English assigment follows the Danish version

Opgavesæt til:

Skriftlig prøve: 20. December 2025

Kursus navn og nr.: **02403 Introduktion til matematisk statistik**

Varighed: 4 timer

Tilladte hjælpemidler: Alle skriftlige hjælpemidler samt lommeregner model TI30XS eller TI30XB

For at besvare spørgsmålene skal du udfylde et særskilt "Answer Sheet".

Opgavesættet består af 30 spørgsmål af "multiple choice" typen, som er fordelt på 16 opgaver. **Du skal kun aflevere dit "Answer Sheet" og ikke hele opgavesættet.**

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.*

Brug af Python til denne eksamen: *Denne eksamen indeholder Python kode. Bemærk at vi bruger de følgende biblioteker og forkortelser:*

```
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
```

Fortsæt på side 2

Opgave I

Du trækker mange tilfældige stikprøver, hver af størrelse $n = 150$, fra en population der er skæv til højre (hvilket betyder, at fordelingen har en hale, der strækker sig langt til højre). Fordelingen har en middelværdi på 40 og en standardafvigelse på 10.

Spørgsmål I.1 (1)

Ifølge den centrale grænseværdisætning (CLT), hvilket af følgende udsagn om fordelingen af stikprøvernes middelværdier er korrekt?

1 ☐ Fordelingen af stikprøvemiddelværdierne vil blive meget skæv til højre.

2 ☐ Middelværdien af stikprøvernes middelværdier vil være meget større end 40 på grund af skævheden.

3 ☐ Fordelingen af stikprøvernes middelværdier vil være omtrent normalfordelt og centreret omkring 40.

4 ☐ Standardafvigelsen for stikprøvernes middelværdier vil være 10, det samme som populationen.

5 ☐ Fordelingen af stikprøvemiddelværdierne vil blive uniform, efterhånden som stikprøvestørrelsen stiger.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 4

Opgave II

Den diskrete stokastiske variabel X har følgende fordeling:

x	1	3	5	6
$f(x)$	0.1	0.4	0.3	0.2

hvor $f(x) = P(X = x)$ er tæthedsfunktionen (pdf).

Spørgsmål II.1 (2)

Hvad er gennemsnittet $\mathbf{E}[X]$ og variansen $\mathbf{V}[X]$?

1 ☐ $\mathbf{E}[X] = 3.75$ og $\mathbf{V}[X] = 1.92$

2 ☐ $\mathbf{E}[X] = 4.00$ og $\mathbf{V}[X] = 1.92$

3 ☐ $\mathbf{E}[X] = 4.40$ og $\mathbf{V}[X] = 2.40$

4 ☐ $\mathbf{E}[X] = 3.75$ og $\mathbf{V}[X] = 2.40$

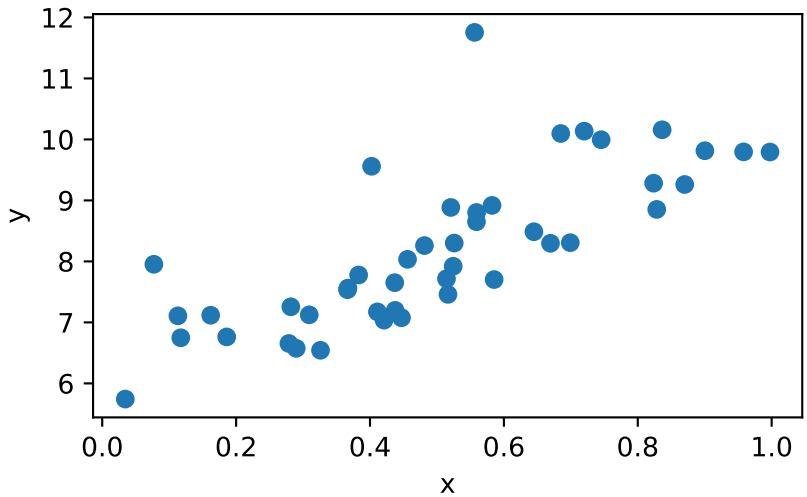
5 ☐ $\mathbf{E}[X] = 4.00$ og $\mathbf{V}[X] = 2.40$

6 ☐ Ved ikke / Intet svar

Fortsæt på side 5

Opgave III

Der er indsamlet noget data, for hvilket vi kalder de observerede værdier "y"og "x". Data er visualiseret i scatter-plottet nedenfor.



Følgende information om data er givet:

$$\bar{x} = 0.5024$$

$$\bar{y} = 8.1964$$

$$Sxx = \sum_{i}^{n} (x_i - \bar{x})^2 = 2.5365$$

$$Sxy = \sum_{i}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 10.2153$$

Hvor n er antallet af observationer i data.

Data blev gemt i Python i en DataFrame kaldet "dat", som indeholder kolonnerne "x"og "y". En simpel lineær regressionsmodel blev fittet til data ved hjælp af følgende kommando i Python:

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

Den resulterende regressionstabel er givet nedenfor (selvom visse værdier er erstattet af bogstaverne A, B, C og D):

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

                            OLS Regression Results
==============================================================================
Dep. Variable:                          y   R-squared:                       0.583
Model:                                OLS   Adj. R-squared:                  0.573
No. Observations:                      45   F-statistic:                     60.12
Covariance Type:                nonrobust   Prob (F-statistic):           1.06e-09
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.1731      0.289     21.388      0.000       5.591       6.755
x                 A          B       7.754      0.000         C           D
==============================================================================
```

Spørgsmål III.1 (3)

Betragt den statistiske model, der blev fittet til data, og de værdier, der er repræsenteret af bogstaverne A og B (indsat i regressionstabellen ovenfor).

Hvilket af følgende udsagn er korrekt?

1 ☐ $A = \hat{\beta}_1$ er estimatet for parameteren β_1 i den statistiske model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, med $\varepsilon_i \sim N(0, \sigma^2)$.
$B = \hat{\sigma}_{\hat{\beta}_1}$ er den estimerede standardfejl for $\hat{\beta}_1$.

2 ☐ $A = \hat{\beta}_0$ er estimatet for parameteren β_0 i den statistiske model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, med $\varepsilon_i \sim N(0, \sigma^2)$.
$B = \hat{\sigma}_{\hat{\beta}_0}$ er den estimerede standardfejl for $\hat{\beta}_0$.

3 ☐ $A = \hat{\beta}_1$ er estimatet for parameteren β_1 i den statistiske model:
$y_i = \beta_1 x_i + \varepsilon_i$, med $\varepsilon_i \sim N(0, \sigma^2)$.
$B = \hat{\sigma}_{\hat{\beta}_1}$ er den estimerede standardfejl for $\hat{\beta}_1$.

4 ☐ $A = \hat{\beta}_0$ er estimatet for parameteren β_0 og $B = \hat{\beta}_1$ er estimatet for parameteren β_1 i den statistiske model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, med $\varepsilon_i \sim N(0, \sigma^2)$.

5 ☐ $A = \hat{\beta}_1$ er estimatet for parameteren β_1 og $B = \hat{\beta}_0$ er estimatet for parameteren β_0 i den statistiske model:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, med $\varepsilon_i \sim N(0, \sigma^2)$.

6 ☐ Ved ikke / Intet svar

Spørgsmål III.2 (4)

Hvad er værdien af A?

1 ☐ $A = 0.5024$

2 ☐ $A = 10.2153$

3 ☐ $A = 5.10765$

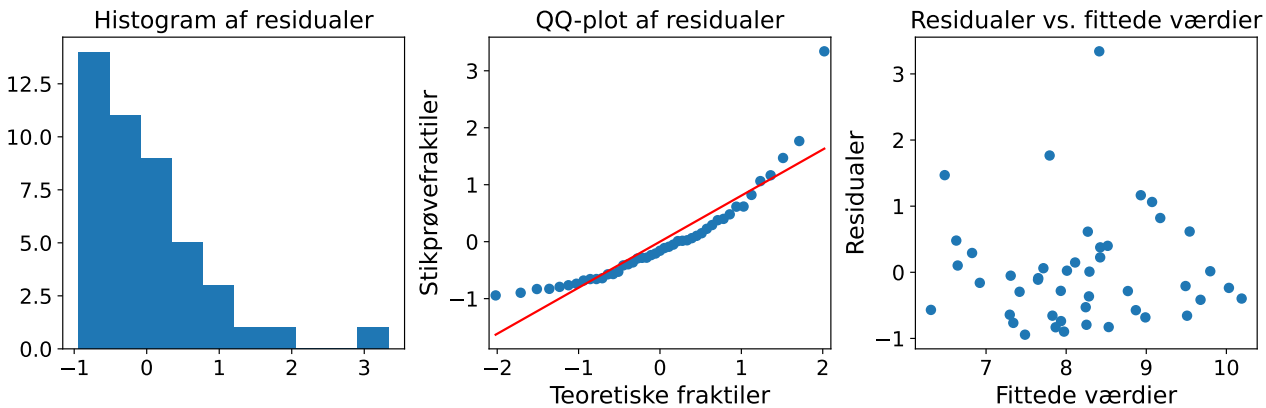
4 ☐ $A = 45/10 = 4.5$

5 ☐ $A = 4.0273$

6 ☐ Ved ikke / Intet svar

Spørgsmål III.3 (5)

For at undersøge, om modelantagelserne holder, blev følgende plots produceret:



Hvilket af følgende udsagn er korrekt?

1 ☐ Histogrammet og qq-plottet af residualerne indikerer, at antagelsen om normalitet ($\varepsilon_i \sim N(0, \sigma^2)$) er brudt - dvs. residualerne ser ikke ud til at følge en normalfordeling.

2 ☐ QQ-plottet af residualerne indikerer, at antagelsen om uafhængighed er brudt - dvs. residualerne ser ikke ud til at være uafhængige.

3 ☐ Plots indikerer ikke et brud på modelantagelserne.

4 ☐ Scatter-plottet af residualerne vs. fittede værdier indikerer, at antagelsen om uafhængighed er brudt - dvs. residualerne ser ikke ud til at være uafhængige.

5 ☐ QQ-plottet af residualerne indikerer, at residualerne følger en normalfordeling med middelværdi nul: $\varepsilon_i \sim N(0, \sigma^2)$.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 8

Opgave IV

En gruppe forskere undersøger celleaktivitet i fire forskellige arter af mus. De indsamler stikprøver med stikprøvestørrelser: $n_1 = 50$, $n_2 = 150$, $n_3 = 150$, $n_4 = 50$ for henholdsvis art 1, 2, 3 og 4. Forskerne estimerer en matematisk model med følgende form:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad i \in \{1, 2, 3, 4\},$$

hvor fejlene ε_{ij} antages at være uafhængige.

Forskerne beregner, at $SS(Tr) = 6.0479$ (her refererer "Tr"til de forskellige arter) og $SST = 163.234$, hvilket giver en p-værdi på 0.0018, når man tester nulhypotesen:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

Spørgsmål IV.1 (6)

Hvilke af følgende udsagn angiver den korrekte værdi af teststørrelsen fra F-testen, og angiver også en korrekt konklusion?

1 ☐ Værdien af teststørrelsen er $F = 5.079$. Nulhypotesen kan ikke forkastes på signifikansniveau $\alpha = 0.01$.

2 ☐ Værdien af teststørrelsen er $F = 5.079$. Nulhypotesen kan ikke forkastes på signifikansniveau $\alpha = 0.001$.

3 ☐ Værdien af teststørrelsen er $F = 2.96$. Nulhypotesen forkastes på signifikansniveau $\alpha = 0.01$.

4 ☐ Værdien af teststørrelsen er $F = 2.96$. Nulhypotesen forkastes på signifikansniveau $\alpha = 0.001$.

5 ☐ Værdien af teststørrelsen er $F = 0.99$. Nulhypotesen forkastes på signifikansniveau $\alpha = 0.05$.

6 ☐ Ved ikke / Intet svar

Spørgsmål IV.2 (7)

Projektionsmatricen hørende til modellen er

$$\boldsymbol{H} = \begin{bmatrix} \frac{1}{n_1}\boldsymbol{E}_{n_1,n_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\boldsymbol{E}_{n_2,n_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{n_3}\boldsymbol{E}_{n_3,n_3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1}{n_4}\boldsymbol{E}_{n_4,n_4} \end{bmatrix},$$

hvor $\boldsymbol{E}_{n_i,n_i}$ er en $n_i \times n_i$ matrice af et-taller. Under nul-hypothesen $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, er projektions matricen

$$\boldsymbol{H}_0 = \frac{1}{n}\boldsymbol{E}_{n,n},$$

hvor $n = n_1 + n_2 + n_3 + n_4$.

Givet at nul-hypothesen er sand, hvad er da fordelingen af

$$Q = \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{Y},$$

hvor σ^2 er den sande (men ukendte) varians af ε_i?

1 ☐ $Q \sim F(4, n-3)$.

2 ☐ $Q \sim \chi^2(3)$.

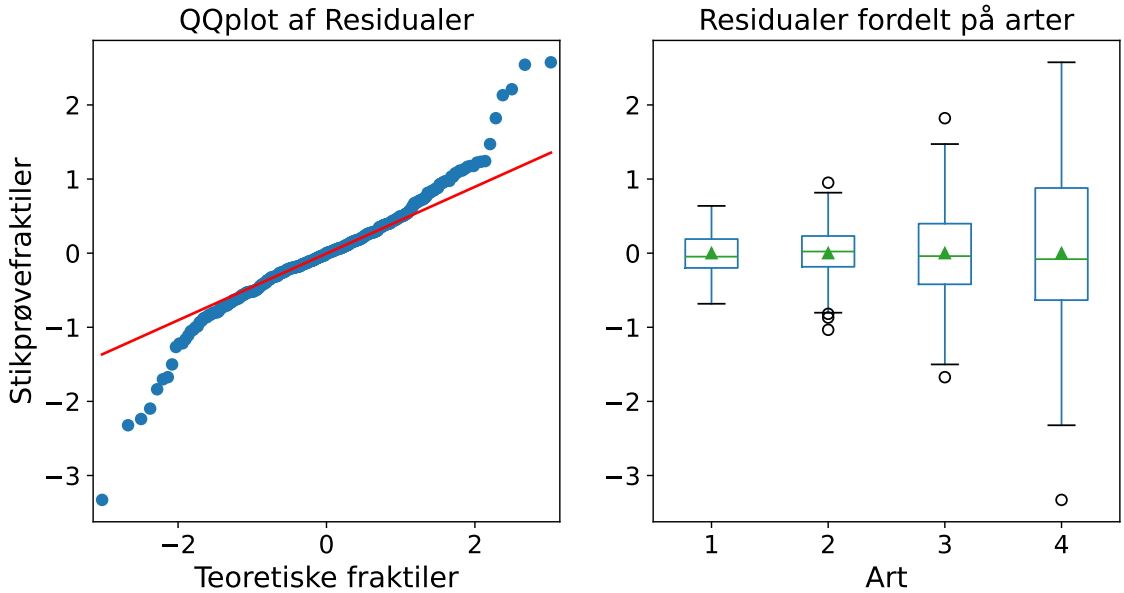
3 ☐ $Q \sim \chi^2(n-4)$.

4 ☐ $Q \sim F(4, n)$.

5 ☐ $Q \sim F(3, n-4)$.

Spørgsmål IV.3 (8)

Forskerne udfører nu modelvalidering (dvs. check af modellen). De producerer de nedenfor viste diagnostiske plots:



Hvilket af følgende udsagn er korrekt (alle argumenter skal være sande)?

1 ☐ QQ-plottet indikerer en overtrædelse af antagelsen om normalfordelte parametre. Boxplots indikerer en ukorrekt estimering af modelparametrene.

2 ☐ QQ-plottet indikerer en overtrædelse af antagelsen om normalfordelte residualer. Boxplots indikerer en overtrædelse af antagelsen om ens residualvarians på tværs af arter.

3 ☐ QQ-plottet indikerer en overtrædelse af antagelsen om normalfordelte residualer. Boxplots indikerer en overtrædelse af antagelsen om, at $\mu_{\varepsilon_{ij}} = 0$.

4 ☐ QQ-plottet indikerer en overtrædelse af antagelsen om ens stikprøvestørrelse inden for hver gruppe. Boxplots indikerer et uforholdsmæssigt antal outliers i dataene.

5 ☐ De diagnostiske plots indikerer ikke nogen overtrædelse af modelantagelserne.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 12

Opgave V

En forsker hævder, at den gennemsnitlige daglige skærmtid til uddannelsesmæssig brug blandt universitetsstuderende er 6 timer. For at teste denne påstand blev en tilfældig stikprøve på $n = 20$ studerende udvalgt, hvilket gav: $\bar{x} = 5.4$ timer og $s = 1.2$ timer. Antag, at skærmtiden er tilnærmelsesvis normalfordelt. En hypotesetest for nulhypotesen $H_0 : \mu = 6$ udføres, og den beregnede p-værdi er 0.03.

Spørgsmål V.1 (9)

Hvilket af følgende udsagn er korrekt?

1 ☐ Ved at bruge et signifikansniveau på $\alpha = 0.05$ forkastes nulhypotesen $H_0 : \mu = 6$. Forskeren konkluderer, at den gennemsnitlige skærmtid til uddannelsesmæssig brug ikke er 6 timer.

2 ☐ Ved at bruge et signifikansniveau på $\alpha = 0.05$ accepteres den alternative hypotese $H_A : \mu = 0$. Forskeren konkluderer, at den gennemsnitlige skærmtid til uddannelsesmæssig brug er signifikant mindre end 6 timer.

3 ☐ Ved at bruge et signifikansniveau på $\alpha = 0.01$ forkastes nulhypotesen $H_0 : \mu = 6$. Forskeren konkluderer, at den gennemsnitlige skærmtid til uddannelsesmæssig brug ikke er 6 timer.

4 ☐ Ved at bruge et signifikansniveau på $\alpha = 0.01$ accepteres den alternative hypotese $H_A : \mu = 0$. Forskeren konkluderer, at den gennemsnitlige skærmtid til uddannelsesmæssig brug er signifikant mindre end 6 timer.

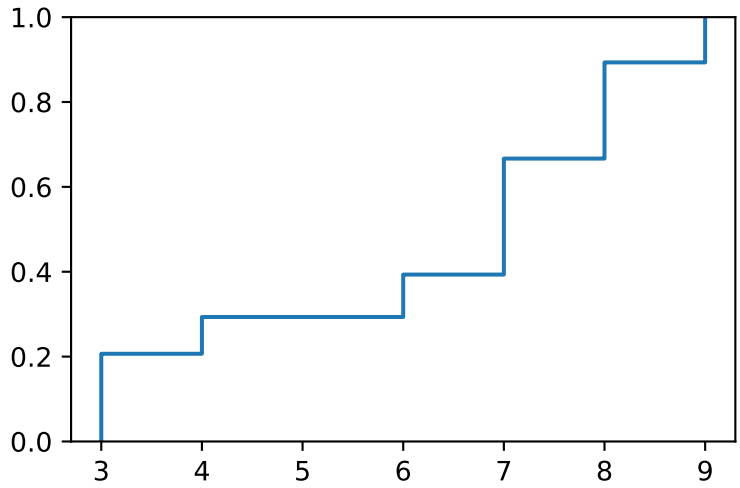
5 ☐ Der er ikke nok information til at træffe en beslutning om H_0.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 13

Opgave VI

150 observationer af en diskret stokastisk variabel er simuleret i Python, og de resulterende værdier er visualiseret i følgende empiriske kumulative fordelingsplot (ecdf-plot):



Spørgsmål VI.1 (10)

Hvilken af følgende Python-koder kunne generere de simulerede observationer?

1 ☐ `np.random.choice(a=[3,4,5,6,7,8,9], size=150, p=[1/7,1/7,1/7,1/7,1/7,1/7,1/7])`

2 ☐ `stats.uniform.rvs(loc=3, scale=6, size=150)`

3 ☐ `stats.uniform.rvs(loc=3, scale=9, size=100)`

4 ☐ `np.random.choice(a=[3,4,6,7,8,9], size=150, p=[2/10,1/10,1/10,3/10,2/10,1/10])`

5 ☐ `stats.norm.rvs(loc=6, scale=2, size=150)`

6 ☐ Ved ikke / Intet svar

Fortsæt på side 14

Opgave VII

En forsker undersøger, om studerendes aktive deltagelse i forelæsninger er relateret til højere kursustilfredshed. I et lille pilotstudie med $n = 15$ studerende blev det konstateret, at omkring 60% af de studerende, der ofte deltog i forelæsninger, vurderede deres samlede kursustilfredshed som "høj". Da pilotstudiet er relativt lille, er estimatet på 60% relateret til en høj grad af usikkerhed. For at planlægge et større studie til det næste semester ønsker forskeren at estimere den sande andel af "høj kursustilfredshed"blandt studerende, der ofte deltager i forelæsninger. Forskeren ønsker at give et 95% konfidensinterval for denne andel, hvor *Margin of Error* (ME) kun er 0.05.

For at besvare følgende spørgsmål skal du muligvis bruge følgende fraktil fra standardnormalfordelingen: $z_{0,975} = 1.96$.

Spørgsmål VII.1 (11)

Hvilken minimumsstikprøvestørrelse n kræves til opfølgningsundersøgelsen?

1 ☐ $n = 78$

2 ☐ $n = 185$

3 ☐ $n = 240$

4 ☐ $n = 369$

5 ☐ $n = 412$

6 ☐ Ved ikke / Intet svar

Spørgsmål VII.2 (12)

Hvis der ikke var noget forudgående estimat på p tilgængeligt, hvad er så den nødvendige stikprøvestørrelse?

1 ☐ $n = 138$

2 ☐ $n = 185$

3 ☐ $n = 240$

4 ☐ $n = 385$

5 ☐ $n = 420$

6 ☐ Ved ikke / Intet svar

Spørgsmål VII.3 (13)

Uden at tage hensyn til resultaterne fra de to foregående spørgsmål, beslutter forskeren at udføre en undersøgelse med en tilfældig stikprøve på 200 studerende. Blandt de 200 studerende blev 50 klassificeret som "Aktive deltagere", og 150 blev klassificeret som "Ikke aktive deltagere". Forskeren ønsker at sammenligne andelen af "høj kursustilfredshed"mellem de to grupper og opnår følgende undersøgelsesresultater:

Aktive deltagere: 30 ud af 50 berettede "høj kursustilfredshed"
Ikke aktive deltagere: 75 ud af 150 berettede "høj kursustilfredshed"

Forskeren udfører en hypotesetest for andele i to stikprøver, hvor han sammenligner andelen af studerende, der beretter "høj kursustilfredshed", inden for de to grupper. Forskeren bruger et signifikansniveau på $\alpha = 0.05$.

Hvilken af følgende konklusioner er korrekt (både beregningen og konklusionen skal være korrekte)?

1 ☐ Den beregnede $z_{obs} = 1.23 < 1.96$, så vi kan ikke forkaste nulhypotesen om lige store andele inden for de to grupper ($H_0 : p_1 = p_2$)

2 ☐ Den beregnede $z_{obs} = 2.01 > 1.96$, så vi forkaster nulhypotesen om lige store andele inden for de to grupper ($H_0 : p_1 = p_2$) og konkluderer, at der er en signifikant forskel mellem de to grupper.

3 ☐ Den beregnede $z_{obs} = 1.19 < 1.96$, så vi kan ikke forkaste nulhypotesen om lige store andele inden for de to grupper ($H_0 : p_1 = p_2$)

4 ☐ Den beregnede $z_{obs} = 1.23 < 1.96$, så vi forkaster nulhypotesen om lige store andele inden for de to grupper ($H_0 : p_1 = p_2$) og konkluderer, at der er en signifikant forskel mellem de to grupper.

5 ☐ Den beregnede $z_{obs} = 2.01 > 1.96$, så vi kan ikke forkaste nulhypotesen om lige store andele inden for de to grupper ($H_0 : p_1 = p_2$).

6 ☐ Ved ikke / Intet svar

Fortsæt på side 16

Spørgsmål VII.4 (14)

I et andet studie ønsker forskeren at undersøge, om de studerendes deltagelsesniveau i forelæsninger er forbundet med deres eksamensresultater. Denne gang er de studerendes deltagelsesniveau klassificeret som "Højt", "Moderat"eller "Lavt", og eksamensresultaterne klassificeres som "Højt", "Mellem"eller "Lavt". I en tilfældig stikprøve på 200 studerende bliver deltagelsesniveau og eksamensresultat registreret og præsenteret i tabellen nedenfor:

	Eksamensresultat:			
Deltagelsesniveau:	Højt	Mellem	Lavt	i alt
Højt	30	15	5	50
Moderat	25	30	15	70
Lavt	10	20	50	80
i alt	65	65	70	200

Forskeren ønsker at teste, om fordelingen af eksamensresultater er uafhængig af deltagelsesniveauet.

Det ønskede signifikans niveau α er blevet gemt i Python i en variabel kaldet "`alpha`". Hvilket af følgende Python-udsagn beregner korrekt den kritiske værdi for den relevante hypotesetest?

1 ☐ `critical_value = stats.chi2.ppf(1 - alpha/2, df=4)`

2 ☐ `critical_value = stats.chi2.ppf(1 - alpha, df=4)`

3 ☐ `critical_value = stats.f.ppf(1 - alpha, dfn=2, dfd=4)`

4 ☐ `critical_value = stats.f.cdf(1 - alpha, dfn=2, dfd=6)`

5 ☐ `critical_value = stats.t.cdf(1 - alpha/2, df=8)`

6 ☐ Ved ikke / Intet svar

Fortsæt på side 17

Opgave VIII

Betragt en multipel lineær regressionsmodel skrevet på matriceformen:

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

for hvilken designmatricen $\mathbf{X}$ er som følger (her vises kun de øverste rækker af matricen):

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & (x_1)^3 \\ 1 & x_2 & (x_2)^3 \\ 1 & x_3 & (x_3)^3 \\ 1 & x_4 & (x_4)^3 \\ 1 & x_5 & (x_5)^3 \\ 1 & x_6 & (x_6)^3 \\ 1 & x_7 & (x_7)^3 \\ 1 & x_8 & (x_8)^3 \\ 1 & x_9 & (x_9)^3 \\ 1 & x_{10} & (x_{10})^3 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Spørgsmål VIII.1 (15)

Hvordan kan den samme regressionsmodel skrives (i ikke-matriceformulering)?

1 ☐ $y_i = \beta_0 + \beta_a a_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

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

3 ☐ $y_i = \beta_0 + \beta_1 x_i + 2 \cdot \beta_2 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

4 ☐ $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^3 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

5 ☐ $y_i = \beta_0 + \beta_1 (x_i + x_i^3), \quad \varepsilon_i \sim N(0, \sigma^2)$

6 ☐ Ved ikke / Intet svar

Spørgsmål VIII.2 (16)

Modelparametrene estimeres ved hjælp af formlen (i matriceformulering):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Hvilket udsagn er korrekt:

1 ☐ Vektoren $\hat{\boldsymbol{\beta}}$ har fire elementer ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}^2$) og værdierne af disse parametre vælges, således at summen af de kvadrerede residualer ($\sum_{i=1}^{n} \varepsilon_i^2$) minimeres.

2 ☐ Vektoren $\hat{\boldsymbol{\beta}}$ har tre elementer ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$) og værdierne af disse parametre vælges, således at den totale kvadratsum ($SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$) minimeres.

3 ☐ Vektoren $\hat{\boldsymbol{\beta}}$ har tre elementer ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$) og værdierne af disse parametre vælges, således at summen af de kvadrerede residualer ($\sum_{i=1}^{n} \varepsilon_i^2$) minimeres.

4 ☐ Vektoren $\hat{\boldsymbol{\beta}}$ har tre elementer ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$) og værdierne af disse parametre vælges, således at variansen af x-værdierne (s_x^2) maksimeres.

5 ☐ Vektoren $\hat{\boldsymbol{\beta}}$ har to elementer ($\hat{\beta}_0$, $\hat{\beta}_1$) og værdierne af disse parametre vælges, således at $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ minimeres.

6 ☐ Ved ikke / Intet svar

Spørgsmål VIII.3 (17)

Lad n angive antallet af observationer i datasættet, og lad videre $\hat{\sigma}^2$ angive den sædvanlige ikke biased (eller centrale) varians estimator for regressions modellen. Følgende tal er nu udregnet

$$Q = \mathbf{y}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$$

Hvad er værdien af Q?

1 ☐ $(n-1)\hat{\sigma}^2$

2 ☐ $(n-3)\hat{\sigma}^2$

3 ☐ $(n-2)\hat{\sigma}^2$

4 ☐ $\hat{\sigma}^2$

5 ☐ $n\hat{\sigma}^2$

6 ☐ Ved ikke / Intet svar

Spørgsmål VIII.4 (18)

Nu overvejes en anden model

$$\mathbf{Y} = \mathbf{X}_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$$

Matricen $\boldsymbol{X}_2$ er konstrueret således at

$$\boldsymbol{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^{-1}\mathbf{X}_2^T = \boldsymbol{H}_2$$

Hvilket af følgende udsagn er (generelt) sandt?

1 ☐ Parameterestimaterne ($\hat{\boldsymbol{\beta}}$) er ens i de to tilfælde, men de tilpassede (fittede) værdier ($\hat{\boldsymbol{Y}}$) kan være forskellige.

2 ☐ Parameterestimaterne ($\hat{\boldsymbol{\beta}}$) og de tilpassede (fittede) værdier ($\hat{\boldsymbol{Y}}$) er lig hinanden i de to tilfælde.

3 ☐ De tilpassede (fittede) værdier ($\hat{\boldsymbol{Y}}$) er ens i de to modeller, men varians estimatet ($\hat{\sigma}^2$) er muligvis forkellig.

4 ☐ De tilpassede (fittede) værdier ($\hat{\boldsymbol{Y}}$) er ens for de to modeller, men de estimerede parametre ($\hat{\boldsymbol{\beta}}$) kan være forskellige.

5 ☐ Designmatricerne er ens ($\boldsymbol{X} = \boldsymbol{X}_2$) og derfor er alle værdier (parameter estimater, tilpassede (fittede) værdier, og varianser) de samme for de to modeller.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 20

Opgave IX

En forsker indsamler en tilfældig stikprøve af størrelsen $n = 10$ fra en normalfordelt population. Stikprøvestandardafvigelsen er $s = 4.2$.

For at besvare dette spørgsmål skal du muligvis bruge følgende fraktiler fra χ^2-fordelingen med ν frihedsgrader:

$\chi^2_{0.025}(\nu = 9) = 2.70, \quad \chi^2_{0.975}(\nu = 9) = 19.02$

Spørgsmål IX.1 (19)

Find 95%-konfidensintervallet for populationens standardafvigelse σ.

1 ☐ $[2.93,\ 7.25]$

2 ☐ $[2.89,\ 7.67]$

3 ☐ $[3.10,\ 7.38]$

4 ☐ $[2.85,\ 7.80]$

5 ☐ $[3.40,\ 8.20]$

6 ☐ Ved ikke / Intet svar

Fortsæt på side 21

Opgave X

En fitnesscoach vil teste, om et nyt 4-ugers træningsprogram har en signifikant effekt på hvilepulsen. Hun måler hvilepulsen (i slag pr. minut) hos 8 deltagere før og efter programmet.

	Hvilepuls:							
Før programmet:	78	85	90	76	88	82	79	84
Efter programmet:	74	80	85	72	83	78	75	80

Antag, at dataene følger normalfordelinger, og at testen udføres ved signifikansniveau $\alpha = 0.05$.

Spørgsmål X.1 (20)

Antag at dataene er blevet læst ind i Python i variabler kaldet `Før` og `Efter`.
Hvilken Python-kommando skal bruges til korrekt at teste, om programmet ændrede den gennemsnitlige hvilepuls signifikant?

1 ☐ `stats.ttest_1samp(Efter, popmean=75)`

2 ☐ `stats.ttest_ind(Før, Efter, equal_var=False)`

3 ☐ `stats.ttest_1samp(Før - Efter, popmean=Før.mean()-Efter.mean())`

4 ☐ `stats.ttest_1samp(Før - Efter, popmean=0)`

5 ☐ `stats.ttest_ind(Før, Efter, equal_var=True)`

6 ☐ Ved ikke / Intet svar

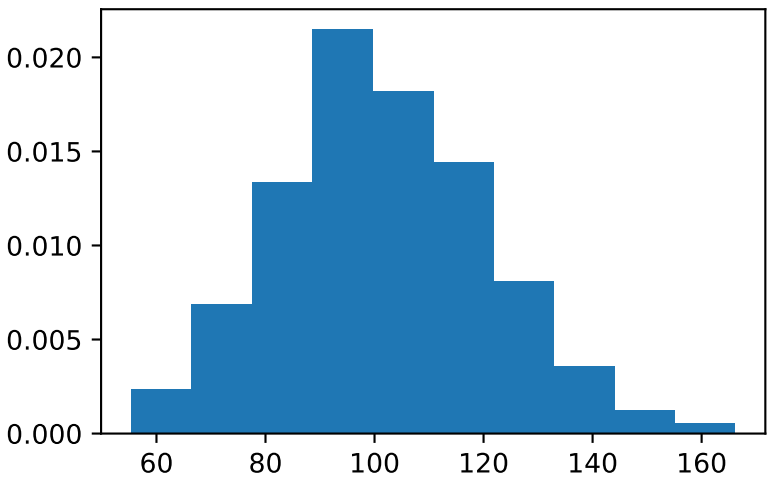
Fortsæt på side 22

Opgave XI

En simulering er blevet udført med følgende Python-kode:

```
A = stats.uniform.rvs(size=500, loc=0, scale = 5)
B = stats.norm.rvs(size=500, loc=10, scale = 1)
C = A + B**2

plt.hist(C, density=True)
plt.show()
```



Spørgsmål XI.1 (21)

Hvilke fordelinger følger de stokastiske variable A og B?

1 ☐ A følger en normalfordeling med middelværdi $\mu_A = 0$ og standardafvigelse $\sigma_A = 5$.
B følger en normalfordeling med middelværdi $\mu_B = 10$ og standardafvigelse $\sigma_B = 1$.

2 ☐ A følger en uniform fordeling med middelværdi $\mu_A = 0$ og standardafvigelse $\sigma_A = 5$.
B følger en normalfordeling med middelværdi $\mu_B = 10$ og standardafvigelse $\sigma_B = 1$.

3 ☐ A er uniformt fordelt mellem $\alpha_A = 0$ og $\beta_A = 5$.
B følger en normalfordeling med middelværdi $\mu_B = 10$ og standardafvigelse $\sigma_B = 1$.

4 ☐ A er uniformt fordelt mellem $\alpha_A = -5$ og $\beta_A = 5$.
B følger en normalfordeling med middelværdi $\mu_B = 10$ og standardafvigelse $\sigma_B = 1$.

5 ☐ Fordelingerne for A og B kan ikke bestemmes ud fra Python-koden.

6 ☐ Ved ikke / Intet svar

Spørgsmål XI.2 (22)

Fra simuleringen ser vi, at $C = A + B^2$.

Antag, at vi udfører en ny simulering, hvor $\mu_A = 2.5$, $\sigma_A = 5/\sqrt{12}$, $\mu_B = 10$ og $\sigma_B = 1$.

Brug fejlophobningsloven til at estimere σ_C. Hvilket af følgende er korrekt?

1 ☐ $\sigma_C = \sqrt{5/\sqrt{12} + 20}$

2 ☐ $\sigma_C = 5/\sqrt{12} + 200$

3 ☐ $\sigma_C = \sqrt{25/12 + 400}$

4 ☐ $\sigma_C = \sqrt{25 + 400}$

5 ☐ $\sigma_C = \sqrt{25/12}$

6 ☐ Ved ikke / Intet svar

Fortsæt på side 24

Opgave XII

En forsker planlægger et eksperiment for at estimere den gennemsnitlige reduktion i blodtrykket, når et specifikt lægemiddel gives. Fra et pilotstudie er den estimerede standardafvigelse i blodtrykket $\sigma = 8$ mmHg. Forskeren ønsker at detektere en gennemsnitlig *effektstørrelse* (forskel i gennemsnitligt blodtryk) på $\mu_0 - \mu_1 = 5$ mmHg med et signifikansniveau på $\alpha = 0.05$ og en styrke på $1 - \beta = 0.80$. Forskeren planlægger at udføre eksperimentet på en stikprøve af n forsøgspersoner, hvor hver forsøgsperson får målt blodtrykket både med og uden lægemidlet.

For at løse denne øvelse skal du muligvis bruge følgende fraktiler fra en standard normalfordeling:

$z_{1-\beta} = z_{0.80} = 0.84$ og $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Spørgsmål XII.1 (23)

Hvad er den nødvendige stikprøvestørrelse til eksperimentet, givet ovenstående information?

1 ☐ $n = 40$

2 ☐ $n = 20$

3 ☐ $n = 41$

4 ☐ $n = 5$

5 ☐ $n = 21$

6 ☐ Ved ikke / Intet svar

Fortsæt på side 25

Opgave XIII

For at få indsigt i anvendelsen af generative AI-værktøjer (såsom ChatGPT eller Copilot) til læringsstøtte, undersøgte et universitetsforskningshold en pilotgruppe på $n = 50$ studerende. Blandt disse beretter $x = 30$ studerende, at de bruger sådanne AI-værktøjer regelmæssigt til at hjælpe med studieopgaver såsom at skrive, kode eller revidere koncepter.

For at løse denne opgave skal du muligvis bruge en af følgende fraktiler:

Fra en standard normalfordeling: $z_{0.95} = 1.64$ og $z_{0.975} = 1.96$.

Fra en t-fordeling (med ν frihedsgrader): $t_{0.95}(\nu = 49) = 1.68$ og $t_{0.975}(\nu = 49) = 2.01$.

Spørgsmål XIII.1 (24)

Hvilket af følgende er det korrekte 95%-konfidensinterval (givet at man anvender metoden beskrevet i lærebogen) for den sande andel (p) af studerende, der regelmæssigt bruger generative AI-værktøjer til læring, og hvad er antagelserne bag beregningen (både konfidensinterval og argument skal være sande)?

1 ☐ 95%-konfidensintervallet for den sande andel er $[0.46; 0.74]$. Beregningen er baseret på antagelsen om, at stikprøvestørrelsen (n) er stor nok til, at stikprøveandelen ($\hat{p}$) er tilnærmelsesvis normalfordelt.

2 ☐ 95%-konfidensintervallet for den sande andel er $[0.46; 0.74]$. Beregningen er baseret på antagelsen om, at stikprøvestørrelsen (n) følger en binomialfordeling.

3 ☐ 95%-konfidensintervallet for den sande andel er $[0.48,\ 0.72]$. Beregningen er baseret på antagelsen om, at stikprøvestørrelsen (n) er stor nok til, at den centrale grænseværdisætning er gyldig.

4 ☐ 95%-konfidensintervallet for den sande andel er $[0.48,\ 0.72]$. Beregningen er baseret på antagelsen om, at stikprøvestørrelsen (n) er stor nok til, at t-fordelingen kan approksimeres godt af en standard normalfordeling.

5 ☐ 95%-konfidensintervallet for den sande andel er $[0.48,\ 0.72]$. Beregningen er baseret på antagelsen om, at stikprøvestørrelsen (n) følger en binomialfordeling.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 26

Opgave XIV

Et forsikringsselskab modtager skadesanmeldelser i henhold til en Poisson-proces med en intensitet på to anmeldelser pr. måned. Dette indebærer, at anmeldelserne ankommer uafhængigt, og at ventetiden mellem to på hinanden følgende anmeldelser følger en eksponentialfordeling med en middelværdi på en halv måned. Dermed følger antallet af anmeldelser, som forsikringsselskabet modtager i løbet af en given måned, en Poissonfordeling med en middelværdi på to anmeldelser pr. måned.

Spørgsmål XIV.1 (25)

Hvad er sandsynligheden for, at forsikringsselskabet modtager færre end to skadesanmeldelser i løbet af en given måned??

1 ☐ 13.5%

2 ☐ 27.1%

3 ☐ 30.3%

4 ☐ 40.6%

5 ☐ 50.5%

6 ☐ Ved ikke / Intet svar

Spørgsmål XIV.2 (26)

Hvilket udtryk beregner korrekt sandsynligheden for, at ventetiden mellem to på hinanden følgende skadesanmeldelser overstiger én måned?

1 ☐ $1 - \exp(-\frac{1}{2})$

2 ☐ $\int_0^1 2\exp(-2x)dx$

3 ☐ $\int_1^\infty 2\exp(-2x)dx$

4 ☐ $\int_1^\infty \frac{1}{2}\exp(-\frac{1}{2}x)dx$

5 ☐ $\int_0^1 \frac{1}{2}\exp(-\frac{1}{2}x)dx$

6 ☐ Ved ikke / Intet svar

Fortsæt på side 27

Opgave XV

Betragt modellen for en tosidet variansanalyse med fire forskellige "Behandlinger"og fem forskellige "Blokke":

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad i \in \{1, ..., 4\}, \quad j \in \{1, ...5\},$$

hvor fejlene ε_{ij} antages at være uafhængige, i refererer til de fire forskellige behandlinger og j refererer til de fem forskellige blokke.

Man har indhentet et datasæt og beregnet følgende gennemsnitsværdier:

Samlet	31.5	Blok 1	31.0
Behandling 1	35.1	Blok 2	32.0
Behandling 2	29.4	Blok 3	32.3
Behandling 3	28.7	Blok 4	30.3
Behandling 4	32.8	Blok 5	31.9

Det oplyses desuden, at kvadratafvigelsessummen $SS(Tr)$ er 134.5 (her refererer "Tr"til behandlingerne).

Lad desuden $\hat{\sigma}$ betegne modellens estimat af σ.

Spørgsmål XV.1 (27)

Hvad er kovarianserne $\text{Cov}(Y_{12}, Y_{13})$ og $\text{Cov}(Y_{22}, Y_{42})$ ifølge modellen?

1 ☐ $\text{Cov}(Y_{12}, Y_{13}) = 0$ og $\text{Cov}(Y_{22}, Y_{42}) = 0$

2 ☐ $\text{Cov}(Y_{12}, Y_{13}) = 0$ og $\text{Cov}(Y_{22}, Y_{42}) = \beta_2$

3 ☐ $\text{Cov}(Y_{12}, Y_{13}) = \alpha_1$ og $\text{Cov}(Y_{22}, Y_{42}) = 0$

4 ☐ $\text{Cov}(Y_{12}, Y_{13}) = \alpha_1$ og $\text{Cov}(Y_{22}, Y_{42}) = \beta_2$

5 ☐ $\text{Cov}(Y_{12}, Y_{13}) = \mu + \alpha_1$ og $\text{Cov}(Y_{22}, Y_{42}) = \mu + \beta_2$

6 ☐ Ved ikke / Intet svar

Spørgsmål XV.2 (28)

En test af nulhypotesen $H_0 : \alpha_i = 0$ for $i = 1, ..., 4$ giver anledning til hvilken værdi af F-teststørrelsen?

1 ☐ $F = \frac{44.83}{\hat{\sigma}^2}$

2 ☐ $F = \frac{44.83}{\hat{\sigma}}$

3 ☐ $F = \frac{134.5}{\hat{\sigma}^2}$

4 ☐ $F = \frac{538}{\hat{\sigma}}$

5 ☐ $F = \frac{538}{\hat{\sigma}^2}$

6 ☐ Ved ikke / Intet svar

Fortsæt på side 29

Opgave XVI

Lad X være en standard normal fordelt stokastisk variabel, og lad $Y_1 \sim \chi^2(5)$ og $Y_2 \sim \chi^2(5)$. Det antages at X, Y_1 og Y_2 er uafhængige.

Spørgsmål XVI.1 (29)

Hvad er middelværdien af

$$Y = \frac{Y_1}{X^2 + Y_2} \quad ?$$

1 ☐ $E[Y] = \frac{5}{4}$

2 ☐ $E[Y] = \frac{6}{4}$

3 ☐ $E[Y] = \frac{6}{4}$

4 ☐ $E[Y] = \frac{5}{6}$

5 ☐ $E[Y] = \frac{5}{3}$

Spørgsmål XVI.2 (30)

Hvad er a, hvis følgende skal gælde

$$P\left(\frac{X}{\sqrt{Y_1}} < a\right) = 0.95 \quad ?$$

1 ☐ $a = t_{0.95}(5)\sqrt{4}$

2 ☐ $a = \frac{t_{0.95}(5)}{\sqrt{5}}$

3 ☐ $a = F_{0.95}(1,5)\sqrt{5}$

4 ☐ $a = t_{0.95}(4)\sqrt{5}$

5 ☐ $a = \frac{F_{0.95}(1,4)}{\sqrt{4}}$

SÆTTET ER SLUT. Nyd ferien!