Danmarks Tekniske Universitet *Side 1 af 27 sider.*

Skriftlig prøve: 15. Dec. 2024

Kursus navn og nr.: **02402 Statistik (Polyteknisk grundlag)**

Varighed: 4 timer

Tilladte hjælpemidler: Alle, undtagen adgang til Internet

Dette sæt er besvaret af

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

Opgavesættet består af 30 spørgsmål af "multiple choice" typen, som er fordelt på 13 opgaver. For at besvare spørgsmålene skal du udfylde "multiple choice" siderne på eksamen.dtu.dk.

Der gives 5 point for et korrekt "multiple choice" svar og -1 point for et forkert svar. KUN følgende 5 svarmuligheder er gyldige: 1, 2, 3, 4 eller 5. Hvis et spørgsmål efterlades blankt eller et ugyldigt svar angives, gives der 0 point for spørgsmålet. Endvidere, hvis mere end et svar angives til det samme spørgsmål, hvilket faktisk er teknisk muligt i online-systemet, gives der 0 point for spørgsmålet. Det antal point der kræves, for at opnå en bestemt karakter eller for at bestå eksamen afgøres endeligt ved censureringen.

Den endelige besvarelse af opgaverne laves ved at udfylde og aflevere online.
Skemaet her er KUN et nød-alternativ til dette.
Husk at angive dit studienummer, hvis du afleverer på papir.

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

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

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

Eksamenssættet består af 27 sider.

Fortsæt på side 2

Brug af Python til denne eksamen: *Denne version er Python versionen af eksamenssættet. Der findes også en R-version.*
Bemærk at vi bruger de følgende Python biblioteker og forkortelser i al Python kode i denne eksamen. Vi anbefaler at du kopierer følgende ind i din egen kode.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.stats.power as smp
import statsmodels.stats.proportion as smprop
```

Vær opmærksom på at særlige tegn ("~", "_", "^", osv.) ikke altid kopieres korrekt hvis du vælger at copy paste fra eksamenssættet. Hvis du får fejl (error messages) i din kode så tjek gerne at særlige tegn står korrekt (du får måske brug for at genindtaste dem manuelt).

Multiple choice opgaver: *Der gøres opmærksom på, at der i hvert spørgsmål er én og kun én korrekt svarmulighed. 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. Husk også, at der kan forekomme små afvigelser mellem resultatet af bogens formler og tilsvarende indbyggede funktioner i Python.*

Opgave I

Et hold af forskere evaluerer en deterministisk simuleringsmodel ved at sammenligne modelsimuleringerne med eksperimentelle resultater. Forskerne overvejer to faktorer: last (kg) og hastighed (knob). Forskerne foreslår følgende model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

hvor fejlene antages at være uafhængige og normalfordelte med $E[\varepsilon_{ij}] = 0$ og $V[\varepsilon_{ij}] = \sigma^2$. I modellen er Y_{ij} forskellen mellem de simulerede og eksperimentelle resultater opnået ved brug af lastniveau i og hastighedsniveau j, og derfor refererer parametrene α_i og β_j til henholdsvis last- og hastighedsvirkningerne (effects). Tabellen nedenfor viser de opnåede forskelle (eksperimentelt resultat minus simuleringsresultat):

	5 knob	10 knob	25 knob	50 knob
100 kg	-33.72	-26.95	29.11	-38.87
200 kg	-5.75	-3.00	-15.41	20.56
300 kg	29.96	-24.77	-12.05	1.52
400 kg	-4.72	5.72	24.39	43.16
500 kg	-22.36	23.99	-24.17	33.36

Data kan indlæses i Python ved hjælp af følgende kode:

```
df = pd.DataFrame({
'y': [-33.72, -26.95,  29.11, -38.87,
       -5.75,  -3.00, -15.41,  20.56,
       29.96, -24.77, -12.05,   1.52,
       -4.72,   5.72,  24.39,  43.16,
      -22.36,  23.99, -24.17,  33.36],
'knob': pd.Categorical([5, 10, 25, 50,
                        5, 10, 25, 50,
                        5, 10, 25, 50,
                        5, 10, 25, 50]),
'load': pd.Categorical([100, 100, 100, 100,
                        200, 200, 200, 200,
                        300, 300, 300, 300,
                        400, 400, 400, 400,
                        500, 500, 500, 500]),
})
```

Spørgsmål I.1 (1)

Hvad er parameterestimatet $\hat{\alpha}_3$ (dvs. for lastniveauet "300 kg")?

1 ☐ -1.335

2 ☐ -0.900

3 ☐ 0.374

4 ☐ 2.705

5 ☐ 17.138

Spørgsmål I.2 (2)

Ifølge modellen er SS(last) 2454.51, SS(hastighed) er 1107.10, og den totale kvadratafvigelsessum er 11867.74. Hvad er middelkvadratafvigelsen for fejlen (MSE)?

1 ☐ 415.3

2 ☐ 692.2

3 ☐ 2076.5

4 ☐ 2768.7

5 ☐ 8306.1

Spørgsmål I.3 (3)

Forskerne forkaster de eksperimentelle resultater på grund af en teknisk fejl. Når de gentager eksperimentet, finder de følgende parameterestimater:

Parameter	α_1	α_2	α_3	α_4	α_5
Estimate	1.00	2.00	3.00	4.00	5.00

Parameter	β_1	β_2	β_3	β_4	μ
Estimate	0.25	1.00	3.13	5.00	0.00

Hvad er MS(last) i henhold til de nye parameterestimater?

1 ☐ 13.75

2 ☐ 35.83

3 ☐ 55.00

4 ☐ 220.00

5 ☐ Størrelsen kan ikke bestemmes uden at kende det komplette datasæt.

Fortsæt på side 5

Opgave II

I et 'bestået'/'ikke bestået' kursus blev en klasse på $n = 30$ studerende evalueret. Resultaterne er vist herunder. En score på 0 indikerer 'ikke bestået' og en score på 1 indikerer 'bestået'.

1	0	1	1	1	0	1	1	1	1
0	0	0	0	1	1	0	1	1	1
1	1	1	1	1	1	0	0	1	1

Data kan indlæses i Python med følgende kode:

```
data = np.array([1,0,1,0,0,1,1,0,1,1,0,1,1,1,1,0,1,1,1,0,0,1,1,0,1,1,1,1,1,1])
```

Spørgsmål II.1 (4)

Hvad er den estimerede sandsynlighed for at bestå kurset inklusiv sandsynlighedens 95% konfidensinterval. Det er forudsat at de sædvanlige antagelser er opfyldt (Bemærk: Resultatet er baseret på formlen i bogen, men hvis konfidensintervallet udregnes ved brug af indbyggede funktioner i Python, kan dette give et lidt anderledes resultat).

1 ☐ $\hat{p} = 0.70$ og $[0.49, 0.91]$

2 ☐ $\hat{p} = 0.70$ og $[0.54, 0.86]$

3 ☐ $\hat{p} = 0.76$ og $[0.57, 0.95]$

4 ☐ $\hat{p} = 0.70$ og $[0.61, 0.79]$

5 ☐ $\hat{p} = 0.76$ og $[0.51, 0.89]$

Spørgsmål II.2 (5)

Hvad er den estimerede standardafvigelse for $\hat{p}$, hvis "Plus-2"tilgangen bliver brugt i beregningen af konfidensintervallet?

1 ☐ $\hat{\sigma}_{\hat{p}} = 0.0786$

2 ☐ $\hat{\sigma}_{\hat{p}} = 0.0802$

3 ☐ $\hat{\sigma}_{\hat{p}} = 0.0868$

4 ☐ $\hat{\sigma}_{\hat{p}} = 0.0883$

5 ☐ $\hat{\sigma}_{\hat{p}} = 0.0918$

Fortsæt på side 6

Opgave III

I et studie, der undersøgte smagen mellem almindelig og koffeinfri kaffe, har en smagstester 4 kopper indeholdende kaffe. Hver kop indeholder enten almindelig eller koffeinfri kaffe. Smagstesteren ved, at der er to kopper af hver. Smagstesteren valgte to kopper tilfældigt.

Spørgsmål III.1 (6)

Hvad er sandsynligheden for, at smagstesteren valgte almindelig kaffe i den ene kop og koffeinfri kaffe i den anden (uden at tage tage hensyn til rækkefølgen)?

1 ☐ 1/4

2 ☐ 1/3

3 ☐ 1/2

4 ☐ 2/3

5 ☐ 3/4

Spørgsmål III.2 (7)

I et andet studie, der undersøger evnen til at opdage smagsforskellen mellem almindelig og koffeinfri kaffe, får 30 deltagere en kop af hver type til at smage. Tidligere undersøgelser tyder på en 85% sandsynlighed ($p = 0.85$) for, at individer kan opdage forskellen mellem almindelig og koffeinfri. Lad Y repræsentere antallet af deltagere ud af 30, der kan skelne mellem de to typer. Hvad er variansen af Y?

1 ☐ $\mathrm{V}(Y) = 5.37$

2 ☐ $\mathrm{V}(Y) = 3.83$

3 ☐ $\mathrm{V}(Y) = 3.11$

4 ☐ $\mathrm{V}(Y) = 2.79$

5 ☐ $\mathrm{V}(Y) = 1.10$

Fortsæt på side 7

Opgave IV

Levetiden for en bestemt type batteri, målt i år, følger en eksponentialfordeling med en middelværdi på 50 år.

Spørgsmål IV.1 (8)

Hvad er sandsynligheden for, at et batteri vil holde mindre end 25 år?

1 ☐ $e^{-\frac{25}{50}}$

2 ☐ $1 - e^{-\frac{50}{25}}$

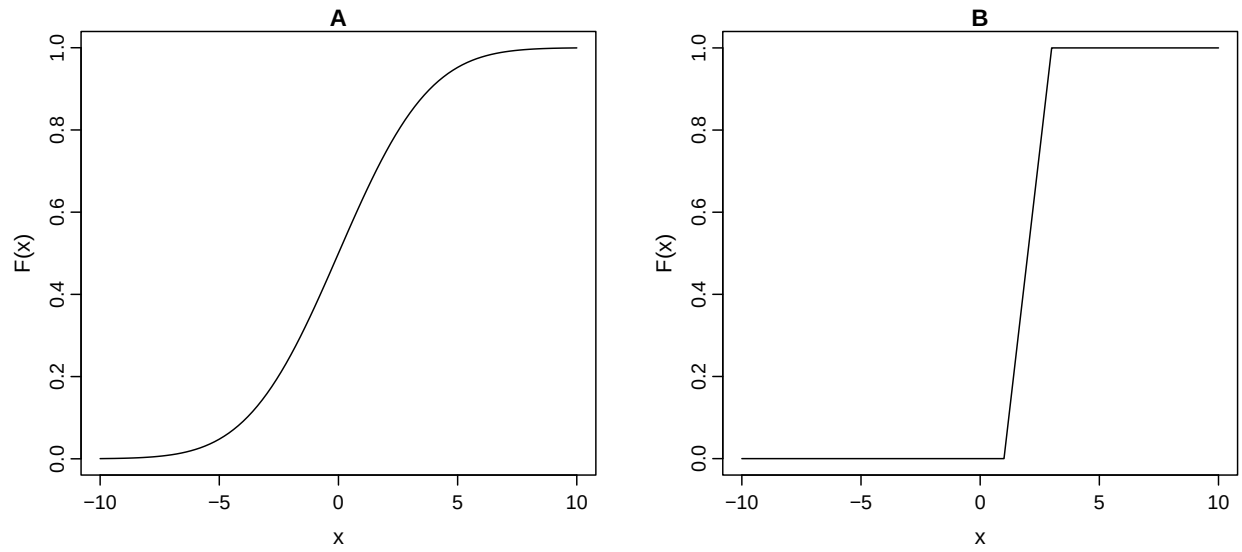
3 ☐ $e^{-\frac{50}{25}}$

4 ☐ $1 - e^{-\frac{25}{50}}$

5 ☐ $e^{-\frac{25}{50}} - e^{-\frac{50}{25}}$

Spørgsmål IV.2 (9)

Nedenfor er to grafer: den ene er en fordelingsfunktion (CDF) for en normalfordeling, og den anden er en fordelingsfunktion (CDF) for en uniform fordeling (ligefordeling).



Et af udsagnene er korrekt, bedømt ud fra graferne, hvilket et er det?

1 ☐ Plot A er en CDF for en uniform fordeling med $a = 3$ og $b = 1$, og plot B er en CDF for en normalfordeling med $\mu = -5$ og $\sigma = 10$.

2 ☐ Plot A er en CDF for en uniform fordeling med $\mu = -5$ og standardafvigelse $\sigma = 10$, og plot B er en CDF for en normalfordeling med $a = 3$ og $b = 1$.

3 ☐ Plot A er en CDF for en normalfordeling med $\mu = -5$ og $\sigma = 10$, og plot B er en CDF for en uniform fordeling med $a = 3$ og $b = 1$.

4 ☐ Plot A er en CDF for en normalfordeling med $\mu = 7$ og $\sigma = 1$, og plot B er en CDF for uniform fordeling med $a = -5$ og $b = 5$.

5 ☐ Plot A er en CDF for en normalfordeling med $\mu = 0$ og $\sigma = 3$, og plot B er en CDF for en uniform fordeling med $a = 1$ og $b = 3$.

Fortsæt på side 9

Opgave V

I et landbrugsstudie undersøger forskere effektiviteten af to forskellige gødninger, A og B, på øgning af afgrødeudbyttet. De vælger tilfældigt 20 jordstykker og anvender gødning A på 10 stykker og gødning B på de resterende 10 stykker. Efter høsten registrerer de udbyttet (i enheden "bushels per acre"$= 6.73g/m^2$) fra hvert jordstykke. Forskerne ønsker at afgøre, om der er en signifikant forskel i middelværdi af udbyttet mellem de to gødninger.

Udbyttedata bliver registreret på følgende måde:

$\text{Fertilizer}_A : 45, 48, 50, 42, 47, 49, 43, 44, 46, 41$
$\text{Fertilizer}_B : 51, 53, 52, 50, 55, 48, 54, 49, 56, 52$

Alle målinger antages at være uafhængige, og udbyttepopulationerne følger normalfordelinger.

Spørgsmål V.1 (10)

Hvad er teststatistikken og 99% konfidensintervallet for forskellen i middelværdi af udbyttet mellem de to gødninger (gødning A minus gødning B) (begge resultater skal være korrekte)?

1 ☐ -5.17, [-8.39, -4.61]

2 ☐ -4.17, [-8.68, -4.32]

3 ☐ -5.76, [-9.14, -3.85]

4 ☐ -5.17, [-9.15, -3.85]

5 ☐ -5.17, [-10.13, -2.87]

Spørgsmål V.2 (11)

Når vi betegner det gennemsnitlige udbytte for gødning A som μ_{A} og det gennemsnitlige udbytte for gødning B som μ_{B}, hvad er så konklusionen for den følgende nulhypotese:

$$H_0 : \mu_{\text{A}} - \mu_{\text{B}} = 0$$

på signifikansniveauet $\alpha = 0.05$ (både konklusion og argument skal være korrekte)?

1 ☐ Nulhypotesen accepteres, da p-værdien er 0.23.

2 ☐ Nulhypotesen afvises, da p-værdien er 0.0023.

3 ☐ Nulhypotesen afvises, da 95% konfidensintervallet indeholder nul.

4 ☐ Nulhypotesen afvises, da 99% konfidensintervallet indeholder nul.

5 ☐ Nulhypotesen afvises, da 95% konfidensintervallet ikke indeholder nul.

Fortsæt på side 10

Opgave VI

En legetøjsbutik sælger glaskugler. Glaskuglerne er ca. samme størrelse med middel diameter (D) 1 cm, men variansen er kun opgivet i forhold til vægt (W): $\sigma_W^2 = 0.03^2$. Glaskuglernes vægt følger en normalfordeling.

Forholdet mellem vægt og diameter er givet ved følgende udtryk

$$W = \rho \cdot \frac{4}{3} \cdot \pi \cdot \left(\frac{D}{2}\right)^3$$

samme forhold kan også skrives på følgende måde

$$D = 2\left(\frac{3W}{4\pi\rho}\right)^{1/3}$$

Hvor $\rho = 2.6$ g/cm^3 er massefylden (lig med massefylden af glas). Du kan bruge $\pi = 3.14$, og $\mu_W = W(\mu_D)$.

Spørgsmål VI.1 (12)

En kunde vil gerne kende standardafvigelsen på diameteren af glaskuglerne (σ_D). Heldigvis har kunden studeret fejlophobning (på engelsk "error propagation") og ved derfor hvordan man approksimerer σ_D fra σ_W. Hvad er standardafvigelsen af diameteren af glaskuglerne?

1 ☐ $\sigma_D = 0.006$ cm

2 ☐ $\sigma_D = 0.086$ cm

3 ☐ $\sigma_D = 0.015$ cm

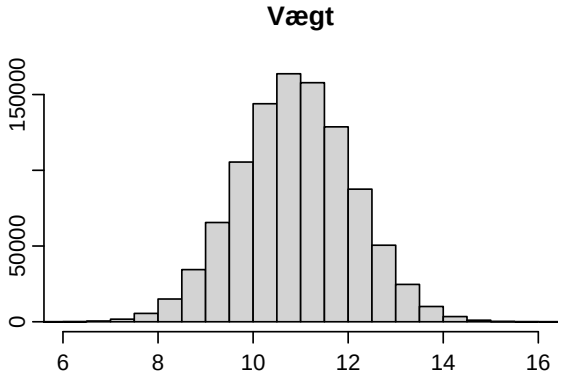
4 ☐ $\sigma_D = 0.007$ cm

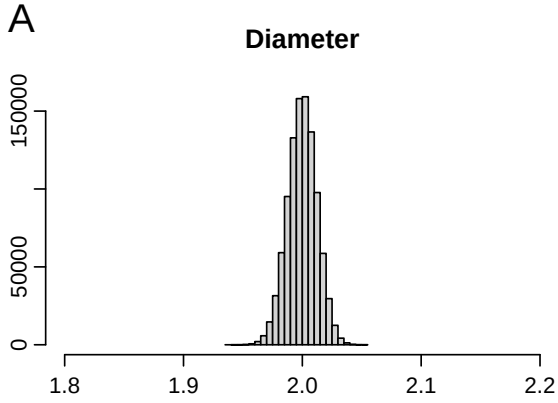
5 ☐ $\sigma_D = 0.04$ cm

Spørgsmål VI.2 (13)

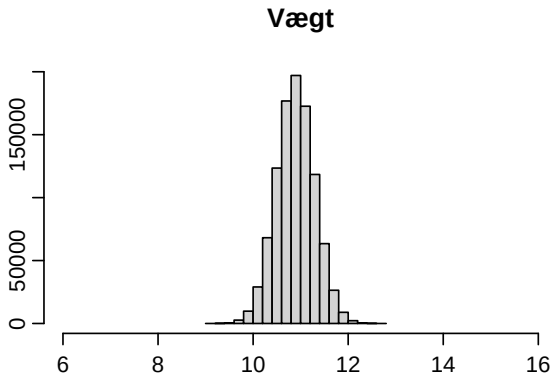
Et andet mærke (på engelsk: "brand") af glaskugler har en gennemsnitlig diameter på 2 cm, $\sigma_W^2 = 0.4^2$ og massefylde $\rho = 2.6$ g/cm^3. Vægten af den enkelte glaskugle kan beregnes som $W = \rho \cdot \frac{4}{3} \cdot \pi \cdot \left(\frac{D}{2}\right)^3$. Hvilket sæt af histogrammer passer til glaskuglerne fra det andet mærke?

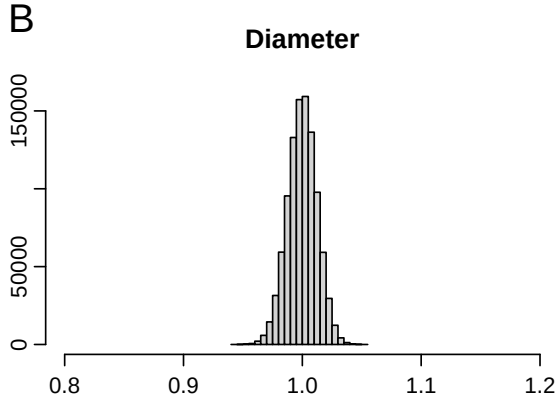
Plot A



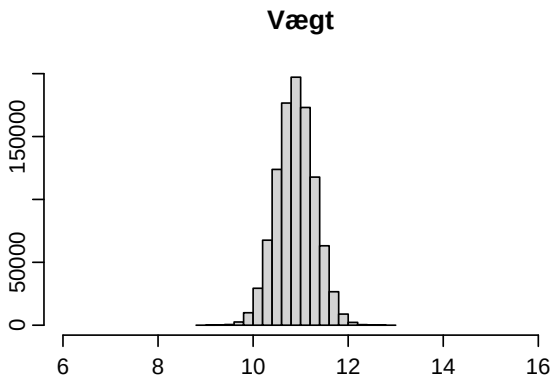


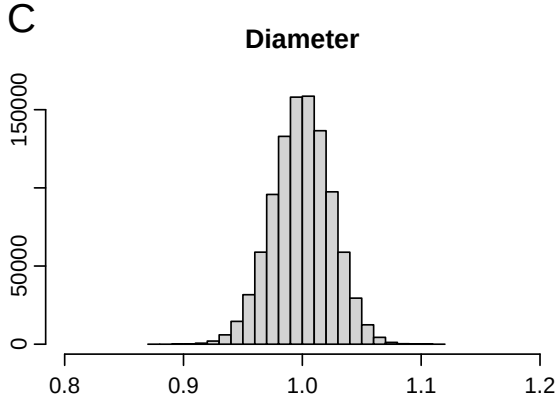
Plot B

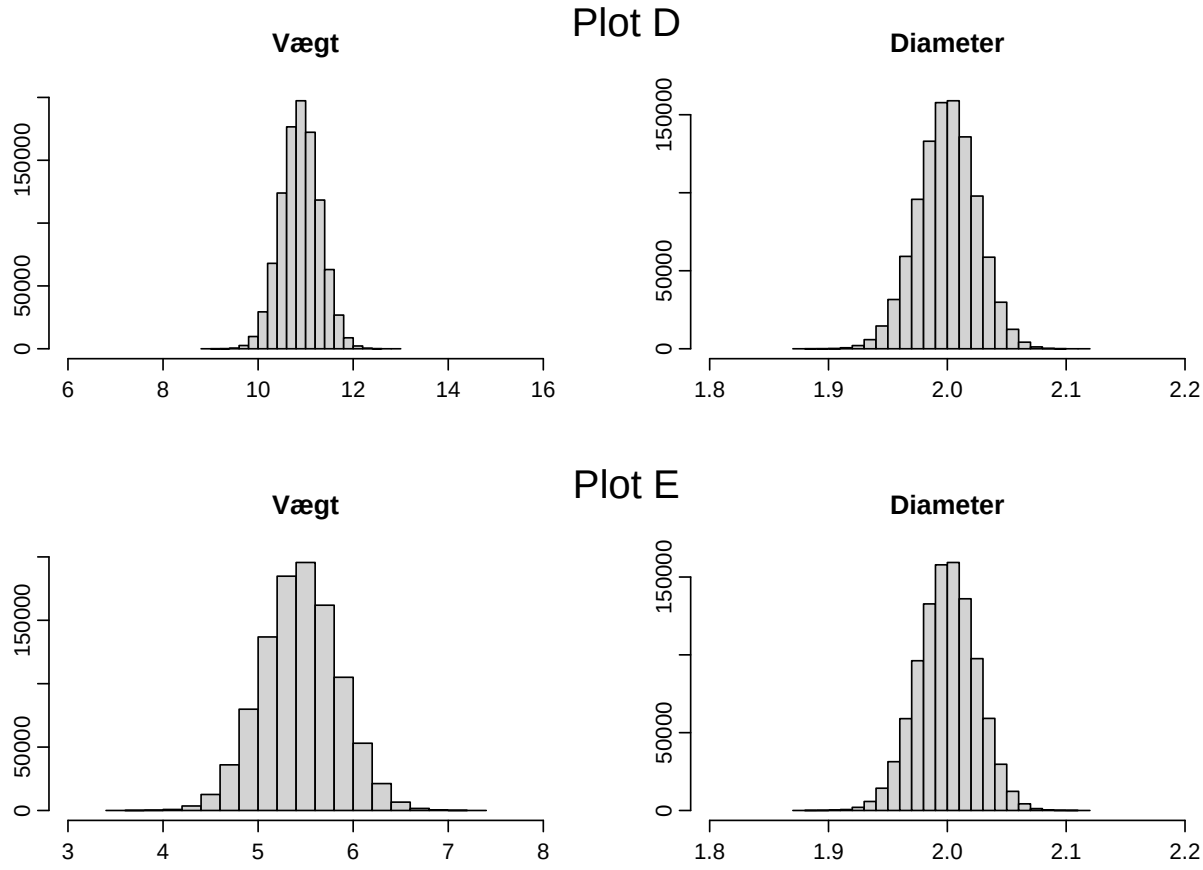




Plot C







1 ☐ Plot A

2 ☐ Plot B

3 ☐ Plot C

4 ☐ Plot D

5 ☐ Plot E

Spørgsmål VI.3 (14)

Legetøjsbutikken modtager nye glaskugler hver måned. Nogle gange opdager de at leveringen indeholder fejlbehæftede glaskugler. Ejeren af legetøjsbutikken beslutter at notere hver gang der er en levering med fejlbehæftede glaskugler. Tiden (målt i måneder) mellem leveringer med fejlbehæftede glaskugler er gemt i variablen x.

```
x = np.array([13, 4, 1, 17, 11, 2, 24, 25, 8, 4, 7, 7, 5, 6, 2, 13, 16, 3, 9, 11])
```

Brug bogens definition af stikprøvekvartiler (på engelsk: "sample quantiles") til at beregne IQR (den interkvartile variationsbredde).

1 ☐ IQR = 4

2 ☐ IQR = 7.5

3 ☐ IQR = 9

4 ☐ IQR = 11

5 ☐ IQR = 12

Spørgsmål VI.4 (15)

Ejeren af legetøjsbutikken beslutter at de vil stoppe med at købe glaskugler fra sælgeren hvis hændelser med fejlbehæftede leveringer bliver for hyppige. Uden at foretage nogen antagelser om fordelingen af tiden mellem fejlbehæftede leveringer, beregner ejeren et ikke-parametrisk 95% bootstrap konfidensinterval for medianen af tiden mellem sådanne hændelser. Hvilken af de følgende Python-koder beregner dette interval for medianen korrekt?

1 ☐
```
simsamples = np.random.choice(x, size=(10000, len(x)))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.05, 0.95], method="averaged_inverted_cdf")
print(quantiles)
```

2 ☐
```
simsamples = np.random.choice(x, size=(10000, len(x)))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.025, 0.975], method="averaged_inverted_cdf")
print(quantiles)
```

3 ☐
```
simsamples = stats.expon.rvs(size=(10000, len(x)), scale=np.mean(x))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.05, 0.95], method="averaged_inverted_cdf")
print(quantiles)
```

4 ☐
```
simsamples = stats.expon.rvs(size=(10000, len(x)), scale=np.mean(x))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.025, 0.975], method="averaged_inverted_cdf")
print(quantiles)
```

5 ☐
```
simsamples = np.random.choice(x, size=(10000, len(x)))
medians = np.median(simsamples, axis=1)
quantiles = np.quantile(medians, [0.005, 0.995], method="averaged_inverted_cdf")
print(quantiles)
```

Spørgsmål VI.5 (16)

Efter noget tid foretager sælgeren en forbedring af kvaliteten ved manuelt at fjerne poser med fejlbehæftede glaskugler. Igen beslutter ejeren af lejetøjsbutikken at notere leveringer med fejlbehæftede glaskugler. Tiden (målt i måneder) mellem leveringer med fejlbehæftede glaskugler er gemt i variablen y.

```
y = np.array([3,2,1,14,23,38,25,4,14,28,6,34,5,25,17,20,11,19,4,9])
```

Herefter foretages de følgende beregninger for at teste om det nye tiltag for at forbedre kvaliteten har hjulpet:

```
simXsamples = stats.expon.rvs(size=(10000, len(x)), scale=np.mean(x))

simYsamples = stats.expon.rvs(size=(10000, len(y)), scale=np.mean(y))

simDiff = np.median(simXsamples, axis=1) - np.median(simYsamples, axis=1)


print(np.percentile(simDiff, [0.5, 99.5], method='averaged_inverted_cdf'))

[-15.8691798    5.10555541]

print(np.percentile(simDiff, [2.5, 97.5], method='averaged_inverted_cdf'))

[-12.40129205   3.13755755]

print(np.percentile(simDiff, [5, 95], method='averaged_inverted_cdf'))

[-10.80677812   1.87399112]
```

Hvilket af følgende udsagn er korrekt?

1 ☐ Analysen laver ingen antagelser om distributionen af x og y. På et $\alpha = 1\%$ signifikansniveau konkluderes det, at der ikke er nogen signifikant forskel mellem medianerne.

2 ☐ Analysen antager at både x og y er normalfordelte. På et $\alpha = 5\%$ signifikansniveau konkluderes det, at der ikke er nogen signifikant forskel mellem medianerne.

3 ☐ Analysen antager at både x og y er exponentielt fordelte. På et $\alpha = 1\%$ signifikansniveau konkluderes det, at der er en signifikant forskel mellem medianerne.

4 ☐ Analysen laver ingen antagelser om distributionen af x og y. På et $\alpha = 5\%$ signifikansniveau konkluderes det, at der er en signifikant forskel mellem medianerne.

5 ☐ Ingen af de ovenstående udsagn er korrekte.

Fortsæt på side 17

Opgave VII

Antag, at vi har indsamlet eksamensresultater fra to grupper:

Gruppe 1: 82, 91, 85, 89, 88

Gruppe 2: 76, 84, 80, 82, 83

Vi antager, at eksamensresultaterne følger normalfordelinger med samme varians i begge grupper. Derudover antager vi, at eksamensresultaterne er uafhængige og ensfordelte (i.i.d.) indenfor hver gruppe.

Spørgsmål VII.1 (17)

Hvad er estimatet af den sammenvægtede varians ("pooled variance")?

1 ☐ 9.00

2 ☐ 27.10

3 ☐ 11.25

4 ☐ 10.00

5 ☐ 8.00

Spørgsmål VII.2 (18)

Hvad er det mindste antal observationer, der kræves i hver gruppe (samme antal observationer i begge grupper) for at opnå en styrke på 99% for at påvise en forskel i middelværdier på mindst 4 mellem de to grupper, hvis variansen er 20 (ens varians i begge grupper) og med et signifikansniveau på 1%?

1 ☐ Mindst 56 (eller 55, afhængigt af beregningsmetode)

2 ☐ Mindst 39 (eller 38, afhængigt af beregningsmetode)

3 ☐ Mindst 62 (eller 61, afhængigt af beregningsmetode)

4 ☐ Mindst 32 (eller 31, afhængigt af beregningsmetode)

5 ☐ Mindst 82 (eller 79, afhængigt af beregningsmetode)

Fortsæt på side 18

Opgave VIII

Til forberedelse af en konference skal arrangørerne planlægge kaffepauser effektivt. De estimerer at antallet af deltagere der har brug for kaffe vil følge en Poisson fordeling og at der i gennemsnit vil være 200 deltagere der har brug for kaffe hver time. Arrangørerne opstiller kaffestationer nok til at servere 240 deltagere per time.

Spørgsmål VIII.1 (19)

Hvad er sandsynligheden for, at antallet af deltagere, der har brug for kaffe, overstiger kapaciteten i løbet af en tilfældigt valgt time?

1 ☐ 0.0027

2 ☐ 0.023

3 ☐ 0.11

4 ☐ 0.24

5 ☐ 0.0045

Fortsæt på side 19

Opgave IX

I en bestemt produktionsvirksomhed følger medarbejdernes produktivitet en normalfordeling. Månedsvis kan 50% af medarbejderne producere 170 enheder eller mere (og dermed kan 50% producere under 170 enheder). Antag, at 68.3% af virksomhedens medarbejdere producerer inden for intervallet 160-180 enheder (og dermed producerer 84.1% under 180 enheder).

Spørgsmål IX.1 (20)

Hvilken procentdel af medarbejderne producerer mindst 190 enheder?

1 ☐ $\approx 2.275\%$

2 ☐ $\approx 1.1\%$

3 ☐ $\approx 4.3\%$

4 ☐ $\approx 0.52\%$

5 ☐ $\approx 0.09\%$

Fortsæt på side 20

Opgave X

En teknologivirksomhed har registreret sine månedlige salgstal over en periode på tre år (36 måneder). De månedlige salgstal er opsummeret i nedenstående tabel, der viser de gennemsnitlige månedlige salg og stikprøve-standardafvigelsen af de månedlige salg for hvert af de tre år.

År	2021	2022	2023
Gennemsnitligt månedligt salg (M DKK)	391.2	402.5	429.4
Standardafvigelsen af de månedlige salg (M DKK)	22.3	27.5	26.7

Herefter formulerer ingeniørerne i virksomheden en ensidet variansanalysemodel for dataene, hvor det månedlige salg bruges som responsvariabel og året som *behandling* (*treatment*).

Spørgsmål X.1 (21)

Hvad er middelkvadratafvigelsen for fejlen (MSE) i variansanalysemodellen?

1 ☐ 25.50

2 ☐ 162.56

3 ☐ 407.70

4 ☐ 655.48

5 ☐ 1966.43

Spørgsmål X.2 (22)

Ingeniørerne planlagde på forhånd at beregne parvise konfidensintervaller for $\mu_{2022} - \mu_{2021}$ og $\mu_{2023} - \mu_{2022}$ med et overordnet signifikansniveau på 10%, hvor μ_i henviser til det gennemsnitlige månedlige salg for år i. Hvilken fraktil fra t-fordelingen skal bruges i beregningerne af konfidensintervallerne?

1 ☐ 90%-fraktilen i t-fordelingen med 33 frihedsgrader

2 ☐ 95%-fraktilen i t-fordelingen med 33 frihedsgrader

3 ☐ 95%-fraktilen i t-fordelingen med 34 frihedsgrader

4 ☐ 97.5%-fraktilen i t-fordelingen med 33 frihedsgrader

5 ☐ 97.5%-fraktilen i t-fordelingen med 34 frihedsgrader

Fortsæt på side 21

Opgave XI

For at studere kriminalitet i Danmark er forskere interesserede i antallet af personer anbragt i varetægtsfængsling efter deres anholdelse. Disse tal er registreret og tilgængelige via Danmarks Statistik. De årlige optællinger fra 2015 til 2022 er kategoriseret i tre aldersgrupper: "Ung" (alder 15-29), "Mellemalder" (alder 30-39) og "Ældre" (alder 40 og over). Data læses ind i Python ved hjælp af følgende kode:

```
tbl = np.array([[2048, 1072, 821],
                [2208, 998, 836],
                [2359, 1092, 853],
                [2138, 1093, 880],
                [1984, 935, 799],
                [1777, 872, 860],
                [1604, 818, 729],
                [1564, 943, 753]])
tbl = pd.DataFrame(tbl,
                   index=['2015','2016','2017','2018',
                          '2019','2020','2021','2022'],
                   columns=['Ung','Mellemalder','Ældre'])
```

Spørgsmål XI.1 (23)

Nulhypotesen, at aldersfordelingen af personer i varetægtsfængsling ikke ændres i årenes løb, skal testes. Hvad er resultatet og konklusionen af den passende test (både argument og konklusion skal være korrekt)?

1 ☐ p-værdien er 0.24 og konklusionen er, at der ikke er en signifikant ændring i fordeling mellem årene.

2 ☐ p-værdien er $0.24 \cdot 10^{-10}$ og konklusionen er, at der er en signifikant ændring i fordeling i hvert eneste år.

3 ☐ p-værdien er $0.24 \cdot 10^{-10}$ og konklusionen er, at der er en signifikant ændring i fordeling i mindst et af årene.

4 ☐ p-værdien er $4.1 \cdot 10^{-10}$ og konklusionen er, at der er en signifikant ændring i fordeling i hvert eneste år.

5 ☐ p-værdien er $4.1 \cdot 10^{-10}$ og konklusionen er, at der er en signifikant ændring i fordeling i mindst et af årene.

Spørgsmål XI.2 (24)

Under nulhypotesen om ingen ændring i fordelingen, hvad er det forventede antal personer anbragt i varetægtsfængsling i "Ung"kategorien, hvis det samlede antal af sådanne placeringer i et bestemt år er 3000?

1 ☐ 978

2 ☐ 1364

3 ☐ 1566

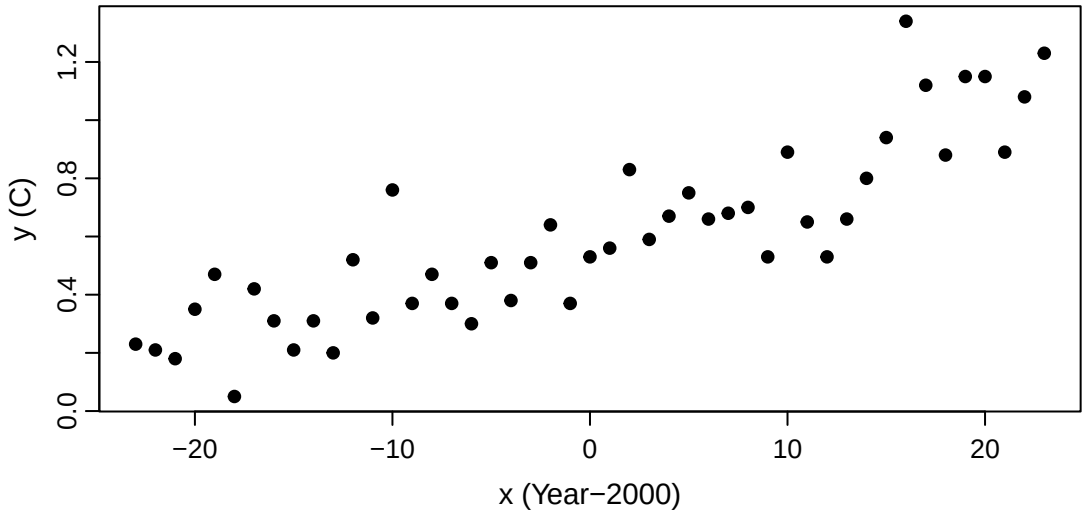
4 ☐ 1960

5 ☐ 2048

Fortsæt på side 23

Opgave XII

Figuren nedenfor viser den gennemsnitlige globale temperaturanomali i [°C], udregnet som temperaturen minus gennemsnitstemperaturen for perioden 1900-2000, som funktion af tiden. Figuren viser udviklingen i perioden 1977 til 2023 (x-aksen er året minus 2000).



I første omgang tilpasses (fittes) en simpel lineær regressionsmodel til dataene

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

I modellen er Y_i temperaturanomalien og x_i året (minus 2000) for observation i. Resultatet er angivet nedenfor:

```
fit = smf.ols(formula = 'y ~ x', data = dat).fit()

print(fit.summary(slim=True))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.754
Model:                            OLS   Adj. R-squared:                  0.748
No. Observations:                  47   F-statistic:                     137.7
Covariance Type:            nonrobust   Prob (F-statistic):           2.76e-15
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.6015      0.023     26.639      0.000       0.556       0.647
x              0.0195      0.002     11.734      0.000       0.016       0.023
==============================================================================
```

```
print(round(np.sqrt(fit.scale),4))
```

```
0.1548
```

```
print(fit.pvalues)
```

```
Intercept    3.420248e-29
x            2.758270e-15
dtype: float64
```

Spørgsmål XII.1 (25)

Hvilket af følgende udsagn om antagelserne for modellen er ikke korrekt?

1 ☐ $\varepsilon_i \sim N(0, \sigma^2)$.

2 ☐ ε_i og ε_j er uafhængige for $i \neq j$.

3 ☐ $V(\varepsilon_i) = V(\varepsilon_j)$ for alle (i, j).

4 ☐ Y_i og ε_i er uafhængige.

5 ☐ $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Spørgsmål XII.2 (26)

Modellen giver anledning til hvilken konklusion (ved brug af signifikansniveauet $\alpha = 0.05$) for forholdet mellem tid (i år) og temperaturen (både konklusionen og argumentet skal være korrekte)?

1 ☐ Temperaturen ændres signifikant med tiden (`x`), da $0.0195 < 0.05$.

2 ☐ Temperaturen ændres signifikant med tiden (`x`), da $0.002 < 0.05$.

3 ☐ Tid (`x`) har en signifikant virkning på temperaturen, da $0.002 < 0.05$.

4 ☐ Temperaturen ændres signifikant med tiden (`x`), da $2.758 \cdot 10^{-15} < 0.05$.

5 ☐ Temperaturen er en funktion af tiden (`x`), da $0.0195 < 0.05$.

Fortsæt på side 25

Spørgsmål XII.3 (27)

I hvilket år estimerer modellen, at den forventede temperatur vil være 1 grad højere end temperaturen i år 2000?

1 ☐ 2051

2 ☐ 2065

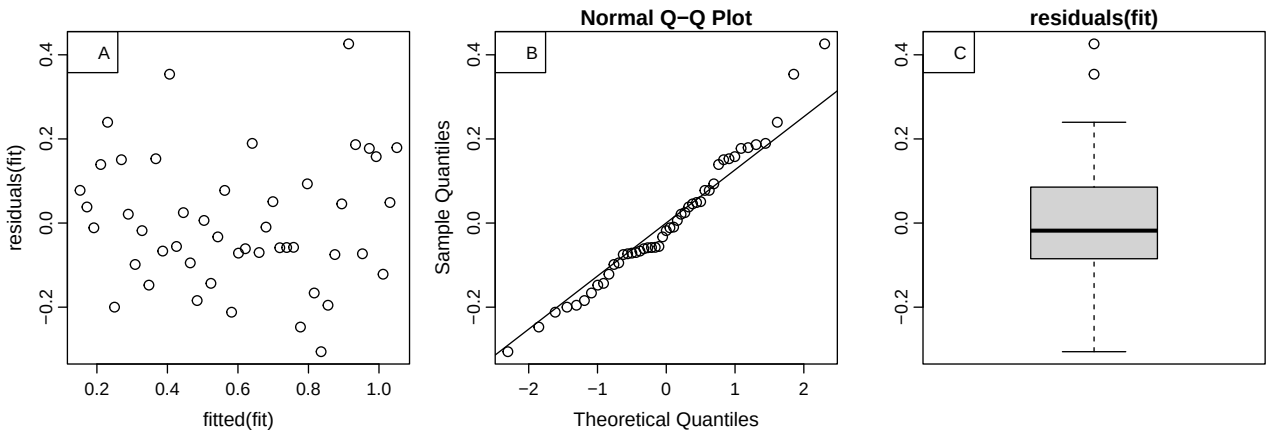
3 ☐ 2075

4 ☐ 2102

5 ☐ 2215

Spørgsmål XII.4 (28)

For at validere modellen er følgende plot af residualerne blevet lavet.



På grundlag af plottene, hvilket af følgende udsagn er korrekt (både konklusionen og figurhenvisningen, hvorfra dette kan konkluderes, skal være korrekte)?

1 ☐ Residualerne er tilsyneladende uafhængige, hvilket ses på Plot B.

2 ☐ Residualerne er tydeligvis ikke ensfordelte, hvilket ses på Plot C.

3 ☐ Der er tilsyneladende ingen systematiske mønstre i residualerne, hvilket ses på Plot A.

4 ☐ Der mangler tydeligvis et kvadratisk led i modellen, hvilket ses på Plot C.

5 ☐ Antagelsen om varianshomogenitet er tydeligvis brudt, hvilket ses på Plot B.

Spørgsmål XII.5 (29)

Uanset konklusionerne i de forrige spørgsmål, besluttes det at formulere en kvadratisk model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

I Python-koden nedenfor repræsenteres $\texttt{x}^2$ med `x2`, medens nogle dele af outputtet fra "summary"er fjernet, og nogle tal er erstattet af bogstaver.

```
fit = smf.ols(formula = 'y ~ x + x2', data = dat).fit()
print(fit.summary(slim=True))

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.779
Model:                            OLS   Adj. R-squared:                  0.769
No. Observations:                  47   F-statistic:                     77.49
Covariance Type:            nonrobust   Prob (F-statistic):           3.82e-15
==============================================================================
                 coef       std err       t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5472833   0.0324647      A          D          -          -
x              0.0195317   0.0015949      B          E          -          -
x2             0.0002946   0.0001315      C          F          -          -
==============================================================================
```

I vurderingen af, om det kvadratiske led skal inkluderes i modellen, hvilken af følgende konklusioner er korrekt på signifikansniveauet $\alpha = 0.05$?

1 ☐ `C`=6.6, og $\hat{\beta}_2$ er signifikant forskellig fra 0, da den kritiske værdi er 2.02.

2 ☐ `C`=2.2, og $\hat{\beta}_2$ er signifikant forskellig fra 0, da den kritiske værdi er 2.02.

3 ☐ `B`=11.7, og $\hat{\beta}_1$ er signifikant forskellig fra 0, da den kritiske værdi er 1.96.

4 ☐ `A`=26.6, og $\hat{\beta}_1$ er signifikant forskellig fra 0, da den kritiske værdi er 1.96.

5 ☐ `C`=2.2, og $\hat{\beta}_2$ er ikke signifikant forskellig fra 0, da den kritiske værdi er 2.02.

Fortsæt på side 27

Opgave XIII

Ti personer registrerede deres systoliske blodtryksniveauer om morgenen den 1. januar og den 1. juli. Ud fra disse målinger har vi til hensigt at undersøge, om der er en signifikant forskel i det systoliske blodtryk mellem vinter og sommer. Det kan antages, at blodtryksmålingerne om vinteren og om sommeren følger en normalfordeling.

Spørgsmål XIII.1 (30)

Hvilken analyse er mest hensigtsmæssig?

1 ☐ Test af forskel mellem to proportioner

2 ☐ t-test med sammenvægtet ("pooled") varians

3 ☐ Welch t-test

4 ☐ Parret t-test

5 ☐ Test med en binomialfordeling

SÆTTET ER SLUT. Nyd ferien!