

English assignment follows the Danish version

Opgavesæt til:

Skriftlig prøve: 20. December 2025

Kursus navn og nr.: **02323 Introduktion til 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 19 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 $\mathbf{V}[X] = 2.40$

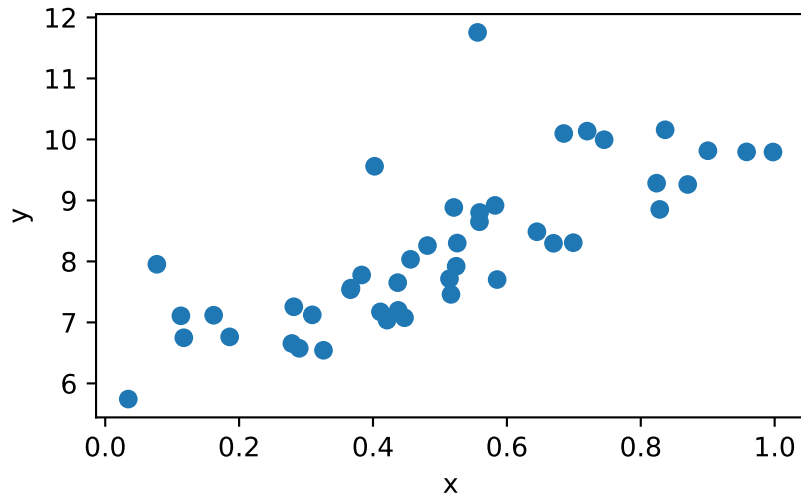
5 ☐ $\mathbf{E}[X] = 4.00$ og $\mathbf{V}[X] = 2.40$

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

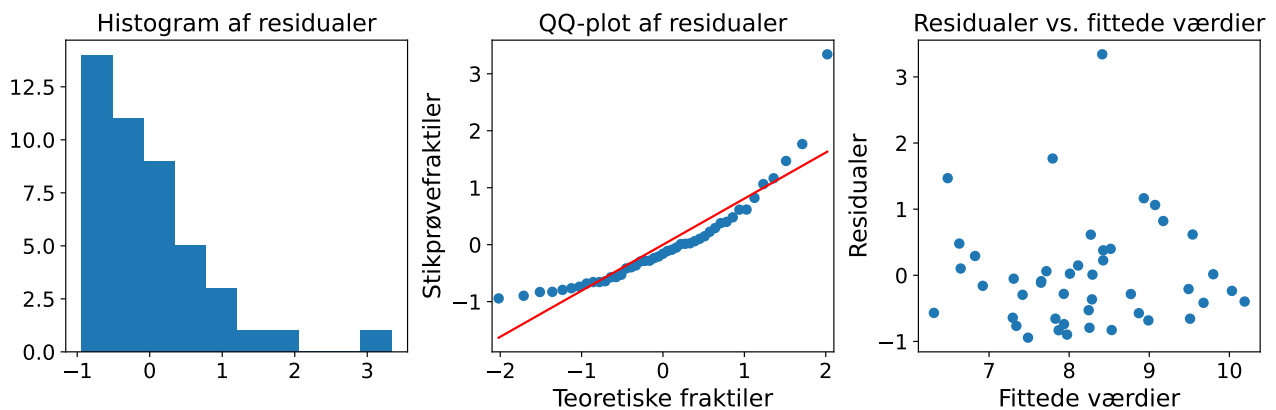
Spørgsmål III.2 (4)

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 industriel ingeniør analyserer en maskines produktions effektivitetsmålinger (på en skala fra 0–100) over 30 på hinanden følgende dage (målingerne kan antages at være uafhængige). Formålet er at vurdere, hvor konsistent maskinen præsterer, samt at identificere mulige outliers med usædvanlig høj eller lav effektivitet. Kvartilerne for de daglige effektivitetsmålinger er følgende:

$$Q_1 = 72, \quad Q_2 = 78, \quad Q_3 = 88.$$

Spørgsmål IV.1 (6)

Hvilket af følgende udsagn om det interkvartile spænd (Inter Quartile Range, IQR) er korrekt?

- 1 ☐ IQR måler den samlede spredning mellem den mindste og største effektivitetsscore i datasættet.
- 2 ☐ IQR beregnes som $Q_1 - Q_3 = 72 - 88 = -16$, hvilket repræsenterer dataenes venstreskævhed.
- 3 ☐ IQR'en er $Q_3 - Q_1 = 88 - 72 = 16$, hvilket repræsenterer spændet af de midterste 50% af effektivitetsscorerne.
- 4 ☐ IQR'en er lig med medianen ($Q_2 = 78$), som deler datasættet i to lige store halvdele.
- 5 ☐ IQR bruges kun på normalfordelte data til at beregne standardafvigelsen.
- 6 ☐ Ved ikke / Intet svar

Fortsæt på side 10

Opgave V

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:

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

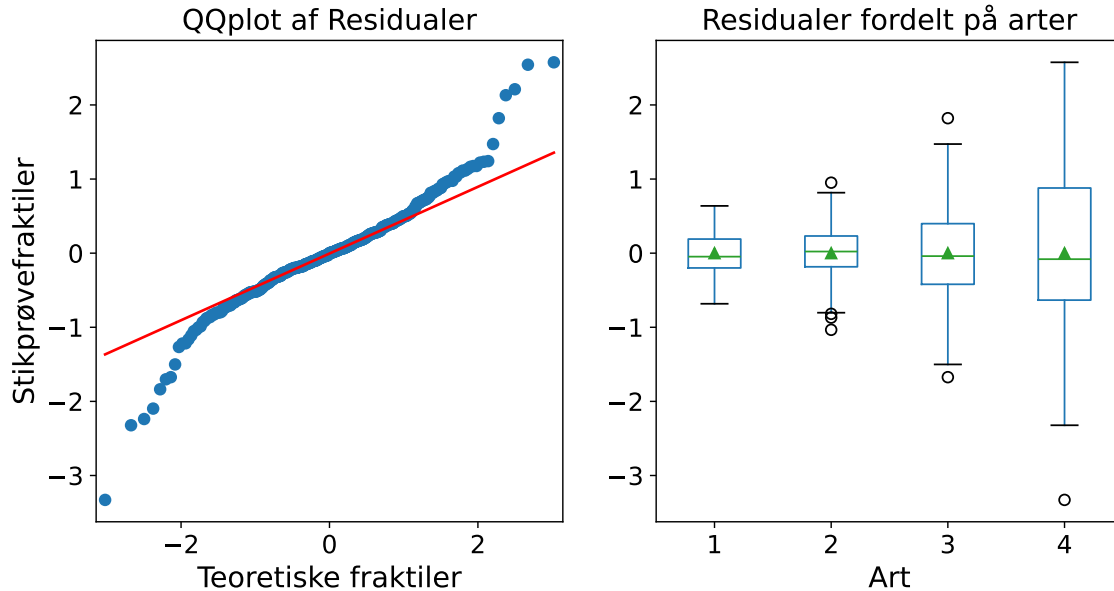
Spørgsmål V.1 (7)

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 V.2 (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 VI

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 VI.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 VII

En underviser indsamler tilfredshedsvurderinger (på en skala fra 1-10) for tre interaktive værktøjer, der bruges under forelæsninger — Kahoot, Vevox og Socrative — for at forbedre de studerendes engagement og konceptuelle forståelse. Statistiske nøgletal (gennemsnit, standardafvigelse og variationskoefficient) for de studerendes vurderinger er vist nedenfor:

Software	Mean	St. Dev.	$CV \times 100\%$
Kahoot	9.0	0.6	6.7%
Vevox	8.2	1.0	12.2%
Socrative	7.0	1.8	25.7%

Spørgsmål VII.1 (10)

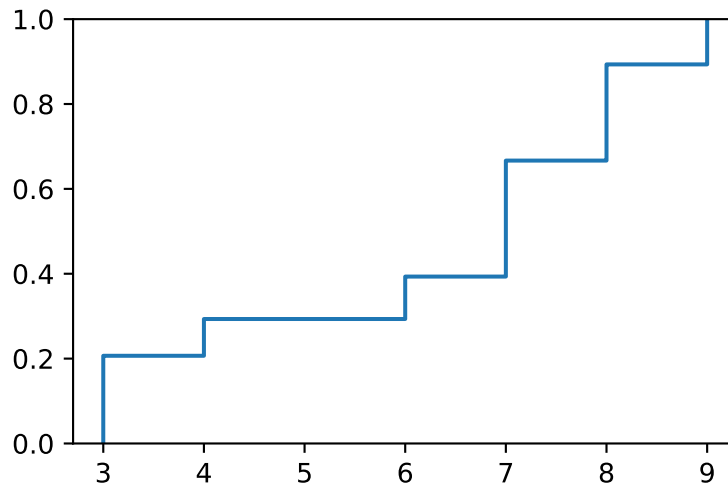
Hvilket af følgende udsagn fortolker disse resultater korrekt?

- 1 ☐ Kahoot har den højeste gennemsnitlige vurdering og laveste relative variation, hvilket indikerer, at eleverne konsekvent vurderede det højt.
- 2 ☐ Vevox har den mindste standardafvigelse, hvilket indikerer, at den fik de bedste vurderinger.
- 3 ☐ Socratives CV på 25,7% tyder på, at de har den højeste studentertilfredshed og de mest ensartede vurderinger.
- 4 ☐ Kahoot har den højeste gennemsnitlige vurdering, men også de største udsving i individuelle vurderinger.
- 5 ☐ CV-værdierne kan ikke sammenlignes, da stikprøvestørrelserne er forskellige.
- 6 ☐ Ved ikke / Intet svar

Fortsæt på side 14

Opgave VIII

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 VIII.1 (11)

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 15

Opgave IX

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 IX.1 (12)

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 IX.2 (13)

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 IX.3 (14)

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 17

Spørgsmål IX.4 (15)

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:

Deltagelsesniveau:	Eksamensresultat:			i alt
	Højt	Mellem	Lavt	
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.

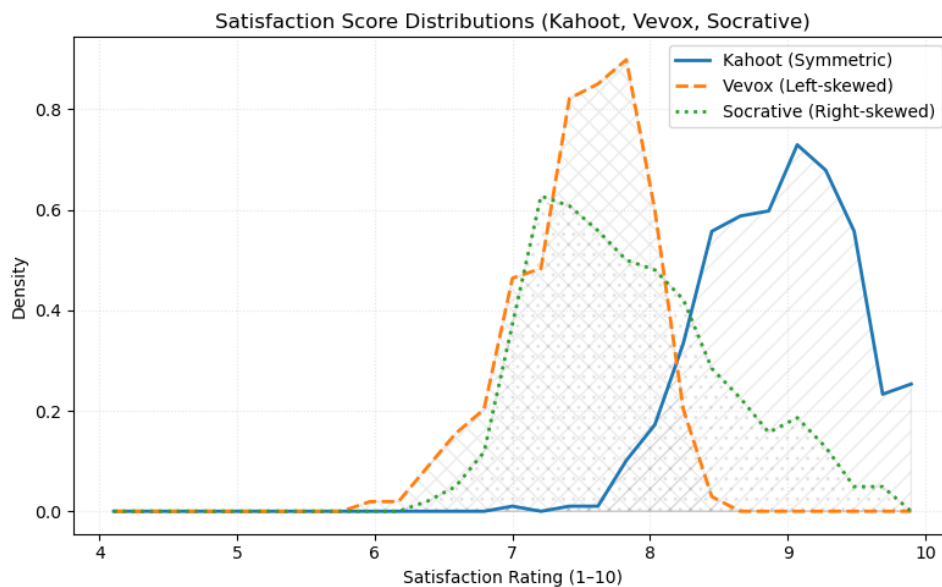
Det ønskede signifikans niveau α er blevet gemt i Python i en variabel kaldet "**alpha**". Hvilket af følgende Python-udsagn beregner korrekt den kritiske værdi for den relevante hypotesetest?

- 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 18

Opgave X

Diagrammet nedenfor viser fordelingen af studerendes tilfredshedsvurderinger (en score på skala 1-10) for tre interaktive læringsværktøjer, der anvendes under forelæsninger: Kahoot, Vevox og Socrative. Hver kurve repræsenterer, hvordan de studerende vurderede værktøjerne med hensyn til engagement og konceptuel forståelse.



Spørgsmål X.1 (16)

Hvilket af følgende udsagn fortolker korrekt forholdet mellem middelværdi- og medianvurderingerne for Vevox og Socrative?

- 1 ☐ Fordelingen af vurderinger for Vevox er venstreskæv, hvilket betyder, at middelværdien er større end medianen, og fordelingen af vurderinger for Socrative er højreskæv, hvilket betyder, at middelværdien er mindre end medianen.
- 2 ☐ Fordelingen af vurderinger for Vevox er venstreskæv, hvilket betyder, at middelværdien er mindre end medianen, og fordelingen af vurderinger for Socrative er højreskæv, hvilket betyder, at middelværdien er større end medianen.
- 3 ☐ For begge fordelinger er middelværdien og medianen identiske, fordi tilfredshedsvurderingerne er ensartet fordelt.
- 4 ☐ Vevox' venstreskæve fordeling betyder, at dens middelværdi er større end Socratives middelværdi.
- 5 ☐ Da begge fordelinger er skæve (ikke symmetriske), kan vi ikke vide, om middelværdierne er større eller mindre end medianen.

6 ☐ Ved ikke / Intet svar

Spørgsmål X.2 (17)

Hvilket af følgende udsagn beskriver korrekt hovedformålet med hver af følgende visualiseringsmetoder: boksplot, histogram, cirkeldiagram, søjlediagram og et kumulativt fordelingsplot?

- 1 ☐ Et boksplot viser kategoriske andele, et histogram viser procentdele af kategorier, et cirkeldiagram illustrerer numerisk spredning, et søjlediagram viser symmetri i data, og et kumulativt fordelingsplot identificerer outliers.
- 2 ☐ Et boksplot opsummerer fordelingen af data ved hjælp af kvartiler og fremhæver outliers; et histogram viser, hvordan kvantitative data er fordelt på intervaller; et cirkeldiagram illustrerer andelen af kategorier i en helhed; et søjlediagram sammenligner kategoriske frekvenser; og et kumulativt fordelingsplot viser andelen af frekvenser op til hver værdi.
- 3 ☐ Et boksplot visualiserer antal; et histogram sammenligner gruppegennemsnit; et cirkeldiagram er bedst til kontinuerlige data; et søjlediagram repræsenterer kvartiler; og et kumulativt fordelingsdiagram viser diskrete tællinger uden akkumulering.
- 4 ☐ Et boksplot og et histogram viser begge kategorisk data; et cirkeldiagram og en søjlediagram bruges til kontinuerte data; og et kumulativt fordelingsplot bruges til at sammenligne grupper.
- 5 ☐ Alle fem teknikker viser de samme oplysninger, men adskiller sig kun i farve og layout.
- 6 ☐ Ved ikke / Intet svar

Fortsæt på side 20

Opgave XI

Lad X og Y være stokastiske variable, som begge følger en kontinuert, uniform fordeling på intervallet $[0, 2]$:

$$X \sim \mathcal{U}(0, 2) \quad \text{og} \quad Y \sim \mathcal{U}(0, 2)$$

Desuden er X og Y uafhængige.

Spørgsmål XI.1 (18)

Lad $Z = X - Y$. Hvad er variansen af Z ?

- 1 ☐ Variansen af Z er: $\mathbf{V}[Z] = 4$
- 2 ☐ Variansen af Z er: $\mathbf{V}[Z] = 2$
- 3 ☐ Variansen af Z er: $\mathbf{V}[Z] = \frac{2}{3}$
- 4 ☐ Variansen af Z er: $\mathbf{V}[Z] = \frac{2}{12}$
- 5 ☐ Variansen af Z kan ikke bestemmes ud fra de tilgængelige oplysninger.
- 6 ☐ Ved ikke / Intet svar

Spørgsmål XI.2 (19)

Lad X^* være den standardiserede version af X :

$$X^* = \frac{X - \mathbf{E}[X]}{\sqrt{\mathbf{V}[X]}}$$

Hvilken fordeling følger X^* ?

- 1 ☐ En kontinuert, uniform fordeling på intervallet $[0, 1]$.
- 2 ☐ En kontinuert, uniform fordeling på intervallet $[-1, 1]$.
- 3 ☐ En kontinuert, uniform fordeling på intervallet $[-\sqrt{3}, \sqrt{3}]$.
- 4 ☐ En standardnormalfordeling.
- 5 ☐ Ingen af de ovenstående fordelinger.
- 6 ☐ Ved ikke / Intet svar

Fortsæt på side 21

Opgave XII

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:

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

Spørgsmål XII.1 (20)

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 22

Opgave XIII

En fitnesscoach vil teste, om et nyt 4-ugers træningsprogram har en signifikant effekt på hvilepuls. Hun måler hvilepuls (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 XIII.1 (21)

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

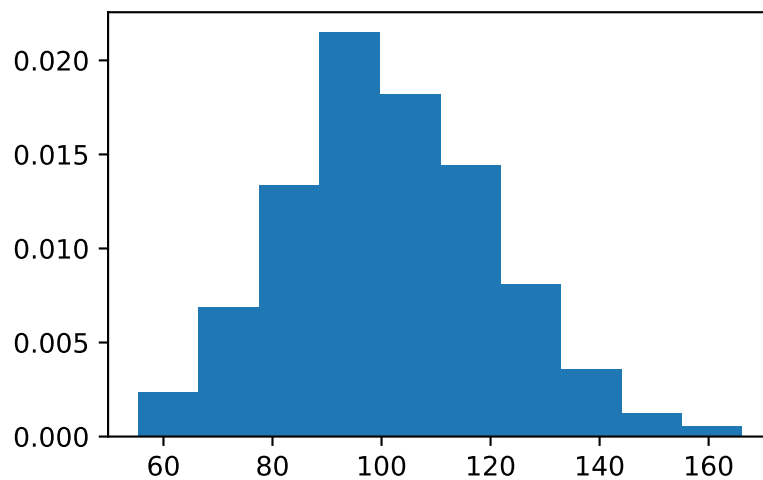
Fortsæt på side 23

Opgave XIV

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 XIV.1 (22)

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 XIV.2 (23)

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 25

Opgave XV

En dataanalytiker i en softwarevirksomhed undersøger forholdet mellem udviklererfaring (målt i år) og deres gennemsnitlige kodekvalitetsscore (bedømt fra 0-100). Stikprøvekorrelationskoefficienten mellem de to størrelser viser sig at være $r = 0.45$, hvilket indikerer en moderat positiv korrelation. Når dataene plottes, observerer analytikeren dog, at både udviklere med meget kort og meget lang erfaring har lavere kodekvalitetsscore, mens dem med moderat erfaring klarer sig bedst – hvilket danner et kurvet (parabolsk) mønster på scatterplottet.

Spørgsmål XV.1 (24)

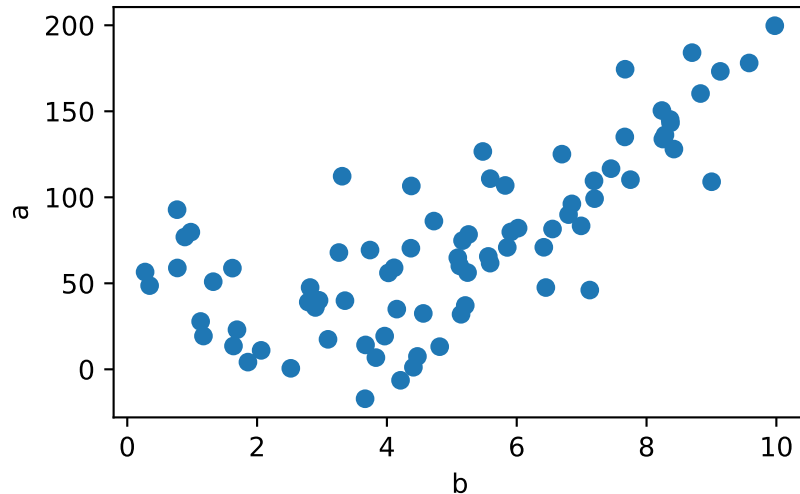
Hvilket af følgende udsagn beskriver bedst denne situation?

- 1 ☐ Stikprøvekorrelationskoefficienten på $r = 0.45$ indfanger korrekt den stærke ikke-lineære sammenhæng mellem erfaring og kodekvalitet.
- 2 ☐ Den moderate korrelation antyder en svag sammenhæng, og scatterplottet bekræfter, at der ikke er noget mønster i dataene.
- 3 ☐ Det kurvede scatterplot indikerer en ikke-lineær sammenhæng, så stikprøvekorrelationskoefficienten er ikke en god estimator for den sande sammenhæng mellem erfaring og kodekvalitet.
- 4 ☐ Stikprøvekorrelationskoefficienten på $r = 0.45$ beviser, at forholdet mellem erfaring og kodekvalitet er lineært og af moderat størrelse.
- 5 ☐ Stikprøvekorrelationskoefficienten bliver meningsløs, medmindre begge variable er kategoriale.
- 6 ☐ Ved ikke / Intet svar

Fortsæt på side 26

Opgave XVI

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



Data blev gemt i Python i en DataFrame kaldet "D", som indeholder kolonnerne "a", "b" og "b2", hvor "b2"-kolonnen indeholder kvadrerede værdier af "b" ($b2 = b^2$).

En lineær regressionsmodel blev tilpasset data ved hjælp af følgende kommando i Python:

```
fit = smf.ols(formula= 'a ~ b + b2', data=D).fit()
```

Den resulterende regressionstabel og nogle ekstra informationer er givet nedenfor:

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

OLS Regression Results						
=====						
Dep. Variable:	a	R-squared:				0.692
Model:	OLS	Adj. R-squared:				0.684
No. Observations:	80	F-statistic:				86.59
Covariance Type:	nonrobust	Prob (F-statistic):				1.99e-20
=====						
	coef	std err	t	P> t	[0.025	0.975]

Intercept	58.9849	11.158	5.286	0.000	36.766	81.204
b	-16.4646	4.894	-3.364	0.001	-26.209	-6.720
b2	3.1350	0.479	6.539	0.000	2.180	4.090
=====						

```
print(fit.scale)
795.9021043987004
```

```
print(fit.pvalues)
Intercept    1.131837e-06
b            1.198447e-03
b2           6.124791e-09
dtype: float64
```

Spørgsmål XVI.1 (25)

Hvilken statistisk model er fittet, og hvad er det tilsvarende estimat for residualstandardafvigelsen?

- 1 ☐ Den tilpassede statistiske model er: $a_i = \mu + b_i + b_i^2 + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ (og ε_i 'erne er uafhængige). Den estimerede residualstandardafvigelse er $\hat{\sigma} = 795.9$.
- 2 ☐ Den tilpassede statistiske model er: $a_i = \beta_0 \cdot b_i + \beta_1 \cdot b_i^2 + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ (og ε_i 'erne er uafhængige). Den estimerede residualstandardafvigelse er $\hat{\sigma} = 28.21$.
- 3 ☐ Den tilpassede statistiske model er: $a_i = \beta_0 + \beta_1 \cdot b_i + \beta_2 \cdot b_i^2 + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ (og ε_i 'erne er uafhængige). Den estimerede residualstandardafvigelse er $\hat{\sigma} = 795.9$.
- 4 ☐ Den tilpassede statistiske model er: $a_i = \beta_0 + \beta_1 \cdot b_i + \beta_2 \cdot b_i^2 + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ (og ε_i 'erne er uafhængige). Den estimerede residualstandardafvigelse er $\hat{\sigma} = 28.21$.
- 5 ☐ Den tilpassede statistiske model er: $a_i = \mu + b_i + b_i^2 + \varepsilon_i$, $\varepsilon_i \sim N(10, \sigma^2)$ (og ε_i 'erne er uafhængige). Den estimerede residualstandardafvigelse er $\hat{\sigma} = 28.21$.
- 6 ☐ Ved ikke / Intet svar

Spørgsmål XVI.2 (26)

Beregn et 99%-konfidensinterval for parameteren, der er angivet i regressionstabellen som koeficient (coef) "b".

For at besvare dette spørgsmål kan du få brug for en af følgende fraktiler fundet ved hjælp af Python:

```
print(stats.t.ppf(0.975, df=80))
1.990063421028384
```

```
print(stats.t.ppf(0.99, df=80))  
2.373868271947045
```

```
print(stats.t.ppf(0.995, df=80))  
2.63869059374035
```

```
print(stats.t.ppf(0.975, df=78))  
1.990847068555052
```

```
print(stats.t.ppf(0.99, df=78))  
2.3751109570702686
```

```
print(stats.t.ppf(0.995, df=78))  
2.6403400123362197
```

```
print(stats.t.ppf(0.975, df=77))  
1.9912543951146038
```

```
print(stats.t.ppf(0.99, df=77))  
2.375756992900897
```

```
print(stats.t.ppf(0.995, df=77))  
2.6411976082359923
```

- 1 ☐ $[-26.21; -6.719]$
- 2 ☐ $[-28.09; -4.844]$
- 3 ☐ $[-29.39; -3.539]$
- 4 ☐ $[-19.11; -13.82]$
- 5 ☐ $[-29.38; -3.551]$
- 6 ☐ Ved ikke / Intet svar

Fortsæt på side 29

Opgave XVII

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* (forskell 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:

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

Spørgsmål XVII.1 (27)

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 30

Opgave XVIII

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 XVIII.1 (28)

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 31

Opgave XIX

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 forsikrings-selskabet modtager i løbet af en given måned, en Poissonfordeling med en middelværdi på to anmeldelser pr. måned.

Spørgsmål XIX.1 (29)

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 XIX.2 (30)

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

SÆTTET ER SLUT. Nyd ferien!