Danmarks Tekniske Universitet *Side 1 af 25 sider.*

Skriftlig prøve: 26.06.2025

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

Varighed: 4 timer

Tilladte hjælpemidler: Alle hjælpemidler - ingen internetadgang

Dette sæt er besvaret af

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

Opgavesættet består af 30 spørgsmål af "multiple choice" typen, som er fordelt på 14 opgaver. For at besvare spørgsmålene skal du udfylde "multiple choice" 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	II.1	III.1	IV.1	IV.2	V.1	V.2	V.3	VI.1
Spørgsmål	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Svar										

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

Opgave	XI.4	XI.5	XI.6	XII.1	XIII.1	XIII.2	XIII.3	XIV.1	XIV.2	XIV.3
Spørgsmål	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)	(29)	(30)
Svar										

Eksamenssættet består af 25 sider.

Fortsæt på side 2

Multiple choice opgaver: *Der gøres opmærksom på, at der i hvert spørgsmål er én og kun én svarmulighed, som er rigtig. Endvidere er det ikke givet, at alle de anførte alternative svarmuligheder er meningsfulde. Husk altid at afrunde dit eget resultat til antallet af decimaler givet i svarmulighederne før du vælger et svar. Husk også, at der kan forekomme små afvigelser mellem resultatet af bogens formler og tilsvarende indbyggede funktioner i Python.*

Opgave I

En ingeniør ønsker at teste, om en ny legering har en trækstyrke med middelværdi 500 MPa. En tilfældig stikprøve på 30 prøver testes, hvilket giver et stikprøvegennemsnit på 510 MPa og en stikprøve standardafvigelse på 20 MPa. Det antages at observationerne er iid og normalfordelt.

Spørgsmål I.1 (1)

Hvad er den tilhørende p-værdi for den relevante hypotesetest med følgende hypoteser:

$$H_0 : \mu = 500, \quad H_A : \mu \neq 500.$$

1 ☐ $p = 0.010$

2 ☐ $p = 0.621$

3 ☐ $p = 0.310$

4 ☐ $p = 0.006$

5 ☐ $p = 0.005$

Ingeniøren planlægger nu et nyt forsøg hvor han gerne vil opnå en "margin of error" på højst 2 MPa. Han bruger den observerede standardafvigelse som scenarie og signifikansniveau $\alpha = 0.05$.

Spørgsmål I.2 (2)

Hvor stor en stikprøve skal der udtages for at opnå en "margin of error" på højst 2 MPa?

1 ☐ ca. 20 observationer

2 ☐ ca. 40 observationer

3 ☐ ca. 1538 observationer

4 ☐ ca. 16 observationer

5 ☐ ca. 385 observationer

Fortsæt på side 3

Opgave II

En træner ønsker at undersøge om der er forskel på forskellige typer af målrettet træning i forhold til at forbedre den tid det tager at løbe op ad trapper. Træneren indsamler data fra 15 forsøgspersoner, der inddeles (tilfældigt) i tre lige store grupper: Gruppe A, Gruppe B og Gruppe C. Træneren sætter forsøgspersonerne til at lave nogle målrettede øvelser over de næste 4 uger. Forsøgspersoner, der er i samme gruppe, laver de samme øvelser, men træneren uddeler forskellige øvelser til de tre grupper. For hver forsøgsperson indsamles data om forbedringen i den tid det tager personen at løbe op ad en trappe i træningscenteret (tidsforbedringen måles i sekunder).

De observerede tidsforbedringer er:

Gruppe:	Tidsforbedring (målt i sekunder):
A	2.1, 2.5, 2.3, 2.4, 2.2
B	2.8, 2.9, 2.7, 3.0, 2.6
C	2.3, 2.4, 2.5, 2.2, 2.1

Den gennemsnitlige tidsforbedring for samtlige 15 forsøgspersoner er $\hat{\mu} = 2.467$ og de gennemsnitlige tidsforbedringer indenfor hver gruppe er givet ved: $\hat{\mu}_A = 2.30$, $\hat{\mu}_B = 2.80$, $\hat{\mu}_C = 2.30$. Det kan antages at alle observationer er uafhængige og normalfordelt.

Spørgsmål II.1 (3)

Hvad er den mest passende statistiske model og analyse, når man ønsker at undersøge om der er forskel på effekten af de forskellige træningstyper?

1 ☐ En passende model kunne være $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ ($\epsilon_{ij} \sim N(0, \sigma^2)$), hvor Y_{ij} er tidsforbedring af person nummer j i gruppe nr i. En relevant analyse ville da være at udføre en t-test, der undersøger nulhypotesen $H_0 : \mu = 0$.

2 ☐ En passende model kunne være $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ ($\epsilon_i \sim N(0, \sigma^2)$), hvor x_i er tidsforbedring af person nummer i. En relevant analyse ville da være at udføre en t-test, der undersøger nulhypotesen $H_0 : \beta_1 = 0$.

3 ☐ En passende model kunne være $Y_{ij} = \mu_i + \epsilon_{ij}$ ($\epsilon_{ij} \sim N(0, \sigma^2)$), hvor Y_{ij} er tidsforbedring af person nummer j i gruppe nr i. En relevant analyse ville da være at udføre en variansanalyse, der undersøger nulhypotesen $H_0 : \mu_A = \mu_B = \mu_C = 0$.

4 ☐ En passende model kunne være $Y_{ij} = \beta_0 + \beta_i x_{ij} + \epsilon_{ij}$ ($\epsilon_{ij} \sim N(0, \sigma^2)$), hvor x_{ij} er tidsforbedring af person nummer j i gruppe nr i. En relevant analyse ville da være en variansanalyse, der undersøger nulhypotesen $H_0 : \beta_i = 0$ (dvs. der udføres i alt 3 test - én for hver gruppe).

5 ☐ En passende model kunne være $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ ($\epsilon_{ij} \sim N(0, \sigma^2)$), hvor Y_{ij} er tidsforbedring af person nummer j i gruppe nr i. En relevant analyse ville da være en variansanalyse, der undersøger nulhypotesen $H_0 : \alpha_A = \alpha_B = \alpha_C = 0$.

Opgave III

Antag at Y følger en exponential fordeling med $E(Y) = 3$.

Spørgsmål III.1 (4)

Hvad er $P(2 < Y < 4)$?

1 ☐ 0.49

2 ☐ 0.25

3 ☐ 0.61

4 ☐ 0.75

5 ☐ 0.0024

Fortsæt på side 5

Opgave IV

En dyreforretning ønsker at undersøge, hvor stor en andel af danske husholdninger der har en hund. De gennemfører en stikprøveundersøgelse blandt 1000 af deres kunder, hvor de bliver spurgt, om de har en hund. Forretningen antager at de 1000 kunder repræsenterer 1000 husholdinger.

Af disse svarer 320, at de har en hund.

Spørgsmål IV.1 (5)

Hvad er den estimerede andel ($\hat{p}$) af husholdninger der har en hund og hvad er usikkerheden (standardfejlen, $s.e._{\hat{p}}$) på denne andel?

1 ☐ $\hat{p} = 0.32$ og $s.e._{\hat{p}} = 0.00022$

2 ☐ $\hat{p} = 0.32$ og $s.e._{\hat{p}} = 0.015$

3 ☐ $\hat{p} = 0.32$ og $s.e._{\hat{p}} = 0.047$

4 ☐ $\hat{p} = 0.32$ og $s.e._{\hat{p}} = 0.32$

5 ☐ $\hat{p} = 0.32$ og $s.e._{\hat{p}} = 0.010$

Spørgsmål IV.2 (6)

Officielle tal siger at ca 20% af danskerne har en hund. Dyreforretningen havde derfor forventet at deres undersøgelse ville resultere i en andel, der var tættere på 0.20. Er det sandsynligt at deres resultat - altså at hele 32% af husholdningerne har hund - skyldes tilfældig variation? Og kan det være rigtigt at den sande andel af danske husholdninger der har hund i virkeligheden (kun) er ca 20%?

1 ☐ Ja, dyreforretningen har ved et tilfælde udvalgt en stikprøve, hvor flere end forventet har en hund. Dette skyldes sandsynligvis tilfældig variation og den sande andel kan godt være ca 20%.

2 ☐ Nej, det er ikke sandsynligt at dyreforretningens resultat skyldes tilfældig variation. p-værdien for den relevante test er 0.0015, så man vil forkaste nulhypotesen om at den sande andel er 0.20. Man må dermed konkludere at den sande andel sandsynligvis ikke er 20%.

3 ☐ Nej, det er ikke sandsynligt at dyreforretningens resultat skyldes tilfældig variation. p-værdien for den relevante test er 0.0015, så man vil forkaste nulhypotesen om at den sande andel er 0.20. Det er dog tvilvsomt om stikprøven er repræsentativ, så den sande andel kan godt være 20%.

4 ☐ Nej, det er ikke sandsynligt at dyreforretningens resultat skyldes tilfældig variation. p-værdien for den relevante test er $2 \cdot 10^{-21}$, så man vil forkaste nulhypotesen om at den sande andel er 0.20. Da stikprøven klart er repræsentativ må man konkludere at den sande andel af husholdninger, der har en hund, sandsynligvis ikke er 20%.

5 ☐ Nej, det er ikke sandsynligt at dyreforretningens resultat skyldes tilfældig variation. p-værdien for den relevante test er $2 \cdot 10^{-21}$, så man vil forkaste nulhypotesen om at den sande andel er 0.20. Det er dog tvilvsomt om stikprøven er repræsentativ, så den sande andel kan godt være 20%.

Fortsæt på side 7

Opgave V

I et studie haves data fra 4 forskellige grupper:

Gruppe 1:	89, 102, 94, 90, 100
Gruppe 2:	78, 46, 65, 72, 69
Gruppe 3:	83, 89, 81, 89, 90
Gruppe 4:	82, 101, 93, 88, 104

Tabellen kan indtastes i Python med følgende kode.

```
y = np.array([89, 102, 94, 90, 100,
              78, 46, 65, 72, 69,
              83, 89, 81, 89, 90,
              82, 101, 93, 88, 104])
Group = pd.Categorical([1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4])
D = pd.DataFrame({'y': y, 'Group': Group})
```

Det kan antages at data kan beskrives med følgende model: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$.

Spørgsmål V.1 (7)

Hvad er variationen mellem grupper, MS(Group), og variationen indenfor grupper, MSE?

1 ☐ MS(Group) = 190.3 og MSE = 1122

2 ☐ MS(Group) = 894.5 og MSE = 70.15

3 ☐ MS(Group) = 2683 og MSE = 1122

4 ☐ MS(Group) = 894.5 og MSE = 2683

5 ☐ MS(Group) = 190.3 og MSE = 70.15

Fortsæt på side 8

Spørgsmål V.2 (8)

Hvilket udsagn om modellen ovenfor er IKKE korrekt?

1 ☐ Y_{ij} er observation nr. j i gruppe nr i. $\hat{\alpha}_i$ er gruppe i's gennemsnitlige afvigelse fra det overordnede gennemsnit $\hat{\mu}$.

2 ☐ Den samlede varians af data (dvs. $\frac{1}{N-1}SST$) kan ikke være større end $\hat{\sigma}^2$.

3 ☐ MSE angiver variansen indenfor hver gruppe, og da vi antager at denne er ens i alle grupper har vi også $MSE = \hat{\sigma}^2$.

4 ☐ Hvis variansen af α_i'erne er stor sammenlignet med MSE betyder dette at der er forskel på grupperne.

5 ☐ I ovenstående data fås (for Gruppe 1, $i = 1$) $\hat{\alpha}_1 = 9.75$.

Hvis vi benævner observationerne fra hver gruppe $\boldsymbol{y}_i$ (eks. $\boldsymbol{y}_1 = [89 \quad 102 \quad 94 \quad 90 \quad 100]^T$), og samlingen af alle observationer ved $\boldsymbol{Y}$, kan modellen skrives som

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \boldsymbol{y}_3 \\ \boldsymbol{y}_4 \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

hvor det antages at $\boldsymbol{\beta}$ er identificerbar.

De tilsvarende "fittede" værdier kan skrives som

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y},$$

hvor $\boldsymbol{H}$ er udregnet fra $\boldsymbol{X}$.

Spørgsmål V.3 (9)

Hvilket af føgende udsagn om $\boldsymbol{X}$ og $\boldsymbol{H}$ er korrekt?

1 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times 5}$, og $Trace(\boldsymbol{H}) = 5$.

2 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times 4}$, og $Trace(\boldsymbol{H}) = 4$.

3 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times 4}$, og $Trace(\boldsymbol{H}) = 20$.

4 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times 20}$, og $Trace(\boldsymbol{H}) = 20$.

5 ☐ $\boldsymbol{X} \in \mathbb{R}^{20\times 5}$, og $Trace(\boldsymbol{H}) = 20$.

Fortsæt på side 9

Opgave VI

Man ønsker at sammenligne gennemsnittet i to stikprøver, A og B. Begge stikprøver indeholder 50 uafhængige målinger der antages normalfordelte. Det oplyses at 95% konfidensintervallet for middelværdien i hver gruppe er:

95% CI for $\hat{\mu}_A = [15.2, 17.8]$

95% CI for $\hat{\mu}_B = [13.0, 15.5]$.

Spørgsmål VI.1 (10)

Hvilket af følgende udsagn er korrekt?

1 ☐ Da konfidensintervallerne overlapper kan de to underliggende populationer godt have samme gennemsnit. Derfor ser vi nemt at forskellen mellem de to stikprøvegennemsnit ikke er signifikant forskellig fra nul (ved signifikansniveau på 5%).

2 ☐ Da konfidensintervallerne overlapper, er forskellen mellem de to stikprøvegennemsnit ikke statistisk signifikant forskellig fra nul (ved signifikansniveau på 5%). Dvs. de to underliggende populationer har samme fordeling.

3 ☐ Stikprøvegennemsnittet i hhv. stikprøve A og B er 16.50 og 14.25, og der er signifikant forskel på disse gennemsnit ved signifikansniveau på 5% (men ikke ved signifikansniveau 1%).

4 ☐ Stikprøvegennemsnittet i hhv. stikprøve A og B er 16.50 og 14.25, og der er signifikant forskel på disse gennemsnit (ved signifikansniveau på 1%).

5 ☐ Stikprøvegennemsnittet i hhv. stikprøve A og B er 16.50 og 14.25, men der er ikke signifikant forskel på disse gennemsnit (ved signifikansniveau på 5%).

Fortsæt på side 10

Opgave VII

Fangst genfangst er en metode hvor et antal individer (dyr) indfanges, mærkes, og sættes ud. Efter en periode indfanges et antal individer og det undersøges hvor mange individer der er mærket. Metoden kan bruges til at estimere populationsstørrelser.

En biolog har indfanget $n_1 = 150$ fisk i en sø, mærket dem og sat dem ud igen. Biologen planlægger nu at vende tilbage og indfange $n_2 = 200$ fisk fra samme sø.

Spørgsmål VII.1 (11)

Hvis vi benævner det totale antal fisk i søen ved N (og antager at N er det samme ved udsætning og genfangst), hvilken fordeling følger da antallet af mærkede fisk (Y) ved genfangsten (det antages at alle mærkede fisk overlever og at det er helt tilfældigt hvilke af de N fisk der indfanges)?

1 ☐ En binomial fordeling med $p = \frac{150}{N}$, og $n = 200$, dvs. $Y \sim B\left(200, \frac{150}{N}\right)$.

2 ☐ En normalfordelig med paramtre $\mu = \frac{200 \cdot 150}{N}$, og $\sigma^2 = \frac{200 \cdot 150}{N}\left(1 - \frac{150}{N}\right)$.

3 ☐ En hypergeometrisk fordeling med parametre $n = 200$, $a = 150$, og N, dvs. $Y \sim H(200, 150, N)$.

4 ☐ En poisson fordeling med parameter $\lambda = \frac{150 \cdot 200}{N}$, dvs. $Y \sim Pois\left(\frac{150 \cdot 200}{N}\right)$.

5 ☐ En exponential fordeling med parameter $\lambda = \frac{N}{150 \cdot 200}$, dvs. $Y \sim Exp\left(\frac{N}{150 \cdot 200}\right)$.

Længden på de fangede fisk måles med henblik på at give et estimat på deres alder. Således klassificeres fisk på mellem 6 og 10 cm. som 1-årige, mens fisk over 10 cm. klassificeres som ældre. Det antages at længden på en et-årig fisk følger en nomal fordeling med middelværdi $\mu = 8$ cm. og standardafvigelse $\sigma = 1$ cm.

Spørgsmål VII.2 (12)

Hvad er sandsynligheden for at en et-årig fisk klassificeres som ældre end et år?

1 ☐ 0.159

2 ☐ 0.5

3 ☐ 0.841

4 ☐ 0.0228

5 ☐ 0.977

Fortsæt på side 11

Opgave VIII

En virksomhed ønsker at undersøge effekten af forhold omkring arbejdsbelysning og musik på medarbejderes produktivitet (målt i antal producerede enheder pr. time).

Der testes to faktorer:

1) Belysning (Faktor A) med to niveauer: Lav og Høj

2) Musik (Faktor B) med tre niveauer: Ingen, Rolig, og Energisk

Alle kombinationer afprøves og medarbejdernes gennemsnitlige produktivitet måles:

	Ingen Musik	Rolig Musik	Energisk Musik
Lav Belysning	20	23	19
Høj Belysning	25	27	22

Man antager nu modellen

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

hvor ϵ_{ij} iid., Y_{ij} (`y`) er produktiviteten, og α_i og β_j er effekten af Musik (`Music`) og Belysning (`Light`). For at undersøge effekterne har man fået følgende resultat fra Python (data fra tabellen overfor er gemt i `D`)

```
fit = smf.ols('y ~ Light + Music', data = D).fit()
sm.stats.anova_lm(fit)

##             df     sum_sq    mean_sq          F    PR(>F)
## Light      1.0  24.000000  24.000000  48.000000  0.020204
## Music      2.0  20.333333  10.166667  20.333333  0.046875
## Residual   2.0   1.000000   0.500000        NaN       NaN
```

Fortsæt på side 12

Spørgsmål VIII.1 (13)

Hvad er konklusionen på de relevante statistiske tests, ved brug af signifikansniveau $\alpha = 0.05$)?

1 ☐ Der ses en signifikant forskel i produktivitet både ift. forskellig belysning og ift. forskellig musik

2 ☐ Der ses en signifikant forskel i produktivitet ift. forskellig belysning, men ikke ift. forskellig musik.

3 ☐ Der ses en signifikant forskel i produktivitet ift. forskellig musik, men ikke ift. forskellig belysning.

4 ☐ Der ses ingen signifikant forskel i produktivitet, hverken ift. forskellig musik eller ift. forskellig belysning.

5 ☐ Der ses en signifikant forskel i produktivitet ift. forskellig musik, men man kan ikke konkludere om der er en effekt af forskellig belysning, da der kun er to niveau'er.

Spørgsmål VIII.2 (14)

For yderligere at undersøge effekten af musik ønsker man at lave parvise sammenligninger af effekterne af musik. Idet man ønsker at korrigere for at der laves flere tests, hvad er mindst signifikante afstand (Least Significant Distance, LSD) for de parvise sammenligninger af effekten af musik (brug signifikansniveau $\alpha = 0.05$)?

1☐ 2.5

2 ☐ 2.0

3 ☐ 1.5

4 ☐ 5.4

5 ☐ 6.2

Fortsæt på side 13

Opgave IX

En forbrugerorganisation ønsker at undersøge, hvor ofte et pakkeleverings firma leverer pakker til den nærmeste pakkeshop. De indsamler data fra 750 pakkeleveringer, fordelt på de 5 regioner. Resultaterne er opsummeret i følgende tabel:

Region	**Leveret til nærmeste pakkeshop**	**Leveret et andet sted**	**Total**
Hovedstaden	40	110	150
Midtjylland	95	55	150
Syddanmark	80	70	150
Nordjylland	85	65	150
Sjælland	70	80	150
Total	370	380	750

Der udføres nu en χ^2-test, for at undersøge om andelen af pakker leveret til nærmeste pakkeshop er den samme i alle 5 regioner.

Spørgsmål IX.1 (15)

Hvad er de - under nulhypotesen - forventede værdier i hver celle i tabellen?

1 ☐

Region	**Leveret til nærmeste pakkeshop**	**Leveret et andet sted**
Hovedstaden	75	75
Midtjylland	75	75
Syddanmark	75	75
Nordjylland	75	75
Sjælland	75	75

2 ☐

Region	**Leveret til nærmeste pakkeshop**	**Leveret et andet sted**
Hovedstaden	80	70
Midtjylland	70	80
Syddanmark	60	90
Nordjylland	50	100
Sjælland	40	110

3 ☐

Region	**Leveret til nærmeste pakkeshop**	**Leveret et andet sted**
Hovedstaden	100	0
Midtjylland	100	0
Syddanmark	100	0
Nordjylland	100	0
Sjælland	100	0

4 ☐

Region	Leveret til nærmeste pakkeshop	Leveret et andet sted
Hovedstaden	74	76
Midtjylland	74	76
Syddanmark	74	76
Nordjylland	74	76
Sjælland	74	76

5☐

Region	Leveret til nærmeste pakkeshop	Leveret et andet sted
Hovedstaden	40	110
Midtjylland	95	55
Syddanmark	80	70
Nordjylland	85	65
Sjælland	70	80

Spørgsmål IX.2 (16)

Der udføres en χ^2-test, for at undersøge om andelen af pakker leveret til nærmeste pakkeshop er den samme i alle 5 regioner, den relevante teststørrelse er udregnet til 47.21. Hvad er p-værdien for den relevante test og hvad er den tilsvarende konklusion (brug signifikansniveau $\alpha = 0.05$)?

1 ☐ p-værdien bliver 0.10 og konklusionen er dermed at der er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner

2 ☐ p-værdien bliver 0.10 og konklusionen er dermed at der ikke er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner

3 ☐ p-værdien bliver 0.05 og konklusionen er dermed at der ikke er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner

4 ☐ p-værdien bliver $1.4 \cdot 10^{-9}$ og konklusionen er dermed at der ikke er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner

5 ☐ p-værdien bliver $1.4 \cdot 10^{-9}$ og konklusionen er dermed at der er forskel i andelen af pakker leveret til nærmeste pakkeshop i de forskellige regioner

Fortsæt på side 15

Opgave X

En sensor måler temperaturen på en maskine, der helst ikke må blive varmere end 80°C. En stikprøve på 40 temperaturmålinger giver et stikprøvegennemsnit på 75.2°C og en stikprøve standardafvigelse på 2.5°C. Antag, at temperaturmålingerne er uafhængige af hinanden, samt at de følger en normalfordeling.

Spørgsmål X.1 (17)

Hvad er et 95% konfidensinterval for den sande middelværdi af temperaturen?

1 ☐ $[72.7, 77.7]$ °C

2 ☐ $[70.1, 80.3]$ °C

3 ☐ $[74.4, 76.0]$ °C

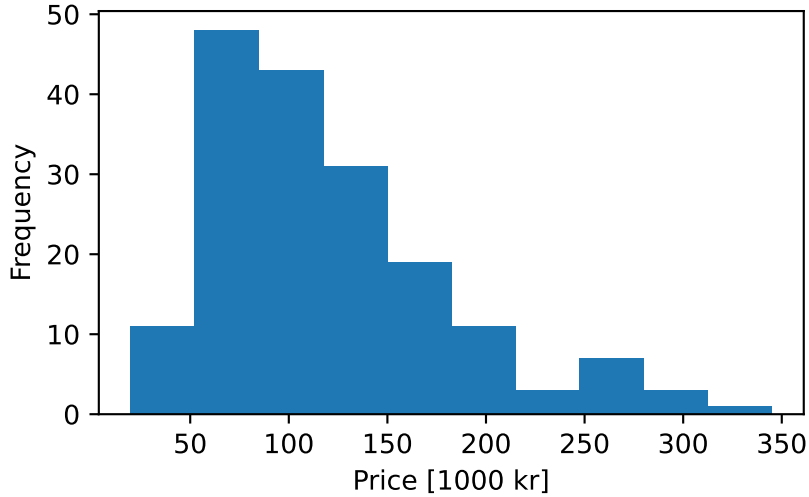
4 ☐ $[75.1, 75.3]$ °C

5 ☐ $[71.0, 79.4]$ °C

Fortsæt på side 16

Opgave XI

En bilejer ønsker at købe en ny (brugt) bil, for at undersøge hvad prisen bør være har hun indsamlet priser på den bil (mærke og model) hun ønsker. Histogrammet herunder viser fordelinger af priser på den bil hun ønsker sig.



I tillæg til histogrammet har hun udregnet gennemsnit og empiriske varians for de observerede priser (`price`) [1000 kr.].

```
np.mean(price)

np.float64(118.94622598870056)

np.var(price,ddof=1)

np.float64(3633.1131006077294)
```

Spørgsmål XI.1 (18)

Baseret på ovenstående, hvilken af følgende antagelser om fordelingen af prisen (Y) er så mest rimelig?

1 ☐ En log-normalfordeling med parametere $\alpha = 4.66$ og $\beta^2 = 0.478$, dvs. $Y \sim LN(4.66, 0.478^2)$.

2 ☐ En normalfordeling med parametere $\mu = 118.94$ og $\sigma^2 = 3633.1^2$, dvs. $Y \sim N(118.9, 3633.1^2)$.

3 ☐ En exponentialfordeling med parameter $\lambda = 118.9$, dvs. $Y \sim Exp(118.9)$.

4 ☐ En log-normal fordeling med parametere $\alpha = 118.9$ og $\beta^2 = 60.28^2$, dvs. $Y \sim LN(118.9, 60.28^2)$.

5 ☐ En normalfordeling med parametere $\mu = 118.94$ og $\sigma^2 = 60.28^2$, dvs. $Y \sim N(118.9, 60.28^2)$.

Bilejeren har også indsamlet data om alder og kilometerstand på bilerne. For at undersøge sammenhængen mellem pris, alder og kilometerstand på bilen har bilejeren kørt følgende kode (`price` [1000 kr.] er prisen, `age` [år] er alderen på bilen, `dist` [1000 km.] er kilometerstanden på bilen og `cars` er datasættet med de indsamlede tal),

```
fit = smf.ols("price ~ age + dist", data = cars).fit()
```

svarende til modellen

$$Y_i = \mu_i + \epsilon_i,$$

hvor formlen for μ_i er bestemt fra inputtet til `smf.ols` og $E(\epsilon_i) = 0$.

Spørgsmål XI.2 (19)

Hvilket af følgende udsagn om modellen eller model antagelserne er korrekt?

1 ☐ ϵ_i'erne er normalfordelt og iid. (uafhængige og identisk fordelte).

2 ☐ Det antages at den observerede korrelation mellem `age` og `dist` er lig nul.

3 ☐ Y_i'erne er normalfordelt og iid. (uafhængige og identisk fordelte).

4 ☐ μ_i'erne følger en normalfordeling og er iid.

5 ☐ $\mu_i = \beta_1 \texttt{age}_i + \beta_2 \texttt{dist}_i$.

Resultaterne af estimationen ovenfor er angivet nedenfor (idet en del tal er erstattet med symboler)

```
fit.summary(slim = True)

                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.793
Model:                            OLS   Adj. R-squared:                  0.791
No. Observations:                 177   F-statistic:                     334.0
Covariance Type:            nonrobust   Prob (F-statistic):           2.65e-60
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    255.8098      5.707         T1      P1         Ql_1        Qu_1
age           -9.6077      0.963         T2      P2         Ql_2        Qu_2
dist          -0.4756      0.055         T3      P3         Ql_3        Qu_3
==============================================================================
```

Spørgsmål XI.3 (20)

Hvilket af følgende udsagn er korrekt ved brug af signifikans niveau $\alpha = 0.05$?

1 ☐ Begge effekter (alder og kilometerstand) er signifikant forskellig fra nul og den forventede pris falder med alder og kilometerstand.

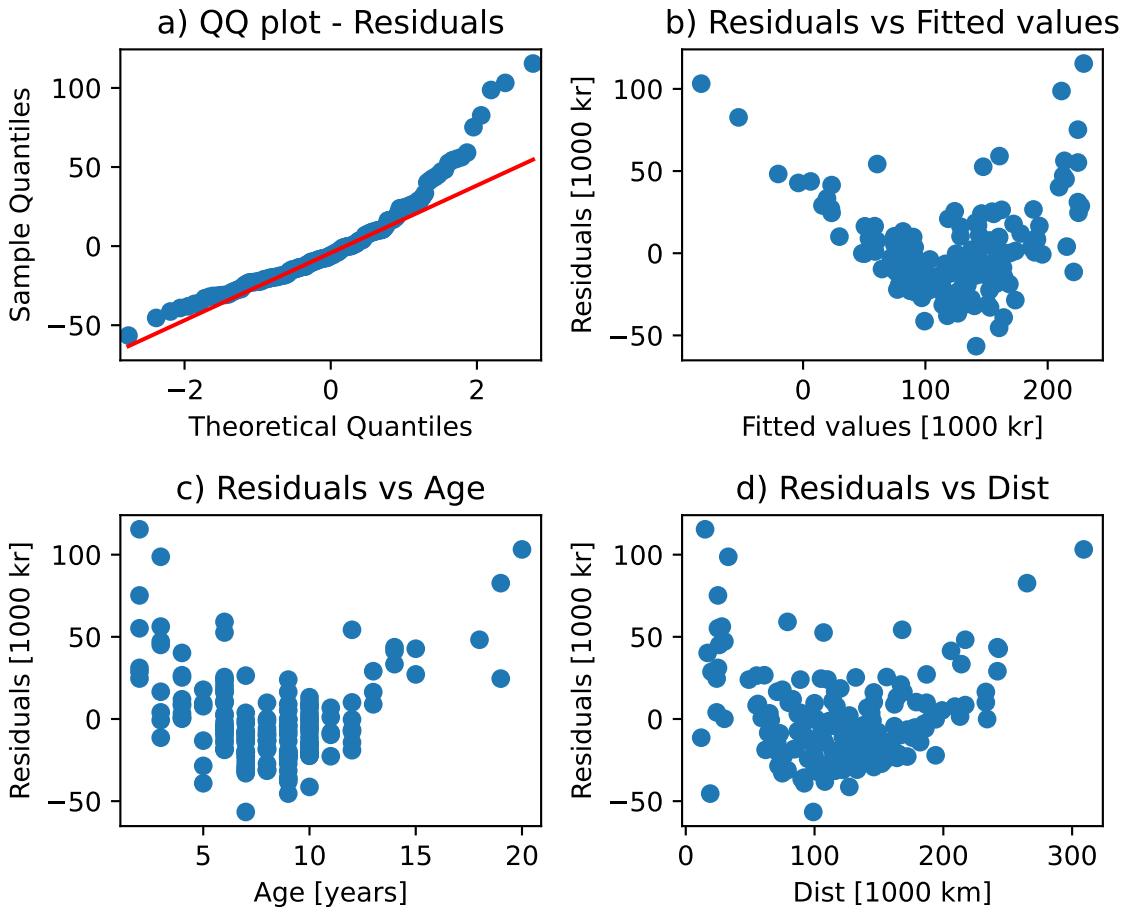
2 ☐ Ingen af effekterne (alder og kilometerstand) er signifikant forskellig fra nul.

3 ☐ Alder på bilen har en signifikant effekt på prisen, mens der ikke kan påvises en effekt af kilometerstand. Prisen stiger når alderen stiger.

4 ☐ Kilometerstand har en signifikant effekt på prisen, mens der ikke kan påvises en effekt af alder. Prisen falder når kilometerstand stiger.

5 ☐ Alder på bilen har en signifikant effekt på prisen, mens der ikke kan påvises en effekt af kilometerstand. Prisen falder når alderen stiger.

For at vurdere modellens gyldighed har bilejeren lavet en række residualplots, som vist herunder.



Fortsæt på side 19

Spørgsmål XI.4 (21)

Hvilket af følgende udsagn er IKKE korrekt (både udsagn og figur reference skal være korrekt)?

1 ☐ Uafhængighedsantagelsen er tydeligvis ikke opfyldt (plot a).

2 ☐ Varians homogenitets antagelsen ser ikke ud til at være opfyldt (plot b).

3 ☐ Et led af formen $\texttt{age}^2$ vil formentlig forbedre modellen (plot c).

4 ☐ Der er klare systematiske effekter i residuals vs. fitted (plot b).

5 ☐ Normalfordelings antagelsen er tydeligvis ikke opfyldt (plot a).

Uanset konklusionen ovenfor beslutter bilejeren at fortsætte analysen af modellen. Modellen svarende til Python summary'et ovenfor (lige efter spørgsmål 19) kan skrives i matrix-vektor notation som

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

hvor $\boldsymbol{X}$ er designmatricen svarende til Python summary'et. Til hjælp for de videre udregninger gives nedenstående Python kode

```
C = np.array([[177,1413,22393], [1413,13099,202256],
              [22393, 202256, 3384553]])
```

her er $\texttt{C}=\boldsymbol{X}^T\boldsymbol{X}$.

Spørgsmål XI.5 (22)

Hvis $\hat{\sigma}$ betegner den sædvanlige standardafvigelse, hvad er så 95% konfidensintervallet for prisen [1000 kr.] på en 10 år gammel bil der har kørt 100,000 km?

1 ☐ $112.2 \pm 0.272 \cdot \hat{\sigma}$

2 ☐ $112.2 \pm 0.0112 \cdot \hat{\sigma}$

3 ☐ $112.2 \pm 0.148 \cdot \hat{\sigma}$

4 ☐ $112.2 \pm 0.0375 \cdot \hat{\sigma}$

5 ☐ $112.2 \pm 1.97 \cdot \hat{\sigma}$

Fortsæt på side 20

Bilejeren ønsker nu at teste om der er en effekt af at inkludere `age` og `dist` i modellen svarende til nulhypotesen

$$\boldsymbol{Y} = \mathbf{1}\mu + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

mod den alternative hypotese (som angivet ovenfor)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}).$$

Projektions matricerne svarende til nul-hypotesen og den alternative hypotese benævnes $\boldsymbol{H}_0$ og $\boldsymbol{H}$ og hun beregner den sædvanlige test størrelse

$$Q = \frac{\frac{1}{df_0} SS_0}{\frac{1}{df_{SSE}} SSE},$$

hvor

$$\begin{aligned} SS_0 =& \boldsymbol{y}^T (\boldsymbol{H}_0 - \boldsymbol{H}) \boldsymbol{y} \\ SSE =& \boldsymbol{y}^T (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{y}. \end{aligned}$$

Spørgsmål XI.6 (23)

Hvad er Q i eksemplet overfor og hvilken fordeling skal Q sammelignes med?

1 ☐ $Q = 0.793$ som skal sammenlignes med en F-fordeling med 1 og 3 frihedsgrader.

2 ☐ $Q = 0.793$ som skal sammenlignes med en F-fordeling med 1 og 2 frihedsgrader.

3 ☐ $Q = 2.65 \cdot 10^{-60}$ som skal sammenlignes med en F-fordeling med 3 og 177 frihedsgrader.

4 ☐ $Q = 334.0$ som skal sammenlingnes med en F-fordeling med 2 og 174 frihedsgrader.

5 ☐ $Q =$ `T1`^{2}`+T2`^{2}`+T3`^{2} som skal sammenlignes med en F-fordeling med 3 og 177 frihedsgrader.

Fortsæt på side 21

Opgave XII

En konsulent har modtaget data for komme- og gå-tider for 35 medarbejdere på en given arbejdsplads. Komme- og gå-tider er registreret for de samme 35 medarbejdere på to forskellige dage - én dag om sommeren og én dag om vinteren. Konsulenten ønsker nu at vurdere om den gennemsnitlige arbejdstid er lige lang på de to dage.

Spørgsmål XII.1 (24)

Hvilken analyse er relevant at udføre?

1 ☐ For hvér komme- og gå-tid udregnes arbejdstiden. Man har nu to uafhængige stikprøver (én for dagen om sommeren og én for dagen om vinteren) med 35 målinger i hver. Gennemsnittene af disse stikprøver sammenlignes med en t-test med nulhypotesen $H_0 : \mu_1 = \mu_2$.

2 ☐ For hvér af de to dage beregnes den gennemsnitlige ankomsttid og den gennemsnitlige gå-tid. Nu udfører man to t-test: én t-test undersøger om der er signifikant forskel på ankomst-tiden og den anden t-test undersøger om der er signifikant forskel på gå-tiden.

3 ☐ For hvér komme- og gå-tid udregnes arbejdstiden. Nu har man to parrede stikprøver (én for dagen om sommeren og én for dagen om vinteren) med 35 målinger i hver. Med en parret t-test undersøges det om den gennemsnitlige forskel i arbejdstid er signifikant forskellig fra nul.

4 ☐ For hvér af de to dage beregnes et 95% konfidens interval for den gennemsnitlige ankomsttid $CI_{\bar{x}_{arrive}}$ og for den gennemsnitlige gå-tid $CI_{\bar{x}_{leave}}$. Hvis de to konfidensintervaller ikke overlapper er der signifikant forskel på den gennemsnitlige arbejdstid på de to dage.

5 ☐ For hvér komme- og gå-tid udregnes en samlet arbejdstid. Nu har man to stikprøver med 35 målinger. Men en one-way ANOVA model undersøges det om der er forskel på den gennemsnitlige arbejdstid mellen de to dage.

Fortsæt på side 22

Opgave XIII

I et experiment har man antaget at sammenhængen mellem et input (x) og et output (y) er givet ved

$$Y_i = f(x_i) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2),$$

og at ϵ_i er iid. Da funktionen f er ukendt har man besluttet at tilpasse (fitte) modellen

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2).$$

Modellens parametre er estimeret, dvs. man har parameter estimater ($\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3]^T$) og en tilsvarende covarians matrice $\boldsymbol{\Sigma}_\beta$ (med $(\boldsymbol{\Sigma}_\beta)_{kl} = \sigma_{kl}$ lig med kovariansen mellem $\hat{\beta}_k$ og $\hat{\beta}_l$). Man ønsker nu at undersøge hvor ekstremum for andengrads polynomiet ligger, svarende til at løse

$$f'(x) = \beta_2 + 2\beta_3 x = 0,$$

for x. Løsningen benævnes $x^*(\boldsymbol{\beta})$, og det antages at $\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_\beta)$.

Spørgsmål XIII.1 (25)

Ved brug af error propagation hvad bliver da approksimationen af variansen af x^*?

1 ☐ $V[x^*] \approx \frac{\hat{\beta}_1^2}{\hat{\beta}_3^2}\sigma_{11} + \frac{1}{4\hat{\beta}_3^2}\sigma_{22} + \frac{\hat{\beta}_2^2}{4\hat{\beta}_3^4}\sigma_{33}$

2 ☐ $V[x^*] \approx \frac{1}{2\hat{\beta}_3^2}\left(\frac{1}{2}\sigma_{22} + \frac{\hat{\beta}_2^2}{2\hat{\beta}_3^2}\sigma_{33}\right)$

3 ☐ $V[x^*] \approx \frac{\hat{\beta}_1^2}{\hat{\beta}_3^2}\sigma_{11} + \frac{1}{4\hat{\beta}_3^2}\sigma_{22} + \frac{\hat{\beta}_2^2}{4\hat{\beta}_3^4}\sigma_{33} - \frac{\hat{\beta}_1\hat{\beta}_2}{\hat{\beta}_3^3}\sigma_{12} + \frac{\hat{\beta}_1}{\hat{\beta}_3^2}\sigma_{13} - \frac{\hat{\beta}_2}{\hat{\beta}_3^3}\sigma_{23}$

4 ☐ $V[x^*] \approx \frac{1}{2\hat{\beta}_3^2}\left(\frac{1}{2}\sigma_{22} + \frac{\hat{\beta}_2^2}{2\hat{\beta}_3^2}\sigma_{33} - \frac{\hat{\beta}_2}{\hat{\beta}_3}\sigma_{23}\right)$

5 ☐ $V[x^*] \approx \frac{\hat{\beta}_1^2}{\hat{\beta}_3^2}\sigma_{11} + \frac{1}{4\hat{\beta}_3^2}\sigma_{22} + \frac{\hat{\beta}_2^2}{4\hat{\beta}_3^4}\sigma_{33} + \frac{\hat{\beta}_1\hat{\beta}_2}{\hat{\beta}_3^3}\sigma_{12} + \frac{\hat{\beta}_2}{\hat{\beta}_3^3}\sigma_{23}$

Fortsæt på side 23

Antag nu at man i et konkret studie har observeret

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}; \quad \boldsymbol{\Sigma}_{\beta} = \begin{bmatrix} 0.2 & 0 & -0.01 \\ 0 & 0.01 & 0 \\ -0.01 & 0 & 0.001 \end{bmatrix}$$

Spørgsmål XIII.2 (26)

Ved brug af simulation fra den antagede fordeling, i hvilket interval ligger da standard afvigelsen for x^* (brug mindst 1000 gentagelser)?

1 ☐ [0.09;0.14]

2 ☐ [0.04;0.08]

3 ☐ [0.01;0.03]

4 ☐ [0.001;0.005]

5 ☐ [0.25;1.25]

Parametriseringen af 2. grads polynomiet ovenfor er ikke unik og kan formuleres som

$$Y_i = \beta_1 p_0(x_i) + \beta_2 p_1(x_i) + \beta_3 p_2(x_i) + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2),$$

hvor $p_j(x)$ er et j'et grads polynomium.

Spørgsmål XIII.3 (27)

Hvis antallet af observationer er $n = 11$ og $[x_1, x_2, ...x_{11}] = [-5, -4, ..., 4, 5]$, hvilket sæt af polynomier giver så et orthogonalt design (dvs. $(\boldsymbol{X}^T\boldsymbol{X})_{ji} = 0$ for $i \neq j$)?

1 ☐ $p_0(x_i) = \frac{1}{11}$, $p_1(x_i) = x_i - 5$, $p_2(x_i) = x_i^2 - x_i - 25$.

2 ☐ $p_0(x_i) = 1$, $p_1(x_i) = x_i - 2.5$, $p_2(x_i) = x_i^2 - 10$.

3 ☐ $p_0(x_i) = 1$, $p_1(x_i) = x_i$, $p_2(x_i) = x_i^2 - 10$.

4 ☐ $p_0(x_i) = 1$, $p_1(x_i) = x_i$, $p_2(x_i) = x_i^2$.

5 ☐ $p_0(x_i) = \frac{1}{11}$, $p_2(x_i) = x_i + 5$, $p_1(x_i) = x_i^2 - 25$.

Fortsæt på side 24

Opgave XIV

Med henblik på at bestemme vægten af et bestemt objekt, vejer to personer objektet et antal gange på en vægt. Det antages at målefejlen på vægten er normalfordelt med middelværdi 0 og varians σ^2, og at alle vejninger er uafhængige. De to personer vejer objektet hhv. 2 og 3 gange og rapporterer gennemsnittet ($\overline{Y}_1$ og $\overline{Y}_2$).

Spørgsmål XIV.1 (28)

Hvis objektets vægt er lig μ, hvad er så middelværdi og varians af $\overline{\overline{Y}} = \frac{1}{2}\overline{Y}_1 + \frac{1}{2}\overline{Y}_2$?

1 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \sigma^2$

2 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \frac{\sigma^2}{2}$

3 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \frac{5\sigma^2}{6}$

4 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \frac{5\sigma^2}{24}$

5 ☐ $E(\overline{\overline{Y}}) = \mu$, $V(\overline{\overline{Y}}) = \frac{\sigma^2}{5}$

Med henblik på at konstruere en teststørrelse regner man et vægtet gennemsnit ($w \in (0,1)$)

$$\widetilde{\overline{Y}} = w\overline{Y}_1 + (1-w)\overline{Y}_2,$$

således at $\widetilde{\overline{Y}}$ og $\overline{Y}_i - \widetilde{\overline{Y}}$ er uafhængige.

Spørgsmål XIV.2 (29)

For hvilket w gælder at $Cov(\overline{Y}_1 - \widetilde{\overline{Y}}, \widetilde{\overline{Y}}) = 0$?

1 ☐ $w = \frac{1}{2}$

2 ☐ $w = \frac{1}{3}$

3 ☐ $w = \frac{2}{5}$

4 ☐ $w = \frac{5}{12}$

5 ☐ $w = \frac{5}{24}$

Fortsæt på side 25

Ved hjælp af resultaterne ovenfor har man konstrueret 2 uafhængige beregningsstørrelser, Q_1 og Q_2, således at $\frac{k_1}{\sigma^2}Q_1 \sim \chi^2(1)$ og $\frac{k_2}{\sigma^2}Q_2 \sim \chi^2(1)$, hvor k_1 og k_2 er konstanter.

Spørgsmål XIV.3 (30)

Hvilket af følgende udsagn er korrekt?

1 ☐ $\frac{1}{2\sigma^2}\frac{k_1Q_1}{k_2Q_2} \sim F(2,1)$.

2 ☐ $\frac{k_1Q_1}{k_2Q_2} \sim F(1,1)$.

3 ☐ $\frac{1}{2}\frac{Q_1/k_1}{Q_2/k_2} \sim F(1,2)$.

4 ☐ $\frac{Q_1}{Q_2} \sim F(1,1)$.

5 ☐ $\frac{Q_1/k_1}{Q_2/k_2} \sim F(1,1)$.