.009567     B   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.05 on 59 degrees of freedom
## Multiple R-squared:  0.7789,Adjusted R-squared:  0.7752
## F-statistic: 207.9 on 1 and 59 DF,  p-value: < 2.2e-16
```

Den relevante teststørrelse, for at undersøge om mængden af cement påvirker styrken, findes til?

1 ☐ $0.009567/0.137930 \approx 0.06936$

2 ☐ $1.96e^{-5}$

3 ☐ $-15.736438/3.389030 \approx -4.643$

4 ☐ $0.137930/0.009567 \approx 14.42$

5 ☐ 6.05

Spørgsmål II.2 (5)

Man har estimeret parametrene i følgende model

$$Strength_i = \beta_0 + \beta_1 \cdot Cement_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

og man ønsker nu at bruge modellen til at forudsige trykstyrken givet mængden af cement.

Forudsigelsen har lavest varians, når cementmængden er:

1 ☐ 339.0 kg/m^3

2 ☐ 344.9 kg/m^3

3 ☐ 540.0 kg/m^3

4 ☐ 0.0 kg/m^3

5 ☐ Kan ikke besvares med de tilgængelige oplysninger.

Spørgsmål II.3 (6)

Man laver en tilsvarende analyse for sammenhængen mellem vandmængden og trykstyrken. Man får følgende R output:

```
## Call:
## lm(formula = Strength ~ Water, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.361  -8.593  -1.161   5.239  28.568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.1345    23.6644   4.908 7.62e-06 ***
## Water        -0.4474     0.1253  -3.570 0.000718 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.67 on 59 degrees of freedom
## Multiple R-squared:  0.1776,Adjusted R-squared:  0.1637
## F-statistic: 12.74 on 1 and 59 DF,  p-value: 0.0007184
```

Hvis residualerne følger de sædvanlige antagelser, så er konklusionen på analysen at:

1 ☐ Vand har ikke en signifikant indflydelse på styrken. Og jo mere vand jo stærkere beton.

2 ☐ Vand har ikke en signifikant indflydelse på styrken. Og jo mindre vand jo stærkere beton.

3 ☐ Vand har en signifikant indflydelse på styrken. Og jo mere vand jo stærkere beton.

4 ☐ Vand har ikke en signifikant indflydelse på styrken. Og man kan derfor ikke sige noget om hvordan mængden af vand påvirker styrken.

5 ☐ Vand har en signifikant indflydelse på styrken. Og jo mindre vand jo stærkere beton.

Spørgsmål II.4 (7)

Dernæst vælger man at estimere en multipel lineær regressionsmodel med kvadratroden af trykstyrken som respons. De tre andre variable benyttes som forklarende variable:

```
## Call:
## lm(formula = sqrt(Strength) ~ Cement + Water + Fine, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31854 -0.12044  0.06208  0.18913  0.73313
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.0623090  1.3125406   3.857 0.000295 ***
## Cement       0.0120327  0.0007024  17.131  < 2e-16 ***
## Water       -0.0244132  0.0037730  -6.470 2.41e-08 ***
## Fine         0.0012022  0.0009038   1.330 0.188774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3293 on 57 degrees of freedom
## Multiple R-squared:  0.9069,Adjusted R-squared:  0.902
## F-statistic: 185.1 on 3 and 57 DF,  p-value: < 2.2e-16
```

Hvilken af følgende konklusionener på ovenstående output er mest korrekt?

1 ☐ Man bør estimere en model, som kun har "Cement" som forklarende variabel

2 ☐ Man bør fjerne "Fine" og estimere parametrene i den reducerede model

3 ☐ Man bør fjerne "(Intercept)" og estimere parametrene i den reducerede model

4 ☐ Man bør fjerne "Water" og estimere parametrene i den reducerede model

5 ☐ Man bør estimere en model, som kun har "Fine" som forklarende variabel

Spørgsmål II.5 (8)

Baseret på de præsenterede R-outputs i opgaven findes det anvendte antal observationer i hver analyse til:

1 ☐ 61

2 ☐ 60

3 ☐ 59

4 ☐ 2

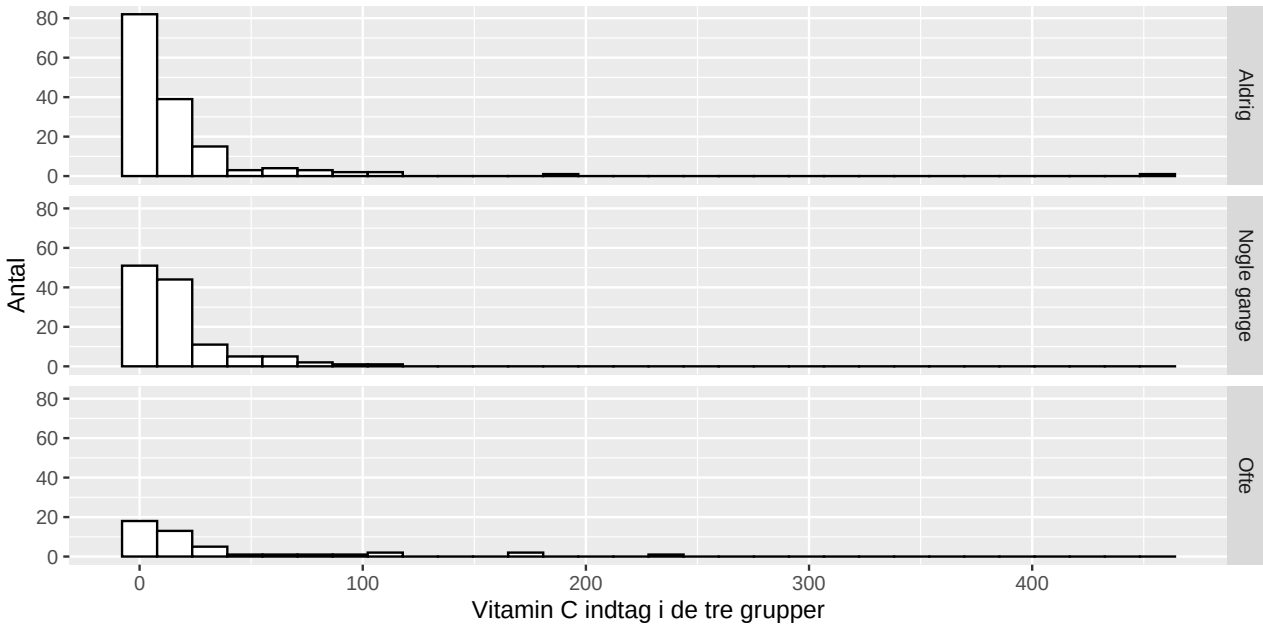
5 ☐ 1

Fortsæt på side 8

Opgave III

Man har lavet en undersøgelse af sammenhængen mellem indtag af vitamin C, og hvor tit man dyrker motion. Nedenfor ses en tabel med deskriptive størrelser for mængden af vitamin C i de tre motionsgrupper og histogrammer fra denne undersøgelse:

	Motion		
	Aldrig	Nogle gange	Ofte
Gennemsnit	20.11	16.71	33.26
Median	7.23	9.27	11.27
SD	44.05	20.55	52.42
n	152	120	45



Man vil gerne undersøge, om der er sammenhæng mellem vitamin C indtag, og hvor tit man dyrker motion. Derfor laver man en ensidet variansanalyse, hvor `z` er vitamin C indtag, og `Gruppe` er en faktor på tre niveauer (Aldrig motion, Nogle gange motion, Ofte motion).

Der er lavet to analyser ved at køre nedenstående R-kode (indlæsning af data er ikke vist):

```
anova(lm(z ~ Gruppe, data=dat))

## Analysis of Variance Table
##
## Response: z
##            Df Sum Sq Mean Sq F value Pr(>F)
## Gruppe      2   9060  4529.8   3.064 0.0481 *
## Residuals 314 464227  1478.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(lm(log(z) ~ Gruppe, data=dat))

## Analysis of Variance Table
##
## Response: log(z)
##            Df Sum Sq Mean Sq F value Pr(>F)
## Gruppe      2   2.94  1.4693   0.918 0.4004
## Residuals 314 502.53  1.6004
```

Spørgsmål III.1 (9)

Hvilket af nedenstående udsagn er korrekt?

1 ☐ Analysen af `z` er den mest korrekte, da man kan se fra tabellen, at der er højere vitamin C indtag i gruppen, som dyrker meget motion

2 ☐ Analysen af `z` er den mest korrekte, fordi man er interesseret i at udtale sig om vitamin C indtag og ikke logaritmen til vitamin C indtag

3 ☐ Analysen af `log(z)` er mest korrekt, for her bliver `Gruppe` ikke statistisk signifikant

4 ☐ Analysen af `log(z)` er den mest korrekte, da man kan se fra histogrammerne, at antagelsen om normalfordelingen ikke kommer til at holde for `z`

5 ☐ Analysen af `z` er den mest korrekte, for her bliver `Gruppe` statistisk signifikant

Fortsæt på side 10

Opgave IV

I et studie af forhøjet blodtryk ville man undersøge, om to forskellige diæter betyder noget for blodtrykket. Studiet blev foretaget på 150 forsøgspersoner, som blev inddelt i 2 grupper á 75 personer. Gruppe I er en kontrolgruppe, som fik normal kost, mens Gruppe II modtog et helsekosttilskud. Resultaterne er vist i nedenstående tabel, idet man definerede et bestemt blodtryk eller derover som 'højt' (ikke nødvendigvis for højt):

Gruppe	Normalt blodtryk	Højt blodtryk	Total
I	55	20	75
II	57	18	75
Total	112	38	150

Spørgsmål IV.1 (10)

Et 95% konfidensinterval for andelen af personer på normal kost (Gruppe I) med højt blodtryk bliver:

1 ☐ $0.49 \pm 1.96\sqrt{\frac{0.49107 \cdot (1-0.49107)}{112}}$ som giver 0.49 ± 0.09

2 ☐ $0.49 \pm 1.645\sqrt{\frac{0.49107 \cdot (1-0.49107)}{55}}$ som giver 0.49 ± 0.11

3 ☐ $0.27 \pm 1.645\sqrt{\frac{0.26667 \cdot (1-0.26667)}{112}}$ som giver 0.27 ± 0.07

4 ☐ $0.27 \pm 1.96\sqrt{\frac{0.49107 \cdot (1-0.49107)}{75}}$ som giver 0.27 ± 0.09

5 ☐ $0.27 \pm 1.96\sqrt{\frac{0.26667 \cdot (1-0.26667)}{75}}$ som giver 0.27 ± 0.10

Spørgsmål IV.2 (11)

Hvilken af følgende beregnede værdier er en sædvanlig teststørrelse for et test af hypotesen om, at andelen af personer med højt blodtryk i gruppe I og II er ens?

1 ☐ $1.96^2 = 3.84$

2 ☐ $\frac{1}{56} + \frac{1}{19} = 0.07$

3 ☐ $\frac{1}{56} + \frac{1}{56} + \frac{1}{19} + \frac{1}{19} = 0.14$

4 ☐ $\frac{55-57}{\sqrt{150}} = 0.16$

5 ☐ $\frac{(112-38)^2}{150} = 36.5$

Spørgsmål IV.3 (12)

Antag, at man havde sammenlignet andelene mellem 3 diætgrupper, altså haft en datatabel på følgende form (f.eks. betegner x_{32} således antallet af personer i Gruppe III, der havde højt blodtryk):

Gruppe	Normalt blodtryk	Højt blodtryk
I	x_{11}	x_{12}
II	x_{21}	x_{22}
III	x_{31}	x_{32}

Hvilken fordeling ville man bruge i det sædvanlige hypotese-test for en sammenligning af de tre andele (dvs. andelene af personer i hver af de 3 grupper med højt blodtryk)?

1 ☐ En χ^2-fordeling med 3 frihedsgrader

2 ☐ En F-fordeling med 2 og 3 frihedsgrader

3 ☐ En F-fordeling med 2 og 5 frihedsgrader

4 ☐ En χ^2-fordeling med 2 frihedsgrader

5 ☐ En χ^2-fordeling med 1 frihedsgrader

Spørgsmål IV.4 (13)

Dette spørgsmål baseres på teksten og tabellen givet i forrige spørgsmål.

Hvad bliver et forudplanlagt 95% konfidensinterval for forskellen mellem andelene med normalt blodtryk i Gruppe I og Gruppe III ?

1 ☐ $\frac{x_{11}}{x_{11}+x_{12}} - \frac{x_{31}}{x_{31}+x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})^3} + \frac{x_{31}\cdot x_{32}}{(x_{31}+x_{32})^3}}$

2 ☐ $\frac{x_{11}}{x_{12}} - \frac{x_{31}}{x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{12})^3} + \frac{x_{31}\cdot x_{32}}{(x_{31}+x_{32})^3}}$

3 ☐ $\frac{x_{11}}{x_{12}} - \frac{x_{31}}{x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{x_{11}+x_{12}} + \frac{x_{31}\cdot x_{32}}{x_{31}+x_{32}}}$

4 ☐ $\frac{x_{11}}{x_{11}+x_{21}+x_{31}} - \frac{x_{31}}{x_{11}+x_{21}+x_{31}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{21}+x_{31})^2} + \frac{x_{31}\cdot x_{32}}{(x_{11}+x_{21}+x_{31})^2}}$

5 ☐ $\frac{x_{11}}{x_{11}+x_{12}} - \frac{x_{31}}{x_{31}+x_{32}} \pm 1.96\sqrt{\frac{x_{11}\cdot x_{12}}{(x_{11}+x_{21}+x_{31})^2} + \frac{x_{31}\cdot x_{32}}{(x_{11}+x_{21}+x_{31})^2}}$

Spørgsmål IV.5 (14)

Et nyt studie planlægges for at undersøge en helt ny diæt. Man kunne godt ønske sig, at 90% konfidensintervallet for andelen af personer med normalt blodtryk får en middelbredde på 0.1. Den samlede omkostning for håndteringen af en enkelt person er 100 kr, og der er foreløbig afsat 25000 kr til dette i budgettet for studiet. Kan ønsket opfyldes inden for det givne budget (Man ved altså ingenting om hvilken størrelsesorden denne andel har)?

1 ☐ Ja, idet $100 \cdot \left(\frac{1.96}{0.05}\right)^2 = 153664 < 25000$

2 ☐ Ja, idet $100 \cdot 0.3 \cdot \cdot 0.7 \cdot \left(\frac{1.645}{0.05}\right)^2 = 22730.61 < 25000$

3 ☐ Nej, idet $\frac{1}{4} \cdot \left(\frac{1.96}{0.05}\right)^2 \approx 384$

4 ☐ Ja, idet $\frac{1}{4} \cdot \left(\frac{1.96}{0.05}\right)^2 \approx 384$

5 ☐ Nej, idet $100 \cdot \frac{1}{4} \cdot \left(\frac{1.645}{0.05}\right)^2 = 27060.25 > 25000$

Fortsæt på side 13

Opgave V

Spørgsmål V.1 (15)

Lad $X_i \sim N(\mu, \sigma^2)$, $i = 1, ..., n$. Hvis $n = 10$ hvad er så

$$P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} > 2\right)$$

hvor S^2 og $\overline{X}$ er hhv. empirisk varians og middelværdi (stokastiske variable)?

1 ☐ $6.8 \cdot 10^{-5}$

2 ☐ 0.038

3 ☐ 0.012

4 ☐ 0.988

5 ☐ 0.962

Fortsæt på side 14

Opgave VI

Spørgsmål VI.1 (16)

En multiple choice eksamen har 30 spørgsmål og 5 svarmuligheder for hvert spørgsmål. Der er et og kun et korrekt svar på hvert spørgsmål. Hvad er sandsynligheden for at svare rigtigt på mindst 15 ud af de 30 spørgsmål, hvis man svarer helt tilfældigt?

1 ☐ 0.000119

2 ☐ 0.00023

3 ☐ 0.835

4 ☐ 0.999

5 ☐ $4.5 \cdot 10^{-9}$

Spørgsmål VI.2 (17)

Hvilken af følgende stokastiske variable bør ikke approksimeres med en normalfordeling?

1 ☐ $\sum_{i=1}^{200} X_i,\ X_i \sim Pois(2)$

2 ☐ $\sum_{i=1}^{5} X_i,\ X_i \sim N(\mu, \sigma^2)$

3 ☐ $\sum_{i=1}^{10} X_i,\ X_i \sim Binom(1000, 0.5)$

4 ☐ $\sum_{i=1}^{100} X_i,\ X_i \sim Exp(1)$

5 ☐ $\sum_{i=1}^{7} X_i,\ X_i \sim Exp(1)$

Fortsæt på side 15

Opgave VII

Levetiden for en bestemt type elektroniske knapper antages at følge en eksponentialfordeling med middelværdi på 5 år.

Spørgsmål VII.1 (18)

Hvad er variansen af levetiden for sådanne knapper?

1 ☐ 1 år

2 ☐ 1/5 år

3 ☐ 5 år

4 ☐ 25 år

5 ☐ 1/25 år

Spørgsmål VII.2 (19)

Hvis 10 af disse knapper er installeret i forskellige systemer (uden indbyrdes påvirkning), hvad er da sandsynligheden for at ingen af disse bryder sammen inden for det første år?

1 ☐ 0.1353

2 ☐ 0.4350

3 ☐ 0.1074

4 ☐ 0.3758

5 ☐ 0.6241

Fortsæt på side 16

Opgave VIII

I et studie ønsker man at undersøge 4 forskellige metoder (A, B, C, D) til at fjerne urenheder i en kemisk proces. Samtidig ønsker man at justere for, at der bruges 3 forskellige reaktorer i studiet. Data ses i nedenstående tabel:

	Reaktor 1	Reaktor 2	Reaktor 3	Sum
Metode A	23.97	29.54	37.91	91.42
Metode B	12.67	17.48	20.28	50.43
Metode C	25.85	40.09	38.00	103.94
Metode D	21.29	23.58	20.19	65.06
Sum	83.78	110.69	116.38	310.85

Man har beregnet kvadratafvigelsessummerne:

	SS
Reaktor	151.61
Metode	593.40
Residual	100.73
Total variation	845.74

Spørgsmål VIII.1 (20)

Vi ønsker nu at undersøge, om det er rimeligt at antage, at de fire metoder i middel renser lige godt. Ved hjælp af ovenstående kvadratafvigelsersummer findes den sædvanlige teststørrelse, her kaldet A, og det kritiske niveau, når vi har signifikansniveau $\alpha = 0.05$ til:

1 ☐ $A = 5.89$ og $A < 8.94$ (d.v.s. ingen signifikant effekt af Metode)

2 ☐ $A = 11.78$ og $A > 4.76$ (d.v.s. en signifikant effekt af Metode)

3 ☐ $A = 441.79$ og $A > 4.76$ (d.v.s. en signifikant effekt af Metode)

4 ☐ $A = 0.70$ og $A < 8.94$ (d.v.s. ingen signifikant effekt af Metode)

5 ☐ $A = 3.91$ og $A < 4.76$ (d.v.s. ingen signifikant effekt af Metode)

Spørgsmål VIII.2 (21)

For det udførte forsøg vil man ofte skrive en to-sidet variansanalyse model op som:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2),\ i = (1, 2, 3, 4),\ j = (1, 2, 3)$$

Her er y_{ij} den observerede renhed for metode i reaktor j, μ middelværdi for alle målinger, α metode-effekt og β er reaktor-effekt.

Hvad er det sædvanlige estimat for effekten af metode B (d.v.s. $\hat{\alpha}_2$) i modellen ovenfor?

1 ☐ $\hat{\alpha}_2 = 50.43$

2 ☐ $\hat{\alpha}_2 = -260.42$

3 ☐ $\hat{\alpha}_2 = 16.81$

4 ☐ $\hat{\alpha}_2 = 4.20$

5 ☐ $\hat{\alpha}_2 = -9.09$

Spørgsmål VIII.3 (22)

Inden forsøget blev udført besluttede man, at man specielt gerne ville sammenligne metode B og C.

Hvad er 95% konfidensintervallet for sammenligningen mellem B og C, hvis man kun har tænkt sig at udføre denne post-hoc sammenligning?

1 ☐ $(50.43 - 103.94)/3 \pm 2.3060\sqrt{\frac{100.73}{12-4}(2/3)} = [-24.52, -11.16]$

2 ☐ $(50.43 - 103.94) \pm 2.4469\sqrt{\frac{100.73}{6}(2/3)} = [-61.70, -45.32]$

3 ☐ $(50.43 - 103.94)/3 \pm 2.4469\sqrt{\frac{100.73}{6}(1/3 + 1/4)} = [-25.49, -10.18]$

4 ☐ $(50.43 - 103.94)/3 \pm 2.4469\sqrt{\frac{100.73}{6}(2/3)} = [-26.02, -9.65]$

5 ☐ $(50.43 - 103.94)/3 \pm 2.3060\sqrt{\frac{100.73}{12-4}(1/3 + 1/4)} = [-24.09, -11.59]$

Spørgsmål VIII.4 (23)

Hvis man alligevel beslutter at lave alle parvise sammenligninger mellem metoderne, hvad bliver så den Bonferroni korrigerede "least significant difference (LSD)"?

1 ☐ $3.8630\sqrt{2 \cdot 16.788/3} = 12.92$

2 ☐ $3.4789\sqrt{2 \cdot 12.591/3} = 10.08$

3 ☐ $2.4469\sqrt{2 \cdot 16.788/6} = 5.79$

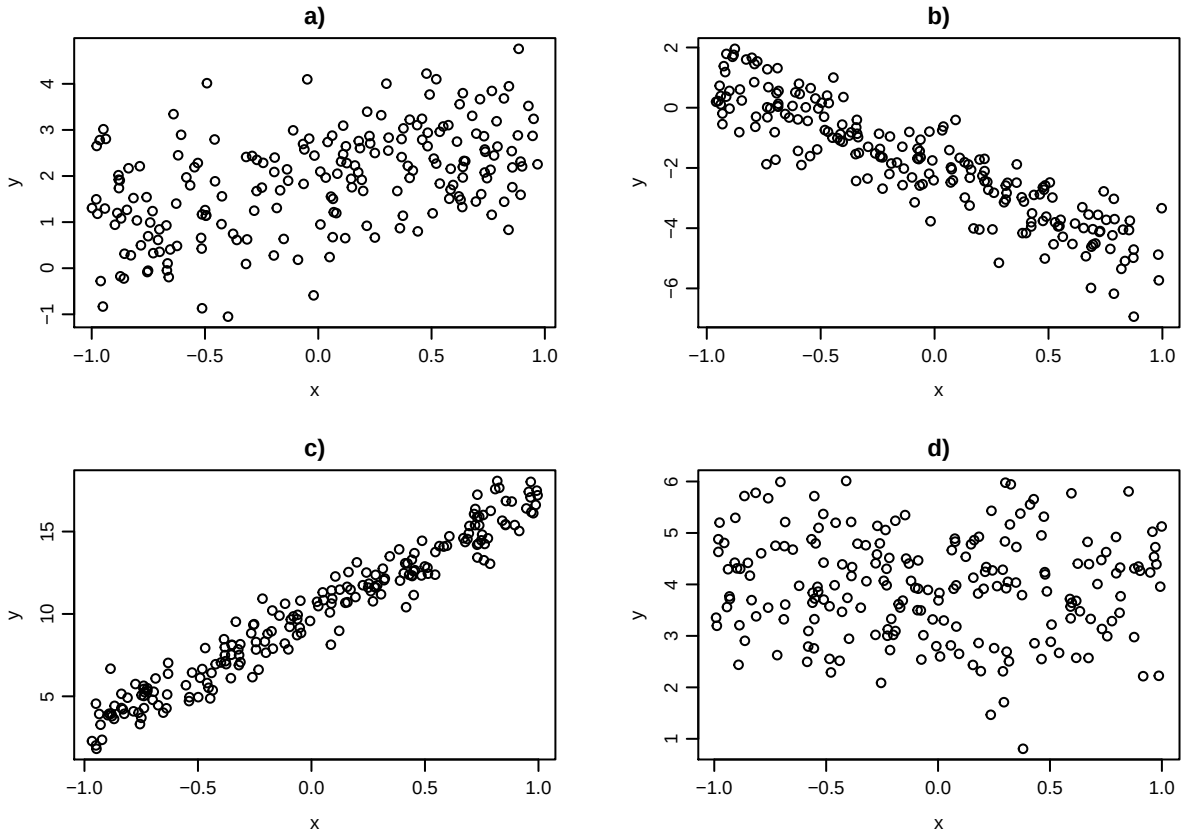
4 ☐ $3.8630\sqrt{2 \cdot 12.591/3} = 11.19$

5 ☐ $2.4469\sqrt{2 \cdot 16.788/3} = 8.19$

Fortsæt på side 18

Opgave IX

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



Spørgsmål IX.1 (24)

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

1 ☐ 0.9, -0.5, 0.65, 0

2 ☐ 0.5, -0.9, 0.97, 0

3 ☐ 0.5, -0.9, 0.65, 0

4 ☐ 0.5, 0.97, 0, -0.9

5 ☐ 0.97, -0.9, 0, 0.5

Spørgsmål IX.2 (25)

Ud fra inspektion af sammenhængen i plot `c)` i figuren, hvilke estimater for parametrene i den sædvanlige simple lineære regressionsmodel $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ hvor $\varepsilon_i \sim N(0, \sigma^2)$ passer da bedst?

1 ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 5$ og $\hat{\sigma} = 10$

2 ☐ $\hat{\beta}_0 = 16$, $\hat{\beta}_1 = -5$ og $\hat{\sigma} = 1$

3 ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 7$ og $\hat{\sigma} = 10$

4 ☐ $\hat{\beta}_0 = -10$, $\hat{\beta}_1 = -7$ og $\hat{\sigma} = 5$

5 ☐ $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 7$ og $\hat{\sigma} = 1$

Fortsæt på side 20

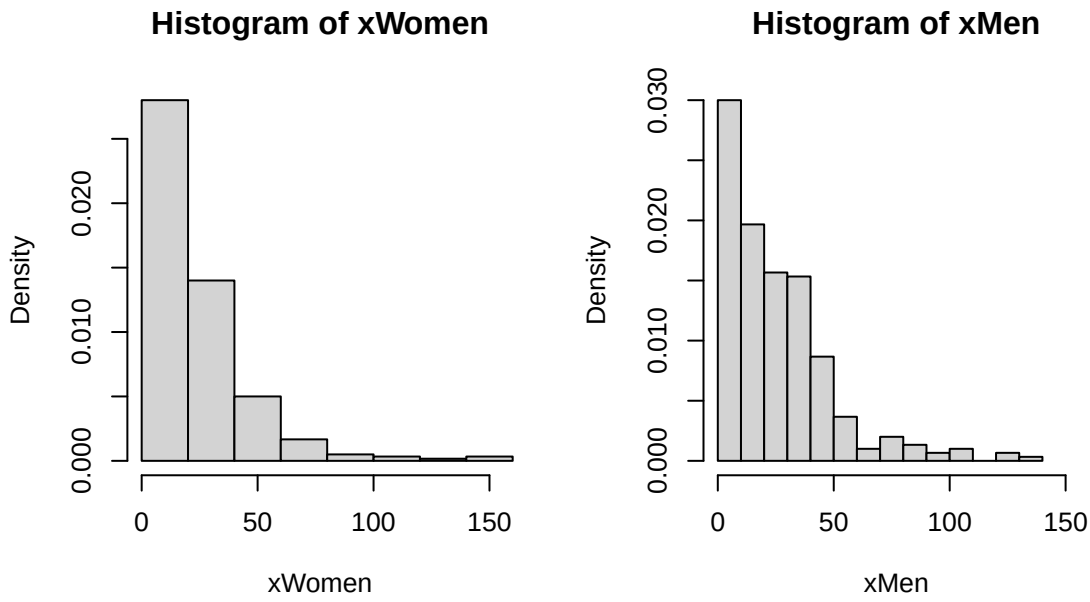
Opgave X

DTU Fødevareinstituttet (DTU Food) overvåger og udarbejder analyser af danskernes kostvaner. Det har de gjort siden 1985 på opfordring fra Miljø- og Fødevareministeriet, og resultaterne indhentes på baggrund af stikprøveundersøgelser. Resultaterne er afsæt til at justere de officielle kostråd og grundlag for til en politisk prioritering af den forebyggende indsats.

I undersøgelserne er respondenter blevet spurgt, hvor meget fisk de spiser og det er opgjort i gram fisk pr. dag. En tilfældig stikprøve omfattende 300 kvindlige respondenter fra kostundersøgelsen 2005-2008, samt desuden en tilfældig stikprøve omfattende 300 mandlige respondenter fra samme kostundersøgelse, er blevet udtaget.

Observationerne er indlæst i en vektor for kvinder `xWomen` og en vektor for mænd `xMen`. Den empiriske tæthed (tæthedshistogram) for observationerne for hvert køn er plottet:

```
par(mfrow=c(1,2), cex=0.8)
hist(xWomen, prob=TRUE, xlim=range(xWomen,xMen))
hist(xMen, prob=TRUE, xlim=range(xWomen,xMen))
```



Det antages at forudsætningen om normalfordeling ikke er opfyldt. Man vil derfor ikke lave nogen fordelingsantagelse.

Spørgsmål X.1 (26)

For at bestemme et 95% konfidensinterval for middelindtag af fisk per dag for kvinder i befolkningen med baggrund i stikprøven for 2005-2008 kostundersøgelsen, har man kørt følgende R-kode (hvoraf alt ikke nødvendigvis er meningsfuldt):

```r
## Antal resamplinger ved simulering
k <- 10000

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
##  19.0  24.3

simsamples <- replicate(k, rnorm(length(xWomen),
                                 mean=log(mean(xWomen)), sd=sd(xWomen)))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
## 0.348 5.777

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, median)
quantile(simprms, c(0.005,0.995))

##  0.5% 99.5%
##  11.1  20.6

simsamples <- replicate(k, sample(xMen, replace=TRUE))
simprms <- apply(simsamples, 2, median)
quantile(simprms, c(0.005,0.995))

##  0.5% 99.5%
##  17.0  25.2

simsamples <- replicate(k, runif(100,0,40))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.025,0.975))

##  2.5% 97.5%
##  17.7  22.3
```

Baseret på ovenstående, hvad er da et korrekt 95% konfidensinterval for middelindtag af fisk per dag for kvinder i befolkningen?

1 ☐ $[19.0, 24.3]$

2 ☐ $[0.348, 5.777]$

3 ☐ $[11.1, 20.6]$

4 ☐ $[17.0, 25.2]$

5 ☐ $[17.7, 22.3]$

Spørgsmål X.2 (27)

Bemærk, at en normalfordelingsantagelse om indtaget af fisk er ikke opfyldt og der findes ikke rimelige transformationer, som sikrer en sådan antagelse.

I det følgende ønsker vi at undersøge om mænd og kvinder spiser lige meget fisk eller om der kan konstateres signifikante forskelle, svarende til følgende hypotese:

$$H_0 : q_{0.5,\text{male}} = q_{0.5,\text{female}}$$
$$H_1 : q_{0.5,\text{male}} \neq q_{0.5,\text{female}}$$

hvor $q_{0.5,gender}$ betegner 50% fraktilen for det angivne køn.

For at teste hypotesen har man kørt følgende R-kode (bemærk alt er muligvis ikke meningsfuldt):

```
t.test(xMen-xWomen)

##
##  One Sample t-test
##
## data:  xMen - xWomen
## t = 2.0508, df = 299, p-value = 0.04115
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.1553667 7.5300188
## sample estimates:
## mean of x
##  3.842693

t.test(xMen, xWomen)

##
##  Welch Two Sample t-test
##
## data:  xMen and xWomen
## t = 1.9908, df = 597.95, p-value = 0.04695
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05193859 7.63344693
## sample estimates:
## mean of x mean of y
##  25.37784  21.53515

simprms <- replicate(k, median(sample(xMen, replace=TRUE))-
                     median(sample(xWomen, replace=TRUE)))
quantile(simprms, c(0.025, 0.975))
```

```
##   2.5%  97.5%
## 0.0354 9.8422

simprms <- replicate(k,
                     mean(rnorm(length(xMen),mean(xMen),sd(xMen))) -
                     mean(rnorm(length(xWomen),mean(xWomen),sd(xWomen))))
quantile(simprms, c(0.025, 0.975))

##    2.5%   97.5%
## -0.0544  7.5434

median(xMen) > median(xWomen)

## [1] TRUE
```

Idet vi benytter signifikansniveau $\alpha = 0.05$, hvad er da konklusionen på hypotesetestet (både konklusion og argument skal være korrekt)?

1 ☐ Da p-værdien $= 0.04115 < 0.05$, kan det afvises at $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

2 ☐ Da $0 \in [-0.0544, 9.842]$ kan det ikke afvises at $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

3 ☐ Da $0 \notin [0.0354, 7.714]$ kan det afvises at $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

4 ☐ Da $0 \in [-0.0544, 7.543]$ kan det ikke afvises at $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

5 ☐ Da $\hat{q}_{0.5,\text{male}} > \hat{q}_{0.5,\text{female}}$ kan det afvises at $q_{0.5,\text{male}} = q_{0.5,\text{female}}$

Spørgsmål X.3 (28)

Blandt de 10 officielle kostråd fra Fødevarestyrelsen er det ene, at man skal spise meget fisk og helst 350 g om ugen. Dette gælder både mænd og kvinder.

I det følgende vil vi undersøge, om stikprøven fra kostundersøgelsen 2005-2008 giver anledning til at vurdere, at kostrådet er opfyldt. Da kostundersøgelsen opgør det daglige indtag af fisk, svarer det lidt forenklet til følgende hypotese:

$$H_0 : \mu = 50$$
$$H_1 : \mu \neq 50$$

Data er ikke normalfordelt, hvilket især skyldes de mange 0-observationer, dvs. respondenter der ikke spiser fisk. For at teste hypotesen har man kørt følgende R-kode (bemærk at alt er muligvis ikke meningsfuldt):

```r
## Antal resamplinger ved simulering
k <- 10000

simsamples <- replicate(k, sample(xMen, replace=TRUE))
simprms <- apply(simsamples, 2, mean)
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
##  22.0  22.8  28.1  29.0

simsamples <- replicate(k, sample(xWomen, replace=TRUE))
simprms <- apply(simsamples, 2, sd) - 50
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
## -31.7 -30.5 -22.0 -20.7

simsamples <- replicate(k, sample(c(xMen,xWomen), replace=TRUE))
simprms <- apply(simsamples, 2, mean) - 50
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
## -29.0 -28.4 -24.6 -23.9

simsamples <- replicate(k, sample(c(xMen,xWomen), replace=TRUE))
simprms <- apply(simsamples, 2, quantile, probs=0.90)
quantile(simprms, c(0.005, 0.025, 0.975, 0.995))

##  0.5%  2.5% 97.5% 99.5%
##  44.7  45.8  54.5  57.4
```

Idet vi benytter signifikansniveau $\alpha = 0.05$, hvad er da konklusionen på, om kostrådet vedr. fisk bliver fulgt (både konklusion og argument skal være korrekt)?

1 ☐ Da $0 \notin [-28.4, -24.6]$, forkastes nulhypotesen, dvs. der er signifikant forskel - altså $\mu \neq 50$ - kostrådet er ikke opfyldt

2 ☐ Da $-50 \notin [-30.5, -22.0]$, forkastes nulhypotesen, dvs. der er signifikant forskel - altså $\mu \neq 50$ - kostrådet er ikke opfyldt

3 ☐ Da $50 \in [45.8, 54.5]$ kan det ikke afvises at $\mu \neq 50$ - dvs. kostrådet er opfyldt

4 ☐ Da $50 \notin [22.0, 29.0]$ forkastes nulhypotesen, dvs. der er signifikant forskel - altså $\mu \neq 50$ og vi kan således slutte, at kostrådet ikke er opfyldt

5 ☐ Idet $44.7 - 50 = -5.3 < 0$ forkastes nulhypotesen, dvs. der er signifikant forskel - altså $\mu \neq 50$ og vi kan således slutte, at kostrådet ikke er opfyldt

Fortsæt på side 25

Opgave XI

En ny scanner til bestemmelse af muskelmassen på kroppen er blevet udviklet. Den er meget nemmere at betjene og den tager kortere tid at udføre en scanning med, end de ellers tilgængelige scannere. Den skal testes i et forsøg for at se, om den får samme resultater, som den normalt anvendte scanner. Til forsøget er muskelmassen for 20 tilfældigt udvalgte kvinder i alderen 20 til 40 år blevet scannet med begge scannere.

De målte muskelmasser i kg er indlæst i R, således at kvindernes rækkefølge er den samme i hver vektor:

```
## Stikprøve fra den nye scanner
x1 <- c(37.6, 31.3, 22.9, 27.1, 41.8, 23.3, 24.5, 24.6, 32.1, 23.8, 33.9, 37.7,
        22.5, 38.6, 31.8, 21.0, 32.2, 17.1, 32.6, 15.5)
## Stikprøve fra den gamle scanner
x2 <- c(35.9, 28.7, 27.9, 29.8, 46.8, 24.2, 28.0, 23.7, 35.2, 26.4, 36.0, 40.9,
        24.8, 42.1, 32.5, 23.7, 36.7, 19.2, 37.7, 16.3)
```

og følgende R kode er kørt:

```
(mean(x1)-mean(x2)) + c(-1,1) * qt(0.975, df=38) * sd(x2-x1)/sqrt(40)
```

```
## [1] -2.9228 -1.5372
```

```
t.test(x1, x2)
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -0.916, df = 37.8, p-value = 0.37
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.1591  2.6991
## sample estimates:
## mean of x mean of y
##    28.595    30.825
```

```
t.test(x1, x2, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  x1 and x2
## t = -4.61, df = 19, p-value = 0.00019
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.2429 -1.2171
## sample estimates:
## mean of the differences
##                   -2.23
```

```r
t.test(x2-mean(x1))
```

```
##
##  One Sample t-test
##
## data:  x2 - mean(x1)
## t = 1.25, df = 19, p-value = 0.23
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.5043  5.9643
## sample estimates:
## mean of x
##      2.23
```

```r
t.test(x1-mean(x2))
```

```
##
##  One Sample t-test
##
## data:  x1 - mean(x2)
## t = -1.35, df = 19, p-value = 0.19
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -5.6965  1.2365
## sample estimates:
## mean of x
##     -2.23
```

Spørgsmål XI.1 (29)

Hvad er det korrekte 95% konfidensinterval for middelforskellen i måling af muskelmasse mellem den gamle og den nye scanner (her afrundet til 3 betydende cifre)?

1 ☐ [-2.92, -1.54]

2 ☐ [-7.16, 2.70]

3 ☐ [-3.24, -1.22]

4 ☐ [-1.50, 5.96]

5 ☐ [-5.70, 1.24]

Spørgsmål XI.2 (30)

Resultatet af en scanning afhænger bl.a. af, hvor meget personen som skal scannes, bevæger sig. I den nye scanner er personen ikke fastspændt, så det er af interesse at finde ud af, hvor stor variationen af resultaterne er. For at undersøge variationen blev de 20 kvinder scannet 2 gange hver med den nye scanner og forskellen i muskelmasse (X_Δ) for hver kvinde mellem de 2 scanninger blev målt til:

i	1	2	3	4	5	6	7	8	9	10
$x_{\Delta,i}$	-0.62	1.12	0.24	2.07	-2.91	2.02	0.36	0.43	-1.77	-0.18
i	11	12	13	14	15	16	17	18	19	20
$x_{\Delta,i}$	1.02	0.85	0.43	1.39	2.82	-4.03	2.84	2.8	1.36	-0.07

Stikprøvegennemsnittet og -standardafvigelsen er beregnet til

$$\bar{x}_\Delta = 0.509$$
$$s_{x_\Delta} = 1.82$$

Hvilket af følgende er et korrekt 99% konfidensinterval for standardafvigelsen på måling af muskelmasse med den nye scanner?

1 ☐ $0.509 \pm 2.86 \cdot \frac{1.35}{\sqrt{20}} = [-0.35, 1.37]$

2 ☐ $\left[\frac{20 \cdot 1.82^2}{38.6},\ \frac{20 \cdot 1.82^2}{6.84}\right] = [1.72, 9.69]$

3 ☐ $0.509 \pm 2.09 \cdot \frac{1.35}{\sqrt{20}} = [-0.12, 1.14]$

4 ☐ $0.509 \pm 2.86 \cdot \frac{1.82}{\sqrt{19}} = [-0.69, 1.70]$

5 ☐ $\left[\sqrt{\frac{19 \cdot 1.82^2}{38.6}},\ \sqrt{\frac{19 \cdot 1.82^2}{6.84}}\right] = [1.28, 3.03]$

SÆTTET ER SLUT. God sensommer!