This exam paper is available in both Danish and English. The English version appears after the Danish version.

Opgavesæt:

Skriftlig prøve: 28.05.2026

Kursusnavn og -nr.: **02323 Introduktion til statistik**

Varighed: 4 timer

Tilladte hjælpemidler: Alle skriftlige hjælpemidler samt lommeregner af typen TI30XS eller TI30XB

De endelige svar skal afleveres ved udfyldelse af det særskilte "Answer Sheet".

Opgavesættet består af 30 multiple choice-spørgsmål fordelt på 11 opgaver. **Kun "Answer Sheet" skal afleveres; selve opgavesættet må ikke afleveres.**

Multiple choice-spørgsmål: *Der er i hvert spørgsmål én og kun én korrekt svarmulighed. Endvidere er det ikke givet, at alle de angivne svarmuligheder er meningsfulde. Ved beregninger skal du altid afrunde dit resultat til det antal decimaler, der er angivet i svarmulighederne, før du vælger dit svar.*

Brug af Python til denne eksamen: *Denne eksamen indeholder Python-kode. Bemærk, at vi anvender 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 3

Opgave I

Denne opgave omhandler basal Python funktionalitet.

Spørgsmål I.1 (1)

Hvilket output giver kommandoen `np.arange(10)`? (`np` henviser til NumPy pakken)

1 ☐ Arrayet `[1 2 3 4 5 6 7 8 9 10]`.

2 ☐ Arrayet `[1 2 3 4 5 6 7 8 9]`.

3 ☐ Arrayet `[0 1 2 3 4 5 6 7 8 9]`.

4 ☐ Arrayet `[0 1 2 3 4 5 6 7 8 9 10]`.

5 ☐ Et array med ti tilfældigt genererede tal mellem nul og et.

6 ☐ Ved ikke / Intet svar

Spørgsmål I.2 (2)

Hvilken funktion i `scipy.stats`-pakken danner tilfældigt genererede tal?

1 ☐ `cdf`-funktionen

2 ☐ `pdf`-funktionen

3 ☐ `pmf`-funktionen

4 ☐ `ppf`-funktionen

5 ☐ `rvs`-funktionen

6 ☐ Ved ikke / Intet svar

Fortsæt på side 4

Opgave II

En studerende ved DTU vil være sikker på, at hun aldrig kommer for sent til en tidlig morgenforelæsning. Over to uger måler hun derfor den tid, det tager hende at cykle til campus. Hun bruger Python og gemmer sine målinger i vektoren `x`.

Spørgsmål II.1 (3)

Hvordan kan hun beregne stikprøvevariansen af sine cykeltider? (`np` henviser til NumPy pakken.)

1 ☐ `np.mean(x)`

2 ☐ `np.std(x)`

3 ☐ `np.std(x, ddof = 1)`

4 ☐ `np.var(x)`

5 ☐ `np.var(x, ddof = 1)`

6 ☐ Ved ikke / Intet svar

Spørgsmål II.2 (4)

Den studerende vil modellere sine cykeltider med en normalfordeling. Hvilket af følgende udsagn omkring en normalfordeling er forkert?

1 ☐ Den har to parametre: middelværdien μ og variansen σ^2.

2 ☐ Den er symmetrisk omkring middelværdien.

3 ☐ Middelværdien er lig med medianen.

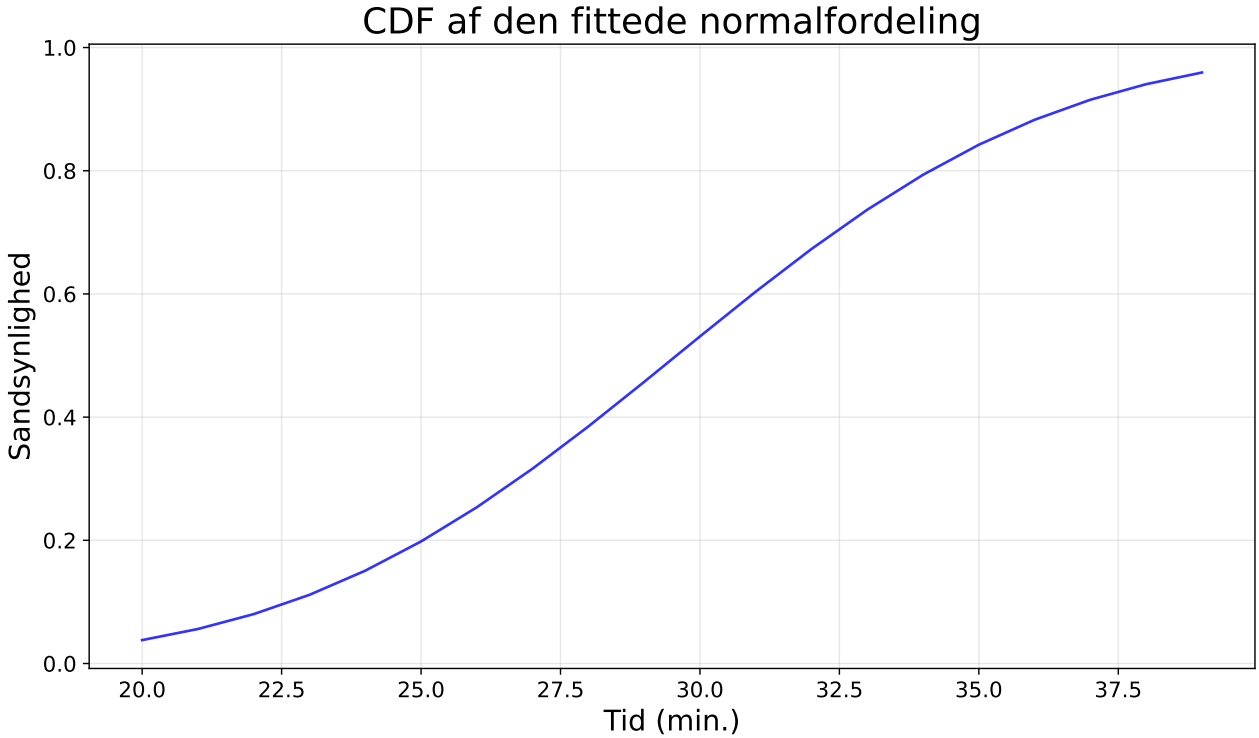
4 ☐ Middelværdien er strengt positivt.

5 ☐ Variansen er strengt positiv.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 5

Den studerende fitter en normalfordeling til sine målte cykeltider og bruger den som model for fremtidige cykeltider. Figuren nedenfor viser fordelingsfunktionen (cdf) for hendes model.



Spørgsmål II.3 (5)

Ifølge modellen, hvad er sandsynligheden for, at hun ankommer før hendes forelæsning starter, hvis hun cykler 35 minutter før den starter?

1 ☐ Omkring 15%.

2 ☐ Omkring 45%.

3 ☐ Omkring 65%.

4 ☐ Omkring 85%.

5 ☐ Omkring 99%.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 6

Opgave III

Lad X og Y være uafhængige stokastiske variable. Forventningsværdierne er $\mathbb{E}[X] = 5$ og $\mathbb{E}[Y] = -3$, mens varianserne er $\mathbb{V}[X] = 9$ og $\mathbb{V}[Y] = 16$.

Spørgsmål III.1 (6)

Hvad er kovariansen mellem X og Y?

1 ☐ -25

2 ☐ -5

3 ☐ 0

4 ☐ 5

5 ☐ 25

6 ☐ Ved ikke / Intet svar

Spørgsmål III.2 (7)

Lad $Z = -4X + 3Y$. Hvad er forventningsværdien af Z?

1 ☐ -29

2 ☐ -27

3 ☐ 0

4 ☐ 27

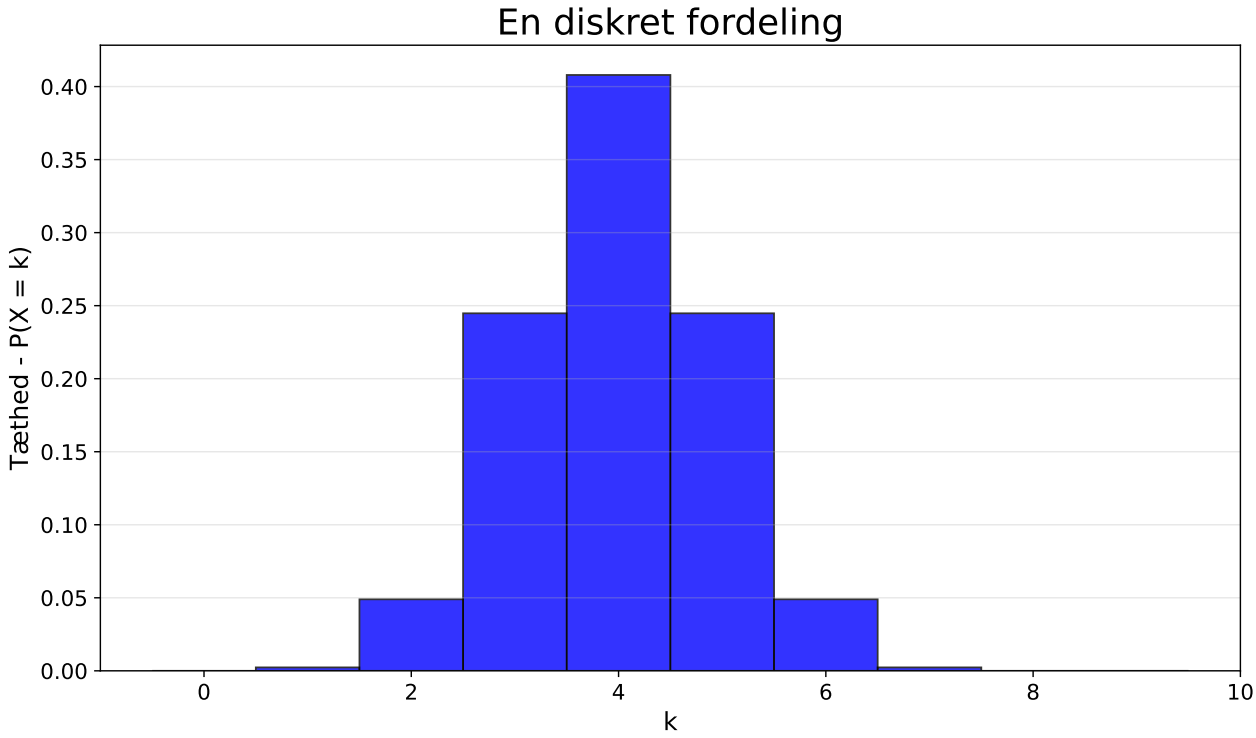
5 ☐ 29

6 ☐ Ved ikke / Intet svar

Fortsæt på side 7

Opgave IV

Betragt den følgende diskrete fordeling på de ikke-negative hele tal repræsenteret ved dens tæthedsfunktion (pdf), og lad X følge denne fordeling.



Spørgsmål IV.1 (8)

Hvilken af følgende muligheder angiver fordelingen korrekt?

1 ☐ Det er en binomialfordeling med $n = 10$ og $p = 0.5$.

2 ☐ Det er en binomialfordeling med $n = 10$ og $p = 0.9$.

3 ☐ Det er en hypergeometrisk fordeling med $n = 7$, $a = 8$ og $N = 14$.

4 ☐ Det er en Poissonfordeling med $\lambda = 4$.

5 ☐ Det er en Poissonfordeling med $\lambda = 10$.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 8

Spørgsmål IV.2 (9)

Hvad er $\mathbb{P}(X < 5)$?

1 ☐ Cirka 0.25

2 ☐ Cirka 0.30

3 ☐ Cirka 0.40

4 ☐ Cirka 0.70

5 ☐ Cirka 0.95

6 ☐ Ved ikke / Intet svar

Spørgsmål IV.3 (10)

Hvad er variansen af X?

1 ☐ Cirka 0.2

2 ☐ Cirka 0.7

3 ☐ Cirka 0.9

4 ☐ Cirka 4.2

5 ☐ Cirka 17.4

6 ☐ Ved ikke / Intet svar

Fortsæt på side 9

Opgave V

I et studie af strafudmåling indsamlede forskere data om fængselsstraffe (målt i år) idømt af forskellige dommere for sammenlignelige forbrydelser. Lad Y_{ij} betegne straf j givet af dommer i. Dataen modelleres som $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, hvor μ er den samlede gennemsnitlige straf på tværs af alle dommere, α_i er afvigelsen mellem den gennemsnitlige straf for dommer i og den samlede gennemsnitlige straf, og ε_{ij} er et stokastisk fejlled. Fejlleddene antages at være uafhængige og ensfordelte med $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, og alle straffe betragtes som uafhængige observationer.

Spørgsmål V.1 (11)

Hvilket udsagn er korrekt baseret på modellen beskrevet ovenfor?

1 ☐ Modellen tillader, at fejlleddene følger en vilkårlig kontinuert fordeling med middelværdi nul.

2 ☐ Modellen tillader, at dommere har forskellige gennemsnitlige straffe og forskellige varianser af straffene.

3 ☐ Modellen tillader, at dommere har forskellige gennemsnitlige straffe, men ikke forskellige varianser af straffene.

4 ☐ Modellen tillader, at dommere har forskellige varianser af straffene, men ikke forskellige gennemsnitlige straffe.

5 ☐ Modellen tillader hverken, at dommere har forskellige gennemsnitlige straffe eller forskellige varianser af straffene.

6 ☐ Ved ikke / Intet svar

Spørgsmål V.2 (12)

Forskerne forventer, at de forskellige dommeres gennemsnitlige straffe varierer meget, men at hver dommer giver meget konsistente straffe (dvs. at hver dommer har en tendens til at give ensartede straffe på tværs af sager). Hvis modelantagelserne er opfyldt, hvilket resultat vil da understøtte forskernes forventninger?

1 ☐ En lille værdi af SS(dommere) og en lille værdi af SSE.

2 ☐ En lille værdi af SS(dommere) og en stor værdi af SSE.

3 ☐ En stor værdi af SS(dommere) og en lille værdi af SSE.

4 ☐ En stor værdi af SS(dommere) og en stor værdi af SSE.

5 ☐ Intet resultat kan understøtte forskernes forventninger, hvis modelantagelserne er opfyldt.

6 ☐ Ved ikke / Intet svar

Modellen fittes til dataen inkluderet i studiet, som er vist i tabellen nedenfor:

Dommer A	Dommer B	Dommer C	Dommer D	Dommer E
4.3	7.8	10.1	4.7	13.2
14.1	8.4	10.5	10.3	17.9
6.9	10.2	10.4	8.1	12.4
20.5	11.3	9.9	7.6	12.5
2.2	9.3	10.1		10.1
10.8		11.0		19.8
		10.2		26.5
		12.6		11.1

Spørgsmål V.3 (13)

Hvilket udsagn om modelantagelserne er korrekt baseret på dataen? (Hint: prøv at visualisere dataen som boksplot.)

1 ☐ Modelantagelserne er opfyldt.

2 ☐ Modelantagelserne er ikke opfyldt, fordi dataen ikke er normalfordelt.

3 ☐ Modelantagelserne er ikke opfyldt, fordi studiet inkluderer færre end fem straffe fra en af dommerne.

4 ☐ Modelantagelserne er ikke opfyldt, fordi dommerne har meget forskellige varianser af straffene.

5 ☐ Modelantagelserne er ikke opfyldt, fordi dommerne har bidraget med forskellige antal straffe.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 11

Opgave VI

Betragt en simpel lineær regressionsmodel:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, ..., n,$$

under de sædvanlige uafhængighedsantagelser.

Spørgsmål VI.1 (14)

Hvilket af følgende udsagn er forkert?

1 ☐ Modellen har tre parametre.

2 ☐ De stokastiske variable ε_i kaldes fejl.

3 ☐ Normalitetsantagelsen tjekkes med et normal QQ-plot af residualerne.

4 ☐ 95%-konfidensintervallet for hældningskoefficienten kan være bredere end 95%-konfidensintervallet for skæringspunktet med y-aksen.

5 ☐ 95%-konfidensintervallet for den forventede respons ved et givet x_0 kan være bredere end 95%-prædiktionsintervallet for en enkelt respons ved x_0.

6 ☐ Ved ikke / Intet svar

Spørgsmål VI.2 (15)

Betragt en situation, hvor modellen er fittet til et datasæt med $n = 4$ observationer.

ID (i)	1	2	3	4
Observation (y_i)	4	6	8	10
Modelforudsigelse ($\hat{y}_i$)	4	5	10	9
Residual (e_i)	0	1	-2	1

Hvad er residualkvadratafvigelsessummen (RSS) for modellen?

1 ☐ Residualkvadratafvigelsessummen er -2.

2 ☐ Residualkvadratafvigelsessummen er 0.

3 ☐ Residualkvadratafvigelsessummen er 1.

4 ☐ Residualkvadratafvigelsessummen er 4.

5 ☐ Residualkvadratafvigelsessummen er 6.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 12

Opgave VII

Hver uge trækker det danske statslotteri 7 tilfældige tal uden tilbagelægning fra tallene 1 til 36 (begge inkluderet). En lottokupon består af 7 forskellige tal, og for at vinde hovedpræmien skal alle 7 tal på en kupon matche de trukne tal. (Bemærk: Rækkefølgen på tallene har ikke nogen betydning).

Spørgsmål VII.1 (16)

Hvad er chancen (sandsynligheden) for at vinde hovedpræmien med en enkelt kupon i lotteriet?

1 ☐ 1 ud af 78.364.164.096

2 ☐ 1 ud af 42.072.307.200

3 ☐ 1 ud af 2.176.782.336

4 ☐ 1 ud af 8.347.680

5 ☐ 1 ud af 5.040

6 ☐ Ved ikke / Intet svar

Opgave VIII

Betragt to uafhængige stokastiske variable $X_1 \sim \mathcal{N}(0,1)$ og $X_2 \sim \mathcal{N}(0,1)$. En teoretisk model foreskriver, at $Y = f(X_1, X_2) = X_1^2 + X_2^2$.

Spørgsmål VIII.1 (17)

Hvad er variansen af Y ifølge den ikke-lineære fejlophobningslov?

1 ☐ Variansen af Y er 0.

2 ☐ Variansen af Y er 2.

3 ☐ Variansen af Y er 4.

4 ☐ Variansen af Y er 8.

5 ☐ Variansen af Y er 16.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 13

Opgave IX

Et universitet undersøger effekterne af AI platforme på studerendes eksamenspræstationer.

Spørgsmål IX.1 (18)

I et stort kursus har en stikprøve på 15 studerende, som har brugt den anbefalede platform, en gennemsnitseksamensscore på 82 point med en standardafvigelse på 6.2 point. Under en normalfordelingsantagelse, hvad er 95%-konfidensintervallet for middeleksamensscoren for de studerende, som bruger den anbefalede platform? (Man kan bruge, at $t_{0.975}(14) = 2.145$.)

1 ☐ $(78.6, 85.4)$

2 ☐ $(79.7, 84.3)$

3 ☐ $(80.0, 84.0)$

4 ☐ $(81.0, 83.0)$

5 ☐ $(77.5, 86.5)$

6 ☐ Ved ikke / Intet svar

Spørgsmål IX.2 (19)

I et andet stort kursus har en stikprøve på 12 studerende, som har brugt en alternativ platform, en gennemsnitsscore på 79 point (standardafvigelsen var 5 point) til eksamen. Universitetet anvender en tosidet t-test til at sammenligne middelscoren for de studerende, der bruger den alternative platform, med den historiske gennemsnitsscore for kurset på 75 point. Testen giver en p-værdi på 0.018 og et 95%-konfidensinterval på $(75.82, 82.18)$. Hvad er den korrekte konklusion på testen ved 5% signifikansniveauet?

1 ☐ Middelscoren for studerende, der bruger den alternative platform, er signifikant lavere end den historiske gennemsnitsscore for kurset.

2 ☐ Middelscoren for studerende, der bruger den alternative platform, er ikke signifikant højere end den historiske gennemsnitsscore for kurset.

3 ☐ Middelscoren for studerende, der bruger den alternative platform, er ikke signifikant forskellig fra den historiske gennemsnitsscore for kurset.

4 ☐ Middelscoren for studerende, der bruger den alternative platform, er signifikant forskellig fra den historiske gennemsnitsscore for kurset.

5 ☐ Der er 1.8% sandsynlighed for, at middelscoren for studerende, der bruger den alternative platform, er signifikant højere end den historiske gennemsnitsscore for kurset.

6 ☐ Ved ikke / Intet svar

Universitetet har udviklet en AI tutor, designet til at hjælpe studerende med at studere og forberede sig til forelæsninger. For at undersøge om de studerende bruger mindre tid på at forberede sig til forelæsninger, når de bruger AI tutoren sammenlignet med traditionelle forberedelsesmetoder, har universitetet fulgt to grupper af studerende, der fulgte de samme forelæsninger: en lille gruppe på 50 studerende, der brugte AI tutoren, og en større gruppe på 300 studerende, der brugte traditionelle forberedelsesmetoder. For hver forelæsning målte universitet den gennemsnitlige forberedelsestid for begge grupper:

Forelæsning	1	2	3	4	$\cdots$	13
Gennemsnitlig forberedelsestid med AI tutor (min.)	30	45	80	110	$\cdots$	95
Gennemsnitlig forberedelsestid uden AI tutor (min.)	30	45	100	120	$\cdots$	100

Universitetet noterer, at den forventede forberedelsestid varierer meget mellem forelæsningerne (jf. tabellen), og at de ulige gruppestørrelser kan føre til forskellige gruppevarianser.

Spørgsmål IX.3 (20)

Hvilken af følgende muligheder repræsenterer den mest velegnede metode til at teste middelforskellen i forberedelsestid mellem de to grupper givet dette forsøgsdesign?

1 ☐ En Welch t-test for to uparrede stikprøver

2 ☐ En t-test for to uparrede stikprøver med sammenvejet varians

3 ☐ En parret t-test

4 ☐ En χ^2-test

5 ☐ En ensidet variansanalysetest

6 ☐ Ved ikke / Intet svar

Fortsæt på side 15

Spørgsmål IX.4 (21)

Universitetet beslutter at bruge en anden type test og vælger et signifikansniveau på 5%. For de givne parametre har testen en statistisk styrke på 90%. Givet at modelantagelserne er opfyldt, og at der er en reel forskel i middelforberedelsestiden mellem de to grupper, hvad er sandsynligheden for, at testen når frem til den korrekte konklusion?

1 ☐ 95%

2 ☐ 90%

3 ☐ 10%

4 ☐ 5%

5 ☐ 2.5%

6 ☐ Ved ikke / Intet svar

Spørgsmål IX.5 (22)

I den alternative test beregnes nogle korrigerede forberedelsestider (som er strengt positive), som testen kræver skal være tilnærmelsesvis normalfordelte. Et histogram viser imidlertid, at de korrigerede forberedelsestider er højreskæve. Hvilken af følgende transformationer bør ikke overvejes, hvis målet er at transformere de korrigerede forberedelsestider, så de bliver tilnærmelsesvis normalfordelte?

1 ☐ Kubikrodstransformationen, dvs. at transformere x til $x^{1/3}$.

2 ☐ Den eksponentielle transformation, dvs. at transformere x til $\exp(x)$.

3 ☐ Logaritmetransformationen, dvs. at transformere x til $\log(x)$.

4 ☐ Reciproktransformationen, dvs. at transformere x til $1/x$.

5 ☐ Kvadratrodstransformationen, dvs. at transformere x til $\sqrt{x}$.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 16

Opgave X

Et universitet vil undersøge sammenhængen mellem studerendes brug af digitale træningsteknologier og deres eksamensscorer. 240 studerende blev tilfældigt tildelt en af tre forskellige digitale træningsteknologier: Videobaseret Læring (VBL), Gamificeret Læringsplatform (GLP) og Interaktive Simuleringer (IS). Deres eksamensscorer blev inddelt i tre niveauer: Under Middel, Middel og Over Middel.

Eksamensscore	VBL	GLP	IS	Rækkesum
Under Middel	18	12	10	40
Middel	32	26	22	80
Over Middel	30	38	52	120
Søjlesum	80	76	84	240

Studerende med Middel eller Over Middel eksamensscorer betragtes som "succesfulde elever". Tabellen nedenfor med fraktiler fra standardnormalfordelingen skal bruges til at løse nogle af spørgsmålene i denne opgave.

Fraktil	$q_{0.90}$	$q_{0.95}$	$q_{0.975}$	$q_{0.99}$
Værdi	1.282	1.645	1.960	2.326

Spørgsmål X.1 (23)

Under nulhypotesen, at fordelingen af eksamensscorer er den samme på tværs af alle tre teknologier, hvad er så det forventede antal elever, som bruger Videobaseret Læring (VBL) med en Under Middel eksamensscore?

1 ☐ 12.90

2 ☐ 13.33

3 ☐ 14.00

4 ☐ 15.12

5 ☐ 18.50

6 ☐ Ved ikke / Intet svar

Fortsæt på side 17

Spørgsmål X.2 (24)

Hvad er 95%-konfidensintervallet for den overordnede andel af succesfulde elever baseret på dataen?

1 ☐ $[0.786, 0.880]$

2 ☐ $[0.701, 0.812]$

3 ☐ $[0.692, 0.833]$

4 ☐ $[0.688, 0.784]$

5 ☐ $[0.622, 0.754]$

6 ☐ Ved ikke / Intet svar

Spørgsmål X.3 (25)

Er der en signifikant forskel i andelen af succesfulde elever mellem VBL og IS grupperne på et 5% signifikansniveau? (Hint: beregn teststørrelsen under $H_0 : p_{VBL} - p_{IS} = 0$, hvor p_{VBL} er andelen af studerende i VBL-gruppen, der betragtes som succesfulde elever, medens p_{IS} er den tilsvarende andel inden for IS-gruppen.)

1 ☐ Der er ingen signifikant forskel, da den observerede teststørrelse $z_{\text{obs}} = -1.56 > -1.96$.

2 ☐ Der er ingen signifikant forskel, da den observerede teststørrelse $|z_{\text{obs}}| = |-2.12| > 1.96$.

3 ☐ Der er ingen signifikant forskel, da den observerede teststørrelse $z_{\text{obs}} = -1.80 > -1.96$.

4 ☐ Der er en signifikant forskel, da den observerede teststørrelse $z_{\text{obs}} = -2.27 < -1.96$.

5 ☐ Der er en signifikant forskel, da den observerede teststørrelse $z_{\text{obs}} = -2.13 < -1.96$.

6 ☐ Ved ikke / Intet svar

Fortsæt på side 18

Spørgsmål X.4 (26)

Hvad er 95%-konfidensintervallet for forskellen i andelen af Over Middel eksamensscorer mellem IS og GLP grupperne?

1 ☐ $[0.113, 0.279]$

2 ☐ $[0.105, 0.293]$

3 ☐ $[0.091, 0.286]$

4 ☐ $[0.082, 0.312]$

5 ☐ $[-0.034, 0.272]$

6 ☐ Ved ikke / Intet svar

Spørgsmål X.5 (27)

I den sædvanlige test af uafhængighed mellem typen af digital træningsteknologi og de studerendes eksamensscore, hvilken fordeling følger teststørrelsen under nulhypotesen om uafhængighed?

1 ☐ En F-fordeling med 3 og 3 frihedsgrader

2 ☐ En F-fordeling med 2 og 2 frihedsgrader

3 ☐ En χ^2-fordeling med 9 frihedsgrader

4 ☐ En χ^2-fordeling med 6 frihedsgrader

5 ☐ En χ^2-fordeling med 4 frihedsgrader

6 ☐ Ved ikke / Intet svar

Fortsæt på side 19

Opgave XI

I klassisk mekanik beskrives bevægelsen af et objekt under konstant acceleration af ligningen

$$v(t) = v_0 + at,$$

hvor $v(t)$ er hastigheden til tiden t, v_0 er starthastigheden (hastigheden til tiden $t = 0$) og a er den konstante acceleration.

I et eksperiment måles et sæt af hastigheder til forskellige tidspunkter for et objekt, der bevæger sig under konstant acceleration. Grundet måleusikkerheder er de observerede hastigheder dog forskellige fra den teoretiske model. For at håndtere dette, indføres et fejlled, hvilket fører til den lineære regressionsmodel:

$$v_i = v_0 + at_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

hvor fejlene antages uafhængige.

Det følgende output er givet:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      v   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
No. Observations:                 100   F-statistic:                3.071e+05
Covariance Type:            nonrobust   Prob (F-statistic):          3.89e-173
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.4695      0.421     10.629      0.000       3.635       5.304
t              4.0064      0.007    554.190      0.000       3.992       4.021
==============================================================================
```

Spørgsmål XI.1 (28)

Hvad er de estimerede parameterværdier?

1 ☐ $\hat{v}_0 = 554.190$ og $\hat{a} = 10.629$

2 ☐ $\hat{v}_0 = 10.629$ og $\hat{a} = 554.190$

3 ☐ $\hat{v}_0 = 0.421$ og $\hat{a} = 0.007$

4 ☐ $\hat{v}_0 = 4.470$ og $\hat{a} = 4.006$

5 ☐ $\hat{v}_0 = 4.006$ og $\hat{a} = 4.470$

6 ☐ Ved ikke / Intet svar

Fortsæt på side 20

De følgende tal kan bruges i det næste spørgsmål:

```
print(np.round(stats.t.ppf(q = [0.995,0.99,0.975,0.95,0.90], df = 100-2),4))
```

```
[2.6269 2.365  1.9845 1.6606 1.2902]
```

Spørgsmål XI.2 (29)

Hvad er 99%-konfidensintervallet for v_0?

1 ☐ $[3.364, 5.575]$

2 ☐ $[3.474, 5.465]$

3 ☐ $[3.635, 5.304]$

4 ☐ $[3.990, 4.023]$

5 ☐ $[3.992, 4.021]$

6 ☐ Ved ikke / Intet svar

Spørgsmål XI.3 (30)

Betragt nulhypotesen $\mathcal{H}_0 : v_0 = 5$ og den tosidede modhypotese $\mathcal{H}_1 : v_0 \neq 5$. Hvad er den observerede teststørrelse (t_{obs}) under nulhypotesen?

1 ☐ -2.360

2 ☐ -1.260

3 ☐ 1.260

4 ☐ 10.629

5 ☐ 554.190

6 ☐ Ved ikke / Intet svar

Opgavesættet er slut.

This exam paper is available in both Danish and English. The English version appears after the Danish version.

Exam paper:

Written examination: 28.05.2026

Course name and number: **02323 Introduction to Statistics**

Duration: 4 hours

Aids allowed: All printed materials and a pocket calculator of type TI30XS or TI30XB

Final answers must be submitted by completing the separate "Answer Sheet".

This examination consists of 30 multiple-choice questions distributed across 11 exercises.
Only the "Answer Sheet" must be submitted; do not hand in the exam paper itself.

Multiple choice questions: *Note that in each question, one and only one of the answer options is correct. Furthermore, not all suggested answer options are necessarily meaningful. When performing calculations, always round your result to the number of decimal places used in the answer options before selecting your answer.*

The use of Python code in this exam: *This examination includes Python code. Note that the following libraries and abbreviations are used:*

```
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
```

Continue on page 3

Exercise I

This exercise is concerned with basic Python functionality.

Question I.1 (1)

What does the command `np.arange(10)` (`np` refers to the NumPy package) output?

1 ☐ The array `[1 2 3 4 5 6 7 8 9 10]`.

2 ☐ The array `[1 2 3 4 5 6 7 8 9]`.

3 ☐ The array `[0 1 2 3 4 5 6 7 8 9]`.

4 ☐ The array `[0 1 2 3 4 5 6 7 8 9 10]`.

5 ☐ An array with ten random numbers between zero and one.

6 ☐ Don't know / No answer

Question I.2 (2)

Which function in the `scipy.stats`-package outputs random numbers?

1 ☐ The `cdf`-function

2 ☐ The `pdf`-function

3 ☐ The `pmf`-function

4 ☐ The `ppf`-function

5 ☐ The `rvs`-function

6 ☐ Don't know / No answer

Continue on page 4

Exercise II

A DTU student wants to ensure that she is never late for an early morning lecture. Over the course of two weeks, she therefore measures the time it takes her to bike to campus. She uses Python and saves her measurements in the vector `x`.

Question II.1 (3)

How can she compute the sample variance of her biking times? (`np` refers to the NumPy package.)

1 ☐ `np.mean(x)`

2 ☐ `np.std(x)`

3 ☐ `np.std(x, ddof = 1)`

4 ☐ `np.var(x)`

5 ☐ `np.var(x, ddof = 1)`

6 ☐ Don't know / No answer

Question II.2 (4)

The student wants to model her transportation times using a normal distribution. Which of the following statements is not true about a normal distribution?

1 ☐ It has two parameters: the mean μ and the variance σ^2.

2 ☐ It is symmetric about the mean.

3 ☐ The mean is equal to the median.

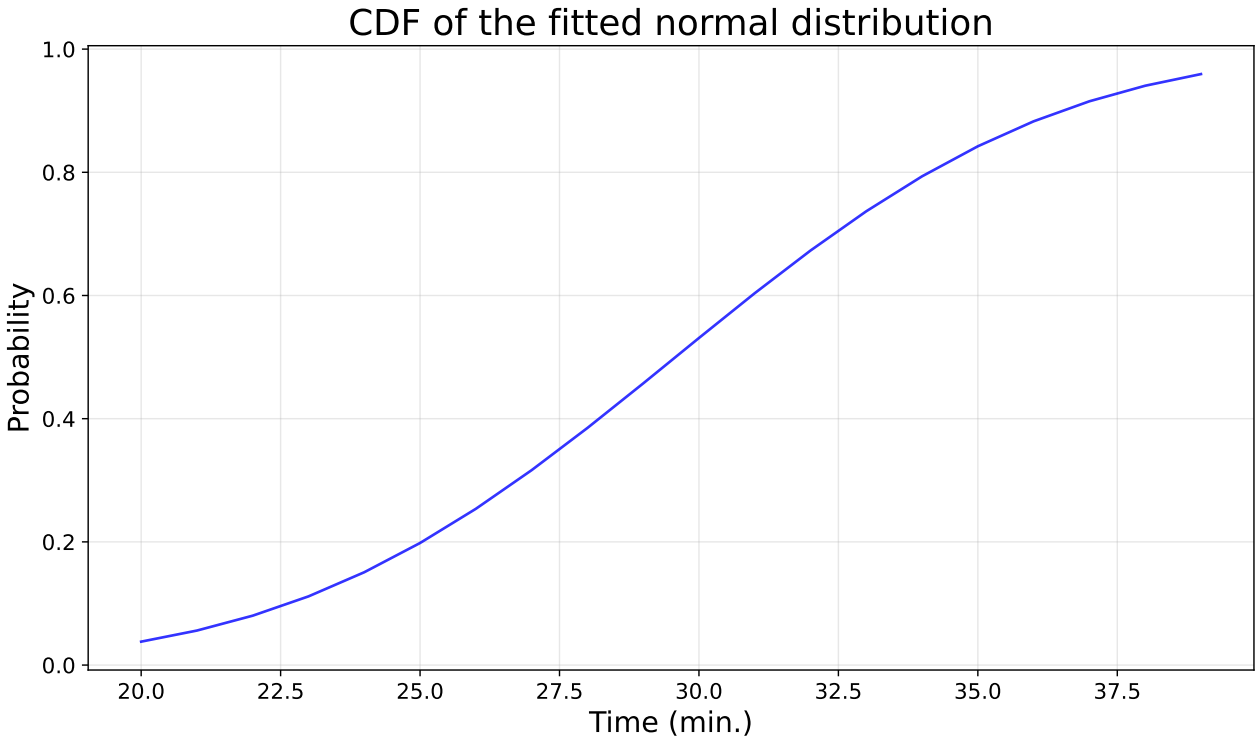
4 ☐ The mean is strictly positive.

5 ☐ The variance is strictly positive.

6 ☐ Don't know / No answer

Continue on page 5

The student fits a normal distribution to her measured biking times and uses this as a model for future biking times. The figure below shows the cumulative distribution function (cdf) of her model.



Question II.3 (5)

According to the model, what is the probability that she arrives in time for her lecture if she leaves 35 minutes before it starts?

1 ☐ Approximately 15%.

2 ☐ Approximately 45%.

3 ☐ Approximately 65%.

4 ☐ Approximately 85%.

5 ☐ Approximately 99%.

6 ☐ Don't know / No answer

Continue on page 6

Exercise III

Let X and Y be independent random variables. The expected values are $\mathbb{E}[X] = 5$ and $\mathbb{E}[Y] = -3$, while the variances are $\mathbb{V}[X] = 9$ and $\mathbb{V}[Y] = 16$.

Question III.1 (6)

What is the covariance between X and Y?

1 ☐ -25

2 ☐ -5

3 ☐ 0

4 ☐ 5

5 ☐ 25

6 ☐ Don't know / No answer

Question III.2 (7)

Let $Z = -4X + 3Y$. What is the expectation of Z?

1 ☐ -29

2 ☐ -27

3 ☐ 0

4 ☐ 27

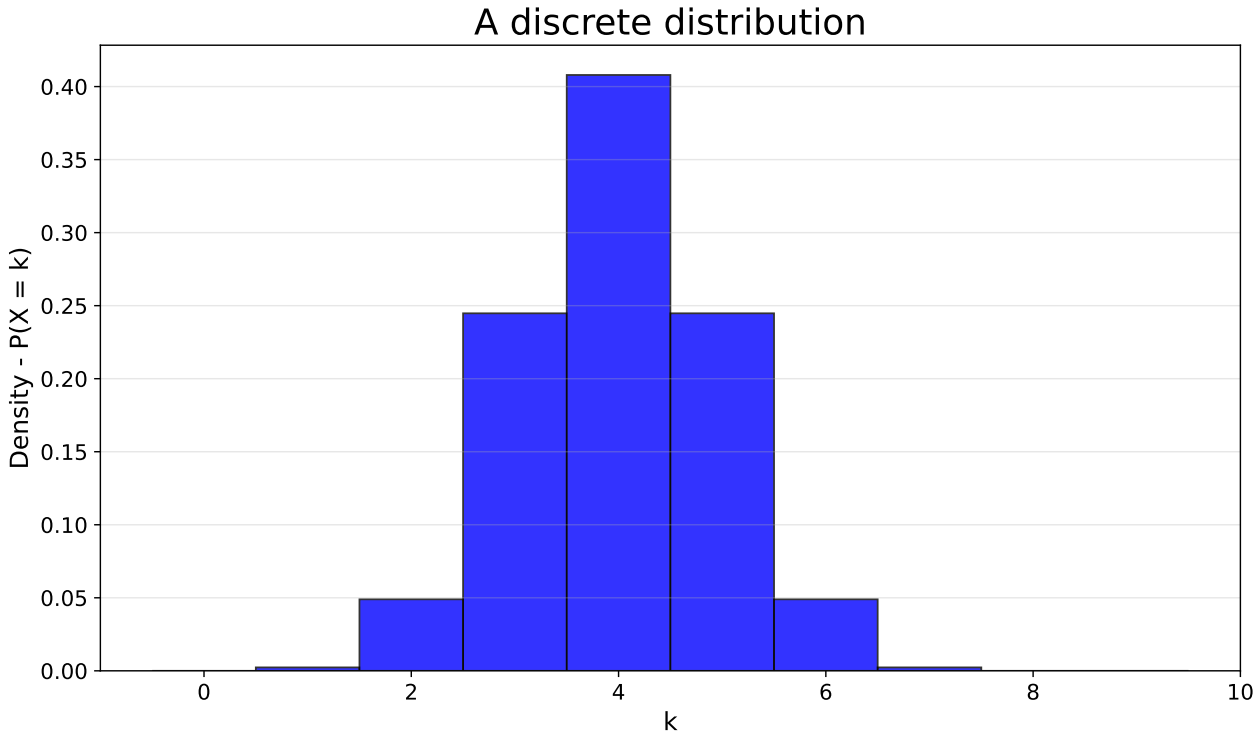
5 ☐ 29

6 ☐ Don't know / No answer

Continue on page 7

Exercise IV

Consider the following discrete distribution on the non-negative integers, represented by its probability density function (pdf), and let X follow this distribution.



Question IV.1 (8)

Which of the following options correctly declares the distribution?

1 ☐ It is a binomial distribution with $n = 10$ and $p = 0.5$.

2 ☐ It is a binomial distribution with $n = 10$ and $p = 0.9$.

3 ☐ It is a hypergeometric distribution with $n = 7$, $a = 8$, and $N = 14$.

4 ☐ It is a Poisson distribution with $\lambda = 4$.

5 ☐ It is a Poisson distribution with $\lambda = 10$.

6 ☐ Don't know / No answer

Continue on page 8

Question IV.2 (9)

What is $\mathbb{P}(X < 5)$?

1 ☐ Approximately 0.25

2 ☐ Approximately 0.30

3 ☐ Approximately 0.40

4 ☐ Approximately 0.70

5 ☐ Approximately 0.95

6 ☐ Don't know / No answer

Question IV.3 (10)

What is the variance of X?

1 ☐ Approximately 0.2

2 ☐ Approximately 0.7

3 ☐ Approximately 0.9

4 ☐ Approximately 4.2

5 ☐ Approximately 17.4

6 ☐ Don't know / No answer

Continue on page 9

Exercise V

In a study of sentencing practices, researchers collected data on prison sentences (measured in years) imposed by different judges for comparable criminal offenses. Let Y_{ij} denote sentence j given by judge i. The data are modeled as $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where μ is the overall average sentence across all judges, α_i is the difference between the average sentence of judge i and the overall average, and ε_{ij} is a random error term. The error terms are assumed to be independent and identically distributed with $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, and all sentences are treated as independent observations.

Question V.1 (11)

Which statement is correct based on the model described above?

1 ☐ The model allows the error terms to follow any continuous distribution with mean zero.

2 ☐ The model allows judges to have different mean sentences and different sentence variances.

3 ☐ The model allows judges to have different mean sentences but not different sentence variances.

4 ☐ The model allows judges to have different sentence variances but not different mean sentences.

5 ☐ The model does not allow judges to have different sentence variances nor different mean sentences.

6 ☐ Don't know / No answer

Question V.2 (12)

The researchers expect the judges' average sentences to differ substantially, but each judge to give very consistent sentences (each judge tends to give very similar sentences across cases). Assuming the model assumptions are satisfied, which findings would support the researchers' expectations?

1 ☐ A small value of SS(judges) and a small value of SSE.

2 ☐ A small value of SS(judges) but a large value of SSE.

3 ☐ A large value of SS(judges) but a small value of SSE.

4 ☐ A large value of SS(judges) and a large value of SSE.

5 ☐ No findings can support the researchers' expectations if the model assumptions are satisfied.

6 ☐ Don't know / No answer

The model is fitted to the data included in the study, which are shown in the table below:

Judge A	Judge B	Judge C	Judge D	Judge E
4.3	7.8	10.1	4.7	13.2
14.1	8.4	10.5	10.3	17.9
6.9	10.2	10.4	8.1	12.4
20.5	11.3	9.9	7.6	12.5
2.2	9.3	10.1		10.1
10.8		11.0		19.8
		10.2		26.5
		12.6		11.1

Question V.3 (13)

Which statement about the model assumptions is correct based on the data? (Hint: Try to visualise the data as boxplots.)

1 ☐ The model assumptions are satisfied.

2 ☐ The model assumptions are not satisfied because the data is not normally distributed.

3 ☐ The model assumptions are not satisfied because the study includes fewer than five sentences from one of the judges.

4 ☐ The model assumptions are not satisfied because the judges exhibit considerably different sentence variances.

5 ☐ The model assumptions are not satisfied because the judges have contributed different numbers of sentences.

6 ☐ Don't know / No answer

Continue on page 11

Exercise VI

Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, ..., n,$$

under the usual independence assumptions.

Question VI.1 (14)

Which one of the following statements is false?

1 ☐ The model has three parameters.

2 ☐ The random variables ε_i are called errors.

3 ☐ The normality assumption is checked with a normal QQ-plot of the residuals.

4 ☐ The 95% confidence interval for the slope can be wider than the 95% confidence interval for the intercept.

5 ☐ The 95% confidence interval for the mean response at a given x_0 can be wider than the 95% prediction interval for an individual response at x_0.

6 ☐ Don't know / No answer

Question VI.2 (15)

Consider a case where the model has been fitted to a dataset with $n = 4$ observations:

ID (i)	1	2	3	4
Observation (y_i)	4	6	8	10
Model prediction ($\hat{y}_i$)	4	5	10	9
Residual (e_i)	0	1	-2	1

What is the residual sum of squares (RSS) for the model?

1 ☐ The residual sum of squares is -2.

2 ☐ The residual sum of squares is 0.

3 ☐ The residual sum of squares is 1.

4 ☐ The residual sum of squares is 4.

5 ☐ The residual sum of squares is 6.

6 ☐ Don't know / No answer

Continue on page 12

Exercise VII

Each week, the Danish state lottery draws 7 numbers at random without replacement from the numbers 1 to 36 (both included). A lottery ticket consists of 7 distinct numbers, and to win the jackpot, all 7 numbers on your ticket must match the drawn numbers. (Note: The order of the numbers on a ticket does not matter.)

Question VII.1 (16)

What is the chance (probability) of winning the jackpot with a single ticket in the lottery?

1 ☐ 1 in 78.364.164.096

2 ☐ 1 in 42.072.307.200

3 ☐ 1 in 2.176.782.336

4 ☐ 1 in 8.347.680

5 ☐ 1 in 5.040

6 ☐ Don't know / No answer

Exercise VIII

Consider two independent random variables $X_1 \sim \mathcal{N}(0,1)$ and $X_2 \sim \mathcal{N}(0,1)$. A theoretical model posits that $Y = f(X_1, X_2) = X_1^2 + X_2^2$.

Question VIII.1 (17)

According to the non-linear error propagation rule, what is the variance of Y?

1 ☐ The variance of Y is 0.

2 ☐ The variance of Y is 2.

3 ☐ The variance of Y is 4.

4 ☐ The variance of Y is 8.

5 ☐ The variance of Y is 16.

6 ☐ Don't know / No answer

Continue on page 13

Exercise IX

A university investigates the effects of generative AI platforms on academic performance.

Question IX.1 (18)

In a large class, a sample of 15 students who used the recommended platform had an average exam score of 82 points with a standard deviation of 6.2 points. Assuming normality of the exam scores, what is the 95% confidence interval for the mean exam score of students using the recommended platform? (You can use that $t_{0.975}(14) = 2.145$.)

1 ☐ $(78.6, 85.4)$

2 ☐ $(79.7, 84.3)$

3 ☐ $(80.0, 84.0)$

4 ☐ $(81.0, 83.0)$

5 ☐ $(77.5, 86.5)$

6 ☐ Don't know / No answer

Question IX.2 (19)

In another large class, a sample of 12 students who used an alternative platform had a mean score of 79 points (SD was 5 points) on the exam. The university applies a two-sided t-test to compare the mean score of the students using the alternative platform with the historical class mean score of 75 points. The test yields a p-value of 0.018 and a 95% confidence interval of $(75.82, 82.18)$. What is the appropriate conclusion of the test at the 5% significance level?

1 ☐ The mean score of students using the alternative platform is significantly lower than the historical class mean score.

2 ☐ The mean score of students using the alternative platform is not significantly higher than the historical class mean score.

3 ☐ The mean score of students using the alternative platform is not significantly different from the historical class mean score.

4 ☐ The mean score of students using the alternative platform is significantly different from the historical class mean score.

5 ☐ There is a 1.8% probability that the mean score of students using the alternative platform is significantly higher than the historical class mean score.

6 ☐ Don't know / No answer

The university has developed an AI tutor designed to help students study and prepare for lectures. To examine whether students spend less time on lecture preparation when using the AI tutor compared with ordinary preparation methods, the university has followed two groups of students attending the same lectures: a small group of 50 students using the AI tutor and a larger group of 300 students using traditional preparation methods. For each lecture, the university has recorded the average preparation time for both groups:

Lecture	1	2	3	4	$\cdots$	13
Average preparation time with AI tutor (min.)	30	45	80	110	$\cdots$	95
Average preparation time without AI tutor (min.)	30	45	100	120	$\cdots$	100

The university notes that the expected preparation time varies substantially across lectures (as evident from the table) and that the unequal group sizes may lead to different group variances.

Question IX.3 (20)

Considering the experimental design, which one of the following options represents the most appropriate test for assessing a mean difference in preparation time between the two groups?

1 ☐ A Welch two-sample t-test

2 ☐ A pooled two-sample t-test

3 ☐ A paired t-test

4 ☐ A χ^2-test

5 ☐ A one-way ANOVA test

6 ☐ Don't know / No answer

Continue on page 15

Question IX.4 (21)

The university eventually decides to use a different type of test and selects a significance level of 5%. For the given parameters, the test has a statistical power of 90%. Assuming that the model assumptions are satisfied and that there truly is a difference in mean preparation time between the two groups, what is the probability that the test reaches the correct conclusion?

1 ☐ 95%

2 ☐ 90%

3 ☐ 10%

4 ☐ 5%

5 ☐ 2.5%

6 ☐ Don't know / No answer

Question IX.5 (22)

In the alternative test, you calculate adjusted preparation times (these remain strictly positive), which the test requires to be approximately normally distributed. However, a histogram of the adjusted preparation times reveals that they are right-skewed. Which of the following transformations should not be considered for making the adjusted preparation times approximately normally distributed?

1 ☐ The cube root transformation i.e., transform x into $x^{1/3}$.

2 ☐ The exponential transformation i.e., transform x into $\exp(x)$.

3 ☐ The logarithmic transformation i.e., transform x into $\log(x)$.

4 ☐ The reciprocal transformation i.e., transform x into $1/x$.

5 ☐ The square root transformation i.e., transform x into $\sqrt{x}$.

6 ☐ Don't know / No answer

Continue on page 16

Exercise X

A university wants to investigate whether the type of digital training technology affects students' exam scores. 240 students were randomly assigned to one of three digital training technologies: Video-based Learning (VBL), Gamified Learning Platform (GLP), and Interactive Simulations (IS). Their exam scores were categorized into three levels: Below Average, Average, and Above Average.

Exam score	VBL	GLP	IS	Row Total
Below Average	18	12	10	40
Average	32	26	22	80
Above Average	30	38	52	120
Column Total	80	76	84	240

Students with Average or Above Average exam scores are considered as "successful learners". The below table of quantiles from the standard normal distribution is needed to solve some of the problems in this exercise.

Quantile	$q_{0.90}$	$q_{0.95}$	$q_{0.975}$	$q_{0.99}$
Value	1.282	1.645	1.960	2.326

Question X.1 (23)

Under the null hypothesis that the distribution of exam scores is the same across all three technologies, what is the expected number of students using Video-based Learning (VBL) with a Below Average exam score?

1 ☐ 12.90

2 ☐ 13.33

3 ☐ 14.00

4 ☐ 15.12

5 ☐ 18.50

6 ☐ Don't know / No answer

Continue on page 17

Question X.2 (24)

What is the 95% confidence interval for the overall proportion of successful learners based on the data?

1 ☐ $[0.786, 0.880]$

2 ☐ $[0.701, 0.812]$

3 ☐ $[0.692, 0.833]$

4 ☐ $[0.688, 0.784]$

5 ☐ $[0.622, 0.754]$

6 ☐ Don't know / No answer

Question X.3 (25)

Is there a significant difference between the proportions of successful learners among students using VBL and students using IS at the 5% significance level? (Hint: Calculate the test statistic under $H_0 : p_{VBL} - p_{IS} = 0$, where p_{VBL} is the proportion of students using VBL who are considered successful learners, and p_{IS} is the corresponding proportion for students using IS.)

1 ☐ There is no significant difference, as the observed test statistic $z_{\text{obs}} = -1.56 > -1.96$.

2 ☐ There is no significant difference, as the observed test statistic $|z_{\text{obs}}| = |-2.12| > 1.96$.

3 ☐ There is no significant difference, as the observed test statistic $z_{\text{obs}} = -1.80 > -1.96$.

4 ☐ There is a significant difference, as the observed test statistic $z_{\text{obs}} = -2.27 < -1.96$.

5 ☐ There is a significant difference, as the observed test statistic $z_{\text{obs}} = -2.13 < -1.96$.

6 ☐ Don't know / No answer

Continue on page 18

Question X.4 (26)

What is the 95% confidence interval for the difference in the proportions of Above Average outcomes for students in the IS and GLP groups?

1 ☐ $[0.113, 0.279]$

2 ☐ $[0.105, 0.293]$

3 ☐ $[0.091, 0.286]$

4 ☐ $[0.082, 0.312]$

5 ☐ $[-0.034, 0.272]$

6 ☐ Don't know / No answer

Question X.5 (27)

In the usual test of independence between the type of digital training technology assigned to students and their exam score, which distribution does the test statistic follow under the null hypothesis of independence?

1 ☐ An F-distribution with 3 and 3 degrees of freedom

2 ☐ An F-distribution with 2 and 2 degrees of freedom

3 ☐ A χ^2-distribution with 9 degrees of freedom

4 ☐ A χ^2-distribution with 6 degrees of freedom

5 ☐ A χ^2-distribution with 4 degrees of freedom

6 ☐ Don't know / No answer

Continue on page 19

Exercise XI

In classical mechanics, the motion of an object under constant acceleration is governed by the equation

$$v(t) = v_0 + at,$$

where $v(t)$ is the velocity at time t, v_0 is the initial velocity (velocity at time $t = 0$), and a is the constant acceleration.

In an experiment, a set of velocity measurements is recorded at different time points for an object moving under constant acceleration. Due to measurement errors, the observed velocities deviate from the theoretical model. To account for this, we introduce an error term, leading to the linear regression model:

$$v_i = v_0 + at_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where the errors are assumed to be independent.

The following output is given:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      v   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
No. Observations:                 100   F-statistic:                 3.071e+05
Covariance Type:            nonrobust   Prob (F-statistic):          3.89e-173
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.4695      0.421     10.629      0.000       3.635       5.304
t              4.0064      0.007    554.190      0.000       3.992       4.021
==============================================================================
```

Question XI.1 (28)

What are the estimated parameter values?

1 ☐ $\hat{v}_0 = 554.190$ and $\hat{a} = 10.629$

2 ☐ $\hat{v}_0 = 10.629$ and $\hat{a} = 554.190$

3 ☐ $\hat{v}_0 = 0.421$ and $\hat{a} = 0.007$

4 ☐ $\hat{v}_0 = 4.470$ and $\hat{a} = 4.006$

5 ☐ $\hat{v}_0 = 4.006$ and $\hat{a} = 4.470$

6 ☐ Don't know / No answer

Continue on page 20

The following numbers can be used in the next question:

```
print(np.round(stats.t.ppf(q = [0.995,0.99,0.975,0.95,0.90], df = 100-2),4))
```

```
[2.6269 2.365  1.9845 1.6606 1.2902]
```

Question XI.2 (29)

What is the 99% confidence interval for v_0?

1 ☐ $[3.364, 5.575]$

2 ☐ $[3.474, 5.465]$

3 ☐ $[3.635, 5.304]$

4 ☐ $[3.990, 4.023]$

5 ☐ $[3.992, 4.021]$

6 ☐ Don't know / No answer

Question XI.3 (30)

Consider the null hypothesis $\mathcal{H}_0 : v_0 = 5$ against a two-sided alternative hypothesis $\mathcal{H}_1 : v_0 \neq 5$. What is the observed test statistic (t_{obs}) under the null hypothesis?

1 ☐ -2.360

2 ☐ -1.260

3 ☐ 1.260

4 ☐ 10.629

5 ☐ 554.190

6 ☐ Don't know / No answer

The exam is finished.